2025 • 29 • 3

Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

VAVILOV JOURNAL OF GENETICS AND BREEDING

Основан в 1997 г.
Периодичность 8 выпусков в год
doi 10.18699/vjgb-25-36

2025 • 29 • 3

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Publication frequency: 8 issues a year
doi 10.18699/vjgb-25-36

Founders

Siberian Branch of the Russian Academy of Sciences

Federal Research Center Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

СОДЕРЖАНИЕ • 2025 • 29 • 3

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII • VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2025 • 29 • 3

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):335-348

doi 10.18699/vjgb-25-37

From cytogenetics to proteogenomics: new horizons in the study of aneuploidies

K.S. Zadesenets [1, 2], N.B. Rubtsov [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
kira_z@bionet.nsc.ru

Abstract. Aneuploidy is defined as the loss or gain of a whole chromosome or its region. Even at early stages of development, it usually leads to fatal consequences, including developmental defects/abnormalities and death. For a long time, it was believed that the disruption of gene balance results in pronounced effects at both the cellular and organismal levels, adversely affecting organism formation. It has been shown that the gene imbalance resulting from aneuploidy leads to proteotoxic and metabolic stress within the cell, reduced cell proliferation, genomic instability, oxidative stress, etc. However, some organisms have exhibited tolerance to aneuploidies, which may even confer adaptive advantages, such as antibiotic resistance in pathogenic fungal strains. A significant factor likely lies in the complexity of the tissue and organ organization of specific species. Polyploid organisms are generally more tolerant of aneuploidy, particularly those that have recently undergone whole-genome duplication. This review places special emphasis on the examination of sex chromosome aneuploidies in humans. In addition to primary effects, or cis effects (changes in the quantity of the transcripts of genes located on the aneuploid chromosome), aneuploidy can induce secondary or trans effects (changes in the expression levels of genes located on other chromosomes). The results of recent studies have prompted a reevaluation of the impact of aneuploidy on the structural-functional organization of the genome, transcriptome, and proteome of both the cell and the entire organism. Despite the fact that, in the cases of aneuploidy, the expression levels for most genes correlate with their altered copy numbers in the cell, there have been instances of dosage compensation, where the transcript levels of genes located on the aneuploid chromosome remained unchanged. The review presents findings from recent studies focused on compensatory mechanisms of dosage compensation that modify gene product quantities at post-transcriptional and post-translational levels, alleviating the negative effects of aneuploidy on cellular homeostasis. It also discusses the influence of extrachromosomal elements on the spatial organization of the genome and the changes in gene expression patterns resulting from their presence. Additionally, the review specifically examines cases of segmental aneuploidy and changes in copy number variants (CNVs) in the genome. Not only the implications of their composition are considered, but also their localization within the chromosome and in various compartments of the interphase nucleus. Addressing these questions could significantly contribute to enhancing cytogenomic diagnostics and establishing a necessary database for accurate interpretation of identified cases of segmental aneuploidy and CNVs in the genome.

Key words: aneuploidy; chromosomal instability; genomic diversity; mosaicism; dosage compensation; differential gene expression; monoallelic expression; protein degradation; ubiquitin-proteasome system; architecture of interphase nucleus

For citation: Zadesenets K.S., Rubtsov N.B. From cytogenetics to proteogenomics: new horizons in the study of aneuploidies. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):335-348. doi 10.18699/vjgb-25-37

Funding. This study was supported by Russian Science Foundation (RSF) under grant project 24-24-00141.

От цитогенетики к протеогеномике: новые горизонты в исследовании анеуплоидий

К.С. Задесенец [1, 2], Н.Б. Рубцов [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
kira_z@bionet.nsc.ru

Аннотация. Анеуплоидией принято считать потерю или приобретение копии целой хромосомы или ее района. Уже на ранних стадиях развития она, как правило, приводит к фатальным последствиям, включая гибель организма и пороки/аномалии развития. Длительное время предполагалось, что именно нарушение баланса генов приводит к выраженным эффектам как на клеточном, так и на организменном уровне, негативно сказываясь на формировании организма. Было показано, что возникший вследствие анеуплоидии дисбаланс генов индуцирует протеотоксический и метаболический стресс в клетке, ее замедленную пролиферацию, нестабильность ее

генома, оксидативный стресс и пр. Однако для некоторых организмов была описана толерантность к анеуплоидиям, которая даже могла способствовать возникновению у них адаптивных преимуществ (например, резистентность к антибиотикам у патогенных штаммов грибов). Вероятно, значимым фактором является сложность тканевой и органной организации особей конкретного вида. К анеуплоидии преимущественно более толерантны полиплоидные организмы и виды, относительно недавно прошедшие полногеномную дупликацию. Особое внимание в обзоре уделено рассмотрению анеуплоидий половых хромосом человека. Помимо первичных эффектов или цис-эффектов (изменение количества транскриптов генов, находящихся на анеуплоидной хромосоме), анеуплоидия может вызывать вторичные или транс-эффекты (изменение уровня экспрессии генов, расположенных на других хромосомах). Результаты исследований последних лет заставили по-новому взглянуть на влияние анеуплоидии на структурно-функциональную организацию генома, транскриптом и протеом как клетки, так и целого организма. Несмотря на то что при анеуплоидии уровень экспрессии для большинства генов коррелирует с измененным числом копий генов в клетке, были описаны случаи дозовой компенсации, при которой уровень транскриптов генов, расположенных на анеуплоидной хромосоме, оставался неизменным. В обзоре приводятся результаты последних исследований, посвященных изучению компенсаторных механизмов дозовой компенсации изменения количества продуктов генов на посттранскрипционных и пострансляционных уровнях, снижающих негативный эффект анеуплоидии на гомеостаз клетки, а также влиянию экстрахромосом на пространственную организацию генома, изменению паттернов экспрессии генов вследствие ее наличия. Кроме того, отдельно обсуждаются варианты сегментных анеуплоидий и изменения числа копий участков генома. Рассмотрено не только значение их состава, но также его локализация в хромосоме и в разных компартментах интерфазного ядра. Решение поднятых вопросов может внести большой вклад в совершенствование цитогеномной диагностики и в создание необходимой базы данных для корректной интерпретации выявленных случаев и сегментной анеуплоидии, и варьирующих по числу копий участков генома.

Ключевые слова: анеуплоидия; хромосомная нестабильность; геномное разнообразие; мозаицизм; дозовая компенсация; дифференциальная экспрессия генов; моноаллельная экспрессия; деградация белков; убиквитин-протеасомная система; архитектоника интерфазного ядра

Introduction

The loss or gain of a copy of a whole chromosome or its part is referred to as aneuploidy (Tang, Amon, 2013). However, whole-genome sequencing and microarray-based comparative genomic hybridization have dramatically increased our understanding of the variation of the human genome; the view of variations in the copy number of genomic regions has become ambiguous (Pinkel et al., 1998). A high level of polymorphism was identified, and variations in the copy number of certain genomic regions (CNV, Copy Number Variant) quite often represented variants of normal genetic diversity. Unfortunately, the principle of describing the human genome based on a separate assembly of the haploid set often does not allow us to give an unambiguous answer to the question of what a particular case of CNV represents. It might be possible to distinguish between normal genomic diversity and its pathological variants more clearly in the future thanks to the generation of a human pangenome (Liao et al., 2023; Miga 2024). However, at present, it is frequently not possible to determine whether a polymorphism is a normal or pathological variant. CNV is the difference in the copy number of DNA segments found by comparing an individual genome to the reference human genome assembly, which is identified through cytogenomic analysis methods. In contrast to CNV, segmental aneuploidy is more frequently associated with a pathogenic effect, as it has a larger size and typically leads to chromosomal changes that can be detected using cytogenetic methods.

In this review, we will consider CNV as one of three types of segmental aneuploidy, differing in the size and structural organization of the corresponding region of the genome: 1) whole-chromosome aneuploidy – aneuploidy of the entire chromosome; 2) segmental aneuploidy – a change in the copy number of large regions of the genome, detected using classical cytogenetic methods; 3) CNV – a change in the copy number of a region of the genome of 1 thousand base pairs (Dürrbaum, Storchová, 2016). The distinction between these aneuploidy types can occasionally be very arbitrary, and they can also be categorized as distinct aneuploidy types at the same time. For instance, aneuploidy in a chromosomal region is caused by the presence of a small supernumerary marker chromosome in humans.

Aneuploidies on an entire chromosome are the result of errors in chromosome segregation. The main cause of these mistakes is the absence or insufficient cohesion of sister chromatids, defects in spindle formation in meiosis or mitosis (multipolar spindle, merotelic kinetochore attachment), and errors in cell cycle checkpoints (Thompson et al., 2010). Segmental aneuploidies often result from the formation of unbalanced gametes in carriers of inversions and balanced translocations. They, like CNVs, can arise due to errors in DNA replication and repair, leading to deletions or amplifications of DNA sequences and structural chromosomal rearrangements (Colnaghi et al., 2011). An abnormal number of chromosomes in the zygote leads to constitutive aneuploidy, the state where all cells are aneuploid. The occurrence of aneuploidy at later stages of organism development leads to somatic mosaicism, which may not have a pathological effect. For example, in some human tissues (brain, liver), a significant number of aneuploid cells are normally detected in the absence of a negative effect on the normal function of these organs (Rehen et al., 2005; Duncan et al., 2012). In this work, we will separately consider the aforementioned variants of aneuploidy, starting with whole-chromosome aneuploidy.

Constitutive whole-chromosome aneuploidy

Both the complexity of the tissue and organ organization and the peculiarities of the structural and functional organization

of the genomes of species belonging to different taxa can cause fundamental differences in the frequency and manifestation of constitutive aneuploidy in different eukaryotic species. Whole-chromosome aneuploidy causes developmental defects, which are frequently severe and fatal, in the majority of species. In some organisms, it may nevertheless represent a variation of the norm. By excluding species with microchromosomes in their karyotypes, we may conclude that whole-chromosome aneuploidy results in gene imbalance, which alters the expression of genes located on the aneuploid chromosome. It is widely accepted that this genetic imbalance affects the development and fitness of an organism at the cellular and organismal level (Torres et al., 2007; Williams, Amon, 2009; Rutledge, Cimini, 2016). The transcriptome studies of aneuploids revealed that the expression levels of genes located on euploid chromosomes also changed, in addition to the number of transcripts of genes directly associated with the aneuploid chromosome (Letourneau et al., 2014; Dürrbaum, Storchová, 2016). Even when the copy number of gene transcripts in a cell changes – for instance, genes with a changed copy number – the amount of their protein product may remain unchanged, which makes evaluating the effect of aneuploidy exceedingly challenging (Muenzner et al., 2024).

Human aneuploidy

In humans, all constitutive variants of autosomal aneuploidy, with the exception of the most common trisomies of autosomes 13, 18, and 21 (Tr13, Tr18, and Tr21, respectively), lead to embryonic mortality. Trisomies of chromosomes 13, 18, and 21 lead to serious developmental abnormalities and are associated with certain clinical phenotypes: Patau syndrome (Tr13), Edwards syndrome (Tr18), and Down syndrome (Tr21) (Lejeune et al., 1959; Edwards et al., 1960; Patau et al., 1960).

Sex chromosome aneuploidies are characterized by different clinical features and outcomes. The most common syndromes are Turner (45,X), Klinefelter (47,XXY), trisomy of the X chromosome (47,XXX), and disomy of the Y chromosome (47,XYY) (Berglund et al., 2020), demonstrating high phenotypic variability with a wide range of clinical manifestations. Clinical phenotypes of patients with various variants of sex chromosome aneuploidies (45,X, 47,XYY, 48,XYYY, 48,XXYY, 49,XXYYY, mos 46,XY/47,XYY, 48,XXYY, 49,XXXYY, 47,XXX, 48,XXXX, 49,XXXXX, and 47,XXY) are described in the atlas of K. Jones et al. (Jones et al., 2022). It is worth noting that monosomy of the X chromosome (45,X) in 99 % of cases leads to the death of the embryo in the early stages of development; a small percentage of the embryos survive, which is probably linked to the mosaic form of the karyotype (Gravholt et al., 2019).

In humans, males are haploid for almost all X-linked genes, which suggests a more stringent natural selection of X chromosome variants during evolution in comparison with autosomes based on the presence of pathogenic gene variants and genes, a change in the copy number of which leads to developmental abnormalities. When considering variants of X-chromosome aneuploidy, one should mention inactivation of one of its copies (XCI, X-Chromosome Inactivation). However, in early human embryogenesis, in the cells of the trophectoderm and inner cell mass, both X chromosomes remain active (Deng X. et al., 2014). In humans, XCI is incomplete, with about 20–25 % of genes remaining active. On the one hand, the result of incomplete XCI can be considered as segmental aneuploidy; on the other hand, the inactivated X chromosome is a heterochromatic extrachromosome, the presence of which can lead to a change in the pattern of the entire cellular transcriptome through epigenetic changes (Deng X. et al., 2014). XCI occurs randomly; the mechanisms by which copies of the X chromosome are selected to be inactivated are unknown. The consequence of such inactivation is the emergence of mosaicism in the expression of allelic variants of genes (i. e., unequal expression of parental alleles) associated with the X chromosome (Werner et al., 2024). Random XCI results in approximately half of the cells having the paternal X chromosome inactivated and the other half having the maternal X chromosome inactivated. However, in some cases, unequal XCI may occur, with different tissues having different ratios of cells with inactivated maternal or paternal X chromosomes. Disturbances in equiprobable X-chromosome inactivation (e. g., a mutant allele of an X-linked gene is expressed in most cells) can lead to the development of X-linked diseases (Minks et al., 2008).

About 12–15 % of X-linked genes remain active in all cells, while for another 8–10 % of X-linked genes, transcription is observed only in some cell types (Carrel, Willard, 2005; Balaton et al., 2015). Altered transcription levels, in addition to mRNA, were also detected for non-coding RNA genes (including microRNA, lncRNA, and circular RNA). The expression level of genes that are not subject to inactivation varies widely (10–95 %) in different cell types. Mosaic aneuploidies of sex chromosomes deserve special mention. Since samples of patients' peripheral blood are most often used for cytogenetic analysis, it is extremely problematic to assess the level of mosaicism of sex chromosomes in different tissues. However, even in the blood of such patients, more than 30 % of mosaic variants were detected for 45,X and 47,XXX karyotypes and a lower level of mosaicism for 47,XXY and 47,XYY (6–7 and 11 %, respectively) (Gravholt et al., 2019; Pavlicek et al., 2022; Tallaksen et al., 2023). The question of to what extent the imbalance in gene copy number is corrected at the proteome level in sex chromosome aneuploidies remains open.

Whole-chromosome aneuploidy in different species of eukaryotes

The negative impact of aneuploidy manifests itself at both the cellular and organismal levels. In most species of eukaryotes, it leads to developmental abnormalities, diseases, and nonviability. However, the frequency of aneuploidy and its effect on the host phenotype can vary greatly among representatives of different taxa. In mammals, autosomal aneuploidies have a pronounced negative effect, including fatal developmental abnormalities. In addition to humans, a high frequency of aneuploidy has been described in frozen embryos and piglet embryos with developmental defects and in calf embryos obtained *in vitro*. Sex chromosome monosomies have been

identified in sterile sheep and cattle (Bouwman et al., 2023). Genome imbalance leads to proteotoxic and metabolic stress in the cell, slow proliferation, genomic instability, oxidative stress, etc. (Stingele et al., 2012).

However, in some species from a number of taxa, tolerance to aneuploidy has been revealed, for example, in plants – salsify, *Tragopogon miscellus* (Chester et al., 2012); in fungi – *Saccharomyces cerevisiae*, *Candida albicans* (Rustchenko, 2007; Kvitek et al., 2008); in protozoa – *Leishmania*, *Giardia*, *Trypanosoma* (Sterkers et al., 2010); and among flatworms – some representatives of the genus *Macrostomum* (Zadesenets et al., 2020). Moreover, in aneuploids of some species, an unbalanced karyotype may likely contribute to adaptation to various environments. For example, in a number of pathogenic yeasts, aneuploidy leads to the formation of genomic diversity and the emergence of antibiotic resistance (Pavelka et al., 2010).

Karyotypes of protozoans of the genus *Leishmania* (the causative agent of leishmaniasis in mammals, including humans) contain from 34 to 36 chromosomes, and a number of studies have shown that aneuploid variants predominate among them (Lachaud et al., 2014). Moreover, an amazing feature in the form of constitutive mosaic aneuploidy was revealed in *L. major*: individuals of the same line, even having the same clonal origin, are mosaic and contain mono-, di-, and trisomic cells on different chromosomes (Sterkers et al., 2011, 2012). Mosaic aneuploidy was later identified in other *Leishmania* species (Lachaud et al., 2014). In addition to pronounced genotypic and karyotypic diversity, *Leishmania* is characterized by maintaining high genetic heterogeneity in a population consisting of homozygous individuals. The authors believe that genomic variability, due to the high plasticity of the karyotype, provides phenotypic diversity and is an adaptive mechanism of *Leishmania* to environmental changes during a complex life cycle (Sterkers et al., 2012).

An unusual variant of aneuploidy was discovered in natural populations and laboratory lines of free-living flatworms of the genus *Macrostomum* (Zadesenets et al., 2016, 2020). The genomes of *M. lignano*, *M. janickei*, and *M. mirumnovem* arose due to a recent whole-genome duplication followed by intensive reorganization of the duplicated genome (chromosomal fusions, inversions, indels, etc.) (Zadesenets et al., 2020, 2023; Zadesenets, Rubtsov, 2021). In these species, the karyotype evolution involved the fusions of all ancestral chromosomes into one large chromosome. In *M. lignano*, the presence of aneuploids with tri- and tetrasomy on a large chromosome, exhibiting no phenotypic and reproductive features, was recorded (Zadesenets et al., 2016). The karyotypic variation in *M. mirumnovem* was so high that a specific nomenclature for the species' chromosomes had to be established. A hypothetical basic karyotype had to be introduced in order to apply the standards for characterizing karyotypes that are recognized in modern cytogenetics (Zadesenets et al., 2020). A large chromosome with extensive paralogous regions that were highly homologous to the chromosomes of the ancestral set was linked to the whole-chromosome aneuploidy observed in the *Macrostomum* species (Zadesenets et al., 2017a, b).

A distinct, less noticeable manifestation of aneuploidy in organisms with an increased genome ploidy is worth discussing separately, in addition to species with a recent whole-genome duplication. Comparing the effects of aneuploidy across species reveals its species specificity, which may be related to varying levels of tissue and organogenesis complexity.

Model systems to study aneuploidy

The presence of species tolerant to aneuploidy would seem to facilitate a simple and effective establishment of experimental models for its study. Indeed, numerous studies performed on aneuploid yeast strains have significantly expanded the fundamental knowledge of the causes and consequences of the effect of aneuploidy on the genome, transcriptome, and proteome of the cell (Torres et al., 2007; Pavelka et al., 2010; Torres, 2023). Moreover, mechanisms for correcting imbalances in gene dosage have been proposed.

In humans, modeling of aneuploidies has naturally been limited to experiments with cell cultures obtained from patients with aneuploidies and/or the creation of aneuploid cells using chromosome engineering methods (MMCT, Microcell-Mediated Chromosome Transfer; targeted chromosome elimination with *Cre/loxP*, CRISPR/Cas9; induction of CIN) (Fournier, Ruddle, 1977; Thomas et al., 2018; Leibowitz et al., 2021; Zhang X.M. et al., 2022; Truong et al., 2023).

Whole-chromosome aneuploidy in cells cultured *in vitro*

Cell cultures and lines maintained *in vitro* are tolerant of aneuploidy. To identify the effect of aneuploidy, their proliferative potential is usually assessed. Whole-chromosome monosomy is rare in cell lines. Note that some authors believe that monosomies, in contrast to whole-chromosome and segmental tri- and tetrasomies, less often lead to chromosomal instability (CIN, Chromosomal Instability) (Taylor et al., 2019). *In vivo*, monosomies are most often associated with hematological malignancies; monosomies on the arms of some chromosomes are also associated with malignant neoplasms (deletion of 1p – neuroblastoma, 3p – lung cancer, 7q or entire HSA7 – myeloid leukemia) (Taylor et al., 2019). This is likely due to loss of heterozygosity for tumor suppressor genes; for example, deletion of 17p in many tumors is associated with the loss of a copy of the *TP53* gene in the absence of its normal allele on the homologous chromosome (Chundury et al., 2021).

The problem of loss of heterozygosity should probably be considered in detail separately, taking into account the possibility of obtaining and maintaining haplodiploid cell cultures. For example, sequencing the genome of cells from one of these cultures allowed for the complete assembly of its haploid set from telomere to telomere (T2T-CHM13), including extended regions of heterochromatin (Nurk et al., 2022). The result was the announcement of the successful completion of the human genome sequencing program (Nurk et al., 2022).

Aneuploid cells *in vitro* typically exhibit a reduced proliferation rate. Taking into account that they undergo excessive protein synthesis, some of which can be leveled by the ubiquitin-proteasome system, such a slowdown in proliferation seems natural but not critical for obtaining and maintaining cell cultures *in vitro*. At the same time, a decreased rate of

cell proliferation during the development of the organism at various stages of ontogenesis can be critical and lead to serious disorders.

It should be noted that when cultivating cell lines *in vitro*, as in cells at an early stage of tumorigenesis *in vivo*, genome doubling (WGD, Whole Genome Duplication) can occur, leading to its tetraploidization. Subsequently, these cells, due to a reduction in the number of chromosomes due to tolerance to chromosome segregation errors in mitosis, become aneuploid (hypotetraploid). This may subsequently induce additional genomic instability. Consequently, for cells to simply proliferate, it is not necessary to maintain a balance in the number of chromosomes; moreover, when they overcome the proliferative barrier (the Hayflick limit) and become malignant, it is often accompanied by CIN. In this case, aneuploid cells are better adapted to environmental conditions and proliferate faster. However, rapid proliferation does not ensure coordinated behavior of cells in the organism. Rather, it leads to the formation of various developmental abnormalities, for example, pathologies in histo- and organogenesis.

Effect of gene dosage on the transcriptome in aneuploidy

The effect of aneuploidy on gene expression has been studied in a variety of experimental models, both cell lines and model organisms. Unfortunately, most studies only assessed the number of transcripts of differentially expressed genes (DEGs). The effect of aneuploidy on the expression of genes localized directly on the chromosome with an altered copy number has been proven in yeast (et al., 2007; Torres et al., 2007), *Arabidopsis* (Huettel et al., 2008; Sheltzer et al., 2012), maize (Birchler et al., 2013), as well as for mouse (Williams et al., 2008) and human (Nawata et al., 2011; Stingele et al., 2012) cell lines. It is worth noting that aneuploid models included variants of the presence of additional copies of chromosomes and not the loss of one of the copies, i. e., tri- and tetrasomy, not monosomy.

To date, experiments conducted in cell cultures and aneuploid model organisms have shown that aneuploidy may have a broader effect on gene expression than previously thought. In addition to the primary cis-effects (changes in the level of transcripts of genes located on the aneuploid chromosome), secondary trans-effects (changes in the expression level of genes located on other chromosomes) were identified (Sheltzer et al., 2012; Birchler, 2013; Dürrbaum, Storchová, 2016).

In *in vivo* models, the trans effect of aneuploidy on gene expression was first described in maize (Guo, Birchler, 1994). Later, using the example of various cellular and organismal model systems, it was shown that in aneuploidy, the list of DEGs is not limited to the genes of aneuploid chromosomes and includes a significant number of genes from euploid chromosomes. This phenomenon was called the aneuploidy-induced transcriptional response (Sheltzer et al., 2012). Trans effects of aneuploidy have been identified in aneuploid cells of yeast, mice, and humans. In yeast, the trans effect of aneuploidy affects about 5–7 % of genes. When comparing euploid human fibroblasts and fibroblasts with trisomy 21, about 88 % of DEGs are not associated with chromosome 21 but are distributed on other chromosomes (Sullivan et al., 2016). In Turner and Klinefelter syndromes, more than 75 % of DEGs were identified in autosomes, while in carriers of karyotypes 46,XXX and 47,XYY, less than 30 % of DEGs were autosomal (Raznahan et al., 2018). The extent to which trans effects of aneuploidy occur varies among species, and the underlying mechanisms are still poorly understood (Li R., Zhu, 2022).

Thus, the physiological and phenotypic effects of aneuploidy may be associated either directly with changes in the copy number of genes located on the aneuploid chromosome or indirectly with changes in the expression of many genes on euploid chromosomes. The result may be additive or synergistic expression and functional effects at the transcriptional and/or posttranscriptional levels (Pavelka et al., 2010). This is consistent with the nonlinear nature of gene dosage effects that determine subsequent biochemical processes in the cell (Veitia et al., 2013; Pires, Conant, 2016). Although many of the biological effects caused by aneuploidy are consistent with the gene dosage balance hypothesis (Birchler, Veitia, 2012; Veitia, Potier, 2015), it is worth considering the impact of the presence of extra chromosomes on the spatial organization of the nucleus and potentially on the genome-wide transcriptional activity of a wide variety of genes.

Separately, it is worth noting that often in studies of the effect of aneuploidy on the transcriptome, non-isogenic lines are used when analyzing DEGs (especially in studies conducted on human cells), which, when conducting a comparative analysis, introduces additional difficulties for correctly assessing the contribution of aneuploidy and the existing genetic diversity. The use of isogenic lines, differing only in the presence of an additional chromosome, could significantly increase the efficiency and reliability of the analysis. Such lines can be obtained by cloning mosaic samples. An alternative approach currently implemented is the comparison of transcriptomes and genomes of individual cells obtained from mosaics based on chromosomal aneuploidies (Wang S. et al., 2024).

Note that the effect of aneuploidy on one chromosome on the cell transcriptome as a whole significantly complicates the analysis and assessment of the effect of aneuploidy. Among the genes listed in the OMIM database (Online Mendelian Inheritance in Man, https://omim.org), only a part of them showed a pathogenic effect when their copy number changed, but the secondary effect of aneuploidy may be an extremely important component of its total pathogenic effect. Nevertheless, it is logical to expect that the more dosage-sensitive genes and genes encoding transcription factors, peptides, proteins, and small RNAs that affect the transcriptional activity of many genes there are in a given chromosome, the stronger the change in the transcriptome and disturbance of homeostasis in the cell, and the more pronounced the pathologies observed during histo- and organogenesis.

Possible outcomes of gene copy number alterations in individual cells

Monoallelic expression in individual cells should be taken into consideration when analyzing transcriptional changes caused by aneuploidy. In contrast to the data of the single-cell transcriptomes, the gene expression patterns obtained

earlier represented averaged data and did not accurately reflect the real gene expression in single human cells. Studies have revealed variability in monoallelic expression for most autosomal genes and gradations in gene expression during parent-of-origin imprinting and X-chromosome inactivation (Borel et al., 2015; Santoni et al., 2017; Garieri et al., 2018). The latter is likely a result of the stochastic and pulsed nature of transcription, in which transcription of each copy of a gene, including its allelic variants, is independently regulated and determines the monoallelic expression of most autosomal genes in a significant proportion of cells (Reinius, Sandberg, 2016; Larsson et al., 2019). Thus, despite aneuploidy, in some cells, transcription may occur from one copy of the gene, but at the same time, there will also be cells with transcription from a larger number of its copies. In cells with trisomy, there is an increase in the proportion of cells with simultaneous transcription from two or more copies of the gene, leading to an increase in the number of transcripts by one and a half times when analyzing the cell pool. Moreover, the picture may differ for different genes in one cell, creating a large diversity in the transcriptome of individual cells (Ramsköld et al., 2024).

Some studies have been devoted to the study of transcriptional bursting in individual cells, in which the frequency of transcriptional bursting (the time between acts of transcriptional bursting), its intensity (the number of transcripts synthesized in one act), and the stability of the synthesized mRNA were assessed (Deng Q. et al., 2014; Stamoulis et al., 2019; Larsson et al., 2021; Ramsköld et al., 2024). In this paper, we only note that the pulsed transcription of different copies of genes in a cell is independent, and at a sufficiently low frequency of the transcription act in a cell containing three copies of a gene, it can occur from one or several copies of the gene (Larsson et al., 2021). In most diploid cells, the expression of only one allele is predominant (monoallelic expression) (Stamoulis et al., 2019), while in a triploid cell, different transcript variants can be formed due to mono- or biallelic gene expression (Larsson et al., 2021). The stochastic determination of the transcription pattern and its time results in a distribution of cells based on the level of transcripts from various copies of the gene. Among cells with trisomy, the distribution includes cells with transcription from one copy, from two, and, rarely, from three copies, which provides an average value of the number of transcripts corresponding to a transcript level one and a half times higher than in a diploid cell.

Therefore, the concept of pulsed transcription assumes variability in the level of transcripts within the cell, as well as a high level of variability in the level of transcripts in aneuploidy. This is characterized by the appearance of cells with a high transcript content, the ability to select cells based on the number of transcripts of the corresponding genes, and the reproduction of variability in the number of transcripts in each subsequent generation of cells. Negative selection of cells by a high transcript level for dosage-dependent genes can lead to a delay of cell cycle progression or even induction of apoptosis. In other words, during ontogenesis, aneuploidy causes a continuous loss of cells involved in the formation of new tissues and organs. In some instances, the instability of the epiblast's development and changes in the development of the hypoblast and trophectoderm are evident at the early stages of development in human embryos with trisomy (Wang S. et al., 2024).

Dosage compensation at the transcriptome and proteome levels

Despite the fact that in aneuploidy the expression level for the majority of genes correlates with the altered number of gene copies in the cell, there have been instances of dosage compensation where the level of gene transcripts of genes on the aneuploid chromosome remains constant (Guo, Birchler, 1994; Birchler et al., 2001; Hose et al., 2015; Gasch et al., 2016). Some studies have shown that in aneuploidy, transcriptional dosage compensation may be provided by autoregulation of gene expression, suppression of mRNA translation, and mRNA decay. For example, in wild yeasts with an additional copy of chromosome 12, autoregulation (overproduction of a certain protein reduces the transcription of its gene) of the *RPL15A* and *RPL22A* genes encoding ribosomal proteins leads to their dosage compensation (Hose et al., 2015). The increased expression of genes encoding certain microRNAs (for example, *miR-155*) and localized on human chromosome 21 in Tr21 may lead to dosage compensation of genes localized on this chromosome or affect the expression level of genes on other chromosomes. For example, an increase in miR-155 can suppress the expression of the transcriptional regulator BACH1 located on chromosome 21 (Li R., Zhu, 2022).

A pronounced effect of post-translational dosage compensation has been described in aneuploids from natural isolates and laboratory strains of *S. cerevisiae* (Muenzner et al., 2024). Despite the fact that 20 % of the studied natural isolates were stable aneuploids, similar aneuploid laboratory-engineered strains were less stable. The transcriptomic profiles of the corresponding pairs of natural isolates and laboratory strains were similar, but while approximately 70 % of proteins encoded on aneuploid chromosomes were corrected to normal levels in natural aneuploid isolates, such a correction in laboratory strains was described for less than 50 % of such proteins. Moreover, if a decrease in the excess amount in laboratory strains was mainly observed for complex protein complexes, then in natural aneuploid isolates, the decrease in the excess amount of proteins affected all classes of proteins (Storchová, 2024). An increased level of ubiquitinylation was detected for proteins encoded on aneuploid chromosomes, and their abundance was reduced via the ubiquitin-proteasome system (UPS, Ubiquitin Proteasome System) (Muenzner et al., 2024).

Therefore, in yeast, the ubiquitin-mediated proteasomal degradation system plays an important, and possibly key, role in maintaining the balance of the proteome of an aneuploid cell (Storchová, 2024). The stability of natural aneuploid isolates of *S. cerevisiae* suggests that in their genome, there is an adaptation to the presence of an additional chromosome, or there is a selection of a genome variant in which aneuploidy not only does not have a negative effect but even has a positive adaptive effect. In addition to the UPS, other proteolytic mechanisms for correcting the proteome (autophagic-lysosomal system, calpain, and caspase enzymes) exist in the cell to regulate

protein homeostasis (Noormohammadi et al., 2018). For example, during proteotoxic stress in aneuploid human cells, the transcription factor TFEB is activated, which regulates the expression of genes involved in the autophagic-lysosomal pathway for the degradation of excess protein aggregates, and an additional mechanism for correcting the abundance of protein products in the cell is triggered (Santaguida et al., 2015).

Obviously, the idea of the pathogenic effect of aneuploidy, caused by a single disturbance in the balance of gene copies localized on the aneuploid chromosome, is too simplified. For instance, clinical manifestations with Tr21 vary significantly, which may likely be due to large differences between personal genomes, which can result in differences in the correction of the abundance of proteins encoded on chromosome 21, similar to what happens in aneuploid yeast.

Studies of the transcriptome and proteome of individual human cells at the stages of early embryogenesis (Wang S. et al., 2024) have significantly expanded the understanding of the role and mechanisms of manifestation of aneuploidy. A transcriptome analysis of about 15 thousand individual cells from 203 eu- and aneuploid human blastocysts (epi- and hypoblasts, polar and mural trophectoderm) showed that changes in the copy number of chromosomes are significant for ~20 % of genes. About 90 dosage-dependent domains have been identified in aneuploid chromosomes. Especially in monosomies, common consequences like apoptosis were found, which helps to explain why autosomal monosomies occur in fewer cells. It is likely that with autosomal monosomies, critical developmental disorders occur even before implantation. Of course, the cause of such disorders may be not only or not so much a change in gene dosage but a loss of heterozygosity, leading to the absence of complete copies of some genes in the cell. In this regard, it is not surprising that the sets of dosage-dependent genes in complementary tri- and monosomies turned out to be different. The downregulation of TGF-β and FGF signaling, which led to deficient trophectoderm maturation, was another lineage-specific consequence that caused unstable epiblast formation in aneuploids (Wang S. et al., 2024).

Aneuploidy and architecture of interphase nuclei

Previously, it was believed that changes in the copy number of chromosomes of the main set have an effect on the phenotype, mainly due to the imbalance of gene copies. However, in humans, the manifestation of a number of syndromes (at least with Tr21) is caused not only by an increased expression of genes from the aneuploid chromosome (Olson et al., 2004). Trans effects of aneuploidy have also been identified, and it has been hypothesized that the disruption of cellular homeostasis is caused by the presence of an extra chromosome (Krivega et al., 2022).

In the interphase nucleus, chromosomes and their regions are not randomly located relative to transcriptionally active and inactive compartments (Cremer T., Cremer C., 2001; Cremer T., Cremer M., 2010; Cremer M. et al., 2020). Moreover, the architectonics of the nucleus and chromosomal territories may differ both at different stages of ontogenesis and in different cell types (Croft et al., 1999; Tanabe et al., 2002). In the nuclei of cells that differ in morphology and tissue affiliation, different principles of spatial localization of chromosomes and chromosomal regions can be implemented (Cremer M. et al., 2003; Mayer et al., 2005; Solovei et al., 2013), determining its functional compartmentalization due to the specific distribution of transcriptionally active and inactive chromatin regions in the nucleus (Meaburn, Misteli, 2007).

The development of 3D genomics (3C, chromosome conformation capture, Hi-C, ChIA-PET, Micro-C, snHi-C, etc.) has significantly expanded the understanding of the levels of hierarchical and spatial organization of chromatin in the nucleus and the dynamics and plasticity of the structural and functional compartments of the nucleus (Dekker et al., 2002; Li G. et al., 2010; van Berkum et al., 2010; Nagano et al., 2013; Hsieh et al., 2020). For the human genome, topologically associated domains (TADs), A/B compartments and their subcompartments (Oji et al., 2024), chromatin loops, lamina-associated domains (LADs), nucleolus-associated domains (NADs), and their variants in different cell types and at different stages of development/differentiation are described in detail. Recent studies have investigated the influence of structural and numerical chromosomal aberrations on the spatial organization of chromatin (Shao et al., 2018; Wang Y. et al., 2023; Zhegalova et al., 2023).

The mechanisms of the influence of aneuploidy on changes in the spatial organization of chromatin in the nucleus are unknown, and in this work, we present only data from studies describing changes in nuclear architecture in aneuploid human cells. Important factors that determine the structural and functional organization of the genome are the connection of its specific sections with the nuclear lamina, the localization of chromatin relative to the nucleolus, and the formation of interchromatin compartments (nuclear bodies) (Razin, Ulianov, 2022). The spatial organization of the nucleus is determined primarily by the anchoring of chromosomal territories on the nucleolus (helped by NADs) and the nuclear lamina (helped by LADs), as well as the presence of interchromatin compartments (nuclear bodies) (Razin, Ulianov, 2022).

Although about a third of the human genome contains potential LADs (van Steensel, Belmont, 2017) in different cell types, only about 30 % of potential LADs are associated with the lamina (Zhegalova et al., 2023). Most genes located in lamina-associated regions are not expressed or expressed at low levels. Alterations in the composition of lamina-associated regions lead to changes in the transcriptome of the cell (van Steensel, Belmont, 2017; Shah et al., 2023). An important role is played by the distribution of LADs along the chromosome; chromosomes with a small proportion of LADs tend to be found medially, in the center of the nucleus; for example, human chromosome 19 is characterized by the highest gene density and has an internal position in the nucleus (Croft et al., 1999). Due to the altered chromosome copy number, the conditions of competition of potential LADs for association with the lamina may change, which can lead to changes in the structural organization of chromatin, and not only that of the aneuploid chromosome. This, in turn, can lead to changes in the transcriptional activity of genes located on different

chromosomes, and such changes can be critical, leading to disorders already at early stages of development (Zhegalova et al., 2023).

As an example, we can consider the organization of chromatin in the nuclei of aneuploid human colonic epithelial cells (HCEC) with trisomy of chromosome 7. 3-D FISH, a whole-chromosome probe that specifically stains the corresponding chromosomal territory did not reveal fundamental changes in the localization of the territory of the aneuploid chromosome in the interphase nucleus. However, Hi-C analysis, in addition to an increase in the frequency of interchromosomal contacts of DNA regions of chromosome 7, revealed changes in A/B compartmentalization and in the boundaries of TADs. Changes in the chromatin of chromosome 4 were detected: a reduction in the number of TADs (from 133 to 109) and movement of the chromatin of a chromosome 14 region (chr14:62.4Mb–63.8Mb) from the active A to the inactive B compartment (Braun et al., 2019).

In human chorionic villi cells at Tr21, changes in the nuclear localization of chromosomal territories of chromosomes 1 and 3 were noted (Kemeny et al., 2018). When studying other trisomies (Tr13, Tr16 in chorion cells; Tr18 in *in vitro* cultured fibroblasts), changes in patterns of interchromosomal contacts were noted for all human chromosomes (Zhegalova et al., 2023). These studies revealed a correlation between the number of loci with altered compaction and the number of LADs in the aneuploid chromosome (Tr13, Tr18). It turned out that the number of LADs in chromosomes 13 and 18 is three times higher than in chromosome 16, which could potentially cause a more pronounced effect on the chromatin-lamin interactome in the nucleus, leading to changes in chromatin compaction. In addition, it turned out that the number of loci with altered compaction in small chromosomes is higher in Tr16 compared to Tr13 and Tr18. The presence of an extra chromosome 16 also significantly reduced the frequency of DNA contacts of small chromosomes (chromosomes 16–22) in chorion cells (up to 20 % for a single pair of chromosomes). The authors suggest that additional copies of small chromosomes, competing with copies of similar small chromosomes, lead to changes in the distribution of their material in the nucleus, reducing the frequency of contacts (Zhegalova et al., 2023). In NPCs (neuronal progenitor cells), an extra copy of chromosome 21 increased the frequency of DNA contacts within the group of small chromosomes HSA16–22. Thus, aneuploidies of different chromosomes can lead to different changes in the spatial organization of chromatin in the interphase nucleus, and such changes can be different in different cell types (Meharena et al., 2022; Zhegalova et al., 2023).

The authors believe that in trisomies, different variants of spatial DNA contacts can be formed in different subpopulations of cells (Zhegalova et al., 2023), and the observed differences in the Hi-C data array may reflect the combined effect of several factors (the presence of an extra chromosome, the proliferative activity and age of the cell, the degree of its differentiation, etc.). Changes in the structural and functional organization of chromatin are probably of critical importance in early embryogenesis and are the cause of formation of multiple abnormalities in different tissues and organs observed in trisomy (Zhegalova et al., 2023). The authors believe that changes in the spatial organization of chromatin, systematic and stochastic, are determined by a combination of many factors, including the size of the chromosome, its LAD coverage, and the density of gene localization in it (Zhegalova et al., 2023). However, it should be recognized that most questions about the effect of aneuploidy on the architecture of the nucleus and the structural and functional organization of chromatin remain unanswered.

The small amount of research conducted, which sheds only a little light on the effect of aneuploidy on the spatial organization of the nucleus in trisomies in human cells, leaves open the question of the presence of features or general patterns in changes in the spatial organization of the genome during chromosomal aberrations.

Mosaic aneuploidy

As a result of errors that occur in mitosis during the proliferation of somatic cells, cells with an altered genome constantly appear in the body. As a result, most organisms are mosaics. In humans, aneuploid cells are present in various tissue types, including hepatocytes (2.2 %), neurons (<5 %), lymphocytes, etc. (Knouse et al., 2014). Aneuploidies of different chromosomes (HSA1, 7, 8, 9, 10, 11, 14, 15–18, 21, and X/Y) have been identified in brain cells (Graham et al., 2019). It is possible that somatic mosaicism contributes to the formation of diversity, in which neurons of the same lineage perform different functions (McConnell et al., 2017). It turned out that somatic mosaicism is more often observed for sex chromosomes than for autosomes (Machiela et al., 2016). In lymphocytes, mosaicism on the Y chromosome associated with its loss (mLOY) is the most common type of aneuploidy (1.7–20 %) (Graham et al., 2019). Several characteristics should be considered when examining mosaicism studies. Accordingly, if the percentage of cells with a different karyotype was at least 5 % when mosaicism was detected using FISH conducted on interphase nuclei, it was deemed significant (Modi et al., 2003; Yurov et al., 2007); however, 1.6 % was already deemed significant when mosaicism on the X chromosome was examined (Guttenbach et al., 1995).

The phenotype of mosaics depends on the proportion of aneuploid cells, which may vary in different tissues and at different stages of development. Analysis of individual cells of embryos at the preimplantation stage (blastocyst) showed the presence of aneuploid and mosaic embryos. According to different studies, the proportion of mosaic embryos varied from 2 to 90 % (Starostik et al., 2020; Rana et al., 2023). It is worth noting that in a number of cases, when analyzing a pool of cells, mosaicism in embryos was not detected, since aneuploidy in the cells was compensatory (trisomy and monosomy on the same chromosome in different cells). The use of methods for analyzing the genome and transcriptome of single cells (scWGS, scRNAseq) of the embryo has made it possible to describe in more detail the levels of mosaicism at different stages of embryonic development. It was found that 100 % of the analyzed embryos were mosaics at the blastocyst stage; during the development of the embryo, at later stages of its development (5–26 weeks of gestation), the proportion of

aneuploid cells decreased. In addition, cases of healthy children being born with a normal karyotype, although aneuploidy was detected during retrospective analysis of their embryonic cells, have been described (Zhai et al., 2024).

Concluding a brief discussion of the problems of mosaicism associated with aneuploidy of different chromosomes, we note that it can occur in cancer cells after WGD in the early stages of tumorigenesis (Lambuta et al., 2023) and/or as a result of CIN, including both numerical and structural chromosome aberrations (Li R., Zhu, 2022). According to recent data, WGD is detected in 30 % of tumors at the early stages of tumorigenesis (Lambuta et al., 2023). Up to 90 % of solid tumors and 70 % of hematopoietic malignancies are associated with aneuploidy (Xiao et al., 2024). An increased frequency of chromosomal abnormalities, including aneuploidy, is also observed in *in vitro* cultured human embryonic stem cells, which may contribute to their potential tumorigenicity (Baker et al., 2007). The phenomenon of CIN, associated with WGD and/or aneuploidy, is often accompanied by genomic instability and manifests itself in the form of diversity of tumor cell karyotypes and high intra- and inter-tumor heterogeneity of the cancer cell genome (Burrell et al., 2013).

Segmental aneuploidy and CNVs

Segmental aneuploidy and CNV might have distinct origins. In carriers of balanced translocations, unbalanced gametes arise, leading to various clinically significant forms of segmental aneuploidy. Despite the 50 % frequency of such gametes, the percentage of children with partial trisomy and partial monosomy in such parents is lower. It is unknown when selection favors carriers of a balanced genome, and it may vary depending on the type of chromosomal rearrangement. Even standard cytogenetic techniques can easily determine a balanced translocation in parents if both translocation-related chromosomal regions are relatively large. Unfortunately, detecting such a balanced translocation can be difficult if one of the chromosomal regions is small and distal. To do this, FISH using DNA probes specific to distal regions of the chromosomes or microarray-CGH is required. The goal of medical cytogenetics is to discover and characterize carriers of these combined partial trisomies and monosomies as well as carriers of balanced chromosomal translocations, whose offspring may also be carriers of combined partial trisomies and monosomies. It should be noted that these combinations have a pathogenic effect.

DNA replication and repair errors result in the loss or gain of chromosome regions, leading to CNVs and segmental aneuploidies. In studying the clinical significance of such segmental aneuploidies and CNVs, researchers have encountered unique challenges. Whole-genome sequencing of thousands of personal human genomes has revealed a huge number of bi- and multiallelic single nucleotide variants (SNVs, Single Nucleotide Variants), biallelic indels, and structural variants (SVs, Structural Variants) of the genome, including large insertions, deletions, inversions, and variations of genomic regions by copy number (The 1000 Genomes Project Consortium et al., 2015). Given this variability, assessing the potential pathogenic significance of variations in a specific genome region's copy number frequently proves to be quite a challenging task. In this section, we will consider cases of appearance of additional copies of genome regions because, in a diploid organism, loss of a chromosomal region leads to haploidization of part of the genome and usually has a pronounced pathogenic effect or a delayed pathogenic effect. However, in the case of a tetraploid genome, the loss of one copy of a genomic fragment may be one of the first stages towards genomic rediploidization, which is a very important stage in genomic evolution, but its consideration is beyond the scope of this review.

A bioinformatics study of an aneuploid chromosome region's composition usually involves considering a number of hypotheses. Due to the enormous genomic diversity in humans, analysis of a large number of patients is required to make a definitive conclusion about the clinical significance of specific CNVs. In addition, since the same CNVs or segmental aneuploidies can manifest themselves in fundamentally different ways in different genomes, analysis of a large number of cases of a particular CNV in relatives may not provide a definitive answer. Finding patients with identical CNVs is often a challenging task because the frequency of each specific CNV is low, and the study of patients and their relatives reduces the ability to assess its clinical significance when found in different genomes. As a result of the analysis of a large sample of patients, CNVs can be classified as either variants without pathogenic influence, or without potential pathogenic influence, or as CNVs with unknown influence on the phenotype, or as CNVs with possible potential pathogenic influence, and finally as CNVs with pathogenic influence (Zhang F. et al., 2009; Auwerx et al., 2022). It should be taken into account that the genomes of people from diverse populations have substantial differences and are well divided into clades (Mallick et al., 2016). Moreover, they may also differ in the presence of DNA that originates from other, long-extinct hominins (Neanderthals, Denisovans, etc.) (Vernot et al., 2016). Thus, it cannot be ruled out that a conclusion drawn for one group of populations will be incorrect for another.

Importance of localization of segmental duplications in the genome

The location of the chromosome's changed copy number region is important. Duplications may occur as a single structural and functional element of the chromosome (TAD) or as a tandem cluster of duplicons, distant from the original sequence, in a human small supernumerary marker chromosome (sSMC), or in an extra chromosome (B chromosome) in other eukaryotic species. If the structural and functional organization of the duplicated region and its localization are in tandem relative to the original region and are preserved, one can expect the presence of transcriptional activity of the genes included in this region.

It is more difficult to assess the impact of additional material in human sSMC due to their variable content. The majority of them are composed of the original chromosome's pericentromeric region, which includes nearby heterochromatin and perhaps euchromatin with a variable number of genes. It has been observed that if the size of the euchromatic region of

human sSMC does not exceed 3–5 Mb, it usually does not have a pathogenic effect. It can be assumed that the absence of negative phenotypic traits in the carrier of such sSMC is associated with inactivation of the sSMC material due to the localization of its domains in the transcriptionally inactive compartment of the interphase nucleus in comparison with the homologous region of the original chromosome. Therefore, conducting a number of studies on the spatial organization of the genome with sSMCs of varying sizes and DNA content is an urgent and highly intriguing task, the resolution of which would enable us to assess the potential pathogenic effect of different sSMCs.

Conclusion

Considering the data on the manifestation of various variants of aneuploidy, it should be noted that there are a huge number of factors that can play a very significant role and influence their manifestation. A significant factor is probably the complexity of the tissue and organ organization of the organism of a particular species. Thus, yeast, like cell cultures, is quite tolerant of chromosomal aneuploidy. Polyploid organisms and species that have relatively recently undergone whole-genome duplication are usually much more resistant to aneuploidy. A special position is occupied by aneuploidy of sex chromosomes, which may be due to the peculiarities of their gene composition formed during the process of evolution.

A special variant of aneuploidy is represented by segmental aneuploidies and CNVs. In these cases, the composition of the additional material, its localization in the chromosome, and its localization in different compartments of the interphase nucleus may be of particular importance. Of particular interest are the mechanisms of dosage compensation for changes in the level of gene product during aneuploidy at the post-transcriptional and post-translational levels.

The study of aneuploidies and their clinical significance is of great interest in light of data on the huge diversity of personal human genomes, including SNVs, SVs, and CNVs. It can make a great contribution to improving cytogenomic diagnostics by creating the necessary database for the correct interpretation of identified cases of CNVs and segmental aneuploidy.

References

Auwerx C., Lepamets M., Sadler M.C., Patxot M., Stojanov M., Baud D., Mägi R.; Estonian Biobank Research Team; Porcu E., Reymond A., Kutalik Z. The individual and global impact of copy-number variants on complex human traits. *Am J Hum Genet*. 2022; 109(4):647-668. doi 10.1016/j.ajhg.2022.02.010

Baker D.E., Harrison N.J., Maltby E., Smith K., Moore H.D., Shaw P.J., Heath P.R., Holden H., Andrews P.W. Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat Biotechnol*. 2007;25(2):207-215. doi 10.1038/nbt1285

Balaton B.P., Cotton A.M., Brown C.J. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ*. 2015;6:35. doi 10.1186/s13293-015-0053-7

Berglund A., Stochholm K., Gravholt C.H. The epidemiology of sex chromosome abnormalities. *Am J Med Genet C Semin Med Genet*. 2020;184(2):202-215. doi 10.1002/ajmg.c.31805

Birchler J.A. Aneuploidy in plants and flies: the origin of studies of genomic imbalance. *Semin Cell Dev Biol*. 2013;24(4):315-319. doi 10.1016/j.semcdb.2013.02.004

Birchler J.A., Veitia R.A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci USA*. 2012;109(37):14746-14753. doi 10.1073/pnas.1207726109

Birchler J.A., Bhadra U., Bhadra M.P., Auger D.L. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol*. 2001;234(2):275-288. doi 10.1006/dbio.2001.0262

Borel C., Ferreira P.G., Santoni F., Delaneau O., Fort A., Popadin K.Y., Garieri M., Falconnet E., Ribaux P., Guipponi M., Padioleau I., Carninci P., Dermitzakis E.T., Antonarakis S.E. Biased allelic expression in human primary fibroblast single cells. *Am J Hum Genet*. 2015;96(1):70-80. doi 10.1016/j.ajhg.2014.12.001

Bouwman A.C., Hulsegge I., Hawken R.J., Henshall J.M., Veerkamp R.F., Schokker D., Kamphuis C. Classifying aneuploidy in genotype intensity data using deep learning. *J Anim Breed Genet*. 2023;140(3):304-315. doi 10.1111/jbg.12760

Braun R., Ronquist S., Wangsa D., Chen H., Anthuber L., Gemoll T., Wangsa D., Koparde V., Hunn C., Habermann J.K., Heselmeyer-Haddad K., Rajapakse I., Ried T. Single chromosome aneuploidy induces genome-wide perturbation of nuclear organization and gene expression. *Neoplasia*. 2019;21(4):401-412. doi 10.1016/j.neo.2019.02.003

Burrell R.A., McClelland S.E., Endesfelder D., Groth P., Weller M.C., Shaikh N., Domingo E., Kanu N., Dewhurst S.M., Gronroos E., Chew S.K., Rowan A.J., Schenk A., Sheffer M., Howell M., Kschischo M., Behrens A., Helleday T., Bartek J., Tomlinson I.P., Swanton C. Replication stress links structural and numerical cancer chromosomal instability. *Nature*. 2013;494(7438):492-496. doi 10.1038/nature11935

Carrel L., Willard H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*. 2005;434(7031): 400-404. doi 10.1038/nature03479

Chester M., Gallagher J.P., Symonds V.V., Cruz da Silva A.V., Mavrodiev E.V., Leitch A.R., Soltis P.S., Soltis D.E. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc Natl Acad Sci USA*. 2012; 109(4):1176-1181. doi 10.1073/pnas.1112041109

Chikashige Y., Tsutsumi C., Okamasa K., Yamane M., Nakayama J., Niwa O., Haraguchi T., Hiraoka Y. Gene expression and distribution of Swi6 in partial aneuploids of the fission yeast *Schizosaccharomyces pombe*. *Cell Struct Funct*. 2007;32(2):149-161. doi 10.1247/csf.07036

Chunduri N.K., Menges P., Zhang X., Wieland A., Gotsmann V.L., Mardin B.R., Buccitelli C., Korbel J.O., Willmund F., Kschischo M., Raeschle M., Storchova Z. Systems approaches identify the consequences of monosomy in somatic human cells. *Nat Commun*. 2021; 12:5576. doi 10.1038/s41467-021-25288-x

Colnaghi R., Carpenter G., Volker M., O'Driscoll M. The consequences of structural genomic alterations in humans: genomic disorders, genomic instability and cancer. *Semin Cell Dev Biol*. 2011;22(8): 875-885. doi 10.1016/j.semcdb.2011.07.010

Cremer M., Küpper K., Wagler B., Wizelman L., Hase J., Weiland Y., Kreja L., Diebold J., Speicher M.R., Cremer T. Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. *J Cell Biol*. 2003;162(5):809-820. doi 10.1083/jcb.200304096

Cremer M., Brandstetter K., Maiser A., Rao S.S.P., Schmid V.J., Guirao-Ortiz M., Mitra N., Mamberti S., Klein K.N., Gilbert D.M., Leonhardt H., Cardoso M.C., Aiden E.L., Harz H., Cremer T. Cohesin depleted cells rebuild functional nuclear compartments after endomitosis. *Nat Commun*. 2020;11(1):6146. doi 10.1038/s41467-020-19876-6

Cremer T., Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*. 2001;2(4): 292-301. doi 10.1038/35066075

Cremer T., Cremer M. Chromosome territories. *Cold Spring Harb Perspect Biol*. 2010;2(3):a003889. doi 10.1101/cshperspect.a003889

Croft J.A., Bridger J.M., Boyle S., Perry P., Teague P., Bickmore W.A. Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol*. 1999;145(6):1119-1131. doi 10.1083/jcb.145.6.1119

Dekker J., Rippe K., Dekker M., Kleckner N. Capturing chromosome conformation. *Science*. 2002;295(5558):1306-1311. doi 10.1126/science.1067799

Deng Q., Ramsköld D., Reinius B., Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*. 2014;343(6167):193-196. doi 10.1126/science.1245316

Deng X., Berletch J., Nguyen D., Disteche C.M. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet*. 2014;15:367-378. doi 10.1038/nrg3687

Duncan A.W., Hanlon Newell A.E., Smith L., Wilson E.M., Olson S.B., Thayer M.J., Strom S.C., Grompe M. Frequent aneuploidy among normal human hepatocytes. *Gastroenterology*. 2012;142(1):25-28. doi 10.1053/j.gastro.2011.10.029

Dürrbaum M., Storchová Z. Effects of aneuploidy on gene expression: implications for cancer. *FEBS J*. 2016;283(5):791-802. doi 10.1111/febs.13591

Edwards J.H., Harnden D.G., Cameron A.H., Crosse V.M., Wolff O.H. A new trisomic syndrome. *Lancet*. 1960;1(7128):787-790. doi 10.1016/s0140-6736(60)90675-9

Fournier R.E., Ruddle F.H. Microcell-mediated transfer of murine chromosomes into mouse, Chinese hamster, and human somatic cells. *Proc Natl Acad Sci USA*. 1977;74(1):319-323. doi 10.1073/pnas.74.1.319

Garieri M., Stamoulis G., Blanc X., Falconnet E., Ribaux P., Borel C., Santoni F., Antonarakis S.E. Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc Natl Acad Sci USA*. 2018;115(51):13015-13020. doi 10.1073/pnas.1806811115

Gasch A.P., Hose J., Newton M.A., Sardi M., Yong M., Wang Z. Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. *eLife*. 2016;5:e14409. doi 10.7554/eLife.14409

Graham E.J., Vermeulen M., Vardarajan B., Bennett D., De Jager P., Pearse R.V. 2nd, Young-Pearse T.L., Mostafavi S. Somatic mosaicism of sex chromosomes in the blood and brain. *Brain Res*. 2019;1721:146345. doi 10.1016/j.brainres.2019.146345

Gravholt C.H., Viuff M.H., Brun S., Stochholm K., Andersen N.H. Turner syndrome: mechanisms and management. *Nat Rev Endocrinol*. 2019;15:601-614. doi 10.1038/s41574-019-0224-4

Guo M., Birchler J.A. Trans-acting dosage effects on the expression of model gene systems in maize aneuploids. *Science*. 1994;266(5193):1999-2002. doi 10.1126/science.266.5193.1999

Guttenbach M., Koschorz B., Bernthaler U., Grimm T., Schmid M. Sex chromosome loss and aging: in situ hybridization studies on human interphase nuclei. *Am J Hum Genet*. 1995;57(5):1143-1150

Hose J., Yong C.M., Sardi M., Wang Z., Newton M.A., Gasch A.P. Dosage compensation can buffer copy-number variation in wild yeast. *eLife*. 2015;4:e05462. doi 10.7554/eLife.05462

Hsieh T.S., Cattoglio C., Slobodyanyuk E., Hansen A.S., Rando O.J., Tjian R., Darzacq X. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol Cell*. 2020;78(3):539-553.e8. doi 10.1016/j.molcel.2020.03.002

Huettel B., Kreil D.P., Matzke M., Matzke A.J. Effects of aneuploidy on genome structure, expression, and interphase organization in *Arabidopsis thaliana*. *PLoS Genet*. 2008;4(10):e1000226. doi 10.1371/journal.pgen.1000226

Jones K.J., Jones M.C., del Campo M. Smith's Recognizable Patterns of Human Malformation. Elsevier Health Sciences, 2022

Kemeny S., Tatout C., Salaun G., Pebrel-Richard C., Goumy C., Ollier N., Maurin E., Pereira B., Vago P., Gouas L. Spatial organization of chromosome territories in the interphase nucleus of trisomy 21 cells. *Chromosoma*. 2018;127(2):247-259. doi 10.1007/s00412-017-0653-6

Knouse K.A., Wu J., Whittaker C.A., Amon A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci USA*. 2014;111(37):13409-13414. doi 10.1073/pnas.1415287111

Krivega M., Stiefel C.M., Storchova Z. Consequences of chromosome gain: a new view on trisomy syndromes. *Am J Hum Genet*. 2022;109(12):2126-2140. doi 10.1016/j.ajhg.2022.10.014

Kvitek D.J., Will J.L., Gasch A.P. Variations in stress sensitivity and genomic expression in diverse *S. cerevisiae* isolates. *PLoS Genet*. 2008;4(10):e1000223. doi 10.1371/journal.pgen.1000223

Lachaud L., Bourgeois N., Kuk N., Morelle C., Crobu L., Merlin G., Bastien P., Pagès M., Sterkers Y. Constitutive mosaic aneuploidy is a unique genetic feature widespread in the *Leishmania* genus. *Microbes Infect*. 2014;16(1):61-66. doi 10.1016/j.micinf.2013.09.005

Lambuta R.A., Nanni L., Liu Y., Diaz-Miyar J., Iyer A., Tavernari D., Katanayeva N., Ciriello G., Oricchio E. Whole-genome doubling drives oncogenic loss of chromatin segregation. *Nature*. 2023;615(7954):925-933. doi 10.1038/s41586-023-05794-2

Larsson A.J.M., Johnsson P., Hagemann-Jensen M., Hartmanis L., Faridani O.R., Reinius B., Segerstolpe Å., Rivera C.M., Ren B., Sandberg R. Genomic encoding of transcriptional burst kinetics. *Nature*. 2019;565(7738):251-254. doi 10.1038/s41586-018-0836-1

Larsson A.J.M., Ziegenhain C., Hagemann-Jensen M., Reinius B., Jacob T., Dalessandri T., Hendriks G.J., Kasper M., Sandberg R. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. *PLoS Comput Biol*. 2021;17(3):e1008772. doi 10.1371/journal.pcbi.1008772

Leibowitz M.L., Papathanasiou S., Doerfler P.A., Blaine L.J., Sun L., Yao Y., Zhang C.Z., Weiss M.J., Pellman D. Chromothripsis as an on-target consequence of CRISPR–Cas9 genome editing. *Nat Genet*. 2021;53(6):895-905. doi 10.1038/s41588-021-00838-7

Lejeune J., Gautier M., Turpin R. Study of somatic chromosomes from 9 mongoloid children. *C R Hebd Seances Acad Sci*. 1959;248(11):1721-1722

Letourneau A., Santoni F., Bonilla X., Sailani M.R., Gonzalez D., Kind J., Chevalier C., Thurman R., Sandstrom R.S., Hibaoui Y., Garieri M., Popadin K., Falconnet E., Gagnebin M., Gehrig C., Vannier A., Guipponi M., Farinelli L., Robyr D., Migliavacca E., Borel C., Deutsch S., Feki A., Stamatoyannopoulos J.A., Herault Y., van Steensel B., Guigo R., Antonarakis S.E. Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature*. 2014;508(7496):345-350. doi 10.1038/nature13200

Li G., Fullwood M.J., Xu H., Mulawadi F.H., Velkov S., Vega V., Ariyaratne P.N., Mohamed Y.B., Ooi H.S., Tennakoon C., Wei C.L., Ruan Y., Sung W.K. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol*. 2010;11(2):R22. doi 10.1186/gb-2010-11-2-r22

Li R., Zhu J. Effects of aneuploidy on cell behaviour and function. *Nat Rev Mol Cell Biol*. 2022;23(4):250-265. doi 10.1038/s41580-021-00436-9

Liao W.W., Asri M., Ebler J., Doerr D., Haukness M., Hickey G., Lu S., … Garrison E., Marschall T., Hall I.M., Li H., Paten B. A draft human pangenome reference. *Nature*. 2023;617(7960):312-324. doi 10.1038/s41586-023-05896-x

Machiela M.J., Zhou W., Karlins E., Sampson J.N., Freedman N.D., Yang Q., Hicks B., … Rothman N., Tucker M., Dean M.C., Yeager M., Chanock S.J. Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat Commun*. 2016;7:11843. doi 10.1038/ncomms11843

Mallick S., Li H., Lipson M., Mathieson I., Gymrek M., Racimo F., Zhao M., … Thangaraj K., Pääbo S., Kelso J., Patterson N., Reich D.

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016;538(7624):201-206. doi 10.1038/nature18964

Mayer R., Brero A., von Hase J., Schroeder T., Cremer T., Dietzel S. Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *BMC Cell Biol*. 2005;6:44. doi 10.1186/1471-2121-6-44

McConnell M.J., Moran J.V., Abyzov A., Akbarian S., Bae T., Cortes-Ciriano I., Erwin J.A., … Kidd J.M., Park P.J., Pevsner J., Vaccarino F.M.; Brain Somatic Mosaicism Network. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science*. 2017;356(6336):eaal1641. doi 10.1126/science.aal1641

Meaburn K.J., Misteli T. Cell biology: chromosome territories. *Nature*. 2007;445(7126):379-781. doi 10.1038/445379a

Meharena H.S., Marco A., Dileep V., Lockshin E.R., Akatsu G.Y., Mullahoo J., Watson L.A., Ko T., Guerin L.N., Abdurrob F., Rengarajan S., Papanastasiou M., Jaffe J.D., Tsai L.H. Down-syndrome-induced senescence disrupts the nuclear architecture of neural progenitors. *Cell Stem Cell*. 2022;29(1):116-130.e7. doi 10.1016/j.stem.2021.12.002

Miga K.H. From complete genomes to pangenomes. *Am J Hum Genet*. 2024;111(7):1265-1268. doi 10.1016/j.ajhg.2024.05.012

Minks J., Robinson W.P., Brown C.J. A skewed view of X chromosome inactivation. *J Clin Invest*. 2008;118(1):20-23. doi 10.1172/JCI34470

Modi D., Berde P., Bhartiya D. Down syndrome: a study of chromosomal mosaicism. *Reprod Biomed Online*. 2003;6(4):499-503. doi 10.1016/s1472-6483(10)62174-8

Muenzner J., Trébulle P., Agostini F., Zauber H., Messner C.B., Steger M., Kilian C., Lau K., Barthel N., Lehmann A., Textoris-Taube K., Caudal E., Egger A.S., Amari F., De Chiara M., Demichev V., Gossmann T.I., Mülleder M., Liti G., Schacherer J., Selbach M., Berman J., Ralser M. Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature*. 2024;630(8015):149-157. doi 10.1038/s41586-024-07442-9

Nagano T., Lubling Y., Stevens T.J., Schoenfelder S., Yaffe E., Dean W., Laue E.D., Tanay A., Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 2013;502(7469):59-64. doi 10.1038/nature12593

Nawata H., Kashino G., Tano K., Daino K., Shimada Y., Kugoh H., Oshimura M., Watanabe M. Dysregulation of gene expression in the artificial human trisomy cells of chromosome 8 associated with transformed cell phenotypes. *PLoS One*. 2011;6(9):e25319. doi 10.1371/journal.pone.0025319

Noormohammadi A., Calculli G., Gutierrez-Garcia R., Khodakarami A., Koyuncu S., Vilchez D. Mechanisms of protein homeostasis (proteostasis) maintain stem cell identity in mammalian pluripotent stem cells. *Cell Mol Life Sci*. 2018;75(2):275-290. doi 10.1007/s00018-017-2602-1

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., … Zook J.M., Schatz M.C., Eichler E.E., Miga K.H., Phillippy A.M. The complete sequence of a human genome. *Science*. 2022;376(6588):44-53. doi 10.1126/science.abj6987

Oji A., Choubani L., Miura H., Hiratani I. Structure and dynamics of nuclear A/B compartments and subcompartments. *Curr Opin Cell Biol*. 2024;90:102406. doi 10.1016/j.ceb.2024.102406

Olson L.E., Richtsmeier J.T., Leszl J., Reeves R.H. A chromosome 21 critical region does not cause specific Down syndrome phenotypes. *Science*. 2004;306(5696):687-690. doi 10.1126/science.1098992

Patau K., Smith D.W., Therman E., Inhorn S.L., Wagner H.P. Multiple congenital anomaly caused by an extra autosome. *Lancet*. 1960;1(7128):790-793. doi 10.1016/s0140-6736(60)90676-0

Pavelka N., Rancati G., Zhu J., Bradford W.D., Saraf A., Florens L., Sanderson B.W., Hattem G.L., Li R. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*. 2010;468(7321):321-325. doi 10.1038/nature09529

Pavlicek J., Soucek O., Vrtel R., Klaskova E., Hana V., Stara V., Adamova K., Fürst T., Hana V., Jr., Kapralova S., Prochazka M., Snajderova M., Tomaskova H., Tüdös Z., Vrbicka D., Vrtel P., Zapletalova J., Tauber Z., Lebl J. Karyotyping of lymphocytes and epithelial cells of distinct embryonic origin does not help to predict the Turner syndrome features. *Horm Res Paediatr*. 2022;95(5):465-475. doi 10.1159/000525823

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C., Zhai Y., Dairkee S.H., Ljung B.M., Gray J.W., Albertson D.G. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*. 1998;20(2):207-211. doi 10.1038/2524

Pires J.C., Conant G.C. Robust yet fragile: expression noise, protein misfolding, and gene dosage in the evolution of genomes. *Annu Rev Genet*. 2016;50:113-131. doi 10.1146/annurev-genet-120215-035400

Ramsköld D., Hendriks G.J., Larsson A.J.M., Mayr J.V., Ziegenhain C., Hagemann-Jensen M., Hartmanis L., Sandberg R. Single-cell new RNA sequencing reveals principles of transcription at the resolution of individual bursts. *Nat Cell Biol*. 2024;26(10):1725-1733. doi 10.1038/s41556-024-01486-9

Rana B., Lambrese K., Mendola R., Xu J., Garrisi J., Miller K., Marin D., Treff N.R. Identifying parental and cell-division origins of aneuploidy in the human blastocyst. *Am J Hum Genet*. 2023;110(4):565-574. doi 10.1016/j.ajhg.2023.03.003

Razin S.V., Ulianov S.V. Genome-directed cell nucleus assembly. *Biology*. 2022;11(5):708. doi 10.3390/biology11050708

Raznahan A., Parikshak N.N., Chandran V., Blumenthal J.D., Clasen L.S., Alexander-Bloch A.F., Zinn A.R., Wangsa D., Wise J., Murphy D.G.M., Bolton P.F., Ried T., Ross J., Giedd J.N., Geschwind D.H. Sex-chromosome dosage effects on gene expression in humans. *Proc Natl Acad Sci USA*. 2018;115(28):7398-7403. doi 10.1073/pnas.1802889115

Rehen S.K., Yung Y.C., McCreight M.P., Kaushal D., Yang A.H., Almeida B.S., Kingsbury M.A., Cabral K.M., McConnell M.J., Anliker B., Fontanoz M., Chun J. Constitutional aneuploidy in the normal human brain. *J Neurosci*. 2005;25(9):2176-2180. doi 10.1523/JNEUROSCI.4560-04.2005

Reinius B., Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat Rev Genet*. 2015;16(11):653-664. doi 10.1038/nrg3888

Rustchenko E. Chromosome instability in *Candida albicans*. *FEMS Yeast Res*. 2007;7(1):2-11. doi 10.1111/j.1567-1364.2006.00150.x

Rutledge S.D., Cimini D. Consequences of aneuploidy in sickness and in health. *Curr Opin Cell Biol*. 2016;40:41-46. doi 10.1016/j.ceb.2016.02.003

Santaguida S., Vasile E., White E., Amon A. Aneuploidy-induced cellular stresses limit autophagic degradation. *Genes Dev*. 2015;29(19):2010-2021. doi 10.1101/gad.269118.115

Santoni F.A., Stamoulis G., Garieri M., Falconnet E., Ribaux P., Borel C., Antonarakis S.E. Detection of imprinted genes by single-cell allele-specific gene expression. *Am J Hum Genet*. 2017;100(3):444-453. doi 10.1016/j.ajhg.2017.01.028

Shah P.P., Keough K.C., Gjoni K., Santini G.T., Abdill R.J., Wickramasinghe N.M., Dundes C.E., Karnay A., Chen A., Salomon R.E.A., Walsh P.J., Nguyen S.C., Whalen S., Joyce E.F., Loh K.M., Dubois N., Pollard K.S., Jain R. An atlas of lamina-associated chromatin across twelve human cell types reveals an intermediate chromatin subtype. *Genome Biol*. 2023;24(1):16. doi 10.1186/s13059-023-02849-5

Shao Y., Lu N., Wu Z., Cai C., Wang S., Zhang L.-L., Zhou F., Xiao S., Liu L., Zeng X., Zheng H., Yang C., Zhao Z., Zhao G., Zhou J.-Q., Xue X., Qin Z. Creating a functional single-chromosome yeast. *Nature*. 2018;560(7718):331-335. doi 10.1038/s41586-018-0382-x

Sheltzer J.M., Torres E.M., Dunham M.J., Amon A. Transcriptional consequences of aneuploidy. *Proc Natl Acad Sci USA*. 2012;109(31):12644-12649. doi 10.1073/pnas.1209227109

Solovei I., Wang A.S., Thanisch K., Schmidt C.S., Krebs S., Zwerger M., Cohen T.V., Devys D., Foisner R., Peichl L., Herrmann H., Blum H., Engelkamp D., Stewart C.L., Leonhardt H., Joffe B. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell*. 2013;152(3):584-598. doi 10.1016/j.cell.2013.01.009

Stamoulis G., Garieri M., Makrythanasis P., Letourneau A., Guipponi M., Panousis N., Sloan-Béna F., Falconnet E., Ribaux P., Borel C., Santoni F., Antonarakis S.E. Single cell transcriptome in aneuploidies reveals mechanisms of gene dosage imbalance. *Nat Commun*. 2019;10:4495. doi 10.1038/s41467-019-12273-8

Starostik M.R., Sosina O.A., McCoy R.C. Single-cell analysis of human embryos reveals diverse patterns of aneuploidy and mosaicism. *Genome Res*. 2020;30(6):814-825. doi 10.1101/gr.262774.120

Sterkers Y., Lachaud L., Crobu L., Bastien P., Pagès M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell Microbiol*. 2011;13(2):274-283. doi 10.1111/j.1462-5822.2010.01534.x

Sterkers Y., Lachaud L., Bourgeois N., Crobu L., Bastien P., Pagès M. Novel insights into genome plasticity in eukaryotes: mosaic aneuploidy in *Leishmania*. *Mol Microbiol*. 2012;86(1):15-23. doi 10.1111/j.1365-2958.2012.08185.x

Stingele S., Stoehr G., Peplowska K., Cox J., Mann M., Storchova Z. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol*. 2012;8:608. doi 10.1038/msb.2012.40

Storchová Z. Cells cope with altered chromosome numbers by enhancing protein breakdown. *Nature*. 2024;630(8015):41-43. doi 10.1038/d41586-024-01360-6

Sullivan K.D., Lewis H.C., Hill A.A., Pandey A., Jackson L.P., Cabral J.M., Smith K.P., Liggett L.A., Gomez E.B., Galbraith M.D., DeGregori J., Espinosa J.M. Trisomy 21 consistently activates the interferon response. *eLife*. 2016;5:e16220. doi 10.7554/eLife.16220

Tallaksen H.B.L., Johannsen E.B., Just J., Viuff M.H., Gravholt C.H., Skakkebæk A. The multi-omic landscape of sex chromosome abnormalities: current status and future directions. *Endocr Connect*. 2023;12(9):e230011. doi 10.1530/EC-23-0011

Tanabe H., Müller S., Neusser M., von Hase J., Calcagno E., Cremer M., Solovei I., Cremer C., Cremer T. Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci USA*. 2002;99(7):4424-4429. doi 10.1073/pnas.072618599

Tang Y.C., Amon A. Gene copy-number alterations: a cost-benefit analysis. *Cell*. 2013;152(3):394-405. doi 10.1016/j.cell.2012.11.043

Taylor A.M.R., Rothblum-Oviatt C., Ellis N.A., Hickson I.D., Meyer S., Crawford T.O., Smogorzewska A., Pietrucha B., Weemaes C., Stewart G.S. Chromosome instability syndromes. *Nat Rev Dis Primers*. 2019;5(1):64. doi 10.1038/s41572-019-0113-0

The 1000 Genomes Project Consortium; Auton A., Brooks L.D., Durbin R.M., Garrison E.P., Kang H.M., Korbel J.O., Marchini J.L., McCarthy S., McVean G.A., Abecasis G.R. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi 10.1038/nature15393

Thomas R., Marks D.H., Chin Y., Benezra R. Whole chromosome loss and associated breakage–fusion–bridge cycles transform mouse tetraploid cells. *EMBO J*. 2018;37(2):201-218. doi 10.15252/embj.201797630

Thompson S.L., Bakhoum S.F., Compton D.A. Mechanisms of chromosomal instability. *Curr Biol*. 2010;20(6):R285-R295. doi 10.1016/j.cub.2010.01.034

Torres E.M. Consequences of gaining an extra chromosome. *Chromosome Res*. 2023;31(3):24. doi 10.1007/s10577-023-09732-w

Torres E.M., Sokolsky T., Tucker C.M., Chan L.Y., Boselli M., Dunham M.J., Amon A. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science*. 2007;317(5840):916-924. doi 10.1126/science.1142210

Truong M.A., Cané-Gasull P., Lens S.M.A. Modeling specific aneuploidies: from karyotype manipulations to biological insights. *Chromosome Res*. 2023;31(3):25. doi 10.1007/s10577-023-09735-7

van Berkum N.L., Lieberman-Aiden E., Williams L., Imakaev M., Gnirke A., Mirny L.A., Dekker J., Lander E.S. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*. 2010;39:1869. doi 10.3791/1869

van Steensel B., Belmont A.S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*. 2017;169(5):780-791. doi 10.1016/j.cell.2017.04.022

Veitia R.A., Potier M.C. Gene dosage imbalances: action, reaction, and models. *Trends Biochem Sci*. 2015;40(6):309-317. doi 10.1016/j.tibs.2015.03.011

Veitia R.A., Bottani S., Birchler J.A. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends Genet*. 2013;29(7):385-393. doi 10.1016/j.tig.2013.04.004

Vernot B., Tucci S., Kelso J., Schraiber J.G., Wolf A.B., Gittelman R.M., Dannemann M., Grote S., McCoy R.C., Norton H., Scheinfeldt L.B., Merriwether D.A., Koki G., Friedlaender J.S., Wakefield J., Pääbo S., Akey J.M. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*. 2016;352(6282):235-239. doi 10.1126/science.aad9416

Wang S., Leng L., Wang Q., Gu Y., Li J., An Y., Deng Q., Xie P., Cheng C., Chen X., Zhou Q., Lu J., Chen F., Liu L., Yang H., Wang J., Xu X., Hou Y., Gong F., Hu L., Lu G., Shang Z., Lin G. A single-cell transcriptome atlas of human euploid and aneuploid blastocysts. *Nat Genet*. 2024;56(7):1468-1481. doi 10.1038/s41588-024-01788-6

Wang Y., Qu Z., Fang Y., Chen Y., Peng J., Song J., Li J., Shi J., Zhou J.-Q., Zhao Y. Chromosome territory reorganization through artificial chromosome fusion is compatible with cell fate determination and mouse development. *Cell Discov*. 2023;9(1):11. doi 10.1038/s41421-022-00511-1

Werner J.M., Hover J., Gillis J. Population variability in X-chromosome inactivation across 10 mammalian species. *Nat Commun*. 2024;15(1):8991. doi 10.1038/s41467-024-53449-1

Williams B.R., Amon A. Aneuploidy: cancer's fatal flaw? *Cancer Res*. 2009;69(13):5289-5291. doi 10.1158/0008-5472.CAN-09-0944

Williams B.R., Prabhu V.R., Hunter K.E., Glazier C.M., Whittaker C.A., Housman D.E., Amon A. Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science*. 2008;322(5902):703-709. doi 10.1126/science.1160058

Xiao R., Xu D., Zhang M., Chen Z., Cheng L., Du S., Lu M., Zhou T., Li R., Bai F., Huang Y. Aneuploid embryonic stem cells drive teratoma metastasis. *Nat Commun*. 2024;15(1):8883. doi 10.1038/s41467-024-53288-0

Yurov Y.B., Vorsanova S.G., Iourov I.Y., Demidova I.A., Beresheva A.K., Kravetz V.S., Monakhov V.V., Kolotii A.D., Voinova-Ulas V.Y., Gorbachevskaya N.L. Unexplained autism is frequently associated with low-level mosaic aneuploidy. *J Med Genet*. 2007;44(8):521-525. doi 10.1136/jmg.2007.049312

Zadesenets K.S., Rubtsov N.B. B chromosomes in free-living flatworms of the genus *Macrostomum* (Platyhelminthes, Macrostomorpha). *Int J Mol Sci*. 2021;22(24):13617. doi 10.3390/ijms222413617

Zadesenets K.S., Vizoso D.B., Schlatter A., Konopatskaia I.D., Berezikov E., Schärer L., Rubtsov N.B. Evidence for karyotype polymorphism in the free-living flatworm, *Macrostomum lignano*, a model organism for evolutionary and developmental biology. *PLoS One*. 2016;11(10):e0164915. doi 10.1371/journal.pone.0164915

Zadesenets K.S., Ershov N.I., Berezikov E., Rubtsov N.B. Chromosome evolution in the free-living flatworms: first evidence of intrachromosomal rearrangements in karyotype evolution of *Macrostomum lignano* (Platyhelminthes, Macrostomida). *Genes* (*Basel*). 2017a;8(11):298. doi 10.3390/genes8110298

Zadesenets K.S., Schärer L., Rubtsov N.B. New insights into the karyotype evolution of the free-living flatworm *Macrostomum lig-*

nano (Platyhelminthes, Turbellaria). *Sci Rep*. 2017b;7(1):6066. doi 10.1038/s41598-017-06498-0

Zadesenets K.S., Jetybayev I.Y., Schärer L., Rubtsov N.B. Genome and karyotype reorganization after whole genome duplication in free-living flatworms of the genus *Macrostomum*. *Int J Mol Sci*. 2020; 21(2):680. doi 10.3390/ijms21020680

Zadesenets K.S., Ershov N.I., Bondar N.P., Rubtsov N.B. Unraveling the unusual subgenomic organization in the neopolyploid free-living flatworm *Macrostomum lignano*. *Mol Biol Evol*. 2023;40(12): msad250. doi 10.1093/molbev/msad250

Zhai F., Kong S., Song S., Guo Q., Ding L., Zhang J., Wang N., Kuo Y., Guan S., Yuan P., Yan L., Yan Z., Qiao J. Human embryos harbor complex mosaicism with broad presence of aneuploid cells during early development. *Cell Discov*. 2024;10(1):98. doi 10.1038/s41421-024-00719-3

Zhang F., Gu W., Hurles M.E., Lupski J.R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451-481. doi 10.1146/annurev.genom.9.081307.164217

Zhang X.M., Yan M., Yang Z., Xiang H., Tang W., Cai X., Wu Q., Liu X., Pei G., Li J. Creation of artificial karyotypes in mice reveals robustness of genome organization. *Cell Res*. 2022;32(11):1026-1029. doi 10.1038/s41422-022-00722-x

Zhegalova I.V., Vasiluev P.A., Flyamer I.M., Shtompel A.S., Glazyrina E., Shilova N., Minzhenkova M., Markova Z., Petrova N.V., Dashinimaev E.B., Razin S.V., Ulianov S.V. Trisomies reorganize human 3D genome. *Int J Mol Sci*. 2023;24(22):6044. doi 10.3390/ijms242216044

Conflict of interest. The authors declare no conflict of interest.

Received November 19, 2024. Revised December 6, 2024. Accepted December 9, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):349-357

doi 10.18699/vjgb-25-38

Generation of the ICGi019-B-1 and ICGi019-B-2 lines via correction of the p.Met659Ile (c.1977G>A) variant in *MYH7* of patient-specific induced pluripotent stem cells using CRISPR/Cas9

A.E. Shulgina, S.V. Pavlova, J.M. Minina, S.M. Zakian, E.V. Dementyeva

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
dementyeva@bionet.nsc.ru

Abstract. The problem of interpretation of the genetic data from patients with inherited cardiovascular diseases still remains relevant. To date, the clinical significance of approximately 40 % of variants in genes associated with inherited cardiovascular diseases is uncertain, which requires new approaches to the assessment of their pathogenetic contribution. A combination of the induced pluripotent stem cell (iPSC) technology and editing the iPSC genome with CRISPR/Cas9 is thought to be the most promising tool for clarifying variant pathogenicity. A variant of unknown significance in *MYH7*, p.Met659Ile (c.1977G>A), was previously identified in several genetic screenings of hypertrophic cardiomyopathy patients. In this study, the single nucleotide substitution was corrected with CRISPR/Cas9 in iPSCs generated from a carrier of the variant. As a result, two iPSC lines (ICGi019-B-1 and ICGi019-B-2) were generated and characterized using a standard set of methods. The iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* possessed a morphology characteristic of human pluripotent cells, expressed markers of the pluripotent state (the OCT4, SOX2, NANOG transcription factors and SSEA-4 surface antigen), were able to give rise to derivatives of three germ layers during spontaneous differentiation, and retained a normal karyotype (46,XY). No CRISPR/Cas9 off-target activity was found in the ICGi019-B-1 and ICGi019-B-2 iPSC lines. The maintenance of the pluripotent state and normal karyotype and the absence of CRISPR/Cas9 off-target activity in the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* allow using the iPSC lines as an isogenic control for further studies of the variant pathogenicity and its impact on the hypertrophic cardiomyopathy development.

Key words: hypertrophic cardiomyopathy; variants of unknown significance; induced pluripotent stem cells; CRISPR/Cas9

For citation: Shulgina A.E., Pavlova S.V., Minina J.M., Zakian S.M., Dementyeva E.V. Generation of the ICGi019-B-1 and ICGi019-B-2 lines via correction of the p.Met659Ile (c.1977G>A) variant in *MYH7* of patient-specific induced pluripotent stem cells using CRISPR/Cas9. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):349-357. doi 10.18699/vjgb-25-38

Funding. This study was supported by the Russian Science Foundation, grant No. 22-15-00271.

Acknowledgements. Karyotype analysis was conducted using resources of the Common Facilities Centre of Microscopic Analysis of Biological Objects supported by the State project of ICG SB RAS (FWNR-2022-0015).

Получение линий ICGi019-B-1 и ICGi019-B-2 посредством исправления с помощью системы CRISPR/Cas9 варианта p.Met659Ile (c.1977G>A) в гене *MYH7* в пациент-специфичных индуцированных плюрипотентных стволовых клетках

А.Е. Шульгина, С.В. Павлова, Ю.М. Минина, С.М. Закиян, Е.В. Дементьева

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
dementyeva@bionet.nsc.ru

Аннотация. Проблема интерпретации результатов генетического анализа пациентов, страдающих наследственными сердечно-сосудистыми заболеваниями, по-прежнему сохраняет свою актуальность. На сегодняшний день клиническое значение около 40 % вариантов в генах, ассоциированных с наследственными сердечно-сосудистыми заболеваниями, остается неясным, что приводит к необходимости использования новых подходов для оценки патогенетического вклада этих вариантов. Совместное применение технологии индуцированных плюрипотентных стволовых клеток и редактирования их генома с помощью системы CRISPR/Cas9 считается наиболее перспективным способом выяснения патогенности генетических вариантов. Ранее в нескольких генетических скринингах пациентов с гипертрофической кардиомиопатией был выявлен

вариант с неясным клиническим значением в гене *MYH7*, p.Met659Ile (c.1977G>A). В настоящем исследовании данная однонуклеотидная замена с помощью системы CRISPR/Cas9 была исправлена в индуцированных плюрипотентных стволовых клетках, полученных от носителя этого генетического варианта. В результате получены и охарактеризованы с использованием стандартного набора методов две линии индуцированных плюрипотентных стволовых клеток (ICGi019-B-1 и ICGi019-B-2). Линии индуцированных плюрипотентных стволовых клеток с исправленным вариантом p.Met659Ile (c.1977G>A) в гене *MYH7* имели характерную для плюрипотентных клеток человека морфологию, экспрессировали маркеры плюрипотентного состояния (транскрипционные факторы OCT4, SOX2, NANOG и поверхностный антиген SSEA-4), были способны давать производные трех зародышевых листков при спонтанной дифференцировке и сохраняли нормальный кариотип (46,XY). В линиях индуцированных плюрипотентных стволовых клеток ICGi019-B-1 и ICGi019-B-2 не обнаружено нецелевой активности системы CRISPR/Cas9. Поддержание плюрипотентного состояния и нормального кариотипа, а также отсутствие нецелевой активности системы CRISPR/Cas9 в линиях индуцированных плюрипотентных стволовых клеток с исправленным вариантом p.Met659Ile (c.1977G>A) в гене *MYH7* позволят использовать полученные линии в качестве изогенного контроля для дальнейших исследований патогенности данного генетического варианта и его влияния на развитие гипертрофической кардиомиопатии.
Ключевые слова: гипертрофическая кардиомиопатия; варианты с неясным клиническим значением; индуцированные плюрипотентные стволовые клетки; CRISPR/Cas9

Introduction

Generation of induced pluripotent stem cells (iPSCs) and their subsequent differentiation into cardiomyocytes is an important tool for modeling, studying, and developing therapy methods for inherited cardiovascular diseases (Parrotta et al., 2020; Funakoshi, Yoshida, 2021; Gähwiler et al., 2021). A combined use of the iPSC-based technology and genome editing methods, e. g. CRISPR/Cas9, also allows generating the so-called isogenic iPSCs by introducing a certain variant into iPSCs of a healthy donor or its correction in patient-specific iPSCs. Examination of cardiomyocytes derived from the isogenic iPSCs can overcome the challenge caused by numerous variants of unknown significance found in genetic screenings of patients with cardiovascular diseases (Guo H. et al., 2021).

Hypertrophic cardiomyopathy (HCM) is one of the most widespread inherited cardiovascular pathologies (overall prevalence is 0.2 %). The disease manifestations include an asymmetric thickness of the left ventricular walls and the interventricular septum, left ventricular outflow tract obstruction, progressive heart failure, and a high risk of atrial or ventricular arrhythmias and sudden cardiac death (Geske et al., 2018). HCM-causing variants can be found in genes encoding proteins involved in sarcomere functioning and regulation of calcium homeostasis. There are also HCM phenocopies that are due to variants in genes associated with metabolic disorders, neuromuscular diseases, and RASopathies. However, the majority of HCM-causing variants (about 80 %) have been found in *MYH7* and *MYBPC3*, encoding sarcomere proteins: β-myosin heavy chain and myosin-binding protein C, respectively (Akhtar, Elliott, 2018; Pasipoularides, 2018).

A variant of unknown significance, p.Met659Ile (c.1977G>A, rs1241603111) in *MYH7*, was found in a number of genetic screenings of HCM patients (Richard et al., 2003; Bashyam et al., 2012; Dementyeva et al., 2020a). The single nucleotide substitution is a rare variant with no frequency in gnomAD v4.1.0 (https://gnomad.broadinstitute.org/) and is located in the actin-binding site of the myosin motor domain that is highly conservative in vertebrates (Hesaraki et al., 2022). The variant is predicted to be pathogenic by multiple *in silico* tools and AlphaMissense (Dementyeva et al., 2020a; Cheng et al., 2023; Pavlova et al., 2024). However, according to the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar), the available evidence is not sufficient to determine the role of the variant in HCM development. Therefore, functional studies are required to find out the pathogenicity of the variant.

In our previous study, CRISPR/Cas9 was used to introduce the variant into iPSCs of a healthy donor (Malakhova et al., 2020) and an iPSC line heterozygous at the single nucleotide substitution was generated. The cardiomyocytes derived from the iPSC line with the introduced p.Met659Ile (c.1977G>A) variant in *MYH7* were characterized by an increased size, an elevated diastolic calcium level, changes in the expression of HCM-related genes, and a decrease in basal oxygen consumption rate compared to the isogenic control, which indicates the pathogenicity of the variant (Pavlova et al., 2024). However, it would be useful to verify the effects of the variant on the cardiomyocyte properties under another genetic background. This study was aimed at correction of the variant in an iPSC line derived earlier from an HCM patient carrying the p.Met659Ile (c.1977G>A) variant in *MYH7* (Dementyeva et al., 2020b) and generation of the second panel of isogenic iPSC lines for studying the impact of the variant on HCM development.

Materials and methods

iPSC lines used. ICGi019-B (https://hpscreg.eu/cell-line/ICGi019-B), an iPSC line derived from an HCM patient who was a carrier of the p.Met659Ile (c.1977G>A) variant in *MYH7* (Dementyeva et al., 2020b). ICGi022-A (https://hpscreg.eu/cell-line/ICGi022-A), an iPSC line derived from a healthy donor (Malakhova et al., 2020).

iPSC cultivation. iPSC lines were cultured at 37 °C in 5 % CO_2 on a layer of mitotically inactivated mouse embryonic fibroblasts (feeder) in KnockOut DMEM supplemented with 15 % KnockOut Serum Replacement, 0.1 mM MEM Non-Essential Amino Acids Solution, 1× penicillin-streptomycin, 1 mM GlutaMAX (all reagents – Thermo Fisher Scientific), 0.05 mM 2-mercaptoethanol (Amresco), and 10 ng/mL bFGF (SCI-store). The iPSC lines were passaged with TrypLE™ Express Enzyme (Thermo Fisher Scientific) at a ratio of 1:10 every 4–5 days.

Correction of the p.Met659Ile (c.1977G>A) variant in *MYH7* of patient-specific iPSCs. 100 pmol of single-guide RNA (Synthego, Table 1) and 20 pmol of Cas9_NLS (NEB) were incubated for 20 min at room temperature. The ribonu-

Table 1. Oligonucleotides and antibodies used in this study

Oligonucleotides			
Application	Gene/locus	Product size	Nucleotide sequence (5′–3′)
Protospacer for single-guide RNA	*MYH7*, exon 18	20 b	GGGATGGGTGGAGCGCAAGT
Single-stranded donor oligonucleotide	*MYH7*, exon 18	89 b	TATTGCATTTTTGGCCACAGGAAAATCTGAACAAGCTGAT GACAAACTTGCGCTCCACCCATCCCCACTTTGTACGTTGTA TCATCCCT
Analysis of editing events	*MYH7*, exon 18	258 bp	TCCTTCCTTCTTCTCCTCTCTT/ GTGGTGGTAGGTAGGGAGAT
Analysis of CRISPR/Cas9 off-target activity	chr4:14217409–14217428	559 bp	TCTGGTAAGAGCCTGACTTCTG/ TCCCACCTGCCATTGGAATA
	chr7:57819587–57819606	378 bp	ACGATACTCAAGGCCCAATCT/ TGGTGTTTCCTCATCCTGGT
	chr8:139343032–139343051	535 bp	GCCAGGAAAGTTCAGTGGTTAG/ CCCTCTCTCTTCCTGCTCTTAT
	chr20:9996001–9996020	575 bp	GACTTGTAATAACTCTCACTCACCTAAA/ CCAGGCAATGTTAAGCCTTCAT
	chr2:241709069–241709088	541 bp	TCCCGTGTGGATTTCTTTAGGT/ TGTAGGCGTTCTGGATCTTCTG
Pluripotency markers (RT-qPCR)	*NANOG*	116 bp	TTTGTGGGCCTGAAGAAAACT/ AGGGCTGTCCTGAATAAGCAG
	SOX2	100 bp	GCTTAGCCTCGTCGATGAAC/ AACCCCAAGATGCACAACTC
Reference gene (RT-qPCR)	*B2M*	90 bp	TAGCTGTGCTCGCGCTACT/ TCTCTGCTGGATGACGTGAG
Mycoplasma detection	16S ribosomal RNA gene	280 bp	GGGAGCAAACAGGATTAGATACCCT/ TGCACCATCTGTCACTCTGTTAACCTC
Antibodies			
Application	Antibody	Dilution	Company, cat. no., RRID
Pluripotency markers	Mouse IgG2b anti-OCT3/4	1:50	Santa Cruz Biotechnology, sc-5279, RRID:AB_628051
	Rabbit IgG anti-SOX2	1:200	Cell Signaling Technology, 3579, RRID:AB_2195767
	Mouse IgG3 anti-SSEA-4	1:200	Abcam, ab16287, RRID: AB_778073
Markers of differentiated derivatives	Mouse IgG2a anti-TUBB3	1:500	BioLegend, 801201, RRID:AB_2313773
	Mouse IgG2a anti-αSMA	1:100	Dako, M0851, RRID:AB_2223500
	Mouse IgG1 anti-CK18	1:100	Abcam, ab668, RRID:AB_305647
Secondary antibodies	Goat anti-Mouse IgG (H + L) Secondary Antibody, Alexa Fluor 568	1:400	Thermo Fisher Scientific, A11031, RRID:AB_144696
	Goat anti-Rabbit IgG (H + L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	1:400	Thermo Fisher Scientific, A11008, RRID:AB_143165
	Goat anti-Mouse IgG3 Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	1:400	Thermo Fisher Scientific, A21151, RRID: AB_2535784
	Goat anti-Mouse IgG1 Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	1:400	Thermo Fisher Scientific, A21121, RRID:AB_2535764

cleoprotein complexes, together with 300 ng of single-stranded donor oligonucleotide (Biolegio, Table 1), were electroporated into 1×10^5 iPSCs of the ICGi019-B line on a Neon Transfection System (Thermo Fisher Scientific), using the program: 1100 V, 30 ms, 1 time. The electroporated cells were transferred to a feeder layer in the iPSC medium without antibiotic and supplemented with 10 ng/ml Y-27632 (Sigma-Aldrich). 48 h later, the cells were subcloned into 96-well plates. iPSC clones were cultivated as described in the previous section.

Analysis of editing events in *MYH7* and CRISPR/Cas9 off-target activity. Genomic DNA was isolated from the iPSC clones using Wizard® Genomic DNA Purification Kit (Promega). Genomic DNA regions contained exon 18 of *MYH7* or CRISPR/Cas9 off-target sites predicted using IDT (https://www.idtdna.com/) were amplified by PCR with BioMaster HS-Taq PCR-Color (2×) (Biolabmix) on a T100 Thermal Cycler (Bio-Rad), using the program: 95 °C – 3 min; 35 cycles: 95 °C – 30 s, 62 °C – 30 s, 72 °C – 30–40 s; 72 °C – 5 min. The primers used are listed in Table 1. Sanger sequencing of the PCR products was performed using the Big Dye Terminator V. 3.1. Cycle Sequencing Kit (Applied Biosystems) and analysis was conducted at the SB RAS Genomics Core Facility (http://www.niboch.nsc.ru/doku.php/corefacility).

Karyotype analysis. iPSC lines were plated at a ratio of 1:4 on a 12-well plate 48 h before metaphase collection. Four different concentrations of Colcemide (from 25 to 50 ng/mL) were added 2.5 h before metaphase collection. Cells were disaggregated with TrypLE™ Express Enzyme (Thermo Fisher Scientific). Hypotonic treatment was conducted for 20 min at 37 °C in 0.28 % KCl. Cells were fixed in Carnoy's solution (methanol–acetic acid 3:1) as described in (Sorogina et al., 2023). Karyotype of the iPSC lines was analyzed on Axioplan 2 (Zeiss) with the ISIS 5 program (MetaSystems). 50 metaphase plates were analyzed for the iPSC lines.

Spontaneous *in vitro* differentiation. iPSCs were treated for 40 min with 0.15 % Collagenase IV (Thermo Fisher Scientific). The resulting cell aggregates were transferred to Petri dishes coated with 1 % agarose and cultivated for 2 weeks in DMEM/F12 (1:1) medium supplemented with 15 % KnockOut Serum Replacement, 0.1 mM MEM Non-Essential Amino Acids Solution, 1× penicillin-streptomycin, 1 mM GlutaMAX (all reagents – Thermo Fisher Scientific). The embryoid bodies formed were plated on 8-well Chambered Coverglasses (Thermo Fisher Scientific) coated with Matrigel (Corning) and cultured for a week in the same medium. The medium was changed every 3 days. The differentiated derivatives were analyzed by immunofluorescence staining.

Immunofluorescence staining. iPSCs or their differentiated derivatives were fixed in 4 % paraformaldehyde (Sigma-Aldrich) for 10 min, permeabilized in 0.4 % Triton-X100 (Sigma-Aldrich) for 10 min, incubated with 1 % bovine serum albumin (VWR) for 30 min (all the steps were conducted at room temperature). In case of SSEA-4, cell treatment was carried out without the permeabilization step. Cells were incubated overnight at 4 °C with primary antibodies and for 1 h at room temperature with secondary antibodies. After each incubation with antibodies, the cells were washed with PBS twice for 15 min. The antibodies used are provided in Table 1. Nuclei were counterstained with DAPI (Sigma-Aldrich). Immunofluorescence staining was analyzed on a Nikon Eclipse Ti-E microscope with NIS Elements Advanced Research software version 4.30 (Nikon).

RT-qPCR. RNA was isolated from iPSC lines with TRIzol Reagent and treated using the Invitrogen™ DNA-free™ DNA Removal Kit (all reagents – Thermo Fisher Scientific). Reverse transcription of 1 μg of RNA was conducted with M-MuLV reverse transcriptase (Biolabmix). RT-qPCR was performed with BioMaster HS-qPCR SYBR Blue 2× (Biolabmix) on a QuantStudio™ 5 Real-Time PCR System (Applied Biosystems), using the program: 95 °C – 5 min; 40 cycles: 95 °C – 10 s, 60 °C – 1 min. CT values were normalized by the ΔΔCT method (Livak, Schmittgen, 2001) using *B2M* as a reference gene. The primers used are listed in Table 1.

Mycoplasma detection. Mycoplasma contamination in iPSC lines was detected by PCR with BioMaster HS-Taq PCR-Color (2×) (Biolabmix) on a T100 Thermal Cycler (Bio-Rad), using the program: 95 °C – 3 min; 35 cycles: 95 °C – 15 s, 67 °C – 15 s, 72 °C – 20 s; 72 °C – 5 min. The primers used are listed in Table 1.

Results

The p.Met659Ile (c.1977G>A) variant in *MYH7* was corrected in the patient-specific ICGi019-B iPSC line via introduction of a double-strand break with CRISPR/Cas9 and subsequent homology-directed repair with single-stranded donor oligonucleotide. The protospacer for the single-guide RNA and Protospacer Adjacent Motif (PAM) were designed to be located as close as possible to the target substitution and to introduce a synonymous substitution in PAM to protect *MYH7* from repetitive editing (Fig. 1*a*). Low off-target activity of the selected single-guide RNA was confirmed using Benchling (https://www.benchling.com/) and IDT (https://www.idtdna.com/). The single-stranded donor oligonucleotide was chosen to correspond to the reference nucleotide sequence of a part of *MYH7* intron 17 and exon 18 and contained the synonymous substitution (Fig. 1*a*).

CRISPR/Cas9 ribonucleoprotein complexes consisting of single-guide RNA and Cas9_NLS, together with the single-stranded donor oligonucleotide, were electroporated into cells of the ICGi019-B line. 84 iPSC clones were generated and analyzed using Sanger sequencing. In 71 (84.52 %) iPSC clones, editing events in *MYH7* exon 18 were found. Non-homologous end joining (indels) occurred in 54 (64.29 %) of the iPSC clones. Homology-directed repair with the single-stranded donor oligonucleotide accompanied by p.Met659Ile (c.1977G>A) variant correction was detected in 17 (20.23 %) iPSC clones. However, 9 iPSC clones with the homology-directed repair in the mutant allele also demonstrated off-target editing events (indels) in the second allele. Thus, 8 iPSC clones with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* have been generated (Fig. 1*b*).

For use in fundamental and applied studies, iPSC lines have to match a number of criteria, including the normal karyotype. Karyotype analysis of the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* showed that two iPSC lines, ICGi019-B-1 and ICGi019-B-2, retained the normal number and structure of chromosomes – 46,XY (Fig. 2*a*). The ICGi019-B-1 and ICGi019-B-2 iPSC lines were chosen for further characterization. No CRISPR/Cas9 off-target activity was found after comparison of the nucleotide

Fig. 1. Correction of the p.Met659Ile (c.1977G>A) variant in *MYH7* of the patient-specific iPSCs using CRISPR/Cas9.

a – design of single-guide RNA and single-stranded donor oligonucleotide for *MYH7* editing. The nucleotide sequence of a fragment of intron 17 and exon 18 is given. The positions of the protospacer for the single-guide RNA, PAM, and the single-stranded donor oligonucleotide are indicated in grey, magenta, and yellow, respectively. The target substitution and synonymous substitution in PAM are shown with red rectangles; *b* – an example of two iPSC clones with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* (ICGi019-B-1 and ICGi019-B-2). Nucleotide sequences of the same region in the patient-specific iPSC line (ICGi019-B) and the iPSC line of the healthy donor (ICGi022-A) are provided for comparison. The target substitution and synonymous substitution in PAM are shown with red rectangles; *c* – absence of CRISPR/Cas9 off-target activity at one of the predicted CRISPR/Cas9 off-target sites in the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* (ICGi019-B-1 and ICGi019-B-2). The nucleotide sequence of the same region in the patient-specific iPSC line used for *MYH7* editing (ICGi019-B) is given for comparison. The CRISPR/Cas9 off-target site and its positions in the human genome (hg38) are indicated in red.

sequences of the top-five CRISPR/Cas9 off-target sites and their surroundings in the iPSC lines and the patient-specific ICGi019-B line used for *MYH7* editing (Fig. 1*c*). Despite *MYH7* editing, the ICGi019-B-1 and ICGi019-B-2 iPSC lines retained their pluripotent properties. The iPSC lines possessed morphology similar to that of human pluripotent stem cells (Fig. 2*b*) and were characterized by expression of the pluripotent state markers such as the OCT4 and SOX2 transcription factors and SSEA-4 surface antigen (Fig. 2*c*). The expression level of pluripotency genes, *NANOG* and *SOX2*, in the ICGi019-B-1 and ICGi019-B-2 iPSC lines was demonstrated to be comparable to that in the original patient-

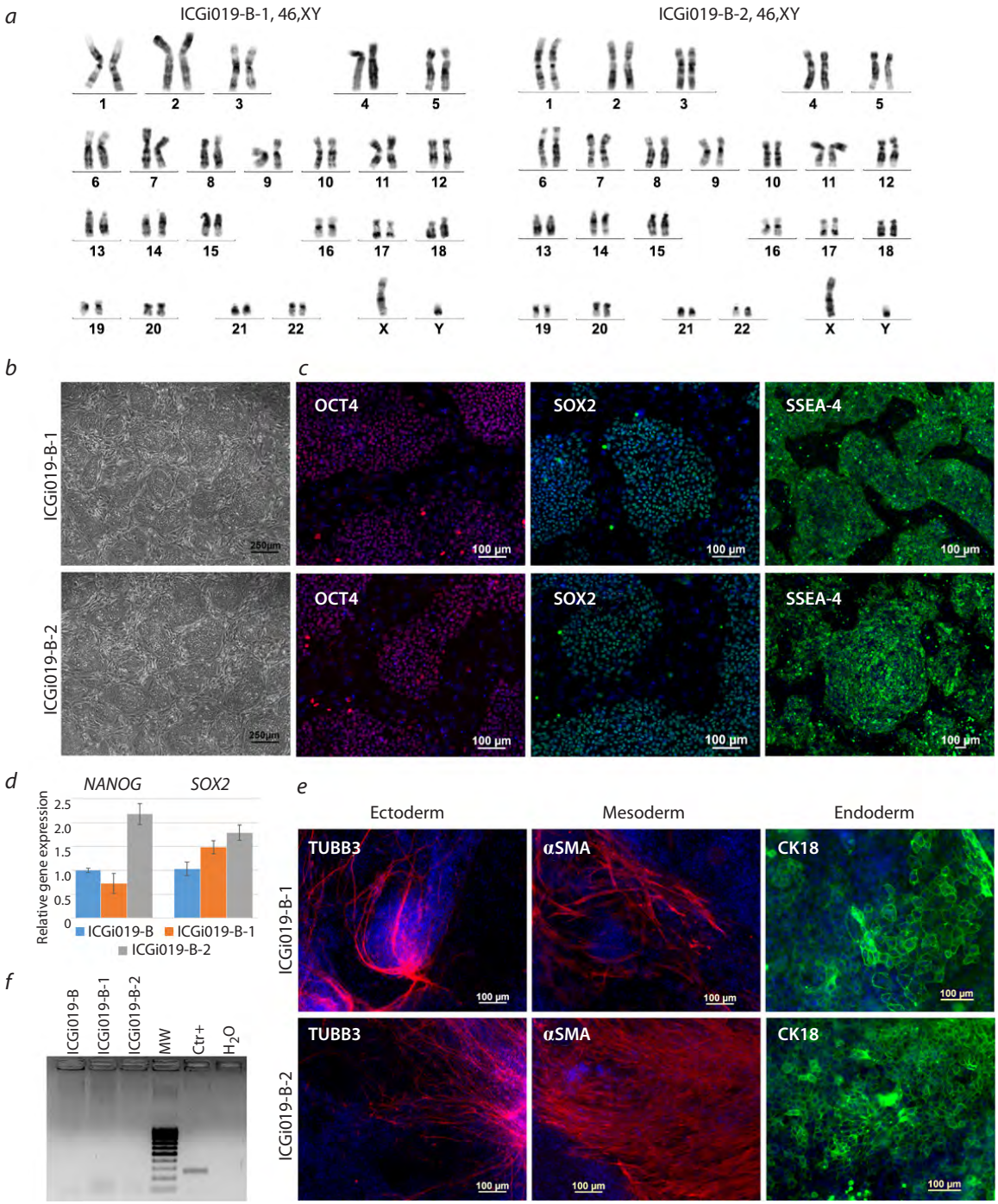


Fig. 2. Characterization of the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7*.

a – karyotype of the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* (ICGi019-B-1 and ICGi019-B-2); *b* – morphology of the ICGi019-B-1 and ICGi019-B-2 iPSC lines. Scale bar 250 µm; *c* – expression of the OCT4 and SOX2 transcription factors and SSEA-4 surface antigen in the ICGi019-B-1 and ICGi019-B-2 iPSC lines. Scale bar 100 µm; *d* – expression of pluripotency genes, *NANOG* and *SOX2*, in the ICGi019-B-1 and ICGi019-B-2 iPSC lines in comparison with the original patient-specific ICGi019-B line. Data are presented as mean ± SEM; *e* – capacity of the ICGi019-B-1 and ICGi019-B-2 iPSC lines to be differentiated into derivatives of three germ layers: ectoderm (TUBB3, β3-tubulin), mesoderm (αSMA, smooth muscle α-actin), and endoderm (CK18, cytokeratin 18). Scale bar 100 µm; *f* – absence of mycoplasma contamination in the ICGi019-B-1 and ICGi019-B-2 iPSC lines. Ctr+, positive control for mycoplasma contamination, H_2O, negative control for mycoplasma contamination.

specific iPSC line (Fig. 2*d*). The iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* were capable to be differentiated into derivatives of three germ layers as was shown by expression of markers of ectoderm, mesoderm, and endoderm in cells obtained under spontaneous differentiation of the iPSC lines in embryoid bodies (Fig. 2*e*). The iPSC lines were also free from mycoplasma contamination (Fig. 2*f*). The ICGi019-B-1 and ICGi019-B-2 iPSC lines were registered in the Human Pluripotent Stem Cell Registry (hPSCreg, https://hpscreg.eu/). Their passport is provided in Table 2.

Table 2. Passport of the ICGi019-B-1 and ICGi019-B-2 iPSC lines

Unique identifier	ICGi019-B-1 ICGi019-B-2
Alternative name of cell line	HCM1f33-wt119 HCM1f33-wt147
Institution	Federal Research Center Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
Type of cell line	iPSCs
Origin	Human
Additional origin information	Age: 38 Sex: M Ethnicity: Caucasian
Cell source	Peripheral blood mononuclear cells
Method of reprogramming	Transgene free episomal plasmid vectors
Reprogramming factors	OCT4, KLF4, L-MYC, SOX2, LIN28
Clonality	Clonal
Evidence for reprogramming transgene elimination	PCR, not detected
Genetic modification	Yes
Type of genetic modification	Correction of genetic variant
Disease	Hypertrophic cardiomyopathy
Gene/locus	*MYH7*:c.1977G>A (p.Met659Ile), rs1241603111
Method of modification/site-specific nuclease used	CRISPR/Cas9
Site-specific nuclease delivery method	Electroporation with ribonucleoprotein complexes of single-guide RNA and Cas9_NLS
Genetic material introduced into the cells	Single-stranded donor oligonucleotide corresponding to the reference nucleotide sequence and containing a synonymous substitution in PAM
Method of introduced modification analysis	Sanger sequencing of PCR-products containing *MYH7* exon 18
Method of off-target activity analysis	Sanger sequencing of PCR-products containing predicted CRISPR/Cas9 off-target sites
Morphology	Monolayer colonies similar to those of human pluripotent cells
Pluripotency	Confirmed by expression of the pluripotent state markers and spontaneous differentiation in derivatives of three germ layers
Karyotype	46,XY
Contaminations	Bacteria, fungi, and mycoplasma were not detected
Potential application	Studying pathogenetic contribution of the p.Met659Ile (c.1977G>A) variant in *MYH7* to hypertrophic cardiomyopathy development
Method of culturing	On a layer of mitotically inactivated mouse fibroblasts (feeder)
Medium	KnockOut DMEM (Thermo Fisher Scientific)
Temperature, °C	37
CO_2 concentration, %	5
O_2 concentration, %	20
Method of passaging	TrypLE™ Express Enzyme (Thermo Fisher Scientific)
Ratio for passaging	1:10
Cryoconservation	90 % FBS, 10 % DMSO
Storage conditions	Liquid nitrogen
Cell line repository/bank	https://hpscreg.eu/cell-line/ICGi019-B-1 https://hpscreg.eu/cell-line/ICGi019-B-2
Date archived/stock date	June 2024

Discussion

iPSC editing with CRISPR/Cas9 has been successfully applied for studying HCM pathogenetic mechanisms caused by known pathogenic variants in sarcomere protein genes (Mosqueira et al., 2018; Smith et al., 2018; Wang et al., 2018; Cohn et al., 2019; Bhagwan et al., 2020; Shafaattalab et al., 2021; Chai et al., 2023; Escribá et al., 2023; Guo G. et al., 2024) and deciphering the pathogenicity of several variants of uncertain significance in HCM-associated genes (Ma et al., 2018; Pavlova et al., 2024).

This study was devoted to generating iPSC lines via correction of a variant of unknown significance, p.Met659Ile (c.1977G>A) in *MYH7,* in patient-specific iPSCs using CRISPR/Cas9. The variant is localized in the actin-binding site of the myosin motor domain where 37 variants have been described according to the ClinVar database. 12 variants are pathogenic and likely pathogenic whereas the remaining 25 (67.6 %) variants have uncertain significance or conflicting classifications of pathogenicity. This fact emphasizes the problem of interpretation of genetic data in clinical practice and makes examining the impact of variants of unknown significance in the functionally important actin-binding region much more relevant.

We previously introduced the p.Met659Ile (c.1977G>A) variant into *MYH7* of the iPSCs from the healthy donor with CRISPR/Cas9 (Pavlova et al., 2024). Comparing the cardiomyocytes derived from the CRISPR/Cas9-edited iPSC line and its healthy isogenic control demonstrated that introduction of the variant resulted in appearance of HCM features such as an increased cardiomyocyte size, an elevated diastolic calcium level, a decreased basal oxygen consumption rate, and changes in expression pattern of HCM-related genes. These findings support the pathogenicity of the variant. However, validating the effects of the p.Met659Ile (c.1977G>A) variant in *MYH7* under another genetic background could reinforce the conclusion on the variant clinical significance.

The variant correction in the patient-specific iPSCs was performed using electroporation with CRISPR/Cas9 ribonucleoprotein complexes and single-stranded donor oligonucleotide. The method of CRISPR/Cas9 delivery was shown to cause a higher rate of editing events and reduced rate of CRISPR/Cas9 off-target activity in comparison with CRISPR/Cas9 delivery via plasmid transfection (Liang et al., 2015). To augment the efficiency of the editing process, we also used chemical modifications, 2′-O-methyl 3′ phosphorothioate and 3′ phosphorothioate bonds between the first three 5′ and 3′ terminal nucleotides, to protect the single-guide RNA and single-stranded donor oligonucleotide, respectively, from degradation and to stabilize the system (Hendel et al., 2015). As a result, a high percentage of the iPSC clones with editing events (84.52 %) and homology-directed repair (20.23 %) was observed after correction of the p.Met659Ile (c.1977G>A) variant in *MYH7* of the patient-specific iPSCs. Moreover, no CRISPR/Cas9 off-target activity was revealed when analyzing the top-5 CRISPR/Cas9 off-target sites in the iPSC clones.

The iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* (ICGi019-B-1 and ICGi019-B-2) matched all the criteria of human pluripotent stem cells. The iPSC lines had an appropriate morphology, expressed the main transcription factors and surface antigens characteristic of the pluripotent state, and gave rise to derivatives of three germ layers during spontaneous differentiation. This fact, together with the maintenance of the normal karyotype, makes the iPSC lines a good isogenic control for further verification of the variant pathogenicity and examination of HCM pathogenetic mechanisms triggered by the p.Met659Ile (c.1977G>A) variant in *MYH7*.

Conclusion

Using CRISPR/Cas9, an HCM-associated variant of unknown significance, p.Met659Ile (c.1977G>A) in *MYH7*, was corrected in the patient-specific iPSCs. Eight iPSC lines with the corrected variant have been generated and two of the iPSC lines (ICGi019-B-1 and ICGi019-B-2) have been characterized in detail. The ICGi019-B-1 and ICGi019-B-2 iPSC lines retained the pluripotent status and normal karyotype and demonstrated no CRISPR/Cas9 off-target activity, which gives an opportunity to use the iPSC lines with the corrected p.Met659Ile (c.1977G>A) variant in *MYH7* for studying the variant pathogenicity and role in HCM pathogenesis.

References

Akhtar M., Elliott P. The genetics of hypertrophic cardiomyopathy. *Glob Cardiol Sci Pract.* 2018;2018(3):36. doi 10.21542/gcsp.2018.36

Bashyam M.D., Purushotham G., Chaudhary A.K., Rao K.M., Acharya V., Mohammad T.A., Nagarajaram H.A., Hariram V., Narasimhan C. A low prevalence of *MYH7*/*MYBPC3* mutations among Familial Hypertrophic Cardiomyopathy patients in India. *Mol Cell Biochem.* 2012;360(1-2):373-382. doi 10.1007/s11010-011-1077-x

Bhagwan J.R., Mosqueira D., Chairez-Cantu K., Mannhardt I., Bodbin S.E., Bakar M., Smith J.G.W., Denning C. Isogenic models of hypertrophic cardiomyopathy unveil differential phenotypes and mechanism-driven therapeutics. *J Mol Cell Cardiol.* 2020;145:43-53. doi 10.1016/j.yjmcc.2020.06.003

Chai A.C., Cui M., Chemello F., Li H., Chen K., Tan W., Atmanli A., McAnally J.R., Zhang Y., Xu L., Liu N., Bassel-Duby R., Olson E.N. Base editing correction of hypertrophic cardiomyopathy in human cardiomyocytes and humanized mice. *Nat Med.* 2023;29(2):401-411. doi 10.1038/s41591-022-02176-5

Cheng J., Novati G., Pan J., Bycroft C., Žemgulyte A., Applebaum T., Pritzel A., … Senior A.W., Jumper J., Hassabis D., Kohli P., Avsec Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science.* 2023;381(6664):eadg7492. doi 10.1126/science.adg7492

Cohn R., Thakar K., Lowe A., Ladha F.A., Pettinato A.M., Romano R., Meredith E., Chen Y.S., Atamanuk K., Huey B.D., Hinson J.T. A contraction stress model of hypertrophic cardiomyopathy due to sarcomere mutations. *Stem Cell Rep.* 2019;12(1):71-83. doi 10.1016/j.stemcr.2018.11.015

Dementyeva E.V., Vyatkin Y.V., Kretov E.I., Elisaphenko E.A., Medvedev S.P., Zakian S.M. Genetic analysis of patients with hypertrophic cardiomyopathy. *Genes Cells.* 2020a;15(3):68-73. doi 10.23868/202011011 (in Russian)

Dementyeva E.V., Kovalenko V.R., Zhiven M.K., Ustyantseva E.I., Kretov E.I., Vyatkin Y.V., Zakian S.M. Generation of two clonal iPSC lines, ICGi019-A and ICGi019-B, by reprogramming peripheral blood mononuclear cells of a patient suffering from hypertrophic cardiomyopathy and carrying a heterozygous p.M659I mutation in *MYH7*. *Stem Cell Res.* 2020b;46:101840. doi 10.1016/j.scr.2020.101840

Escribá R., Larrañaga-Moreira J.M., Richaud-Patin Y., Pourchet L., Lazis I., Jiménez-Delgado S., Morillas-García A., … de la Pompa J.L., Brugada R., Monserrat L., Barriales-Villa R., Raya A. iPSC-based modeling of variable clinical presentation in hypertrophic cardio-

myopathy. *Circ Res.* 2023;133(2):108-119. doi 10.1161/circresaha.122.321951

Funakoshi S., Yoshida Y. Recent progress of iPSC technology in cardiac diseases. *Arch Toxicol.* 2021;95(12):3633-3650. doi 10.1007/s00204-021-03172-3

Gähwiler E.K.N., Motta S.E., Martin M., Nugraha B., Hoerstrup S.P., Emmert M.Y. Human iPSCs and genome editing technologies for precision cardiovascular tissue engineering. *Front Cell Dev Biol.* 2021;9:639699. doi 10.3389/fcell.2021.639699

Geske J.B., Ommen S.R., Gersh B.J. Hypertrophic cardiomyopathy: clinical update. *JACC Heart Fail.* 2018;6(5):364-375. doi 10.1016/j.jchf.2018.02.010

Guo G., Wang L., Li X., Fu W., Cao J., Zhang J., Liu Y., … Liu G., Zhang Y., Dong J., Tao H., Zhao X. Enhanced myofilament calcium sensitivity aggravates abnormal calcium handling and diastolic dysfunction in patient-specific induced pluripotent stem cell-derived cardiomyocytes with *MYH7* mutation. *Cell Calcium.* 2024;117:102822. doi 10.1016/j.ceca.2023.102822

Guo H., Liu L., Nishiga M., Cong L., Wu J.C. Deciphering pathogenicity of variants of uncertain significance with CRISPR-edited iPSCs. *Trends Genet.* 2021;37(12):1109-1123. doi 10.1016/j.tig.2021.08.009

Hendel A., Bak R.O., Clark J.T., Kennedy A.B., Ryan D.E., Roy S., Steinfeld I., … Bacchetta R., Tsalenko A., Dellinger D., Bruhn L., Porteus M.H. Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat Biotechnol.* 2015;33(9):985-989. doi 10.1038/nbt.3290

Hesaraki M., Bora U., Pahlavan S., Salehi N., Mousavi S.A., Barekat M., Rasouli S.J., Baharvand H., Ozhan G., Totonchi M. A novel missense variant in actin binding domain of *MYH7* is associated with left ventricular noncompaction. *Front Cardiovasc Med.* 2022;9:839862. doi 10.3389/fcvm.2022.839862

Liang X., Potter J., Kumar S., Zou Y., Quintanilla R., Sridharan M., Carte J., Chen W., Roark N., Ranganathan S., Ravinder N., Chesnut J.D. Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J Biotechnol.* 2015;208:44-53. doi 10.1016/j.jbiotec.2015.04.024

Livak K.J., Schmittgen T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods.* 2001;25(4):402-408. doi 10.1006/meth.2001.1262

Ma N., Zhang J.Z., Itzhaki I., Zhang S.L., Chen H., Haddad F., Kitani T., Wilson K.D., Tian L., Shrestha R., Wu H., Lam C.K., Sayed N., Wu J.C. Determining the pathogenicity of a genomic variant of uncertain significance using CRISPR/Cas9 and human-induced pluripotent stem cells. *Circulation.* 2018;138(23):2666-2681. doi 10.1161/circulationaha.117.032273

Malakhova A.A., Grigor'eva E.V., Pavlova S.V., Malankhanova T.B., Valetdinova K.R., Vyatkin Y.V., Khabarova E.A., Rzaev J.A., Zakian S.M., Medvedev S.P. Generation of induced pluripotent stem cell lines ICGi021-A and ICGi022-A from peripheral blood mononuclear cells of two healthy individuals from Siberian population. *Stem Cell Res.* 2020;48:101952. doi 10.1016/j.scr.2020.101952

Mosqueira D., Mannhardt I., Bhagwan J.R., Lis-Slimak K., Katili P., Scott E., Hassan M., … Williams P.M., Gaffney D., Eschenhagen T., Hansen A., Denning C. CRISPR/Cas9 editing in human pluripotent stem cell-cardiomyocytes highlights arrhythmias, hypocontractility, and energy depletion as potential therapeutic targets for hypertrophic cardiomyopathy. *Eur Heart J.* 2018;39(43):3879-3892. doi 10.1093/eurheartj/ehy249

Parrotta E.I., Lucchino V., Scaramuzzino L., Scalise S., Cuda G. Modeling cardiac disease mechanisms using induced pluripotent stem cell-derived cardiomyocytes: progress, promises and challenges. *Int J Mol Sci.* 2020;21(12):4354. doi 10.3390/ijms21124354

Pasipoularides A. Challenges and controversies in hypertrophic cardiomyopathy: clinical, genomic and basic science perspectives. *Rev Esp Cardiol* (*Engl Ed*). 2018;71(3):132-138. doi 10.1016/j.rec.2017.07.003

Pavlova S.V., Shulgina A.E., Zakian S.M., Dementyeva E.V. Studying pathogenetic contribution of a variant of unknown significance, p.M659I (c.1977G>A) in *MYH7*, to the development of hypertrophic cardiomyopathy using CRISPR/Cas9-engineered isogenic induced pluripotent stem cells. *Int J Mol Sci.* 2024;25(16):8695. doi 10.3390/ijms25168695

Richard P., Charron P., Carrier L., Ledeuil C., Cheav T., Pichereau C., Benaiche A., … Desnos M., Schwartz K., Hainque B., Komajda M., EUROGENE Heart Failure Project. Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. *Circulation.* 2003;107(17):2227-2232. doi 10.1161/01.CIR.0000066323.15244.54

Shafaattalab S., Li A.Y., Gunawan M.G., Kim B., Jayousi F., Maaref Y., Song Z., Weiss J.N., Solaro R.J., Qu Z., Tibbits G.F. Mechanisms of arrhythmogenicity of hypertrophic cardiomyopathy-associated troponin T (*TNNT2*) variant I79N. *Front Cell Dev Biol.* 2021;9:787581. doi 10.3389/fcell.2021.787581

Smith J.G.W., Owen T., Bhagwan J.R., Mosqueira D., Scott E., Mannhardt I., Patel A., Barriales-Villa R., Monserrat L., Hansen A., Eschenhagen T., Harding S.E., Marston S., Denning C. Isogenic pairs of hiPSC-CMs with hypertrophic cardiomyopathy/LVNC-associated ACTC1 E99K mutation unveil differential functional deficits. *Stem Cell Rep.* 2018;11(5):1226-1243. doi 10.1016/j.stemcr.2018.10.006

Sorogina D.A., Grigor'eva E.V., Malakhova A.A., Pavlova S.V., Medvedev S.P., Vyatkin Y.V., Khabarova E.A., Rzaev J.A., Zakian S.M. Creation of induced pluripotent stem cells ICGi044-B and ICGi044-C using reprogramming of peripheral blood mononuclear cells of a patient with Parkinson's disease associated with c.1492T>G mutation in the *GLUD2* gene. *Russ J Dev Biol.* 2023;54(1):104-111. doi 10.1134/S1062360423010125

Wang L., Kim K., Parikh S., Cadar A.G., Bersell K.R., He H., Pinto J.R., Kryshtal D.O., Knollmann B.C. Hypertrophic cardiomyopathy-linked mutation in troponin T causes myofibrillar disarray and pro-arrhythmic action potential changes in human iPSC cardiomyocytes. *J Mol Cell Cardiol.* 2018;114:320-327. doi 10.1016/j.yjmcc.2017.12.002

Conflict of interest. The authors declare no conflict of interest.

Received November 29, 2024. Revised December 9, 2024. Accepted December 9, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):358-369

doi 10.18699/vjgb-25-39

Modern methods in peach (*Prunus persica*) genome research

I.V. Rozanova [1, 2], E.A. Vodiasova [2]

[1] Federal Research Center the N.I. Vavilov All-Russian Institute of Plant Genetic Resources (VIR), St. Petersburg, Russia
[2] The Nikitsky Botanical Gardens – National Scientific Centre of RAS, Nikita, Yalta, Republic of Crimea, Russia
i.rozanova@vir.nw.ru

Abstract. Peach (*Prunus persica* (L.) Batsch) is one of the main agricultural stone fruit crops of the family Rosaceae. Modern breeding is aimed at improving the quality of the fruit, extending the period of its production, increasing its resistance to unfavorable environmental conditions and reducing the total cost of production of cultivated varieties. However, peach breeding is an extremely long process: it takes 10–15 years from hybridization of the parental forms to obtaining fruit-bearing trees. Research into peach varieties as donors of desirable traits began in the 1980s. The first version of the peach genome was presented in 2013, and its appearance contributed to the identification and localization of loci, followed by the identification of candidate genes that control the desired trait. The development of NGS has accelerated the development of methods based on the use of diagnostic DNA markers. Approaches that allow accelerating classical breeding processes include marker-oriented selection (MOS) and genomic selection. In order to develop DNA markers associated with the traits under investigation, it is necessary to carry out preliminary mapping of loci controlling economically desirable traits and to develop linkage maps. SNP-chip approaches and genotyping by sequencing (GBS) methods are being developed. In recent years, genome-wide association analysis (GWAS) has been actively used to identify genomic loci associated with economically important traits, which requires screening of large samples of varieties for hundreds and thousands of SNPs. Study on the pangenome has shown the need to analyze a larger number of samples, since there is still not enough data to identify polymorphic regions of the genome. The aim of this review was to systematize and summarize the major advances in peach genomic research over the last 40 years: linkage and physical map construction, development of different molecular markers, full genome sequencing for peach, and existing methods for genome-wide association studies with high-density SNP markers. This review provides a theoretical basis for future GWAS analysis in order to identify high-performance markers of economically valuable traits for peach and to develop genomic selection of this crop.

Key words: *Prunus persica*; GWAS; selection; genotyping; SNP

For citation: Rozanova I.V., Vodiasova E.A. Modern methods in peach (*Prunus persica*) genome research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):358-369. doi 10.18699/vjgb-25-399

Funding. The study is supported by the Kurchatov Genomic Centre of the NBG–NSC (No. 075-15-2019-1670).

Современные методы в исследованиях генома персика (*Prunus persica*)

И.В. Розанова [1, 2], Е.А. Водясова [2]

[1] Федеральный исследовательский центр Всероссийский институт генетических ресурсов растений им. Н.И. Вавилова (ВИР), Санкт-Петербург, Россия
[2] Никитский ботанический сад – Национальный научный центр РАН, п.г.т. Никита, Ялта, Республика Крым, Россия
i.rozanova@vir.nw.ru

Аннотация. Персик (*Prunus p*ersica (L.) Batsch) – одна из основных сельскохозяйственных плодовых косточковых культур семейства розоцветных. Современная селекция направлена на улучшение качества плодов, расширение сроков их получения, создание сортов с устойчивостью к неблагоприятным условиям среды и снижение общих затрат на производство культивируемых сортов. При этом селекция персика – долгий процесс: от гибридизации родительских форм до получения плодоносящих деревьев проходит 10–15 лет. Исследования сортов персика как доноров желаемых признаков начались с 1980-х годов. Первый вариант генома персика был представлен в 2013 г., и его появление способствовало определению и локализации локусов, с последующим обнаружением генов-кандидатов, под чьим контролем находится хозяйственно ценный признак. Развитие методов NGS ускорило продвижение подходов, основанных на применении диагностических ДНК-маркеров. К таким подходам, позволяющим ускорять процессы классической селекции, относятся маркер-ориентированная селекция и геномная селекция. Для того чтобы разработать ДНК-маркеры, ассоциированные с изучаемыми свойствами, необходимо провести предварительное картирование локусов, контролирующих хозяйственно ценные признаки, создать карты сцепления. Работы по пангеному показали необходимость анализировать

большее количество образцов, так как до сих пор не хватает данных для нахождения полиморфных областей генома. Развиваются подходы использования SNP-чипов и методов генотипирования через секвенирование (GBS, genotyping-by-sequencing). В последние годы для обнаружения локусов генома, ассоциированных с хозяйственно ценными признаками, активно применяется метод полногеномного анализа ассоциаций (GWAS, genome-wide association studies), для использования которого требуется скрининг больших выборок сортов по сотням и тысячам SNP. Цель настоящего обзора – систематизация и обобщение основных достижений в области геномных исследований персика за последние 40 лет: построение карт сцепления и физических карт, получение различных молекулярных маркеров, полногеномное секвенирование для персика, а также описание существующих работ полногеномных исследований ассоциаций с маркерами SNP высокой плотности. Этот обзор обеспечивает теоретическую основу для проведения GWAS с целью выявления высокоэффективных маркеров хозяйственно ценных признаков для персика и развития геномной селекции этой культуры.
Ключевые слова: *Prunus persica*; GWAS; селекция; генотипирование; SNP

Introduction

Peach (*Prunus persica* L.) is one of the main agricultural stone crops of the temperate zone, consumed fresh and processed, contains high amounts of vitamins, minerals, fiber and antioxidant compounds, while being low in calories and therefore excellent for dietary menus. As a species, it originated about 2.5 million years ago in the southwestern Tibetan Plateau region of China, from where its domestication began 4,000–5,000 years ago (Yu Y. et al., 2018) (see the Figure).

The current peach gene pool is divided into three groups that are characterized by differing climatic growing conditions (see the Figure). These groups originated in China in different geographical regions. The Southern group originated in climates with mild winters and hot, humid summers. These peaches have a flat shape and are characterized by a slightly acidic "honey" flavor. The Northern group included peach genotypes found in regions with cold winters and hot, dry summers. These peaches are generally resistant to drought and cold, but are not adapted for growth in southern areas. The third group is found in the arid northwest of China. It includes nectarines and peaches with yellow flesh, in contrast to the white-flesh peaches typical of the rest of China (Scorza, 1991).

The crop spread to Europe more than 2,000 years ago, along ancient trade routes through Persia (Hesse, 1975; Byrne et al., 2012). Peach was introduced to the Americas by Spanish and Portuguese settlers 400 years ago (Hesse, 1975; Scorza, Okie, 1990; Faust, Timon, 1995).

By now, peach is the second most important temperate fruit crop after apple. More than 1,000 varieties of *P. persica* with different phenotypic variations for various traits such as shape, fruit size, flavor, flower type, etc. have been produced worldwide. According to FAO data (https://www.fao.org/faostat/en/#data/QCL), peach is now grown almost everywhere on all continents except Greenland, northern regions of Europe and some regions in central Africa (see the Figure).

There is little information in the literature about existing peach collections in the world. In the United States, the first

Distribution and origin of *P. persica*.

The arrows show the distribution routes of the peach. Countries where the peach currently grows and its production level are shown according to the color scheme. Origin of peach in China: the Southern group spans the Yangtze River; the Northern group is along the Yellow River; the third group originated in northwest China.

breeding programs appeared in the late 18th century (Hesse, 1975; Scorza, Okie, 1990; Faust, Timon, 1995). Since the founding of the USDA (United States Department of Agriculture) in 1889, more than 2,100 clones or seeds of peaches and nectarines were imported into America from China and other parts of the world. The collection housed at the USDA Plant Introduction Garden at Chico, in California, had about 700 unique peach accessions. It was used by most breeders but focused primarily on genotypes derived from crosses with a member of the southern group of Chinese cultivars, 'Chinese Cling' (Scorza et al., 1985). This variety was widely used for crop improvement. During the 1950–1960s, the collection gradually declined and was almost eliminated with the closure of the station in the late 1960s. Only 60 varieties survived and were transferred to farms in Byron, Ga., and Beltsville, Md. farms. These varieties have unique traits not found in the gene pool descending from 'Chinese Cling' (Werner, Okie, 1998). Currently, the gene pool of the peach population in the U.S. is considered to be the most impoverished.

In order to establish a peach cultivar collection in China, peach germplasm collection was initiated at the Zhengzhou Fruit Research Institute (ZFRI) in the 1960s. In 1986, the National Peach Collection was established, which consisted of more than 600 accessions by 2000 (Wang et al., 2001). To date, more than 1,200 peach accessions have been collected from around the world, including wild species, ancient cultivar populations, and modern cultivars (Lirong et al., 2020).

Two peach collections are located in northern Spain: the National Peach Collection Gene Pool "Centro de Investigación y Tecnología Agroalimentaria de Aragón" (CITA) and "Estación Experimental de Aula Dei" (EEAD-CSIC) (https://cita-aragon.es/en/history-mission-vision-and-aims/). The quantitative characteristics of the collections are not presented on the website.

In Russia, the largest peach collection is located in the Nikita Botanical Garden in the Crimea and has 790 peaches and 85 nectarines (Smykov et al., 2021).

Peach is diploid ($2n = 16$), self-pollinating, with a base chromosome number of eight and belongs to the Rosaceae family, subfamily Prunoideae (Bassi, Monet, 2008). It has a lower level of genetic variability compared to other *Prunus* cultivars. On the one hand, the ability to self-pollinate is a limiting factor in breeding programs, on the other hand, in combination with such biological features as small genome size (265 Mb) (Arumuganathan, Earle, 1991) and diploid set of chromosomes, and as a result of its economic value, peach is an excellent model for genomic studies of stone fruits of the Rosaceae family (Monet et al., 1996; Abbott et al., 2002). The genomes of different *Prunus* species are highly conserved (Dirlewanger et al., 2004), allowing many of the major genes and loci of peach and other *Prunus* species to be placed on the same genetic map (Abbott et al., 2008).

Traditional peach seedling breeding is a labor-intensive process that takes 10–15 years from the initial crossing to the emergence of a new cultivar (Bliss, 2010; Ru et al., 2015). In addition, peach breeding programs require significant acreage due to the large size of the trees, as well as financial resources to cover the ongoing costs for technical treatments such as spraying herbicides, insecticides and fungicides, planting, pruning, thinning and watering. With the development of genetics, studies on the genetic diversity of the crop began (Herrero et al., 1964), the use of diagnostic DNA markers was developed (Callahan et al., 1991; Lambert et al., 2016; Demirel et al., 2024), and the advent of NGS sequencing techniques (Micali et al., 2015; Kim et al., 2021) allowed the generation of high-quality whole genome sequences for genomic breeding approaches.

The aim of this review is to summarize the results of genetic and breeding works for the *P. persica* culture based on the application of NGS methods.

Methods before NGS: isoenzymes, DNA markers, first linkage maps

Despite the significant progress made by peach breeders over the past hundred years, traditional seedling breeding is a labor-intensive process since, in temperate climates, peach trees require at least three years to reach fruiting maturity before progeny fruit quality can be evaluated (Bliss, 2010). In the late 1980s, it was recognized that markers, the alleles of which have distinct differences at the phenotypic level, could be useful in the analysis of complex traits (Monet, 1988). And markers, which are closely linked to traits that appear late in development, can be valuable for early tree selection (Chaparro et al., 1994).

Research on peach varieties as donors of desirable traits began in the 1980s (see the Table). Protein markers, or isozymes (isoenzymes), were the first to be used as potential markers to identify particularly valuable hybrids. Isoenzymes are different variants of an enzyme (different amino acid sequence isoforms of the same enzyme) that differ in electrophoretic mobility. Isoenzyme analysis could be used to distinguish hybrids between plum and peach from plum offspring (Parfitt et al., 1985), peach and almond hybrids (Arulsekar et al., 1986a; Chaparro et al., 1987). G.E. Jr. Carter with colleagues showed that differences in protein structure were sufficient to distinguish each peach cultivar (Carter, Brock, 1980) or to reveal their similarities (Arulsekar et al., 1986b).

Over time, a sufficient number of morphological markers, the localization of which on the chromosome is known, have become available. In this case, by analyzing F_2 populations from crosses, it is possible to determine the chromosomal positions of isozyme loci relative to morphological markers. R.E. Durham with colleagues identified the presence of separate, independently inherited loci by examining isoenzymes such as diaphorase, malate dehydrogenase and peroxidase (Durham et al., 1987). A total of four phenotypic trait linkage groups with isoenzymes were known in the early 1990s (Bailey, French, 1949; Monet et al., 1985; Monet, 1988; Monet, Gibault, 1991). However, low frequency and other drawbacks have prevented the widespread use of protein markers in breeding programs.

With the development of sequencing methods, a technological breakthrough occurred, and new methods of DNA polymorphism analysis appeared, which led to the emergence of molecular markers characterized by high frequency of occurrence in the genome. A molecular marker (DNA marker) corresponds to a gene or genomic region having different variants (alleles) and associated with different phenotypic manifestations (Khlestkina, 2014). Based on their use, approaches have emerged that complement classical breeding methods by

History of the genetic and genomic *P. persica* research

Year	Significant studies	Reference
1980–1990	Morphological traits localization on peach chromosomes using isoenzyme analysis techniques	Bailey, French, 1949; Carter, Brock, 1980; Monet et al., 1985; Monet, 1988; Monet, Gibault, 1991
1992	Generation of the first map showing linkage of markers to peach traits (35 RFLP markers for five bonding groups)	Belthoff et al.,1993
1995	Creation of a reference genetic map (T × E)	Foolad et al., 1995
2012	The DNA chip IPSC 9K SNP v1 creation	Verde et al., 2012
2013	Creation of the first version of the peach reference genome Peach v 1.0 (dihaploid 'Lovell' variety)	International Peach Genome Initiative et al., 2013
2015	Application of the GBS method on the peach	Bielenberg et al., 2015
2017	Updated peach reference genome Peach v 2.0 (dihaploid 'Lovell' variety)	Verde et al., 2017
2019	The 18K SNP array v2 DNA chip creation	Gasic et al., 2019
2020	Peach pangenome establishment (100 accessions)	Cao et al., 2020
2021	The 'Chinese Cling' variety sequencing (the variety central to the development of cultivated peaches)	Cao et al., 2021

searching for DNA markers associated with valuable traits in order to accelerate the breeding process.

Economic traits such as productivity, quality, maturity, tolerance to biotic and abiotic stresses are quantitative traits and are polygenic. The search and labeling of polymorphic loci associated with quantitative traits, or, in other words, Quantitative Trait Loci (QTL), is an agronomically important task. By accumulating information on molecular markers, it is possible to create genetic maps, the purpose of which is to identify neutrally inherited markers in close proximity to genetic determinants (loci or genes) that control the manifestation of certain traits, including quantitative traits (Chesnokov, Artem'eva, 2011).

QTLs for traits such as ripening period, fruit weight, size and texture, pH, titratable acidity and soluble solids have been found in peach (Quarta et al., 2002). Fruit quality QTLs tend to cluster in several genomic regions, especially in linkage groups 4, 5 and 6 (Dirlewanger et al., 2009). Similarly, most QTL and disease resistance genes are also clustered (Pflieger et al., 1999). This observation suggests that (1) a small number of Mendelian factors can explain most of the genetic variability in fruit quality traits and (2) traits of different characteristics often share common QTLs (Dirlewanger et al., 1999). Consequently, common QTLs usually correspond to different closely related genes or to a single gene with pleiotropic effects on many traits influenced by the same physiological process (Quilot et al., 2004).

Linkage maps approximate the genomic position and genetic distances between markers using linkage analysis of genetic data (Paterson, 1996; Jones et al., 1997; Collard et al., 2005). The construction of a genetic linkage map is based on meiotic events, where genetic recombination occurs, leading to the development of recombinant genotypes. The lower the recombination frequency between molecular markers, the more likely they are to be linked and in the same linkage group (Paterson, 1996; Jones et al., 1997; Collard et al., 2005). Markers are called unlinked if their recombination frequency is greater than 50 % and, thus, they are located in different linkage groups. Recently, such calculations have been done using software, e. g. MapChart 2.2 (Voorrips, 2002), Mapmaker (Lander et al., 1987), TMAP (Cartwright et al., 2007), MapQTL (Van Ooijen, 2009), Joinmap (Van Ooijen, 2006).

The first maps showing the association between peach phenotypic traits and markers appeared in 1992 (Belthoff et al., 1992). L.E. Belthoff and colleagues (1992) developed a genetic map containing five linkage groups using 35 RFLP markers. J.X. Chaparro and colleagues developed a linkage map for peach using 83 RAPD markers and two isozymes (Chaparro et al., 1994). Then, by investigating the F_2 generation obtained from crosses between almond (cultivar 'Texas') and peach (cultivar 'Early Gold'), a genetic map (T × E) was created, which was later used as a *Prunus* reference map (Foolad et al., 1995; Joobeur et al., 1998; Pozzi, Vecchietti, 2009). It included all eight clutch groups and covered a total distance of 491 cM. Microsatellite markers (Simple Sequences Repeats, SSR) have been widely used as diagnostic DNA markers in peach research since the 2000s (Sosinski et al., 2000; Dirlewanger et al., 2002; Hong et al., 2013).

The *Prunus* 'T × E' reference map contains 1,947 anchor markers (i. e., evenly distributed throughout the *Prunus* genome) with known map locations (Howad et al., 2005; Dirlewanger et al., 2007; Pozzi, Vecchietti, 2009), which allowed comparisons between the peach genomes and the rest of the *Prunus* species. This facilitated the subsequent development of intraspecific maps for peach and other maps of interspecific relationships in *Prunus* (Howad et al., 2005; Dirlewanger et al., 2007; Pozzi, Vecchietti, 2009).

Currently, 70 linkage maps have been generated for peach and related interspecific hybrids, which can be found in the Rosaceae Genome Database (GDR; http://www.rosaceae.org/) (Jung et al., 2008, 2014) as well as in J.A. Salazar et al. (2013).

These genetic linkage maps continue to serve as effective tools for comparison with the *Prunus* 'T × E' reference map.

Peach genomic studies

To identify and localize loci, with the following identification of candidate genes under the control of an economically valuable qualitative or quantitative peach trait, it is necessary to obtain the complete genome of *P. persica* (Tanksley et al., 1989; Winter, Kahl, 1995; Paterson, 1996; Jones et al., 1997; Collard et al., 2005).

The first Peach v1.0 genome variant was submitted in 2013 (International Peach Genome Initiative et al., 2013). The target for the full genome sequencing was a dihaploid genotype of the cultivar 'Lovell' (Plov2_2N), which was read by the Sanger method with 8.5-fold coverage. Eight pseudomolecules, reflecting the eight peach chromosomes, were assembled. The resulting genome assembly (Peach v1.0), spanning 227.3 Mb, of which 218.4 Mb (96 %) was decoded, contained 27,852 annotated genes, with an average density of 1.22 genes per 10 Kb. In 2017, an updated Peach v2.0 reference map was published (Verde et al., 2017), constructed by repeated NGS sequencing of the 'Lovell' dihaploid sample on the Illumina platform. The sequence totaled 227.4 Mb, which is only slightly longer than the first variant, but deeper resequencing resulted in 225.7 Mb (99.2 %) of the sequence being decoded. The approximate positions of the centromeric regions of the chromosomes were determined based on repetitive regions with low gene concentration and low recombination frequency. Based on the reference genome sequence, researchers studied the evolutionary history of the peach fruit (Cao et al., 2014, 2020; Yu Y. et al., 2018), identified domestication regions (Cao et al., 2014; Akagi et al., 2016; Li Y. et al., 2019), and genes controlling economically valuable traits (Cao et al., 2014, 2016, 2019).

The Peach v2.0 physical map was correlated with four genetic maps: (1) 67 forms from a mapping population of an interspecific cross between almond and peach 'Texas' × 'Earligold' F_2 (T × E (Joobeur et al., 1998)); (2) 242 forms from a mapping population derived from IF7310828 × Ferganensis BC1 (P × F (Dettori et al., 2001)); (3) 305 seedlings from the mapping population 'Contender' × 'Ambra' F_2 (C × A (Eduardo et al., 2011)); (4) 62 hybrids from the cross 'Maria Dolce' × SD81 F_1 (MD × SD). The mapping strategy involved an approach using SSR and SNP markers targeting specific regions of the peach genome and a full-genome approach using the IPSC 9K SNP v1 chip (Verde et al., 2012). The loci derived from the genetic maps were mapped to physical positions using the MareyMap package (Siberchicot et al., 2017). For each linkage map used in this study (T × E, C × A, P × F and MD × SD), recombination rates were estimated as the ratio between genetic (cM) and physical (million base pairs) distances.

However, a single reference assembly does not reflect intraspecific variability, so there is a need to investigate the genetic variation of different peach cultivars and their wild relatives using pangenomic analysis. Similar work on pangenome construction has been carried out for many crops such as soybean (Li Y.H. et al., 2014; Liu Y. et al., 2020), rice (Zhao et al., 2018), sunflower (Hübner et al., 2019), tomato (Gao et al., 2019), barley (Jayakodi et al., 2020). In 2014, several accessions' genomes were comparatively analyzed. To assess the process of peach domestication, 11 peach accessions (including the dihaploid 'Lovell' PLov2-2N used for reference assembly as a control) and one each of *P. ferganensis, P. kansuensis, P. davidiana* and *P. mira* were re-sequenced. *P. ferganensis* is considered a wild undomesticated peach or, more likely, represents an intermediate variant in the peach domestication. Using a set of 953,357 high-quality SNPs identified in *P. persica* and *P. ferganensis* samples, nucleotide sequence diversity was assessed for eight collected chromosomes (International Peach Genome Initiative et al., 2013).

In the first chromosome, the number of genetic variants at polymorphic loci was minimal. The greatest diversity among SNPs was observed in the distal region of the short arm of chromosome 2 and in the distal region of the long arm of chromosome 4. The density of genes encoding receptor proteins from the family of conserved nucleotide-binding leucine-rich proteins (R-proteins), which are involved in immunity, was 5-fold higher on chromosome 2 than in the rest of the genome (Dodds, Rathjen, 2010). Immunity-related regions are rapidly evolving, so the diversity detected is natural. It is known from the literature that genes associated with fruit ripening are located on chromosome 4 (Eduardo et al., 2011; Dirlewanger et al., 2012).

Since the study included samples with different ripening times, there is a high level of variability in the region associated with ripening due to the given sampling parameters. However, genotypes of *P. kansuensis, P. davidiana* or *P. mira* mature at the same optimal time and no SNP diversity between regions was found in these species (International Peach Genome Initiative et al., 2013). Similarly, studies by sequencing of six related peach (*P. persica*) accessions were also conducted (Guan et al., 2019). The genomic variations identified showed that the comparison of different crop genotypes is effective for the development of DNA markers. These works support the need to analyze more accessions, as there are still insufficient data to identify the polymorphic regions of the genome.

The first peach pangenome, consisting of 100 sequenced samples of *P. persica*, was obtained in 2020 (Cao et al., 2020). Also, in this work, the *de novo* genomes of four wild peach relatives, *P. mira, P. davidiana, P. kansuensis* and *P. ferganensis*, were assembled. When the sequenced peach accessions were compared with the reference genome (Verde et al., 2017), an average of 3.4 % of reads in each accession failed to match the reference genome, and these reads were assembled *de novo* by the researchers. In total, an additional 2.52 Mb of new sequences containing 2,833 contigs (>500 bp) of potential significance were obtained. Additionally, 923 new genes were identified in the newly assembled sequences (Cao et al., 2020). The total number of genes in the pangenome was 27,796. Genes were divided into conserved genes which were common to all 100 samples (24,971, 89.9 %), and variable genes (2,803, 10.1 %), the presence of which was detected in less than 99 % of the samples (Cao et al., 2020).

Pangenomic analysis revealed the presence of resistance genes (R-genes) among the variable gene set. A similar situation has been observed in soybean (Li Y.H. et al., 2014) and rice (Zhao et al., 2018). It is hypothesized that variations in resistance (R) gene copy number may help explain differences in resistance between wild and cultivated accessions (Li Y.H.

et al., 2014). Also, using peach pangenome, we found that 63 % of ornamental, 88 % of local, and 91 % of improved cultivars had a set of "optional" four genes encoding geraniol-8-hydroxylases, which are involved in the biosynthesis of terpenes, which play an important role in plant life and have anticarcinogenic, antiseptic and antimicrobial effects. These genes may have been under positive selection pressure both during domestication and during the breeding process.

When comparing *P. persica* with four wild species of the genus *Prunus* collected *de novo* by K. Cao and colleagues, it was found that 34.7 % of all genes found based on homology for encoded proteins were represented in all five species. At the same time, species-specific genes were found in *P. mira* (543 specific genes), *P. davidiana* (485), *P. kansuensis* (194), *P. ferganensis* (197), and *P. persica* (320). Such studies allow the identification of genes that confer species-specific properties. For example, a nematode resistance gene was identified in *P. kansuensis* (Cao et al., 2020). Such work makes it possible to identify differences in the genomes of closely related species and varieties, which is necessary for the identification of genes responsible for valuable qualities and traits of plants.

In 2021, K. Cao and colleagues (Cao et al., 2021) sequenced the 'Chinese Cling' cultivar, which is very important historically and central to the cultivated peach development in Europe (Byrne et al., 2000), Japan (Yamamoto et al., 2003) and the USA (Aranzana et al., 2010). The assembled genome contained 247.33 million base pairs, representing 99.8 % of the putative genome. Its comparison with the 'Lovell' reference genome revealed 685,407 novel SNPs, 162,655 insertions and deletions, and 16,248 copy number variation (CNV) structural variants. Gene family analysis revealed a reduction in gene families involved in the biosynthesis of flavones, flavonols, flavonoids and monoterpenoids compared to the 'Lovell' variety genome.

Thus, the genomic approach allows the comparative analysis of varieties and identification of variable genes (or loci in the genome) that may be responsible for different varietal traits. Such studies remain essential for further development of genomic selection in peach.

New approaches in peach research with NGS

Polymorphism analysis methods have evolved from rather labor-intensive isoenzyme-related and RFLP methods to high-throughput sequencing methods. Comparative studies of genetic distances between peach accessions estimated using SNP and SSR markers have been conducted. In the early 2000s, methods using SSR markers became dominant. M.T. Hamblin and colleagues showed that 89 SSR markers did a better job of clustering the samples of the study sample of 259 maize inbred lines than a set of 847 SNP markers. The researchers concluded that a large number of polymorphic single nucleotide loci are needed for qualitative analysis using SNP markers (Hamblin et al., 2007). Currently, tens of thousands of SNP loci are being analyzed. J.M. Yu and colleagues (2009) calculated that the power of 1,000 SNPs is similar to that of 100 SSRs for estimating population structure and relatedness. At the same time, SSR markers remain a major option for screening plant genetic resource collections (Nybom, Lācis, 2021) and for passporting samples (Trifonova et al., 2021).

SNP markers have a higher distribution frequency in the genome compared to SSR markers, which makes them more functional when polymorphisms within specific genes are required for targeted studies. The first SNP detection technologies were *in silico* search for SNPs by analyzing EST databases followed by PCR-based validation (Batley et al., 2003), and SNP detection by resequencing transcripts using the Sanger method (Morozova, Marra, 2008). However, these methods were unable to detect SNPs in intergenic and non-coding regions. The advent of GBS approaches and the development of DNA chips have overcome the problems associated with the low throughput ability and high cost of SNP detection (Mardis, 2008) and now allow cost-effective and time-efficient detection of SNPs at significant loci. More and more diagnostic SNP markers are now being used in breeding programs. The use of insertions/deletions as markers is also common, but their reproducibility is lower than that of SNPs.

One common approach to SNP determination is genotyping using microarrays, or DNA chips. SNPs on the chip have been developed in such a way that it is possible to differentiate the samples under investigation in the pool of samples. DNA chips have been developed for many commercially important crops on two different platforms: the Illumina Infinium platform (6K for cherry (Peace et al., 2012), 8K for apple (Chagné et al., 2012), 18K for grape (Laucou et al, 2018) and 6K for avocado (Kuhn et al., 2019)) and the Axiom platform (480K SNPs for apple (Bianco et al., 2016), 68K for rose (Koning-Boucoiran et al., 2015), 700K for walnut (Marrano et al., 2019), and 70K and 200K for pear (Montanari et al., 2019; Li X. et al., 2019)).

In order to establish the medium-density Infinium SNP platform suitable for genotyping the peach gene pool, 56 breeding-significant peach accessions spanning the crop gene pool were selected. The samples selected were those used in international peach breeding programs, contributing to the breeding gene pool according to pedigree records, and based on parentage estimates from SSR studies showing genetic diversity. Over 1 million SNPs were obtained and tested, of which exactly 9,000 passed quality control, were genetically informative and formed the platform for genotyping, the first International Peach SNP Consortium (IPSC) peach 9K SNP array v1 chip (Verde et al., 2012). SNPs on the chip were distributed evenly across all eight chromosomes and the average spacing was 26.7 bp (Verde et al., 2012).

Platform validation was performed on 709 peach accessions comprising two independent evaluation samples: 232 accessions from the European Union and 479 accessions from the USA. The EU panel included 229 peach cultivars, and three wild species of the genus *Prunus* or their hybrids with peach. The US panel included pedigree varieties, breeding lines, and seedlings. Overall, the sampled material consisted of 45 % cultivars, 4 % improved breeding lines, and 51 % seedlings. Specimens clearly related to either peach or almond accounted for 82 and 2 %, respectively, while 16 % of the genotyped material was of interspecific origin (with almond).

In the next step, the peach 9K SNP array v1 platform was extended to 18K. The new chip included 9,000 SNPs from the previous version and 7,206 SNPs identified by sequencing 49 samples and uniformly distributed across all peach

chromosomes (Gasic et al., 2019). The uniform distribution of polymorphisms selected for the chip throughout the genome (the number of gaps smaller than 0.3 million base pairs reduced to 2 on the chromosomes 3 and 8) allows finding associations linked to the traits of interest.

Currently, genotyping by sequencing (GBS) has become the most common method of analyzing SNP markers for genome research. The term "GBS" is already used as an umbrella term for various NGS-based high-throughput genotyping methods under development (Rasheed et al., 2017). In plants, this method was first described by R.J. Elshire et al. in 2011 (2011).

Genotyping methods are used both for sequence determination and to identify associations between phenotype and genotype. Since the peach genome has now been sequenced, the identification of genomic regions associated with a trait can be performed immediately to search for candidate genes. The GBS method has been applied in peach research since 2015 (Bielenberg et al., 2015). Research in quantitative genetics is conducted equally using GBS (Cao et al., 2016, 2019; Guan et al., 2019; Li Y. et al., 2019; Meng et al., 2019; Guajardo et al., 2020; Thurow et al., 2020; Huang et al., 2021; Liu J. et al., 2021; Tan et al., 2021; Li X. et al., 2022, 2023), as well as using SNP chips (Micheletti et al., 2015; Akagi et al., 2016; Font i Forcada et al., 2019; Cirilli et al., 2021; da Silva Linge et al., 2021; Fu et al., 2021; Mas-Gómez et al., 2021, 2022).

Both GBS and SNP-chip genotyping have their advantages and disadvantages. For example, the diversity of biallelic SNPs collected at chip creation is limited, while the GBS method can cover and identify significant SNPs associated with a trait that are, however, not included in the chip set. Conversely, GBS often includes a large amount of missing data and coverage must be high enough to ensure reproducibility between the samples studied (Nybom, Lācis, 2021). GBS is currently used more frequently than SNP chips because this approach can be applied to crops for which the reference genome has not yet been sequenced. At least 96 samples are required for large-scale genotyping with GBS or with SNP chips (Zurn et al., 2020).

Analysis of associations between genomic loci and phenotypic traits

Today, modern technologies make it possible to perform genome-wide association studies (GWAS), the results of which are effectively used in breeding programs because they allow simultaneous genomic analysis of several hundred varieties for tens of thousands of loci, comparing the associations between different alleles and the trait of interest. By creating an appropriate sample, GWAS can identify loci for several economically valuable traits at once. This step expands the ability to select markers for agronomically important traits. In the future, the use of molecular markers will allow the selection of desired genotypes among breeding hybrids, which is actively used in marker-assisted breeding (MAB) programs (Khlestkina, 2014). The identification of significant associations facilitates the development of new markers, which can be used to set the required criteria for the variety to be developed.

In peach populations, due to the low level of genetic diversity, association mapping must consider linkage disequilibrium (LD), which is the non-random relationship between two alleles that causes certain allelic combinations to occur most frequently. The method is sensitive to the presence of a large number of related samples in the population structure, leading to spurious associations between phenotypes and marker loci (Mariette et al., 2010). Thus, if a particular combination of alleles confers an adaptive advantage, its frequency will increase relative to the frequency expected under random assignment. Several studies using SSR markers have been conducted in peach in varieties with different genetic backgrounds, and their results indicate that linkage disequilibrium is quite high in this crop.

Kinship between varieties and selection increase the level of linkage disequilibrium. It has been found to range from 6.01 to 20 cM (Aranzana et al., 2010; Cao et al., 2012; Font i Forcada et al., 2013). One strategy to deal with high linkage disequilibrium is to use SNPs that are not correlated with each other for analyses (e. g., taking $r^2 = 0.20$ as a measure of allelic association). Several algorithms exist to prune SNPs in this way or to reduce the degree of linkage disequilibrium between SNPs. Popular pruning strategies are implemented in PLINK 1.07/1.9 (Purcell et al., 2007), which sequentially scan the genome for correlated SNP pairs using only allele counting. As a result, only one representative SNP is retained for each region where highly correlated SNPs are present (Joiret et al., 2019).

The GWAS method has now identified genomic regions associated with many peach traits. Agronomic traits such as maturation, fruit pubescence, flesh colour, texture, flesh colour around the stone, fruit weight and soluble solids content are being studied (Micheletti et al., 2015; Cao et al., 2016; Elsadr, 2016; Font i Forcada et al., 2019; Li Y. et al., 2019; Liu H. et al., 2019; Thurow et al., 2020; Cirilli et al., 2021; da Silva Linge et al., 2021; Mas-Gómez et al., 2021, 2022; Tan et al., 2021; Li X. et al., 2023), as well as seed characteristics (kernel flavor) (Cao et al., 2016), pollen fertility traits (Huang et al., 2021), flower characteristics (Micheletti et al., 2015; Cao et al., 2016; Elsadr, 2016; Meng et al., 2019; Tan et al., 2021). There are works on peach resistance to various diseases (Fu et al., 2021; Li X. et al., 2022), cold and drought tolerance (Li Y. et al., 2019; Tan et al., 2021).

The above works demonstrate the potential value of the GWAS method for identifying new genomic regions associated with phenotypic traits of agricultural importance. This method can also be used to refine data on previously discovered QTLs (e. g., to more accurately determine the size of the locus under study) and facilitate the discovery of genes controlling the trait under investigation.

Conclusion

With the development of NGS approaches, several peach cultivars have been sequenced, providing a basis for whole-genome association studies. The large diversity of cultivars in existing collections allows not only to assess the diversity of the crop's gene pool, but also to search for marker-trait associations. Modern genotyping methods using GBS and SNP chips allow the identification of new markers that enrich the peach database. On the one hand, these new associations are of fundamental interest, contributing to the identification of peculiarities of genome evolution, individual development

of the peach tree and mechanisms of response to various environmental stimuli, and on the other hand, they are the basis for applied work aimed at developing effective markers and their use in obtaining new peach varieties with specified characteristics. This approach makes it possible to accelerate the breeding time of this stone fruit.

However, difficulties remain in the field of association mapping in peach breeding programs. This is mainly due to the fact that the number of samples in the collections studied should be at least 100 to reflect the degree of efficiency. In addition, the relatedness of the varieties and hybrids under study should be assessed beforehand when compiling the sample set. Thus, when working with peach collections, preliminary analysis of genetic diversity and relatedness is necessary, which is also better performed using SNPs.

References

Abbott A.G., Georgi L., Yvergniaux D., Wang Y., Blenda A., Reighard G., Inigo M., Sosinski B. Peach: the model genome for *Rosaceae*. *Acta Hortic*. 2002;575:145-155. doi 10.17660/ActaHortic.2002.575.14

Abbott A.G., Arús P., Scorza R. Genetic engineering and genomics. In: Layne D., Bassi D. (Eds) The Peach Botany, Production and Uses. London: CAB International, 2008;85-105. doi 10.1079/9781845933869.0085

Akagi T., Hanada T., Yaegaki H., Gradziel T.M., Tao R. Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. *DNA Res*. 2016;23(3):271-282. doi 10.1093/dnares/dsw014

Aranzana M.J., Abbassi E.K., Howad W., Arús P. Genetic variation, population structure and linkage disequilibrium in peach commercial varieties. *BMC Genet*. 2010;11:69. doi 10.1186/1471-2156-11-69

Arulsekar S., Parfitt D.E., Kester D.E. Comparison of isozyme variability in peach and almond cultivars. *J Hered*. 1986a;77(4):272-274. doi 10.1093/oxfordjournals.jhered.a110235

Arulsekar S., Parfitt D.E., Beres W., Hansche P.E. Genetics of malate dehydrogenase isozymes in the peach. *J Hered*. 1986b;77(1):49-51. doi 10.1093/oxfordjournals.jhered.a110166

Arumuganathan K., Earle E.D. Nuclear DNA content of some important plant species. *Plant Mol Biol Rep*.1991;9:208-218. doi 10.1007/BF02672069

Bailey J.S., French A.P. The Inheritance of Certain Fruit and Foliage Characters in the Peach. Amherst, MA: University of Massachusetts Press, 1949

Bassi D., Monet R. Botany and taxonomy. In: Layne D.R., Bassi D. (Eds) The Peach: Botany, Production and Uses. Wallingford: CAB International, 2008;1-36. doi 10.1079/9781845933869.0001

Batley J., Barker G., O'Sullivan H., Edwards K.J., Edwards D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol*. 2003;132(1):84-91. doi 10.1104/pp.102.019422

Belthoff L.E., Ballard R., Abbott A., Morgens P., Callahan A., Scorza R., Baird W.V., Monet R. Development of a saturated linkage map of *Prunus persica* using molecular based marker systems. *Acta Hortic*. 1993;336:51-56. doi 10.17660/ActaHortic.1993.336.5

Bianco L., Cestaro A., Linsmith G., Muranty H., Denancé C., Théron A., Poncet C., … Davassi A., Laurens F., Velasco R., Durel C.E., Troggio M. Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant J*. 2016;86(1):62-74. doi 10.1111/tpj.13145

Bielenberg D.G., Rauh B., Fan S., Gasic K., Abbott A.G., Reighard G.L., Okie W.R., Wells C.E. Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. *PloS One*. 2015;10(10):e0139406. doi 10.1371/journal.pone.0139406

Bliss F.A. Marker-assisted breeding in horticultural crops. *Acta Hort*. 2010;859:339-350. doi 10.17660/ActaHortic.2010.859.40

Byrne D.H., Sherman W.B., Bacon T.A. Stone fruit genetic pool and its exploitation for growing under warm winter conditions. In: Erez A. (Ed.) Temperate Fruit Crops in Warm Climates. Dordrecht: Springer, 2000;157-230. doi 10.1007/978-94-017-3215-4_8

Byrne D.H., Bassols M., Bassi D., Piagnani M., Gasic K., Reighard G., Moreno M., Pérez S. Peach. In: Badenes M.L., Byrne D.H. (Eds) Fruit Breeding. New York: Springer Science, 2012;505-570. doi 10.1007/978-1-4419-0763-9_14

Callahan A., Scorza R., Morgens P., Mante S., Cordts J., Cohen R. Breeding for cold hardiness: searching for genes to improve fruit quality in cold-hardy peach germplasm. *HortScience*. 1991;26(5):522-526. doi 10.21273/HORTSCI.26.5.522

Cao K., Wang L., Zhu G., Fang W., Chen C., Luo J. Genetic diversity, linkage disequilibrium, and association mapping analyses of peach (*Prunus persica*) landraces in China. *Tree Genet Genomes*. 2012;8(5):975-990. doi 10.1007/s11295-012-0477-8

Cao K., Zheng Z., Wang L., Liu X., Zhu G., Fang W., Cheng S., … Li Y., Li H., Guo J., Xu X., Wang J. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol*. 2014;15:415. doi 10.1186/s13059-014-0415-1

Cao K.E., Zhou Z., Wang Q., Guo J., Zhao P., Zhu G., Fang W., Chen C., Wang X., Wang X., Tian Z., Wang L. Genome-wide association study of 12 agronomic traits in peach. *Nat Commun*. 2016;7(1):13246. doi 10.1038/ncomms13246

Cao K., Li Y., Deng C.H., Gardiner S.E., Zhu G., Fang W., Chen C., Wang X., Wang L. Comparative population genomics identified genomic regions and candidate genes associated with fruit domestication traits in peach. *Plant Biotechnol J*. 2019;17(10):1954-1970. doi 10.1111/pbi.13112

Cao K., Peng Z., Zhao X., Li Y., Liu K., Arus P., Zhu G., Deng S., Fang W., Chen C., Wang X., Wu J., Fei Z., Wang L. Pan-genome analyses of peach and its wild relatives provide insights into the genetics of disease resistance and species adaptation. *BioRxiv*. 2020. doi 10.1101/2020.07.13.200204

Cao K., Yang X., Li Y., Zhu G., Fang W., Chen C., Wang X., Wu J., Wang L. New high-quality peach (*Prunus persica* L. Batsch) genome assembly to analyze the molecular evolutionary mechanism of volatile compounds in peach fruits. *Plant J*. 2021;108(1):281-295. doi 10.1111/tpj.15439

Carter G.E. Jr., Brock M.M. Identification of peach cultivars through protein analysis. *HortScience*. 1980;15(3):292-293

Cartwright D.A., Troggio M., Velasco R., Gutin A. Genetic mapping in the presence of genotyping errors. *Genetics*. 2007;176(4):2521-2527. doi 10.1534/genetics.106.063982

Chagné D., Crowhurst R.N., Troggio M., Davey M.W., Gilmore B., Lawley C., Vanderzande S., … Wilhelm L., Van de Weg E., Gardiner S.E., Bassil N., Peace C. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One*. 2012;7(2):e31745. doi 10.1371/journal.pone.0031745

Chaparro J.X., Durham R.E., Moore G.A., Sherman W.B. Utilization of isozyme techniques to identify peach × 'Nonpareil' almond hybrids. *HortScience*. 1987;22(2):300-302. doi 10.21273/HORTSCI.22.2.300

Chaparro J.X., Werner D.J., O'Malley D., Sederoff R.R. Targeted mapping and linkage analysis of morphological isozyme, and RAPD markers in peach. *Theor Appl Genet*. 1994;87(7):805-815. doi 10.1007/BF00221132

Chesnokov Yu.V., Artem'eva A.M. Association mapping in plants (review). *Sel'skokhozyaystvennaya Biologiya = Agricultural Biology*. 2011;46(5):3-16 (in Russian)

Cirilli M., Baccichet I., Chiozzotto R., Silvestri C., Rossini L., Bassi D. Genetic and phenotypic analyses reveal major quantitative loci associated to fruit size and shape traits in a non-flat peach collection

(*P. persica* L. Batsch). *Hortic Res.* 2021;8:232. doi 10.1038/s41438-021-00661-5

Collard B.C.Y., Jahufer M.Z.Z., Brouwer J.B., Pang E.C.K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*. 2005;142:169-196. doi 10.1007/s10681-005-1681-5

da Silva Linge C., Cai L., Fu W., Clark J., Worthington M., Rawandoozi Z., Byrne D.H., Gasic K. Multi-locus genome-wide association studies reveal fruit quality hotspots in peach genome. *Front Plant Sci.* 2021;12:644799. doi 10.3389/fpls.2021.644799

Demirel S., Pehluvan M., Aslantaş R. Evaluation of genetic diversity and population structure of peach (*Prunus persica* L.) genotypes using inter-simple sequence repeat (ISSR) markers. *Genet Resour Crop Evol*. 2024;71(3):1301-1312. doi 10.1007/s10722-023-01691-9

Dettori M.T., Quarta R., Verde I. A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome.* 2001; 44(5):783-790. doi 10.1139/g01-065

Dirlewanger E., Moing A., Rothan C., Svanella L., Pronier V., Guye A., Plomion C., Monet R. Mapping QTLs controlling fruit quality in peach (*Prunus persica* (L.) Batsch). *Theor Appl Genet*. 1999;98: 18-31. doi 10.1007/s001220051035

Dirlewanger E., Cosson P., Tavaud M., Aranzana M., Poizat C., Zanetto A., Arús P., Laigret F. Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). *Theor Appl Genet.* 2002;105(1):127-138. doi 10.1007/s00122-002-0867-7

Dirlewanger E., Graziano E., Joobeur T., Garriga-Calderé F., Cosson P., Howad W., Arús P. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proc Natl Acad Sci USA.* 2004;101(23): 9891-9896. doi 10.1073/pnas.0307937101

Dirlewanger E., Cosson P., Boudehri K., Renaud C., Capdeville G., Tauzin Y., Laigret F., Moing A. Development of a second-generation genetic linkage map for peach [*Prunus persica* (L.) Batsch] and characterization of morphological traits affecting flower and fruit. *Tree Genet Genomes.* 2007;3:1-13. doi 10.1007/s11295-006-0053-1

Dirlewanger E., Claverie J., Iezzoni A.F., Wünsch A. Sweet and sour cherries: linkage maps, QTL detection and marker assisted selection. In: Folta K.M., Gardiner S.E. (Eds) Genetics and Genomics of Rosaceae. Plant Genetics and Genomics: Crops and Models. Vol. 6. New York, NY: Springer, 2009;291-313. doi 10.1007/978-0-387-77491-6_14

Dirlewanger E., Quero-García J., Le Dantec L., Lambert P., Ruiz D., Dondini L., Illa E., Quilot-Turion B., Audergon J.M., Tartarini S., Letourmy P., Arús P. Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three *Prunus* species: peach, apricot and sweet cherry. *Heredity.* 2012;109(5): 280-292. doi 10.1038/hdy.2012.38

Dodds P.N., Rathjen J.P. Plant immunity: towards an integrated view of plant pathogen interactions. *Nat Rev Genet.* 2010;11(8):539-548. doi 10.1038/nrg2812

Durham R.E., Moore G.A., Sherman W.B. Isozyme banding patterns and their usefulness as genetic markers in peach. *J Am Soc Hortic Sci.* 1987;112(6):1013-1018. doi 10.21273/JASHS.112.6.1013

Eduardo I., Pacheco I., Chietera G., Bassi D., Pozzi C., Vecchietti A., Rossini L. QTL analysis of fruit quality traits in two peach intraspecific populations and importance of maturity date pleiotropic effect. *Tree Genet Genomes*. 2011;7:323-335. doi 10.1007/s11295-010-0334-6

Elsadr H. A genome wide association study of flowering and fruit quality traits in peach [(*Prunus persica* (L.) Batsch]: Doctoral dissertation. University of Guelph, 2016

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. A robust, simple genotyping-by-sequencing (GBS) aproach for high diversity species. *PloS One.* 2011;6(5): e19379. doi 10.1371/journal.pone.0019379

Faust M., Timon B. Origin and dissemination of the peach. In: Janick J. (Ed.) Horticultural Reviews. John Wiley & Sons, Inc., 1995;331-379. doi 10.1002/9780470650585.ch10

Font i Forcada C., Oraguzie N., Igartua E., Moreno M.Á., Gogorcena Y. Population structure and marker-trait associations for pomological traits in peach and nectarine cultivars. *Tree Genet Genomes.* 2013;9:331-349. doi 10.1007/s11295-012-0553-0

Font i Forcada C., Guajardo V., Chin-Wo S.R., Moreno M.Á. Association mapping analysis for fruit quality traits in *Prunus persica* using SNP markers. *Front Plant Sci*. 2019;9:2005. doi 10.3389/fpls.2018.02005

Foolad M.R., Arulsekar S., Becerra V., Bliss F.A. A genetic map of *Prunus* based on an interspecific cross between peach and almond. *Theor Appl Genet*. 1995;91:262-269. doi 10.1007/BF00220887

Fu W., da Silva Linge C., Gasic K. Genome-wide association study of brown rot (*Monilinia* spp.) tolerance in peach. *Front Plant Sci.* 2021;12:635914. doi 10.3389/fpls.2021.635914

Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D.M., Burzynski-Chang E.A., … van der Knaap E., Huang S., Klee H.J., Giovannoni J.J., Fei Z. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet.* 2019;51(6):1044-1051. doi 10.1038/s41588-019-0410-2

Gasic K., Da Silva Linge C., Bianco L., Troggio M., Rossini L., Bassi D., Aranzana M.J., Arus P., Verde I., Peace C., Iezzoni A. Development and evaluation of a 9K SNP addition to the peach IPSC 9K SNP array v1. *HortScience*. 2019;54(9S):S188

Guajardo V., Solís S., Almada R., Saski C., Gasic K., Moreno M.Á. Genome-wide SNP identification in *Prunus* rootstocks germplasm collections using Genotyping-by-Sequencing: phylogenetic analysis, distribution of SNPs and prediction of their effect on gene function. *Sci Rep.* 2020;10(1):1467. doi 10.1038/s41598-020-58271-5

Guan L., Cao K., Li Y., Guo J., Xu Q., Wang L. Detection and application of genome-wide variations in peach for association and genetic relationship analysis. *BMC Genet.* 2019;20(1):101. doi 10.1186/s12863-019-0799-8

Hamblin M.T., Warburton M.L., Buckler E.S. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One.* 2007;2(12): e1367. doi 10.1371/journal.pone.0001367

Herrero J., Cambra M., Tabuenca M.C. Cartografía de Frutales de Hueso y Pepita. Zaragoza: Estación Experimental de Aula Dei (EEAD-CSIC), 1964

Hesse C.O. Peaches. In: Janick J., Moore J.N. (Eds) Advances in Fruit Breeding. West Lafayette, Ind.: Purdue University Press, 1975; 285-335

Hong J.H., Yi S.I., Kwon Y.S., Kim Y., Choi K.J. Genetic diversity analysis of peach [*Prunus persica* (L.) Batsch] varieties using SSR markers. *Korean J Breed Sci.* 2013;45(3):201-211. doi 10.9787/KJBS.2013.45.3.201

Howad W., Yamamoto T., Dirlewanger E., Testolin R., Cosson P., Cipriani G., Monforte A.J., Georgi L., Abbott A.G., Arus P. Mapping with a few plants: using selective mapping for microsatellite saturation of the Prunus reference map. *Genetics.* 2005;171(3):1305-1309. doi 10.1534/genetics.105.043661

Huang Z., Shen F., Chen Y., Cao K., Wang L. Preliminary identification of key genes controlling peach pollen fertility using genome-wide association study. *Plants*. 2021;10(2):242. doi 10.3390/plants10020242

Hübner S., Bercovich N., Todesco M., Mandel J.R., Odenheimer J., Ziegler E., Lee J.S., ... Kubach T., Muños S., Langlade N.B., Burke J.M., Rieseberg L.H. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants*. 2019;5(1):54-62. doi 10.1038/s41477-018-0329-0

International Peach Genome Initiative; Verde I., Abbott A.G., Scalabrin S., Jung S., Shu S., Marroni F., … Silva H., Salamini F., Schmutz J., Morgante M., Rokhsar D.S. The high-quality draft ge-

nome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet.* 2013; 45(5):487-494. doi 10.1038/ng.2586

Jayakodi M., Padmarasu S., Haberer G., Bonthala V.S., Gundlach H., Monat C., Lux T., ... Mayer K.F.X., Spannagl M., Li C., Mascher M., Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. 2020;588(7837):284-289. doi 10.1038/s41586-020-2947-8

Joiret M., Mahachie John J.M., Gusareva E.S., Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min*. 2019;12:11. doi 10.1186/s13040-019-0199-7

Jones N., Ougham H., Thomas H. Markers and mapping: we are all geneticists now. *New Phytol*. 1997;137(1):165-177. doi 10.1046/j.1469-8137.1997.00826.x

Joobeur T., Viruel M.A., de Vicente M.C., Jáuregui B., Ballester J., Dettori M.T., Verde I., Truco M.J., Messeguer R., Batlle I., Quarta R., Dirlewanger E., Arús P. Construction of a saturated linkage map for *Prunus* using an almond × peach F2 progeny. *Theor Appl Genet.* 1998;97:1034-1041. doi 10.1007/s001220050988

Jung S., Staton M., Lee T., Blenda A., Svancara R., Abbott A., Main D. GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res*. 2008; 36:D1034-D1040. doi 10.1093/nar/gkm803

Jung S., Ficklin S.P., Lee T., Cheng C.-H., Blenda A., Zheng P., Yu J., Bombarely A., Cho I., Ru S., Evans K., Peace C., Abbott A.G., Mueller L.A., Olmstead M.A., Main D. The genome database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res*. 2014;42: D1237-D1244. doi 10.1093/nar/gkt1012

Khlestkina E.K. Molecular markers in genetic studies and breeding. *Russ J Genet Appl Res.* 2014;4;236-244. https://link.springer.com/article/10.1134/S2079059714030022#citeas

Kim J.S., Ku Y.S., Park S.G., Kim S.H., Park H.W., Won S.Y. Anticipated polymorphic SSRs and their application based on next generation sequencing of *Prunus persica*. *Korean J Breed Sci*. 2021;53(4): 350-360. doi 10.9787/KJBS.2021.53.4.350

Koning-Boucoiran C.F., Esselink G.D., Vukosavljev M., van't Westende W.P., Gitonga V.W., Krens F.A., Voorrips R.E., van de Weg W.E., Schulz D., Debener T., Maliepaard C., Arens P., Smulders M.J. Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.). *Front Plant Sci*. 2015;6:249. doi 10.3389/fpls.2015.00249

Kuhn D.N., Livingstone D.S., Richards J.H., Manosalva P., Van den Berg N., Chambers A.H. Application of genomic tools to avocado (*Persea americana*) breeding: SNP discovery for genotyping and germplasm characterization. *Sci Hortic*. 2019;246:1-11. doi 10.1016/j.scienta.2018.10.011

Lambert P., Campoy J.A., Pacheco I., Mauroux J.B., Da Silva Linge C., Micheletti D., Bassi D., ... Pascal T., Troggio M., Aranzana M.J., Patocchi A., Arús P. Identifying SNP markers tightly associated with six major genes in peach [*Prunus persica* (L.) Batsch] using a high-density SNP array with an objective of marker-assisted selection (MAS). *Tree Genet Genomes.* 2016;12:121. doi 10.1007/s11295-016-1080-1

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E., Newburg L. Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics.* 1987;1(2):174-181. doi 10.1016/0888-7543(87)90010-3

Laucou V., Launay A., Bacilieri R., Lacombe T., Adam-Blondon A.-F., Bérard A., Chauveau A., … Maul E., Ponnaiah M., Töpfer R., Péros J.P., Boursiquot J.M. Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide SNPs. *PLoS One.* 2018;13(2):e0192540. doi 10.1371/journal.pone.0192540

Li X., Singh J., Qin M., Li S., Zhang X., Zhang M., Khan A., Zhang S., Wu J. Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (*Pyrus*). *Plant Biotechnol J*. 2019;17(8):1582-1594. doi 10.1111/pbi.13085

Li X., Wang J., Su M., Zhou J., Zhang M., Du J., Zhou H., ... Fang W., Wang L., Jia H., Gao Z., Ye Z. Single nucleotide polymorphism detection for peach gummosis disease resistance by genome-wide association study. *Front Plant Sci.* 2022;12:763618. doi 10.3389/fpls.2021.763618

Li X., Wang J., Su M., Zhang M., Hu Y., Du J., Zhou H., Yang X., Zhang X., Jia H., Gao Z., Ye Z. Multiple-statistical genome-wide association analysis and genomic prediction of fruit aroma and agronomic traits in peaches. *Hortic Res.* 2023;10(7):uhad117. doi 10.1093/hr/uhad117

Li Y.H., Zhou G., Ma J., Jiang W., Jin L.G., Zhang Z., Guo Y., ... Chang R.Z., Jiang Z., Jackson S.A., Li R., Qiu L.J. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol.* 2014;32(10):1045-1052. doi 10.1038/nbt.2979

Li Y., Cao K.E., Zhu G., Fang W., Chen C., Wang X., Zhao P., Guo J., Ding T., Guan L., Zhang Q., Guo W., Fei Z., Wang L. Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biol.* 2019;20(1):36. doi 10.1186/s13059-019-1648-9

Lirong W., Yong L., Gengrui Z., Weichao F., Changwen C., Ke C., Xinwei W. Peach genomics and breeding programs at Zhengzhou Fruit Research Institute, CAAS. *Acta Hortic*. 2020;1282:1-6. doi 10.17660/ActaHortic.2020.1282.1

Liu H., Cao K., Zhu G., Fang W., Chen C., Wang X., Wang L. Genome-wide association analysis of red flesh character based on resequencing approach in peach. *J Am Soc Hortic Sci.* 2019;144(3):209-216. doi 10.21273/JASHS04622-18

Liu J., Bao Y., Zhong Y., Wang Q., Liu H. Genome-wide association study and transcriptome of olecranon-type traits in peach (*Prunus persica* L.) germplasm. *BMC Genomics.* 2021;22(1):702. doi 10.1186/s12864-021-08017-y

Liu Y., Du H., Li P., Shen Y., Peng H., Liu S., Zhou G.A., … Wang Z., Zhu B., Han B., Liang C., Tian Z. Pan-genome of wild and cultivated soybeans. *Cell.* 2020;182(1):162-176. doi 10.1016/j.cell.2020.05.023

Mardis E.R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008;9(1):387-402. doi 10.1146/annurev.genom.9.081307

Mariette S., Tavaud M., Arunyawat U., Capdeville G., Millan M., Salin F. Population structure and genetic bottleneck in sweet cherry estimated with SSRs and the gametophytic self-incompatibility locus. *BMC Genet.* 2010;11:77. doi 10.1186/1471-2156-11-77

Marrano A., Martínez-García P.J., Bianco L., Sideli G.M., Di Pierro E.A., Leslie C.A., Stevens K.A., Crepeau M.W., Troggio M., Langley C.H., Neale D.B. A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density AxiomTM *J. regia* 700K SNP genotyping array. *Plant Biotechnol J.* 2019; 17(6):1027-1036. doi 10.1111/pbi.13034

Mas-Gómez J., Cantín C.M., Moreno M.Á., Prudencio Á.S., Gómez-Abajo M., Bianco L., Troggio M., Martínez-Gómez P., Rubio M., Martínez-García P.J. Exploring genome-wide diversity in the national peach (*Prunus persica*) germplasm collection at CITA (Zaragoza, Spain). *Agronomy*. 2021;11(3):481. doi 10.3390/agronomy11030481

Mas-Gómez J., Cantín C.M., Moreno M.Á., Martínez-García P.J. Genetic diversity and genome-wide association study of morphological and quality traits in peach using two Spanish peach germplasm collections. *Front Plant Sci.* 2022;13:854770. doi 10.3389/fpls.2022.854770

Meng G., Zhu G., Fang W., Chen C., Wang X., Wang L., Cao K. Identification of loci for single/double flower trait by combining genome-wide association analysis and bulked segregant analysis in peach (*Prunus persica*). *Plant Breed.* 2019;138(3):360-367. doi 10.1111/pbr.12673

Micali S., Vendramin E., Dettori M.T., Verde I. Genetics and genomics of stone fruits. In: Agricultural and Food Biotechnologies of *Olea europaea* and Stone Fruits. Bentham, 2015;243-307. doi 10.2174/9781608059935115010008

Micheletti D., Dettori M.T., Micali S., Aramini V., Pacheco I., Da Silva Linge C., Foschi S., ... Rossini L., Verde I., Laurens F., Arús P., Aranzana M.J. Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. *PloS One*. 2015;10(9):e0136803. doi 10.1371/journal.pone.0136803

Monet R. Peach genetics: past present and future. *Acta Hortic.* 1988;254:49-58. doi 10.17660/ActaHortic.1989.254.8

Monet R., Gibault B. Polymorphisme de l'alpha-amylase chez le pecher. Etude genetique. *Agronomie* (*France*). 1991;11(5):353-358

Monet R., Bastard Y., Gibault B. Genetic studies on the breeding of flat peaches. *Agronomie* (*France*). 1985;5(8):727-731

Monet R., Guye A., Roy M., Dachary N. Peach mendelian genetics: a short review and new results. *Agronomie*. 1996;16(5):321-329. doi 10.1051/agro:19960505

Montanari S., Bianco L., Allen B.J., Martínez-García P.J., Bassil N.V., Postman J., Knäbel M., … Langley C.H., Evans K., Dhingra A., Troggio M., Neale D.B. Development of a highly efficient Axiom™ 70 K SNP array for *Pyrus* and evaluation for high-density mapping and germplasm characterization. *BMC Genomics*. 2019;20(1):331. doi 10.1186/s12864-019-5712-3

Morozova O., Marra M.A. Aplications of next-generation sequencing technologies in functional genomics. *Genomics.* 2008;92(5):255-264. doi 10.1016/j.ygeno.2008.07.001

Nybom H., Lācis G. Recent large-scale genotyping and phenotyping of plant genetic resources of vegetatively propagated crops. *Plants.* 2021;10(2):415. doi 10.3390/plants10020415

Parfitt D.E., Arulsekar S., Ramming D.W. Identification of plum× peach hybrids by isoenzyme analysis. *HortScience*. 1985;20(2):246-248

Paterson A.H. Making genetic maps. In: Paterson A.H. Genome Mapping in Plants. Academic Press, 1996;23-39

Peace C., Bassil N., Main D., Ficklin S., Rosyara U.R., Stegmeir T., Sebolt A., Gilmore B., Lawley C., Mockler T.C., Bryant D.W., Wilhelm L., Iezzoni A. Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLoS One.* 2012;7(12):e48305. doi 10.1371/journal.pone.0048305

Pflieger S., Lefebvre V., Caranta C., Blattes A., Goffinet B., Palloix A. Disease resistance gene analogs as candidates for QTLs involved in pepper-pathogen interactions. *Genome.* 1999;42(6):1100-1110

Pozzi C., Vecchietti A. Peach structural genomics. In: Folta K.M., Gardiner S.E. (Eds) Genetics and Genomics of Rosaceae. Plant Genetics and Genomics: Crops and Models. Vol. 6. New York, NY: Springer, 2009;235-257. https://link.springer.com/book/10.1007/978-0-387-77491-6

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. PLINK: A tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007;81(3):559-575. doi 10.1086/519795

Quarta R., Cedrola C., Dettori M.T., Verde I. QTL analysis of agronomic traits in a BC1 peach population. *Acta Hortic.* 2002;592:291-297. doi 10.17660/ActaHortic.2002.592.41

Quilot B., Wu B.H., Kervella J., Génard M., Foulongne M., Moreau K. QTL analysis of quality traits in an advanced backcross between *Prunus persica* cultivars and the wild relative species *P. davidiana*. *Theor Appl Genet.* 2004;109(4):884-897. doi 10.1007/s00122-004-1703-z

Rasheed A., Hao Y., Xia X., Khan A., Xu Y., Varshney R.K., He Z. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol Plant*. 2017;10(8):1047-1064. doi 10.1016/j.molp.2017.06.008

Ru S., Main D., Evans K., Peace C. Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. *Tree Genet Genomes.* 2015;11:8. doi 10.1007/s11295-015-0834-5

Salazar J.A., Ruiz D., Campoy J.A., Sánchez-Pérez R., Crisosto C.H., Martínez-García P.J., Blenda A., Jung S., Main D., Martínez-Gómez P., Rubio M. Quantitative trait loci (QTL) and Mendelian trait loci (MTL) analysis in *Prunus*: a breeding perspective and beyond. *Plant Mol Biol Rep*. 2013;32:1-18. doi 10.1007/s11105-013-0643-7

Scorza R. Gene transfer for the genetic improvement of perennial fruit and nut crops. *HortScience*. 1991;26(8):1033-1035

Scorza R., Okie W.R. Peaches (*Prunus*). *Acta Hortic.* 1991;290:177-234. doi 10.17660/ActaHortic.1991.290.5

Scorza R., Mehlenbacher S.A., Lightner G.W. Inbreeding and coancestry of freestone peach cultivars of the eastern United States and implications for peach germplasm improvement. *J Am Soc Hortic Sci.* 1985;110(4):547-552. doi 10.21273/JASHS.110.4.547

Siberchicot A., Bessy A., Gueguen L., Marais G.A. Mareymap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. *Genome Biol Evol.* 2017;9(10):2506-2509. doi 10.1093/gbe/evx178

Smykov A., Shoferistov E., Korzin V., Mesyats N., Saplev N. Promising directions in the selection of peach, apricot and nectarine. *E3S Web Conf.* 2021;254:01010. doi 10.1051/e3sconf/202125401010

Sosinski B., Gannavarapu M., Hager L.D., Beck L.E., King G.J., Ryder C.D., Rajapakse S., Baird W.V., Ballard R.E., Abbott A.G. Characterization of microsatellite markers in peach [*Prunus persica* (L.) Batsch]. *Theor Appl Genet.* 2000;101:421-428. doi 10.1007/s001220051499

Tan Q., Li S., Zhang Y., Chen M., Wen B., Jiang S., Chen X., Fu X., Li D., Wu H., Wang Y., Xiao W., Li L. Chromosome-level genome assemblies of five *Prunus* species and genome-wide association studies for key agronomic traits in peach. *Hortic Res.* 2021;8(1):213. doi 10.1038/s41438-021-00648-2

Tanksley S.D., Young N.D., Paterson A.H., Bonierbale M.W. RFLP mapping in plant-breeding – new tools for an old science. *Nat Biotechnol.* 1989;7:257-264. doi 10.1038/nbt0389-257

Thurow L.B., Gasic K., Bassols Raseira M.C., Bonow S., Marques Castro C. Genome-wide SNP discovery through genotyping by sequencing, population structure, and linkage disequilibrium in Brazilian peach breeding germplasm. *Tree Genet Genomes.* 2020;16:10. doi 10.1007/s11295-019-1406-x

Trifonova A.A., Boris K.V., Mesyats N.V., Tsiupka V.A., Smykov A.V., Mitrofanova I.V. Genetic diversity of peach cultivars from the collection of the Nikita Botanical Garden based on SSR markers. *Plants.* 2021;10(12):2609. doi 10.3390/plants10122609

Van Ooijen J.W. Joinmap® 4. Software for the calculation of genetic linkage maps in experimental populations. ScienceOpen, Inc., 2006

Van Ooijen J.W. MapQTL® 6. Software for the mapping of quantitative trait loci in experimental populations of diploid species. ScienceOpen, Inc., 2009

Verde I., Bassil N., Scalabrin S., Gilmore B., Lawley C.T., Gasic K., Micheletti D, ... Aranzana M.J., Arús P., Iezzoni A., Morgante M., Peace C. Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PloS One.* 2012;7(4):e35668. doi 10.1371/journal.pone.0035668

Verde I., Jenkins J., Dondini L., Micali S., Pagliarani G., Vendramin E., Paris R., ... Shu S., Grimwood J., Tartarini S., Dettori M.T., Schmutz J. The Peach v2. 0 release: high-resolution linkage map-

ping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics.* 2017;18(1):225. doi 10.1186/s12864-017-3606-9

Voorrips R.E. MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered*. 2002;93(1):77-78. doi 10.1093/jhered/93.1.77

Wang L., Zhu G., Fang W. Peach germplasm and breeding programs at Zhengzhou in China. *Acta Hortic.* 2001;592:177-182. doi 10.17660/ActaHortic.2002.592.25

Werner D.J., Okie W.R. A history and description of the *Prunus persica* plant introduction collection. *HortScience.* 1998;33(5):787-793. doi 10.21273/HORTSCI.33.5.787

Winter P., Kahl G. Molecular marker technologies for plant improvement. *World J Microbiol Biotechnol*. 1995;11(4):438-448. doi 10.1007/BF00364619

Yamamoto T., Mochida K., Hayashi T. Shanhai Suimitsuto, one of the origins of Japanese peach cultivars. *J Japan Soc Hortic Sci*. 2003; 72(2):116-121

Yu J.M., Zhang Z.W., Zhu C.S., Tabanao D.A., Pressoir G., Tuinstra M.R., Kresovich S., Todhunter R.J., Buckler E.S. Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. *Plant Genome.* 2009; 2(1):63-77. doi 10.3835/plantgenome2008.09.0009

Yu Y., Fu J., Xu Y., Zhang J., Ren F., Zhao H., Tian S., … Wang G., Ma R., Jiang Q., Wei J., Xie H. Genome re-sequencing reveals the evolutionary history of peach fruit edibility. *Nat Commun.* 2018; 9(1):5404. doi 10.1038/s41467-018-07744-3

Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q, ... Xu Q., Wang Z.X., Wei X., Han B., Huang X. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet.* 2018;50(2):278-284. doi 10.1038/s41588-018-0041-z

Zurn J.D., Nyberg A., Montanari S., Postman J., Neale D., Bassil N. A new SSR fingerprinting set and its comparison to existing SSR- and SNP-based genotyping platforms to manage *Pyrus* germplasm resources. *Tree Genet Genomes.* 2020;16:72. doi 10.1007/s11295-020-01467-7

Conflict of interest. The authors declare no conflict of interest.

Received July 25, 2024. Revised January 24, 2025. Accepted January 27, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):370-379

doi 10.18699/vjgb-25-40

Variability of carotenoid synthesis and degradation genes in Russian durum wheat cultivars

A.A. Trifonova [1] ✉, K.V. Boris [1], L.V. Dedova [1], P.N. Malchikov [2], A.M. Kudryavtsev [1]

[1] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
[2] Samara Scientific Research Agriculture Institute named after N.M. Tulajkov – Branch of Samara Federal Research Scientific Center of the Russian Academy of Sciences, Bezenchuk, Samara region, Russia
✉ aichka89@mail.ru

Abstract. Yellow index is an important quality parameter of durum wheat cultivars, associated with carotenoid pigment content in grain and the level of carotenoid degradation during processing, and determining the yellow color of products made from durum wheat. Molecular markers of genes that influence carotenoid content can be used for fast identification of valuable genotypes and development of new high-quality durum wheat cultivars. The aim of the study was to investigate the domestic durum wheat gene pool using molecular markers of the yellow pigment synthesis (*Psy-A1*) and degradation (*Lpx-B1*) genes. Using two markers of the phytoene synthase *Psy-A1* gene (PSY1-A1_STS and YP7A-2) and three markers of the lipoxygenase *Lpx-B1* locus (Lpx-B1.1a/1b, Lpx-B1.1c and Lpx-B1.2/1.3), 54 durum wheat cultivars were studied for the first time. For 38 cultivars, yellow pigment content in grain was also assessed. The detected allelic variation of the phytoene synthase *Psy-A1* and lipoxygenase *Lpx-B1* genes was rather low. The most common *Psy-A1* alleles among the studied cultivars were *Psy-A1l* for the PSY1-A1_STS marker and *Psy-A1d* for the YP7A-2 marker, identified in 51 cultivars and associated with high carotenoid content. According to the markers of the *Lpx-B1* locus, haplotype II, associated with medium lipoxygenase activity, identified in 43 cultivars, was predominant. Haplotype III, associated with low enzyme activity, was identified in only three winter durum wheat cultivars (Donchanka, Gelios and Leucurum 21). Despite the predominance of allelic variants associated with increased carotenoid content and moderate lipoxygenase activity, the studied cultivars had different levels of yellow pigment content in grain, from low to high.

Key words: yellow pigment; yellow index; molecular markers; phytoene synthase; lipoxygenase; genetic diversity

For citation: Trifonova A.A., Boris K.V., Dedova L.V., Malchikov P.N., Kudryavtsev A.M. Variability of carotenoid synthesis and degradation genes in Russian durum wheat cultivars. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):370-379. doi 10.18699/vjgb-25-40

Funding. The study was funded by the Russian Science Foundation project number 23-76-01079, https://rscf.ru/project/23-76-01079/.

Разнообразие отечественных сортов твердой пшеницы по генам синтеза и деградации каротиноидов в зерне

А.А. Трифонова [1] ✉, К.В. Борис [1], Л.В. Дедова [1], П.Н. Мальчиков [2], А.М. Кудрявцев [1]

[1] Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия
[2] Самарский научно-исследовательский институт сельского хозяйства им. Н.М. Тулайкова – филиал Самарского федерального исследовательского центра Российской академии наук, пгт Безенчук, Самарская область, Россия
✉ aichka89@mail.ru

Аннотация. Индекс желтизны – важный параметр качества сортов твердой пшеницы, связанный с содержанием каротиноидов в зерне и уровнем их деградации в процессе его переработки и определяющий желтый цвет продуктов, получаемых из твердой пшеницы. Применение молекулярно-генетических маркеров генов, влияющих на содержание каротиноидов, позволяет быстро идентифицировать ценные для селекции генотипы для ускоренного создания новых высококачественных отечественных сортов твердой пшеницы. Целью работы стало изучение отечественного генофонда твердой пшеницы с помощью молекулярных маркеров генов синтеза (*Psy-A1*) и деградации (*Lpx-B1*) желтых пигментов в зерне. С использованием двух маркеров гена фитоенсинтазы *Psy-A1* (PSY1-A1_STS и YP7A-2) и трех маркеров локуса липоксигеназы *Lpx-B1* (Lpx-B1.1a/1b, Lpx-B1.1c и Lpx-B1.2/1.3) впервые были исследованы 54 сорта твердой пшеницы, 38 из которых охарактеризованы по уровню содержания желтых пигментов в зерне. Аллельное разнообразие изученных отечественных сортов твердой пшеницы по генам фитоенсинтазы *Psy-A1* и липоксигеназы *Lpx-B1* оказалось достаточно низким. Наиболее распространенными в выборке были аллельные варианты *Psy-A1l* по маркеру

PSY1-A1_STS и *Psy-A1d* по маркеру YP7A-2, выявленные у 51 образца и ассоциированные с высокими значениями индекса желтизны. По маркерам локуса *Lpx-B1* в выборке преобладал гаплотип II, связанный со средней активностью липоксигеназы, который был идентифицирован у 43 образцов. Гаплотип III, ассоциированный с низкой активностью фермента, выявлен только у трех озимых сортов (Дончанка, Гелиос и Леукурум 21). Несмотря на преобладание аллельных вариантов, связанных с повышенным содержанием каротиноидов и средней активностью липоксигеназы, исследуемые образцы имели различный уровень содержания желтых пигментов – от низкого до высокого.

Ключевые слова: желтые пигменты; индекс желтизны; молекулярные маркеры; фитоенсинтаза; липоксигеназа; генетическое разнообразие

Introduction

Durum wheat (*Triticum durum* Desf.) is an important cereal crop. Hardness, amber-yellow color and high content of protein and gluten in durum wheat grain allow making high-quality pasta, as well as semolina, bulgur and couscous (Shevchenko et al., 2018). In Russia, about 650–700 thousand tons of durum wheat are produced annually. Currently, the domestic market's demand for this crop is growing and is estimated at 1.5 million tons (Natoli et al., 2021). At the same time, our country has the capacity to meet the growing need for durum wheat, as well as exports. There is enough arable land, and the conditions of the Volga, Siberia and Urals steppe regions allow to produce a sufficient amount of high-quality durum wheat grain (Shevchenko et al., 2018; Natoli et al., 2021). Currently, the State Register of Varieties and Hybrids of Agricultural Plants Admitted for Usage (National list) (2024) includes 71 spring and 37 winter durum wheat cultivars, adapted to various growing regions. Developing new domestic cultivars with high quality parameters for pasta production that follow international standards will help to satisfy the growing demand of processing companies.

Yellow index is one of the main quality parameters of durum wheat grain affecting the yellow color of pasta, which is important to consumers (Colasuonno et al., 2019; Requena-Ramirez et al., 2022). Yellow index largely depends on the genotype, so developing domestic cultivars with high yellow index is justified and relevant (Malchikov, Myasnikova, 2020).

Yellow index is a complex trait that is associated with the content of yellow pigments, mainly carotenoids, in grain and the level of their degradation during processing (Colasuonno et al., 2019; Parada et al., 2020). Carotenoids not only provide the yellow color of the grain and its end products, but are also important for human nutrition, as precursors of vitamin A (Ficco et al., 2014). There is a significant positive correlation between yellow index and yellow pigment content in grain, and these indicators are often used to characterize the color of durum wheat grain and end products (Digesu et al., 2009; Blanco et al., 2011; Campos et al., 2016).

In durum wheat breeding in Russia, there has been a significant increase in the yellow index, especially in recently released cultivars (Vasil'chuk, 2001; Malchikov, Myasnikova, 2020). However, at present, breeding centers working on increasing carotenoid concentration in grain, semolina and end products mainly use traditional breeding methods. To accelerate the breeding process, it is necessary to use modern molecular genetic methods, e. g., for the identification of alleles associated with high yellow index.

Phytoene synthase (PSY, EC 2.5.1.32) is the major enzyme of carotenoid accumulation in the endosperm, which catalyzes the first stage of carotenoid biosynthesis (Gallagher et al., 2004). Of the three known PSY isoforms, PSY-1, which is active in maturing grain as well as in young leaves, plays the most important role. As previously shown, the *Psy-A1* and *Psy-B1* genes encoding PSY-1 are located on chromosomes 7A and 7B respectively and are linked to the major QTLs associated with yellow pigment content in durum wheat. Of the two genes, *Psy-A1* has a greater influence on carotenoid content, explaining up to 50 % of phenotypic variability (Colasuonno et al., 2019). Several allelic variants of the *Psy-A1* gene associated with insertions/deletions in the third and fourth introns and related to different carotenoid content in grain have been identified in common and durum wheat (He et al., 2008, 2009a; Singh et al., 2009).

Various markers (YP7A, YP7A-2, PSY1-A1_STS, Psy-A1SSR) have been developed to identify alleles of the *Psy-A1* gene (He et al., 2008, 2009a, b; Singh et al., 2009; Patil et al., 2018). These markers were previously used to study the allelic diversity of the phytoene synthase gene in landraces and modern foreign wheat cultivars (Singh et al., 2009; Campos et al., 2016; Parada et al., 2020), as well as in durum wheat breeding lines (Campos et al., 2016; Patil et al., 2018). The association of the identified allelic variants with different levels of yellow index was confirmed, and effectiveness of these markers for breeding was shown (Campos et al., 2016).

One of the main enzymes leading to the degradation of carotenoids during durum wheat grain processing and the bleaching of the end products is lipoxygenase (LOX, EC 1.13.11.12), which catalyzes the oxidation of polyunsaturated fatty acids (Verlotta et al., 2010; Colasuonno et al., 2019). Of the loci encoding various lipoxygenase isoforms in durum wheat (*Lpx-1*, *Lpx-2*, *Lpx-3*), the *Lpx-B1* locus plays the major role at the final stages of grain maturation, accounting for 36 to 54 % of the enzyme activity variation (Carrera et al., 2007; Verlotta et al., 2010; Parada et al., 2020). The *Lpx-B1* locus is located on the short arm of chromosome 4B and includes three related genes: *Lpx-B1.1*, *Lpx-B1.2* and *Lpx-B1.3* (Verlotta et al., 2010). The differences between these genes and their allelic variants are due to the presence of DNA transposon of the MITE (Miniature Inverted-Repeat Transposable Element) group (Hessler et al., 2002; Carrera et al., 2007), the transposition of which led to a large deletion in the sequence of the *Lpx-B1.1* gene and a significant decrease in lipoxygenase activity (Carrera et al., 2007; Verlotta et al., 2010). Several molecular markers have been developed to identify the genes of the *Lpx-B1* locus and their allelic variants (Verlotta et al., 2010; Parada et al., 2020). In previous studies of foreign durum wheat cultivars using these markers, several different combinations between the alleles and genes of the *Lpx-B1* locus (haplotypes) associated

with different levels of lipoxygenase activity were reported (Verlotta et al., 2010; Parada et al., 2020).

The use of the mentioned markers of phytoene synthase and lipoxygenase genes to study domestic durum wheat material will allow to characterize its allelic diversity for the first time and to assess its potential for breeding. The use of appropriate markers for the selection of breeding material and the involvement of genotypes with target alleles into the breeding process will significantly accelerate the development of durum wheat cultivars with high-quality grain.

The aim of the work was to study domestic durum wheat cultivars differing in the level of yellow pigment content using molecular markers of the *Psy-A1* gene and the *Lpx-B1* locus and to compare the results with data on the variability of the foreign durum wheat gene pool.

Materials and methods

Plant material. For the study, 54 spring and winter durum wheat cultivars from the collections of the Samara Federal Research Scientific Center, Russian Academy of Sciences and Vavilov Institute of General Genetics of the Russian Academy of Sciences were selected (Table 1). Of the selected cultivars, 44 (two foreign and 42 domestic cultivars from various breeding centers) are included in the State Register of Varieties and Hybrids of Agricultural Plants Admitted for Usage (National list) (2024). Cultivars Langdon and Giusto were used as references.

DNA was isolated from five-day-old seedlings according to the standard CTAB protocol (Doyle J.J., Doyle J.L., 1990) with minor modifications. For each cultivar, two DNA samples from individual plants were obtained, and further analysis was carried out with two repetitions.

Phytoene synthase (*Psy-A1*) and lipoxygenase (*Lpx-B1*) gene markers. Genotyping of the studied cultivars was carried out using SCAR markers of the *Psy-A1* and *Lpx-B1* genes. The primer sequences and annealing temperatures are presented in Table 2.

PCR reactions were performed in a GeneAmp 9700 (Applied Biosystems, USA) thermal cycler. PCR reaction mixture 15 µl in volume contained 20 ng of genomic DNA, 0.3 µM of each primer (Syntol, Russia), 0.16 mM dNTPs, 1.6 mM $MgCl_2$, 1 u Taq polymerase and 1x standard PCR buffer (Dialat LTD., Russia). To determine PCR fragment sizes, GeneRuler 100 bp DNA ladder (Thermo Fisher Scientific, USA) was used. After amplification, PCR products were separated in 1.5 % agarose gels, stained with ethidium bromide, analyzed on a UV-light box and photographed.

Yellow pigment content. For 38 cultivars from the Samara Federal Research Scientific Center of the Russian Academy of Sciences, total yellow pigment content in grain was assessed (Table 1). To assess the yellow pigment content, cultivars were grown for three years, from 2021 to 2023, in the Samara Scientific Research Agriculture Institute field. The evaluation of yellow pigment content was made by extraction of total pigment in water-saturated *n*-butanol followed by photocolorimetric quantification of the absorbance of extract at 440–450 nm using a KFK-3 M spectrophotometer. For each sample, 7.0 g of semolina were taken, placed in a 20 × 220 mm tube with a stopper, which was then filled with 35 ml of water-saturated *n*-butanol, shaken vigorously for one minute and left for extraction in a darkened room for 18 hours at room temperature. Then the solution was filtered through a pleated filter into clean tubes. The yellow pigment content was evaluated using a spectrophotometer in a cuvette with a working distance of 10 mm. The pigment content in parts per million parts of semolina (ppm) was calculated by multiplying the obtained value by a coefficient of 16.632. For convenience, the obtained value was converted into microgram percent by multiplying it by 100 (100 µg% = 1 ppm) (Methods for Assessing…, 1971). Measurements were taken for each of the two plants of one cultivar, and then the average value was determined. The yellow pigment content was considered high if it was more than 500 µg%, intermediate – 401–500 µg%, and low – 200–400 µg%.

The influence of environmental conditions in different years was assessed based on the average value of pigment content in grain in the experiment. According to this principle, the years were arranged in the following order: 2022 with the maximum (546.9 µg%), 2021 with the intermediate (476.2 µg%), 2023 with the minimum pigment content (402.1 µg%). The results were analyzed by the two-ways analysis of variance (ANOVA) using MS Excel. The parameters of general, specific adaptability (GAC_i, SAC_i) and stability (S_{gi}) of the trait were calculated according to the method of A.V. Kilchevsky, L.V. Khotyleva (1997). The regression coefficient (b_i) that measures the response of the cultivar to varying environments was determined following S.A. Eberhart, W.A. Russell (1966) as presented by A.V. Kilchevsky, L.V. Khotyleva (1997).

Results

In the present study, 54 durum wheat cultivars were analyzed using two markers of the *Psy-A1* phytoene synthase gene: PSY1-A1_STS and YP7A-2, and three markers of the *Lpx-B1* lipoxygenase locus: Lpx-B1.1a/1b, Lpx-B1.1c and Lpx-B1.2/1.3 (Table 2). Clear and reproducible results were obtained for all samples, coinciding for the two studied samples of each cultivar.

The PSY1-A1_STS marker identifies alleles *Psy-A1a* (1,776 bp), *Psy-A1l* (1,089 bp) and *Psy-A1o* (897 bp). Cultivar Langdon, for which the presence of the *Psy-A1l* allele was previously shown (Singh et al., 2009), was used as a reference.

Using the PSY1-A1_STS marker, the *Psy-A1l* allele was detected in 51 studied cultivars, including Langdon. The *Psy-A1o* allele was identified in two cultivars Krasnokutka 13 and Donchanka, and the *Psy-A1a* allele in cultivar Kurant (Fig. 1, Table 3). In cultivars Krasnokutka 13, Donchanka and Kurant, an additional ~1,100 bp fragment was amplified with the PSY1-A1_STS marker, which was not taken into account in further analysis (Fig. 1). It was previously shown that the presence of an additional ~1,100 bp fragment together with the *Psy-A1o* or *Psy-A1a* alleles occurs due to cross-amplification of the *Psy-B1n* allele of the *Psy-B1* locus (Singh et al., 2009; Campos et al., 2016).

The YP7A-2 marker allows detection of the *Psy-A1d* (1,001 bp) and *Psy-A1e* (1,686 bp) alleles. Cultivar Langdon, for which the presence of the *Psy-A1d* allele was previously shown (He et al., 2009b), was used as a reference.

With the YP7A-2 marker, the *Psy-A1d* allele was identified in 51 cultivars, including Langdon, the *Psy-A1e* allele

Table 1. Durum wheat cultivars used in the study and data on yellow pigment content in grain

No.	Cultivar	Breeding center**	Form	Yellow pigment content, μg%***
1	Aksinit	Agrarian Scientific Center Donskoy	Winter	–
2	Alejskaya	Federal Altai Scientific Center for Agrobiotechnology	Spring	–
3	Altajskaya niva*			334.7
4	Altajskij yantar*			347.3
5	Amazonka	Agrarian Scientific Center Donskoy	Winter	–
6	Annushka	Federal Center of Agriculture Research of the South-East Region	Spring	380.7
7	Bezenchukskaya 139*	Samara Federal Research Scientific Center	Spring	357.3
8	Bezenchukskaya 182	Samara Federal Research Scientific Center and Federal Research Centre of Biological Systems and Agrotechnologies	Spring	417.7
9	Bezenchukskaya 205	Samara Federal Research Scientific Center	Spring	503.0
10	Bezenchukskaya 209			473.0
11	Bezenchukskaya 210			566.0
12	Bezenchukskaya zolotistaya	Samara Federal Research Scientific Center and VolgaSemMarket LLC	Spring	687.7
13	Bezenchukskaya krepost	Samara Federal Research Scientific Center	Spring	656.0
14	Bezenchukskaya niva			534.0
15	Bezenchukskaya stepnaya			539.7
16	Bezenchukskaya yubilejnaya			498.7
17	Bezenchukskij vektor*			–
18	Bezenchukskij podarok			517.0
19	Burbon	Agroliga Plant Selection Center Ltd. and Agroliga Semena Ltd.	Spring	468.0
20	Valentina	Federal Center of Agriculture Research of the South-East Region	Spring	487.7
21	Volnodonskaya	Federal Rostov Agrarian Scientific Center	Spring	407.3
22	Galla*	Federal Center of Agriculture Research of the South-East Region	Spring	–
23	Gelios	Agrarian Scientific Center Donskoy	Winter	–
24	Donskaya elegiya	Federal Rostov Agrarian Scientific Center	Spring	370.3
25	Donchanka	Agrarian Scientific Center Donskoy	Winter	–
26	Zhemchuzhina Sibiri	Omsk Agrarian Scientific Center	Spring	572.0
27	Zolotaya*	Samara Federal Research Scientific Center	Spring	553.0
28	Krasnokutka 13	Federal Center of Agriculture Research of the South-East Region	Spring	383.0
29	Kurant	Agrarian Scientific Center Donskoy	Winter	–
30	Leucurum 21	National Grain Center P.P. Lukyanenko	Winter	–
31	Luch 25	Federal Center of Agriculture Research of the South-East Region	Spring	436.0
32	Lyudmila	Federal Center of Agriculture Research of the South-East Region and Saraktashkhleboprodukt LLC	Spring	–
33	Marina	Samara Federal Research Scientific Center	Spring	452.3
34	Nikolasha	National Grain Center P.P. Lukyanenko and Federal Center of Agriculture Research of the South-East Region	Spring	383.7
35	Oasis	Federal Altai Scientific Center for Agrobiotechnology	Spring	458.3
36	Omskij izumrud	Omsk Agrarian Scientific Center	Spring	539.7
37	Omskij korund			–
38	Omskaya stepnaya			–

Table 1 (end)

No.	Cultivar	Breeding center**	Form	Yellow pigment content, μg%***
39	Orenburgskaya 21	Federal Research Centre of Biological Systems and Agrotechnologies	Spring	–
40	Pamyati Chekhovicha*	Samara Federal Research Scientific Center	Spring	649.0
41	Pamyati Yanchenko	Federal Altai Scientific Center for Agrobiotechnology and EkoNiva-Semena LLC	Spring	413.7
42	Sladunica*	Federal Center of Agriculture Research of the South-East Region and National Grain Center P.P. Lukyanenko	Spring	–
43	Salyut Altaya	Federal Altai Scientific Center for Agrobiotechnology and EkoNiva-Semena LLC	Spring	358.7
44	Saratovskaya zolotistaya	Federal Center of Agriculture Research of the South-East Region	Spring	564.7
45	SY Atlante	SYNGENTA CROP PROTECTION AG (Switzerland)	Spring	446.3
46	Taganrog	Agroliga Plant Selection Center Ltd. and Agroliga Semena Ltd.	Spring	606.7
47	Tessadur	SAATBAU LINZ EGEN (Austria)	Spring	492.3
48	Triada	Samara Federal Research Scientific Center and Agroliga Plant Selection Center Ltd.	Spring	415.7
49	Harkovskaya 46	Ukrainian Research Institute of Plant Growing, Breeding and Genetic	Spring	433.3
50	Yadrica	National Grain Center P.P. Lukyanenko	Spring	367.0
51	Yarina			442.0
52	Yasenka			540.3
53	Langdon*	USA	Spring	–
54	Giusto*	Italy	Spring	–

* Not included in the State Register of Varieties and Hybrids of Agricultural Plants Admitted for Usage (National list). Data on the breeding center provided by the Samara Federal Research Scientific Center of the Russian Academy of Sciences.
** According to the State Register of Varieties and Hybrids of Agricultural Plants Admitted for Usage (National list).
*** Average value for 2021–2023.

Table 2. Phytoene synthase and lipoxygenase gene markers used in this study

Marker	Primer sequences	Gene/Allele	Fragment size, bp	Annealing temperature, °C	Reference
PSY1-A1_STS	F-GTGGATATTCCCTGTCAGCATC	*Psy1-A1o* – *Psy1-A1l* – *Psy1-A1a*	897 – 1,089 – 1,776	56	Singh et al., 2009
	R-GCCTCCTCGAAGAACATCCTC				
YP7A-2	F-GCCAGCCCTTCAAGGACATG	*Psy1-A1d* – *Psy1-A1e*	1,001 – 1,686	60	He et al., 2009a
	R-CAGATGTCGCCACACTGCCA				
Lpx-B1.1a/1b	F-GCAGGCGCTGGAAAGCAACAGGC	*Lpx-B1.1a* – *Lpx-B1.1b*	1,320 – 1,246	68	Verlotta et al., 2010
	R-GCGCTCTAACTCCGCGTACTCG				
Lpx-B1.1c	F-CCAAGATGATACTGGGCGGGC	*Lpx-B1.1c*	1,558	67	Verlotta et al., 2010
	R-CGCCGCCTTGCCGTGGTTGG				
Lpx-B1.2/1.3	F-GAACCGAGAGGTGAGAGCGTGCCTGATC	*Lpx-B1.2* – *Lpx-B1.3*	1,785 – 1,709	62	Parada et al., 2020
	R-GTGGTCGGAGGTGTTGGGGTAGAGC				

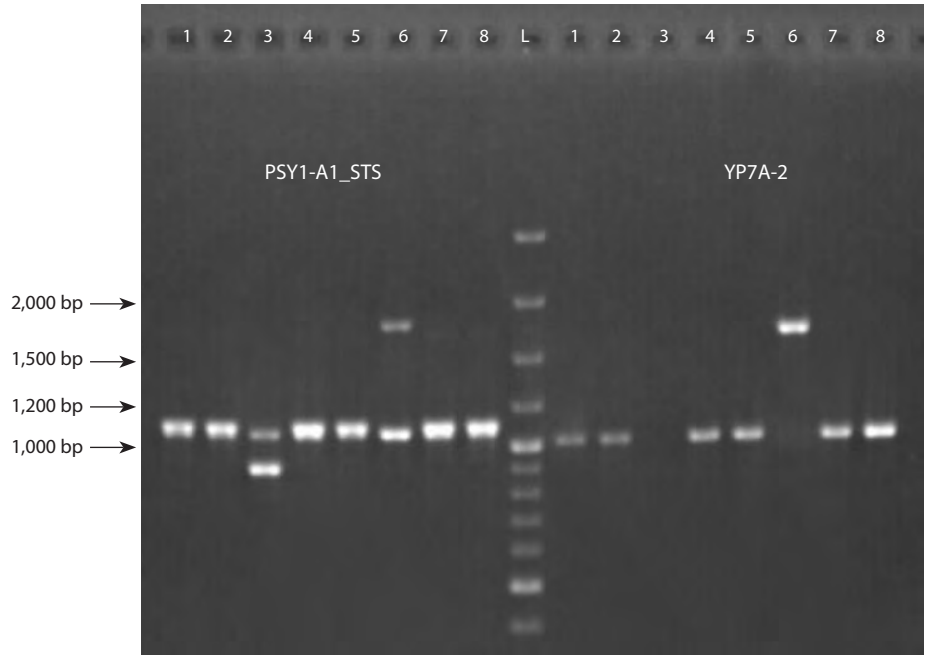


Fig. 1. Results of the *Psy-A1* alleles identification with markers PSY1-A1_STS and YP7A-2 in durum wheat cultivars: 1 – Aksinit; 2 – Alejskaya; 3 – Donchanka; 4 – Zhemchuzhina Sibiri; 5 – Zolotaya; 6 – Kurant; 7 – Leucurum 21; 8 – Langdon; L – marker GeneRuler 100 bp Plus.

Table 3. Alleles of the *Psy-A1* gene identified in the studied durum wheat cultivars

Allele	Number of cultivars	Frequency, %	Cultivars
Marker PSY1-A1_STS			
Psy-A1a (1,776 bp)	1	1.85	Kurant
Psy-A1l (1,089 bp)	51	94.45	51 cultivars
Psy-A1o (897 bp)	2	3.70	Krasnokutka 13, Donchanka
Marker YP7A-2			
Psy-A1d (1,001 bp)	51	94.45	51 cultivars
Psy-A1e (1,686 bp)	1	1.85	Kurant
No amplification	2	3.70	Krasnokutka 13, Donchanka

was identified in Kurant, and the absence of amplification products was detected in Krasnokutka 13 and Donchanka (Fig. 1, Table 3).

Analysis of the *Lpx-B1.1*, *Lpx-B1.2* and *Lpx-B1.3* lipoxygenase genes variability in 54 durum wheat cultivars was also performed. The allelic state of the *Lpx-B1.1* gene was analyzed using two markers: Lpx-B1.1a/1b was used to distinguish between the *Lpx-B1.1a* (1,320 bp) and *Lpx-B1.1b* (1,246 bp) alleles, and Lpx-B1.1c, to identify the *Lpx-B1.1c* allele (1,558 bp) (Verlotta et al., 2010). To identify the *Lpx-B1.2* (1,785 bp) and *Lpx-B1.3* (1,709 bp) genes, the Lpx-B1.2/1.3 marker (Parada et al., 2020) was used (Table 2). Cultivar Giusto, for which the presence of the *Lpx-B1.2* gene and the *Lpx-B1.1c* allele was previously shown (Verlotta et al., 2010), was used as a reference.

Allele *Lpx-B1.1a* was identified in 45 cultivars, *Lpx-B1.1b*, in five cultivars (Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha and Saratovskaya zolotistaya), and *Lpx-B1.1c*, in four cultivars (Gelios, Donchanka, Leucurum 21 and Giusto) (Fig. 2, Table 4).

Using the Lpx-B1.2/1.3 marker, the *Lpx-B1.2* gene was detected in 47 cultivars, and the *Lpx-B1.3* gene, in seven cultivars (Alejskaya, Altajskaya niva, Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha, and Saratovskaya zolotistaya) in the studied collection (Fig. 2, Table 4).

Using markers of the *Lpx-B1* locus to analyze foreign durum wheat cultivars and breeding lines, five haplotypes with different combinations of the *Lpx-B1.1* gene alleles and one of

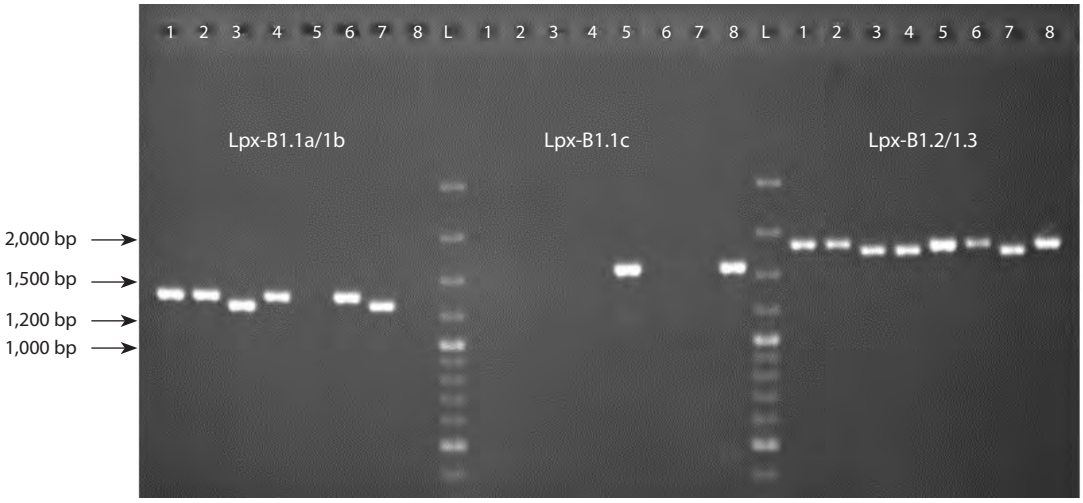


Fig. 2. Results of the *Lpx-B1* locus genes and alleles identification with markers Lpx-B1.1a/1b, Lpx-B1.1c and Lpx-B1.2/1.3 in durum wheat cultivars: 1 – Bezenchukskaya 209; 2 – Bezenchukskaya 210; 3 – Bezenchukskaya zolotistaya; 4 – Alejskaya; 5 – Donchanka; 6 – Zhemchuzhina Sibiri; 7 – Pamyati Chekhovicha; 8 – Giusto; L – marker GeneRuler 100 bp Plus.

Table 4. Genes and alleles of the *Lpx-B1* locus identified in the studied durum wheat cultivars

Allele/gene	Number of cultivars	Frequency, %	Cultivars
		***Lpx-B1.1* gene**	
		Markers Lpx-B1.1a/1b, Lpx-B1.1c	
Lpx-B1.1a (1,320 bp)	45	83.33	45 cultivars
Lpx-B1.1b (1,246 bp)	5	9.26	Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha, Saratovskaya zolotistaya
Lpx-B1.1c (1,558 bp)	4	7.41	Donchanka, Gelios, Leucurum 21, Giusto
		***Lpx-B1.2* and *Lpx-B1.3* genes**	
		Marker Lpx-B1.2/1.3	
Lpx-B1.2 (1,785 bp)	47	87.04	47 cultivars
Lpx-B1.3 (1,709 bp)	7	12.96	Alejskaya, Altajskaya niva, Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha, Saratovskaya zolotistaya
		***Lpx-B1* haplotypes**	
Haplotype I (*Lpx-B1.1b* + *Lpx-B1.3*)	5	9.26	Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha, Saratovskaya zolotistaya
Haplotype II (*Lpx-B1.1a* + *Lpx-B1.2*)	43	79.63	43 cultivars
Haplotype III (*Lpx-B1.1c* + *Lpx-B1.2*)	4	7.41	Donchanka, Gelios, Leucurum 21, Giusto
Haplotype V (*Lpx-B1.1a* + *Lpx-B1.3*)	2	3.70	Alejskaya, Altajskaya niva

the *Lpx-B1.2* or *Lpx-B1.3* genes were found (Verlotta et al., 2010; Parada et al., 2020).

Four of the five known *Lpx-B1* locus haplotypes were identified in the studied cultivars. Haplotype I (*Lpx-B1.1b* + *Lpx-B1.3*) was identified in Bezenchukskaya zolotistaya, Bezenchukskij vector, Bezenchukskij podarok, Pamyati Chekhovicha and Saratovskaya zolotistaya, haplotype III (*Lpx-B1.1c* + *Lpx-B1.2*) – in Donchanka, Gelios, Leucurum 21 and Giusto, haplotype V (*Lpx-B1.1a* + *Lpx-B1.3*) – in Alejskaya and Altajskaya niva and haplotype II (*Lpx-B1.1a* + *Lpx-B1.2*) – in the remaining 43 cultivars (Table 4).

Yellow pigment content in grain was determined for 38 studied spring durum wheat cultivars. It varied from 334.7 μg% to 687.7 μg% with an average value of 475.1 μg%.

In the studied set, 14 cultivars had high (more than 500 μg%), 15 cultivars had medium (400–500 μg%), and nine cultivars had low yellow pigment content (200–400 μg%) (Table 1).

The relative influence of genotype, environmental conditions (in this experiment, conditions of the year) and their interaction on the accumulation of yellow pigment in grain was determined using 38 spring durum wheat genotypes in a 3-year (2021–2023) experiment at the Samara Federal Research Scientific Center of the Russian Academy of Sciences. As a result, significant effects of all factors were established using the two-way analysis of variance. The contributions of genotype, environment and their interaction to the total variance were 65.3, 28.0 and 6.3 %, respectively (Supplementary Materials, Table S1)[1].

On average, for the groups of cultivars with medium and high values, the parameters of general and specific adaptability (GAC_i, SAC_i), responsiveness to the environment (by the regression coefficient – b_i) of the trait "yellow pigment content in grain" significantly exceeded similar parameters for the group with a low value of the trait. Judging by the regression coefficient, the most effective assessment of the phenotype can be given in favorable environmental conditions. At the same time, no significant differences were observed between the groups for the relative stability parameter (S_{gi}) (Table S2).

The rank correlation coefficients between the cultivars' arrangement in the variability rows by the content of yellow pigment by years and between the ranks of cultivars by average values for three years and for each year varied within 0.83–0.96, which is significant at the 1.0 % level. These results suggest that the studied set of spring durum wheat genotypes differs significantly in the accumulation of yellow pigment in grain, the differences between cultivars are stable under different environmental conditions, and this is the result of the functioning of the corresponding genetic systems.

Discussion

The analysis of 54 durum wheat cultivars using markers of the *Psy-A1* phytoene synthase gene and the *Lpx-B1* lipoxygenase locus allowed evaluating their variability in the studied collection.

The markers were used for the first time to analyze domestic durum wheat cultivars. Fragments of the expected size were obtained with all markers and allelic variants previously described when analyzing foreign material were identified. The results were clear, reproducible, and coincided for the two studied samples of each cultivar, which indicates the effectiveness of using these markers to analyze domestic durum wheat cultivars.

Analysis of the *Psy-A1* phytoene synthase gene polymorphism

To analyze the *Psy-A1* gene, encoding a key enzyme of carotenoid synthesis, two SCAR markers, PSY1-A1_STS and YP7A-2, were used, in order to identify differences between alleles having indels in the third and fourth introns associated with the level of yellow pigment content (He et al., 2009a; Singh et al., 2009).

[1] Tables S1 and S2 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Trifonova_Engl_29_3.pdf

The study of the collection using these markers showed an extremely low level of its diversity. The *Psy-A1l* allele (PSY1-A1_STS marker) prevailed, as well as the *Psy-A1d* allele (YP7A-2 marker). These alleles were noted in 51 cultivars studied; their frequency was 94.45 %. Only three cultivars had other *Psy-A1* alleles. In cultivars Krasnokutka 13 and Donchanka, the *Psy-A1o* allele was identified with the PSY1-A1_STS marker, and no amplification with the YP7A-2 marker was noted, and in the cultivar Kurant, the *Psy-A1a* allele was detected with the PSY1-A1_STS marker, and the *Psy-A1e*, with the YP7A-2 marker (Table 3).

The combined use of the PSY1-A1_STS and YP7A-2 markers showed the correspondence of the detected allelic variants, which was also noted in previous studies (Campos et al., 2016; Patil et al., 2018). Thus, samples having the *Psy-A1l* allele identified with the PSY1-A1_STS marker had the *Psy-A1d* allele detected with the YP7A-2 marker, and samples having the *Psy-A1a* allele identified with the PSY1-A1_STS marker had the *Psy-A1e* allele detected with the YP7A-2 marker. When *Psy-A1o* was detected with the PSY1-A1_STS marker, there was no amplification with the YP7A-2 marker. Thus, both markers allow to detect the 688 bp indel in the fourth intron of the *Psy-A1* gene, which can distinguish the *Psy-A1l* and *Psy-A1d* alleles from *Psy-A1a* and *Psy-A1e*. Using the PSY1-A1_STS marker, an additional *Psy-A1o* allele can be identified, which is not detected using the YP7A-2 marker due to a 198 bp deletion in the third intron, which results in the absence of the binding site for the forward primer of the YP7A-2 marker (Campos et al., 2016).

Previously, when studying durum wheat collections using *Psy-A1* gene markers, an association of the identified alleles with the level of yellow pigment content was shown. Alleles *Psy-A1d* and *Psy-A1e*, identified using the YP7A-2 marker, were associated with high and low yellow index, respectively (He et al., 2009b). Alleles *Psy-A1l* and *Psy-A1o*, identified using the PSY1-A1_STS marker, were associated with high or intermediate, and *Psy-A1a*, with low content of yellow pigment (Singh et al., 2009; Campos et al., 2016).

Thus, in the studied collection, *Psy-A1* alleles associated with high and intermediate yellow pigment content (*Psy-A1l*, *Psy-A1o* and *Psy-A1d*) predominate. Alleles associated with low yellow index (*Psy-A1a* and *Psy-A1e*) were identified only in one cultivar.

Similar results were shown in studies of the foreign durum wheat gene pool. So, in the collections of foreign cultivars released in different periods, as well as in breeding lines studied using the PSY1-A1_STS marker, the *Psy-A1l* allele prevailed with a 68 to 97 % frequency (Singh et al., 2009; Campos et al., 2016; Parada et al., 2020). In the study of 100 durum wheat breeding lines from the CIMMYT collection using the YP7A-2 marker, the prevalence of the *Psy-A1d* allele was revealed (99 % frequency). Allele *Psy-A1o* was quite common in Mediterranean landraces, but was rare in modern cultivars, despite its association with a high or intermediate yellow index (Campos et al., 2016). Alleles associated with low yellowness were also rare in the foreign gene pool (Singh et al., 2009; Campos et al., 2016; Parada et al., 2020).

The predominance of allelic variants associated with high yellow pigment content may be the result of a long selection

process that led to the rejection of samples with alleles that negatively affect the trait.

Analysis of the *Lpx-B1* lipoxygenase locus polymorphism

Using three SCAR markers Lpx-B1.1a/1b, Lpx-B1.1c and Lpx-B1.2/1.3, haplotypes of the *Lpx-B1* locus were determined for all cultivars studied. It was previously shown that of the five *Lpx-B1* haplotypes, only haplotype III is associated with low lipoxygenase activity (Verlotta et al., 2010; Parada et al., 2020).

Four of the five previously reported haplotypes were identified in the studied cultivars, with haplotype II, associated with an intermediate level of lipoxygenase activity, being the most common and occurring with 79.63 % frequency (Table 4). Among foreign durum wheat cultivars, this haplotype was also quite common; for example, in Mediterranean landraces, the frequency of this haplotype was 54 % (Parada et al., 2020), and in cultivars of different breeding periods cultivated in Italy, 42 % (Verlotta et al., 2010).

The most valuable for breeding is haplotype III (the *Lpx-B1.1c* allele and the *Lpx-B1.2* gene). Due to the MITE transposition, a large deletion occurred in the sequence of the *Lpx-B1.1c* allele, which led to the loss of gene function and a significant decrease in lipoxygenase activity (Carrera et al., 2007; Verlotta et al., 2010). Haplotype III was identified in only three studied winter cultivars: Donchanka, Gelios, and Leucurum 21, and was not found among spring cultivars. Previously, in a study of 85 predominantly Italian durum wheat genotypes released in different breeding periods (before 1971; 1971–1990; 1991–2005), this haplotype was noted in 41 cultivars, 32 of which were released after 1991 (Verlotta et al., 2010). Among Italian cultivars of an earlier breeding period, haplotype III was much less common (Verlotta et al., 2010), and the frequency of this haplotype was also low in Mediterranean landraces (Parada et al., 2020).

Haplotype I, associated with high lipoxygenase activity, was identified in four cultivars from the Samara Federal Research Scientific Center of the Russian Academy of Sciences: Bezenchukskaya zolotistaya, Bezenchukskij podarok, Bezenchukskij vector, Pamyati Chekhovicha and Saratovskaya zolotistaya from the Federal Center of Agriculture Research of the South-East Region. Among Mediterranean landraces, the frequency of this haplotype was 39 %, and in cultivars grown in Italy, this haplotype was found mainly in the material released before the 1970s, and was not found in modern cultivars (Verlotta et al., 2010; Parada et al., 2020).

Cultivars Alejskaya and Altajskaya niva from the Federal Altai Scientific Center for Agrobiotechnology had haplotype V. This haplotype is associated with high lipoxygenase activity and is quite rare in foreign cultivars (Parada et al., 2020).

In general, according to previous studies, in the foreign gene pool, the proportion of the *Lpx-B1* haplotype III, valuable for breeding, increases in modern cultivars and breeding lines, which indicates targeted selection of cultivars with a low level of lipoxygenase activity. At the same time, in the domestic gene pool, the frequency of haplotype III is still quite low. The use of *Lpx-B1* locus markers for the analysis of domestic breeding material will contribute to the effective selection of genotypes with haplotype III.

Association between yellow pigment content and identified alleles of the *Psy-A1* gene and haplotypes of the *Lpx-B1* locus

The studied cultivars varied significantly in the accumulation of yellow pigment in grain (Table 1). Most of the cultivars had medium and high content of yellow pigment. At the same time, according to molecular markers, most cultivars, including those with low yellow pigment content, were found to have *Psy-A1* alleles that determine high yellowness (*Psy-A1d* and *Psy-A1l*), as well as haplotype II of the *Lpx-B1* locus, associated with an intermediate level of lipoxygenase activity. Such a discrepancy may be due to the fact that the yellow index is a complex, polygenic trait that depends on the interaction of various enzymes, controlling both carotenoid synthesis and degradation (Colasuonno et al., 2019). Furthermore, the haplotype of the lipoxygenase locus has a greater influence on the trait at post-harvest stages and during pasta manufacturing (flour and pasta yellow index) (Parada et al., 2020).

Also, according to the data obtained, 65.3 % of the trait variance was determined by the genotype. The significant prevalence of the genotype effect over the influence of the environment and the genotype–environment interaction confirms data on the high heritability of the yellow pigment accumulation processes in durum wheat grain with the predominance of additive effects of genes (Blanco et al., 2011; Roncallo et al., 2012; Schulthess, Schwember, 2013).

Conclusion

Thus, using molecular markers, the allelic diversity of the *Psy-A1* phytoene synthase and *Lpx-B1* lipoxygenase genes in Russian durum wheat cultivars was studied for the first time, and turned out to be quite low. In the studied cultivars, allelic variants of the *Psy-A1* gene associated with high yellow pigment content predominate, as in most modern foreign durum wheat cultivars. Haplotype III of the *Lpx-B1* locus, valuable for breeding, associated with low lipoxygenase activity, was detected only in three winter cultivars (Donchanka, Gelios and Leucurum 21), while among foreign cultivars, especially modern ones, the proportion of this haplotype is significantly higher. The obtained results confirmed the dependence of the yellow pigment content on the genotype; however, the presence of the *Psy-A1* and *Lpx-B1* alleles associated with high carotenoid content did not always determine their high content in the grain of the studied cultivars, which is most likely due to the influence of other genes of yellow pigment metabolism. Nevertheless, the studied markers can be used for breeding new durum wheat cultivars with a high yellow index.

References

Blanco A., Colasuonno P., Gadaleta A., Mangini G., Schiavulli A., Simeone R., Digesu A.M., De Vita P., Mastrangelo A.M., Cattivelli L. Quantitative trait loci for yellow pigment concentration and individual carotenoid compounds in durum wheat. *J Cereal Sci.* 2011;54(2):255-264. doi 10.1016/j.jcs.2011.07.002

Campos K.M., Royo C., Schulthess A., Villegas D., Matus I., Ammar K., Schwember A.R. Association of phytoene synthase *Psy1-A1* and *Psy1-B1* allelic variants with semolina yellowness in durum wheat (*Triticum turgidum* L. var. *durum*). *Euphytica*. 2016;207:109-117. doi 10.1007/s10681-015-1541-x

Carrera A., Echenique V., Zhang W., Helguera M., Manthey F., Schrager A., Picca A., Cervigni G., Dubcovsky J. A deletion at the *Lpx-B1*

locus is associated with low lipoxygenase activity and improved pasta color in durum wheat (*Triticum turgidum* ssp. *durum*). *J Cereal Sci*. 2007;45(1):67-77. doi 10.1016/j.jcs.2006.07.001

Colasuonno P., Marcotuli I., Blanco A., Maccaferri M., Condorelli G.E., Tuberosa R., Parada R., de Camargo A.C., Schwember A.R., Gadaleta A. Carotenoid pigment content in durum wheat (*Triticum turgidum* L. var. *durum*): an overview of quantitative trait loci and candidate genes. *Front Plant Sci*. 2019;10:1347. doi 10.3389/fpls.2019.01347

Digesu A.M., Platani C., Cattivelli G., Blanco A. Genetic variability in yellow pigment components in cultivated and wild tetraploid wheats. *J Cereal Sci*. 2009;50(2):210-218. doi 10.1016/j.jcs.2009.05.002

Doyle J.J., Doyle J.L. Isolation of plant DNA from fresh tissue. *Focus*. 1990;12(1):13-15

Ficco D.B.M., Mastrangelo A.M., Trono D., Borrelli G.M., De Vita P., Fares C., Beleggia R., Platani C., Papa R. The colours of durum wheat: a review. *Crop Pasture Sci*. 2014;65(1):1-15. doi 10.1071/CP13293

Gallagher C.E., Matthews P.D., Li F., Wurtzel E.T. Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses. *Plant Physiol*. 2004;135(3):1776-1783. doi 10.1104/pp.104.039818

He X.Y., Zhang Y.L., He Z.H., Wu Y.P., Xiao Y.G., Ma C.X., Xia X.C. Characterization of phytoene synthase 1 gene (*Psy1*) located on common wheat chromosome 7A and development of a functional marker. *Theor Appl Genet*. 2008;116(2):213-221. doi 10.1007/s00122-007-0660-8

He X.Y., He Z.H., Ma W., Appels R., Xia X.C. Allelic variants of phytoene synthase 1 (*Psy1*) genes in Chinese and CIMMYT wheat cultivars and development of functional markers for flour colour. *Mol Breed*. 2009a;23:553-563. doi 10.1007/s11032-009-9255-1

He X., Wang J., Ammar K., Pena R.J., Xia X., He Z. Allelic variants at the *Psy-A1* and *Psy-B1* loci in durum wheat and their associations with grain yellowness. *Crop Sci*. 2009b;49(6):2058-2064. doi 10.2135/cropsci2008.11.0651

Hessler T.G., Thomson M.J., Benscher D., Nachit M.M., Sorrells M.E. Association of a lipoxygenase locus, *Lpx-B1*, with variation in lipoxygenase activity in durum wheat seeds. *Crop Sci*. 2002;42(5):1695-1700. doi 10.2135/cropsci2002.1695

Kilchevsky A.V., Khotyleva L.V. Environment-oriented Plant Breeding. Minsk: Tekhnalogiya Publ., 1997 (in Russian)

Malchikov P.N., Myasnikova M.G. The content of yellow pigments in durum wheat (*Titicum durum* Desf.) grains: biosynthesis, genetic control, marker selection. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2020;24(5):501-511. doi 10.18699/VJ20.642 (in Russian)

Methods for Assessing Technological Qualities of Grain. Moscow: Academic Council for Grain Quality, 1971 (in Russian)

Natoli V., Malchikov P., De Vita P., Shevchenko S., Dolaberidze S. Genetic improvement for gluten strength in Russian spring durum wheat genotypes. In: Antipova T. (Ed.) Comprehensible Science. ICCS 2020. Lecture Notes in Networks and Systems. Vol. 186. Springer, 2021;301-312. doi 10.1007/978-3-030-66093-2_29

Parada R., Royo C., Gadaleta A., Colasuonno P., Marcotuli I., Matus I., Castillo D., de Camargo A.C., Araya-Flores J., Villegas D., Schwember A.R. Phytoene synthase 1 (*Psy-1*) and lipoxygenase 1 (*Lpx-1*) genes influence on semolina yellowness in wheat Mediterranean germplasm. *Int J Mol Sci*. 2020;21(13):4669. doi 10.3390/ijms21134669

Patil R., Oak M., Deshpande A., Tamhankar S. Development of a robust marker for *Psy-1* homoeologs and its application in improvement of yellow pigment content in durum wheat. *Mol Breed*. 2018;38(11):136. doi 10.1007/s11032-018-0895-x

Requena-Ramirez M.D., Rodríguez-Suarez C., Flores F., Hornero-Méndez D., Atienza S.G. Marker-trait associations for total carotenoid content and individual carotenoids in durum wheat identified by genome-wide association analysis. *Plants*. 2022;11(15):2065. doi 10.3390/plants11152065

Roncallo P.F., Cervigni G.L., Jensen C., Miranda R., Carrera A.D., Helguera M., Echenique V. QTL analysis of main and epistatic effects for flour color traits in durum wheat. *Euphytica*. 2012;185:77-92. doi 10.1007/s10681-012-0628-x

Schulthess A., Schwember A.R. Improving durum wheat (*Triticum turgidum* L. var *durum*) grain yellow pigment content through plant breeding. *Cienc Inv Agr.* 2013;40(3):475-490. doi 10.4067/S0718-16202013000300002

Shevchenko S.N., Malchikov P.N., Myasnikova M.G., Natoli V., De Vita P., Giuliani M. Genetic methods of improving wheat quality of durum cultivars adapted to climate conditions of Russia with special accent on commercial characteristics of grain. *Izvestia of Samara Scientific Center of the Russian Academy of Sciences*. 2018;20(2/2):220-230. doi 10.24411/1990-5378-2018-00064 (in Russian)

Singh A., Reimer S., Pozniak C.J., Clarke F.R., Clarke J.M., Knox R.E., Singh A.K. Allelic variation at *Psy1-A1* and association with yellow pigment in durum wheat grain. *Theor Appl Genet*. 2009;118(8):1539-1548. doi 10.1007/s00122-009-1001-x

State Register of Varieties and Hybrids of Agricultural Plants Admitted for Usage (National list): official publication. Moscow: Rosinformagrotech Publ., 2024;19-21 (in Russian)

Vasil'chuk N.S. Spring Durum Wheat Breeding. Saratov: Novaya Gazeta Publ., 2001 (in Russian)

Verlotta A., De Simone V., Mastrangelo A.M., Cattivelli L., Papa R., Trono D. Insight into durum wheat *Lpx-B1*: a small gene family coding for the lipoxygenase responsible for carotenoid bleaching in mature grains. *BMC Plant Biol*. 2010;10:263. doi 10.1186/1471-2229-10-263

Conflict of interest. The authors declare no conflict of interest.

Received October 2, 2024. Revised January 16, 2025. Accepted February 12, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):380-391

doi 10.18699/vjgb-25-41

The IIIVmrMLM method uncovers new genetic variants associated with resistance to Fusarium wilt in flax

M.A. Duk [1, 2], A.A. Kanapin [1], A.A. Samsonova [1], M.P. Bankin [1], M.G. Samsonova [1]

[1] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia
[2] Ioffe Institute of the Russian Academy of Sciences, St. Petersburg, Russia
m.g.samsonova@gmail.com

Abstract. Flax (*Linum usitatissimum*) is an important agricultural crop grown for fiber and oil production, playing a key role in various industries such as production of paints, linoleum, food, clothes and composite materials. Fusarium wilt caused by the fungus *Fusarium oxysporum* f. sp. *lini* is a reason of significant economic damage in flax cultivation. The spores of the fungus can persist in the soil for a long time, so obtaining resistant varieties is important. Here we used data on the resistance of 297 flax accessions from the collection of the Federal Center for Bast Crops in Torzhok (Russian Federation) to infection by a highly virulent isolate of the fungus MI39 in 2019–2021. Genotype resistance to infection was assessed by calculating the DSI index, a normalized proportion of genotypes with the same disease symptoms. The IIIVmrMLM program in Single_env mode was used to search for regions of the flax genome associated with resistance. The IIIVmrMLM model was designed to address methodological shortcomings in identifying all types of interactions between alleles, genes and environment, and to unbiasedly estimate their genetic effects. Being a multilocus MLM model, it estimates the effects of all genes as well as the effects of all interactions simultaneously. A total of 111 QTNs were found, of which 34 fell within the body of a known gene or were located in flanking regions within 1,000 bp. The genes into which the detected variants fell were associated with resistance to abiotic and biotic stresses, root, shoot and flower growth and development. Ten of the QTNs found mapped to regions of previously identified QTLs controlling the synthesis of palmitic, oleic, and other fatty acids. QTN Chr1_1706865/Chr1_1706872 and QTN Chr8_22542741 mark regions identified previously in an association search by the GAPIT program. The allelic effect was confirmed for all the QTNs found: a Mann–Whitney test was performed, which confirmed significant differences between the DSI index value in carriers of the reference and alternative allele. An increase in the number of alleles with negative effects in the genotype leads to a statistically significant decrease in the DSI value for all three years of testing. The groups of varieties with a large number of alleles reducing the DSI index had the best resistance. A total of 5 varieties were selected from the collection for which the number of alleles reducing the DSI index value did not exceed the number of alleles with the opposite effect for all three years. These varieties can be used further in breeding programs.

Key words: flax; *Linum usitatissimum*; GWAS; Fusarium wilt; *Fusarium oxysporum* f. sp. *lini*

For citation: Duk M.A., Kanapin A.A., Samsonova A.A., Bankin M.P., Samsonova M.G. The IIIVmrMLM method uncovers new genetic variants associated with resistance to Fusarium wilt in flax. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):380-391. doi 10.18699/vjgb-25-41

Funding. This research was funded by Russian Science Foundation, grant number 23-16-00037.

Acknowledgements. The authors would like to thank Peter the Great St. Petersburg Polytechnic University Centre for Supercomputing (scc.spbstu.ru) for providing excellent computational resources and support for this project.

Метод IIIVmrMLM обнаруживает новые генетические варианты, связанные с устойчивостью к фузариозному увяданию у льна

М.А. Дук [1, 2], А.А. Канапин [1], А.А. Самсонова [1], М.П. Банкин [1], М.Г. Самсонова [1]

[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия
[2] Физико-технический институт им. А.Ф. Иоффе Российской академии наук, Санкт-Петербург, Россия
m.g.samsonova@gmail.com

Аннотация. Лен (*Linum usitatissimum*) – важная сельскохозяйственная культура, выращиваемая для получения волокна и масла. Лен используют для производства красок, линолеума, в пищевой промышленности, для производства одежды и композитных материалов. Значительный экономический ущерб при выращивании льна наносит фузариозное увядание, вызываемое грибом *Fusarium oxysporum* f. sp. *lini*. Споры гриба могут долгое время сохраняться в почве, поэтому получение устойчивых к заражению сортов имеет большое значение.

Здесь мы использовали данные об устойчивости 297 образцов льна из коллекции Федерального научного центра лубяных культур в Торжке (Россия) к заражению сильно вирулентным изолятом гриба MI39 в 2019–2021 гг. Устойчивость генотипа к заражению оценивали путем вычисления индекса DSI – нормализованной пропорции генотипов с одинаковыми симптомами болезни. Для поиска районов генома льна, ассоциированных с устойчивостью, использовали программу IIIVmrMLM в режиме Single_env. Модель IIIVmrMLM была разработана для устранения методологических недостатков в выявлении всех типов взаимодействий между аллелями, генами и средой и для несмещенной оценки их генетических эффектов. Поскольку это мультилокусная MLM-модель, она оценивает эффекты всех генов, а также эффекты всех взаимодействий одновременно. Всего было найдено 111 QTN, из которых 34 были локализованы в последовательности известного гена или расположены во фланкирующих районах на расстоянии, не превышающем 1 т. п. н. Гены, в которые попадали обнаруженные варианты, были связаны с устойчивостью к абиотическим и биотическим стрессам, с ростом и развитием корня, побега и цветка. Десять из найденных QTN картировались в областях ранее идентифицированных QTL, контролирующих синтез пальмитиновой, олеиновой и других жирных кислот. QTN Chr1_1706865/Chr1_1706872 и QTN Chr8_22542741 маркируют районы, идентифицированные нами ранее при поиске ассоциаций программой GAPIT. Для всех найденных QTN был подтвержден аллельный эффект: произведен тест Манна–Уитни, который подтвердил значимые различия между значением DSI у носителей референсного и альтернативного аллеля. Увеличение в генотипе числа аллелей с негативным эффектом приводит к статистически значимому уменьшению величины DSI для всех трех лет тестирования. Группы сортов с большим количеством аллелей, уменьшающих индекс DSI, имели наилучшую устойчивость. Всего из коллекции было выбрано пять сортов, для которых число аллелей, уменьшающих величину DSI, не превышало число аллелей с обратным эффектом по всем трем годам. Эти сорта могут быть использованы в дальнейшем в селекционных программах.
Ключевые слова: лен; *Linum usitatissimum*; GWAS; фузариозное увядание; *Fusarium oxysporum* f. sp. *lini*

Introduction

Flax (*Linum usitatissimum*) is an important crop grown for both fiber and oil. Flaxseed oil is used in the food industry as a source of unsaturated fatty acids and is also used as the main component of varnishes, paints and linoleum. Flax fiber is used in textiles, composites and insulation materials (Goudenhooft et al., 2019). Fusarium wilt caused by the fungus *Fusarium oxysporum* f.sp. *lini* limits flax production (Dean et al., 2012). This disease lowers fiber quality and can lead to yield loss in the absence of proactive measures.

Primary fungal infection occurs through the roots. The pathogen enters the xylem and blocks the flow of water and nutrients, causing wilting, stem damage and eventually plant death. The spores of the fungus can persist in infested soil for up to 50 years and are very difficult to eliminate (Houston, Knowles, 1949).

Control of Fusarium wilt is possible through various agricultural practices, such as the use of pesticides (Rashid, Kenaschuk, 1993), but the possible harmfulness of pesticides to human health leads to the preference of using varieties resistant to infection, which is an alternative option to control yield loss caused by *F. oxysporum* (Ondrej, 1993; Rozhmina, Loshakova, 2016).

Resistance to the disease has been acquired through breeding, but the mechanisms of resistance remain incompletely understood. Modern flax varieties have high to medium resistance to Fusarium wilt (Rozhmina, Loshakova, 2016; Rozhmina, 2017). However, co-evolution of pathogen and plant can lead to the emergence of strains with higher aggressiveness or to the loss of resistance in varieties, determined by a small number of genes. Therefore, breeding new varieties with different combinations of genes determining resistance is important for long-term effects. Transcriptomics experiments have shown that cell wall components, transcription factors, secondary metabolites and antioxidants play a prominent role in the response of flax to infection by *F. oxysporum* f.sp. *lini* (Galindo-González, Deyholos, 2016; Dmitriev et al., 2017; Boba et al., 2021).

The search for new genomic variants associated with disease resistance and the identification of new genes affecting resistance to fungal infection play a key role in breeding programs. The use of classical GWAS identified QTNs (Quantitative Trait Nucleotides) associated with resistance to Fusarium wilt (Kanapin et al., 2021) and located mainly on chromosome 1, as well as on chromosomes 8 and 13. A large number of QTNs are localized on chromosome 1 within 640 kb (Kanapin et al., 2021; Cloutier et al., 2024).

Plant resistance to disease may also be determined by multiple indirect factors related not only to resistance to fungal infection but also to other plant characteristics, e. g. fatty acids in plants are known to be involved in defense mechanisms against various stressors, including fungal infection (Kachroo et al., 2008; He, Ding, 2020).

Classical MLM-type models help eliminate effects introduced by population structure and sample relatedness, but suffer from the Bonferroni correction for multiple testing, which is too stringent to detect associations with complex traits (Zhang Y.M. et al., 2019). To address this problem, multi-locus MLM models have been proposed that can detect QTNs with marginal effects for which the significance threshold set by the Bonferroni correction is too stringent.

One such method using multi-locus models is the mrMLM method (Zhang Y.-M. et al., 2020) implemented in the IIIVmrMLM package (Li M. et al., 2022). In this paper, we applied this method to search for genomic associations with resistance assessed in infected plants in 2019, 2020 and 2021, which allowed us to identify novel genetic variants not previously detected by classical methods. These variants fell within resistance-related genes as well as within quantitative trait loci (QTL, Quantitative Trait Locus) published previously and related to fatty acid production (You, Cloutier, 2020).

Materials and methods

297 flax samples from the collection of the Federal Scientific Centre for Bast Crops were grown in Torzhok, Russia. 180 accessions were fiber flax varieties, and 117 belonged to oilseed flax. Of the oilseed samples, 98 belonged to the intermediate type, 4, to large-seeded varieties, and 15, to the crown type.

Resistance of accessions to *F. oxysporum* f. sp. *lini* was evaluated under infection-provocation nursery conditions with controlled irrigation but not controlled temperature. Evaluations were conducted in 2019, 2020 and 2021 (Rozhmina et al., 2022). Each variety was replicated 16 times by sowing all seeds in cross rows of containers. The dimensions of the containers were 550 × 85 × 20 cm. Two genotypes, AP5 and I-7, were used as susceptible and resistant genotypes to control Fusarium wilt. The infection background was established by introducing 400 g of pure culture of *F. oxysporum* f. sp. *lini* strain MI39. Seeds were planted on the 12th day after inoculation with pure culture of the fungus.

Pure culture was prepared by preliminary cultivation of strain MI39 on agar-agar medium with beer wort and subsequent incubation on oat grain substrate (grain/water ratio 1 to 1.75) for 3–4 weeks, until complete infection of oats by the fungus, after which the pathogen was introduced into the soil. The indicator of reliability of the infection background was the reference varieties (resistant and susceptible genotypes), which were sown at the edges and in the middle of each container (16 seeds each). Disease severity was assessed using the Disease Severity score (DSS). The DSS scores ranged from 0 to 3, where 0 was a healthy plant, 1 was a partially blighted plant or stem blight on one side, 2 was a completely blighted plant with seed pods, and 3 was a completely blighted plant that died before pod formation. Based on the DSS, disease severity index (DSI) was calculated using the formula adopted in phytopathology (Guidelines for the Phytopathological Assessment, 2000): $DSI = (\Sigma ab/3A) \times 100$ %, where a is the number of plants with the same DSS, b is the DSS score; A is the total number of plants, and 3 is the highest DSS score.

DNA was isolated from leaves using the DNeasy Plant Mini Kit (Qiagen). Whole-genome sequencing of DNA was performed in BGI using the Illumina protocol, which generates paired-end reads of 150 base pairs in length. Comparison with the NCBI ASM22429v2 reference genome assembly (Wang Z. et al., 2012) was performed using bwa-mem (Li H., Durbin, 2009). Variant prediction was performed using NGSEP (Tello et al., 2019) version 4.0; from the 3,416,829 SNPs obtained, 72,526 SNPs were retained after filtering by MAF = 0.05 and conditioning on the presence of the variant in at least 85 % of genotypes. An annotation of the flax genome with the indicated Arabidopsis orthologous genes was provided by the Cloutier group (You, Cloutier, 2020).

Using the IIIVmrMLM package (Li M. et al., 2022) in Single_env mode, GWAS analyses were performed on genetic data filtered by MAF = 0.05. TASSEL (Bradbury et al., 2007) and PLINK (Purcell et al., 2007) with standard settings were used for the necessary data transformation.

The additive effect calculated by the IIIVmrMLM package was used to identify genotypes with high performance. An allele with a negative effect led to a decrease in the DSI in its carriers, while an allele with a positive effect increased the DSI. Varieties were selected in which the number of negative-effect alleles exceeded the number of positive-effect alleles.

Linkage disequilibrium decay (LD) was estimated using the square of the Pearson correlation coefficient (r^2). PopLDdecay version 3.4.1 (Zhang C. et al., 2019) was run to calculate r^2 in a 300 kb window. LD decay was calculated based on r^2 and distance for each SNP pair using the R script.

Results

Environmental characteristics may influence disease development. Plants were grown under the infection-provocation nursery conditions with regular irrigation but not controlled temperature. According to the weather station at the growing site, the temperature in the first decade of May was above average in 2019 and 2020 and below average in 2021 (Table 1). In the second decade of May, the temperature was above average in 2019 and 2021 and below average in 2020.

The analysis of variance showed that the Fusarium wilt infection depends on the year of cultivation and genotype (Table 2). When considering the influence of temperatures, it was found that only the temperature in the 1st decade of May has a significant influence on the variation ($F > 1$, $Pr(>F) < 0.05$); moreover, its influence on the Fusarium wilt infection is almost identical to the influence of the year, as can be seen from the values of the root mean square of the residuals in the analysis of variance in Table 2, whereas other environmental characteristics made only a small contribution.

On average, the difference between the maximum and minimum DSI values for genotype in different years is 25.9. Nevertheless, the differences in the DSI for the whole population under consideration from one year to another do not show sufficient significance: when comparing the 2019 and 2020 data, the p-value was 0.996, the 2019 and 2021 data, $p = 0.113$, the 2020 and 2021 data, $p = 0.12$.

In other words, despite the large influence of growing conditions, the main interest of the study continues to be the effect of variety (genotype) on disease resistance.

GWAS identified 111 QTNs (Supplementary Materials, Table S1)[1] associated with the DSI in different years, of which 35 were associated with 2019 data, 37, with 2020 data, and 40, with 2021 data. QTNs associated with data from different years are located on all chromosomes, of which 44 fell within known QTLs (You, Cloutier, 2020; Cloutier et al., 2024) or appeared to be localized in the gene sequence or less than 1,000 bp away from genes (Fig. 1*a–c*). The distribution of all found QTNs in the genome is shown in Figure 1*d*. The allelic effect was confirmed for all found QTNs: a Mann–Whitney test was performed, which confirmed significant differences between the DSI value in carriers of the reference and alternative allele (Table S1).

The largest number of QTNs found for each year's infection data were located on chromosomes 1, 2, 8 and 15. In a previous study that used the GAPIT package to find associations with resistance to Fusarium wilt, QTNs were also located on chromosomes 1 and 8 (Kanapin et al., 2021). In total, all the

[1] Tables S1, S2 and Figure S1 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Duk_Engl.xlsx

Table 1. Average temperature of the growing seasons 2019–2021 (according to Torzhok meteorological station)

Month	Decade	Average temperature, °C			Average, long-term
		2019	2020	2021	
May	1	10.1	11.3	7.9	9.9
	2	14.5	8.0	16.7	11.5
	3	16.3	11.1	12.6	12.9
June	1	18.8	15.8	15.8	15.3
	2	17.9	19.6	19.2	15.6
	3	16.0	18.9	22.5	16.5
July	1	14.2	18.2	20.5	16.9
	2	13.8	15.8	23.2	17.4
	3	17.1	16.8	17.9	17.5
August	1	13.3	17.7	18.4	17.2
	2	16.1	–	–	15.2
	3	15.7	–	–	13.9

QTNs explain more than 50 % of the variation, at most one QTN explains about 5 % of the variation for one year, as can be seen in Table 3.

Only one QTN was found in the data of two years and three pairs of QTNs found for different years appeared to be located quite close to each other, as shown in Table 4. Table 4 also presents the mean non-normalized DSI values for carriers of the reference and alternative allele for a given QTN. It can be seen that carriers of the alternative allele for all the indicated QTNs showed much lower DSI values than the reference allele carriers. However, of the mentioned QTNs, only one QTN common to the 2020 and 2021 data fell within the sequence of a gene, the function of which, however, is not known, while the other three pairs were located more than 1 kb away from the nearest genes. It can also be noted that QTNs Chr1_1706865/Chr1_1706872 fell into the previously identified region on chromosome 1, with coordinates 1213418–1854337, associated with resistance to Fusarium wilt (Kanapin et al., 2021).

Of the 111 QTNs associated with Fusarium wilt resistance in different years, 34 were localized within the gene body or were located at a distance of less than 1 kb from the gene (Table 5).

Within the protein-coding genes and their 1-kb flanking regions, we found 34 QTNs (Table 5), of which 12 had an alternative allele with an effect of decreasing the value of the DSI and 22 with an effect of increasing this value.

10 QTNs fell within the QTLs published previously in (You, Cloutier, 2020), of which two were near a known gene (marked as ** in Table 5). In addition, one QTN fell within a region associated with resistance to Fusarium wilt on chromosome 1 (Table 6), published in (Kanapin et al., 2021; Cloutier et al., 2024).

Eleven QTNs whose positions overlap with previously identified QTLs from (You, Cloutier, 2020; Cloutier et al., 2024) are shown in Table 6. Most of these QTL are associated with the production of fatty acids: palmitic acid, oleic acid, linolenic acid, etc., and only two QTNs, Chr1_17552378 and Chr1_2540379, fell into QTLs associated with plant immunity. In nine out of eleven cases, the presence of the alternative QTN allele in the plant resulted in an increase in the DSI value, and only in two cases the alternative allele resulted in a decrease in the DSI value compared to the reference allele carriers.

Table 2. Dispersion analysis

Source of variance	Mean Sq	F	Pr(>F)
DSI ~ genotype + year			
Genotype	2931.5	10.016	<2e–16
Year	1927.7	6.586	0.00148
Residuals	292.7		
DSI ~ genotype + temperature in the 1st decade of May			
Genotype	2931.0	10.01	<2e–16
Temperature in the 1st decade of May	3396.0	11.590	0.000707
Residuals	293.0		

To assess variety performance, the number of alleles with a negative effect (reduction of the DSI value) and with a positive effect was counted among the QTNs found from each year's data (Table S2). The number of negative and positive alleles affecting the DSI for each year is different, but an increase in the total number of alleles with a negative effect in the varieties leads to a statistically significant decrease in the DSI value for all three years, as can be seen in Figure 2.

Table 7 shows the varieties for which the number of alleles that increased the DSI value did not exceed the number of alleles with the opposite effect in all three years.

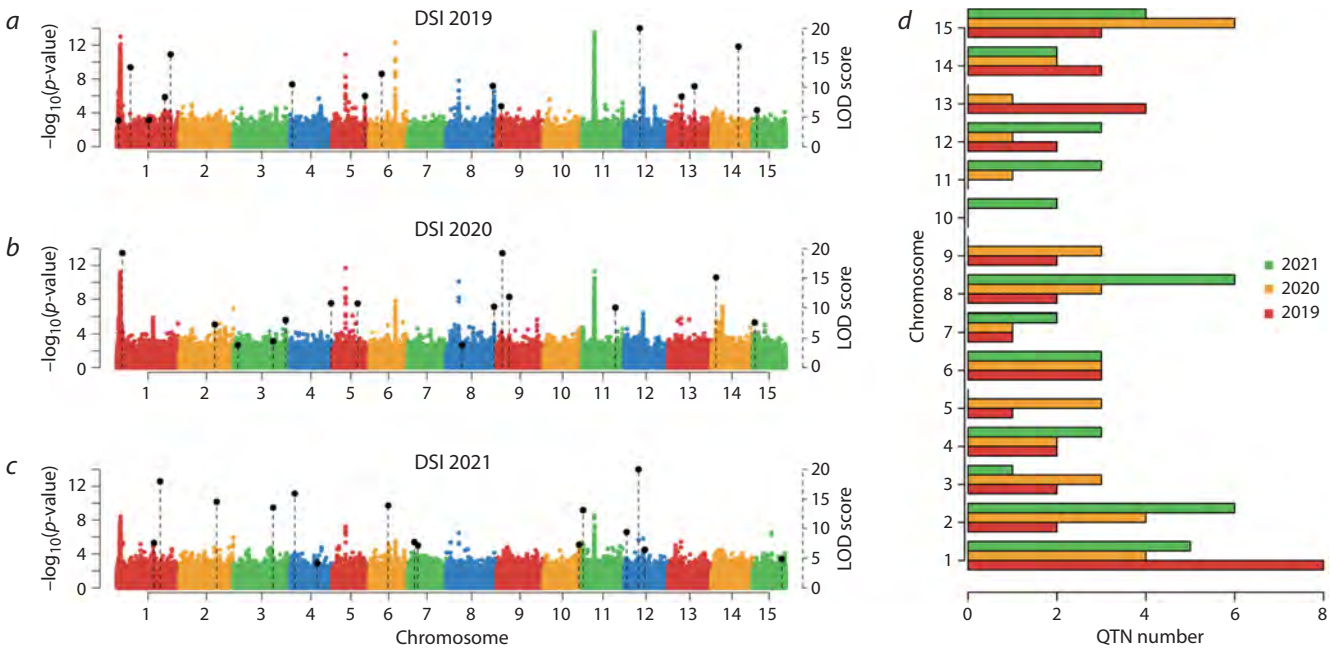


Fig. 1. Location of QTNs associated with fusarium wilt relative to chromosomes in flax.

a –c – Manhattan plots of GWAS results using the IIIVmrMLM package; black shows QTNs that fell in the QTL or were located near genes along with their LOD score value, which is used in IIIVmrMLM to assess significance; *d* – distribution of QTNs found for the DSI for the three years data, by chromosome.

Table 3. Cumulative percentage of variation in each year's data explained by QTNs

Data	Total r^2, %	QTNs with the largest r^2, %	Largest r^2, %
DSI 2019	55.99	Chr3_18720497	5.31
DSI 2020	58.50	Chr2_15253612	4.43
DSI 2021	69.84	Chr12_10144355	5.35

Note. QTN names are formed as ChrX_N, where X is the chromosome number and N is the position in the chromosome.

Table 4. Co-localized QTNs across the years

QTN	Chr3_18671763	Chr1_1706865/ Chr1_1706872	Chr15_7067724/ Chr15_7067662	Chr2_25600109/ Chr2_25600116
Year	2020/2021	2019/2021	2020/2021	2020/2021
Distance between QTNs	0 bp	7 bp	62 bp	7 bp
r^2, %	1.14/3.47	4.24/3.17	1.05/0.75	4.25/4.66
Average DSI for REF	41.55/36.92	43.12/38.47	41.24/36.31	41.85/38.7
Average DSI for ALT	25.18/21.02	10.55/9.52	24.75/20.56	14.53/15.06
p-value of the Mann–Whitney test	0.0014/0.0021	5.42e–19/1.16e–13	0.0026/0.0027	1.56e–06/2.92e–10
Nearest gene	*Lus10033807*	*Lus10025819*	*Lus10001477*	*Lus10003500*
QTN location relative to the gene	Within the gene	1,026/1,033 bp downstream	3,605/3,543 bp downstream	23,760/23,753 bp upstream
Gene annotation	Protein with an unknown function (DUF1664)	2-Oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein	Remorin family protein	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Ortholog in *Arabidopsis*	*AT1G04960.1*	*AT3G21360.1*	*AT5G23750.2*	*AT3G21330.1*

Note. The corresponding lines show data for different years separated by "/"; bp – base pairs; REF – reference allele; ALT – alternative allele.

Table 5. QTNs located within protein-coding genes and their 1-kb flanking regions

QTN	r^2, %	Average DSI for REF / Average DSI for ALT	QTN position relative to a gene	Annotation	Ortholog in *Arabidopsis*
2019					
Chr1_740951	0.51	32.67 / 46.32*	*Lus10036050*, gene body	Calcium-dependent protein kinase 34	*AT5G19360.1*
Chr1_6391647	2.21	35.67 / 43.75*	*Lus10034284*, gene body	Sodium/calcium exchanger family protein	*AT5G17850.1*
Chr1_15073726	0.90	41.12* / 31.36	*Lus10015586*, gene body	Prolyl oligopeptidase family protein	*AT1G50380.1*
Chr1_22688905	0.64	41.63* / 8.42	*Lus10014640*, 573 bp downstream	Major facilitator superfamily protein	*AT2G39210.1*
Chr1_25377570	3.15	41.97* / 30.21	*Lus10027990*, gene body	Oxidoreductase, 2OG-Fe(II) oxygenase family protein	*AT4G02940.1* (Duan et al., 2017)
Chr4_1087234	1.82	35.08 / 46.28*	*Lus10030349*, 155 bp upstream	DZC (Disease resistance/ zinc finger/chromosome condensation-like region) domain-containing protein	*AT1G31880.1* (Depuydt et al., 2013; Rodriguez-Villalon et al., 2014)
Chr5_15553508 **(QPAL-Lu5.2, PAL)	1.24	33.95 / 45.51*	*Lus10024055*, gene body	(Ortholog *Arabidopsis*: nitric oxide synthase interacting protein)	*AT5G65030.1*
Chr6_5732293	1.86	36.49 / 48.47*	*Lus10036674*, 132 bp downstream	Homeobox 1	*AT3G01470.1* (Aoyama et al., 1995)
Chr9_2114668	0.68	35.09 / 52.45*	*Lus10017493*, 464 bp upstream	–	*AT4G34630.1*
Chr13_6339069	1.71	35.05 / 53.25*	*Lus10002083*, 88 bp downstream	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	*AT5G08790.1* (Delessert et al., 2005; Wang X. et al., 2009; Wang X., Culver, 2012)
Chr13_12427556	1.21	34.20 / 49.92*	*Lus10010801*, 894 bp upstream	Cytochrome P450, family 721, subfamily A, polypeptide 1	*AT1G75130.1*
Chr14_12732300	1.08	41.14* / 21.40	*Lus10008367*, 263 bp upstream	ARM repeat superfamily protein	*AT3G08960.1* (Jia et al., 2023)
Chr15_2097827	0.88	33.09 / 49.75*	*Lus10007320*, gene body	RING/FYVE/PHD-type zinc finger family protein	*AT1G29800.1* (Kim et al., 2023)
2020					
Chr1_2540379 **(Lu1_2500703, DSI)	2.02	36.98 / 58.99*	*Lus10025924*, 539 bp upstream	Sec14p-like phosphatidylinositol transfer family protein	*AT3G24840.1*
Chr2_16849610	1.01	35.50 / 50.85*	*Lus10016310*, gene body	Cytochrome P450, family 721, subfamily A, polypeptide 1	*AT1G75130.1*
Chr3_1992356	1.34	35.70 / 58.40*	*Lus10037255*, gene body	Solute:sodium symporters, urea transmembrane transporters	*AT5G45380.1* (Liu et al., 2003; Kojima et al., 2007)
Chr3_24632490	1.27	43.84* / 28.25	*Lus10037741*, gene body	Lipoamide dehydrogenase 1	*AT3G16950.2* (Lutziger, Oliver, 2000)

Table 5 (end)

QTN	r^2, %	Average DSI for REF / Average DSI for ALT	QTN position relative to a gene	Annotation	Ortholog in *Arabidopsis*
2020 г.					
Chr4_19469341	0.79	40.45* / 21.40	*Lus10039825*, 945 bp upstream	–	*AT4G28290.1*
Chr8_7404794	0.81	40.23* / 15.74	*Lus10021849*, gene body	Cysteine-rich RLK (receptor-like protein kinase) 8	*AT4G23160.1*
Chr9_2644458	3.16	36.05 / 54.10*	*Lus10010491*, gene body	Immunoglobulin E-set superfamily protein	*AT3G07880.1* (Carol et al., 2005)
Chr11_15810790	0.29	37.57 / 64.13*	*Lus10023622*, gene body	ADC synthase superfamily protein	*AT1G74710.1* (Wildermuth et al., 2001; Strawn et al., 2007)
Chr14_2163238	1.74	42.23* / 23.94	*Lus10025537*, gene body	PAZ domain-containing protein/piwi domain-containing protein	*AT5G21030.1*
Chr15_1044247	0.95	34.39 / 43.97*	*Lus10011210*, 33 bp upstream	F-box and associated interaction domains-containing protein	*AT1G32420.1*
2020, 2021					
Chr3_18671763	1.14	41.55* / 25.18	*Lus10033807*, gene body	Protein of unknown function (DUF1664)	*AT1G04960.1*
2021					
Chr1_20417569	0.57	35.92* / 14.00	*Lus10015886*, gene body	Nucleotidyl-transferase family protein	*AT4G00060.1*
Chr2_17726495	2.95	30.63 / 45.43*	*Lus10033187*, gene body	K-box region and MADS-box transcription factor family protein	*AT3G54340.1* (Krizek, Meyerowitz, 1996)
Chr4_2301676	3.27	29.94 / 48.51*	*Lus10029444*, gene body	Peptide chain release factor 1	*AT3G62910.1* (Motohashi et al., 2007)
Chr4_12925693	2.22	32.24 / 55.99*	*Lus10015799*, gene body	Leucine-rich repeat protein kinase family protein	*AT1G67510.1*
Chr6_8828608	2.89	31.06 / 44.86*	*Lus10036278*, 568 bp downstream	RNA-directed DNA polymerase (reverse transcriptase)	*AT5G04050.1*
Chr7_3147157 **(QPAL-Lu7.3, PAL)	0.87	31.48 / 48.22*	*Lus10023551*, 387 bp upstream	–	*AT5G66440.1*
Chr10_16815632	1.08	29.87 / 43.13*	*Lus10022764*, 483 bp upstream	ABI five binding protein 3	*AT3G29575.1* (Garcia et al., 2008)
Chr11_575034	0.49	32.80 / 59.28*	*Lus10027253*, gene body	Ortholog *Arabidopsis*: GPI- anchor protein	*AT3G18050.1*
Chr12_9853001	1.96	36.04* / 20.76	*Lus10024259*, gene body	Aldehyde dehydrogenase 2C4	*AT3G24503.1* (Nair et al., 2004)
Chr15_13834579	0.87	35.60* / 16.58	*Lus10037970*, gene body	Plant U-box 14, flowering regulation	*AT3G54850.1* (Andersen et al., 2004)

Note. REF – reference allele; ALT – alternative allele. * The largest of the mean DSI values in carriers of the reference or alternative allele; ** QTNs localized both in the gene body and known QTL.

Table 6. QTNs located within previously identified QTLs

QTN	r^2, %	Average DSI for REF Average DSI for ALT	QTL	Trait	QTL position
Chr5_15553508	1.24	33.95 45.51*	QPAL-Lu5.2	PAL	13796740–15667804
Chr8_21862725	2.47	36.27 51.36*	QOLE-Lu8.1	OLE	21781910–23526575
Chr12_7449738	1.06	36.35 60.51*	QOIL-Lu12.6	OIL	4591134–7490902
Chr1_2540379	2.02	36.98 58.99*	Lu1_ 2500703	DSI	2500703–2636369
Chr5_12086840	1.77	37.19 53.09*	QPAL-Lu5.1	PAL	12061283–12181348
Chr8_22542741	2.77	41.64* 23.63	QOLE-Lu8.1	OLE	21781910–23526575
Chr1_17552378	0.90	28.86 41.23*	QPM-crc-LG1	PM	16920407–18739647
Chr7_3147157	0.87	31.48 48.22*	QPAL-Lu7.3	PAL	624439–5423600
Chr7_4787639	1.57	35.91* 18.53	QPAL-Lu7.3	PAL	624439–5423600
Chr12_1240570	0.81	32.95 55.79*	QIOD-Lu12.3/QLIN-Lu12.3/QLIO-Lu12.3	IOD/LIN/LIO	489561–2981562
Chr12_6862107	3.27	32.31 55.52*	QOIL-Lu12.6	OIL	4591134–7490902

Note. REF – carriers of the reference allele, ALT – carriers of the alternative allele. * The largest of the mean DSI values in carriers of the reference or alternative allele. Abbreviations of trait names from (You, Cloutier, 2020): PAL (Palmitic %) – palmitic acid content; OLE (Oleic %) – oleic acid content; OIL (Oil content %) – oil content, PM (Powdery mildew rating) – powdery mildew rate; IOD (Iodine value) – iodine content; LIN (Linoleic %) – linoleic acid content; LIO (Linolenic %) – linolenic acid content.

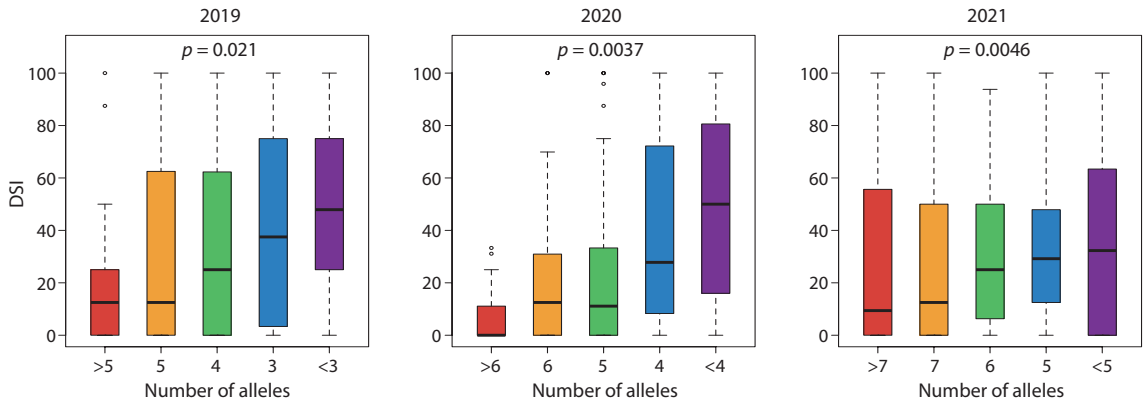


Fig. 2. Distribution of DSI values in different years for accessions containing different numbers of alleles that have a negative effect on the DSI value.

The upper part of the graphs shows the *p*-value of the statistical test.

Table 7. Varieties that had the best combination of alleles with positive and negative effects in all three years

DSI			Morphotype	Country	Breeding	Name
2019	2020	2021				
0	8.3	0	Fiber	Japan	Line	Honkei35, k-5396
0	0	0	Intermediate	Czechia	Line	AGT987, k-7225
0	8.3	0		France	Cultivar	Eolle, k-7034
19	8.3	0		Russia	Line	VNIIL, LM92, k-6672
0	0	0			Cultivar	Voronezhskij 1308/138, k-3052

Discussion

In this paper, we used the IIIVmrMLM program to find genomic regions controlling resistance to Fusarium wilt in flax. A total of 111 QTNs associated with the disease severity index (DSI) value in data from different years were found.

QTNs Chr1_1706865, Chr1_1706872, and Chr8_22542741 apparently point to regions on chromosomes 1 and 8 previously found using the GAPIT package, as published in (Kanapin et al., 2021), being less than the average LD for the corresponding chromosomes, which was 16 and 45 kb for chromosomes 1 and 8, respectively (Fig. S1).

Many of the QTNs fall into or near genes with important functions, and it is possible that these genes are casual (Table 5). Among the QTNs found to fall into genes, there are QTNs that have a favorable effect on a trait. For example, the alternative allele of the QTN Chr1_25377570 in the *Lus10027990* gene decreases the DSI which is a favorable effect for this value. The orthologue of this gene in *Arabidopsis AT4G02940.1* encodes a dioxygenase that demethylates m^6A in mRNA. Mutations in this gene affect the mRNA stability of the flowering time regulators FT, SPL3, and SPL9 and delay the transition from vegetative growth to flowering (Duan et al., 2017). The alternative allele of QTN Chr3_1992356 (Table 5) located in the *Lus10037255* gene also increases the DSI value. The *Lus10037255* ortholog encodes the urea proton symporter DUR3, which is involved in urea transport across the plasma membrane into root cells (Liu et al., 2003; Kojima et al., 2007). Since *F. oxysporum* infects plants via roots, the transport of metabolites in roots may influence the susceptibility of the plant to infection.

Some QTNs fall into genes associated with plant immunity (Table 5). For example, QTN Chr15_2097827 with a positive effect (ALT allele increases the DSI) is localized in the *Lus10007320* gene, the orthologue of which in *Arabidopsis* regulates autophagy (Kim et al., 2023). In contrast the alternative alleles of QTNs in the genes *Lus10021849, Lus10008367*, and *Lus10024259* decrease the DSI value. *Lus10021849* is an orthologue of *Arabidopsis CRK8*, which encodes a receptor-like protein kinase. The *Arabidopsis* orthologue *Lus10008367* encodes the effector Ran KA120. This effector prevents autoimmune activation in the absence of pathogens and restricts the activity of the *SNC* gene, which encodes a TIR-NB-LRR-like receptor involved in the salicylic acid-mediated immune response (Jia et al., 2023). The *Lus10024259* orthologue in *Arabidopsis* is involved in the biosynthesis of ferulic and synapic acids (Nair et al., 2004), which are important for plant resistance to biotic and abiotic stresses.

In many cases, the presence of the alternative allele resulted in an increased DSI value in its carriers. Many of the genes that harbored such QTNs were associated with root or leaf growth. For example, *Lus10030349* (Table 5) (orthologue *AT1G31880.1*) encodes the BREVIS RADIX protein, which regulates cell elongation and differentiation in the root and shoot (Depuydt et al., 2013; Rodriguez-Villalon et al., 2014). *Lus10036674* (orthologue *AT3G01470.1*) encodes the HAT5 protein with homeobox and leucine zipper domains that is involved in the mechanism of leaf growth regulation (Aoyama et al., 1995). *Lus10010491* (orthologue *AT3G07880.1*) encodes RhoGDI, an inhibitor of GDP dissociation from Rho GTPase. This inhibitor spatially restricts the sites of growth to a single point on the trichoblast and regulates activation of the RHD2/AtrbohC NADPH oxidase, which is required for root hair growth (Carol et al., 2005).

Mutations in genes related to plant immunity and stress response can also have a negative effect on plant resistance to Fusarium wilt (Table 5). For example, *Lus10002083* (orthologue of *AT5G08790.1*) encodes the ATAF2 protein, which is involved in the regulation of basal defense responses of the host plant against viral infection (Delessert et al., 2005; Wang X. et al., 2009; Wang X., Culver, 2012). *Lus10023622* (ortholog *AT1G74710.1*) encodes chloroplast isochorismate synthase 1, which is involved in the synthesis of salicylic acid, essential for plant defense against pathogens (Wildermuth et al., 2001; Strawn et al., 2007). The *AT1G67510.1* orthologue, *Lus10015799*, encodes an RLK protein kinase rich in leucine repeats. Many RLK kinases are involved in cell response processes to pathogens and abiotic stresses (Lease et al., 1998; Gish, Clark, 2011; Yan et al., 2023). The orthologue of *AT3G29575.1*, *Lus10022764*, acts as a negative regulator of abscisic acid (ABA) and stress response (Garcia et al., 2008).

Also, some QTNs are located in genes related to energy metabolism and flower growth. For example, *AT3G16950.2*, the orthologue of the *Lus10037741* gene (Table 5), encodes a dehydrogenase that is a component of the plastid pyruvate dehydrogenase complex (PDC) (Lutziger, Oliver, 2000). This complex is involved in glycolysis. *Lus10037741* contains the Chr3_24632490 QTN, in which the alternative allele reduces the DSI value (Table 5). Conversely, the alternative QTN alleles Chr4_2301676 and Chr2_17726495 in the genes *Lus10029444* and *Lus10033187* increase the DSI value (Table 5). *AT3G62910.1*, the orthologue of the *Lus10029444* gene, encodes the chloroplast peptide chain release factor APG3, which is required for normal chloroplast development (Motohashi et al., 2007). *AT3G54340.1,* the orthologue of

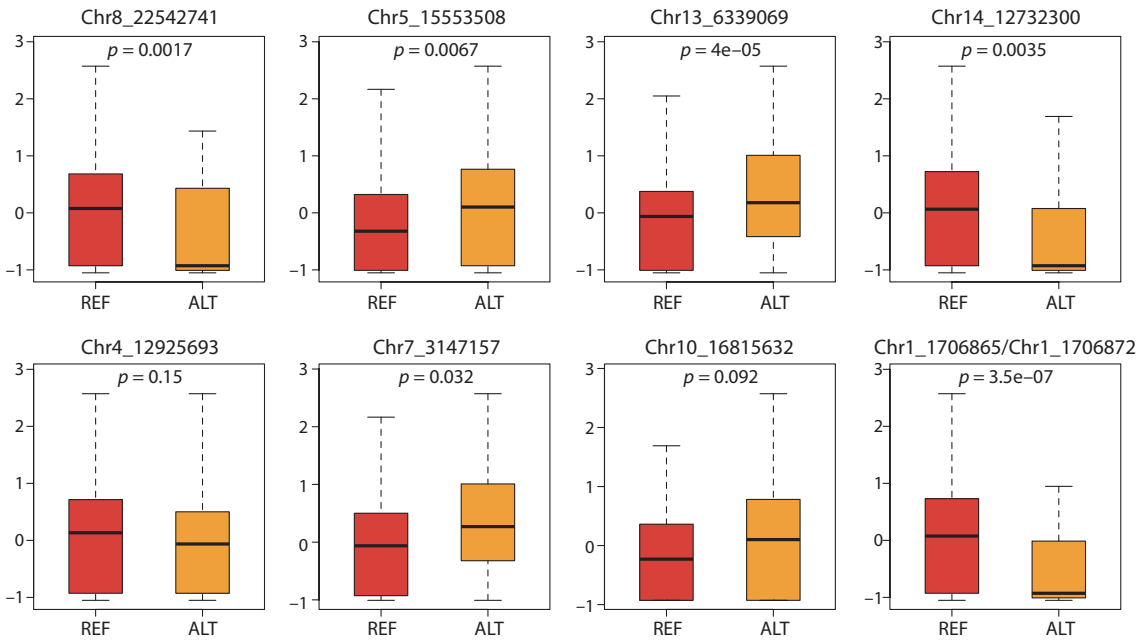


Fig. 3. Normalized DSI values in carriers of the reference (REF) and alternative (ALT) alleles in an independent dataset of 100 samples for some QTNs common to both datasets.

The *p*-values of the Mann–Whitney test are shown.

Lus10033187, encodes the homeobox protein APETALA 3, which regulates flower development (Krizek, Meyerowitz, 1996). On the other hand, QTN Chr15_13834579 in the *Lus10037970* gene, orthologous to the flowering regulator *AT3G54850.1*, has a positive effect on resistance to *F. oxysporum*, reducing the DSI value in carriers of the alternative allele.

It is also interesting to note that 11 of the QTNs found overlapped with previously published functional QTL regions, but only two of these regions were associated with plant immunity, while the rest were related to fatty acid production (Table 6). Fatty acids in plants act as a defense against pathogens and abiotic stresses (Kachroo et al., 2008; He, Ding, 2020); in addition, palmitic acid has been shown to reduce *Fusarium* infection in other plants (Ma et al., 2021). Thus, QTNs located in regions associated with fatty acid production may influence plant resistance to Fusarium wilt. Four QTNs fell into regions associated with palmitic acid (Table 6), which may indicate an important role of this acid in defense against Fusarium wilt in flax. The Chr8_22542741 QTN overlapped with the QOLE-Lu8.1 QTL associated with oleic acid production, and the Chr8_2256060236 and Chr8_2256060290 QTNs previously found with the GAPIT package (Kanapin et al., 2021) also fell within this region, indicating the possible importance of oleic acid production in protecting the plant against Fusarium wilt. One QTN, Chr1_2540379, also overlapped with a recently published region associated with flax resistance to Fusarium wilt (Cloutier et al., 2024).

We also tested on an independent dataset of 100 accessions the validity of the detected associations between QTNs and the DSI value (Fig. 3). This dataset grown under the same conditions was previously sequenced separately from the dataset under consideration and does not overlap with the dataset used in this study. It can be noted that the Chr5_15553508 and Chr7_3147157 QTNs, which fell into the palmitic acid-related regions, and the Chr8_22542741 QTN, which fell into the oleic acid-related region, demonstrate a significant difference in the DSI value between carriers of the reference and alternative alleles in this dataset (Fig. 3). Also, a significant allelic effect is seen in QTNs located in genes involved in plant immunity and stress response (Tables 5 and 6): Chr13_6339069 (*Lus10002083*), Chr14_12732300 (*Lus10008367*), Chr4_12925693 (*Lus10015799*), and Chr10_16815632 (*Lus10022764*). This suggests that these genes may also be involved in the defense of flax plants against infection.

We identified five varieties with the largest number of alleles decreasing the DSI (Table 7). The DSI of these varieties is much lower than the average DSI value, which for 2019, 2020 and 2021 was 38.7, 38.9 and 34.4, respectively. These varieties can be integrated into modern breeding programs.

Conclusion

As a result of application of the new multilocus model IIIVmrMLM to search for genomic associations with flax resistance to Fusarium wilt wilt, new genomic variants located in important regulatory regions were identified. Varieties with these variants showed greater resistance to the disease and can be used in breeding programs.

References

Andersen P., Kragelund B.B., Olsen A.N., Larsen F.H., Chua N.H., Poulsen F.M., Skriver K. Structure and biochemical function of a prototypical Arabidopsis U-box domain. *J Biol Chem.* 2004; 279(38):40053-40061. doi 10.1074/jbc.M405057200

Aoyama T., Dong C.H., Wu Y., Carabelli M., Sessa G., Ruberti I., Morelli G., Chua N.H. Ectopic expression of the Arabidopsis transcriptional activator Athb-1 alters leaf cell fate in tobacco. *Plant Cell.* 1995;7(11):1773-1785. doi 10.1105/tpc.7.11.1773

Boba A., Kostyn K., Kozak B., Zalewski I., Szopa J., Kulma A. Transcriptomic profiling of susceptible and resistant flax seedlings after *Fusarium oxysporum* lini infection. *PLoS One*. 2021;16:e0246052. doi 10.1371/journal.pone.0246052

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007;23(19):2633-2635. doi 10.1093/bioinformatics/btm308

Carol R.J., Takeda S., Linstead P., Durrant M.C., Kakesova H., Derbyshire P., Drea S., Zarsky V., Dolan L. A RhoGDP dissociation inhibitor spatially regulates growth in root hair cells. *Nature*. 2005; 438(7070):1013-1016. doi 10.1038/nature04198

Cloutier S., Edwards T., Zheng C., Booker H.M., Islam T., Nabetani K., Kutcher H.R., Molina O., You F.M. Fine-mapping of a major locus for Fusarium wilt resistance in flax (*Linum usitatissimum* L.). *Theor Appl Genet*. 2024;137(1):27. doi 10.1007/s00122-023-04528-2

Dean R., Van Kan J.A., Pretorius Z.A., Hammond-Kosack K.E., Di Pietro A., Spanu P.D., Rudd J.J., Dickman M., Kahmann R., Ellis J., Foster G.D. The Top 10 fungal pathogens in molecular plant pathology. *Mol Plant Pathol*. 2012;13(4):414-430. doi 10.1111/j.1364-3703.2011.00783.x

Delessert C., Kazan K., Wilson I.W., Van Der Straeten D., Manners J., Dennis E.S., Dolferus R. The transcription factor ATAF2 represses the expression of pathogenesis-related genes in Arabidopsis. *Plant J*. 2005;43(5):745-757. doi 10.1111/j.1365-313X.2005.02488.x

Depuydt S., Rodriguez-Villalon A., Santuari L., Wyser-Rmili C., Ragni L., Hardtke C.S. Suppression of *Arabidopsis* protophloem differentiation and root meristem growth by CLE45 requires the receptor-like kinase BAM3. *Proc Natl Acad Sci USA*. 2013;110(17): 7074-7079. doi 10.1073/pnas.1222314110

Dmitriev A.A., Krasnov G.S., Rozhmina T.A., Novakovskiy R.O., Snezhkina A.V., Fedorova M.S., Yurkevich O.Y., Muravenko O.V., Bolsheva N.L., Kudryavtseva A.V., Melnikova N.V. Differential gene expression in response to *Fusarium oxysporum* infection in resistant and susceptible genotypes of flax (*Linum usitatissimum* L.) *BMC Plant Biol*. 2017;17(Suppl.2):253. doi 10.1186/s12870-017-1192-2

Duan H.C., Wei L.H., Zhang C., Wang Y., Chen L., Lu Z., Chen P.R., He C., Jia G. ALKBH10B is an RNA N^6-methyladenosine demethylase affecting Arabidopsis floral transition. *Plant Cell*. 2017;29(12): 2995-3011. doi 10.1105/tpc.16.00912

Galindo-González L., Deyholos M.K. RNA-seq transcriptome response of flax (*Linum usitatissimum* L.) to the pathogenic fungus *Fusarium oxysporum* f. sp. *lini*. *Front Plant Sci*. 2016;7:1766. doi 10.3389/fpls.2016.01766

Garcia M.E., Lynch T., Peeters J., Snowden C., Finkelstein R. A small plant-specific protein family of ABI five binding proteins (AFPs) regulates stress response in germinating *Arabidopsis* seeds and seedlings. *Plant Mol Biol*. 2008;67(6):643-658. doi 10.1007/s11103-008-9344-2

Gish L.A., Clark S.E. The RLK/Pelle family of kinases. *Plant J*. 2011; 66(1):117-127. doi 10.1111/j.1365-313X.2011.04518.x

Goudenhooft C., Bourmaud A., Baley C. Flax (*Linum usitatissimum* L.) fibers for composite reinforcement: exploring the link between plant growth, cell walls development, and fiber properties. *Front Plant Sci*. 2019;10:411. doi 10.3389/fpls.2019.00411

Guidelines for the Phytopathological Assessment of the Resistance of Fiber Flax to Diseases. Moscow, 2000 (in Russian)

He M., Ding N.Z. Plant unsaturated fatty acids: multiple roles in stress response. *Front Plant Sci*. 2020;11:562785. doi 10.3389/fpls.2020.562785

Houston B.R., Knowles P.F. Fifty-years survival of flax fusarium wilt in the absence of flax culture. *Plant Dis Rep*. 1949;33:38-39

Jia M., Chen X., Shi X., Fang Y., Gu Y. Nuclear transport receptor KA120 regulates molecular condensation of MAC3 to coordinate plant immune activation. *Cell Host Microbe*. 2023;31(10):1685-1699.e7. doi 10.1016/j.chom.2023.08.015

Kachroo A., Fu D.Q., Havens W., Navarre D., Kachroo P., Ghabrial S.A. An oleic acid-mediated pathway induces constitutive defense signaling and enhanced resistance to multiple pathogens in soybean. *Mol Plant Microbe Interact*. 2008;21(5):564-575. doi 10.1094/MPMI-21-5-0564

Kanapin A., Bankin M., Rozhmina T., Samsonova A., Samsonova M. Genomic regions associated with Fusarium wilt resistance in flax. *Int J Mol Sci*. 2021;22(22):12383. doi 10.3390/ijms222212383

Kim J.H., Jung H., Song K., Lee H.N., Chung T. The phosphatidylinositol 3-phosphate effector FYVE3 regulates FYVE2-dependent autophagy in *Arabidopsis thaliana*. *Front Plant Sci*. 2023;14:1160162. doi 10.3389/fpls.2023.1160162

Kojima S., Bohner A., Gassert B., Yuan L., von Wirén N. AtDUR3 represents the major transporter for high-affinity urea transport across the plasma membrane of nitrogen-deficient Arabidopsis roots. *Plant J*. 2007;52(1):30-40. doi 10.1111/j.1365-313X.2007.03223.x

Krizek B.A., Meyerowitz E.M. The *Arabidopsis* homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the B class organ identity function. *Development*. 1996;122(1):11-22. doi 10.1242/dev.122.1.11

Lease K., Ingham E., Walker J.C. Challenges in understanding RLK function. *Curr Opin Plant Biol*. 1998;1(5):388-392. doi 10.1016/s1369-5266(98)80261-6

Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi 10.1093/bioinformatics/btp324

Li M., Zhang Y.W., Xiang Y., Liu M.H., Zhang Y.M. IIIVmrMLM: the R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol Plant*. 2022; 15(8):1251-1253. doi 10.1016/j.molp.2022.06.002

Liu L.H., Ludewig U., Frommer W.B., von Wirén N. AtDUR3 encodes a new type of high-affinity urea/H^+ symporter in Arabidopsis. *Plant Cell*. 2003;15(3):790-800. doi 10.1105/tpc.007120

Lutziger I., Oliver D.J. Molecular evidence of a unique lipoamide dehydrogenase in plastids: analysis of plastidic lipoamide dehydrogenase from *Arabidopsis thaliana*. *FEBS Lett*. 2000;484(1):12-16. doi 10.1016/s0014-5793(00)02116-5

Ma K., Kou J., Khashi U., Rahman M., Du W., Liang X., Wu F., Li W., Pan K. Palmitic acid mediated change of rhizosphere and alleviation of Fusarium wilt disease in watermelon. *Saudi J Biol Sci*. 2021; 28(6):3616-3623. doi 10.1016/j.sjbs.2021.03.040

Motohashi R., Yamazaki T., Myouga F., Ito T., Ito K., Satou M., Kobayashi M., Nagata N., Yoshida S., Nagashima A., Tanaka K., Takahashi S., Shinozaki K. Chloroplast ribosome release factor 1 (AtcpRF1) is essential for chloroplast development. *Plant Mol Biol*. 2007;64(5):481-497. doi 10.1007/s11103-007-9166-7

Nair R.B., Bastress K.L., Ruegger M.O., Denault J.W., Chapple C. The *Arabidopsis thaliana REDUCED EPIDERMAL FLUORESCENCE1* gene encodes an aldehyde dehydrogenase involved in ferulic acid and sinapic acid biosynthesis. *Plant Cell*. 2004;16(2):544-554. doi 10.1105/tpc.017509

Ondrej M. Evaluation of flax genepool according to resistance to Fusarium wilt of flax and to mildew. *Plant Genet. Resour*. 1993;92:54-58

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi 10.1086/519795

Rashid K.Y., Kenaschuk E.O. Effect of trifluralin on fusarium wilt in flax. *Can J Plant Sci*. 1993;3:893-901. doi 10.4141/cjps93-117

Rodriguez-Villalon A., Gujas B., Kang Y.H., Breda A.S., Cattaneo P., Depuydt S., Hardtke C.S. Molecular genetic framework for protophloem formation. *Proc Natl Acad Sci USA*. 2014;111(31):11551-11556. doi 10.1073/pnas.1407337111

Rozhmina T.A. Identification of effective genes of resistance to Fusarial wilt at variety of fibre-flax. *Biology in Agriculture*. 2017;4: 10-12 (in Russian)

Rozhmina T.A., Loshakova N.I. New sources of effective resistance genes to Fusarium wilt in flax (*Linum usitatissimum* L.) depending on temperature. *Sel'skokhozyaistvennaya Biologiya = Agric Biol.* 2016;51(3):310-317. doi 10.15389/agrobiology.2016.3.310eng

Rozhmina T., Samsonova A., Kanapin A., Samsonova M. An account of Fusarium wilt resistance in flax *Linum usitatissimum*: the disease severity data. *Data Brief*. 2022;41:107869. doi 10.1016/j.dib.2022.107869

Strawn M.A., Marr S.K., Inoue K., Inada N., Zubieta C., Wildermuth M.C. *Arabidopsis* isochorismate synthase functional in pathogen-induced salicylate biosynthesis exhibits properties consistent with a role in diverse stress responses. *J Biol Chem*. 2007;282(8):5919-5933. doi 10.1074/jbc.M605193200

Tello D., Gil J., Loaiza C.D., Riascos J.J., Cardozo N., Duitama J. NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics*. 2019;35(22):4716-4723. doi 10.1093/bioinformatics/btz275

Wang X., Culver J.N. DNA binding specificity of ATAF2, a NAC domain transcription factor targeted for degradation by Tobacco mosaic virus. *BMC Plant Biol*. 2012;12:157. doi 10.1186/1471-2229-12-157

Wang X., Goregaoker S.P., Culver J.N. Interaction of the Tobacco mosaic virus replicase protein with a NAC domain transcription factor is associated with the suppression of systemic host defenses. *J Virol*. 2009;83(19):9720-9730. doi 10.1128/JVI.00941-09

Wang Z., Hobson N., Galindo L., Zhu S., Shi D., McDill J., Yang L., Hawkins S., Neutelings G., Datla R., Lambert G., Galbraith D.W., Grassa C.J., Geraldes A., Cronk Q.C., Cullis C., Dash P.K., Kumar P.A., Cloutier S., Sharpe A.G., Wong G.K., Wang J., Deyholos M.K. The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *Plant J*. 2012;72(3):461-473. doi 10.1111/j.1365-313X.2012.05093.x

Wildermuth M.C., Dewdney J., Wu G., Ausubel F.M. Isochorismate synthase is required to synthesize salicylic acid for plant defence. *Nature*. 2001;414(6863):562-565. doi 10.1038/35107108

Yan J., Su P., Meng X., Liu P. Phylogeny of the plant receptor-like kinase (RLK) gene family and expression analysis of wheat RLK genes in response to biotic and abiotic stresses. *BMC Genomics*. 2023;24(1):224. doi 10.1186/s12864-023-09303-7

You F., Cloutier S. Mapping quantitative trait loci onto chromosome-scale pseudomolecules in flax. *Methods Protoc*. 2020;3(2):28. doi 10.3390/mps3020028

Zhang C., Dong S.-S., Xu J.-Y., He W.-M., Yang T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*. 2019;35(10):1786-1788. doi 10.1093/bioinformatics/bty875

Zhang Y.M., Jia Z., Dunwell J.M. Editorial: the applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front Plant Sci*. 2019;10:100. doi 10.3389/fpls.2019.00100

Zhang Y.-W., Tamba C.L., Wen Y.-J., Li P., Ren W.-L., Ni Y.-L., Gao J., Zhang Y.-M. mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies. *Genomics Proteomics Bioinformatics*. 2020;18(4):481-487. doi 10.1016/j.gpb.2020.06.006

Conflict of interest. The authors declare no conflict of interest.

Received September 8, 2024. Revised October 23, 2024. Accepted October 23, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):392-400

doi 10.18699/vjgb-25-42

Anthocyanins and phenolic compounds in colored wheat grain

E.V. Chumanova [1], T.T. Efremova [1], K.V. Sobolev [1, 2], E.A. Kosyaeva [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State Agrarian University, Novosibirsk, Russia
chumanova@bionet.nsc.ru

Abstract. Wheat is an extremely important and preferred source of human nutrition in many regions of the world. The production of biofortified colored-grain wheat varieties, which are known to contain a range of biologically active compounds, including anthocyanins, phenolic compounds, vitamins and minerals, reflects a worldwide trend toward increasing dietary diversity and improving diet quality through the development and introduction of diverse functional foods. The present work describes the genetic systems that regulate the biosynthesis and accumulation of anthocyanins in the pericarp and aleurone layer, the presence of which imparts purple, blue and black grain color. The review is devoted to the systematization of available information on the peculiarities of qualitative and quantitative content of anthocyanins, soluble and insoluble phenolic acids in wheat grain of different color, as well as on indicators of antioxidant activity of alcoholic extracts of grain depending on the content of anthocyanins and phenolic compounds. A huge number of studies have confirmed that these compounds are antioxidants, have anti-inflammatory activity and their consumption makes an important contribution to the prevention of a number of socially significant human diseases. Consumption of colored cereal grain products may contribute to an additional enrichment of bioactive compounds in human diet along with the usual sources of antioxidants. Special attention in the review is paid to the description of achievements of Russia's breeders in developing promising varieties and lines with colored grain, which will be a key factor in expanding the opportunities of the domestic and international grain market.

Key words: wheat; blue, purple, black grain; anthocyanins; phenolic compounds; antioxidant activity

For citation: Chumanova E.V., Efremova T.T., Sobolev K.V., Kosyaeva E.A. Anthocyanins and phenolic compounds in colored wheat grain. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):392-400. doi 10.18699/vjgb-25-42

Funding. This work was supported by grant No. 24-26-20028 from the Russian Science Foundation and the Ministry of Science and Innovation Policy of the Novosibirsk Region (No. p-99).

Антоцианы и фенольные соединения в окрашенном зерне пшеницы

Е.В. Чуманова [1], Т.Т. Ефремова [1], К.В. Соболев [1, 2], Е.А. Косяева [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский государственный аграрный университет, Новосибирск, Россия
chumanova@bionet.nsc.ru

Аннотация. Пшеница является чрезвычайно важным и предпочтительным источником питания человека во многих регионах земного шара. Получение биофортифицированных сортов мягкой пшеницы с окрашенным зерном, которое, как известно, содержит целый ряд биологически активных соединений, в том числе антоцианы, фенольные соединения, витамины и минералы, отражает общемировую тенденцию на увеличение разнообразия и повышение качества рациона путем разработки и внедрения разнообразных продуктов функционального питания. В настоящей работе описаны генетические системы, регулирующие биосинтез и накопление антоцианов в перикарпе и алейроновом слое, присутствие которых придает зерну фиолетовую, голубую и черную окраску. Обзор посвящен систематизации информации об особенностях качественного и количественного содержания антоцианов, растворимых и нерастворимых фенольных кислот в зерне пшеницы с различной окраской, а также показателях антиоксидантной активности спиртовых экстрактов зерна в зависимости от содержания антоцианов и фенольных соединений. Огромным количеством исследований подтверждено, что данные соединения являются антиоксидантами и соединениями с противовоспалительной активностью и их употребление вносит важный вклад в профилактику ряда социально значимых заболеваний человека. Употребление продуктов из окрашенного зерна злаков может способствовать дополнительному обогащению рациона людей биологически активными соединениями наряду с привычными источниками антиоксидантов. Отдельное внимание в обзоре уделено описанию достижений отечественных

селекционеров, усилия которых в этой области позволили получить ряд перспективных сортов и линий с окрашенным зерном, которые могут послужить основой создания рынка биофортифицированных диетических продуктов питания в России и увеличения экспортного потенциала рынка зерна.
Ключевые слова: пшеница; голубая, фиолетовая, черная окраска зерна; антоцианы; фенольные соединения; антиоксидантная активность

Introduction

Wheat occupies an important place in the structure of world consumption. Over the last two decades, the biofortification associated with increasing the nutritional value of food products from wheat grain has become an actual trend in breeding, in particular, much attention of researchers and breeders is focused on obtaining colored-grain wheat varieties, rich in anthocyanins. Depending on the type and accumulation of anthocyanins in different layers of the grain, wheat grain can have purple (in the pericarp, controlled by *Pp* genes), blue (in the aleurone layer, *Ba* genes) and dark purple (black) color (in both layers at the same time, *Ba* + *Pp* genes).

The value of colored wheat is a more diverse composition of flavonoids with important biological properties (Wang et al., 2020; Razgonova et al., 2021). In addition, many researchers have shown that colored-grain wheat has a higher content of protein and essential amino acids (Tian et al., 2018; Garg et al., 2022), a number of macro- and microelements: Zn, Fe, Mg, K, Ca, Se, Cu an Mn (Ficco et al., 2014; Sharma S. et al., 2018; Tian et al., 2018; Dhua et al., 2021; Shamanin et al., 2024), vitamins B1, B2, B9 and E (Granda et al., 2018) compared to red and white grains. Anthocyanins and phenolic compounds have great antioxidant potential, protecting cells from free radical damage. As well as these compounds have anti-inflammatory and antibacterial activity, preventing the development of diabetes, cardiovascular, neurodegenerative diseases and cancer (Laddomada et al., 2017; Francavilla, Joye, 2020; Mohammadi et al., 2024).

Colored wheat in China, India, Singapore, Canada and Austria is used to produce functional food products from whole wheat flour containing high amounts of antioxidants: different types of whole wheat breads, bakery products, cookies, pasta, pancakes, crackers (Garg et al., 2022; Gamel et al., 2023). However, the content of anthocyanins and phenolic compounds decreases when exposed to high temperatures. According to the literature data, the loss of anthocyanins during bread baking varies between 10–73 % and during the preparation of noodles, pasta, tortillas, biscuits, the content of anthocyanins and phenolic compounds decreases by 29–74 and 26–80 %, respectively (Garg et al., 2022). It has been shown that bakery products from colored grains are not inferior or are even superior to products from uncolored flour in terms of baking and organoleptic properties, and their shelf life increases (Khlestkina et al., 2017).

In recent years in our country, the direction towards the production of anthocyanin-biofortified wheat has been actively developed (Khlestkina et al., 2017; Vasilova et al., 2021; Rubets et al., 2022; Gordeeva et al., 2022; Shamanin et al., 2022, 2024), which forms the idea of a healthy lifestyle and nutrition, since wheat is an important food crop for Russia.

This review compiles available information on genetic factors regulating the accumulation of anthocyanins in colored wheat grain, peculiarities of anthocyanins and phenolic acids content and antioxidant activity (AOA) in colored grain, and summarizes information on the achievements of Russian scientists in obtaining promising colored-grain wheat lines and varieties.

Genetic control of the synthesis and accumulation of anthocyanins in wheat grain

The wheat grain consists of an embryo and endosperm densely surrounded by epidermis and a seed coat (Fig. 1). The fruit sheath (pericarp or pericarpium: etymologically derived from two Greek words, i. e., peri: around and carpos: fruit), consisting of several layers: epidermis, hypodermis, remnants of thin-walled cells, intermediate, transverse and tubular cells, surrounds the grain and plays a protective role. Seed coat cells (pigment layer, testa) of red-grain wheat contain proanthocyanidins that increase grain resistance to preharvest germination (Himi et al., 2011). The aleurone layer of the grain is the outer layer of the endosperm, consisting of a single layer of cells, square or slightly oblong in shape. It derives its name from the content of aleurone grains, which are protein storage structures.

Colored wheat is known to exist in three different forms: blue, purple and dark purple (black), depending upon the types and position of the anthocyanins in kernel layers. The bluish-gray color of wheat is because of the synthesis of anthocyanins in the aleurone layer. The presence of purple color, in turn, is due to the accumulation of anthocyanins in pericarp cells. The black grain results from the accumulation of anthocyanins simultaneously in the pericarp and aleurone layer (Fig. 1).

Anthocyanin biosynthesis in the aleurone layer is under the control of dominant alleles of *Ba* genes (Blue aleurone) localized in chromosomes of the fourth homeologous group of some cereal species. *Ba1* (syn. *Ba*(*b*)) is localized in the long arm of chromosome 4E (formerly 4Ag) of *Thinopyrum ponticum* (Podp.) Barkworth & D.R. Dewey (*Agropyron elongatum* L.; *Lophopyrum ponticum* (Podb.) Love; *Elytrigia pontica* (Podp.) Holub) (Zheng et al., 2006). *Ba2* (syn. *Ba*(*a*)) is localized in the long arm of chromosome $4A^{bo}$ *Triticum boeoticum* Boiss. or $4A^{m}$ *T. monococcum* L. (Singh et al., 2007). *BaThb* (syn. *Ba*(*c*)) is localized in the chromosome 4J *Th. bessarabicum* (Săvul. & Rayss) Á. Lőve (Shen et al., 2013).

ThMyc4E, encoding a MYC-type transcription factor with a bHLH domain is a *Ba1* candidate gene (Li N. et al., 2017). *TbMyC4A*, encoding a bHLH transcription factor containing three regulatory domains (bHLH-MYC_N, HLH and ACT-like) (Liu X. et al., 2021) is considered as a likely *Ba*(*a*) candidate gene. It is suggested that the *BaThb* and *Ba1* genes may have a common origin (Burešová et al., 2015), as *Th. bessarabicum* is a probable donor species of the E^{b} genome of most polyploid wheatgrass species, including *Th. ponticum*.

Ba genes were introgressed into the common wheat genome by producing substitution, addition and translocation lines. V.S. Arbuzova et al. (2012) and E.I. Gordeeva et al. (2022)

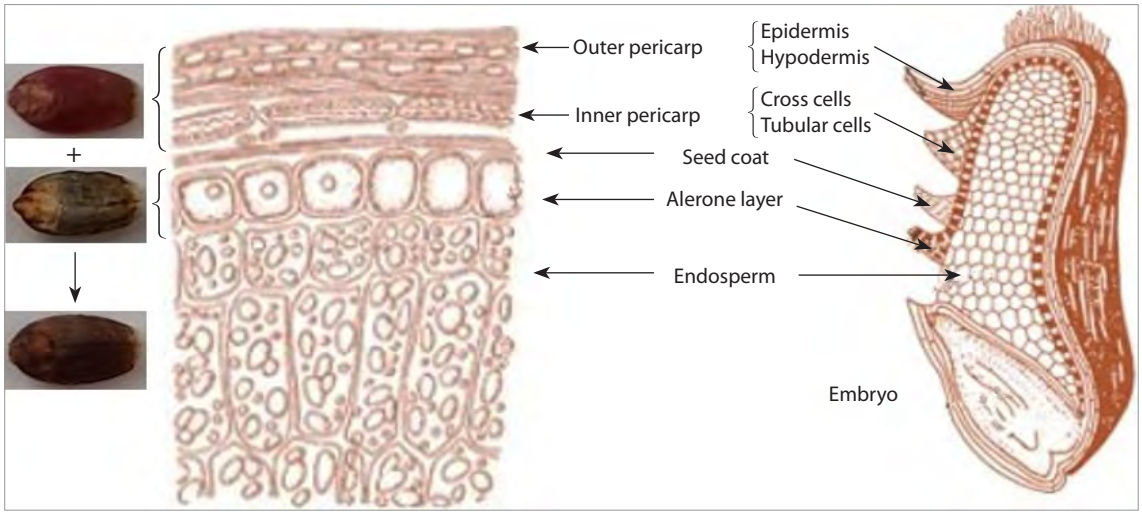


Fig. 1. Internal structure of wheat grain: longitudinal and transverse section.
Adapted from (Laddomada et al., 2015).

obtained substitution lines of spring bread wheat Saratovskaya 29 (S29) with the replacement of chromosome 4B or 4D with chromosome 4E *Th. ponticum*. Liu Xin et al. (2020) obtained six substitution lines $4A^{bo}$(4B) with the *Ba2* gene: Z18-1150, Z18-1195, Z18-1223, Z18-1244, Z18-1289, and Z18-3816. The chromosomal composition of several wheat lines with blue grain was described using *in situ* hybridization (Burešová et al., 2015). The results showed that six different types of *Th. ponticum* chromatin introgressions were detected: ditelosomic additions (Blue Norco), ditelosomic substitution (Blue Baart), T4BS.4AgL (UC66049), and different translocations of the distal parts of chromosomal arms of *Th. ponticum* (Sebesta Blue 1-3). Y. Shen et al. (2013) obtained 157 lines derived from the cross between *T. aestivum* cv. Chinese Spring and a *T. aestivum-Th. bessarabicum* amphiploid: they isolated monosomic and disomic addition lines with chromosomes 4J, 4JL, and 4JS, as well as T4DS.4DL-4JL carrying a fragment of chromosome 4J.

A significant problem in blue wheat lines selection is caused by the negative influence of wheatgrass genes linked to blue aleurone genes (Garg et al., 2016). Therefore, it is preferable to involve in breeding purple-grain donors, which are devoid of such a disadvantage.

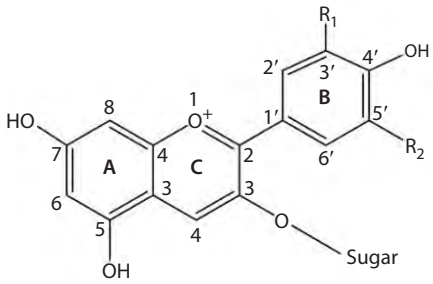
The purple color of wheat is because of anthocyanin synthesis in the pericarp layer. The first samples of tetraploid wheat *T. aethiopicum* Jakubz. (*T. turgidum* L. subsp. *abyssinicum* Vavilov) with purple-grain genes were collected by Wittmack in Abyssinia (Northern Ethiopia) in the early 1870s and brought to Europe, from where they further spread to different countries (Eticha et al., 2011). It is noteworthy that landraces of purple wheat are still cultivated in Ethiopia.

Anthocyanin synthesis in pericarp is controlled by the complementary interaction of *Pp* genes (Purple pericarp): *Pp-1* (*TaPpm1*) and *Pp3* (*TaPpb1*/*TaMyc1*), which encode different types of transcription factors that activate transcription of structural anthocyanin biosynthesis genes (Jiang W. et al., 2018). *Pp-1* is a MYB-like transcription factor with an R2R3 regulatory domain. A set of *Pp-1* homoeologous genes in chromosomes of the seventh homeologous group is currently known: *Pp-A1* in 7AS (*T. aestivum*) (Gordeeva et al., 2015), *Pp-B1* in 7BS (7B in *T. durum*, 7S in *Aegilops speltoides* Tausch.) (Khlestkina et al., 2010), and *Pp-D1* in 7DS (*T. aestivum*) (Tereshchenko et al., 2012).

The dominant allele of the *Pp3* gene, localized in the centromeric region of chromosome 2A, encodes a transcription factor with a bHLH regulatory domain (Shoeva et al., 2014). Tissue-specific transcriptional activity of the dominant *TaMyc1* allele, which is a likely candidate for *Pp3*, was shown in the colored pericarp of grains with lower expression levels in coleoptile, scales, and leaves. At the same time, *Pp-1* is expressed in many plant tissues (Shoeva et al., 2014; Jiang W. et al., 2018). It was found that *TaMyc1* has at least four copies in common wheat. In addition to *TaMyc1*, three copies, *TaMyc2–4* are localized in 2AL, 2BL and 2DL, respectively; however, none of these extra copies are transcribed in the pericarp. Comparison of *TaMyc1* expression in near-isogenic lines carrying different combinations of dominant and recessive alleles of *Pp-1* and *Pp3* showed that the dominant allele *Pp-D1* partially suppressed the transcription of *TaMyc1* in the pericarp (Shoeva et al., 2014).

Four allelic variants were found in the *TaPpm1* coding region: *TaPpm1a* (dominant, in purple wheat) and *TaPpm1b*, *TaPpm1c*, and *TaPpm1d*, which are nonfunctional due to differently sized insertions that cause frameshift or premature transcription termination (in uncolored wheat). There were six 261-bp tandem repeats in the promoter region of *TaPpb1* in the purple-grained varieties (*TaPpb1a* allele), while there was only one repeat unit present in the uncolored wheat varieties (*TaPpb1b*) (Jiang W. et al., 2018).

The expression of structural genes involved in the anthocyanin biosynthesis pathway is regulated by the MBW complex, which includes the R2R3-MYB, bHLH, and WD40 proteins. The allelic variations of *TaPpm1* influence anthocyanin pigmentation by altering the binding ability with bHLH, whereas variations in the *TaPpb1* promoter alter its expression level (Jiang W. et al., 2018).



Aglycon	R_1	R_2	Color
Cyanidin	OH	H	Purple
Delphinidin	OH	OH	Blue
Malvidin	OCH_3	OCH_3	Reddish purple
Pelargonidin	H	H	Orange-red
Peonidin	OCH_3	H	Bluish purple
Petunidin	OCH_3	OH	Purple

Fig. 2. Structure of major anthocyanidins found in wheat grain.

Anthocyanins content in colored wheat grain

Anthocyanins are water-soluble pigments related to flavonoids that give color to various parts of plants. The basic structure of anthocyanins is shown in Figure 2. Anthocyanidin (aglycone) is the base of the anthocyanin molecule. For most anthocyanins, the sugar moieties, most often glucose, galactose, arabinose and rutinose, are usually connected to anthocyanidins through O-glycosidic bonds at C3 position, sometimes at C3 and C5 positions. In addition, sugars can be acylated by aliphatic and aromatic acids (Francavilla, Joye, 2020). Cyanidin, delphinidin, malvidin, pelargonidin, peonidin, and petunidin are well-known anthocyanidins, which differ from each other in the number of hydroxyl or methoxyl groups (Fig. 2).

The total anthocyanin content (TAC) of wheat widely varies from 10 to 305 μg/g in purple grain, from 17 to 211 μg/g in blue grain and from 56 to 198 μg/g in black grain, but in general, black-grain wheat has a higher anthocyanin content. White and red grain wheat genotypes have the lowest TAC (7–10 μg/g) (Abdel-Aal, Hucl, 2003; Abdel-Aal et al., 2006; Varga et al., 2013; Garg et al., 2016; Kumari et al., 2020; Wang et al., 2020; Iannucci et al., 2022). In addition, whole wheat flour has a lower anthocyanin content compared to the bran fraction (Siebenhandl et al., 2007; Iannucci et al., 2022) (Supplementary Materials, Table S1)[1]. The anthocyanin content of blue wheat, on average, is higher than that of purple wheat, but some purple wheat contains much more anthocyanins than blue wheat (Abdel-Aal et al., 2016). It has been suggested that in blue grain and black grain, the location of the pigment in deeper layers leads to increased stability of anthocyanins (Garg et al., 2016). In addition, anthocyanins located in the aleurone layer are less firmly bound to cellular components than those in the pericarp, allowing them to be more easily extracted.

It was also found that the qualitative composition of anthocyanins differs between blue- grain and purple-grain wheat. In addition, the main anthocyanins in the genotypes with purple and blue grains are different. It is considered that purple-grain wheat varieties have a more complex anthocyanin composition than blue-grain wheat varieties but a lower TAC. Cyanidin-3-glucoside is the dominant anthocyanin in purple wheat grain. The most abundant anthocyanins in purple wheat along with cyanidin-3-glucoside are cyanidin-3-galactoside, cyanidin-3-rutinoside, cyanidin-3-(6″-malonyl glucoside), as well as delphinidin-3-galactoside, malvidin-3-glucoside, peonidin-3-glucoside, petunidin-3-glucoside, peonidin-3-(6″-malonyl glucoside) (Hosseinian et al., 2008; Abdel-Aal et al., 2018; Jiang Y. et al., 2024; Shamanin et al., 2024).

E.S.M. Abdel-Aal et al. (2006) found eight anthocyanins in purple grain such as cyanidin-3-glucoside, cyanidin-3,5-di-glucoside, peonidin-3-glucoside, and malonyl and succinyl derivatives of cyanidin and peonidin. E.S.M. Abdel-Aal et al. (2018) found a number of other anthocyanins in purple grain: delphinidin-3-rutinoside, malvidin-3-rutinoside, malvidin succinylglucoside, pelargonidin-3-(6″-malonylglucoside), peonidin-3-rutinoside, petunidin-3-(6″-malonylglucoside). F.S. Hosseinian et al. (2008) found the presence of 13 anthocyanins, which included arabinoside derivatives of cyanidin, delphinidin, pelargonidin and peonidin, glucoside derivatives of malvidin, pelargonidin and petunidin, and delphinidin-3-galactoside.

P. Bartl et al. (2015) identified a number of cyanidin glycosides with a hexose acetylated with malonic and/or acetic acid, delphinidin with a hexose acetylated with coumaric acid, peonidin with a hexose/rhamnose acetylated with malonic and/or acetic acid, and petunidin with two or three sugar moieties (hexose and rhamnose) acetylated with caffeic or coumaric acid. Y. Jiang et al. (2024) found the presence of 26 anthocyanin glycosides, including 12 acylated ones (acetyl-, malonyl-, and succinyl- derivatives). It was also shown that the TAC and content of individual anthocyanin glycosides increased with size reduction of the flour particle (coarse, fine and superfine flour samples).

The main anthocyanins in blue grains are delphinidin-3-rutinoside, delphinidin-3-glucoside, malvidin-3-glucoside (Ficco et al., 2014), delphinidin-3-rutinoside, cyanidin-3-glucoside, cyanidin-3-rutinoside (Abdel-Aal et al, 2006), malvinidin-3-glucoside, delphinidin-3-galactoside, cyanidin-3-glucoside (Sharma N. et al., 2020), delphinidin-3-rutinoside, delphinidin-3-glucoside, petunidin-3-glucoside (Iannucci et al., 2022). Blue grain has a high concentration of delphinidin-3-glucoside (9.9–56.5 μg/g) and delphinidin-3-rutinoside (35.9–72.5 μg/g) (Abdel-Aal et al., 2006; Ficco et al., 2014; Iannucci et al., 2022) or delphinidin-3-galactoside and malvidin-3-glucoside (Sharma N. et al., 2020). E.S.M. Abdel-Aal et al. (2006) found the presence of eight anthocyanins in blue-grain wheat: cyanidin-3-glucoside, cyanidin-3-rutinoside, delphinidin-3-glucoside, delphinidin-3-rutinoside, malvidin-3-rutinoside, peonidin-3-rutinoside, petunidin-3-glucoside, and petunidin-3-rutinoside. D.B.M. Ficco et al. (2014) showed the presence of eight anthocyanins and identified

[1] Tables S1–S5 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Chuman_Engl_29_3.pdf

new peonidine derivatives. P. Bartl et al. (2015) identified cyanidin, delphinidin, malvidin, peonidin and petunidin derivatives with 1, 2 or 3 hexose moieties with rhamnose or coumaric acid.

Black-grain wheat has not only a higher total anthocyanin content, but also a more diverse composition of anthocyanin glycosides. For example, N. Sharma et al. (2020) found 10 different anthocyanins in blue grain, 6 in purple grain and 11 in black grain. M. Garg et al. (2016) identified 22 different anthocyanins in blue wheat, 23 in purple wheat and 26 in black wheat including cyanidin-3-(6″-succinylglucoside), cyanidin-3-(2G-xylosylrutinoside), cyanidin-3-(3″,6″-dimalonylglucoside), cyanidin-3-(6″-feruloylglucoside)-5-glucoside, cyanidin-3-rutinoside-3'-glucoside, delphinidin-3-caffeoylglucoside, delphinidin-3-sambubioside, malvidin-3-rutinoside-5-glucoside, malvidin-3-(6″-p-caffeoylglucoside), pelargonidin-3-(6″-malonylglucoside), peonidin-3-rutinoside-5-glucoside, peonidin-3,5-diglucoside, petunidin-3-rutinoside-5-glucoside. It was found that black wheat has a high concentration of cyanidin-3-glucoside, cyanidin-3,5-di-glucoside, delphinidin-3-glucoside, delphinidin-3-galactoside, and malvidin-3-glucoside (Sharma N. et al., 2020; Shamanin et al., 2024). Red and white wheat varieties contain cyanidin, delphinidin, malvidin and peonidin anthocyanin derivatives in small concentrations, with the highest content of cyanidin-3-glucoside (Ficco et al., 2014; Garg et al., 2016; Sharma N. et al., 2020). More detailed information on the qualitative and quantitative content of anthocyanin glycosides in grains with different coloration is presented in Table S2.

It is assumed that qualitative and quantitative differences in anthocyanin composition may be due to genetic characteristics of the analyzed samples, as well as differences in the equipment used for grain grinding, extraction technologies and quantitative analysis of anthocyanins. Genetic characters cause variation in the qualitative and quantitative composition of anthocyanins in wheat. Each variety has an individual anthocyanin profile (Abdel-Aal, Hucl, 2003). The anthocyanin content is affected by environmental factors like temperature during grain filling period, drought, disease damage (Garg et al., 2022). E.S.M. Abdel-Aal and P. Hucl (2003) studied the anthocyanin content over three crop years. Blue wheat exhibited a reduced effect of environmental factors on anthocyanin content as compared to purple wheat, perhaps due to the location of the anthocyanins in different grain layers. D.V. Bustos et al. (2012) found that anthocyanin content increases rapidly during grain development and decreases before maturity. TAC decreases in the distal position of grains in the spike and when plants are shaded before tillering. On the contrary, TAC increases by halving the spikelet number per spike. Magnesium fertilization and early harvesting increases TAC in purple wheat by 65 and 39 %, respectively. X. Fan et al. (2020) showed that anthocyanin accumulation in purple wheat increases when grown under nitrogen-deficient conditions. According to R. Beleggia et al. (2021), late sowing dates of wheat increases TAC.

Content of phenolic compounds in colored wheat grain

Phenolic compounds are secondary metabolites that play an important role in the mechanisms of plant defense against UV radiation, pathogen suppression and ensuring the structural integrity of the cell wall. Phenolic acids are the most common class of phenolic compounds (Laddomada et al., 2017), the molecules of which consist of a phenolic ring and a carboxylic acid functional group. There are mainly two groups of phenolic acids in wheat: hydroxybenzoic (vanillic, syringic, p-hydroxybenzoic, gallic, salicylic, protocatechuic, ellagic, and gentisic acid), which have a C6-C1 structure, and hydroxycinnamic acid derivatives (ferulic, cinnamic, coumaric, caffeic, and sinapic acid), which are aromatic compounds with a three-carbon side chain (C6-C3) (Table S3). Individual compounds in each of these groups differ from each other by the presence and structure of side radicals.

Phenolic acids in wheat grains can take the following forms: insoluble, bound by ether and ether-ether bonds to cell wall components such as cellulose, arabinoxylan, lignin, and proteins (about 50–70 % on average); soluble, conjugated to sugars or other low molecular weight components (13–20 %); and soluble, free (0.5–2 %) (Menga et al., 2023).

A lot of studies have shown that the total content of soluble and insoluble phenolic compounds in colored wheat grain increases in the following order: white < purple < blue < black (Kumari et al., 2020; Paznocht et al., 2020) with up to 4–6 times higher content in colored grain compared to uncolored grain (Sharma S. et al., 2018; Kumari et al., 2020; Wang et al., 2020; Garg et al., 2022; Shamanin et al., 2022; Sahu et al., 2023). In general, regardless of the grain color, the grain shells that are removed during milling have the highest content of phenolic acids, as well as anthocyanins and only by using whole wheat flour products you can get all the benefits possible. The quantitative content of soluble and insoluble phenolic compounds in wheat grain with different colors is presented in Table S4.

Ferulic acid is the most abundant compound in wheat grain (65.0–94.9 % of all insoluble bound phenolic compounds) (Ma et al., 2021). It has been shown that the quantitative content of individual phenolic acids in free and bound forms can vary widely depending on the genotype. According to D. Ma et al. (2016), purple grain has higher contents of soluble and insoluble phenolic acids including ferulic acid, vanillin and caffeic acid than blue and red grain.

The study of free phenolic acids content in bran fractions showed that purple grain had maximum TPC (636–1,134 μg/g GAE (gallic acid equivalent)), while in blue and black grain the TPC was from 476 to 874 μg/g and from 495 to 590 μg/g, respectively. Gallic acid (29–33 μg/g), ferulic acid (and isoferulic acid) (59–66 μg/g) and salicylic acid (30–65 μg/g) had the highest content in all samples (Zhang et al., 2018). Among the bound phenolic acids, ferulic (from 1,726 μg/g in black grain to 2,620 μg/g in blue grain) and salicylic acids (from 535 μg/g in blue grain to 906.02 μg/g in black grain, respectively) had the highest content. Overall, phenolic acid content in both free and bound forms as well as AOA gradually decreased in the following order: outer bran > coarse bran > shorts.

V.P. Shamanin et al. (2022) showed that the total phenolic compound content ranged from 446 to 708 mg GAE/100 g in red wheat varieties (189–271 and 227–487 mg GAE/100 g free and bound, respectively), 457 and 767 mg GAE/100 g in blue-grain wheat lines (204 and 247 mg GAE/100 g; 253 and 520 mg GAE/100 g), 353–772 mg GAE/100 g in purple-grain

wheat lines (164–248 and 190–432 mg GAE/100 g) and 476–520 mg GAE/100 g (190–218 and 259–323 mg GAE/100 g) in black-grain wheat lines. In the bound fractions of some purple-grain wheat genotypes, as well as in F_4 black wheat hybrids, ferulic and sinapic acids had the highest content: 307–582 and 277–619 μg/g (in purple wheat); 257–424 and 272–450 μg/g (in black wheat); in some genotypes, ellagic or protocatechuic acids were predominant (31–89 and 90–157 μg/g). The free fraction was dominated by gallic, protocatechuic, and – in a number of purple-grain wheat samples – ellagic acid.

According to M. Bueno-Herrera and S. Pérez-Magariño (2020), vanillic (20.3–34.2 μg/g) and trans-ferulic (8.4–20.2 μg/g) prevailed in the free fraction, while cis- and trans-ferulic (245.1–304.6 μg/g), p-coumaric (8.8–9.9 μg/g) and vanillic acids (6.5–7.2 μg/g) were predominant in the bound fraction. A higher content of phenolic compounds was characteristic of the fine bran fraction (with a particle diameter of 200–800 μm) than in the coarse bran (with a particle diameter of 800–2,000 μm) and flour fractions. Ö.G. Geyik et al. (2023) showed that blue wheat had higher ferulic acid content in the bran fraction than purple and red wheat (2,264, 1,945 and 988 μg/100g, respectively). In the free fractions, p-coumaric acid (11.5 μg/100 g) had the highest content in red wheat and ellagic acid (14.7 and 11.5 μg/100 g, respectively) in purple wheat and black wheat.

D. Ma et al. (2016) studied the accumulation of phenolic acids in white, red, and purple wheat grains. They concluded that the maximum accumulation of ferulic and syringic acids was observed 14 days after flowering, while the levels of p-coumaric and caffeic acids reached the maximum level 7 days after flowering, and the levels of vanillic acid increased gradually during grain filling and reached the maximum level at the ripening stage (35 days after flowering). White wheat had higher phenolic acid contents and relatively high phenolic acid biosynthesis pathway genes (*TaPAL1*, *TaPAL2*, *TaC3H1*, *TaC3H2*, *TaC4H*, *Ta4CL1*, *Ta4CL2*, *TaCOMT1* and *TaCOMT2*) expression at the early stage, while purple wheat had the highest phenolic acid content and gene expression levels at later stages.

Antioxidant activity of colored wheat grains

Antioxidants are known to have the ability to neutralize and destroy free radicals that cause damage to cellular structures. The AOA of wheat is caused by anthocyanins, phenolic acids, flavones and flavonols. *In vitro* and *in vivo* grain AOA is assessed using a number of methods. The DPPH method is based on the registration of DPPH (2,2-diphenyl-1-picrylhydrazyl) radical reduction upon interaction with antioxidants. Other methods are also used, such as ABTS (decrease in the intensity of absorption by cations of the ABTS radical (2,2′-azino-bis-(3-ethylbenzothiazoline-6-sulfonic acid)), ORAC (oxygen radicals absorbance capacity) – ability to intercept peroxyl radicals, FRAP (ferric reducing antioxidant power) – reduction of trivalent iron complex ion (TPTZ (2, 4,6-3(2-pyridyl)-1,3,5-triazine)) concentration, CUPRAC (cupric reducing antioxidant capacity) – change in optical density in the reduction reaction of Cu^{2+} to Cu^{+}, PCL – chemiluminescence registration (Ma et al., 2016; Abdel-Aal et al., 2018; Sharma S. et al., 2018; Shamanin et al., 2024). The AOA of colored wheat grain compared to uncolored grain is mainly due to its higher anthocyanin content. C. Hu et al. (2007) found that 69 % of the total free radical scavenging capacity of blue wheat is determined by the content of anthocyanins, while 19 % is attributed to phenolic acids, and the contribution of bound ones is much higher than free ones (Zhang et al., 2018; Shamanin et al., 2022). Cyanidin-3-glucoside has the strongest AOA among anthocyanins – 3.5 times stronger than Trolox (vitamin E analog).

Several studies have shown that blue, purple and black wheat have higher AOA values compared to red and white grain wheat (Ficco et al., 2014; Ma et al., 2016; Sharma S. et al., 2018; Kumari et al., 2020; Wang et al., 2020). The highest AOA values (ABTS and DPPH) are characteristic of black wheat, then decrease in the following order: blue > purple > white, as well as TAC and TPC (Kumari et al., 2020; Sharma A. et al., 2023). AOA values of grains with different coloring determined using different methods are given in Table S5. Since compounds with antioxidant properties are found predominantly in the bran fraction, the AOA of bran is significantly higher than that of whole grain (Siebenhandl et al., 2007; Abdel-Aal et al., 2018; Iannucci et al., 2022; Saini et al., 2023). Y. Jiang et al. (2024) showed that the AOA of superfine flour was 1.18 and 1.62 times higher than that of coarse flour (ABTS and ORAC). A positive correlation was shown between the TPC in flour and AOA (r = 0.769 (ABTS) and r = 0.984 (FRAP)) (Li Y. et al., 2015), between TPC and ABTS (r = 0.97) (Ficco et al., 2014), between soluble phenolic compounds and DPPH (r = 0.65) (Sharma S. et al., 2018). Significant positive correlations were also observed between TAC and AOA (PCL) (r = 0.9) (Sharma S. et al., 2018), individual anthocyanins and ABTS (r = 0.65–0.91) (Shamanin et al., 2024).

Breeding achievements in Russia in obtaining colored-grain wheat varieties

In Russia, several research institutions are actively working on obtaining promising colored-grain wheat breeding lines and varieties. To date, three purple-grain wheat varieties have passed competitive variety testing and have been included in the register of breeding achievements: Nadira (FRC Kazan Scientific Center of RAS, 2022), Pamyati Konovalov (FSC of Legumes and Groat Crops and Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, 2023), and EF 22 (Omsk State Agrarian University named after P.A. Stolypin, 2024). These varieties are characterized by high levels of TAC and AOA, and therefore can be used for the production of functional foods. The variety Nadira was obtained by individual selection from the hybrid population F_3 L.22-95 / Kommissar (Vasilova et al., 2021). L.22-95, which was obtained at the Siberian Research Institute of Agriculture, was a donor of purple grain color. Nadira is recommended for cultivation in the Volga-Vyatka, Middle Volga and Ural regions. It is a medium-maturing variety. The average yield in the competitive variety trial for 2016–2018 was 4.8 t/ha (the standard Yoldyz yield was 4.7 t/ha). The variety is resistant to loose smut, moderately susceptible to leaf rust and powdery mildew. Drought tolerance of the Nadira variety is at the level of standard varieties. The variety contains 13.8 % protein in grain, 25.5 % crude gluten and has baking qualities corresponding to valuable varieties.

The variety Pamyati Konovalov was obtained by individual selection from a hybrid population: (Laval 19 × Granny) × Granny. Laval 19 (Canada) is the donor of purple color. This medium-maturing variety is recommended for cultivation in the Moscow region. Grain yield per plot in 2020–2021 was 451 and 284 g/m^2 (the standard variety Zlata yield was 593 and 408 g/m^2, respectively). Baking qualities of the variety are satisfactory and good, it is a good filler. The variety is resistant to lodging, septoriosis and fusariosis, moderately resistant to leaf rust (Rubets et al., 2022).

The variety Ivolga fioletovaya (Russian State Agrarian University – Moscow Timiryazev Agricultural Academy) is an isogenic purple wheat line of the variety Ivolga. Grain yield per plot in 2020–2021 was 417 and 358 g/m^2, which is lower than that of the Zlata variety. Ivolga fioletovaya is a medium-maturing variety, resistant to lodging, leaf rust and powdery mildew, but susceptible to fusarium and septoriosis (Rubets et al., 2022).

The efforts of scientists of OmSAU and ICIG SB RAS resulted in a number of promising lines from crossing lines of the S29 variety with *Ba* and *Pp* genes with Siberian varieties such as Element 22 (with Zn content more than 50 mg/kg), Aina, Tobolskaya and line BW49880 (CIMMYT, high Zn content) (Gordeeva et al., 2020; Shamanin et al., 2022). These lines are characterized by high TAC (maximum values of 254 and 326 μg/100 g in F_4 purple- and black-grain hybrids obtained based on BW 49880), phenolic compounds (767 and 599–772 mg GAE/100 g; 520 and 427–566 mg GAE/100 g total and bound phenolic acids in the blue-grain Blue 10 line (s:C29 4Th(4D)/Element 22) and in purple-grain BC_1F_4–BC_1F_5 hybrids of the Element 22 variety, respectively), AOA (CUPRAC) (482–494 mg TE/100 g (trolox equivalent) in the BC_1F_8 purple-grain lines of the Tobolskaya variety and F_4 hybrids of Element 22 and BW 49880), as well as Zn content from 44.5 to 56.5 mg/kg and Fe content from 53.5 to 65.5 mg/kg (Shamanin et al, 2022, 2024).

The EF 22 variety was created by marker-assisted selection using SSR markers flanking the *Pp-D1* and *Pp3* genes (Gordeeva et al., 2020) over six years. The donor of purple grain was the i:S29^{PF} line (introgression fragments from Purple Feed) (Arbuzova et al., 1998). EF 22 is a valuable medium-late variety, which has been included in the State Register for the Ural and West Siberian regions. The average yield for 2016–2020 was 3.12 t/ha, while the Element 22 variety had an average yield of 3.89 t/ha (Pototskaya et al., 2022). A study of the bread characteristics from whole wheat flour of EF 22, breeding lines Blue 10 and Purple 8 (Element 22*2/i:C29^{PF}) showed higher content of phenolic compounds and AOA and lower glycemic index compared to white wheat bread (Koksel et al., 2023).

Conclusion

The presented review summarized information on the genetic control of the regulation of anthocyanin accumulation and biosynthesis in the pericarp and aleurone layer by the *Ba*, *Pp-1* and *Pp3* genes. Information on anthocyanin content, phenolic compounds and AOA levels in wheat with different grain coloration is presented. Purple, blue and black wheat has higher TAC, TPC and AOA than uncolored wheat, and TAC, soluble and insoluble phenolic compounds and AOA values increase in the following order: purple > blue > black wheat. Purple-grain wheat as well as black-grain wheat have a more diverse anthocyanin compositions compared to blue-grain wheat. Colored-grain wheat varieties obtained by Russian breeders are a source of bioactive compounds that play an important role in disease prevention and can serve as a basis for the development of the anthocyanin-biofortified food industry in the internal market and increase the value of exported grain products.

References

Abdel-Aal E.S.M., Hucl P. Composition and stability of anthocyanins in blue-grained wheat. *J. Agric. Food Chem.* 2003;51(8):2174-2180. doi 10.1021/jf021043x

Abdel-Aal E.S.M., Young J.C., Rabalski I. Anthocyanin composition in black, blue, pink, purple, and red cereal grains. *J. Agric. Food Chem.* 2006;54(13):4696-4704. doi 10.1021/jf0606609

Abdel-Aal E.S.M., Hucl P., Shipp J., Rabalski I. Compositional differences in anthocyanins from blue- and purple-grained spring wheat grown in four environments in central Saskatchewan. *Cereal Chem.* 2016;93(1):32-38. doi 10.1094/cchem-03-15-0058-R

Abdel-Aal E.S.M., Hucl P., Rabalski I. Compositional and antioxidant properties of anthocyanin-rich products prepared from purple wheat. *Food Chem.* 2018;254:13-19. doi 10.1016/j.foodchem.2018.01.170

Arbuzova V.S., Maystrenko O.I., Popova O.M. Development of near-isogenic lines of the common wheat cultivar 'Saratovskaya 29'. *Cereal Res. Commun.* 1998;26:39-46. doi 10.1007/BF03543466

Arbuzova V.S., Badaeva E.D., Efremova T.T., Osadchaia T.S., Trubacheva N.V., Dobrovol'skaia O.B. A cytogenetic study of the blue-grain line of the common wheat cultivar Saratovskaya 29. *Russ. J. Genet.* 2012;48(8):785-791. doi 10.1134/S102279541205002X

Bartl P., Albreht A., Skrt M., Tremlová B., Ošádalová M., Šmejkal K., Vovk I., Ulrih N.P. Anthocyanins in purple and blue wheat grains and in resulting bread: quantity, composition, and thermal stability. *Int. J. Food Sci. Nutr.* 2015;66(5):514-519. doi 10.3109/09637486.2015.1056108

Beleggia R., Ficco D.B.M., Pecorella I., De Vita P., Nigro F.M., Giovanniello V., Colecchia S.A. Effect of sowing date on bioactive compounds and grain morphology of three pigmented cereal species. *Agronomy*. 2021;11(3):591. doi 10.3390/agronomy11030591

Bueno-Herrera M., Pérez-Magariño S. Validation of an extraction method for the quantification of soluble free and insoluble bound phenolic compounds in wheat by HPLC-DAD. *J. Cereal Sci.* 2020;93:102984. doi 10.1016/j.jcs.2020.102984

Burešová V., Kopecký D., Bartoš J., Martinek P., Watanabe N., Vyhnánek T., Doležel J. Variation in genome composition of blue-aleurone wheat. *Theor. Appl. Genet.* 2015;128(2):273-282. doi 10.1007/s00122-014-2427-3

Bustos D.V., Riegel R., Calderini D.F. Anthocyanin content of grains in purple wheat is affected by grain position, assimilate availability and agronomic management. *J. Cereal Sci.* 2012;55(3):257-264. doi 10.1016/j.jcs.2011.12.001

Dhua S., Kumar K., Kumar Y., Singh L., Sharanagat V.S. Composition, characteristics and health promising prospects of black wheat: a review. *Trends Food Sci. Technol.* 2021;112:780-794. doi 10.1016/j.tifs.2021.04.037

Eticha F., Grausgruber H., Siebenhandl-Ehn S., Berghofer E. Some agronomic and chemical traits of blue aleurone and purple pericarp wheat (*Triticum* L.). *J. Agric. Sci. Technol.* 2011;1:48-58

Fan X., Xu Z., Wang F., Feng B., Zhou Q., Cao J., Ji G., Yu Q., Liu X., Liao S., Wang T. Identification of colored wheat genotypes with suitable quality and yield traits in response to low nitrogen input. *PLoS One*. 2020;15(4):0229535. doi 10.1371/journal.pone.0229535

Ficco D.B.M., De Simone V., Colecchia S.A., Pecorella I., Platani C., Nigro F., Finocchiaro F., Papa R., De Vita P. Genetic variability in anthocyanin composition and nutritional properties of blue, purple,

and red bread (*Triticum aestivum* L.) and durum (*Triticum turgidum* L. ssp. *turgidum* convar. *durum*) wheats. *J. Agric. Food Chem.* 2014;62(34):8686-8695. doi 10.1021/jf5003683

Francavilla A., Joye I.J. Anthocyanins in whole grain cereals and their potential effect on health. *Nutrients*. 2020;12(10):2922. doi 10.3390/nu12102922

Gamel T.H., Muhammad S., Saeed G., Ali R., Abdel-Aal E.S.M. Purple wheat: food development, anthocyanin stability, and potential health benefits. *Foods*. 2023;12(7):1358. doi 10.3390/foods12071358

Garg M., Chawla M., Chunduri V., Kumar R., Sharma S., Sharma N.K., Kaur N., Kumar A., Mundey J.K., Saini M.K., Singh S.P. Transfer of grain colors to elite wheat cultivars and their characterization. *J. Cereal Sci.* 2016;71:138-144. doi 10.1016/j.jcs.2016.08.004

Garg M., Kaur S., Sharma A., Kumari A., Tiwari V., Sharma S., Kapoor P., Sheoran B., Goyal A., Krishania M. Rising demand for healthy foods-anthocyanin biofortified colored wheat is a new research trend. *Front. Nutr.* 2022;9:878221. doi 10.3389/fnut.2022.878221

Geyik Ö.G., Tekin-Cakmak Z.H., Shamanin V.P., Karasu S., Pototskaya I.V., Shepelev S.S., Chursin A.S., Morgounov A.I., Yaman M., Sagdic O., Koksel H. Effects of phenolic compounds of colored wheats on colorectal cancer cell lines. *Qual. Assur. Saf. Crop. Foods*. 2023;15(4):21-31. doi 10.15586/qas.v15I4.1354

Gordeeva E.I., Shoeva O.Y., Khlestkina E.K. Marker-assisted development of bread wheat near-isogenic lines carrying various combinations of purple pericarp (*Pp*) alleles. *Euphytica*. 2015;203(2):469-476. doi 10.1007/S10681-014-1317-8/FIGURES/2

Gordeeva E., Shamanin V., Shoeva O., Kukoeva T., Morgounov A., Khlestkina E. The strategy for marker-assisted breeding of anthocyanin-rich spring bread wheat (*Triticum aestivum* L.) cultivars in Western Siberia. *Agronomy*. 2020;10(10):1603. doi 10.3390/agronomy10101603

Gordeeva E., Shoeva O., Mursalimov S., Adonina I., Khlestkina E. Fine points of marker-assisted pyramiding of anthocyanin biosynthesis regulatory genes for the creation of black-grained bread wheat (*Triticum aestivum* L.) lines. *Agronomy*. 2022;12(12):2934. doi 10.3390/agronomy12122934

Granda L., Rosero A., Benešová K., Pluháčková H., Neuwirthová J., Cerkal R. Content of selected vitamins and antioxidants in colored and nonpigmented varieties of quinoa, barley, and wheat grains. *J. Food Sci.* 2018;83(10):2439-2447. doi 10.1111/1750-3841.14334

Himi E., Maekawa M., Miura H., Noda K. Development of PCR markers for *Tamyb10* related to *R-1*, red grain color gene in wheat. *Theor. Appl. Genet.* 2011;122(8):1561-1576. doi 10.1007/s00122-011-1555-2

Hosseinian F.S., Li W., Beta T. Measurement of anthocyanins and other phytochemicals in purple wheat. *Food Chem.* 2008;109(4):916-924. doi 10.1016/j.foodchem.2007.12.083

Hu C., Cai Y.Z., Li W., Corke H., Kitts D.D. Anthocyanin characterization and bioactivity assessment of a dark blue grained wheat (*Triticum aestivum* L. cv. Hedong Wumai) extract. *Food Chem.* 2007;104(3):955-961. doi 10.1016/j.foodchem.2006.12.064

Iannucci A., Suriano S., Cancellaro S., Trono D. Anthocyanin profile and main antioxidants in pigmented wheat grains and related milling stream fractions. *Cereal Chem.* 2022;99(6):1282-1295. doi 10.1002/cche.10591

Jiang W., Liu T., Nan W., Jeewani D.C., Niu Y., Li C., Wang Y., Shi X., Wang C., Wang J., Li Y., Gao X., Wang Z. Two transcription factors *TaPpm1* and *TaPpb1* co-regulate anthocyanin biosynthesis in purple pericarps of wheat. *J. Exp. Bot.* 2018;69(10):2555-2567. doi 10.1093/jxb/ery101

Jiang Y., Qi Z., Li J., Gao J., Xie Y., Henry C.J., Zhou W. Role of superfine grinding in purple-whole-wheat flour. Part I: Impacts of size reduction on anthocyanin profile, physicochemical and antioxidant properties. *LWT*. 2024;197:115940. doi 10.1016/j.lwt.2024.115940

Khlestkina E.K., Röder M.S., Börner A. Mapping genes controlling anthocyanin pigmentation on the glume and pericarp in tetraploid wheat (*Triticum durum* L.). *Euphytica*. 2010;171(1):65-69. doi 10.1007/s10681-009-9994-4

Khlestkina E.K., Usenko N.I., Gordeeva E.I., Stabrovskaya O.I., Sharfunova I.B., Otmakhova Y.S. Evaluation of wheat products with high flavonoid content: justification of importance of marker-assisted development and production of flavonoid-rich wheat cultivars. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2017;21(5):545-553. doi 10.18699/VJ17.25-o (in Russian)

Koksel H., Cetiner B., Shamanin V.P., Tekin-Cakmak Z.H., Pototskaya I.V., Kahraman K., Sagdic O., Morgounov A.I. Quality, nutritional properties, and glycemic index of colored whole wheat breads. *Foods*. 2023;12(18):3376. doi 10.3390/foods12183376

Kumari A., Sharma S., Sharma N., Chunduri V., Kapoor P., Kaur S., Goyal A., Garg M. Influence of biofortified colored wheats (purple, blue, black) on physicochemical, antioxidant and sensory characteristics of chapatti (indian flatbread). *Molecules*. 2020;25(21):5071. doi 10.3390/molecules25215071

Laddomada B., Caretto S., Mita G. Wheat bran phenolic acids: bioavailability and stability in whole wheat-based foods. *Molecules*. 2015;20(9):15666-15685. doi 10.3390/molecules200915666

Laddomada B., Durante M., Mangini G., D'Amico L., Lenucci M.S., Simeone R., Piarulli L., Mita G., Blanco A. Genetic variation for phenolic acids concentration and composition in a tetraploid wheat (*Triticum turgidum* L.) collection. *Genet. Resour. Crop Evol.* 2017;64(3):587-597. doi 10.1007/s10722-016-0386-z

Li N., Li S., Zhang K., Chen W., Zhang B., Wang D., Liu D., Liu B., Zhang H. *ThMYC4E*, candidate *Blue aleurone 1* gene controlling the associated trait in *Triticum aestivum*. *PLoS One*. 2017;12(7):0181116. doi 10.1371/journal.pone.0181116

Li Y., Ma D., Sun D., Wang C., Zhang J., Xie Y., Guo T. Total phenolic, flavonoid content, and antioxidant activity of flour, noodles, and steamed bread made from different colored wheat grains by three milling methods. *Crop J.* 2015;3(4):328-334. doi 10.1016/j.cj.2015.04.004

Liu X., Zhang M., Jiang X., Li H., Jia Z., Hao M., Jiang B., Huang L., Ning S., Yuan Z., Chen Xuejiao, Chen Xue, Liu D., Liu B., Zhang L. *TbMYC4A* is a candidate gene controlling the blue aleurone trait in a wheat-triticum boeoticum substitution line. *Front. Plant Sci.* 2021;12:762265. doi 10.3389/fpls.2021.762265

Liu Xin, Feng Z., Liang D., Zhang M., Liu Xiaojuan, Hao M., Liu D., Ning S., Yuan Z., Jiang B., Chen Xuejiao, Chen Xue, Zhang L. Development, identification, and characterization of blue-grained wheat – *Triticum boeoticum* substitution lines. *J. Appl. Genet.* 2020;61(2):169-177. doi 10.1007/s13353-020-00553-9

Ma D., Li Y., Zhang J., Wang C., Qin H., Ding H., Xie Y., Guo T. Accumulation of phenolic compounds and expression profiles of phenolic acid biosynthesis-related genes in developing grains of white, purple, and red wheat. *Front. Plant Sci.* 2016;7:185202. doi 10.3389/fpls.2016.00528

Ma D., Wang C., Feng J., Xu B. Wheat grain phenolics: a review on composition, bioactivity, and influencing factors. *J. Sci. Food Agric.* 2021;101(15):6167-6185. doi 10.1002/JSFA.11428

Menga V., Giovanniello V., Savino M., Gallo A., Colecchia S.A., De Simone V., Zingale S., Ficco D.B.M. Comparative analysis of qualitative and bioactive compounds of whole and refined flours in durum wheat grains with different year of release and yield potential. *Plants*. 2023;12(6):1350. doi 10.3390/plants12061350

Mohammadi N., Farrell M., O'Sullivan L., Langan A., Franchin M., Azevedo L., Granato D. Effectiveness of anthocyanin-containing foods and nutraceuticals in mitigating oxidative stress, inflammation, and cardiovascular health-related biomarkers: a systematic review of animal and human interventions. *Food Funct.* 2024;15(7):3274-3299. doi 10.1039/d3fo04579j

Paznocht L., Kotíková Z., Burešová B., Lachman J., Martinek P. Phenolic acids in kernels of different coloured-grain wheat genotypes. *Plant Soil Environ.* 2020;66(2):57-64. doi 10.17221/380/2019-PSE

Pototskaya I.V., Nardin D.S., Yurkinson A.V., Pototskaya A.A., Shamanin V.P. Prospects of "colored wheat" for functional nutrition. In: Abstracts from the Int. conf. "Vavilov Readings – 2022". Nov. 22–25, 2022. Saratov, Russia, 2022;190-194 (in Russian)

Razgonova M.P., Zakharenko A.M., Gordeeva E.I., Shoeva O.Y., Antonova E.V., Pikula K.S., Koval L.A., Khlestkina E.K., Golokhvast K.S. Phytochemical analysis of phenolics, sterols, and terpenes in colored wheat grains by liquid chromatography with tandem mass spectrometry. *Molecules*. 2021;26(18):5580. doi 10.3390/molecules26185580

Rubets V.S., Voronchikhina I.N., Igonin V.N., Sidorenko V.S., Voronchikhin V.V. Characteristics of violet-green variety of spring soft wheat in the conditions of the central region of the Non-Chernozem zone of Russia. *Mezhdunarodnyi Sel'skokhoziaystvennyi Zhurnal = Int. Agric. J*. 2022;5:525-529. doi 10.55186/25876740_2022_65_5_525 (in Russian)

Sahu R., Mandal S., Das P., Ashraf G.J., Dua T.K., Paul P., Nandi G., Khanra R. The bioavailability, health advantages, extraction method, and distribution of free and bound phenolics of rice, wheat, and maize: a review. *Food Chem. Adv.* 2023;3:100484. doi 10.1016/j.focha.2023.100484

Saini P., Kumar N., Kumar S., Panghal A., Attkan A.K. Analysis of engineering properties, milling characteristics, antioxidant potential, and nutritional benefits of purple wheat and its bran. *Food Bioeng.* 2023;2(4):406-419. doi 10.1002/fbe2.12073

Shamanin V.P., Tekin-Cakmak Z.H., Gordeeva E.I., Karasu S., Pototskaya I., Chursin A.S., Pozherukova V.E., Ozulku G., Morgounov A.I., Sagdic O., Koksel H. Antioxidant capacity and profiles of phenolic acids in various genotypes of purple wheat. *Foods*. 2022;11(16):2515. doi 10.3390/foods11162515

Shamanin V.P., Tekin-Cakmak Z.H., Karasu S., Pototskaya I.V., Gordeeva E.I., Verner A.O., Morgounov A.I., Yaman M., Sagdic O., Koksel H. Antioxidant activity, anthocyanin profile, and mineral compositions of colored wheats. *Qual. Assur. Saf. Crop. Foods*. 2024;16(1):98-107. doi 10.15586/qas.v16I1.1414

Sharma A., Yadav M., Tiwari A., Ali U., Krishania M., Bala M., Sharma P., Goudar G., Roy J.K., Navik U., Garg M. A comparative study of colored wheat lines across laboratories for validation of their phytochemicals and antioxidant activity. *J. Cereal Sci.* 2023;112:103719. doi 10.1016/j.jcs.2023.103719

Sharma N., Tiwari V., Vats S., Kumari A., Chunduri V., Kaur S., Kapoor P., Garg M. Evaluation of anthocyanin content, antioxidant potential and antimicrobial activity of black, purple and blue colored wheat flour and wheat-grass juice against common human pathogens. *Molecules*. 2020;25(24):5785. doi 10.3390/molecules25245785

Sharma S., Chunduri V., Kumar A., Kumar R., Khare P., Kondepudi K.K., Bishnoi M., Garg M. Anthocyanin bio-fortified colored wheat: nutritional and functional characterization. *PLoS One*. 2018;13(4):0194367. doi 10.1371/journal.pone.0194367

Shen Y., Shen J., Dawadondup, Zhuang L., Wang Y., Pu J., Feng Y., Chu C., Wang X., Qi Z. Physical localization of a novel blue-grained gene derived from *Thinopyrum bessarabicum*. *Mol. Breed.* 2013;31(1):195-204. doi 10.1007/s11032-012-9783-y

Shoeva O.Y., Gordeeva E.I., Khlestkina E.K. The regulation of anthocyanin synthesis in the wheat pericarp. *Molecules*. 2014;19(12):20266-20279. doi 10.3390/molecules191220266

Siebenhandl S., Grausgruber H., Pellegrini N., Del Rio D., Fogliano V., Pernice R., Berghofer E. Phytochemical profile of main antioxidants in different fractions of purple and blue wheat, and black barley. *J. Agric. Food Chem.* 2007;55(21):8541-8547. doi 10.1021/jf072021j

Singh K., Ghai M., Garg M., Chhuneja P., Kaur P., Schnurbusch T., Keller B., Dhaliwal H.S. An integrated molecular linkage map of diploid wheat based on a *Triticum boeoticum* × *T. monococcum* RIL population. *Theor. Appl. Genet.* 2007;115:301-312. doi 10.1007/s00122-007-0543-z

Tereshchenko O., Gordeeva E., Arbuzova V., Börner A., Khlestkina E. The D genome carries a gene determining purple grain colour in wheat. *Cereal Res. Commun.* 2012;40(3):334-341. doi 10.1556/crc.40.2012.3.2

Tian S., Chen Z., Wei Y. Measurement of colour-grained wheat nutrient compounds and the application of combination technology in dough. *J. Cereal Sci.* 2018;83:63-67. doi 10.1016/j.jcs.2018.07.018

Varga M., Bánhidy J., Cseuz L., Matuz J. The anthocyanin content of blue and purple coloured wheat cultivars and their hybrid generations. *Cereal Res. Commun.* 2013;41(2):284-292. doi 10.1556/crc.41.2013.2.10

Vasilova N.Z., Askhadullin D.F., Askhadullin D.F., Bagavieva E.Z., Tazutdinova M.R., Khusainova I.I. Violet-green variety of spring soft wheat Nadira. *Zernobobovye i Krupânye Kul'tury = Legumes Groat Crops*. 2021;4(40):66-75. doi 10.24412/2309-348X-2021-4-66-75 (in Russian)

Wang X., Zhang X., Hou H., Ma X., Sun S., Wang H., Kong L. Metabolomics and gene expression analysis reveal the accumulation patterns of phenylpropanoids and flavonoids in different colored-grain wheats (*Triticum aestivum* L.). *Food Res. Int.* 2020;138:109711. doi 10.1016/j.foodres.2020.109711

Zhang J., Ding Y., Dong H., Hou H., Zhang X. Distribution of phenolic acids and antioxidant activities of different bran fractions from three pigmented wheat varieties. *J. Chem.* 2018;1:459243. doi 10.1155/2018/6459243

Zheng Q., Li B., Mu S., Zhou H., Li Z. Physical mapping of the blue-grained gene(s) from *Thinopyrum ponticu*m by GISH and FISH in a set of translocation lines with different seed colors in wheat. *Genome*. 2006;49(9):1109-1114. doi 10.1139/g06-073

Conflict of interest. The authors declare no conflict of interest.

Received June 3, 2024. Revised September 9, 2024. Accepted September 11, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):401-413

doi 10.18699/vjgb-25-43

The effects of *Non3* mutations on chromatin organization in *Drosophila melanogaster*

A.A. Yushkova #, A.A. Ogienko #, E.N. Andreyeva, A.V. Pindyurin, A.E. Letiagina, E.S. Omelina

Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
omelina@mcb.nsc.ru

Abstract. The nucleolus is a large membraneless subnuclear structure, the main function of which is ribosome biogenesis. However, there is growing evidence that the function of the nucleolus extends beyond this process. While the nucleolus is the most transcriptionally active site in the nucleus, it is also the compartment for the location and regulation of repressive genomic domains and, like the nuclear lamina, is the hub for the organization of inactive heterochromatin. Studies in human and *Drosophila* cells have shown that a decrease in some nucleolar proteins leads to changes in nucleolar morphology, heterochromatin organization and declustering of centromeres. This work is devoted to the study of the effects of *Novel nucleolar protein 3* (*Non3*) gene mutations in *D. melanogaster* on the organization of chromatin in the nucleus. Previously, it was shown that partial deletion of the *Non3* gene leads to embryonic lethality, and a decrease in NON3 causes an extension of ontogenesis and formation of a *Minute*-like phenotype in adult flies. In the present work, we have shown that mutations in the *Non3* gene suppress the position effect variegation (PEV) and increase the frequency of meiotic recombination. We have analyzed the classical heterochromatin markers in *Non3* mutants and shown that the amount of the HP1 protein as well as the modification of the histone H3K9me2 do not change significantly in larval brains and salivary glands compared to the control in Western blot analysis. Immunostaining with antibodies to HP1 and H3K9me2 did not reveal a significant reduction or change in the localization patterns of these proteins in the pericentromeric regions of salivary gland polytene chromosomes either. We analyzed the localization of the HP1 protein in *Non3* mutants using DNA adenine methyltransferase identification (DamID) analysis and did not find substantial differences in protein distribution compared to the control. In hemocytes of *Non3* mutants, we observed changes in the morphology of the nucleolus and in the size of the region detected by anti-centromere antibodies, but this was not accompanied by declustering of centromeres and their untethering from the nucleolar periphery. Thus, the NON3 protein is important for the formation/function of the nucleolus and is required for the correct chromatin packaging, but the exact mechanism of NON3 involvement in these processes requires further investigations.

Key words: nucleolus; NON3; HP1; CID; H3K9me2; chromatin; pericentromeric regions of chromosome; PEV; *Su(var)205*; *Su(var)3-9*; *Drosophila*

For citation: Yushkova A.A., Ogienko A.A., Andreyeva E.N., Pindyurin A.V., Letiagina A.E., Omelina E.S. The effects of *Non3* mutations on chromatin organization in *Drosophila melanogaster*. *Vavilovskii Zhurnal Genetiki i Selektsii* = *Vavilov J Genet Breed*. 2025;29(3):401-413. doi 10.18699/vjgb-25-43

Funding. This research was funded by the Russian Science Foundation, grant no. 23-24-00619 (https://rscf.ru/en/project/23-24-00619/).

Acknowledgements. We thank Prof. Peter Verrijzer for anti-HP1 antibodies; Harald Saumweber, for anti-β-Tubulin antibodies; and Prof. Gunter Reuter, for *Su(var)205*^{5} and *Su(var)3-9*^{6} fly stocks.

Влияние мутаций гена *Non3* на организацию хроматина у *Drosophila melanogaster*

А.А. Юшкова #, А.А. Огиенко #, Е.Н. Андреева, А.В. Пиндюрин, А.Е. Летягина, Е.С. Омелина

Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
omelina@mcb.nsc.ru

Аннотация. Ядрышко – это крупная безмембранная субъядерная структура, где происходит биогенез рибосом. Однако известно все больше данных о том, что ядрышко выполняет и другие функции в клетке. Помимо того, что в ядрышке происходят активный синтез и процессинг рРНК, оно является компартментом, на периферии которого локализованы репрессированные участки генома. Таким образом, наряду с ядерной ламиной ядрышко выступает центром организации гетерохроматина. Исследования на клетках человека и дрозофилы показали, что снижение количества отдельных ядрышковых белков приводит к изменениям

# These authors contributed equally.

в морфологии ядрышка, организации гетерохроматина и декластеризации центромер. Данная работа посвящена изучению влияния мутаций в гене *Novel nucleolar protein 3* (*Non3*) *D. melanogaster* на организацию хроматина в ядре. Ранее было показано, что делеция части гена *Non3* приводит к эмбриональной летальности, а снижение количества белка NON3 – к замедлению онтогенеза и формированию *Minute*-подобного фенотипа у взрослых мух. В настоящей работе мы продемонстрировали, что мутации в гене *Non3* супрессируют эффект положения мозаичного типа и увеличивают частоту мейотической рекомбинации. Общее количество классических маркеров гетерохроматина, белка HP1 и модификации гистона H3K9me2, в мозгах и слюнных железах личинок мутантов *Non3* существенно не отличается от контроля, согласно оценке вестерн-блот анализом. Иммуноокрашивание антителами к HP1 и H3K9me2 также не выявило значительного изменения в количестве и паттерне локализации этих белков в прицентромерных районах политенных хромосом слюнных желез. Изучив локализацию белка HP1 у мутантов по гену *Non3* с помощью метода DamID (DNA adenine methyltransferase identification), мы также не обнаружили значимых отличий в распределении белка по сравнению с контролем. В гемоцитах мутантов *Non3* мы наблюдали изменение морфологии ядрышка и размера области, выявляемой антицентромерными антителами, но это не сопровождалось декластеризацией центромер и отхождением их от периферии ядрышка. Таким образом, белок NON3 важен для формирования/функционирования ядрышка и необходим для правильной упаковки хроматина, однако точный механизм участия NON3 в данных процессах требует дальнейшего изучения.
Ключевые слова: ядрышко; NON3; HP1; CID; H3K9me2; хроматин; прицентромерные районы хромосом; эффект положения; *Su(var)205*; *Su(var)3-9*; дрозофила

Introduction

The nucleolus is a membraneless organelle that forms through phase separation in the nucleus. It is formed around nucleolus organizer regions (NORs), which contain ribosomal gene (rDNA) clusters encoding rRNAs (Pavlakis et al., 1979; Smirnov et al., 2016; Trinkle-Mulcahy, 2018). The main function of the nucleolus is the synthesis and processing of rRNA, production of the small 40S and large 60S ribosome subunits and ribosome assembly (Panse, Johnson, 2010). During interphase, the nucleolus can be divided into three compartments: the fibrillar center (FC), the dense fibrillar component (DFC), and the outer granular component (GC). The FC border is responsible for rDNA transcription, while the DFC and GC are involved in rRNA processing and the assembly of ribosomes, respectively. The FC is enriched in components of the RNA pol I machinery, such as the transcription factor UBF, whereas the DFC harbors various RNA-modifying enzymes and pre-rRNA processing factors including Fibrillarin (Boisvert et al., 2007; Boulon et al., 2010; Hernandez-Verdun et al., 2010; Razin, Ulianov, 2022). While the nucleolus is the most transcriptionally active site in the nucleus, it is also the compartment for the location and regulation of repressive genomic domains and, like the nuclear lamina, represents the hub for the organization of heterochromatin (Janssen et al., 2018; Quinodoz et al., 2018; Bersaglieri, Santoro, 2019; Iarovaia et al., 2019). Indeed, a shell of perinucleolar heterochromatin composed of silent rDNA, repetitive satellite DNA, heterochromatic regions from non-NOR-bearing chromosomes and pericentromeric/centromeric regions of chromosomes is often located close to the GC of the nucleolus (Németh, Längst, 2011).

Large segments of the eukaryotic genome are packaged in heterochromatin domains characterized by late replication and a low level of meiotic recombination. These domains containing arrays of repetitive sequences and transposable elements are enriched in H3K9me2/3 and harbor a small number of essential protein-coding genes. About one third of the *Drosophila* genome is considered heterochromatic, including the entire Y chromosome, most of the small chromosome 4 and the pericentric regions that cover 40 and 20 % of the X chromosome and the large autosomes, respectively (Grewal, Jia, 2007; Smith et al., 2007; Elgin, Reuter, 2013; Allshire, Madhani, 2018; Janssen et al., 2018). The best studied non-histone components of heterochromatin are HP1 encoded by the *Su(var)205* gene (Lu et al., 2000) and Su(var)3-9 methyltransferase encoded by the gene of the same name (Schotta et al., 2002). HP1 is mainly found in the chromocenter, telomeres, chromosome 4, and in some sites on the chromosome arms (Meyer-Nava et al., 2020). Su(var)3-9 performs di- and trimethylation of H3K9, which is necessary for the specific binding of the HP1 protein (Rea et al., 2000; Schotta et al., 2002). Su(var)3-9 associates with the histone deacetylase HDAC1 (Czermin et al., 2001) and concerted histone deacetylation and methylation by a Su(var)3-9/HDAC1-containing complex leads to permanent silencing of transcription in particular regions of the genome (Czermin et al., 2001). HP1 binding recruits additional Su(var)3-9 to methylate the adjacent nucleosome, which provides another binding site for HP1 in a self-propagating process (Schotta et al., 2003; Sentmanat, Elgin, 2012).

Centromeres are specialized domains of heterochromatin that provide the foundation for the kinetochore (Bloom, 2014). These multiprotein structures play an essential role during cell division by connecting chromosomes to spindle microtubules in mitosis and meiosis to mediate accurate chromosome segregation (Heun et al., 2006; Kyriacou, Heun, 2023). Centromeres are marked by the histone H3 variant centromere protein A (CENP-A, also called centromere identifier (CID) in *Drosophila*), which is necessary and sufficient for kinetochore activity (Bloom, 2014; Chang et al., 2019). Clustering and positioning of centromeres near the nucleolus is essential for the stable organization of pericentric heterochromatin in *Drosophila* (Padeken et al., 2013). Studies on human and *Drosophila* cells have provided evidence that a decrease in some nucleolar proteins can cause the repositioning of heterochromatin away from the nucleolar periphery and declustering of centromeres during interphase (Padeken et al., 2013; Rodrigues et al., 2023). The depletion of nucleolar

proteins such as Nucleolin and Nucleophosmin led to changes in nucleolar morphology and heterochromatin organization (including decreased levels of H3K9me3 and HP1 foci at perinucleolar regions), as well as to mitotic defects (Olausson et al., 2014; Bizhanova, Kaufman, 2021). The decrease of *Drosophila* nucleolar protein Modulo (Nucleolin orthologue) led to declustering and untethering of centromeres from the nucleolar periphery (Padeken et al., 2013).

Previously, we demonstrated that *Drosophila* protein NON3 localizes to the nucleolus in larval brain cells. Null allele of the *Non3* gene ($Non3^{\Delta 600}$) causes early larval lethality, whereas viable combinations of hypomorphic alleles ($Non3^{G4706}/Non3^{259}$, $Non3^{G4706}/Non3^{197}$, $Non3^{G4706}/Non3^{310}$) result in *Minute*-like phenotype, which is manifested as prolonged development, poor viability and fertility, as well as abnormally short and thin bristles (Andreyeva et al., 2019). The NON3 protein belongs to the group of Brix domain-containing proteins, which are highly conserved from archaea to humans (Eisenhaber et al., 2001; Maekawa et al., 2018). The Brix domain is supposed to function as a structural hub for interactions with both proteins and RNA, mediated by its N- and C-terminal halves, respectively (Maekawa et al., 2018). The NON3 orthologous proteins are Ribosome production factor 2 (Rpf2) in *S. cerevisiae* (Morita et al., 2002), ARPF2 in *A. thaliana* (Maekawa et al., 2018) and RPF2 in humans (Hirano et al., 2009). All orthologues localize in the nucleolus. Human RPF2 and NON3 exhibit 66 % similarity and 47 % sequence identity (Gramates et al., 2017). Despite the fact that human RPF2 is an rRNA-interacting protein involved in pre-rRNA processing, it was isolated together with 172 proteins embedded in heterochromatic H3K9me3 domains in the course of proteomic analysis of purified H3K9me3-marked heterochromatin in human fibroblasts (Becker et al., 2017). The fact that RNA-binding proteins remain strongly enriched in H3K9me3-marked chromatin provides strong support for these proteins having a role in heterochromatin maintenance (Becker et al., 2017).

Here, we describe the role of the conserved NON3 protein in position-effect variegation (PEV) and show that *Non3* mutations are weak suppressors of PEV. We also show that $Non3^{\Delta 600}$ background slightly enhances meiotic recombination. However, neither immunostaining for HP1 nor genome-wide DamID-seq mapping of HP1 binding to salivary gland polytene chromosomes reveals any substantial changes between the control and *Non3* mutants. Finally, we provide evidence that *Non3* mutations affect the size of the nucleolus and the region detected by anti-centromere antibodies in larval hemocytes, but do not affect the clustering of centromeres and their positioning relative to the nucleolus. Identification of new functions of nucleolar proteins may provide new insights into the functions of the nucleolus.

Materials and methods

Fly stocks. All fly lines used in this work are presented in Table 1. Flies were raised and crossed on standard cornmeal agar media at 25 °C unless otherwise stated.

Generation of the *sgs3*-FLP transgenic construct and *Drosophila* germline transformation. To make the *sgs3*-FLP construct, a cassette consisting of a 1345-bp genomic DNA fragment [chr3L:11510890–11512234, but with 11510939A>G, 11511517G>T, 11511567A>C and 11511624T>A substitutions; the coordinates are from Release 6 of the *D. melanogaster* genome assembly (Hoskins et al., 2015)] spanning the salivary gland-specific *Sgs3* gene promoter (Biyasheva et al., 2001; McPherson et al., 2024; Suárez Freire et al., 2024) and the FLP recombinase coding sequence were cloned upstream of the SV40 poly(A) sequence of the pattB vector (DGRC Stock 1420; https://dgrc.bio.indiana.edu//stock/1420; RRID:DGRC_1420). Details of plasmid construction are available upon request. The *sgs3*-FLP construct was integrated into the genome at the attP154 site (chromosome 3R) (Petersen, Stowers, 2011) mainly as described previously (Bischof et al., 2007) using the fly stock BDSC #36347.

Eye pigment analysis. Red eye pigment extraction and analysis were performed as described previously (Connolly et al., 1969) with the following modifications. Adult flies were aged for 3 days at 18 °C before measurement. For analysis, we took thirty heads of each sex per genotype. The optical density was measured at 480 nm using a Multimode Microplate Reader (Tecan SPARK® 10M).

Meiotic recombination analysis. We counted crossing-over frequencies along chromosome 3 using two different fly strains (##306, 620) carrying recessive marker mutations and the $Non3^{\Delta 600}$ mutation. $Non3^{ex}$ was used as control. For each studied chromosomal region, the frequency of meiotic recombination between markers was calculated by dividing the number of recombinant progeny by the total number of flies analyzed.

Western blotting. Immunoblotting was performed as described earlier (Andreyeva et al., 2019). The following primary antibodies were used: mouse anti-β-Tubulin (1:800; BX69 (Tavares et al., 1996), kindly provided by Prof. Harald Saumweber), mouse anti-Non3 (1:5,000 (Andreyeva et al., 2019)), mouse anti-HP1 (1:800, Developmental Studies Hybridoma Bank (DSHB) C1A9), mouse anti-H3K9me2 antibody (1:400, Abcam 1220). The primary antibodies were detected with HRP-conjugated goat anti-mouse IgG (1:3,500; Life Technology G-21040) and images were captured using an Amersham Imager 600 System (GE Healthcare). Band intensities were analyzed using ImageJ. The intensity of each band was normalized with the intensity of the corresponding loading control.

Immunostaining and microscopy. Indirect immunofluorescence (IF) staining of polytene chromosomes, whole-mount salivary glands and hemocytes was carried out as described previously (Andreyeva et al., 2017; Tracy, Krämer, 2017; Meyer-Nava et al., 2021). The following primary antibodies were used: rabbit anti-HP1 (1:100, kindly provided by Prof. Peter Verrijzer), mouse anti-NON3 (1:50 (Andreyeva et al., 2019)), mouse anti-H3K9me2 (1:100, Abcam 1220), mouse anti-Fibrillarin antibody (38F3; 1:100, Thermo MA1-22000), rabbit anti-CID (1:200, Abcam 10887). The primary antibodies were detected with goat anti-rabbit IgG (H+L) highly cross-adsorbed secondary antibody, Alexa Fluor™ 568 (1:500, Thermo Scientific A-11036), and goat anti-mouse IgG (H+L) cross-adsorbed secondary antibody, Alexa Fluor™ 488 (1:500, Thermo Scientific A-11001). Samples were imaged using a Zeiss Axio Imager M2 (Carl Zeiss) and a confocal microscope

Table 1. List of the used fly stocks

Genotype	Source	Stock number	Description	Reference
$y^1 w^{67c23}$	BDSC	6599		
y^1 *w**; M{w^{+mC}=hs.min(FRT.STOP1)dam}ZH-51C. Hereafter, STOP#1-Dam		65433	Expresses Myc-tagged *E. coli* DNA adenine methyltransferase under the control of a minimal *hsp70* promoter after FLP-mediated excision of the STOP1 transcription termination cassette	Pindyurin et al., 2016
y^1 *w**; M{w^{+mC}=hs.min(FRT.STOP1)dam-HP1} ZH-51C. Hereafter, STOP#1-Dam-HP1		65436	Expresses Myc-tagged *E. coli* DNA adenine methyltransferase fused to HP1 under the control of a minimal *hsp70* promoter after FLP-mediated excision of the STOP1 transcription termination cassette	Pindyurin et al., 2018
y w; M{Dam[intein@L127C]-LAM}ZH51C-M2. Hereafter, Dam(intein)-LAM		65430	Expresses Myc-tagged 4-HT-intein-containing *E. coli* DNA adenine methyltransferase fused to LAM under the control of the full-length *hsp70* promoter	Pindyurin et al., 2016
w^{1118}; P{w^{+mC}= EP} $Non3^{G4706}$/TM6C, Sb^1. Hereafter, $Non3^{G4706}$		30094	Strong hypomorphic mutation of the *Non3* gene caused by the insertion of *P*-element-based transgene	Andreyeva et al., 2019
$In(1)w^{m4h}$, y^1 ac^1. Hereafter, $In(1)w^{m4h}$		76618	Chromosomal rearrangement associated with variegation of the $white^+$ gene expression	
Oregon $R^{modENCODE}$. Hereafter, *Oregon R*		25211	The reference wild-type strain	
y^1 *w**; P{$y^{+t7.7}$ w^{+mC} =iav-QF.P}attP154		36347	attP docking sites located at 97D2 on chromosome 3R (between the *CG14247* and *Tl* genes)	Markstein et al., 2008
$Diap1^{th\text{-}1}$ st^1 cp^1 in^1 $kni^{ri\text{-}1}$ p^p		620	Carries a set of recessive markers on chromosome 3	
ru^1 hry^1 $Diap1^{th\text{-}1}$ st^1 cu^1	Laboratory stock collection	306	Carries a set of recessive markers on chromosome 3	Lindsley, Zimm, 1992
$Non3^{ex}$		–	Precise excision of *P*{EP} transposon from the $Non3^{G4706}$ allele. Used as a control	Andreyeva et al., 2019
$Non3^{259}$		–	Strong hypomorphic mutation of the *Non3* gene carrying remnants of *P*{EP} transposon ends	
$Non3^{\Delta 600}$		–	Partial deletion of the *Non3* gene coding region, null allele	
P[*rescue*]		–	2.76-kb genomic DNA fragment carrying a full-length copy of the wild-type *Non3* allele	
$Su(var)205^5$	Laboratory of Prof. Gunter Reuter	–	Loss of function mutation of the gene encoding HP1	Westphal, Reuter, 2002
$Su(var)3\text{-}9^6$		–	6-kb insertion in the gene encoding the Su(var)3-9 histone methyltransferase	Schotta et al., 2002; Westphal, Reuter, 2002

Note. BDSC – Bloomington Drosophila Stock Center (Bloomington, IN, USA; flystocks.bio.indiana.edu). All *Non3* mutations were balanced with the T(2;3)TSTL, *CyO*:TM6B, Tb^1.

LSM 710 (Carl Zeiss). Optical sections were combined using the LSM Image Browser version 4.2 software (Carl Zeiss).

Image analysis. To measure relative fluorescence intensity of the HP1 protein and H3K9me2 histone modification on polytene chromosomes and whole-mount salivary glands, the fluorescent signals recorded separately as grayscale digital images were pseudocolor-coded and merged using the ImageJ program. In the case of whole-mount salivary glands, we analyzed only corpuscular cells. Hemocyte image analysis was done using Zeiss LSM Image Browser 4.2.0.121 software. Centromere clusters were defined by anti-CID antibodies with an individual center of gravity. Briefly, a sum projection over 4 optical z-sections (0.35 μm each) was created for each centromere foci and nucleolus centered around the brightest pixel of the structure. Distances of centromeres to the nearest nucleolus were measured by drawing a line from the center of the centromere to the edge of the nucleolus. Areas of the nucleolus, centromeres and nuclei were measured by outlin-

ing their boundaries and calculating the areas of the resulting polygons. 35 hemocytes of *Oregon R*, 39 hemocytes of $Non3^{\Delta 600}/Non3^{259}$ and 39 hemocytes of $Non3^{\Delta 600}/Non3^{G4706}$ third-instar larvae were analyzed.

DamID-seq procedure. Fly genotypes used for DamID experiments are listed below in the relevant section. Each experiment was performed in two technical replicates with 60 salivary glands in each replicate dissected from third-instar larvae. Isolation of genomic DNA from the collected material and the entire DamID procedure were performed as previously described (Pindyurin, 2017). To remove DamID adapters from the PCR-amplified Dam-methylated DNA fragments, the latter were digested with DpnII restriction enzyme. After that, the size of the DNA fragments was reduced to a range of 150–450 bp by ultrasonic fragmentation using the Bioruptor Pico Sonication system (Diagenode). Libraries for NGS were prepared using the TruSeq protocol (Illumina).

Illumina NGS and data analysis. Sequencing of the samples was carried out on the Illumina MiSeq 2 × 75 bp platform using the MiSeq Reagent Kit v3 150 cycles (Illumina). The obtained fastq files contained ~1–2 million reads for each sample. The quality analysis of the raw data was performed using the FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequent bioinformatic analysis of DamID-seq data was done as described earlier with minor modifications (Pindyurin et al., 2018). Briefly, sequencing reads from two technical replicates of Dam or Dam-HP1 samples were adapter clipped and uniquely mapped to the dm6 genomic assembly by "bowtie2" (Langmead et al., 2009). Reads were counted by "HTSeqcount" software (Anders et al., 2015) in GATC genomic fragments. Next, read counts were merged between the replicates, as they were highly correlated (the Pearson correlation coefficient = 0.93–0.97). The resulting read counts of Dam or Dam-HP1 samples were converted to reads per million (RPM), and then Dam-HP1 values were normalized to those of the Dam and $\log_2$ transformed. Finally, quantile normalization was applied.

Results

We investigated the role of NON3 in nucleolar morphology, heterochromatin organization, and centromere localization. First, we examined whether *Non3* mutants could modify PEV. In *Drosophila*, PEV assay has been extensively employed to study heterochromatin formation (Elgin, Reuter, 2013). We used $In(1)w^{m4}$ inversion, in which the normally euchromatic $white^{+}$ gene responsible for eye pigmentation is placed close to the pericentric heterochromatin due to chromosomal inversion and becomes silent in some cells (Cooper, 1959). To test whether *Non3* mutations modify PEV, we genetically combined inversion $In(1)w^{m4h}$ with the following *Non3* alleles: $Non3^{ex}$ (control), $Non3^{259}$ (strong hypomorphic mutation), and $Non3^{\Delta 600}$ (null allele of the *Non3* gene) (Andreyeva et al., 2019). The $Su(var)205^{5}$ and $Su(var)3\text{-}9^{6}$ mutations, known PEV suppressors (Eissenberg et al., 1992; Schotta et al., 2002), were used as references. The *yw*/*yw*; $Non3^{ex}$/+ and *Oregon R* flies were used as negative and positive controls for absence/presence of eye pigment, respectively (Fig. 1*a*). PEV can be modified by a variety of factors. The temperature of development and the amount of heterochromatin within the genome were the first factors shown to affect the extent of variegation (Elgin, Reuter, 2013). We carried out experiments separately for males and females at 18 °C. The visual inspection of fly eyes showed that both the $Non3^{259}$ and the $Non3^{\Delta 600}$ mutations combined with $In(1)w^{m4h}$ lead to an increase in $white^{+}$ expression compared to control (Fig. 1*a*).

To quantitatively measure the effects, we performed extraction and measurement of the eye pigment. For example, the $Non3^{259}$ mutation ($In(1)w^{m4h}/yw$; $Non3^{259}$/+) resulted in a 1.32- (for females) and 1.06- (for males) fold increase in the eye pigmentation level compared to control ($In(1)w^{m4h}/yw$; $Non3^{ex}$/+). The $Non3^{\Delta 600}$ mutation ($In(1)w^{m4h}/yw$; $Non3^{\Delta 600}$/+) resulted in a 1.67- (for females) and 1.47- (for males) fold increase in the eye pigmentation level compared to control ($In(1)w^{m4h}/yw$; $Non3^{ex}$/+) (Fig. 1*b*). Taken together, these results demonstrate that *Non3* mutations are suppressors of PEV, although their influence is significantly lower (~4 times for both females and males) than that of the $Su(var)205^{5}$ and $Su(var)3\text{-}9^{6}$ mutations. Generally, the results confirmed the visual observations of fly eyes (Fig. 1*a*).

Next, we decided to perform meiotic recombination analysis within the euchromatin and pericentromeric regions in *Non3* mutants. Normally, recombination in pericentromeric heterochromatin is almost absent and strongly suppressed in adjacent euchromatic regions (Baker, 1958; Westphal, Reuter, 2002). However, the dominant effects of suppressors of PEV on crossing-over in the pericentromeric regions were shown for some mutations. For example, for double mutants $Su(var)205^{5}$ and $Su(var)3\text{-}9^{6}$, the meiotic recombination frequency in the pericentromeric regions between the marker genes *kni* and *p* was increased (Westphal, Reuter, 2002). We used two different strains (##306, 620) carrying viable recessive genetic markers on chromosome 3 (Fig. 2*a*) and measured the crossing-over frequencies between them in a *Non3* mutant background (Fig. 2*b*, Supplementary Tables S1, S2)[1]. In Figure 2, we presented crossings for strain #306 but not for strain #620, since it was similar to #306. We analyzed 2,380 crossover flies for strain #306 and 3,079 flies for strain #620 (Table 2, Supplementary Tables S1, S2). For strain #306, we showed that the presence of one copy of the $Non3^{\Delta 600}$ allele in the genome leads to a statistically significant 1.24- and 1.71-fold increase in recombination frequency in the euchromatin region between the marker genes *ru* and *hry* and in the pericentromeric regions between the marker genes *st* and *cu*, respectively. For strain #620, we observed a statistically significant 1.95-fold increase in recombination frequency in the pericentromeric region between the marker genes *kni* and *p* (Fig. 2*c*, Table 2). The results suggest a possible role of NON3 in maintenance of integrity and stability of the euchromatin and pericentromeric regions.

We sought to understand whether NON3 is required for chromosomal localization of the heterochromatin components. For that, the *Non3* mutant and wild-type *Oregon R* squash preparations of polythene chromosomes were immunostained with anti-HP1 and anti-H3K9me2 antibodies. No difference in the protein binding patterns at the chromocenter between the mutant and the control background was found, but the intensity of the signals was reduced by 27.6 and 23.0 % for HP1 and H3K9me2, respectively, in *Non3* mutants ($N = 41$) compared to control ($N = 81$) (Fig. 3*a*, *b*). However, when whole-mount

[1] Supplementary Tables S1, S2 and Figures S1–S3 are available at: https://vavilovj-icg.ru/download/pict-2025-29/appx14.pdf

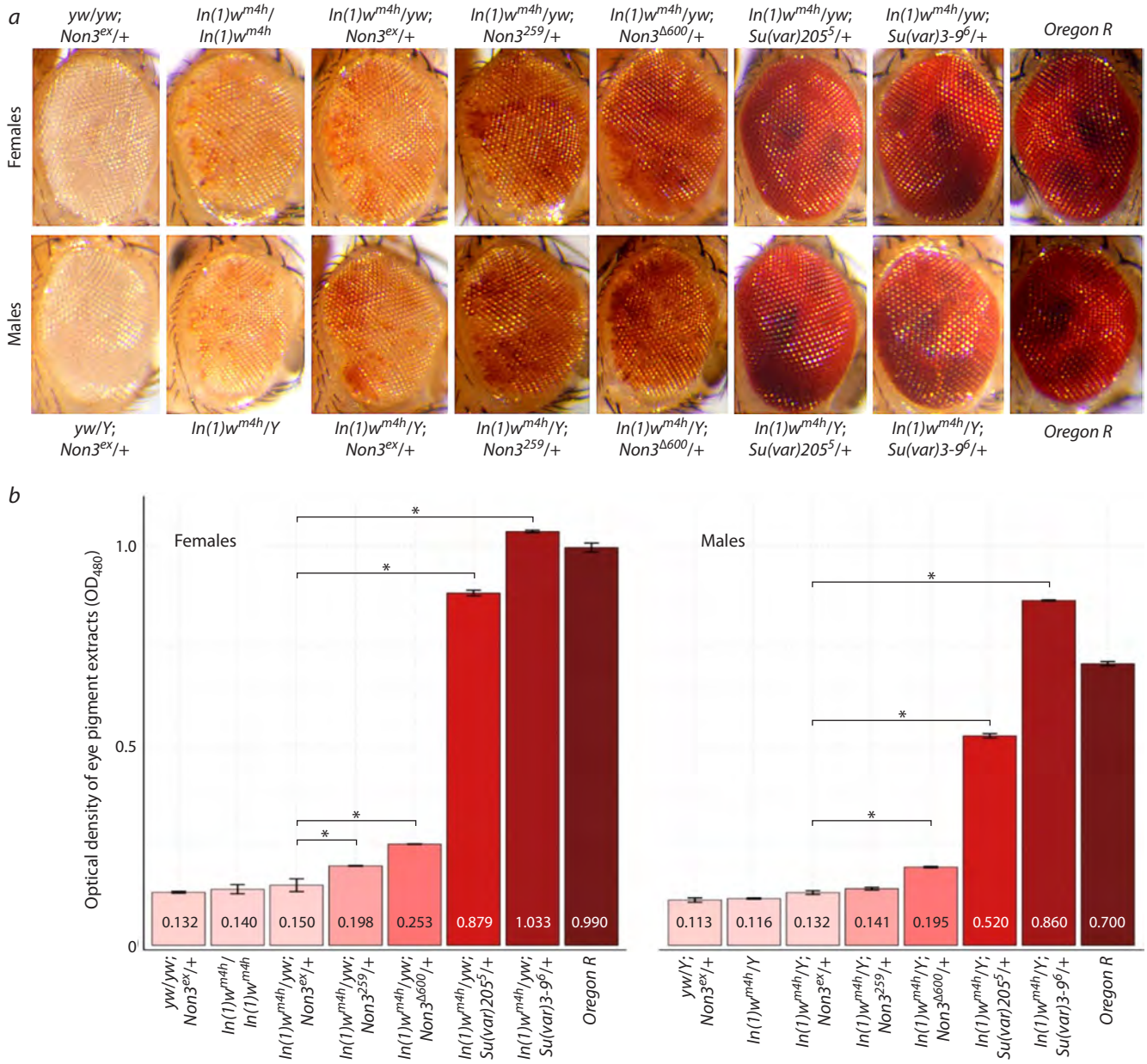


Fig. 1. Suppression of PEV by the *Non3* mutations at 18 °C.

a – Eye pigmentation of control and mutant flies. The most characteristic images for each of the indicated genotypes are provided. *b* – Quantification of PEV phenotype of adult flies based on the concentration of red eye pigment. Data are graphically represented as a histogram for three measurement points for each genotype. The *y*-axis reflects the optical density (OD_{480}) of the eye pigment extracts from the flies of the indicated genotypes. Numbers inside the columns indicate the average pigment optical density for each genotype. * Significance level $p < 0.05$, pairwise *t*-test.

salivary glands were immunostained with anti-HP1 antibodies, we even found an increase in the HP1 intensity of 36.0 % in *Non3* mutants ($N = 82$) in comparison with control ($N = 75$) (Fig. 3*c*, *d*). The data of whole-mount salivary glands' immunostaining was in accordance with Western blotting analysis of total protein levels from larval salivary glands and brains with adjacent imaginal discs, which showed that the total levels of the HP1 protein (Fig. 4*a*) and H3K9me2 histone modification (Fig. 4*b*) were slightly increased in $Non3^{\Delta 600}/Non3^{259}$ mutants compared to control.

Next, we investigated the distribution of the HP1 protein in the salivary gland polytene chromosomes of *Non3* mutants using the DamID approach. To generate DamID profiles of HP1, we used the FLP-inducible STOP#1-Dam system (Pindyurin et al., 2016). Expression of Dam only or Dam-HP1 construct in larval salivary glands was activated by FLP recombinase expressed under the control of the *sgs3* promoter, which is specifically active in this tissue (Biyasheva et al., 2001; McPherson et al., 2024; Suárez Freire et al., 2024). To achieve that, the *sgs3*-FLP transgene was integrated at the 97D2 region and its activity was indeed detected in larval salivary glands but not in whole adult flies (Supplementary Fig. S1). Next, we combined the *sgs3*-FLP transgene with the $Non3^{259}$ mutation on the same chromosome. Then, we generated larvae of the following genotypes: STOP#1-Dam (-HP1)/+; *sgs3*-FLP, $Non3^{259}/Non3^{\Delta 600}$ or STOP#1-Dam

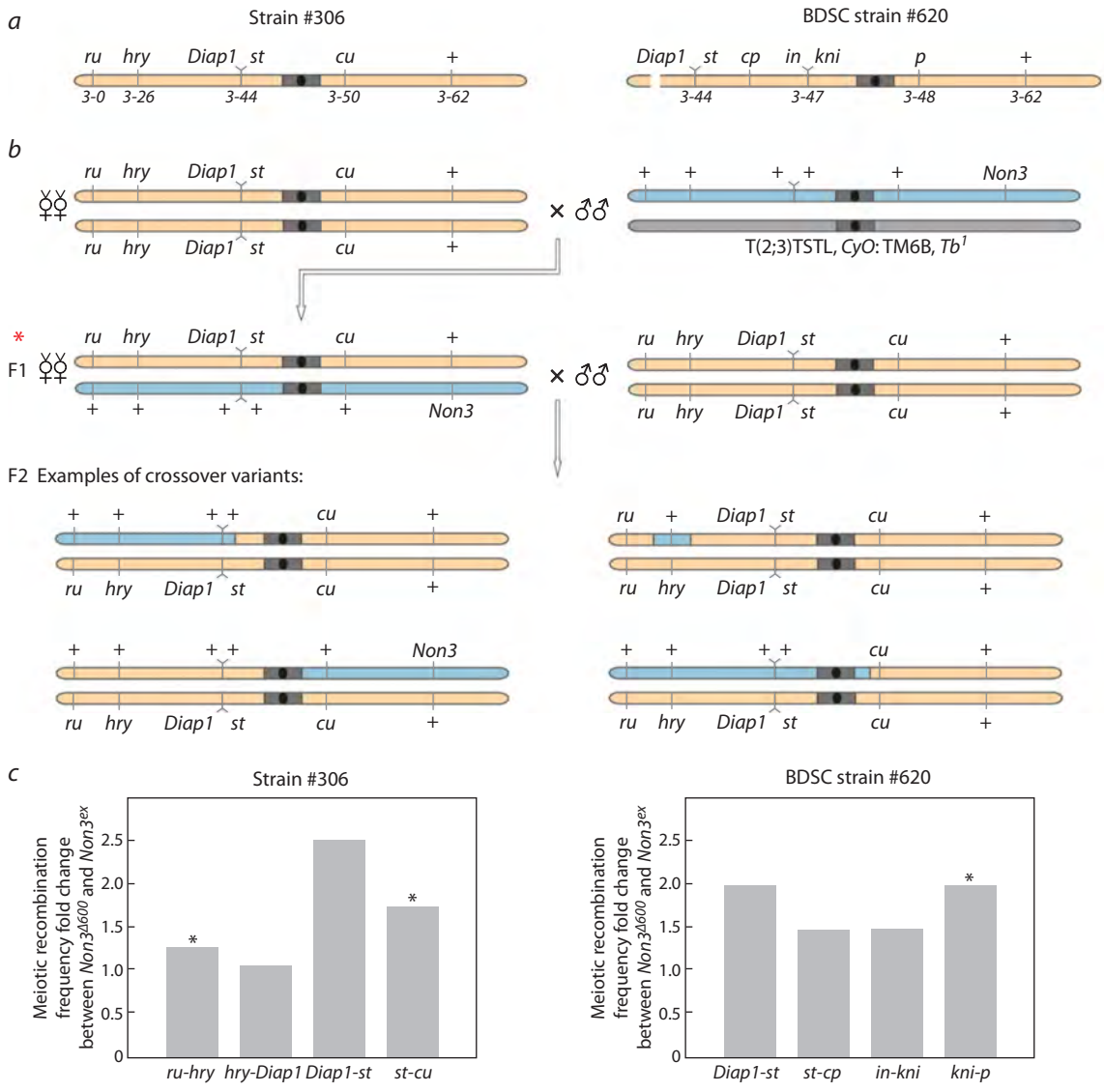


Fig. 2. Mutations in the *Non3* gene increase the meiotic recombination frequency.

a – Schematic presentation of strains ##306 and 620 carrying a set of recessive mutations in chromosome 3 and used for analysis of meiotic recombination frequencies. The numbers indicate the localization of recessive mutations on the genetic map according to (Gramates et al., 2017). Pericentromeric heterochromatin and the centromere are shown by a gray rectangle and a black circle, respectively. *b* – Experimental setup. The #306 strain is shown as an example. The initial cross involves the homozygous strain #306 and the strain carrying any of the *Non3* alleles ($Non3^{ex}$ or $Non3^{\Delta 600}$). Virgin F1 females (red asterisk), in the gonads of which crossing-over takes place, are crossed with males of the strain #306. Several examples of F2 crossovers are shown. *c* – Effect of the heterozygous $Non3^{\Delta 600}$ mutation on the frequency of meiotic recombination at the studied chromosomal regions as compared to the $Non3^{ex}$ control. * Significance level $p < 0.01$ by Fisher's exact test.

Table 2. The number of F2 crossover flies analyzed for meiotic recombination frequency

Crossovers	Number of flies		Fisher's exact test *p*-value	Crossovers	Number of flies		Fisher's exact test *p*-value
	$Non3^{ex}$	$Non3^{\Delta 600}$			$Non3^{ex}$	$Non3^{\Delta 600}$	
Strain #306				BDSC strain #620			
ru–hry	299	371	0.001	*Diap1–st*	6	8	0.276
hry–Diap1	90	94	0.818	*st–cp*	14	14	0.338
Diap1–st	4	10	0.178	*in–kni*	2	2	1.000
st–cu	42	72	0.005	*kni–p*	27	36	0.009
Total	1,190	1,190		Total	1,827	1,252	

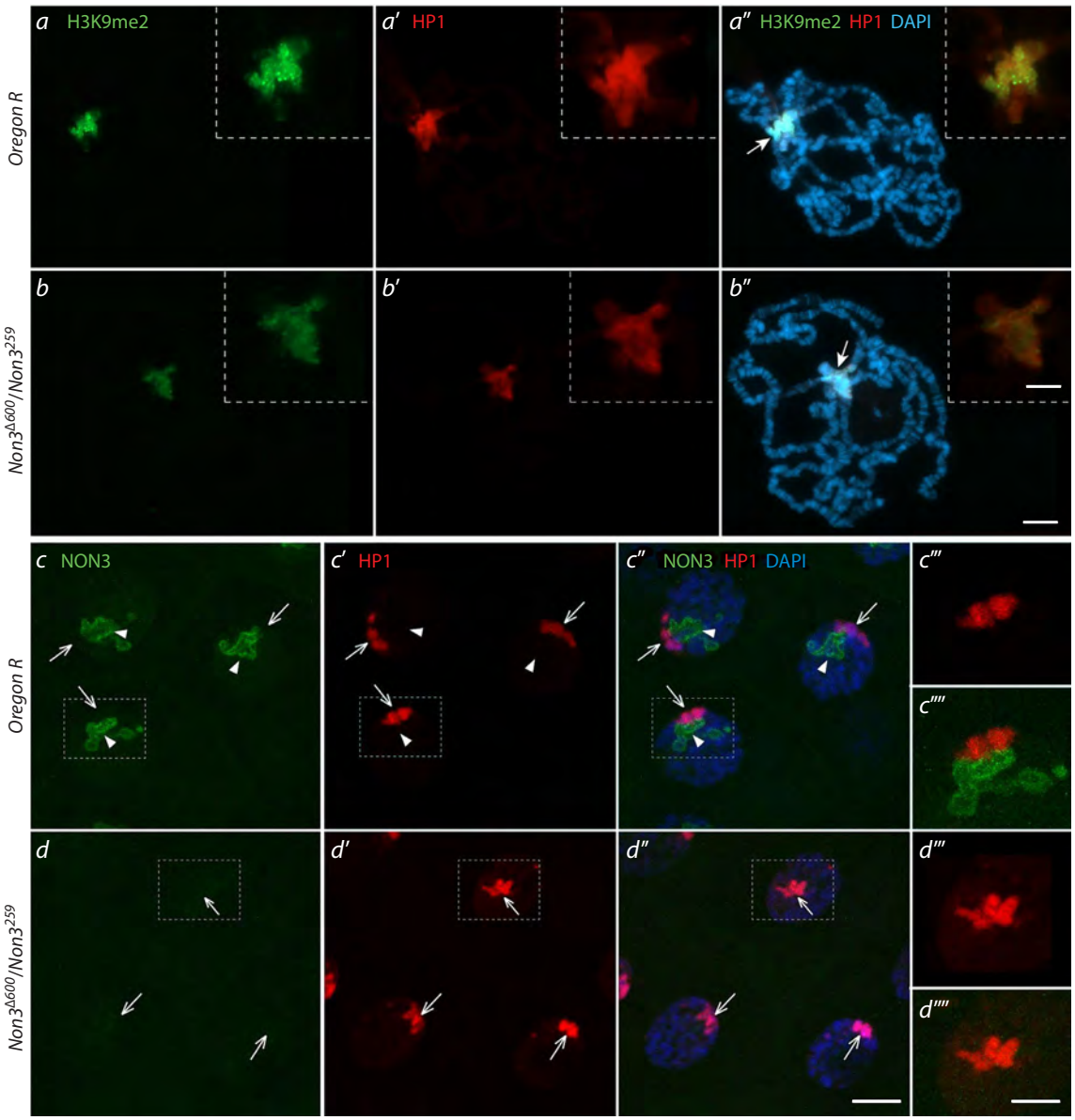


Fig. 3. The levels of the HP1 protein and H3K9me2 histone modification in larval salivary glands of *Non3* mutants.

a and *b* – Immunofluorescence images of polytene chromosome staining from wild-type *Oregon R* (*a*–*a*″) and $Non3^{\Delta 600}/Non3^{259}$ mutant (*b*–*b*″) larvae co-stained with anti-HP1 and anti-H3K9me2 antibodies. The intensities of HP1 and H3K9me2 signals are slightly reduced in mutants compared to control. Arrows with a filled arrowhead indicate the chromocenter. Images in dotted frames represent magnification of the chromocenter. *c* and *d* – Confocal microscopy (maximum projection) of salivary gland nuclei from wild-type *Oregon R* (*c*–*c*″) and $Non3^{\Delta 600}/Non3^{259}$ mutant (*d*–*d*″) larvae co-stained with anti-HP1 and anti-NON3 antibodies. There is no detectable NON3 signal in $Non3^{\Delta 600}/Non3^{259}$ mutants compared to control, but some increase in the fluorescence level of the HP1 protein in *Non3* mutants *vs* control is detected. Arrows indicate the HP1 signal, arrowheads show nucleoli visualized by anti-NON3 antibodies. Fragments marked by a dotted rectangle are shown with larger magnification in (*c*‴, *c*⁗, *d*‴, *d*⁗). DNA is visualized by DAPI (blue), nucleolus by anti-NON3 antibodies (green), H3K9me2 histone modification, by the corresponding antibodies (green), the HP1 protein by the corresponding antibodies (red). Scale bar for all images except magnification is 20 μm, for enlarged fragments of microphotographs, 10 μm.

(-HP1)/+; *sgs3*-FLP/+ or STOP#1-Dam (-HP1)/*P*[*rescue*]; *sgs3*-FLP, $Non3^{259}/Non3^{\Delta 600}$ (Fig. 5*a*). Subsequent amplification of Dam-methylated fragments of the salivary gland genome was performed as previously described (Pindyurin et al., 2017). The high specificity of the amplification procedure was confirmed by gel electrophoresis showing substantially more mePCR products in experimental samples compared to negative controls (Fig. 5*b*). DamID-derived libraries were subjected to Illumina sequencing and analyzed as described previously (Pindyurin et al., 2018). The correlation between control, *Non3* mutants carrying one copy of the rescue construct (*P*[*rescue*]) and *Non3* mutant-only samples across the entire genome showed no significant difference (Fig. 5*c*). All three samples demonstrated the increased binding of HP1 protein to chromosome X in comparison with autosomes (Fig. 5*d*), which is in good agreement with previous reports for larval brains, neurons, glia, fat body and Kc167 cells (Pindyurin et al., 2018). Comparison of DamID profiles obtained for all three samples did not reveal substantial changes either at the pericentromeric regions or at any other parts of

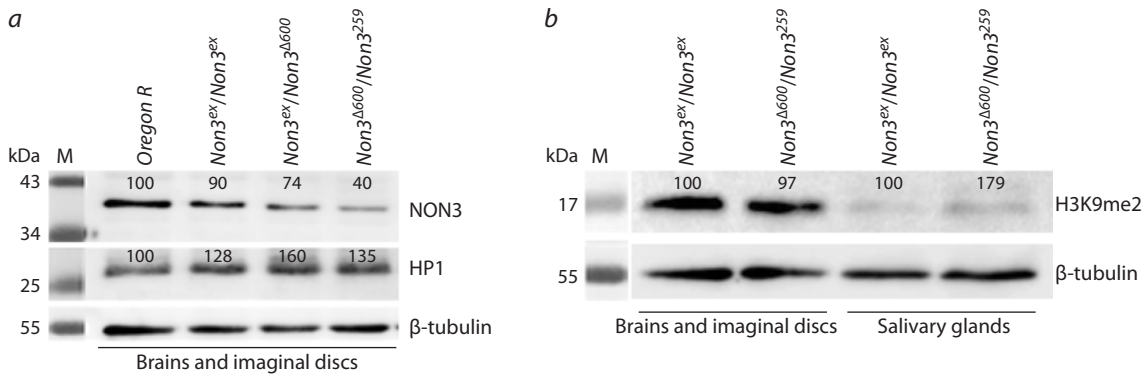


Fig. 4. The levels of the HP1 and NON3 proteins, as well as H3K9me2 histone modification in larval tissues from *Non3* mutants.

a – Western blot from larval brains with adjacent imaginal discs showing that in $Non3^{\Delta 600}/Non3^{259}$ mutants, the level of the NON3 protein is substantially reduced compared to the *Oregon R* and $Non3^{ex}/Non3^{ex}$ controls. *b* – Western blot from larval brains with adjacent imaginal discs and salivary glands. H3K9me2 histone modification is not reduced in both tissue types in $Non3^{\Delta 600}/Non3^{259}$ mutants compared to the $Non3^{ex}/Non3^{ex}$ controls. The numbers show the intensity of each band normalized to the intensities of the corresponding loading control taken as 100 %. M – Prestained Protein Ladder. The protein level of β-tubulin is shown as a loading control.

the chromosomes (Fig. 5*e*, Supplementary Figs. S2, S3), suggesting that NON3 is not essential for the binding of HP1 to chromatin.

To understand if NON3 has a role in the tethering and clustering of centromeres, we isolated hemocytes from *Drosophila* third-instar larvae and analyzed the number of centromeres per cell and their localization relatively to the nucleolus. *Drosophila* diploid larval hemocytes possess 8 centromeres revealed as 2–3 individual centromere foci (Padeken et al., 2013), which have been described to cluster together and associate with the periphery of the nucleolus (Fig. 6*a*) (Padeken, Heun, 2013; Padeken et al., 2013). We observed that in $Non3^{\Delta 600}/Non3^{259}$ and $Non3^{\Delta 600}/Non3^{G4706}$ mutant interphase hemocytes, the size of the nucleolus was increased (Fig. 6*a*, *b*), while the size of the nucleus was not changed. In wild-type hemocytes, the mean size of the nucleolus was 1.56 ± 0.10 μm^2, in $Non3^{\Delta 600}/Non3^{259}$, 5.82 ± 0.32 μm^2, and in $Non3^{\Delta 600}/Non3^{G4706}$, 3.22 ± 0.19 μm^2. In mutant hemocytes, we did not detect any increase in individual centromere foci per cell or any significant dissociation of centromeres from the nucleolar periphery, but observed an increase in the size of the regions detected by anti-centromere antibodies (Fig. 6*b*). In wild-type hemocytes, the mean size of the analyzed regions was 0.07 ± 0.01 μm^2, while in $Non3^{\Delta 600}/Non3^{259}$ mutants, it was 0.140 ± 0.02 μm^2, and in $Non3^{\Delta 600}/Non3^{G4706}$ mutants, 0.216 ± 0.02 μm^2.

Discussion

Increasing evidence suggests that the function of the nucleolus goes beyond ribosome biogenesis. Among many other functions, the nucleolus is considered as the hub for the organization of inactive chromatin in the cell (Quinodoz et al., 2018). In this study, we aimed to understand the role of *Drosophila Non3* mutations in chromatin organization. We have found that *Non3* is a weak suppressor of PEV ($In(1)w^{m4h}$) implicating NON3 directly or indirectly in the maintenance of normal chromatin structure during eye development. The involvement of nucleolar proteins in chromatin compaction is not surprising. Previously, it was shown that mutants of *modulo* display a suppressor effect on PEV. The Modulo protein binds DNA directly and may serve to anchor multimeric complexes, promoting chromatin compaction and silencing (Garzino et al., 1992). NON3 does not have a predicted DNA-binding domain (https://www.uniprot.org/uniprotkb/Q9VEB3/entry); therefore, we suggest that it might be associated with some unknown factors to form condensed chromatin.

In addition to the effect on PEV, we have shown that the frequency of meiotic recombination is increased when one copy of *Non3* is missing from the genome. In 2002, T. Westphal and G. Reuter conducted a large-scale screen to assess the effects of PEV suppressor genes on crossing-over frequency in the pericentromeric regions of chromosomes. It was shown that 16 mutations in the *Su(var)* genes have a significant effect on increasing the frequency of meiotic recombination in the region of chromosome 3 between the *kni–p* markers. Heterozygous combination of the $Su(var)205^5/+$; $Su(var)3\text{-}9^6/+$ mutations increased recombination at this chromosomal region by 1.4 times compared to control (Westphal, Reuter, 2002). These results confirmed that the frequency of crossover events can be controlled at the level of chromatin structure. We showed that the deletion of one functional copy of *Non3* in the genome increased the frequency of meiotic recombination in the *kni–p* and *st–cu* regions compared to the control $Non3^{ex}$ allele. Moreover, an increased frequency of meiotic recombination was also observed in the euchromatic *ru–hry* region of chromosome 3 located far away from the centromere (Table 2). Altogether, our results indicate that a single functioning copy of *Non3* is not enough for maintenance of normal chromatin structure; it is especially evident in the pericentromeric regions of chromosome 3.

Since *Non3* mutants suppress PEV and enhance meiotic recombination in the pericentromeric regions of chromosome 3, we wondered whether the localization pattern of the main heterochromatin components, HP1 and H3K9me2, is somehow affected in *Non3* mutants. No substantial changes were found in their localization patterns on salivary gland polytene chromosomes, but we noticed a slight decrease in HP1 and H3K9me2 signals in a *Non3* mutant background using acetic acid fixation (Fig. 3*a*, *b*). However, immunostaining with formaldehyde fixation and/or immunoblotting

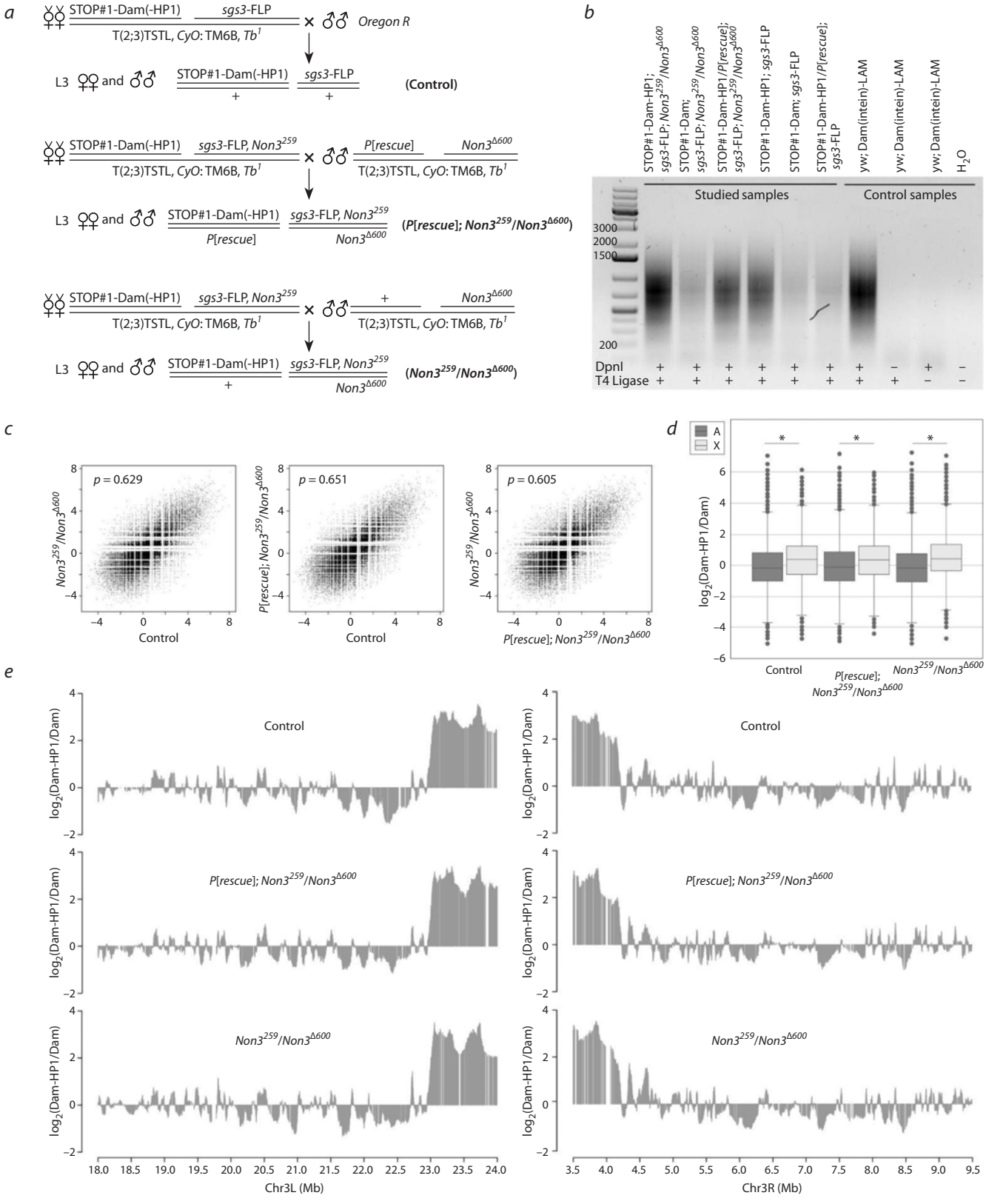


Fig. 5. FLP-inducible *Drosophila* DamID system in salivary glands of the *Non3* mutant and control third-instar larvae.

a – Genetic crosses used to activate STOP#1-Dam(-HP1)-containing transgenes. *b* – Methylation detected in genomic DNA isolated from larval salivary glands. Specificity of amplification of the methylated GATC fragments was confirmed by the '-DpnI' and '-T4 DNA ligase' control reactions. Dam(intein)-LAM flies were used as a positive control; the banded pattern is derived from mitochondrial DNA. *c* – Genome-wide correlation between the studied datasets (*p* – Pearson's correlation coefficient). *d* – Box plots showing distributions of $\log_2$(Dam-HP1/Dam) values in the non-repetitive parts of the X chromosome (light gray) and autosomes (dark gray) in the studied samples. Wilcoxon rank sum test was used for pairwise comparison of distributions on the X chromosome *vs* autosomes, * *p*-value < $2.2 \cdot 10^{-16}$. *e* – Mutation in the *Non3* gene does not substantially affect the HP1 binding profile. Representative 6.0-Mb fragments of chromosomal arms 3L and 3R are shown. A running mean algorithm (a sliding window of 50 GATC fragments, one fragment per step) was applied to the HP1 binding data.

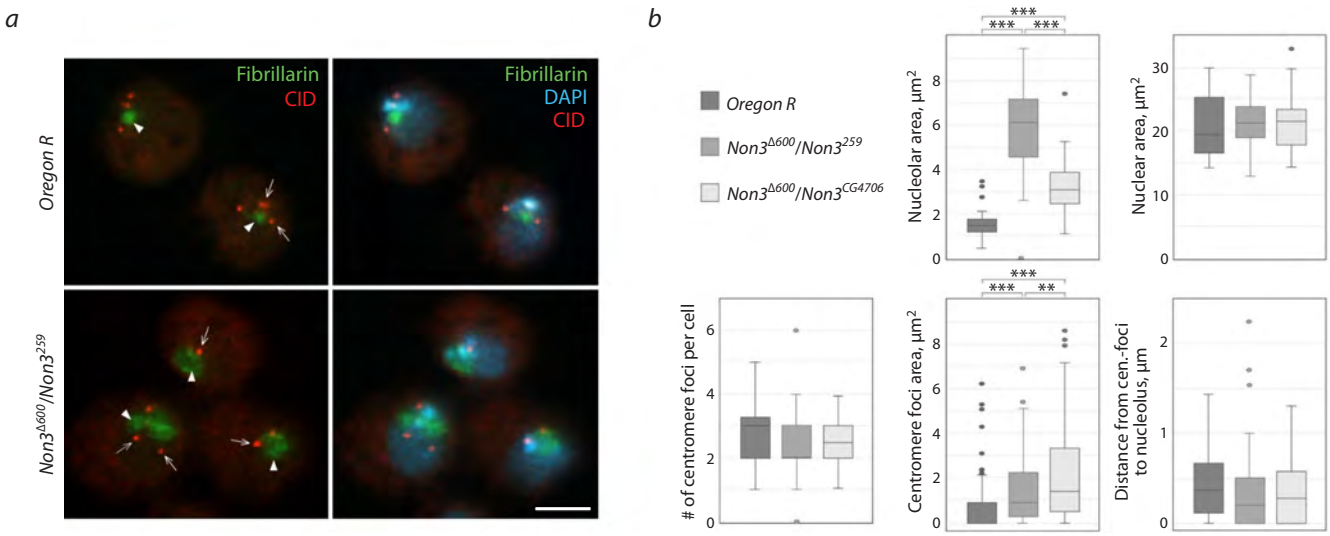


Fig. 6. The lack of the NON3 protein in *Non3* mutants leads to an increase in the size of the nucleolus and centromeres.

a – Immunofluorescence images of fixed *Drosophila* hemocytes during interphase, isolated from *Oregon R* (control) and $Non3^{\Delta 600}/Non3^{259}$ mutants, show the relative position of the centromere (CID) and nucleolus (Fibrillarin) in cells. Arrows indicate the centromeres, arrowheads, the nucleolus. *b* – Quantification of the nucleolar area, number of centromere foci per cell, distance from centromere foci to the nucleolus, the centromere foci area and nuclear area in hemocytes of *Oregon R*, $Non3^{\Delta 600}/Non3^{259}$ and $Non3^{\Delta 600}/Non3^{CG4706}$ third-instar larvae. ** Significance level $p < 0.01$, *** $p < 0.001$, Student's *t*-test. Nuclear DNA is visualized by DAPI (blue), nucleolus by anti-Fibrillarin antibodies (green), and centromere by anti-CID (red). Scale bar for all images: 5 µm.

of whole salivary glands did not detect a decrease in amounts of these proteins within the cells (Fig. 3*c*, *d*, Fig. 4). The difference in HP1 signal intensities seen using various fixation methods can be explained by the fact that acetic acid is prone to extract histones from the tissues (Dick, Johns, 1968; Johansen et al., 2009). Since NON3 seems to play some role in maintenance of integrity and stability of the pericentromeric regions, acetic acid fixation may extract histones more extensively in *Non3* mutant tissues. This may be the reason for the observed differences.

We also examined HP1 binding in the polytene chromosomes of *Non3* mutants with higher resolution using the DamID approach, but did not detect substantial differences compared to control animals either in the pericentromeric regions of chromosome 3 or somewhere else in the genome (Fig. 5*e*, Supplementary Figs. S2, S3). Thus, the effects on PEV and meiotic recombination in pericentromeric chromosomal regions observed in *Non3* mutants do not seem to be due to changes in HP1 localization pattern.

Since there is a connection between clustering and positioning of centromeres near the nucleolus and stable organization of pericentric heterochromatin (Padeken et al., 2013), we analyzed the number and localization of centromeres using *Drosophila* third-instar larval hemocytes. We found that in *Non3* mutants, the number of centromeres (CID foci) in interphase cells is not different from control. We also did not detect any untethering centromeres from the periphery of the nucleolus in *Non3* mutants. However, we observed an increase in the size of the nucleolus and centromeres in $Non3^{\Delta 600}/Non3^{259}$ and $Non3^{\Delta 600}/Non3^{G4706}$ mutant hemocytes (Fig. 6). Previously, it was shown that reduction of the NON3 orthologue protein ARPF2 leads to the redistribution of Fibrillarin and Nucleolin from the nucleolus to the nucleoplasm (Choi et al., 2020). In another study, FCs and DFCs were delocalized to the periphery of the GC upon a significant decrease in the levels of Nucleolin protein in HeLa cells (Ugrinova et al., 2007). In the case of *Non3* mutants, we cannot exclude that there is no enlargement of the nucleolus but Fibrillarin has moved beyond the DFC compartment. It is interesting to note that the disappearance of Fibrillarin was earlier observed in *Non3* mutant larval brain cells (Andreyeva et al., 2019). Such a discrepancy with the findings of the present study might be caused either by cell type-specific peculiarities or differences in the fixation methods used. We also observed an increase in centromere size in *Non3* mutants (Fig. 6*b*). Centromeres are generally flanked by heterochromatin (Kapoor et al., 2015) and it was shown earlier that flanking heterochromatin is a prerequisite for maintaining centromeres (Henikoff et al., 2001). Therefore, we suggest that the increase in centromere size in *Non3* mutants may be associated with the role of NON3 in maintenance of integrity and stability of the pericentromeric regions of chromosomes.

Conclusions

Thus, we analyzed the effects of *Non3* mutants on chromatin organization in the nuclei of various *Drosophila* tissues. We have shown that *Non3* mutants suppress PEV, enhance meiotic recombination in the euchromatin and pericentromeric regions of chromosome 3, however, this does not accompanied by any significant changes in the amount or distribution of classical heterochromatin markers: the HP1 protein as well as the modification of the histone H3K9me2. In *Non3* mutants, we observed an increased size of both, the nucleoli and the region detected by anti-centromere antibodies. However, we did not detect centromere declustering or their detachment from the nucleolar periphery. Thus, we suggest that the NON3 protein is important for the formation/function of the nucleolus and is required for the correct chromatin packaging, but the exact mechanism of NON3 involvement in these processes requires further study.

References

Allshire R.C., Madhani H.D. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol*. 2018;19(4):229-244. doi 10.1038/nrm.2017.119

Anders S., Pyl P.T., Huber W. HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi 10.1093/bioinformatics/btu638

Andreyeva E.N., Bernardo T.J., Kolesnikova T.D., Lu X., Yarinich L.A., Bartholdy B.A., Guo X., Posukh O.V., Healton S., Willcockson M.A., Pindyurin A.V., Zhimulev I.F., Skoultchi A.I., Fyodorov D.V. Regulatory functions and chromatin loading dynamics of linker histone H1 during endoreplication in *Drosophila*. *Genes Dev.* 2017;31(6):603-616. doi 10.1101/gad.295717.116

Andreyeva E.N., Ogienko A.A., Yushkova A.A., Popova J.V., Pavlova G.A., Kozhevnikova E.N., Ivankin A.V., Gatti M., Pindyurin A.V. *Non3* is an essential *Drosophila* gene required for proper nucleolus assembly. *Vavilov J Genet Breed*. 2019;23(2):190-198. doi 10.18699/VJ19.481

Baker W.K. Crossing over in heterochromatin. *Am Nat.* 1958;92(862):59-60. doi 10.1086/282010

Becker J.S., McCarthy R.L., Sidoli S., Donahue G., Kaeding K.E., He Z., Lin S., Garcia B.A., Zaret K.S. Genomic and proteomic resolution of heterochromatin and its restriction of alternate fate genes. *Mol Cell*. 2017;68(6):1023-1037.e1015. doi 10.1016/j.molcel.2017.11.030

Bersaglieri C., Santoro R. Genome organization in and around the nucleolus. *Cells*. 2019;8(6):579. doi 10.3390/cells8060579

Bischof J., Maeda R.K., Hediger M., Karch F., Basler K. An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc Natl Acad Sci USA*. 2007;104(9):3312-3317. doi 10.1073/pnas.0611511104

Biyasheva A., Do T.V., Lu Y., Vaskova M., Andres A.J. Glue secretion in the *Drosophila* salivary gland: a model for steroid-regulated exocytosis. *Dev Biol*. 2001;231(1):234-251. doi 10.1006/dbio.2000.0126

Bizhanova A., Kaufman P.D. Close to the edge: heterochromatin at the nucleolar and nuclear peripheries. *Biochim Biophys Acta Gene Regul Mech*. 2021;1864(1):194666. doi 10.1016/j.bbagrm.2020.194666

Bloom K.S. Centromeric heterochromatin: the primordial segregation machine. *Annu Rev Genet*. 2014;48:457-484. doi 10.1146/annurev-genet-120213-092033

Boisvert F.M., van Koningsbruggen S., Navascués J., Lamond A.I. The multifunctional nucleolus. *Nat Rev Mol Cell Biol.* 2007;8(7):574-585. doi 10.1038/nrm2184

Boulon S., Westman B.J., Hutten S., Boisvert F.M., Lamond A.I. The nucleolus under stress. *Mol Cell*. 2010;40(2):216-227. doi 10.1016/j.molcel.2010.09.024

Chang C.H., Chavan A., Palladino J., Wei X., Martins N.M.C., Santinello B., Chen C.C., Erceg J., Beliveau B.J., Wu C.T., Larracuente A.M., Mellone B.G. Islands of retroelements are major components of *Drosophila* centromeres. *PLoS Biol*. 2019;17(5):e3000241. doi 10.1371/journal.pbio.3000241

Choi I., Jeon Y., Yoo Y., Cho H.S., Pai H.S. The *in vivo* functions of ARPF2 and ARRS1 in ribosomal RNA processing and ribosome biogenesis in Arabidopsis. *J Exp Bot*. 2020;71(9):2596-2611. doi 10.1093/jxb/eraa019

Connolly K., Burnet B., Sewell D. Selective mating and eye pigmentation: an analysis of the visual component in the courtship behavior of *Drosophila melanogaster*. *Evolution*. 1969;23(4):548-559. doi 10.1111/j.1558-5646.1969.tb03540.x

Cooper K.W. Cytogenetic analysis of major heterochromatic elements (especially Xh and Y) in *Drosophila melanogaster*, and the theory of "heterochromatin". *Chromosoma*. 1959;10:535-588. doi 10.1007/BF00396588

Czermin B., Schotta G., Hülsmann B.B., Brehm A., Becker P.B., Reuter G., Imhof A. Physical and functional association of SU(VAR)3-9 and HDAC1 in *Drosophila*. *EMBO Rep*. 2001;2(10):915-9. doi 10.1093/embo-reports/kve210

Dick C., Johns E.W. The effect of two acetic acid containing fixatives on the histone content of calf thymus deoxyribonucleoprotein and calf thymus tissue. *Exp Cell Res*. 1968;51(2-3):626-632. doi 10.1016/0014-4827(68)90150-x

Eisenhaber F., Wechselberger C., Kreil G. The Brix domain protein family – a key to the ribosomal biogenesis pathway? *Trends Biochem Sci*. 2001;26(6):345-347. doi 10.1016/s0968-0004(01)01851-5

Eissenberg J.C., Morris G.D., Reuter G., Hartnett T. The heterochromatin-associated protein HP-1 is an essential protein in Drosophila with dosage-dependent effects on position-effect variegation. *Genetics*. 1992;131(2):345-352. doi 10.1093/genetics/131.2.345

Elgin S.C., Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb Perspect Biol*. 2013;5(8):a017780. doi 10.1101/cshperspect.a017780

Garzino V., Pereira A., Laurenti P., Graba Y., Levis R.W., Le Parco Y., Pradel J. Cell lineage-specific expression of modulo, a dose-dependent modifier of variegation in Drosophila. *EMBO J*. 1992;11(12):4471-4479. doi 10.1002/j.1460-2075.1992.tb05548.x

Gramates L.S., Marygold S.J., Santos G.D., Urbano J.M., Antonazzo G., Matthews B.B., Rey A.J., Tabone C.J., Crosby M.A., Emmert D.B., Falls K., Goodman J.L., Hu Y., Ponting L., Schroeder A.J., Strelets V.B., Thurmond J., Zhou P.; the FlyBase Consortium. FlyBase at 25: looking to the future. *Nucleic Acids Res.* 2017;45(D1):D663-D671. doi 10.1093/nar/gkw1016

Grewal S.I., Jia S. Heterochromatin revisited. *Nat Rev Genet*. 2007;8(1):35-46. doi 10.1038/nrg2008

Henikoff S., Ahmad K., Malik H.S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*. 2001;293(5532):1098-1102. doi 10.1126/science.1062939

Hernandez-Verdun D., Roussel P., Thiry M., Sirri V., Lafontaine D.L. The nucleolus: structure/function relationship in RNA metabolism. *Wiley Interdiscip Rev RNA*. 2010;1(3):415-431. doi 10.1002/wrna.39

Heun P., Erhardt S., Blower M.D., Weiss S., Skora A.D., Karpen G.H. Mislocalization of the *Drosophila* centromere-specific histone CID promotes formation of functional ectopic kinetochores. *Dev Cell*. 2006;10(3):303-315. doi 10.1016/j.devcel.2006.01.014

Hirano Y., Ishii K., Kumeta M., Furukawa K., Takeyasu K., Horigome T. Proteomic and targeted analytical identification of BXDC1 and EBNA1BP2 as dynamic scaffold proteins in the nucleolus. *Genes Cells*. 2009;14(2):155-166. doi 10.1111/j.1365-2443.2008.01262.x

Hoskins R.A., Carlson J.W., Wan K.H., Park S., Mendez I., Galle S.E., Booth B.W., Pfeiffer B.D., George R.A., Svirskas R., Krzywinski M., Schein J., Accardo M.C., Damia E., Messina G., Méndez-Lago M., de Pablos B., Demakova O.V., Andreyeva E.N., Boldyreva L.V., Marra M., Carvalho A.B., Dimitri P., Villasante A., Zhimulev I.F., Rubin G.M., Karpen G.H., Celniker S.E. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res*. 2015;25(3):445-458. doi 10.1101/gr.185579.114

Iarovaia O.V., Minina E.P., Sheval E.V., Onichtchouk D., Dokudovskaya S., Razin S.V., Vassetzky Y.S. Nucleolus: a central hub for nuclear functions. *Trends Cell Biol.* 2019;29(8):647-659. doi 10.1016/j.tcb.2019.04.003

Janssen A., Colmenares S.U., Karpen G.H. Heterochromatin: guardian of the genome. *Annu Rev Cell Dev Biol*. 2018;34:265-288. doi 10.1146/annurev-cellbio-100617-062653

Johansen K.M., Cai W., Deng H., Bao X., Zhang W., Girton J., Johansen J. Polytene chromosome squash methods for studying transcription and epigenetic chromatin modification in *Drosophila* using antibodies. *Methods*. 2009;48(4):387-397. doi 10.1016/j.ymeth.2009.02.019

Kapoor S., Zhu L., Froyd C., Liu T., Rusche L.N. Regional centromeres in the yeast *Candida lusitaniae* lack pericentromeric heterochromatin. *Proc Natl Acad Sci USA*. 2015;112(39):12139-12144. doi 10.1073/pnas.1508749112

Kyriacou E., Heun P. Centromere structure and function: lessons from *Drosophila*. *Genetics*. 2023;225(4):iyad170. doi 10.1093/genetics/iyad170

Langmead B., Trapnell C., Pop M., Salzberg S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi 10.1186/gb-2009-10-3-r25

Lindsley D.L., Zimm G.G. The Genome of *Drosophila melanogaster*. New York: Academic Press, 1992

Lu B.Y., Emtage P.C., Duyf B.J., Hilliker A.J., Eissenberg J.C. Heterochromatin protein 1 is required for the normal expression of two heterochromatin genes in *Drosophila*. *Genetics*. 2000;155(2):699-708. doi 10.1093/genetics/155.2.699

Maekawa S., Ueda Y., Yanagisawa S. Overexpression of a Brix domain-containing ribosome biogenesis factor ARPF2 and its interactor ARRS1 causes morphological changes and lifespan extension in *Arabidopsis thaliana*. *Front Plant Sci*. 2018;9:1177. doi 10.3389/fpls.2018.01177

Markstein M., Pitsouli C., Villalta C., Celniker S.E., Perrimon N. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat Genet*. 2008;40(4):476-483. doi 10.1038/ng.101

McPherson W.K., Van Gorder E.E., Hilovsky D.L., Jamali L.A., Keliinui C.N., Suzawa M., Bland M.L. Synchronizing *Drosophila* larvae with the salivary gland reporter Sgs3-GFP for discovery of phenotypes in the late third instar stage. *Dev Biol*. 2024;512:35-43. doi 10.1016/j.ydbio.2024.05.002

Meyer-Nava S., Nieto-Caballero V.E., Zurita M., Valadez-Graham V. Insights into HP1a-chromatin interactions. *Cells*. 2020;9(8):1866. doi 10.3390/cells9081866

Meyer-Nava S., Zurita M., Valadez-Graham V. Immunofluorescent staining for visualization of heterochromatin associated proteins in *Drosophila* salivary glands. *J Vis Exp*. 2021;174:e62408. doi 10.3791/62408

Morita D., Miyoshi K., Matsui Y., Toh-e A., Shinkawa H., Miyakawa T., Mizuta K. Rpf2p, an evolutionarily conserved protein, interacts with ribosomal protein L11 and is essential for the processing of 27 SB pre-rRNA to 25 S rRNA and the 60 S ribosomal subunit assembly in *Saccharomyces cerevisiae*. *J Biol Chem*. 2002;277(32):28780-28786. doi 10.1074/jbc.M203399200

Németh A., Längst G. Genome organization in and around the nucleolus. *Trends Genet*. 2011;27(4):149-156. doi 10.1016/j.tig.2011.01.002

Olausson H.K., Nistér M., Lindström M.S. Loss of nucleolar histone chaperone NPM1 triggers rearrangement of heterochromatin and synergizes with a deficiency in DNA methyltransferase DNMT3A to drive ribosomal DNA transcription. *J Biol Chem*. 2014;289(50):34601-34619. doi 10.1074/jbc.M114.569244

Padeken J., Heun P. Centromeres in nuclear architecture. *Cell Cycle*. 2013;12(22):3455-3456. doi 10.4161/cc.26697

Padeken J., Mendiburo M.J., Chlamydas S., Schwarz H.J., Kremmer E., Heun P. The nucleoplasmin homolog NLP mediates centromere clustering and anchoring to the nucleolus. *Mol Cell*. 2013;50(2):236-249. doi 10.1016/j.molcel.2013.03.002

Panse V.G., Johnson A.W. Maturation of eukaryotic ribosomes: acquisition of functionality. *Trends Biochem Sci*. 2010;35(5):260-266. doi 10.1016/j.tibs.2010.01.001

Pavlakis G.N., Jordan B.R., Wurst R.M., Vournakis J.N. Sequence and secondary structure of *Drosophila melanogaster* 5.8S and 2S rRNAs and of the processing site between them. *Nucleic Acids Res*. 1979;7(8):2213-2238. doi 10.1093/nar/7.8.2213

Petersen L.K., Stowers R.S. A Gateway MultiSite recombination cloning toolkit. *PLoS One*. 2011;6(9):e24531. doi 10.1371/journal.pone.0024531

Pindyurin A.V. Genome-wide cell type-specific mapping of in vivo chromatin protein binding using an FLP-inducible DamID system in *Drosophila*. In: Kaufmann M., Klinger C., Savelsbergh A. (Eds) Functional Genomics. Methods in Molecular Biology. Vol. 1654. New York: Humana Press, 2017;99-124. doi 10.1007/978-1-4939-7231-9_7

Pindyurin A.V., Pagie L., Kozhevnikova E.N., van Arensbergen J., van Steensel B. Inducible DamID systems for genomic mapping of chromatin proteins in *Drosophila*. *Nucleic Acids Res*. 2016;44(12):5646-5657. doi 10.1093/nar/gkw176

Pindyurin A.V., Ilyin A.A., Ivankin A.V., Tselebrovsky M.V., Nenasheva V.V., Mikhaleva E.A., Pagie L., van Steensel B., Shevelyov Y.Y. The large fraction of heterochromatin in *Drosophila* neurons is bound by both B-type lamin and HP1a. *Epigenetics Chromatin*. 2018;11(1):65. doi 10.1186/s13072-018-0235-8

Quinodoz S.A., Ollikainen N., Tabak B., Palla A., Schmidt J.M., Detmar E., Lai M.M., Shishkin A.A., Bhat P., Takei Y., Trinh V., Aznauryan E., Russell P., Cheng C., Jovanovic M., Chow A., Cai L., McDonel P., Garber M., Guttman M. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*. 2018;174(3):744-757.e724. doi 10.1016/j.cell.2018.05.024

Razin S.V., Ulianov S.V. Genome-directed cell nucleus assembly. *Biology* (*Basel*). 2022;11(5):708. doi 10.3390/biology11050708

Rea S., Eisenhaber F., O'Carroll D., Strahl B.D., Sun Z.W., Schmid M., Opravil S., Mechtler K., Ponting C.P., Allis C.D., Jenuwein T. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*. 2000;406(6796):593-599. doi 10.1038/35020506

Rodrigues A., MacQuarrie K.L., Freeman E., Lin A., Willis A.B., Xu Z., Alvarez A.A., Ma Y., White B.E.P., Foltz D.R., Huang S. Nucleoli and the nucleoli-centromere association are dynamic during normal development and in cancer. *Mol Biol Cell*. 2023;34(4):br5. doi 10.1091/mbc.E22-06-0237

Schotta G., Ebert A., Krauss V., Fischer A., Hoffmann J., Rea S., Jenuwein T., Dorn R., Reuter G. Central role of *Drosophila* SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J*. 2002;21(5):1121-1131. doi 10.1093/emboj/21.5.1121

Schotta G., Ebert A., Reuter G. SU(VAR)3-9 is a conserved key function in heterochromatic gene silencing. *Genetica*. 2003;117(2-3):149-158. doi 10.1023/a:1022923508198

Sentmanat M.F., Elgin S.C. Ectopic assembly of heterochromatin in *Drosophila melanogaster* triggered by transposable elements. *Proc Natl Acad Sci USA*. 2012;109(35):14104-14109. doi 10.1073/pnas.1207036109

Smirnov E., Cmarko D., Mazel T., Hornáček M., Raška I. Nucleolar DNA: the host and the guests. *Histochem Cell Biol*. 2016;145(4):359-372. doi 10.1007/s00418-016-1407-x

Smith C.D., Shu S., Mungall C.J., Karpen G.H. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science*. 2007;316(5831):1586-1591. doi 10.1126/science.1139815

Suárez Freire S., Perez-Pandolfo S., Fresco S.M., Valinoti J., Sorianello E., Wappner P., Melani M. The exocyst complex controls multiple events in the pathway of regulated exocytosis. *eLife*. 2024;12:RP92404. doi 10.7554/eLife.92404

Tavares A.A., Glover D.M., Sunkel C.E. The conserved mitotic kinase polo is regulated by phosphorylation and has preferred microtubule-associated substrates in *Drosophila* embryo extracts. *EMBO J*. 1996;15(18):4873-4883. doi 10.1002/j.1460-2075.1996.tb00868.x

Tracy C., Krämer H. Isolation and infection of *Drosophila* primary hemocytes. *Bio Protoc*. 2017;7(11):e2300. doi 10.21769/BioProtoc.2300

Trinkle-Mulcahy L. Nucleolus: the consummate nuclear body. In: Lavelle C., Victor J.-M. (Eds) Nuclear Architecture and Dynamics. Academic Press, 2018;257-282. doi 10.1016/B978-0-12-803480-4.00011-9

Ugrinova I., Monier K., Ivaldi C., Thiry M., Storck S., Mongelard F., Bouvet P. Inactivation of nucleolin leads to nucleolar disruption, cell cycle arrest and defects in centrosome duplication. *BMC Mol Biol*. 2007;8:66. doi 10.1186/1471-2199-8-66

Westphal T., Reuter G. Recombinogenic effects of suppressors of position-effect variegation in Drosophila. *Genetics*. 2002;160(2):609-621. doi 10.1093/genetics/160.2.609

Conflict of interest. The authors declare no conflict of interest.
Received September 5, 2024. Revised December 2, 2024. Accepted December 4, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):414-422

doi 10.18699/vjgb-25-44

The key role of heterochromatin in the phenotypic manifestation of the *In(1)sc*^{8} inversion disrupting the *achaete-scute* complex in *Drosophila melanogaster*

T.D. Kolesnikova [1, 2], M.N. Balantaeva[2], G.V. Pokholkova [1], O.V. Antonenko[1], I.F. Zhimulev [1]

[1] Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
kolesnikova@mcb.nsc.ru

Abstract. The *achaete-scute* complex (*AS-C*) is a locus approximately 90 kbp in length, containing multiple enhancers. The local expression of the *achaete* and *scute* genes in proneural clusters of *Drosophila melanogaster* imaginal discs results in the formation of a well-defined pattern of macrochaetae in adult flies. A wide variety of easily analyzed phenotypes, along with the direct connection between individual regulatory elements and the development of specific setae make this locus a classic model in developmental genetics. One classic *AS-C* allele is *sc*^{8}, which arose as a result of the *In(1)sc*^{8} inversion. One breakpoint of this inversion lies between the *ac* and *sc* genes, while the second is in the pericentromeric heterochromatin of chromosome X, within satellite block 1.688. The heterochromatic position of the breakpoint raised the question of whether position effect variegation contributes to the disruption of normal locus function in the *In(1)sc*^{8} flies. However, conflicting results were obtained. Previously, we found that a secondary inversion, *In(1)19EHet*, arose spontaneously in one of the stocks of the *In(1)sc*^{8} BDSC line, transferring most of the heterochromatin from the *ac* gene to the 19E region of the X chromosome. Here, we demonstrate that the *In(1)19EHet* inversion leads to complete rescue of the number of posterior supraalar (PSA) and partial rescue of the number of dorsocentral (DC) macrochaetes observed in the original *In(1)sc*^{8} line. The same rescue of the macrochaetes pattern was observed when the *In(1)sc*^{8} inversion was introduced into a strain with the *Su(var)3-9*^{06} position effect modifier. Combining the inversion with the *Rif1*^{1} mutation, a conserved factor determining late replication and underreplication, does not restore the normal pattern of bristles. Our data indicate that the phenotype of flies carrying the *In(1)sc*^{8} inversion, associated with a disturbance in bristle development, is determined by the effect of heterochromatin on the distal part of the locus. This model can be used to test the influence of various factors on the position effect variegation caused by heterochromatin. Another phenotypic manifestation of *In(1)sc*^{8}, a decreased proportion of males in the offspring, was independent of the proximity of the distal part of *AS-C* to heterochromatin and was not affected by the *Rif1*^{1} mutation.

Key words: *achaete-scute* complex; *AS-C*; position effect; position effect modifiers; heterochromatin; inversions; *Drosophila melanogaster*; *Rif1*; *Su(var)3-9*

For citation: Kolesnikova T.D., Balantaeva M.N., Pokholkova G.V., Antonenko O.V., Zhimulev I.F. The key role of heterochromatin in the phenotypic manifestation of the *In(1)sc*^{8} inversion disrupting the *achaete-scute* complex in *Drosophila melanogaster*. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):414-422. doi 10.18699/vjgb-25-44

Funding. The work was supported by the Russian Science Foundation (grant No. 24-14-00133).

Acknowledgements. The authors thank D.P. Furman for the discussion of results and assistance in the literature search, and V.E. Grafodatskaya for pointing out the disappearance of bristle pattern abnormalities characteristic of *In(1)sc*^{8} flies in flies with the double inversion *In(1)sc*^{8}*+19EHet*, while supporting the collections of *Drosophila* laboratory lines at the IMCB SB RAS.

Определяющая роль гетерохроматина в фенотипическом проявлении инверсии *In(1)sc*^{8}, разрывающей *achaete-scute* комплекс *Drosophila melanogaster*

Т.Д. Колесникова [1, 2], М.Н. Балантаева[2], Г.В. Похолкова [1], О.В. Антоненко[1], И.Ф. Жимулев [1]

[1] Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
kolesnikova@mcb.nsc.ru

Аннотация. Локус *achaete-scute* (*achaete-scute* complex, *AS-C*) занимает около 90 т. п. н. и содержит множественные энхансеры. Локальная экспрессия генов *achaete* и *scute* в пронейральных кластерах имагинальных дисков *Drosophila melanogaster* приводит к формированию детерминированного рисунка макрохет у взрослых мух. Большое многообразие легко анализируемых видимых фенотипов, прямая связь между отдельными регуляторными элементами и развитием конкретных щетинок сделали этот локус классическим модельным объектом генетики развития. Одним из самых известных *AS-C* является аллель *sc*^{8}, возникший в результате инверсии *In(1)sc*^{8}. Дистальная точка разрыва этой инверсии лежит между генами *ac* и *sc*, проксимальная – в прицентромерном гетерохроматине хромосомы X, в блоке сателлита 1.688. Гетерохроматиновое положение точки разрыва поднимало вопрос о роли эффекта положения мозаичного типа в нарушении нормальной работы локуса у носителей инверсии, но были получены противоречивые результаты. Ранее мы обнаружили, что в одном из стоков линии, несущей *In(1)sc*^{8}, спонтанно возникла вторичная инверсия *In(1)19EHet*, которая переносит большую часть гетерохроматина от гена *ac* в локус, соответствующий району 19E политенной X-хромосомы. В настоящей статье мы показали, что инверсия *In(1)19EHet* приводит к полному восстановлению числа задних супраалярных и частичному – дорзо-центральных щетинок, наблюдаемых у мух исходной линии *In(1)sc*^{8}. Точно такое же восстановление паттерна щетинок мы увидели при введении в линию с инверсией *In(1)sc*^{8} модификатора эффекта положения – мутации *Su(var)3-9*^{06}. Введение же в линию с инверсией мутации по гену *Rif1* – консервативного фактора, определяющего позднюю репликацию и недорепдикацию ДНК в клетках *D. melanogaster*, не приводит к восстановлению нормального паттерна щетинок. Наши данные указывают на то, что фенотип мух – носителей инверсии *In(1)sc*^{8}, связанный с нарушением развития щетинок, определяется эффектом гетерохроматина на дистальную часть локуса и может использоваться для проверки влияния различных факторов на вызываемый гетерохроматином эффект положения гена. Еще одно фенотипическое проявление *In(1)sc*^{8} – снижение доли самцов в потомстве – оказалось независимым от соседства дистальной части *AS-C* с гетерохроматином. Этот фенотип также не восстанавливался на фоне мутации в гене *Rif1*.
Ключевые слова: *achaete-scute* complex; *AS-C*; эффект положения; модификаторы эффекта положения; гетерохроматин; инверсии; *Drosophila melanogaster*; *Rif1*; *Su(var)3-9*

Introduction

External mechanoreceptors in *Drosophila* are represented by bristles of varying sizes – macro- and microchaetae. Macrochaetae form a stable structural composition known as the bristle pattern, in which each macrochaeta occupies a strictly defined position. The formation of the bristle pattern begins with the establishment of its precursor in the imaginal disc. The specificity of the future mechanoreceptor positions is determined by the local expression of two proneural genes – *achaete* (*ac*) and *scute* (*sc*) – that are part of the *AS-C* complex (Modolell, Campuzano, 1998; Gómez-Skarmeta et al., 2003; Bukharina, Furman, 2015; Troost et al., 2015; Furman, Bukharina, 2019). The *AS-C* occupies approximately 90 kilobase pairs of DNA and consists of four genes (*achaete*, *scute*, *lethal of scute*, and *asense*) that encode transcription factors involved in the regulation of nervous system development. Multiple enhancers have been identified and characterized within the locus, each of which determines the function of the complex genes in specific proneural clusters, giving rise to the corresponding macrochaeta (Fig. 1*a*, *b*). Under normal conditions, the *ac* and *sc* genes are regulated by the same enhancers, and their products are produced in the same cells. Furthermore, the functions of these genes are partially redundant (Modolell, Campuzano, 1998).

The inversion *In(1)sc*^{8} (Fig. 1*c*, *d*) splits the *AS-C* locus apart between the *ac* and *sc* genes and connects both parts to 1.688 satellite blocks in pericentric heterochromatin (Miller et al., 2016). In this case, the distal enhancers remain with the portion of the complex carrying *ac*, while the others are translocated by the inversion along with the *sc* gene (Fig. 1*d*). It has been demonstrated that in this scenario, one part of the proneural clusters expresses only *ac*, while the complementary part expresses only *sc*, and their effects complement each other (Gómez-Skarmeta et al., 1995). This likely explains the weak phenotype observed in the inversion-bearing flies, as phenotypic changes involve only a few groups of macrochaetae. In inversion carriers, a reduction in the number of supra-alars (SA) may occur, and additional bristles can be found on the scutellum and in the dorso-central region (García-Bellido, 1979; Lindsley, Zimm, 1992).

The inversion *In(1)sc*^{8} was obtained in the laboratory of A.S. Serebrovsky through irradiation of flies with the w^a genotype (Sidorov, 1931). The line *In(1)sc*^{8}, sc^{8} y^{31d} w^{a} was transferred to the Bloomington Drosophila Stock Center (BDSC) in 1986 and assigned the number #798. In 2012, this line was split into two independent sublines (#798 main copy and #798 backup copy). In 2020, after replacing the second chromosome in the #798 main copy line with a chromosome carrying the *Rif1*^{1} mutation, which completely suppresses the underreplication of heterochromatic sequences in polytene chromosomes, an additional inversion was discovered. This finding occurred during the analysis of polytene chromosome preparations. The breakpoints of the new inversion were characterized cytologically and molecularly (Kolesnikova et al., 2022). The new inversion was named "*In(1)19EHet*", and the complex chromosomal rearrangement involving both inversions was designated "*In(1)sc*^{8}*+19EHet*". Schematic diagrams illustrating the positions of the breakpoints are shown in Figure 1 (*c*, *d*).

Numerous genetic and environmental factors are involved in bristle development. The phenotype depends on the genetic background, developmental temperature of the flies, and their sex (Child, 1935; Furman, Ratner, 1977). The existence of two lines with a common origin, similar genetic backgrounds (both lines are descendants of flies that were split from a single tube in 2012), and differing by the *In(1)19EHet* inversion, which removes a large portion of heterochromatin from one of the breakpoints of the *In(1)sc*^{8} inversion, provides a unique opportunity to investigate the influence of heterochromatin on the phenotype of inversion carriers.

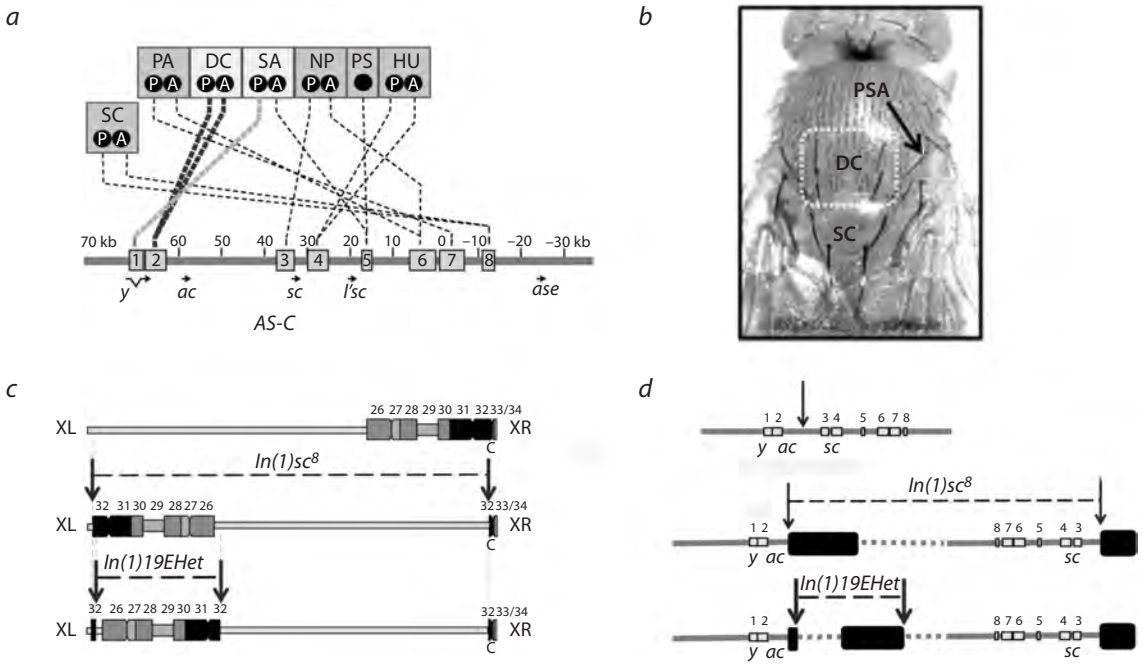


Fig. 1. The double inversion *In(1)sc*^{8}*+19EHet* splits the *AS-C* complex into two parts, attaching each part to a block of pericentric heterochromatin, resulting in altered spatial arrangement of the locus genes and their regulators, as well as a potential position effect variegation.

a – the relationship between the regulatory elements of the *AS-C* complex and the positions of macrochaetae (only macrochaetae on the notum are shown), the development of which is governed by the corresponding elements.
Notations: P (posterior); A (anterior); SC (scutellars); PA (post-alars); DC (dorsocentrals); SA (supra-alars); NP (notopleurals); PS (presutural); HU (humerals) – bristles. At the bottom, the positions of the four genes of the *AS-C* complex are indicated: *achaete* (*ac*), *scute* (*sc*), l*ethal of scute* (*l'sc*), and *asense* (*ase*), as well as the *yellow* gene (*y*). Dashed lines represent experimentally established connections between specific enhancers (gray rectangles numbered 1–8) in the locus and the macrochaetae, the development of which is controlled by these enhancers. The enhancers determine in which proneural clusters the expression of *ac* and *sc* is activated (after (Held, 2021), with modifications); *b* – the position of macrochaetae PSA and DC on the notum of *D. melanogaster*, determined by the distal enhancers of the *AS-C* locus, as well as the scutellar macrochaetae (SC); *c* – a schematic representation of the localization of the breakpoints of inversions *In(1)sc*^{8} and *In(1)19EHet* relative to the *D. melanogaster* heterochromatin map (Gatti, Pimpinelli, 1992); *d* – the location of the breakpoints of the rearrangements *In(1)sc*^{8} and *In(1)19EHet* relative to the elements of the *AS-C* complex of *D. melanogaster*. The inversion *In(1)19EHet* translocates a large portion of the heterochromatin attached to the distal part of the *AS-C* cluster from the breakpoint of *In(1)sc*^{8} to a region corresponding to section 19E of the polytene X chromosome (Kolesnikova et al., 2022).

Materials and methods

Flies were cultured at temperatures of 18 or 25 ºC, avoiding overcrowding, on a standard medium composed of: agar-agar – 10 g, pressed yeast – 100 g, cornmeal – 50 g, sugar – 20 g, and raisins – 40 g per 1 liter of water. The Table lists the fly lines used in the study and the sample sizes.

Flies were examined under a binocular loupe. For each fly, information was recorded in the table regarding the number of posterior supraalar bristles (2 (normal), 1, or 0), and the presence/absence of abnormalities in the number of dorsocentral and scutellar bristles (normal, extra/missing bristles).

Statistical analysis was conducted using Excel, and the Chi-square test was employed to assess the significance of the differences (*p*-value).

Results

Based on literature data regarding the phenotype of *In(1)sc*^{8}, we selected three groups of macrochaetae for detailed analysis: posterior supra-alar (PSA), dorsocentral (DC), and scutellar (SC). We analyzed the number of these bristles on the notum of flies from the lines *In(1)sc*^{8}, *y*^{31} *sc*^{8} *w*^{a} (hereinafter referred to for simplicity as *In(1)sc*^{8}), and *In(1)sc*^{8} *In(1)19EHet*, *sc*^{8} *y*^{31d} *w*^{a} (hereinafter *In(1)sc*^{8}*+19EHet*). As control lines, for which mutations disrupting macrochaeta development have not been described, we used the wild-type line Oregon R and several lines from the laboratory of molecular cytogenetics, IMCB SB RAS.

Additionally, we included the line *In(1)sc*^{8}*; Su(var)3-9*^{06} in our analysis, where the *In(1)sc*^{8} inversion occurs against a strong mosaic position effect modifier. In the control lines Oregon R and *In(1)w*^{m4}*; SuUR*, we did not observe any abnormalities in the number of PSA bristles (n = 236 and 240, respectively). The proportion of flies with additional DC bristles did not exceed 2 % in the Oregon R line and 4 % in the *In(1)w*^{m4}*; SuUR*. Additionally, in the Oregon R line, the proportion of flies with abnormalities in the number of scutellar bristles reached 6 %, while in the *In(1)w*^{m4}*; SuUR* line, it did not exceed 1 %.

In flies of the *In(1)sc*^{8} line, it is characteristic to observe the absence of one or both PSA bristles (Fig. 2*a*). Only 14–15 % of females reared at both 25 and 18 ºC had both PSA bristles present (Fig. 2*b*).

The phenotype of the absence of the posterior supraalar bristle is fully rescued in flies with the double inversion (Fig. 2). The most distal enhancer of the *AS-C* locus is responsible for the development of PSA bristles (Fig. 1*a*). The

The *D. melanogaster* lines used in the study and the sizes of the corresponding samples

Short name	Genotype	Origin	Sample, pcs			
			25 °C		18 °C	
			females	males	females	males
In(1)sc^{8}	*In(1)sc*^{8}, *sc*^{8} *y*^{31d} *w*^{a}	#798 (BDSC)	368	258	370	220
In(1)sc^{8}+19EHet	*In(1)sc*^{8}, *In(1)19EHet*, *sc*^{8} *y*^{31d} *w*^{a}	#94727 (BDSC)	247	151	411	254
In(1)sc^{8}; *Su(var)3-9*^{06}	*In(1)sc*^{8}, *sc*^{8} *y*^{31d} *w*^{a}; *Su(var)3-9*^{06}	(Demakova et al., 2007)	204	186	230	163
Oregon R	Wild-type line	From the Institute of Molecular and Cellular Biology collection	47	60	66	59
In(1)w^{m4}; *SuUR*	The line does not carry annotated mutations in *AS-C*	From the Institute of Molecular and Cellular Biology collection	34	31	96	79
In(1)sc^{8}; *Rif1*^{1}	*In(1)sc*^{8}, *y*^{31} *sc*^{8} *w*^{a}; *Rif1*^{1}	The line obtained in this work by replacing chromosome 2 in line #798 backup copy with the second chromosome of line *Rif1*^{1}	–	–	121	49
In(1)sc^{8}+19EHet; *Rif1*^{1}	*In(1)sc*^{8}, *In(1)19EHet*, *y*^{31} *sc*^{8} *w*^{a}; *Rif1*^{1}	(Kolesnikova et al., 2022)	–	–	130	44
Rif1^{1}	*w*^{1118}; *Rif1*^{1}	The line kindly provided by J. Nordman (Munden et al., 2018)	–	–	62	55

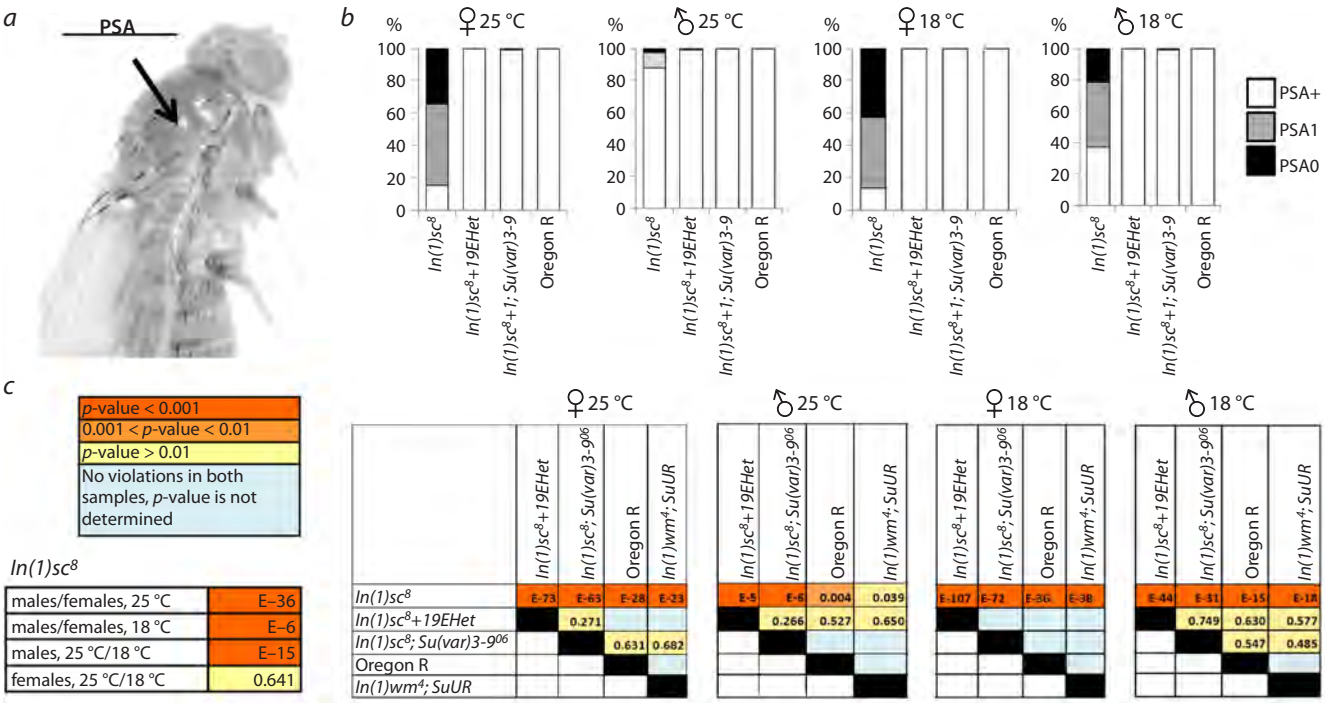


Fig. 2. The effect of heterochromatin removing and the position effect variegation modifier *Su(var)3-9*^{06} on the development of posterior supra-alar bristles in *In(1)sc*^{8} flies.

a – a typical phenotype of the *In(1)sc*^{8} line showing the absence of the posterior supraalar bristle (arrow indicates the position where the PSA is normally located); *b* – the proportion of flies with abnormalities in the number of posterior supraalar bristles (PSA+ – normal phenotype, PSA1 – presence of one bristle, PSA0 – absence of bristles) in the lines *In(1)sc*^{8}, *In(1)sc*^{8}+19EHet, *In(1)sc*^{8}; *Su(var)3-9*^{06}, and two control lines Oregon R and *In(1)w*^{m4}; *SuUR*. Results for males and females are provided separately, as well as for flies reared at different temperatures; *c* – pairwise comparison of the proportions of flies with altered bristle counts among different genotypes. The *p*-value for each comparison, calculated using the Chi-square test, is indicated.

complete rescue of the phenotype in flies with the secondary inversion indicates that the phenotype of the developmental anomalies of these bristles is caused not by the break in *AS-C* itself, but by the effect of heterochromatin on the distal part of the locus. Additional confirmation of the predominant effect of heterochromatin on the PSA phenotype is the complete restoration of the normal PSA phenotype in flies carrying *In(1)sc*^{8} and the position effect modifier *Su(var)3-9*^{06} (n = 783).

Interestingly, the proportion of males with abnormalities in the number of PSA bristles is significantly lower, at only 12 % at 25 ºC, although it rises to 42 % at 18 ºC, in accordance with the classical understanding of the enhancement of the

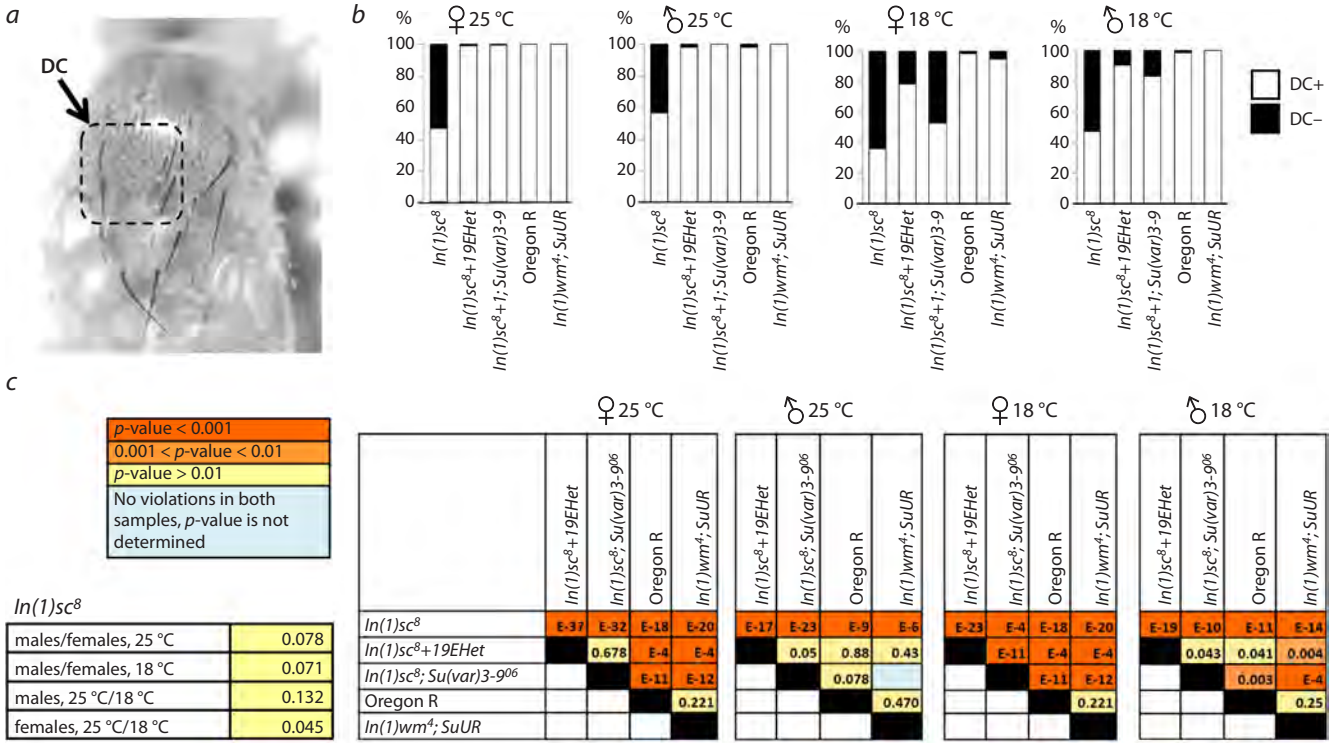


Fig. 3. The effect of heterochromatin removing and the position effect variegation modifier *Su(var)3-9*^{06} on the development of dorsocentral bristles in *In(1)sc*^{8} flies.

a – additional dorsocentral bristles in flies of the *In(1)sc*^{8} line; *b* – the proportion of flies with abnormalities in the number of dorsocentral bristles in the analyzed lines; *c* – pairwise comparison of the proportions of flies with altered bristle counts among different genotypes. The *p*-value for each comparison, calculated using the Chi-square test, is indicated.

heterochromatin position effect at lower temperatures (Elgin, Reuter, 2013). This potentially explains why PSA disruption is not documented for flies with *In(1)sc*^{8} in A. García-Bellido's (1979) article. In that work, only hemizygous males of *In(1)sc*^{8} were analyzed, which is mentioned in a separate comment. The authors may have aimed to emphasize the weak phenotype in carriers of this inversion compared to other rearrangements affecting *AS-C*. A more pronounced influence of the inversion *sc*^{8} on the PSA phenotype in females was also observed in the work by D.P. Furman and V.A. Ratner (1977).

Another phenotype described in the literature for *In(1)sc*^{8} flies is the appearance of additional bristles in the dorsocentral zone (García-Bellido, 1979; Lindsley, Zimm, 1992) (Fig. 3*a*). We observed this phenotype in both females and males of *In(1)sc*^{8} (Fig. 3*b*). The additional bristles most often formed orderly rows with pairs of PDC and ADC bristles. Instances of absence of individual dorsocentral bristles were also noted. In flies of *In(1)sc*^{8}*+19EHet* as well as *In(1)sc*^{8}*; Su(var)3-9*^{06}, reared at 25 °C, we saw almost complete restoration of the phenotype. When culturing flies at 18 °C, the proportion of carriers with the mutant phenotype was higher in the *In(1)sc*^{8} line and was only partially, but significantly reduced against the background of the position effect modifier and the *19EHet* inversion. Dorsocentral bristles develop from proneural clusters, where *AS-C* expression is regulated by enhancer 2, a distal enhancer that is closer to the break point of the inversion than the enhancer controlling the development of PSA. It can be assumed that the effect of heterochromatin on this enhancer is stronger. According to the data from nanopore sequencing of *In(1)sc*^{8}*+19EHet* flies, at least 30 kb of satellite DNA 1.688 continues to flank the break point of the *In(1)sc*^{8} inversion (Kolesnikova et al., 2022).

In some literature sources, it is noted that mutants *In(1)sc*^{8} are characterized by abnormalities in the number, thickness, and length of scutellar bristles (Lindsley, Zimm, 1992; Belyaeva et al., 2003; Golovin et al., 2003). However, we did not observe a significant increase in the proportion of flies with additional or absent scutellar bristles in *In(1)sc*^{8} compared to control lines – under certain conditions, the proportion of flies with scutellar bristle abnormalities was higher in the control than in *In(1)sc*^{8} flies (Fig. 4).

Another phenotype presumably associated with the heterochromatin effect on *AS-C* in flies with the *In(1)sc*^{8} inversion is the decrease in the proportion of males in the offspring. This phenotype is most pronounced in the absence of the Y chromosome, which serves as the primary evidence that it is related to a position effect (Lindsley, Zimm, 1992; Belyaeva et al., 2003). All lines carrying both *In(1)sc*^{8} and the double inversion exhibited a significant decrease in the male proportion in the offspring (Fig. 5). This ratio did not change in response to the removal of heterochromatin by the secondary inversion, nor did it depend on temperature; however, at a temperature of 25 °C, it was restored in the presence of *Su(var)3-9*^{06}. Thus, this phenotype is not associated with the distal part of *AS-C*.

Since the mutant phenotype associated with the absence of PSA bristles is observed in a large proportion of flies in the

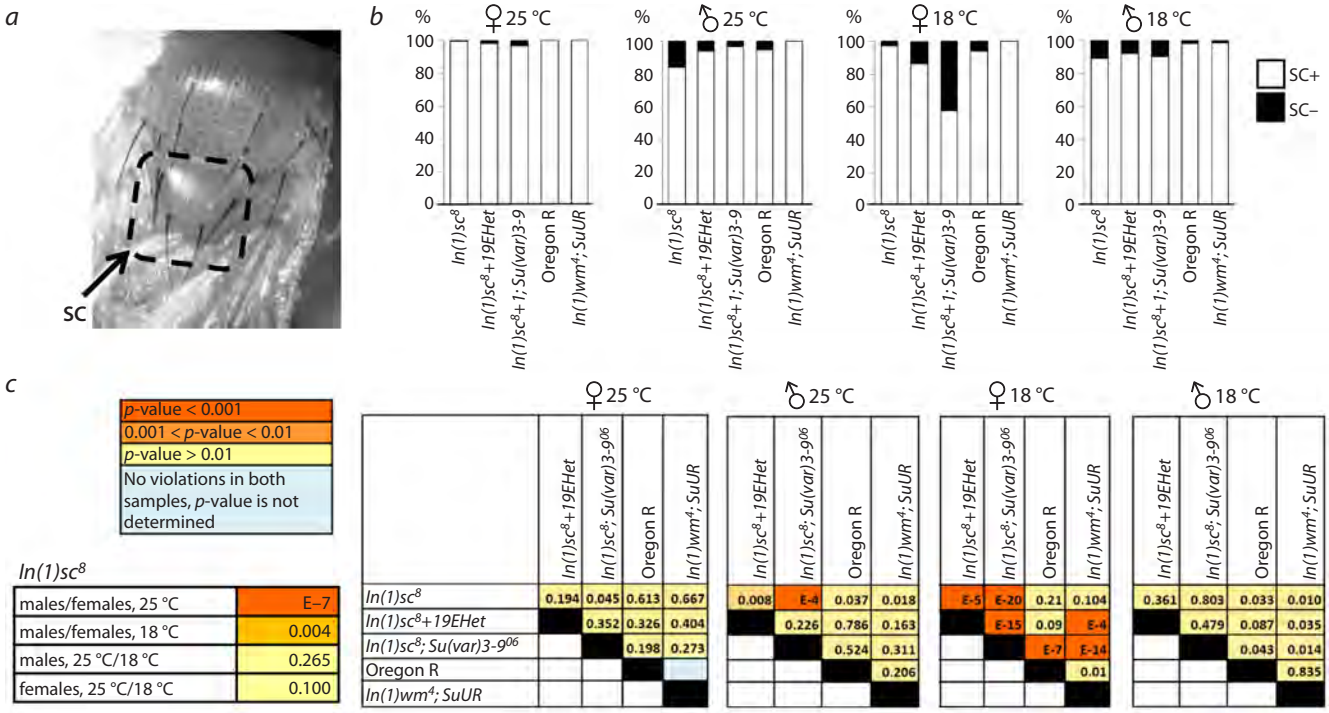


Fig. 4. Analysis of the effect of heterochromatin removing and the position effect variegation modifier *Su(var)3-9*^{06} on the development of scutellar bristles in *In(1)sc*^{8} flies.

a – additional scutellar bristles in the *In(1)sc*^{8} line; *b* – the proportion of flies with abnormalities in the number of scutellar bristles in the lines *In(1)sc*^{8}, *In(1)sc*^{8}*+19EHet*, *In(1)sc*^{8}*; Su(var)3-9*^{06} and two control lines, Oregon R and *In(1)wm*^{4}*; SuUR*; *c* – pairwise comparison of the proportions of flies with altered bristle counts among different genotypes. The *p*-value was calculated using the Chi-square test.

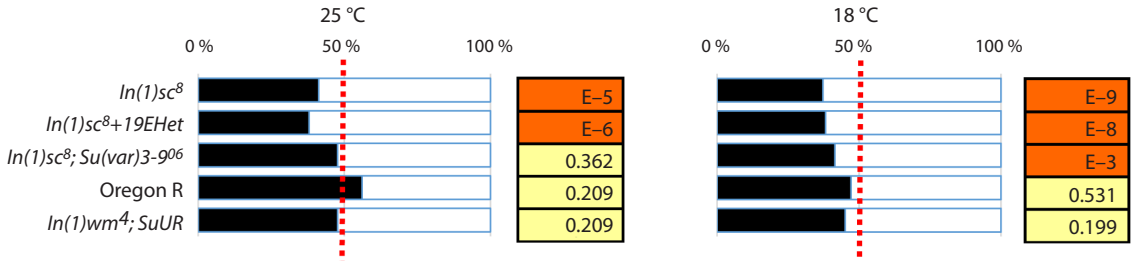


Fig. 5. Decrease in the proportion of males in lines carrying the *In(1)sc*^{8} inversion.

The ratio of females to males is shown for flies reared at temperatures of 25 and 18 °C in lines harboring *In(1)sc*^{8} and in control lines. For each comparison, the *p*-value calculated using the Chi-square test is indicated.

In(1)sc^{8} line, and the probability of its manifestation depends on modifier factors affecting position effects variegation, such as temperature and *Su(var)3-9*, we decided to use *In(1)sc*^{8} as a model system to test whether a mutation in the *Rif1* gene acts as a modifier of position effect. The Rap1 interacting factor 1 (Rif1) protein is an evolutionarily conserved protein that participates in various processes, including telomere length regulation, DNA repair, and establishing the temporal order of replication origin activation (Richards et al., 2022). In *D. melanogaster*, Rif1 is involved in establishing the late replication program of satellite sequences during embryogenesis (Sreesankar et al., 2015; Seller, O'Farrell, 2018) and is responsible for the underreplication of heterochromatin, including satellite DNA, in polytene chromosomes (Munden et al., 2018; Kolesnikova et al., 2020). To date, there are no data on whether this protein can influence the heterochromatin effect on the expression of genes positioned near blocks of satellite DNA due to chromosomal rearrangements (i. e., whether it acts as a modifier of position effect variegation).

We compared females of the lines *In(1)sc*^{8}, *In(1)sc*^{8}*; Rif1*^{1}, *In(1)sc*^{8}*+19EHet; Rif1*^{1}, and *Rif1*^{1}, cultured at 18 °C (Fig. 6). Replacing chromosome 2 in the *In(1)sc*^{8} line with chromosome 2 carrying the *Rif1*^{1} mutation does not restore the normal number of PSA bristles. Moreover, a slight enhancement of the phenotype is observed. A similar small enhancement of the phenotype was noted in terms of the decrease in the proportion of males. Comparing with the effect of *Su(var)3-9*^{06}, we can conclude that *Rif1*^{1} is not a suppressor of the position effect related to the influence of satellite 1.688 on *AS-C* in *In(1)sc*^{8} mutants.

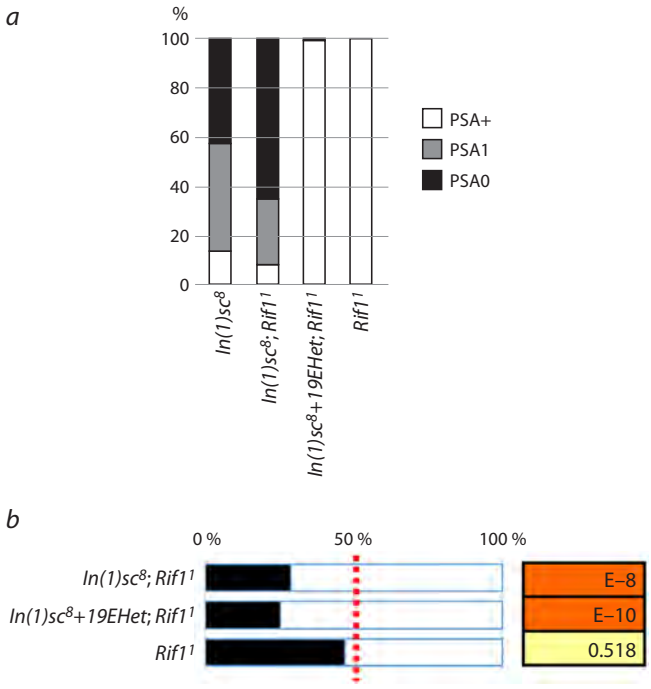


Fig. 6. Analysis of the effect of the *Rif1*^{1} mutation on the phenotypes of flies with the *In(1)sc*^{8} inversion.

a – the results of the comparison of the ratio of normal flies to flies with disrupted PSA bristles for females reared at 18 °C in the lines *In(1)sc*^{8}, *In(1)sc*^{8}; *Rif1*^{1}, *In(1)sc*^{8}*+19EHet; Rif1*^{1}, and *Rif1*^{1} are presented; *b* – the *Rif1*^{1} mutation does not rescue the phenotype of decreased male proportion in flies carrying *In(1)sc*^{8}.

Discussion

The *AS-C* locus is a classic model system for studying various aspects of genetic regulation of development in multicellular organisms. The vast array of easily analyzable visible phenotypes and the direct connection between individual regulatory elements and the development of specific bristles have made this locus highly attractive to researchers since the 1930s, and interest in it remains strong to this day. A deep understanding of how this locus is structured and regulated, as well as its role in the development of *Drosophila*, has been achieved (Modolell, Campuzano, 1998; Gómez-Skarmeta et al., 2003; Furman, Bukharina, 2019; Bukharina et al., 2023). However, some patterns and peculiarities in the behavior of mutant alleles discovered in the 1930s to 1970s have only recently become clear. A vivid example is the discovery of the hypermorphic allele of the *Notch* gene found in lines carrying the *w*^{a} mutation (Rice et al., 2015): it turned out that all balancer X chromosomes carrying *In(1)sc*^{8} differ significantly in the expression of the *sc* phenotype depending on the presence of the *w*^{a} allele and the linked *opa33b* allele of the *Notch* gene in the chromosome. Indirect evidence of the importance of the *Notch* gene status is the significant difference in the manifestation of the mutant phenotype in flies carrying the *In(1)sc*^{V2} and *In(1)sc*^{8} inversions, which have closely spaced breakpoints (Rice et al., 2015): in flies of the *In(1)sc*^{V2} line, the disruptions in the bristle pattern are significantly stronger, affecting more bristles. Such a strong effect of the genetic background likely complicated the interpretation of the observed patterns in the manifestation of mutations in the *AS-C* locus, as many *sc* alleles used in genetic studies were created based on the X chromosome carrying *w*^{a} (Sidorov, 1931; Furman, Ratner, 1977).

The heterochromatic position of the breakpoints raises the question of the role of position effects in the manifestation of inversion phenotypes. Molecular analysis of the inversion breakpoints showed that both parts of the cluster were adjacent to large blocks of satellite 1.688 (Miller et al., 2016). Studies from the 1970s failed to reach a conclusive understanding of the role of heterochromatin (Ratner, Furman, 1978), although it was noted that in lines with the *In(1)sc*^{V2} and *In(1)sc*^{8} inversions, a temperature and sex effect was observed, which differed from that in other alleles (Furman, Ratner, 1977) and supported the hypothesis of heterochromatin's effect on the *AS-C* locus. In our work, these patterns were confirmed.

The position effect of heterochromatin on the phenotype of *In(1)sc*^{8}, most prominently expressed in males of the X0 genotype, was described in a study by E.S. Belyaeva et al. (2003), which analyzed the effect on scutellar bristles. In our work, we did not observe a pronounced phenotype associated with scutellar bristle disruptions. It is possible that the differences are related to the criteria for disruption adopted for the analysis: we considered as a disruption only the excess or absence of bristles, while the work of E.S. Belyaeva et al. mentioned changes in their thickness and length. By applying stricter criteria to identify disruptions, we did not find abnormalities in bristle number, for which, according to the literature, proximal enhancers of *AS-C* are required. The work of A.K. Golovin et al. (2003) shows that the phenotype of flies with mutations *In(1)sc*^{V2} and *In(1)sc*^{8} is significantly enhanced by mutations in the genes *su(Hw)* or *mod(mdg4)*, and this enhancement affects the scutellar bristles. The authors conclude that under normal circumstances, the effect of heterochromatin on the proximal part of *AS-C* in *In(1)sc*^{V2} and *In(1)sc*^{8} mutants is blocked by an unannotated insulator. The presence of a well-studied insulator localized between the regulatory region of *AS-C* and the *yellow* gene (Golovin et al., 2003) explains the weak effect of heterochromatin on the gene *yellow* – the effect is observed only in X0 males (Lindsley and Zimm, 1992; Belyaeva et al., 2003).

Two *sc* alleles – *In(1)sc*^{V2} and *In(1)sc*^{8} – are associated with inversions, one of the breakpoints of which is located within the *AS-C* locus between the *ac* and *sc* genes, while the second is in the pericentric heterochromatin (Miller et al., 2016). Elegant studies analyzing the expression of the *ac* and *sc* genes in *In(1)sc*^{8} mutants showed that the locus, split into two parts, can continue to function normally because the functions of these genes largely duplicate each other. Under normal conditions, both proximal and distal enhancers influence the expression of each of the *ac* and *sc* genes, resulting in both proteins being detected in all proneural clusters during immunostaining of imaginal discs with antibodies to the Ac and Sc proteins. In carriers of *In(1)sc*^{8}, the corresponding regulatory element in each proneural cluster of the imaginal disc activates the expression of only one of the genes (either *ac* or *sc*), but this is sufficient to form a nearly normal bristle pattern (Gómez-Skarmeta et al., 1995). Moreover, this is sufficient for the complete restoration of the phenotype when the block of heterochromatin is removed from the distal part of the cluster or when there is a strong modifier of the position

effect variegation. We obtained direct evidence that, with a reduction in the effect of heterochromatin, the rearranged *AS-C* locus can provide a normal phenotype in flies. This observation is interesting from the perspective of the evolution of loci with complex regulatory systems: the protein-coding genes that arose as a result of duplication with redundant functions may, due to chromosomal rearrangements, divide their functions and start evolving along independent trajectories. It is known that the *ac* and *sc* genes are the result of relatively recent duplication; outside of the Drosophila group, the homologs of *ac* and *sc* are represented by a single gene (Negre, Simpson, 2009).

Using the contrasting effect of heterochromatin in flies carrying *In(1)sc*^{8} on the PSA phenotype, we decided to test whether the Rif1 protein can modify this effect. In *D. melanogaster*, the Rif1 protein is involved in establishing the late replication program of satellite sequences during embryogenesis (Seller, O'Farrell, 2018). In the polytene chromosomes, mutations in the *Rif1* gene completely suppress underreplication in heterochromatic regions, including the replication of satellite DNA (Kolesnikova et al., 2020).

In cells with polytene chromosomes, Rif1 interacts with the suppressor of underreplication (SuUR) protein (Nordman et al., 2018), which is a weak modifier of the position effect in *D. melanogaster* (Belyaeva et al., 2003). Therefore, one would expect that in *Rif1*^{1} mutants, the properties of heterochromatin could significantly differ from the norm; particularly, the effect of heterochromatin on the transcription of adjacent genes in chromosomal rearrangements could change. We did not observe a suppressive effect of the *Rif1*^{1} mutation on the position effect associated with the satellite 1.688 effect on the distal part of the *AS-C* cluster in the *In(1)sc*^{8} inversion, nor on the sex ratio in the offspring of flies carrying this inversion. Moreover, we detected a weak enhancer effect, the evidence of which requires further verification.

Conclusion

In summary, we can conclude that the phenotype associated with the disruption of the bristle pattern in *In(1)sc*^{8} mutants is primarily caused not by the splitting of the *AS-C* locus into two parts, but by the effect of heterochromatin on the distal part of the cluster. This can be used to test the influence of various factors on heterochromatin-induced position effect variegation.

References

Belyaeva E.S., Boldyreva L.V., Volkova E.I., Nanayev R.A., Alekseyenko A.A., Zhimulev I.F. Effect of the *Suppressor of Underreplication* (*SuUR*) gene on position-effect variegation silencing in *Drosophila melanogaster*. *Genetics*. 2003;165(3):1209-1220. doi 10.1093/genetics/165.3.1209

Bukharina T.A., Furman D.P. The mechanisms determining bristle pattern in *Drosophila melanogaster*. *Russ J Dev Biol*. 2015;46(3):99-110. doi 10.1134/S1062360415030029

Bukharina T.A., Golubyatnikov V.P., Furman D.P. The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis. *Vavilovskii Zhurnal Genet Selektsii = Vavilov J Genet Breed*. 2023;27(7):746-754. doi 10.18699/VJGB-23-87

Child G. Phenogenetic studies on *scute-1* of *Drosophila melanogaster*. I. The associations between the bristles and the effects of genetic modifiers and temperature. *Genetics*. 1935;20(2):109-126. doi 10.1093/genetics/20.2.109

Demakova O.V., Pokholkova G.V., Kolesnikova T.D., Demakov S.A., Andreyeva E.N., Belyaeva E.S., Zhimulev I.F. The SU(VAR)3-9/HP1 complex differentially regulates the compaction state and degree of underreplication of X chromosome pericentric heterochromatin in *Drosophila melanogaster*. *Genetics*. 2007;175(2):609-620. doi 10.1534/genetics.106.062133

Elgin S.C.R., Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harb Perspect Biol*. 2013;5(8):a017780. doi 10.1101/cshperspect.a017780

Furman D.P., Bukharina T.V. The bristle pattern development in *Drosophila melanogaster*: the prepattern and *achaete-scute* genes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2018;22(8):1046-1054. doi 10.18699/VJ18.449

Furman D.P., Ratner V.A. Investigation of the genetic topography of the *scute* locus in *Drosophila melanogaster*. II. Thermal effect of mutation manifestation in homozygotes. *Genetika = Genetics* (*Moscow*). 1977;13(4):667-680 (in Russian)

García-Bellido A. Genetic analysis of the *achaete-scute* system of *Drosophila melanogaster*. *Genetics*. 1979;91(3):491-520. doi 10.1093/genetics/91.3.491

Gatti M., Pimpinelli S. Functional elements in *Drosophila melanogaster* heterochromatin. *Annu Rev Genet*. 1992;26:239-275. doi 10.1146/annurev.ge.26.120192.001323

Golovnin A., Biryukova I., Romanova O., Silicheva M., Parshikov A., Savitskaya E., Pirrotta V., Georgiev P. An endogenous Su(Hw) insulator separates the *yellow* gene from the *Achaete-scute* gene complex in *Drosophila*. *Development*. 2003;130(14):3249-3258. doi 10.1242/dev.00543

Gómez-Skarmeta J.L., Rodríguez I., Martínez C., Culí J., Ferrés-Marcó D., Beamonte D., Modolell J. Cis-regulation of *achaete* and *scute*: shared enhancer-like elements drive their coexpression in proneural clusters of the imaginal discs. *Genes Dev*. 1995;9(15):1869-1882. doi 10.1101/gad.9.15.1869

Gómez-Skarmeta J.L., Campuzano S., Modolell J. Half a century of neural prepatterning: the story of a few bristles and many genes. *Nat Rev Neurosci*. 2003;4(7):587-598. doi 10.1038/nrn1142

Held L.I., Jr. Animal Anomalies: What Abnormal Anatomies Reveal about Normal Development. Cambridge: Cambridge University Press, 2021. doi 10.1017/9781108876612

Kolesnikova T.D., Kolodyazhnaya A.V., Pokholkova G.V., Schubert V., Dovgan V.V., Romanenko S.A., Prokopov D.Y., Zhimulev I.F. Effects of mutations in the *Drosophila melanogaster Rif1* gene on the replication and underreplication of pericentromeric heterochromatin in salivary gland polytene chromosomes. *Cells*. 2020;9(6):1501. doi 10.3390/cells9061501

Kolesnikova T.D., Klenov M.S., Nokhova A.R., Lavrov S.A., Pokholkova G.V., Schubert V., Maltseva S.V., Cook K.R., Dixon M.J., Zhimulev I.F. A spontaneous inversion of the X chromosome heterochromatin provides a tool for studying the structure and activity of the nucleolus in *Drosophila melanogaster*. *Cells*. 2022;11(23):3872. doi 10.3390/cells11233872

Lindsley D., Zimm G. The genome of *Drosophila melanogaster*. San Diego, CA: Academic Press, 1992

Miller D.E., Cook K.R., Yeganeh Kazemi N., Smith C.B., Cockrell A.J., Hawley R.S., Bergman C.M. Rare recombination events generate sequence diversity among balancer chromosomes in *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 2016;113(10):E1352-E1361. doi 10.1073/pnas.1601232113

Modolell J., Campuzano S. The *achaete-scute* complex as an integrating device. *Int J Dev Biol*. 1998;42(3):275-282. doi 10.1387/IJDB.9654009

Munden A., Rong Z., Sun A., Gangula R., Mallal S., Nordman J.T. Rif1 inhibits replication fork progression and controls DNA copy number in *Drosophila*. *eLife*. 2018;7:e39140. doi 10.7554/eLife.39140

Negre B., Simpson P. Evolution of the *achaete-scute* complex in insects: convergent duplication of proneural genes. *Trends Genet*. 2009;25(4):147-152. doi 10.1016/j.tig.2009.02.001

Ratner V.A., Furman D.P. Investigation of the genetic topography of the scute locus in *Drosophila melanogaster*. VI. Possible role of chromosome rearrangements and the position effect. *Genetika = Genetics* (*Moscow*). 1978;14(9):1662-1664 (in Russian)

Rice C., Beekman D., Liu L., Erives A. The nature, extent, and consequences of genetic variation in the *opa* repeats of *Notch* in *Drosophila*. *G3* (*Bethesda*). 2015;5(11):2405-2419. doi 10.1534/g3.115.021659

Richards L., Das S., Nordman J.T. Rif1-dependent control of replication timing. *Genes* (*Basel*). 2022;13(3):550. doi 10.3390/genes13030550

Seller C.A., O'Farrell P.H. Rif1 prolongs the embryonic S phase at the *Drosophila* mid-blastula transition. *PLoS Biol*. 2018;16(5):e2005687. doi 10.1371/journal.pbio.2005687

Sidorov B.N. Study of step-allelomorphism in *Drosophila melanogaster*. Emergence of an allelomorph for scute producing simultaneously hairy wing characters (mutation scute-8). *Zhurnal Eksperimental'noy Biologii = J Experim Biol*. 1931;7(1):28-40 (in Russian)

Sreesankar E., Bharathi V., Mishra R.K., Mishra K. *Drosophila* Rif1 is an essential gene and controls late developmental events by direct interaction with PP1-87B. *Sci Rep*. 2015;5:10679. doi 10.1038/srep10679

Troost T., Schneider M., Klein T. A re-examination of the selection of the sensory organ precursor of the bristle sensilla of *Drosophila melanogaster*. *PLoS Genet*. 2015;11:e1004911. doi 10.1371/journal.pgen.1004911

Compliance with Ethical Standards. In accordance with Article 3 of Chapter 1 of Directive 2010/63/EU of September 22, 2010, on the protection of animals used for scientific purposes, the requirements of bioethics do not apply to the subject of this research.

Conflict of interest. The authors declare no conflict of interest.

Received November 27, 2024. Revised December 9, 2024. Accepted December 16, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):423-432

doi 10.18699/vjgb-25-45

Great Bolgar's historical genetics: a genomic study of individuals from burials close to the Greek Chamber in the 14th century

T.V. Andreeva [1, 2, 3], A.D. Soshkina [3], S.S. Kunizheva [1, 3], A.D. Manakhov [1, 3], D.V. Pezhemsky [4], E.I. Rogaev [1, 5]

[1] Research Centre for Genetics and Life Sciences, Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia
[2] Centre of Genetics and Genetic Technologies, Lomonosov Moscow State University, Moscow, Russia
[3] Vavilov Institute of General Genetics of the Russian Academy of Sciences, Moscow, Russia
[4] Research Institute and Museum of Anthropology, Lomonosov Moscow State University, Moscow, Russia
[5] Department of Psychiatry, UMass Chan Medical School, Shrewsburry, MA, USA

andreeva@rogaevlab.ru; evivrecc@gmail.com

Abstract. Bolgar was one of the most significant mediaeval cities in Eastern Europe. Before the Mongol conquest, it served as a major administrative centre of Volga Bulgaria, and after 1236, it temporarily functioned as the capital of the Golden Horde. Historical, archaeological, and paleoanthropological evidence indicates a mixed population of this city during the 13th–15th centuries; however, the contributions of exact ethnic groups into its genetic structure remain unclear. To date, there are no genetic data for this medieval group. For the first time, using massive parallel sequencing methods, we determined whole-genome sequences for three individuals from Bolgar who were buried in the early 14th century close to the so-called "Greek Chamber". The average coverage of the studied genomes ranged from x0.5 to x1.5. We identified the genetic sex of the people (two men and one woman), and performed a population genetic analysis. The authenticity of the DNA studied and the low level of contamination were confirmed, and the mitochondrial DNA haplogroups of all three individuals as well as the Y-chromosome haplogroups of two male individuals were determined. We used more than 2.7 thousand DNA samples from representatives of ancient and modern populations that had been previously published to perform a comparative population-genetic analysis. Whole-genome data analysis employing uniparental markers (mitochondrial DNA and Y chromosome) and autosomal markers revealed genetic heterogeneity in this population. Based on PCA and *f4*-statistics analysis, a genetic connection was identified between one of the individuals (female) and modern Finno-Ugric peoples of the Volga-Ural region. Genomic analysis of the other two individuals suggests their Armenian origin and indicates migrant influx from the Caucasus or Anatolia. The results align well with archaeological and paleoanthropological findings and significantly enhance them by reconstructing the contributions of the indigenous population to the formation of the mediaeval Bolgar population structure.

Key words: ancient DNA; genome; massive parallel sequencing; paleoanthropology; Bolgar; Greek Chamber

For citation: Andreeva T.V., Soshkina A.D., Kunizheva S.S., Manakhov A.D., Pezhemsky D.V., Rogaev E.I. Great Bolgar's historical genetics: a genomic study of individuals from burials close to the Greek Chamber in the 14th century. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):423-432. doi 10.18699/vjgb-25-45

Funding. This work was supported by the grant of the state program of the "Sirius" Federal Territory "Scientific and technological development of the 'Sirius' Federal Territory", Agreement No. 18-03 date September 10, 2024 (TVA).

Acknowledgements. We express our gratitude to Alexandra Buzhilova from the Research Institute and Museum of Anthropology, Lomonosov Moscow State University for selecting and providing the paleoanthropological material used in the work.

К исторической генетике Великого Болгара: геномный анализ людей из погребений XIV века у Греческой палаты

Т.В. Андреева [1, 2, 3], А.Д. Сошкина [3], С.С. Кунижева [1, 3], А.Д. Манахов [1, 3], Д.В. Пежемский [4], Е.И. Рогаев [1, 5]

[1] Научный центр генетики и наук о жизни, Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия
[2] Центр генетики и генетических технологий, Московский государственный университет им. М.В. Ломоносова, Москва, Россия
[3] Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия
[4] Научно-исследовательский институт и Музей антропологии им. Д.Н. Анучина Московского государственного университета им. М.В. Ломоносова, Москва, Россия
[5] Медицинская школа Чан Массачусетского университета, департамент психиатрии, Шрусбери, США

andreeva@rogaevlab.ru; evivrecc@gmail.com

Аннотация. Великий Болгар – один самых значительных средневековых городов Восточной Европы, до монгольского завоевания был крупным административным центром Волжской Булгарии, а после 1236 г. некоторое время выполнял функции столицы Золотой Орды. Исторические, археологические и палеоантропологи-

ческие данные свидетельствуют о смешанном составе населения этого города в XIII–XV вв., однако вклад тех или иных этнических групп в его генетическую структуру остается не вполне ясным, и до сих пор отсутствуют генетические данные для этой группы средневекового населения. Мы впервые с применением методов масштабного параллельного секвенирования определили полногеномные последовательности у трех жителей Великого Болгара, погребенных в первой половине XIV в. у так называемой Греческой палаты. Среднее покрытие исследованных геномов составило от x0.5 до x1.5, что позволило определить генетический пол индивидов (двое мужчин и одна женщина) и провести популяционно-генетический анализ. Были подтверждены аутентичность исследованной ДНК и низкий уровень контаминации, а также определены гаплогруппы митохондриальной ДНК всех трех индивидов и гаплогруппы Y-хромосомы двух индивидов мужского пола. Для проведения сравнительного популяционно-генетического анализа нами задействованы опубликованные ранее данные геномного секвенирования более 2.7 тыс. образцов ДНК представителей современных и древних популяций. Анализ полногеномных данных трех индивидов из Великого Болгара с использованием однородительских маркеров (митохондриальной ДНК и Y-хромосомы) и аутосомных маркеров показал генетическую гетерогенность исследованной группы населения. По результатам геномного анализа с применением аутосомных маркеров и метода главных компонент (PCA) и *f4*-статистики нами выявлена генетическая связь одного индивида (женщины) с современными финно-угорскими народами Волго-Уральского региона. Геномный анализ двух других индивидов предполагает их армянское происхождение и свидетельствует о существовании потока мигрантов с территории Кавказа или Анатолии. Полученные результаты хорошо согласуются с данными археологии и палеоантропологии, а также существенно дополняют их в части реконструкции вклада автохтонного населения в формирование популяционной структуры средневековых жителей Болгара.

Ключевые слова: древняя ДНК; геном; параллельное секвенирование; палеоантропология; Болгар; Греческая палата

Introduction

Great Bolgar was a significant administrative center of two sequential medieval states in Eastern Europe, and one of the key cities of Volga Bulgaria, where Islam was adopted in 922. After the Mongol invasion and devastation in 1236, it was restored as the first capital of the Golden Horde. Volga Bulgaria was a polyethnic polity where Finno-Ugric peoples, Slavs, and the Turkic tribes (Bulgars) – who migrated to the interfluve of the Kama and the Great Cheremshan from the Azov region and the territory of Krasnodar Krai –coexisted. As part of the Golden Horde, the Volga region population maintained its complex ethnic composition, which was particularly characteristic of major urban centers such as Bolgar.

In its history and economy, due to its advantageous location on the main waterway of the East European plain – both during the Bulgar period and especially during the Golden Horde period – considerable importance was attributed to craft production and transcontinental trade. This circumstance contributed to the formation of a polyethnic urban population, as a significant portion of its inhabitants consisted of foreign merchants (Smirnov, 1951, 1972, 1974; Gening, 1989; Khalikov, 1989; Iskhakov, Izmailov, 2000; Bulgarica, 2012; Sitdikov, Bocharov, 2024).

Archaeological data as well as paleoanthropological materials indicate a mixed population composition in Great Bolgar throughout all stages of its existence (Trofimova, 1956; Postnikova, 1970, 1973; Efimova, 1991; Gazimzyanov, 2000, 2015). The contribution of steppe Turkic and indigenous Finno-Ugric populations to the anthropological makeup of its inhabitants is considered proven. The question of the contribution of the Azov Bulgars, who migrated to the Volga region in the early Middle Ages, as well as the participation of Slavic groups, remains a topic of discussion. During the Golden Horde period, new components that increased the ethnic diversity of the medieval Bolgar population emerged. A striking example of this is a series of skulls from burial sites at the "Greek Chamber", which were identified by T.A. Trofimova as "Armenoid" upon initial study (Trofimova, 1949).

Although the population of Great Bolgar was predominantly Muslim, as indicated by archaeological data, there were also diasporas of other faiths present. For instance, the burial ground at Babiy Bugor is interpreted by several researchers as a Christian cemetery. However, the most vivid illustration of this is a site known as the "Greek Chamber", which was a rectangular stone structure measuring 12.6 by 16.4 meters, oriented along the West-East axis. Still well visible at the beginning of the 18th century, it was scientifically documented for the first time during that period. Archaeological excavations of the "Greek Chamber" were carried twice: once in 1916 by V.F. Smolin and again in the mid-1940s by A.P. Smirnov. The structural and dimensional characteristics of the building itself, along with the gravestones featuring inscriptions in Armenian, described in a timely manner, allowed it to be identified as a small temple, similar to the construction from 1339 in Noravank (Armenia) or Armenian churches from the 14th century in Feodosia and Old Crimea (Smirnov, 1951). Temples of this type were typically two-storied, with the lower level serving as a burial chamber. The deciphered and repeatedly published epitaphs date from 1308 to 1335. This enabled A.P. Smirnov to reasonably assert that a necropolis of the Armenian colony, formed around a commemorative temple, had been discovered in Bolgar (Smirnov, 1951, 1958).

Further extensive paleoanthropological research in Bolgar has shown that the population buried at the "Greek Chamber" finds considerable analogies in the multidimensional craniometric space with groups interred in various Muslim mausoleums of the city, as well as near the Small Minaret, at the soil cemeteries of Excavations 45 and 191 (Gazimzyanov, 2000, 2015). This emphasizes the importance of employing

paleogenetic methods as well as new paleoanthropological analyses.

While genomic methods are already widely employed to study ancient peoples, populations, and individuals, genetic data related to the population of Great Bolgar are still lacking. Such information would greatly improve our understanding of its ethnic makeup provided by archaeological and paleoanthropological studies. A particular goal is to assess the contribution of the medieval population to the formation of the present-day peoples of the Volga region. In this study, for the first time, we have applied genomic analysis methods to investigate individuals buried in the territory of Great Bolgar.

Materials and methods

Paleoanthropological material (inventory numbers 8964, 8973, 8977) from the collections of the Research Institute and the Museum of Anthropology named after D.N. Anuchin at Moscow State University was used. The remains originate from burials dating to the first half of the 14th century, which were archaeologically examined by a Joint expedition of the Institute of Material Culture History and the Museum of the Tatar ASSR, led by A.P. Smirnov, in 1945 and 1947 in Bolgar, at the so-called "Greek Chamber". This site was located approximately 150 meters west of the city wall, in an area currently occupied by the Bolgar grain terminal, and is not discernible on the surface today.

Excavations of the ruins of the "Greek Chamber" and the surrounding territory to the south and southeast allowed A.P. Smirnov and A.M. Efimova to identify 113 Christian burials. From an archaeological perspective, these burials are characterized as "homogeneous/uniform", contained within wooden coffins (the wood was found to be decayed, with numerous iron nails discovered), which were placed in rectangular pits measuring 2 × 0.8 meters, with rounded corners, vertical walls, and flat bottoms. The remains were found in an extended position on their backs, with heads facing west, faces upward, hands folded on their chests, and accompanied by a small number of personal items. However, it is noteworthy that some burials with gravestones and their fragments, as well as burials containing rich silk textiles embroidered with gold and silver threads, were also recorded; additionally, a temporal gold ring was discovered. A.P. Smirnov believed that his excavations uncovered a significant portion of this necropolis, estimating that it likely contained no more than 150 interments. Interestingly, while characterizing female ornaments, particularly temporal rings, A.P. Smirnov found analogies for some of them among the Slavic burial mounds of the Smolensk and Tver regions, attributed others to Bulgar prototypes from the 10th to 12th centuries, and associated some with artifacts from burials dated to the 12th to 14th centuries found in the Northern Caucasus (Smirnov, 1951).

Currently, considering our understanding of the multicomponent nature of Bulgar material culture, which is predominantly urban and where many elements lose their "ethno-defining" characteristics upon contact, one might question the significance of these observations. Nevertheless, based on paleoanthropology and paleogenetics, they have the potential to significantly enhance our understanding of the ethnic composition of the medieval population of Bolgar.

In 1948, the collection of skulls from the burials at the "Greek Chamber" was transferred to the Museum of Anthropology at Moscow State University. These specimens were first measured and reported by T.A. Trofimova, and subsequently re-examined by M.M. Gerasimova and published later by I.R. Gazimzyanov (Gazimzyanov, 2000; Trofimova, 1956).

Genetic analysis was performed using fragments of teeth from three individuals with the best-preserved anthropological material. Genomic DNA was extracted from fragments weighing 100–200 mg in clean rooms, where studies of modern materials had not been conducted, using a previously published method (Andreeva et al., 2022). The DNA extract and blank control were tested using the High Sensitivity DNA reagent kit (Agilent) on a Bioanalyzer 2100 (Agilent). Fragmented genomic libraries were prepared according to a single-stranded DNA-based protocol (Gansauge et al., 2017) and sequenced firstly on Illumina MiSeq (paired-end reads mode of 76+76 cycles) and then on Illumina NovaSeq 6000 in single-end read mode of 56 cycles.

We used AdapterRemoval v2 (Schubert et al., 2016) for trimming the adapter sequences of the raw reads. Short reads were mapped using the BWA program (Li, Durbin, 2009) with parameters adapted for short fragments of ancient DNA (Schubert et al., 2012) to the human reference genome (hg19/GRCh37 assembly), and to the mtDNA Cambridge Reference Sequence (NC_012920.1). The authenticity of the ancient DNA was evaluated using the MapDamage2 program (Jónsson et al., 2013). To determine the genetic sex of individuals, we calculated the ratio of reads mapped to the X and Y chromosomes to reads mapped to autosomes; reads with a mapping quality (MQ) greater than 30 were used for the assessment.

We used contamMix (Fu et al., 2013) to estimate the contamination of samples by mtDNA heterozygosity. For male individuals, the contamination level was additionally estimated by X-chromosome heterozygosity (Rasmussen et al., 2011).

Mitochondrial haplogroup was determined using Haplogrep 2 (Weissensteiner et al., 2016). The Y-chromosome haplogroup was determined using the Yhaplo program (Poznik, 2016). Then, in order to clarify the Y-chromosomal lineage, the markers of the corresponding haplogroup and all its derived branches were visually checked using the IGV browser (Robinson et al., 2011). Both Y-chromosomal haplogroup markers presented in the ISOGG database (version 15.73) (International Society…, 2020) and markers from the Yfull database (YFull, 2024) were used. To conduct a phylogeographic analysis of mtDNA using the BLAST service, we selected all mtDNA sequences close to the sequences of the studied samples (identity >99.98 %), as well as samples from the YFull and AmtDB databases (Ehler et al., 2019) belonging to the identified haplogroup, and used them to construct a phylogenetic tree using the mtPhyl program (Eltsov, Volodko, 2016).

For the analysis of genomic markers, pseudohaploid genotypes corresponding to the AADAR panel (version v54.1.p1), which includes 1,240 thousand genetic markers (Mallick et al., 2024), were obtained using the PileupCaller program with the "--randomHaploid" parameter. Population analysis was performed using principal component analysis (PCA). For this purpose, the genotypes of the studied samples from the burials at the "Greek Chamber" were projected onto the genetic variability of 2,775 representatives of 80 present-day European and Caucasian populations from the Human Origin panel (Lazaridis et al., 2016). PCA analysis was performed using smartpca from the EIGENSOFT software package (Patterson et al., 2006). To assess genetic similarity with ancient populations, *f*4-statistics were calculated. Pseudohaploid genotypes for the panel of 1,240 thousand genetic markers for the populations included in the analysis were obtained from the AADR database (version v54.1.p1). For the calculation, the ADMIXTOOLS v.7.0.1 (Patterson et al., 2012) and admixr v.0.9.1 (Petr et al., 2019) software packages were used. The "inbreed=YES" parameter was applied in the calculation. The results were visualized using the R/4.2 package (R Core Team, 2021).

Results

Genomic DNA obtained from the teeth of three individuals (AB188, AB189, AB190) was used for genomic library preparation and sequencing (Table 1, Fig. 1). Bioinformatics analysis of the obtained short reads confirmed the authenticity of the DNA for each sample (Fig. 2). The proportion of reads mapped to the human reference genome ranged from 20 % to 66 %, indicating the high quality of the examined bone material and its suitability for whole-genome analysis (Table 2). According to the results of the assessment of the ratio of the average coverage of sex chromosomes to autosomes, it was observed that two samples (AB188 and AB190) belong to males, while one (AB189), to a female. The genetic sex of all samples coincided with their phenotypic sex, previously determined by biological and anthropological methods.

Based on the sequencing data, complete mitochondrial sequences of the studied individuals were determined, as well as their mitochondrial haplogroups. The mtDNA sequences of all three individuals belong to three different mitochondrial lineages (I5c3, A+152+16362, and H78); therefore, these individuals are not maternally related (Table 3).

Table 1. Archaeological and anthropological data of the samples

Sample ID	Museum ID	Region	Archaeological site, burial number	Anthropological sex
AB188	8964	Tatarstan	Bolgar, "Greek chamber", grave 14	Male
AB189	8973	Tatarstan	Bolgar, "Greek chamber", grave 93	Female
AB190	8977	Tatarstan	Bolgar, "Greek chamber", grave 66	Male

Fig. 1. Anthropological material (teeth) used in the study.

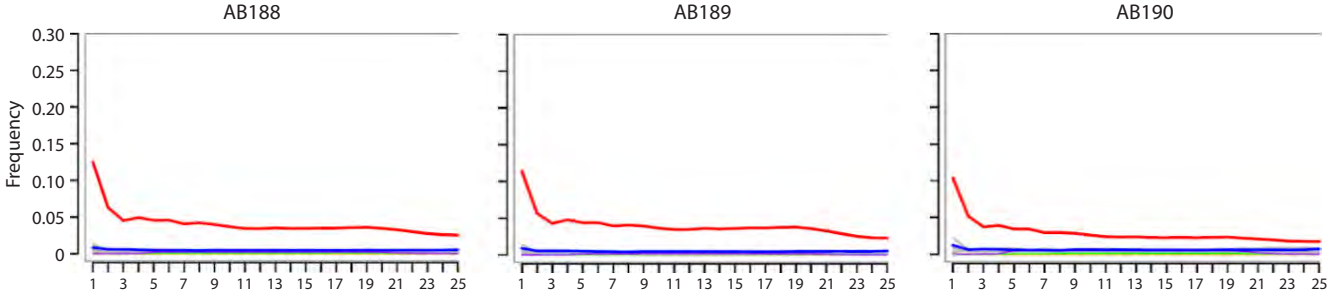


Fig. 2. The profile of nucleotide substitutions for reads mapped to the reference mtDNA sequence, calculated with MapDamage2 (Jónsson et al., 2013). C>T transitions specific to ancient DNA are indicated by a red line. The X axis denotes the nucleotide position from the 5' end of the DNA fragments.

Table 2. DNA sequencing statistics, mtDNA and Y-chromosome haplogroups

Sample ID	Total reads	Mapped reads, %	Contamination, %		Mean coverage		Genetic sex	Haplogroup	
			mtDNA*	X-chromosome**	genome	mtDNA		mtDNA	Y-chromosome
AB188	475,898,786	20.27	2.4	7.2	x0.39	x35.1	XY (Male)	I5c3	R1b1a1b1~ R-L23 ->R-Z2103
AB189	451,469,305	66.21	1.5	–	x1.1	x34.7	XX (Female)	A+152+16362	–
AB190	606,729,041	51.10	2.0	5.2	x1.51	x51.3	XY (Male)	H78	G2a2b1a1~ FGC5083/Y2724

* Based on contamMix (Fu et al., 2013).
** ANGSD was used to determine X-chromosome contamination in men (Korneliussen et al., 2014).

Table 3. Variants of the mtDNA sequences found in the tested samples from the burials at the "Greek Chamber"

Sample ID	Database ID (origin or ethnic group, period)	The smallest number of single nucleotide differences from the test sample	Sample ID	Database ID (origin or ethnic group, period)	The smallest number of single nucleotide differences from the test sample
AB188	RISE408 (Armenia, Iron Age) KJ690072 (Bulgaria) OP642525 (Russia, Chechen) MF362879 (Turkey, Armenian) MK491434 (Turkey, Armenian) MF362904 (Turkey, Armenian)	MF362904 (Turkey, Armenian) – 2	AB190*	FJ999540 (Germany) MK217231 (Assyrian) AY739001 (Italy) MZ920899 (Spain) AY339402 (Finland) AY738987 (Spain) JX153287 (Italy) EU600330 (Israel, Druze) EU600333 (Israel, Druze) MH001823 (Finland) KC763393 (Finland) MZ920486 (Spain) MZ920710 (Spain) MN595895 (Pakistan, Pashtun) ON597638 (Italy) KR858867 (Finland)	7 and more single nucleotide differences
AB189	AP010745 (Japan) EF397559 (Czechia) MH981888 (Paraguay) MF522991 (Pamir) MF523016 (Pamir) HM036549 (Himalayas) MH449268 (Vietnam) AP013161 (Japan) 59 (Russia, Besermyan) 68 (Russia, Udmurt) 95 (Russia, Udmurt)	59 (Russia, Besermyan) 68 (Russia, Udmurt) – 4			

* Several randomly selected mtDNA sequences from those selected by the percentage identity parameter >99.98 % are presented. The mtDNA sequences found are widely distributed across Europe and Western Asia.

For two male individuals, different Y-chromosome haplogroups were identified. The Y chromosome of AB188 belongs to haplogroup R1b1a1b1~ (R-Z2103). Individual AB190 is a carrier of haplogroup G2a2b1a1~ (FGC5083/Y2724). Therefore, the studied males are not related to each other through the paternal line.

A PCA indicated that both male samples (AB188 and AB190) cluster within present-day Caucasian populations, in close proximity to samples from present-day Armenia and Turkey. The female sample AB189 is projected near present-day populations of the Volga-Ural region (Fig. 3).

For testing the potential genetic contribution and similarity to ancient populations, we performed a calculation of the *f*4-statistic of the form *f*4(Test, Mbuti; AB188 and AB190, AB189), where Test represents one of the tested ancestral populations, and the African Mbuti group was used as an outgroup. The results of the analysis (Tables 4, 5, Fig. 4) confirm a greater similarity of individual AB189 with ancient groups that formed the genetic substrate of the contemporary Finno-Ugric population (Russia_Karelia_HG and Russia_Krasnoyarsk_BA) and with modern Siberian groups (the Besermyans, Udmurts, and Nganasans). In contrast, individuals AB188 and AB190 exhibit significantly more shared alleles with the population group from the Kura-Araxes culture of the Bronze Age in Armenia (Armenia_EBA_Kura_Araxes), which has made a substantial contribution to the genetic structure of contemporary Armenians, as well as with modern representatives of present-day Turks, and Iranians (Fig. 4).

Discussion

Previous studies on Bolgar's craniological series have shown that the local Finno-Ugric component is significantly represented in the population. Notably, the male skulls from the burials next to the "Greek Chamber" have frequently been

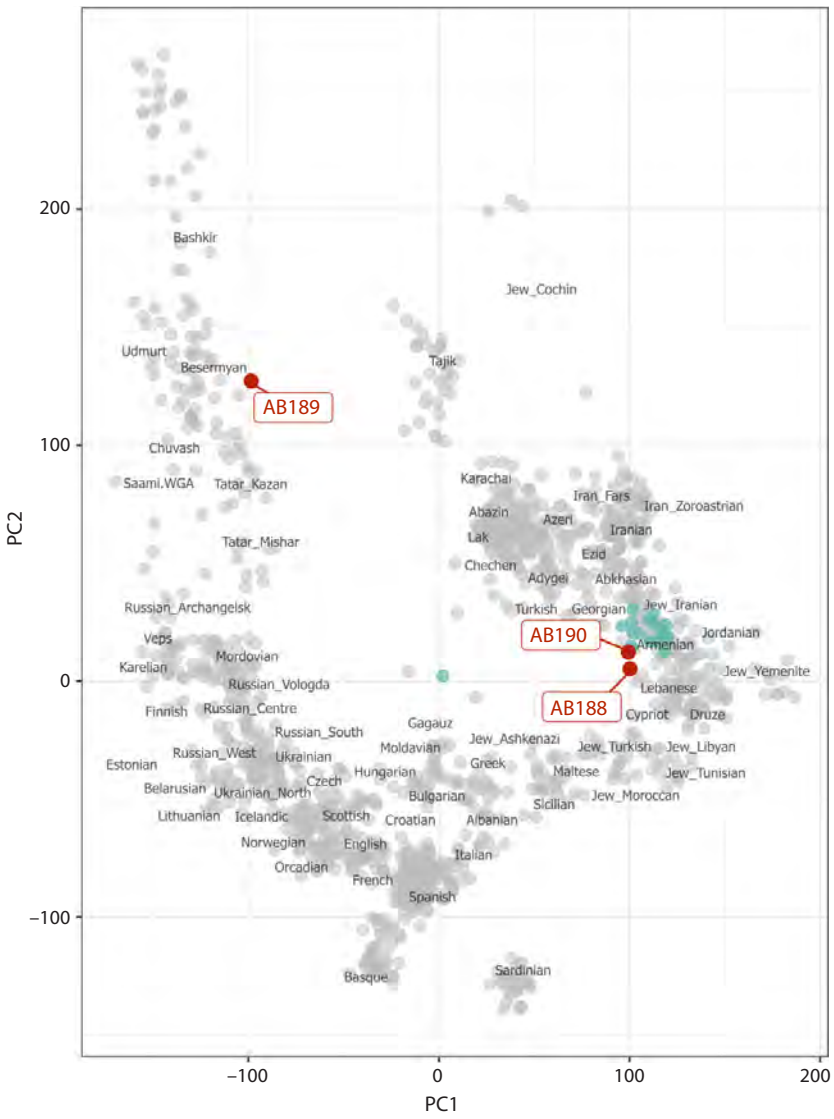


Fig. 3. PCA plot visualizing Volga Bulgar samples (red dots) projected onto the first two (PC1 and PC2) components of genetic variability of present-day individuals from the Human Origin dataset (grey dots); present-day Armenians are marked with green dots.

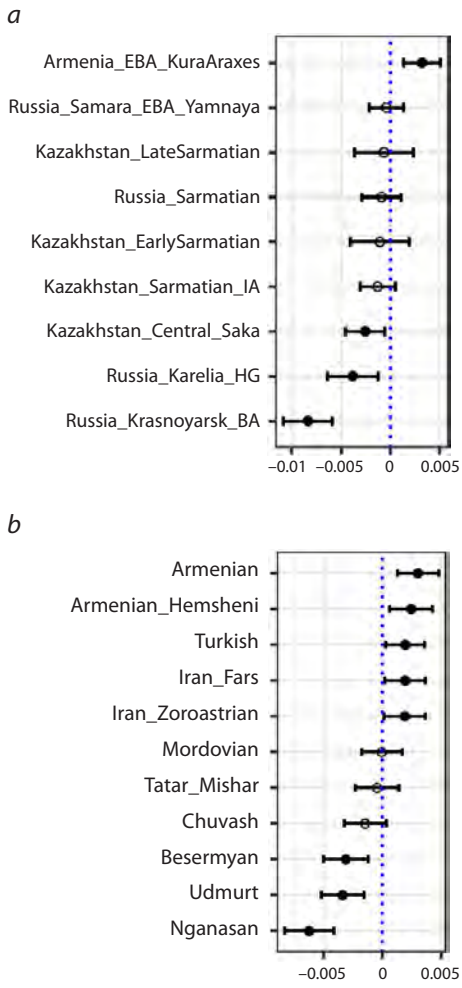


Fig. 4. *f4*-statistics of the form of *f4*(Test, Mbuti; AB188 and AB190, AB189).

Test – ancient (*a*) and present-day (*b*) populations.

Table 4. *f4*-statistics of the form *f4*(W,X;Y,Z) for the ancient samples from the AADR database

W*	X*	Y	Z	*f4*	stderr	Zscore	BABA	ABBA	nsnps
Armenia_EBA_KuraAraxes	Mbuti	AB188_AB190	AB189	0.003216	0.000621	5.179	6,803	6,386	129,387
Russia_Karelia_HG	Mbuti	AB188_AB190	AB189	−0.00382	0.000858	−4.447	6,136	6,612	124,604
Russia_Samara_EBA_Yamnaya	Mbuti	AB188_AB190	AB189	−0.00041	0.000583	−0.699	7,422	7,480	143,943
Russia_Sarmatian	Mbuti	AB188_AB190	AB189	−0.00088	0.000669	−1.314	6,674	6,790	131,859
Kazakhstan_EarlySarmatian	Mbuti	AB188_AB190	AB189	−0.00106	0.001	−1.062	4,006	4,090	79,364
Kazakhstan_LateSarmatian	Mbuti	AB188_AB190	AB189	−0.00065	0.001004	−0.645	4,080	4,132	80,499
Kazakhstan_Sarmatian_IA	Mbuti	AB188_AB190	AB189	−0.00129	0.000594	−2.177	7,136	7,318	140,913
Russia_Krasnoyarsk_BA	Mbuti	AB188_AB190	AB189	−0.00836	0.000836	−9.988	6,951	8,190	148,263

Note. Here and in Table 5: W – ancient samples from the AADR database; X – Mbuti; Y – AB188 and AB190; Z – AB189; *f4* – calculated *f4*-statistic; stderr – standard error; Zscore – calculated Z-score; BABA/ABBA – the number of ABBA/BABA patterns of allele sharing among tested populations; nsnps – allele number.
* The labels for the ancient groups correspond to the AADR database (version v54.1.p1) of 1,240 K genetic markers.

Table 5. *f4*-statistics of the form *f4*(W,X;Y,Z) for the present-day samples from the AADR database

W*	X*	Y	Z	*f4*	stderr	Zscore	BABA	ABBA	nsnps
Iran_Fars.HO	Mbuti	AB188_AB190	AB189	0.001956	0.000573	3.414	3,313	3,173	71,262
Iran_Zoroastrian.HO	Mbuti	AB188_AB190	AB189	0.001924	0.000588	3.271	3,310	3,173	71,262
Armenian_Hemsheni.HO	Mbuti	AB188_AB190	AB189	0.002469	0.000608	4.061	3,366	3,191	70,835
Armenian.HO	Mbuti	AB188_AB190	AB189	0.003041	0.000588	5.168	3,392	3,172	72,211
Turkish.HO	Mbuti	AB188_AB190	AB189	0.001966	0.000556	3.535	3,357	3,215	72,211
Udmurt.HO	Mbuti	AB188_AB190	AB189	–0.00338	0.000607	–5.57	3,178	3,418	70,835
Besermyan.HO	Mbuti	AB188_AB190	AB189	–0.00311	0.000632	–4.922	3,191	3,411	70,835
Chuvash.HO	Mbuti	AB188_AB190	AB189	–0.00147	0.000595	–2.465	3,252	3,358	72,211
Nganasan.HO	Mbuti	AB188_AB190	AB189	–0.00624	0.000697	–8.953	3,043	3,493	72,211

* The labels for the population groups correspond to the Human Origin set of the AADR database (version v54.1.p1) with 600 K genetic markers.

likened to the skulls of present-day Armenians. In contrast, the female skulls exhibited closer alignment with the cranial characteristics typical of the local Finno-Ugric population (Trofimova, 1956; Efimova, 1991). The high-quality whole-genomic data we obtained from all three samples of individuals representing the medieval Bulgarian population enabled us to analyze both uniparental markers (mtDNA and the Y-chromosome) and autosomal markers, thereby providing insights into the probable origins of the studied individuals.

According to the population analysis utilizing autosomal genetic markers (Fig. 3), the two male samples and the female sample (AB189) from grave 93 next to the "Greek Chamber" differ significantly. According to the projection of the first two principal components, the female sample is most similar to modern-day Besermyans, who speak a Finno-Ugric language, and to Chuvash and Kazan Tatars, who speak Turkic languages. All these groups have a significant amount of autochthonous substrate in their genetic makeup. These modern communities are descendants of a Finno-Ugric people that lived in regions that were once part of Volga Bulgaria and the Golden Horde.

The maternal lineage (mtDNA) of individual AB189 belongs to the East Eurasian haplogroup A+152+16362. This mitochondrial haplogroup is distributed mainly in East Asia and among the indigenous population of America. Among the present-day European populations, haplogroup A is most prevalent in Tatars and Bashkirs of the Volga-Ural region, where it accounts for up to 3.6 % (Malyarchuk et al., 2010).

In the databases of present-day and ancient DNA (GenBank, AmtDB, Yfull), we searched for complete mitochondrial sequences that are most similar to the mtDNA sequence of sample AB189 (percent identity >99.94). The analysis of the phylogenetic tree constructed using the mtPhyl program revealed that sample AB189 has a substitution at position 93, which allows it to form a common clade with samples from modern Udmurts and Besermyans (Fig. 5). Thus, the results of the genetic analysis indicate that the woman we studied has a genetic profile similar to that of the contemporary Finno-Ugric population of the Volga-Ural region. Due to the current lack of genomic data for the medieval population of the Volga region, conducting a comparative analysis of the examined individuals with the ancient inhabitants of this area is challenging. Nevertheless, paleoanthropological data suggest similarities between the medieval and modern indigenous populations (Efimova, 1991), which allows us to hypothesize that the woman from burial 93 next to the "Greek Chamber" was a representative of the local Finno-Ugric tribes.

The two male samples are positioned on the PCA plot close to the present-day populations of the eastern Mediterranean (Turks, Jews, Libyans, and Cypriots) and the Caucasus (Georgians, Armenians), and they are quite far from the samples that belong to the Volga-Ural region's populations (Fig. 3).

The mtDNA of male AB188 belongs to haplogroup I5c3. Sequences of contemporary mtDNA samples of this haplogroup, as represented in the GenBank database, have been identified exclusively among individuals from Caucasian populations, predominantly among Armenians. Notably, among ancient samples, haplogroup I5c, which is ancestral to haplogroup I5c3, was also found in an individual who lived in the territory of present-day Armenia and belonged to the Lchashen-Metsamor culture (1209–1009 BCE) (Allentoft et al., 2015). The Y chromosome of male AB188 belongs to haplogroup R1b1a1b1~ (R-Z2103). Due to insufficient genome coverage, we were unable to determine subsequent markers within this Y clade. This haplogroup is part of the larger Y-chromosomal clade R1b, which is widespread in modern populations of Western Europe. However, haplogroup R-Z2103 identified in individual AB188 belongs to the Eastern European lineage R1b-L23, which is most prevalent in the Caucasus, Turkey, and the Ural region, where its frequency reaches up to 10 % (Myres et al., 2011).

According to the YFull database, the largest number of contemporary representatives of haplogroup R-Z2103 originates from Armenia. Some estimates suggest that between 19 and 23 % of the modern population of Armenia belong to various branches of this haplogroup (FamilyTreeDNA; Hovhannisyan

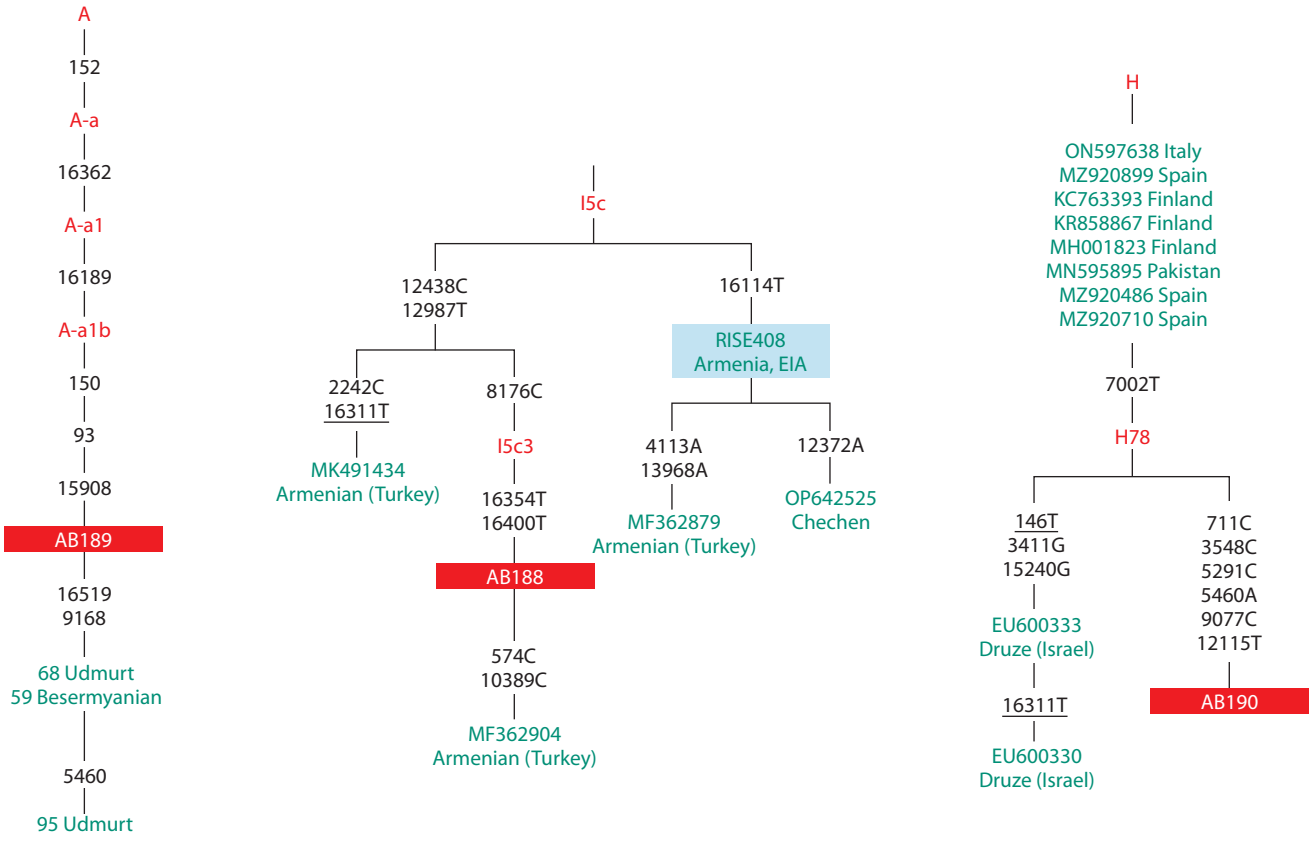


Fig. 5. Fragments of phylogenetic trees of haplogroups A, I5c3, and H78, constructed using the mtPhyl software (Eltsov, Volodko, 2016).

The positions and nucleotide substitutions relative to the reference mtDNA sequence are specified. Mitochondrial haplogroups are highlighted in red. For each sample, its identification number in the database, origin, and ethnic affiliation are provided. The ancient individual is highlighted in blue; EIA – Erly Iron Age. The studied samples are enclosed in red rectangles.

et al., 2025). The results of the genomic analysis of individual AB188 from burial 14, based on data obtained from autosomal markers as well as maternal and paternal lineage markers, indicate that the studied male from the burial site in the "Greek Chamber" in Volga Bulgaria had Armenian origins.

The maternal lineage of the second male individual (AB190) belongs to mitochondrial haplogroup H78. This mitochondrial haplogroup originated from an ancestral branch of haplogroup H due to a single substitution at position 7002. Phylogenetic analysis based on sequences available in databases that are similar to the mtDNA of individual AB190 (percent identity >99.93) revealed that only two sequences in the databases contain the same substitution (7002), both belonging to contemporary representatives of the Druze ethnic group residing in northern Israel (Fig. 5). However, it is important to note that the sequence of AB190 differs from the sequences of these two individuals by at least nine positions in the mtDNA, whereas the differences between mtDNA of AB190 and the sequences of the root haplogroup H, the most common haplogroup in Europe, as well as numerous sequences of sister lineages to H78, amount to only 7–8 substitutions. Thus, drawing conclusions about the likely geographical or ethnic origins of the maternal lineage of individual AB190 is challenging, although one might hypothesize the existence of a common ancient ancestor with the Druze of Israel.

The analysis of the male lineage of this man from Great Bulgaria showed that his Y-chromosomal haplogroup belongs to haplogroup G2a2b1a1~ (FGC5083/Y2724), which is predominantly found in modern populations of the Caucasus and is most widely distributed in contemporary Turkey. Recent data on the most likely origin of modern Armenian groups from ancient Anatolia (Hovhannisyan et al., 2025) indicate a high level of genetic similarity between Armenians and the inhabitants of Turkey. Thus, the second man we studied (AB190, burial 101) was also likely a migrant from the territory of Armenia or Anatolia.

Conclusion

The data we obtained are in line with the historical and archaeological evidence regarding the existence of a segment of the population in Great Bulgaria that was represented by migrants or merchants of Armenian descent. Previously, based on craniological data, a hypothesis was proposed suggesting that Armenians who migrated to Bulgaria took local women as wives (Trofimova, 1956). The results of our analysis partially support the hypothesis of different origins of the men and women buried at the "Greek Chamber". However, to provide evidence for such marital practices, it is necessary to increase the sample size and include potential descendants from such mixed marriages.

References

Allentoft M.E., Sikora M., Sjögren K.-G., Rasmussen S., Rasmussen M., Stenderup J., Damgaard P.B., … Sicheritz-Pontén T., Brunak S., Nielsen R., Kristiansen K., Willerslev E. Population genomics of Bronze Age Eurasia. *Nature.* 2015;522(7555):167-172. doi 10.1038/nature14507

Andreeva T.V., Manakhov A.D., Gusev F.E., Patrikeev A.D., Golovanova L.V., Doronichev V.B., Shirobokov I.G., Rogaev E.I. Genomic analysis of a novel Neanderthal from Mezmaiskaya Cave provides insights into the genetic relationships of Middle Palaeolithic populations. *Sci Rep*. 2022;12(1):13016. doi 10.1038/s41598-022-16164-9

Bulgarica. Time and Space of Bulgarian Civilization: Atlas. Moscow; Kazan: Feoria Publ., 2012 (in Russian)

Efimova S.G. Paleoanthropology of the Volga and Cis-Ural Regions. Moscow: Moscow State University Publ., 1991 (in Russian)

Ehler E., Novotný J., Juras A., Chyleński M., Moravčík O., Pačes J. AmtDB: a database of ancient human mitochondrial genomes. *Nucleic Acids Res.* 2019;47(D1):D29-D32. doi 10.1093/nar/gky843

Eltsov N., Volodko N. MtPhyl: Software tool for human mtDNA analysis and phylogeny reconstruction. 2016 (Available online: http://eltsov.org или Available from: https://sites.google.com/site/mtphyl/home)

FamilyTreeDNA. Available online: https://www.familytreedna.com/

Fu Q., Mittnik A., Johnson P.L.F., Bos K., Lari M., Bollongino R., Sun C., Giemsch L., Schmitz R., Burger J., Ronchitelli A.M., Martini F., Cremonesi R.G., Svoboda J., Bauer P., Caramelli D., Castellano S., Reich D., Pääbo S., Krause J. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol*. 2013;23(7):553-559. doi 10.1016/j.cub.2013.02.044

Gansauge M.-T., Gerber T., Glocke I., Korlevic P., Lippik L., Nagel S., Riehl L.M., Schmidt A., Meyer M. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res*. 2017;45:e79-e79. doi 10.1093/nar/gkx033

Gazimzyanov I.R. Golden Horde and ethnogenetic processes in the Middle Volga region. In: Ethnic Groups of Russia: from the Past to the Present. Ser.: Anthropology. Part II. Moscow, 2000;189-216 (in Russian)

Gazimzyanov I.R. The medieval Bolgar population according to craniology data. Preliminary results according to the materials of 2010–2013 excavations. *Povolzhskaya Arkheologiya = The Volga River Region Archaeology.* 2015;3:112-124 (in Russian)

Gening V. Some issues of periodization of the ethnic history of ancient Bulgarians. In: Early Bulgarians in Eastern Europe. Kazan, 1989; 4-15 (in Russian)

Hovhannisyan A., Delser P.M., Hakobyan A., Jones E.R., Schraiber J.G., Antonosyan M., Margaryan A., Xue Z., Jeon S., Bhak J., Hrechdakian P., Sahakyan H., Saag L., Khachatryan Z., Yepiskoposyan L., Manica A. Demographic history and genetic variation of the Armenian population. *Am J Hum Genet*. 2025;112(1):11-27. doi 10.1016/j.ajhg.2024.10.022

International Society of Genetic Genealogy (ISOGG v15.73), 2020. Available online: https://isogg.org/tree/index.html

Iskhakov D.M., Izmailov I.L. Ethnopolitical History of Tatars in the 6th – First Quarter of 15th Centuries. Kazan: Iman Publ., 2000 (in Russian)

Jónsson H., Ginolhac A., Schubert M., Johnson P.L.F., Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*. 2013;29(13):1682-1684. doi 10.1093/bioinformatics/btt193

Khalikov A.H. The Tatar People and their Ancestors. Kazan, 1989 (in Russian)

Korneliussen T.S., Albrechtsen A., Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014;15(1):356-356. doi 10.1186/s12859-014-0356-4

Lazaridis I., Nadel D., Rollefson G., Merrett D.C., Rohland N., Mallick S., Fernandes D., … Yengo L., Hovhannisyan N.A., Patterson N., Pinhasi R., Reich D. Genomic insights into the origin of farming in the ancient Near East. *Nature.* 2016;536(7617):419-424. doi 10.1038/nature19310

Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760. doi 10.1093/bioinformatics/btp324

Mallick S., Micco A., Mah M., Ringbauer H., Lazaridis I., Olalde I., Patterson N., Reich D. The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci Data*. 2024; 11(1):182. doi 10.1038/s41597-024-03031-7

Malyarchuk B., Derenko M., Denisova G., Kravtsova O. Mitogenomic diversity in Tatars from the Volga-Ural region of Russia. *Mol Biol Evol*. 2010;27(10):2220-2226. doi 10.1093/molbev/msq065

Myres N.M., Rootsi S., Lin A.A., Järve M., King R.J., Kutuev I., Cabrera V.M., Khusnutdinova E.K., Pshenichnov A., Yunusbayev B., Balanovsky O., Balanovska E., Rudan P., Baldovic M., Herrera R.J., Chiaroni J., Di Cristofaro J., Villems R., Kivisild T., Underhill P.A. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur J Hum Genet*. 2011;19(1): 95-101. doi 10.1038/ejhg.2010.146

Patterson N., Price A.L., Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2(12):e190. doi 10.1371/journal.pgen.0020190

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. Ancient admixture in human history. *Genetics.* 2012;192(3):1065-1093. doi 10.1534/genetics.112.145037

Petr M., Vernot B., Kelso J. *admixr* – R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics.* 2019;35(17):3194-3195. doi 10.1093/bioinformatics/btz030

Postnikova N.M. On the anthropology of "Chetyrehugol'nik" medieval burial ground. In: The Volga Region in the Middle Ages. Materials and Research on the Archeology of the USSR. Moscow: Nauka Publ., 1970;164:24-38 (in Russian)

Postnikova N.M. On the anthropology of the population of Volga Bulgaria: anthropological materials from Minaret burial ground, 14th–15th centuries. *Sovetskaya Arkheologiya = Soviet Archeology.* 1973;3:203-211 (in Russian)

Poznik G.D. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*. 2016. doi 10.1101/088716

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2021

Rasmussen M., Guo X., Wang Y., Lohmueller K.E., Rasmussen S., Albrechtsen A., Skotte L., … Sicheritz-Pontén T., Villems R., Nielsen R., Wang J., Willerslev E. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*. 2011; 334(6052):94-98. doi 10.1126/science.1211177

Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26. doi 10.1038/nbt.1754

Schubert M., Ginolhac A., Lindgreen S., Thompson J.F., Al-Rasheid K.A.S., Willerslev E., Krogh A., Orlando L. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*. 2012;13:178. doi 10.1186/1471-2164-13-178

Schubert M., Lindgreen S., Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes.* 2016;9:88. doi 10.1186/s13104-016-1900-2

SequenceTools, 2025. Available online: https://github.com/stschiff/sequenceTools

Sitdikov A.G., Bocharov S.G. Archaeological Investigation of Bolghar. Kazan, 2024 (in Russian)

Smirnov A.P. Volga Bulgarians. Moscow: State Historical Museum, 1951 (in Russian)

Smirnov A.P. Armenian colony of the city of Bolghar. In: Works of the Kuibyshev Archaeological Expedition. Vol. II (Materials and Research on the Archaeology of the USSR). Moscow, 1958;61:330-359 (in Russian)

Smirnov A.P. On the ethnic composition of Volga Bulgaria. In: News in Archaeology. Moscow: Moscow State University Publ., 1972;302-307 (in Russian)

Smirnov A.P. New data on the historical and social topography of the city of Bolghar. In: Cities of the Volga Region in Middle Ages. Moscow: Nauka Publ., 1974;4-13 (in Russian)

Trofimova T.A. Ethnogenesis of Volga Tatars in Light of Anthropological Data. Moscow; Leningrad: Publishing House of the USSR Academy of Sciences, 1949 (in Russian)

Trofimova T.A. Anthropological composition of the population of the city of Bolghar in the 10–15th centuries. In: Proceedings of the Institute of Ethnography of the USSR Academy of Sciences. Vol. 33. Moscow: Publishing House of the USSR Academy of Sciences, 1956;73-145 (in Russian)

Weissensteiner H., Pacher D., Kloss-Brandstätter A., Forer L., Specht G., Bandelt H.J., Kronenberg F., Salas A., Schönherr S. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 2016;44:W58-W63. doi 10.1093/nar/GKW233

YFull, 2024. Available online: https://www.yfull.com/

Conflict of interest. The authors declare no conflict of interest.

Received November 19, 2024. Revised December 26, 2024. Accepted December 28, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):433-439

doi 10.18699/vjgb-25-46

Expression of DNA repair and cell cycle control genes in HPV infection

E.V. Mashkina, V.V. Volchik, E.S. Muzlaeva, E.G. Derevyanchuk

Southern Federal University, Rostov-on-Don, Russia
lenmash@mail.ru

Abstract. One of the main etiological factors in the development of cervical cancer is infection with human papillomavirus (HPV). At the same time, the risk of developing a malignant process increases with an increase in viral load. The aim of this study was to investigate the transcription level of DNA repair and cell cycle control genes in the cervical epithelial cells of women with a clinically significant HPV viral load. The material for the study was DNA and RNA samples isolated from cervical epithelial cells in women. A total of 107 samples were analyzed. 55 women were HPV-positive (with a clinically significant viral load – more than 10^3 HPV genomes per 100 thousand human cells); the control group consisted of 52 HPV-negative women. All women were over 30 years old. The transcription level of the *APEX1*, *ERCC2*, *CHEK2*, *TP53*, *TP73*, *CDKN2A*, *SIRT1* genes was determined using RT-PCR. It was shown that the detection frequency of the *APEX1* and *ERCC2* gene transcripts was increased in the group of women with a clinically significant viral load. The transcription level of all the studied genes did not differ between the control group and the group with clinically significant HPV concentrations. However, the transcription level of the *TP53* and *TP73* genes decreased with increasing viral load. In the control, a correlation between the transcription levels of genes involved in the functioning of the p53 protein was revealed. An increase in viral load during HPV infection is associated with a change in the coexpression of DNA repair and cell cycle control genes.

Key words: human papillomavirus; cell cycle control; gene transcription; coexpression genes

For citation: Mashkina E.V., Volchik V.V., Muzlaeva E.S., Derevyanchuk E.G. Expression of DNA repair and cell cycle control genes in HPV infection. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):433-439. doi 10.18699/vjgb-25-46

Funding. The study was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation within the state assignment framework in the field of scientific activity No. FENW-2023-0018.

Экспрессия генов системы репарации и контроля клеточного цикла при ВПЧ-инфекции

Е.В. Машкина, В.В. Вольчик, Е.С. Музлаева, Е.Г. Деревянчук

Южный федеральный университет, Ростов-на-Дону, Россия
lenmash@mail.ru

Аннотация. Одним из основных этиологических факторов развития рака шейки матки является инфицирование вирусом папилломы человека (ВПЧ). При этом риск развития злокачественного процесса увеличивается с ростом вирусной нагрузки. Целью данного исследования было изучение уровня транскрипции генов репарации ДНК и контроля клеточного цикла в эпителиальных клетках цервикального канала женщин с клинически значимой вирусной нагрузкой при ВПЧ-инфекции. Материалом для исследования послужили образцы ДНК и РНК, выделенные из эпителиальных клеток. Всего было проанализировано 107 образцов: 55 женщин были ВПЧ-положительными (с клинически значимой вирусной нагрузкой – более 10^3 геномов ВПЧ на 100 тыс. клеток человека), контрольную группу составили 52 ВПЧ-отрицательные женщины. Все женщины были старше 30 лет. Уровень транскрипции генов *APEX1*, *ERCC2*, *CHEK2*, *TP53*, *TP73*, *CDKN2A*, *SIRT1* определяли с помощью ОТ-ПЦР. Показано, что частота обнаружения транскриптов генов *APEX1* и *ERCC2* была повышена в группе женщин с клинически значимой вирусной нагрузкой. Уровень транскрипции всех исследованных генов не различался между контрольной группой и группой ВПЧ-инфицированных женщин. Однако установлено, что уровень транскрипции генов *TP53* и *TP73* снижается с ростом вирусной нагрузки. В контроле выявлена корреляция между уровнями транскрипции генов, участвующих в обеспечении функционирования белка p53. Увеличение вирусной нагрузки при инфицировании ВПЧ связано с изменением коэкспрессии генов репарации ДНК и контроля клеточного цикла.

Ключевые слова: вирус папилломы человека; контроль клеточного цикла; транскрипция генов; коэкспрессия генов

Introduction

Human papillomavirus (HPV) is a major risk factor for invasive cervical cancer (Bava et al., 2016). Most cases of HPV infections (70–80 %) are transient and asymptomatic, in which the virus disappears from the body or becomes undetectable within two years of infection (Kim et al., 2012). However, in some cases, infected women develop low- and high-grade squamous intraepithelial lesions or cervical cancer (Chansaenroj et al., 2013). The International Agency for Research on Cancer has documented that a high viral load in HPV infection is a risk factor for cervical cancer (Ylitalo et al., 2000; van der Weele et al., 2016).

The life cycle of human papillomavirus is determined by its influence on intracellular processes, primarily cell cycle control, the activity of factors of the DNA repair system, and factors of the immune system. Normally, differentiation of urogenital tract epithelial cells is accompanied by their exit from the cell cycle. In infected cells, viral DNA is amplified to thousands of copies per cell, which is accompanied by the preservation of the ability of epithelial cells to divide (Longworth, Laimins, 2004; Münger et al., 2004).

Infection of epithelial cells and amplification of viral DNA can lead to damage to the human genome. The response to DNA damage plays a crucial role in maintaining genome stability by coordinating the course of the cell cycle with DNA repair. The main stages of excision repair include recognition of DNA damage, unwinding of the DNA region with a damaged site (ERCC2 is one of the participants), cutting of the DNA chain and cutting out the damaged DNA region (for example, with the participation of APX1), ligation of a new chain. The implementation of repair processes requires stopping the cell cycle, which is carried out due to the functioning of proteins such as CHEK2, CDKN2A, p53, p73, SIRT1.

Aberrant expression of key factors changes the capacity of the repair system, affecting genome stability and integrity, which increases the likelihood of altered/damaged cells preservation and the development of a malignant process (Kushwah et al., 2023). Research data show that the expression level of the DNA repair and cell cycle control systems genes changes in cervical cancer, including that of *APEX1* (Li et al., 2021; Zhang et al., 2023), *ERCC2* (Ye et al., 2012; Bajpai et al., 2013), *TP53* (Ngan et al., 2001; Zhou et al., 2015), *TP73* (Liu et al., 2004; Choi et al., 2007), *CDKN2A* (Hafkamp et al., 2009). However, there are almost no data on the effect of clinically significant viral load in HPV infection on transcription of the repair system genes before the development of severe forms of epithelial cell dysplasia.

The aim of this study was to investigate the expression level of the DNA repair and cell cycle control systems genes at a clinically significant concentration of human papillomavirus in the epithelial cells of the urogenital tract of women.

Materials and methods

The material for the study was DNA and RNA samples isolated from epithelial cells of the urogenital tract of women. Among them, there were 55 women infected with HPV (with a clinically significant viral load – more than 3 lg (lg – common logarithm, 3 lg – 1000 HPV genomes per 100 thousand cells)) and 52 HPV-negative women (control group). The average viral load in infected women was 5.03 ± 0.13 lg (minimum – 3.4 lg, maximum – 8.1 lg). Women with a viral load less than 3 lg were excluded from the study.

Among 55 women infected with HPV, 5.4 % had normal cytological characteristics of the epithelium (NILM); 36.4 % had atypical squamous cells of undetermined significance (ASCUS); low-grade dysplasia (LSIL) was detected in 58.2 % of women.

All women included in the study were over 30 years old. The average age of women in the control group was 37 ± 0.79 years, in the group of women with clinically significant HPV concentrations – 38 ± 1.25 years. All collected samples of epithelial cells scrapings from the urogenital tract of the women were provided by the diagnostic laboratory "Nauka" (Rostov-on-Don, Russia).

The ethnic component was also considered – only women of the Caucasian race were included in the study. In this case, Russians made up 86 %, Armenians, 9 %, other nationalities of the Caucasian race, 5 %.

Informed written consent was obtained from all women. The research received approval from the Bioethics Committee of the Academy of Biology and Biotechnology of the Southern Federal University on March 29, 2016 (Protocol No. 2). All experimental procedures adhere to the standards and ethical guidelines of the World Medical Association (Helsinki Declaration).

Samples of cervical epithelial cells scrapings were used. Total DNA was extracted using the DNA-sorb-AM reagent kit according to protocol (NextBio, Russia). DNA for high-risk HPV types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) was quantified using the protocol AmpliSens-HPV HCR screen-titre-FRT (Interlabservice, Russia) (AmpliSens… Manual, 2018). An assessment of viral load was made based on clinical reports published in the manufacturer's kit manual. Specifically, a viral load less than 3 lg per 10^5 human cells was taken to signify "low clinical significance", while a load greater than 3 lg per 10^5 human cells was equated to a clinically significant probability of dysplasia.

Total RNA isolation was performed using the GeneJET RNA Purification Kit (Thermo Scientific) according to the manufacturer's instructions. The reverse transcription reaction was performed at 45 °C for 50 minutes followed by MMLV-RT inactivation at 92 °C for 5 minutes. The transcription level of the *GAPDH*, *APEX1*, *ERCC2*, *CHEK2*, *TP53*, *TP73*, *CDKN2A*, *SIRT1* genes in epithelial cells of the cervical canal of the women was determined by Real-time PCR using fluorescent gene-specific probes or SYBR Green. The sequence of primers and probes is presented in Table 1. The amplification reaction was performed in two replicates for each sample according to the following program: 94 °C – 10 minutes, 60 °C – 50 seconds, 94 °C – 15 seconds. The last two steps were repeated 40 times.

Statistical analysis. The target mRNA level was normalized to the level of mRNA *GAPDH*. The level of transcription was determined using the ΔCt approach, where ΔCt is the difference in the target and the housekeeping gene threshold cycles. Statistical analysis of data of gene expression was performed by the $2^{-\Delta\Delta Ct}$ method by Livak and Schmittgen (2001). It shows the multiplicity of changes in gene expression level in the samples compared. The average ΔCt values for two groups were compared using Student's *t*-test. Pearson's Rank

Table 1. Sequence of primers and probes used to determine the level of gene transcription

Gene	5′–3′ sequence of primers and probes
GAPDH	F 5′-AGGTCGGAGTCAACGGATTT-3′ R 5′-ATCGCCCCACTTGATTTTGG-3′ FAM-GGCGCCTGGTCACCAGGGCT-BHQ1
APEX1	F 5′-AAAGTTTCTTACGGCATAGGCGAT-3′ R 5′-CTGTTACCAGCACAAACGAGTCA-3′ FAM-ATCACCCGGCCTTCCTGATCATGCTCC-BHQ1
TP53	F 5′-GCCCTCAACAAGATGTTTTGCCAACT-3′ R 5′-CAACCTCCGTCATGTGCTGTGACTG-3′ FAM-TTGATTCCACACCCCCGCCCGGCACC-BHQ1
ERCC2	F 5′-CGAGGCCCACAACATTGACAAC-3′ R 5′-TCTTTGATCCTGAGCACCGTCT-3′ FAM-AACCTCACCCGCCGGACCCTTGACC-BHQ1
CHEK2	F 5′-AGCCCAGCCTTCTACTAGTCGAAA-3′ R 5′-GTTCAAACCACGGAGTTCACAACACA-3′ FAM-TCGGCACCCTCGGCTTCCCCTTCACGG-BHQ1
TP73	F 5′-CACACCATCACCATCCCCAA-3′ R 5′-CCTCCGTGAACTCCTCCTTG-3′ FAM-GGACTTCGGCTTCGACCT-BHQ1
CDKN2A	F 5′-CTTCCTGGACACGCTGGTG-3′ R 5′-ATGGTTACTGCCTCTGGTGC-3′ FAM-GACCTGGCTGAGGAGCTG-BHQ1
SIRT1	F 5′-CTTCACCACCAGATTCTTCAG-3′ R 5′-TTCAGCAATACTTTCAACATTCC-3′

Correlation Coefficient was used to assess whether there were statistical correlations between the expression level and the selected independent variables (the viral load and women's age). GraphPad InStat software (version 3.05) was used for all statistical analyses.

Results

Among 55 women infected with high-risk HPV types, type 16 was detected in 32.7 % of cases. HPV type 18 was the second most frequently detected type (9.1 %). Women infected with either type 31 or type 33 accounted for 7.3 % each. HPV type 51 was detected in three women (5.4 %). In addition, isolated cases of monoinfection with HPV types 45, 52, 58 or 59 were also detected. However, 29.1 % of women were co-infected with two HPV types. No relationship was found between the type of human papillomavirus and the viral load, as well as the transcription level of the studied protein-coding genes.

As a result of the study, it was found that *TP53* and *SIRT1* transcripts are detected in 100 % of the samples; *CHEK2* transcripts are detected in 96 % of the samples. Most samples were characterized by transcription of *TP73* (69 and 73 % in the control and HPV infection groups) and *CDKN2A* (58 and 62 % in the control and HPV infection groups). It was found that in a clinically significant viral load, the part of epithelial cell samples with transcribed *APEX1* and *ERCC2* genes was 78 and 40 %, respectively, which is 20 % higher than in the control group ($p < 0.05$). It should be noted that for the *ERCC2* gene, the transcript detection frequency increased with increasing viral load. So, at an HPV concentration of 3–5 lg, the proportion of samples with *ERCC2* transcripts was 34.4 %, and at a viral load above 5 lg, it was 47.8 %. This may reflect an increase in the intensity of human DNA damage when the HPV concentration in epithelial cells increases. The highest level of transcription was typical of the *CHEK2* and *SIRT1* genes.

The transcription level of all the genes studied in the control and at clinically significant HPV concentrations did not differ ($p > 0.05$) (Fig. 1).

However, the mRNA level of two genes depended on HPV concentrations in the cells of infected women. Thus, the transcript level of the *CHEK2* gene in the cells of women with a viral load of 3–5 lg did not differ from the control. At the same time, at HPV concentrations of more than 5 lg, the relative level of *CHEK2* transcripts was 3 times lower compared to the control ($2^{-\Delta\Delta Ct} = 0.34$, $p = 0.03$). Another dependence was revealed for the *TP73* gene: with a viral load of 3–5 lg HPV genomes, the level of *TP73* transcripts was 2.7 times higher than in the control ($p = 0.03$); with a viral load of more than 5 lg, the level of *TP73* gene transcripts did not differ from the control values (Fig. 2).

A negative correlation was found between the mRNA level of *TP73* or *TP53* and the viral load in HPV infection (Table 2).

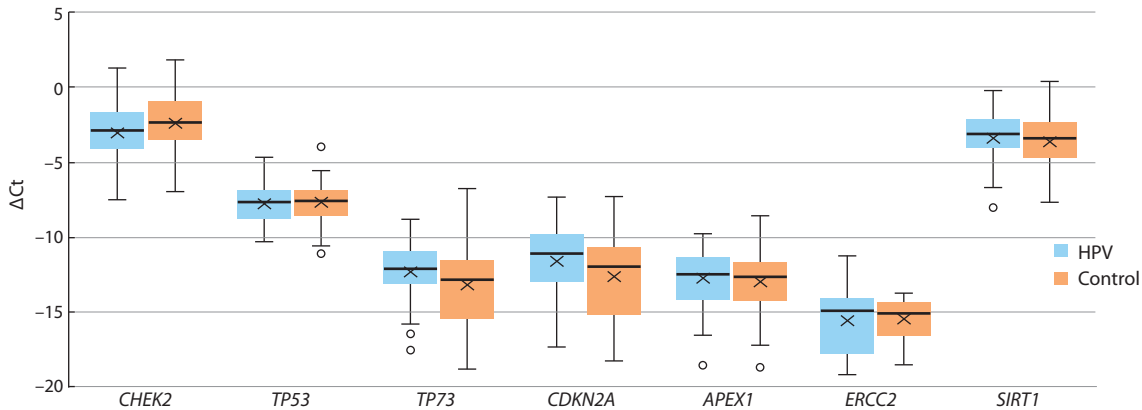


Fig. 1. The level of the genes' transcription in the epithelial cells of the urogenital tract of women relative to the expression of *GAPDH* in the control and at clinically significant HPV concentrations.

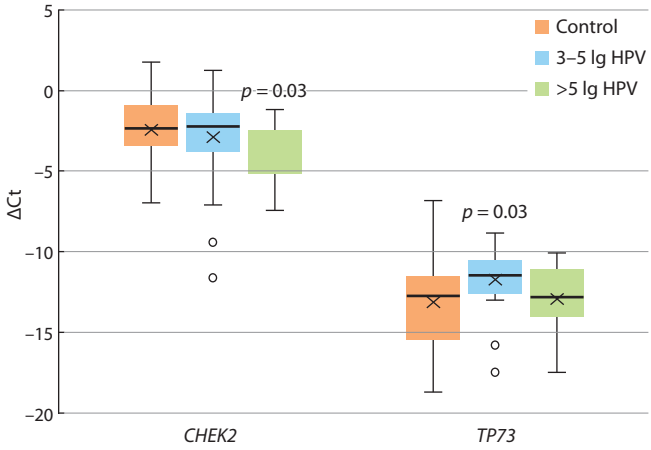


Fig. 2. The level of *CHEK2* and *TP73* transcription in the epithelial cells of the urogenital tract of women relative to the expression of *GAPDH* depending on the viral load in HPV infection (comparison with the corresponding control).

In the control, the transcription level of *CHEK2* correlates with the transcription levels of *SIRT1*, *TP53* and *TP73* (Table 3). The level of *SIRT1* transcripts correlates with the mRNA levels of the *TP53*, *TP73* and *APEX1* genes. *TP53* transcription activity is also associated with *CDKN2A* transcription levels. And the transcription level of *CDKN2A*, in turn, correlates with the transcription level of *ERCC2*.

At a clinically significant viral load in HPV infection, the dependence of the *TP53* and *TP73* genes' transcription and also that of *SIRT1* and *ERCC2* genes has been shown. The transcription of *APEX1* correlated with other genes, except for *CDKN2A* (Table 4). At the same time, unlike the control, the mRNA level of the *CDKN2A* gene does not depend on the *TP53* gene transcription level (Table 4).

Discussion

In the work, the level of gene transcription in the cervical canal cells of women over 30 years old in the control group and in the group with an HPV viral load of more than 10^3 DNA

Table 2. Correlation analysis of the genes' transcription level with viral load in HPV infection

SIRT1	*TP53*	*TP73*	*CDKN2A*	*ERCC2*	*APEX1*	*CHEK2*
$r = 0.15$ $p = 0.27$	**$r = -0.27$** **$p = 0.045$**	**$r = -0.46$** **$p = 0.003$**	$r = -0.13$ $p = 0.47$	$r = -0.05$ $p = 0.81$	$r = 0.23$ $p = 0.14$	$r = -0.15$ $p = 0.29$

Table 3. Correlation of genes' transcription in cervical epithelial cells in the control

Gene	*SIRT1*	*TP53*	*TP73*	*CDKN2A*	*ERCC2*	*APEX1*
CHEK2	**$r = 0.57$** **$p < 0.0001$**	**$r = 0.64$** **$p < 0.0001$**	**$r = 0.48$** **$p = 0.003$**	$r = 0.27$ $p = 0.15$	$r = -0.14$ $p = 0.7$	$r = 0.23$ $p = 0.21$
SIRT1	–	**$r = 0.51$** **$p = 0.0001$**	**$r = 0.54$** **$p = 0.0006$**	$r = 0.34$ $p = 0.07$	$r = -0.12$ $p = 0.74$	**$r = 0.52$** **$p = 0.003$**
TP53	–	–	$r = 0.29$ $p = 0.08$	**$r = 0.45$** **$p = 0.01$**	$r = 0.19$ $p = 0.6$	**$r = 0.39$** **$p = 0.033$**
TP73	–	–	–	$r = 0.38$ $p = 0.07$	$r = 0.19$ $p = 0.6$	$r = 0.29$ $p = 0.15$
CDKN2A	–	–	–	–	**$r = 0.85$** **$p = 0.015$**	$r = 0.15$ $p = 0.53$
ERCC2	–	–	–	–	–	$r = -0.52$ $p = 0.19$

Table 4. Correlation of genes' transcription in cervical epithelial cells in HPV-infection

Gene	*SIRT1*	*TP53*	*TP73*	*CDKN2A*	*ERCC2*	*APEX1*
CHEK2	**$r = 0.35$** **$p = 0.016$**	**$r = 0.43$** **$p = 0.0013$**	**$r = 0.44$** **$p = 0.005$**	$r = 0.28$ $p = 0.12$	$r = 0.27$ $p = 0.26$	**$r = 0.34$** **$p = 0.03$**
SIRT1	–	**$r = 0.42$** **$p = 0.021$**	**$r = 0.37$** **$p = 0.023$**	$r = 0.34$ $p = 0.09$	**$r = 0.53$** **$p = 0.029$**	**$r = 0.53$** **$p = 0.003$**
TP53	–	–	**$r = 0.75$** **$p < 0.0001$**	$r = 0.05$ $p = 0.76$	$r = 0.23$ $p = 0.31$	**$r = 0.63$** **$p < 0.0001$**
TP73	–	–	–	$r = 0.17$ $p = 0.4$	$r = -0.01$ $p = 0.97$	**$r = 0.38$** **$p = 0.03$**
CDKN2A	–	–	–	–	**$r = -0.6$** **$p = 0.018$**	$r = -0.14$ $p = 0.46$
ERCC2	–	–	–	–	–	**$r = 0.47$** **$p = 0.05$**

copies of HPV per 10^5 human cells was investigated. This level of viral load is considered clinically significant and increases the likelihood of dysplastic changes in epithelial cells (AmpliSens… Manual, 2018). The change in the epithelial cell structure is the result of metabolic processes changes in infected cells. The cellular events underlying the transition from normal cervical tissue to cervical dysplasia or cancer at HPV infection remain unexplored. Specifically, it is unknown why some HPV infections persist and progress to cancer, whereas other HPV infections are cleared or precancer tissues return to normal states. In addition, each stage has its own specific HPV-infected epithelial transcription properties.

Cells of the normal cervix highly express multiple tumor suppressors (for example *SLC5A8*, *DERL3*), thereby suppressing cell proliferative, migratory and invasive capacities (Guo et al., 2023).

HPV-positive cells with histological changes consistent with CIN I have approximately 20 % of differentially expressed genes compared to normal keratinocytes. Genes of DNA repair and cell cycle (*ATM*, *ATRX*) are upregulated. Genes of epithelial differentiation and epidermal development are downregulated. In addition, genes involved in the immune response are downregulated (for example *IL-6*, *STAT1*, *IFNβ*) (Templeton, Laimins, 2024).

At HSIL, high cellular motor capacity is manifested, specifically expressing genes related to cell adhesion (*CDH16*, *CDH17* and *VSIG1*) and extracellular matrix degradation. It can promote the expansion of atypical cells in intraepithelial neoplasia progression. *BRCA1*, which plays a crucial role in DNA replication, DNA repair and genomic stability maintenance, is also active in the cells with HSIL. Cancer cells manifest high expression levels of genes related to carcinogenic pathways such as epithelial-to-mesenchymal transition, tumor cell proliferation, migration, invasion and angiogenesis (Guo et al., 2023).

We have evaluated the effect of viral load in HPV infection on the transcription level of the DNA repair and cell cycle control systems genes. A relationship has been revealed between the transcription of the studied genes in uninfected cells. The activity of *CHEK2* transcription is associated with the mRNA level of the *TP53*, *TP73*, and *SIRT1* genes (Table 3, Fig. 3). At the same time, the *TP53* gene transcription is associated with the activity of the *APEX1* and *CDKN2A* genes. The last is co-expressed with the *ERCC2* gene.

The *CHEK2* gene is an oncosuppressor and encodes serine threonine kinase, which reacts to DNA damage and plays a key role in maintaining genome integrity (Bartek et al., 2001). In response to DNA damage, Chk2 kinase is activated, which triggers a protein phosphorylation cascade. The spectrum of phosphorylation substrates includes proteins involved in the cell cycle control, apoptosis and DNA repair, including tumor suppressor p53, cyclin-dependent kinase CDC25C, transcription factors E2F1 and FOXM1, proteins BRCA1 and BRCA2 (Magni et al., 2014; Zannini et al., 2014). In addition, Chk2 participates in the processes of DNA structure modification and cell progression through the cell cycle.

Activation of p53 is also possible under the influence of APEX1. Another protein that participates in the regulation of p53 activity is sirtuin 1 encoded by the *SIRT1* gene (Yang et al., 2015; Chen et al., 2021). Sirtuin 1 (SIRT1) is a deacetylase,

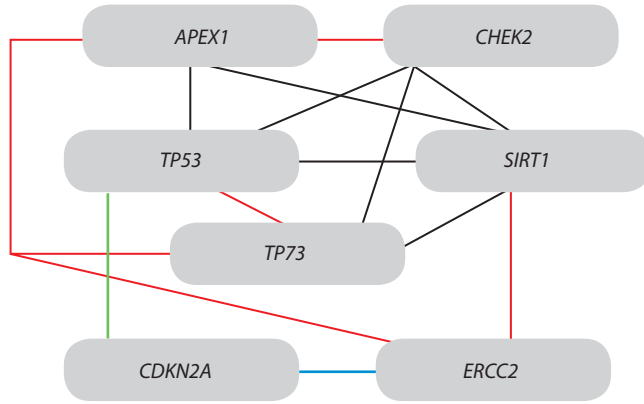


Fig. 3. Coexpression of DNA repair and cell cycle control systems genes in cervical epithelial cells.

The red lines show genes' transcription correlation, which was absent in the control; the blue line shows changing directly to negative correlation; the green line – correlation was present only in the control; the black line – correlation was present in both control and HPV infection.

the activity of which affects gene expression, cell division, and DNA repair.

The $p14^{ARF}$ protein takes part in the regulation of p53 degradation processes. $p14^{ARF}$ is one of the transcription products of the *CDKN2A* gene. This gene encodes two proteins: $p16^{INK4a}$ and $p14^{ARF}$, which are involved in the regulation of the cell cycle, apoptosis, and cell proliferation (Chan et al., 2021). Protein p16 is a cyclin-dependent kinase inhibitor that is targeted on CDK4 and CDK6 and limits their interaction with cyclin D1 (Giacinti, Giordano, 2006). Protein p14 prevents p53 ubiquitination mediated by ubiquitin-protein ligase E3 MDM2. So, it prevents p53 degradation and ensures p21 expression. The p21 protein inhibits the cyclin/CDK complexes formation (Bieging et al., 2014).

Thus, the coexpression of genes involved in the regulation of the p53 protein was revealed in the cervical epithelial cells of women uninfected by human papillomavirus.

In HPV infection, E6 and E7 virus oncoproteins can interact with human proteins of DNA repair and cell cycle control systems. It has been shown that the E6 and E7 proteins interact with CHEK2, which alters its binding efficiency to human DNA and promotes the localization of CHEK2 in HPV DNA replication zones (Gillespie et al., 2012; Bruyere et al., 2023). Also, the E6 protein can bind p53 (Thomas et al., 1999). The decrease in p53 concentration can be partially compensated by p73 activity. p73 activates promoters of several p53-sensitive genes involved in cell cycle control, DNA repair, and apoptosis, and inhibits cell growth in a p53-like manner, inducing apoptosis or cell cycle arrest at the G1 stage (Chellappan et al., 1992; Flores et al., 2002). Our study showed that epithelial cells infected with HPV were characterized by coexpression of the *TP53* and *TP73* genes, which were absent in control (Fig. 3).

However, the E6 protein is also able to interact with p73. At the same time, the more viral proteins there are, the more pronounced the decrease in p73 activity (Park et al., 2001). We revealed a negative correlation between viral load and the *TP53* or *TP73* genes transcription level (Table 2). This may

serve as a molecular basis for increasing risk of epithelial cell dysplasia and cancer development.

We did not detect changes in gene transcription levels during HPV infection, including those of *SIRT1*. However, a few studies have shown that the SIRT1 protein level in human cells increases under the influence of the E7 oncoprotein (Allison et al., 2009; Langsfeld et al., 2015). In infected cells, SIRT1 initiates the assembly of a multi-protein viral DNA replication complex (Langsfeld et al., 2015; Das et al., 2019). In addition, SIRT1 can influence the course of the infectious process by regulating the p53 protein level and activity. SIRT1 can bind, and deacetylate activated p53 (Vaziri et al., 2001). We have previously shown that the intergenic interactions of the *TP53*, *TP73*, *CDKN2A*, and *SIRT1* genes polymorphic loci affect the risk of clinically significant viral load formation in HPV infection, which may be due to the influence on the cell cycle control and apoptosis processes (AlBosale, Mashkina, 2022; Mashkina et al., 2023).

The literature data on the *CDKN2A* gene transcription level in carcinogenesis are contradictory. Overexpression of p16INK4a was established in all cervical intraepithelial neoplasm and invasive cervical cancers (Klaes et al., 2001). In several bioinformatic studies analyzing RNA sequencing data of cervical cancer samples, it was found that *CDKN2A* is a kind of "nodal gene" of the tumor process, since it interacts with various transcription factors, signaling molecules and microRNAs (for example, miR-424-5p and miR-9-5p), and moreover, its overexpression in cervical carcinoma was noted (Zhao et al., 2018). Bioinformatic analysis showed that a change in the *CDKN2A* gene transcription occurs already at the stage of epithelial cell dysplasia (Kulaeva et al., 2024).

However, at the same time, another study of the *CDKN2A* expression in cervical cancer cell lines showed that it was reduced; moreover, the authors concluded that CDKN2A inhibits cell proliferation and invasion in cervical cancer through the lactate dehydrogenase-mediated ACT-mTOR pathway (Luan et al., 2021).

We did not detect changes in the *CDKN2A* transcription level with a significant viral load. However, if in the control, the coexpression of the *TP53* and *CDKN2A* genes transcription levels were revealed, then this dependence disappears in HPV-infected cells. This fact may reflect a change in the p53 protein degradation processes both with the participation of the E6 and E7 proteins, and in a ubiquitin-dependent manner. At the same time, a direct relationship between the mRNA levels of the *CDKN2A* and *ERCC2* genes in the control changes to an inverse relationship at a clinically significant HPV concentration (Table 3). This, on the one hand, may be due to an increase in the number of cells in which the *ERCC2* gene is expressed in HPV infection. But, on the other hand, it may be associated with an increase in the amount of DNA damage, which is accompanied by the associated activity of the repair system proteins.

Conclusion

Thus, an increased frequency of the *APEX1* and *ERCC2* genes transcription at clinically significant HPV concentrations was established. The reverse dependence of the *TP53* and *TP73* genes transcription level on the viral load, as well as a change in the coexpression pattern of the studied genes in HPV infection can lead to a change of cell cycle control and apoptosis. These factors can create conditions for the preservation of HPV-infected cells and contribute to an increased risk of epithelial cells' dysplasia.

References

AlBosale A., Mashkina E. Association between *TP53*, *MDM2* and *NQO1* gene polymorphisms and viral load among women with human papillomavirus. *Vavilov J Genet Breed.* 2022;26(1):59-64. doi 10.18699/VJGB-22-09

Allison S., Jiang M., Milner J. Oncogenic viral protein HPV E7 upregulates the SIRT1 longevity protein in human cervical cancer cells. *Aging* (*Albany NY*). 2009;1(3):316-227. doi 10.18632/aging.100028

AmpliSens HPV HCR screen-titre-FRT PCR kit Instruction Manual. 2018. URL: https://www.pcrdiagnostics.eu/data/machines/hpv-hcr-screen-titre-frt-250321.pdf

Bajpai D., Banerjee A., Pathak S., Jain S.K., Singh N. Decreased expression of DNA repair genes (*XRCC1*, *ERCC1*, *ERCC2*, and *ERCC4*) in squamous intraepithelial lesion and invasive squamous cell carcinoma of the cervix. *Mol Cell Biochem.* 2013;377(1-2):45-53. doi 10.1007/s11010-013-1569-y

Bartek J., Falck J., Lukas J. Chk2 kinase – a busy messenger. *Nat Rev Mol Cell Biol.* 2001;2(12):877-886. doi 10.1038/35103059

Bava S., Thulasidasan A., Sreekanth C., Anto R. Cervical cancer: a comprehensive approach towards extermination. *Ann Med*. 2016;48(3):149-161. doi 10.3109/07853890.2016.1145796

Bieging K., Mello S., Attardi L. Unravelling mechanisms of p53-mediated tumour suppression. *Nat Rev Cancer.* 2014;14(5):359-370. doi 10.1038/nrc3711

Bruyere D., Roncarati P., Lebeau A., Lerho T., Poulain F., Hendrick E., Pilard C., … Hubert P., McBride A., Gillet N., Masson M., Herfs M. Human papillomavirus E6/E7 oncoproteins promote radiotherapy-mediated tumor suppression by globally hijacking host DNA damage repair. *Theranostics.* 2023;13(3):1130-1149. doi 10.7150/thno.78091

Chan S., Chiang J., Ngeow J. *CDKN2A* germline alterations and the relevance of genotype-phenotype associations in cancer predisposition. *Hered Cancer Clin Pract.* 2021;19(1):21. doi 10.1186/s13053-021-00178-x

Chansaenroj J., Theamboonlers A., Junyangdikul P., Swangvaree S., Karalak A., Chinchai T., Poovorawan Y. Polymorphisms in TP53 (rs1042522), p16 (rs11515 and rs3088440) and NQO1 (rs1800566) genes in Thai cervical cancer patients with HPV 16 infection. *Asian Pac J Cancer Prev.* 2013;14(1):341-346. doi 10.7314/apjcp.2013.14.1.341

Chellappan S., Kraus V., Kroger B., Munger K., Howley P., Phelps W., Nevins J. Adenovirus E1A, simian virus 40 tumor antigen, and human papillomavirus E7 protein share the capacity to disrupt the interaction between transcription factor E2F and the retinoblastoma gene product. *Proc Natl Acad Sci USA.* 1992;89(10):4549-4553. doi 10.1073/pnas.89.10.4549

Chen J., Chen H., Pan L. SIRT1 and gynecological malignancies (Review). *Oncol Rep*. 2021;45(4):43. doi 10.3892/or.2021.7994

Choi Y., Bae S., Kim Y., Lee H., Kim Y., Park T., Ro D., Shin J., Shin S., Seo J.-S., Ahn W. Gene expression profiles in squamous cell cervical carcinoma using array-based comparative genomic hybridization analysis. *Int J Gynecol Cancer.* 2007;17(3):687-696. doi 10.1111/j.1525-1438.2007.00834.x

Das D., Bristol M., Smith N., James C., Wang X., Pichierri P., Morgan I. Werner helicase control of Human papillomavirus 16 E1-E2 DNA replication is regulated by SIRT1 deacetylation. *mBio*. 2019;10(2):e00263-19. doi 10.1128/mBio.00263-19

Flores E., Tsai K., Crowley D., Sengupta S., Yang A., McKeon F., Jacks T. p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature.* 2002;416(6880):560-564. doi 10.1038/416560a

Giacinti C., Giordano A. RB and cell cycle progression. *Oncogene*. 2006;25(38):5220-5227. doi 10.1038/sj.onc.1209615

Gillespie K., Mehta K., Laimins L., Moody C. Human papillomaviruses recruit cellular DNA repair and homologous recombination factors to viral replication centers. *J Virol*. 2012;86(17):9520-9526. doi 10.1128/JVI.00247-12

Guo C., Qu X., Tang X., Song Y., Wang J., Hua K., Qiu J. Spatiotemporally deciphering the mysterious mechanism of persistent HPV-induced malignant transition and immune remodelling from HPV-infected normal cervix, precancer to cervical cancer: integrating single-cell RNA-sequencing and spatial transcriptome. *Clin Transl Med*. 2023;13(3):e1219. doi 10.1002/ctm2.1219

Hafkamp H., Mooren J., Claessen S., Klingenberg B., Voogd A., Bot F., Klussmsnn J., Hopman A., Manni J., Kremer B., Ramaekers F., Speel E. $P21^{Cip1/WAF1}$ expression is strongly associated with HPV-positive tonsillar carcinoma and a favorable prognosis. *Mod Pathol*. 2009;22(5):686-698. doi 10.1038/modpathol.2009.23

Kim J., Song S., Jin C., Lee J., Lee N., Lee K. Factors affecting the clearance of high-risk human papillomavirus infection and the progression of cervical intraepithelial neoplasia. *J Int Med Res.* 2012;40(2):486-496. doi 10.1177/147323001204000210

Klaes R., Friedrich T., Spitkovsky D., Ridder R., Rudy W., Petry U., Dallenbach-Hellweg G., Schmidt D., Doeberitz M. Overexpression of $p16^{INK4A}$ as a specific marker for dysplastic and neoplastic epithelial cells of the cervix uteri. *Int J Cancer.* 2001;92(2):276-284. doi 10.1002/ijc.1174

Kulaeva E., Muzlaeva E., Mashkina E. mRNA-lncRNA gene expression signature in HPV-associated neoplasia and cervical cancer. *Vavilov J Genet Breed.* 2024;28(3):342-350. doi 10.18699/vjgb-24-39

Kushwah A., Srivastava K., Banerjee M. Differential expression of DNA repair genes and treatment outcome of chemoradiotherapy (CRT) in cervical cancer. *Gene*. 2023;868:147389. doi 10.1016/j.gene.2023.147389

Langsfeld E., Bodily J., Laimins L. The deacetylase Sirtuin 1 regulates Human papillomavirus replication by modulating histone acetylation and recruitment of DNA damage factors NBS1 and Rad51 to viral genomes. *PLoS Pathog*. 2015;11(9):e1005181. doi 10.1371/journal.ppat.1005181

Li Q., Zhou Z., Duan W., Qian C., Wang S., Deng M., Zi D., Wang J.-M., Mao C.-Y., Song G., Wang D., Westover K., Xu C.-X. Inhibiting the redox function of APE1 suppresses cervical cancer metastasis via disengagement of ZEB1 from E-cadherin, in EMT. *J Exp Clin Cancer Res*. 2021;40(1):220. doi 10.1186/s13046-021-02006-5

Liu S., Leung R., Chan K., Chiu P., Cheung A., Tam K., Ng T.-Y., Wong L.-C., Ngan H. p73 expression is associated with the cellular radiosensitivity in cervical cancer after radiotherapy. *Clin Cancer Res*. 2004;10(10):3309-3316. doi 10.1158/1078-0432.CCR-03-0119

Livak K., Schmittgen T. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *Methods.* 2001;25(4):402-408. doi 10.1006/meth.2001.1262

Longworth M., Laimins L. Pathogenesis of human papillomaviruses in differentiating epithelia. *Microbiol Mol Biol Rev.* 2004;68(2):362-372. doi 10.1128/MMBR.68.2.362-372.2004

Luan Y., Zhang W., Xie J., Mao J. CDKN2A inhibits cell proliferation and invasion in cervical cancer through LDHA-mediated AKT/mTOR pathway. *Clin Transl Oncol.* 2021;23(2):222-228. doi 10.1007/s12094-020-02409-4

Magni M., Ruscica V., Buscemi G., Kim J., Nachimuthu B., Fontanella E., Delia D., Zannini L. Chk2 and REGγ-dependent DBC1 regulation in DNA damage induced apoptosis. *Nucleic Acids Res*. 2014;42(21):13150-13160. doi 10.1093/nar/gku1065

Mashkina E., Muzlaeva E., Volchik V., Shulga A., Derevyanchuk E. Association of *SIRT1*, *CDKN2A*, *TP73* genes polymorphisms with the risk of viral load increase in women infected with human papillomavirus. *Russ J Genet.* 2023;59(Suppl.2):S184-S190. doi 10.1134/S1022795423140089

Münger K., Baldwin A., Edwards K., Hayakawa H., Nguyen C., Owens M., Grace M., Huh K. Mechanisms of human papillomavirus-induced oncogenesis. *J Virol.* 2004;78(21):11451-11460. doi 10.1128/JVI.78.21.11451-11460.2004

Ngan H., Cheung A., Liu S., Cheng D., Ng T., Wong L. Abnormal expression of pan-ras, c-myc and tp53 in squamous cell carcinoma of cervix: correlation with HPV and prognosis. *Oncol Rep.* 2001;8(3):557-561. doi 10.3892/or.8.3.557

Park J., Kim E., Lee J., Sin H., Namkoong S., Um S. Functional inactivation of p73, a homolog of p53 tumor suppressor protein, by human papillomavirus E6 proteins. *Int J Cancer*. 2001;91(6):822-827. doi 10.1002/1097-0215(200002)9999:9999<::aid-ijc1130>3.0.co;2-0

Templeton C., Laimins L. HPV induced R-loop formation represses innate immune gene expression while activating DNA damage repair pathways. *PLoS Pathog.* 2024;20(8):e1012454. doi 10.1371/journal.ppat.1012454

Thomas M., Kalita A., Labrecque S., Pim D., Banks L., Matlashewski G. Two polymorphic variants of wild-type p53 differ biochemically and biologically. *Mol Cell Biol.* 1999;19(2):1092-1100. doi 10.1128/MCB.19.2.1092

van der Weele P., van Logchem E., Wolffs P., van den Broek I., Feltkamp M., de Melker H., Meijer C., Boot H., King A. Correlation between viral load, multiplicity of infection, and persistence of HPV16 and HPV18 infection in a Dutch cohort of young women. *J Clin Virol.* 2016;83:6-11. doi 10.1016/j.jcv.2016.07.020

Vaziri H., Dessain S., Eaton N., Imai S., Frye R., Pandita T., Guarente L., Weinberg R. *hSIR2*^{SIRT1} functions as an NAD-dependent p53 deacetylase. *Cell.* 2001;107(2):149-159. doi 10.1016/s0092-8674(01)00527-x

Yang H., Bi Y., Xue L., Wang J., Lu Y., Zhang Z., Chen X., Chu Y., Yang R., Wang R., Liu G. Multifaceted modulation of SIRT1 in cancer and inflammation. *Crit Rev Oncog.* 2015;20(1-2):49-64. doi 10.1615/critrevoncog.2014012374

Ye F., Jiao J., Zhou C., Cheng Q., Chen H. Nucleotide excision repair gene subunit XPD is highly expressed in cervical squamous cell carcinoma. *Pathol Oncol Res.* 2012;18(4):969-975. doi 10.1007/s12253-012-9527-7

Ylitalo N., Sørensen P., Josefsson A., Magnusson P., Andersen P., Pontén J., Adami H., Gyllensten U., Melbye M. Consistent high viral load of human papillomavirus 16 and risk of cervical carcinoma in situ: a nested case-control study. *Lancet.* 2000;355(9222):2194-2198. doi 10.1016/S0140-6736(00)02402-8

Zannini L., Delia D., Buscemi G. CHK2 kinase in the DNA damage response and beyond. *J Mol Cell Biol.* 2014;6(6):442-457. doi 10.1093/jmcb/mju045

Zhang Q., Yang L., Gao H., Kuang X., Xiao H., Yang C., Cheng Y., Zhang L., Guo X., Zhong Y., Li M. APE1 promotes non-homologous end joining by initiating DNA double-strand break formation and decreasing ubiquitination of artemis following oxidative genotoxic stress. *J Transl Med.* 2023;21(1):183. doi 10.1186/s12967-023-04022-9

Zhao L., Zhang Z., Lou H., Liang J., Yan X., Li W., Xu Y., Ou R. Exploration of the molecular mechanisms of cervical cancer based on mRNA expression profiles and predicted microRNA interactions. *Oncol Lett.* 2018;15(6):8965-8972. doi 10.3892/ol.2018.8494

Zhou R., Wei C., Liu J., Luo Y., Tang W. The prognostic value of p53 expression for patients with cervical cancer: a meta analysis. *Eur J Obstet Gynecol Reprod Biol.* 2015;195:210-213. doi 10.1016/j.ejogrb.2015.10.006

Conflict of interest. The authors declare no conflict of interest.
Received November 1, 2024. Revised January 28, 2025. Accepted March 4, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):440-447

doi 10.18699/vjgb-25-47

Methylation index of the *DLK1* and *MKRN3* genes in precocious puberty

E.A. Sazhenova [1] ✉, O.Yu. Vasilyeva [1], D.A. Fedotov [1], M.B. Kankanam Pathiranage [2], A.D. Lobanov [3], A.Yu. Sambyalova [4], E.E. Khramova [4], L.V. Rychkova [4], S.A. Vasilyev [1, 2], I.N. Lebedev [1, 3]

[1] Scientific Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
[2] Tomsk State University, Tomsk, Russia
[3] Siberian State Medical University, Tomsk, Russia
[4] Scientific Center for Family Health and Human Reproduction Problems, Irkutsk, Russia
✉ elena.sazhenova@medgenetics.ru

Abstract. Precocious puberty (PP, OMIM 176400, 615346) is an autosomal dominant disorder caused by the premature reactivation of the hypothalamic-pituitary-gonadal axis. Genetic, epigenetic, and environmental factors play a decisive role in determining the timing of puberty. In recent years, genetic variants in the *KISS1*, *KISS1R*, *MKRN3*, and *DLK1* genes have been identified as genetic causes of PP. The *MKRN3* and *DLK1* genes are imprinted, and therefore epigenetic modifications, such as DNA methylation, which alter the expression of these genes, can also contribute to the development of PP. The aim of this study is to determine the methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes in girls with a clinical presentation of PP. The methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes was analyzed in a group of 45 girls (age 7.2 ± 1.9 years) with a clinical presentation of PP and a normal karyotype using targeted massive parallel sequencing after sodium bisulfite treatment of DNA. The control group consisted of girls without PP ($n = 15$, age 7.9 ± 1.6 years). No significant age differences were observed between the groups ($p > 0.8$). Analysis of the methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes revealed no significant differences between patients with PP and the control group. However, in the group of patients with isolated adrenarche, an increased methylation index of the imprinting center of the *MKRN3* gene was observed (72 ± 7.84 *vs* 56.92 ± 9.44 %, $p = 0.005$). In the group of patients with central PP, 3.8 % of patients showed a decreased methylation index of the imprinting center of the *DLK1* gene, and 11.5 % of probands had a decreased methylation index of the imprinting center of the *MKRN3* gene. Thus, this study demonstrates that not only genetic variants but also alterations in the methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes can contribute to the development of PP.

Key words: precocious puberty; gonadotropin-releasing hormone (GnRH); hypothalamic-pituitary-gonadal axis (HPG); genomic imprinting; *DLK1*; *MKRN3*

For citation: Sazhenova E.A., Vasilyeva O.Yu., Fedotov D.A., Kankanam Pathiranage M.B., Lobanov A.D., Sambyalova A.Yu., Khramova E.E., Rychkova L.V., Vasilyev S.A., Lebedev I.N. Methylation index of the *DLK1* and *MKRN3* genes in precocious puberty. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):440-447. doi 10.18699/vjgb-25-47

Funding. The study was supported by the Russian Science Foundation, grant number 23-25-00359.

Индекс метилирования генов *DLK1* и *MKRN3* при преждевременном половом созревании

Е.А. Саженова [1] ✉, О.Ю. Васильева [1], Д.А. Федотов [1], М.Б. Канканам Патиранаге [2], А.Д. Лобанов [3], А.Ю. Самбялова [4], Е.Е. Храмова [4], Л.В. Рычкова [4], С.А. Васильев [1, 2], И.Н. Лебедев [1, 3]

[1] Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
[2] Национальный исследовательский Томский государственный университет, Томск, Россия
[3] Сибирский государственный медицинский университет Министерства здравоохранения Российской Федерации, Томск, Россия
[4] Научный центр проблем здоровья семьи и репродукции человека, Иркутск, Россия
✉ elena.sazhenova@medgenetics.ru

Аннотация. Преждевременное половое созревание (ППС, OMIM 176400, 615346) – заболевание, которое вызвано преждевременной реактивацией гипоталамо-гипофизарно-гонадной оси. В определении сроков полового созревания ведущую роль играют генетические, эпигенетические и экологические факторы. В последние годы варианты в генах *KISS1*, *KISS1R*, *MKRN3* и *DLK1* были идентифицированы как генетические причины ППС.

Гены *MKRN3* и *DLK1* являются импринтированными, в связи с чем эпигенетические модификации, такие как метилирование ДНК, изменяющее экспрессию данных генов, также могут рассматриваться в качестве причины ППС. Цель настоящего исследования – определение индекса метилирования центров импринтинга генов *DLK1* и *MKRN3* у девочек с клинической картиной ППС. Анализ индекса метилирования центров импринтинга генов *DLK1* и *MKRN3* проводили в группе из 45 девочек (возраст 7.2 ± 1.9 года) с клинической картиной ППС и нормальным кариотипом методом таргетного массового параллельного секвенирования после обработки ДНК бисульфитом натрия. Контрольная группа состояла из девочек без ППС (n = 15, возраст 7.9 ± 1.6 года). Различий по возрасту между группами не выявлено (p > 0.8). Анализ индекса метилирования центров импринтинга генов *DLK1* и *MKRN3* не показал различий между пациентами с ППС и контрольной группой. Группа пациентов с изолированным адренархе имела повышенный индекс метилирования центра импринтинга гена *MKRN3* (72 ± 7.84 против 56.92 ± 9.44 %, p = 0.005). В группе пациентов с центральным ППС 3.8 % пациентов имели пониженный индекс метилирования центра импринтинга гена *DLK1* и 11.5 % – гена *MKRN3*. Таким образом, показано, что не только генетические варианты, но и нарушение индекса метилирования центров импринтинга генов *DLK1* и *MKRN3* могут быть причиной ППС.

Ключевые слова: преждевременное половое созревание; гонадотропин-рилизинг-гормон (ГнРГ); гипоталамо-гипофизарно-гонадная ось (ГПГ); геномный импринтинг; *DLK1*; *MKRN3*

Introduction

Adolescence is one of the key stages of personality development, characterized by complex changes in the neuroendocrine system and other biological processes that lead to physical and sexual maturation. The appearance of secondary sexual characteristics before the age of 8 in girls and 9 in boys is defined as precocious puberty (PP), with an incidence of approximately 3.7 cases per 10,000 individuals. Recently, the prevalence of this condition has been increasing (Chebotareva et al., 2022; Alghamdi, 2023). According to the International Classification of Diseases (ICD-10), PP is categorized into E22.8 (conditions of pituitary hyperfunction, central precocious puberty) and E30.9 (unspecified disorder of puberty, including isolated thelarche and isolated adrenarche).

The timing of the onset of puberty is significantly influenced by the child's gender, race, genetic predisposition, environmental factors, diet, and socioeconomic status (Sazhenova et al., 2023). For example, obesity and exogenous hormone intake can have adverse effects (Peterkova et al., 2021; Micangeli et al., 2023). However, over the past decade, several genes have been identified that are part of a complex network of inhibitory, activating, and regulatory neuroendocrine factors critical for controlling the onset of puberty. These include *KISS1* (1q32.1) and its receptor *KISS1R* (*GPR54*, 19p13.3), as well as two imprinted genes that are normally expressed only from the paternal allele: *DLK1* (14q32) and *MKRN3* (15q11.2) (Roberts, Kaiser, 2020; Faienza et al., 2022). The primary mechanism of monoallelic expression of imprinted genes is allele-specific DNA methylation, which establishes differential methylation patterns on the two parental chromosomes.

DLK1 (OMIM 176290) encodes an EGF-like membrane-bound protein that belongs to the epidermal growth factor family, participates in the Notch signaling pathway, and regulates preadipocyte differentiation. It is expressed in neuroendocrine tissues, particularly in the adrenal cortex (Gomes et al., 2019; Macedo, Kaiser, 2019). The *MKRN3* gene (OMIM 603856) belongs to the makorin family and plays a role in regulating the onset of puberty by inhibiting the release of GnRH from the hypothalamus, thereby delaying the oncet of puberty (Abreu et al., 2020). The *MKRN3* gene encodes a protein containing a zinc finger RING domain, which is characteristic of most E3 ubiquitin ligases involved in intracellular protein degradation via the ubiquitin-proteasome pathway. The *MKRN3* gene may interact with proteins associated with puberty, insulin signaling, RNA metabolism, and intercellular adhesion (Li C. et al., 2021).

The *DLK1* and *MKRN3* genes, like most imprinted genes, are regulated by imprinting centers. The *DLK1* gene has two imprinting centers: the germline *MEG3/DLK1*:IG-DMR and the secondary *MEG3*:TSS-DMR, which is established after fertilization. *MKRN3* is regulated by the germline imprinting center *SNURF*:TSS-DMR and is directly controlled by the somatic *MKRN3*:TSS-DMR. These imprinting centers are methylated exclusively on the paternal allele in somatic tissues, such as leukocytes and skin fibroblasts (Okae et al., 2014).

Imprinted genes are known to play a crucial role in the development of both the brain and the placenta, the organ responsible for nourishing the embryo. Disruptions in the hypothalamic-pituitary system, which regulates the endocrine activity of the brain during embryonic development, can adversely affect the formation of the fetal endocrine system (Tucci et al., 2019). The fetal pituitary gland produces hormones such as somatotropic, follicle-stimulating, luteinizing, and thyroid-stimulating hormones, which are essential for fetal growth and the regulation of puberty. Dysfunction in the production of these hormones can lead to intrauterine growth restriction or PP after birth (Canton et al., 2021). It is possible that some cases of PP are associated with disruptions in the imprinted state of the *DLK1* and *MKRN3* genes.

The aim of this study is to determine the methylation index of the imprinted regions of the *DLK1* and *MKRN3* gene imprinting centers in girls with a clinical presentation of PP.

Material and methods

The molecular genetic analysis included 45 girls with PP and a normal karyotype, aged 7.2 ± 1.9 years. This group was divided into two subgroups: girls with hyperpituitary function (PP of central origin, ICD-10: E22.8, n = 26, age 7.6 ± 1.4 years) and those with unspecified PP (ICD-10: E30.1, n = 19, age 6.9 ± 0.8 years). The latter subgroup was further divided into girls with isolated thelarche (n = 11, age 7.4 ± 1.2 years) and isolated adrenarche (n = 8, age 6.8 ± 1.4 years). The control group consisted of girls without PP (n = 15, age

7.9 ± 1.6 years). No significant age differences were observed between the groups ($p > 0.8$). The patient cohort was recruited from the Scientific Center of Family Health and Human Reproduction Problems, Irkutsk. The study was conducted in accordance with the principles of the World Medical Association's Helsinki Declaration. The study protocol was approved by the bioethics committee of the Scientific Center of Family Health and Human Reproduction Problems, Irkutsk (Protocol No. 1.1, dated January 12, 2023). Informed consent for participation in the study and DNA analysis was obtained from the parents or legal guardians of all participants.

Description of patient subgroups:

- Girls with the isosexual gonadotropin-dependent form of PP, under 8 years old, exhibiting accelerated physical development (height SDS +1 or more), with their sexual development corresponding to Tanner stages 2–4, levels of pituitary gonadotropic hormones corresponding to pubertal values, and a positive buserelin test. Additionally, they have enlarged mammary glands and uterus, as confirmed by ultrasound, and their biological age does not match their chronological (passport) age.
- Girls with isolated enlargement of the mammary glands (thelarche), under 8 years old, with either accelerated or normal physical development (height SDS +1 or more), advanced sexual development corresponding to Tanner stage 2, levels of pituitary gonadotropic hormones corresponding to prepubertal values, and a negative buserelin test. Additionally, they exhibit enlarged mammary glands and uterus, which is confirmed by ultrasound.
- Girls with isolated adrenarche, under 8 years old, with either accelerated or normal physical development (height SDS +1 or more), advanced sexual development corresponding to Tanner stage 2–3, levels of pituitary gonadotropic hormones corresponding to prepubertal values, and a negative buserelin test. Additionally, they exhibit enlarged mammary glands and uterus, which is confirmed by ultrasound.

All probands underwent standard cytogenetic analysis, which showed a normal karyotype in all cases. Karyotyping was performed using a research-grade microscope AxioImager (Carl Zeiss, Germany).

Genomic DNA was isolated from venous blood by phenol-chloroform extraction. Bisulfite modification of DNA was performed using the EZ DNA Methylation-Direct Kit (Zymo Research, USA) according to the manufacturer's protocol. During bisulfite conversion, unmethylated cytosine is modified to uracil, which is replaced by thymine during further PCR, and methylated cytosine is not modified. Methylation index analysis was performed using targeted bisulfite massive parallel sequencing.

To create libraries, specially designed oligonucleotide primers were used that allow amplification of target genome regions from bisulfite-converted DNA. Primers were selected for imprinted regions containing CpG dinucleotides of the *MEG3/DLK1*:IG-DMR and *MKRN3*:TSS-DMR imprinting centers, which control the expression of the *DLK1* and *MKRN3* genes, respectively. The UCSC genome browser (University of California, Santa Cruz), which contains information on genome sequences (GRCh38), was used to obtain the nucleotide sequence. The obtained nucleotide sequence was then used to select primers using the MethPrimer bioinformatics program (Li L.C., Dahiya, 2002). Vector NTI Advance 11.5 was used to test the thermodynamic properties of the primers.

The imprinted *DLK1* gene is located on chromosome 14, at locus 14q32.2. Expression of this gene is regulated by the imprinting center *MEG3/DLK1*:IG-DMR, position 100,809,090–100,811,721 (GRCh38). This genomic region contains 52 CpG dinucleotides, changes in the methylation index of which can affect the expression of this gene. Expression of the imprinted *MKRN3* gene, located on chromosome 15, at locus 15q11.2, is regulated by the imprinting center *MKRN3*:TSS-DMR, position 23,561,939–23,567,348 (GRCh38). This genomic region contains 26 CpG dinucleotides, changes in the methylation index of which affect the expression of this gene (see the Table).

Amplification of target fragments was carried out using the HS-Taq PCR kit (2×) (Biolabmix, Russia) according to the manufacturer's protocol with the following PCR conditions: 95 °C for 5 min; 36 cycles: 95 °C for 20 s, 66 °C for 30 s, 72 °C for 40 s. The concentration of the target fragments was determined using a Qubit 4.0 fluorimeter (Thermo Fisher Scientific, USA). The reaction products were purified using Sephadex G50 solution (Sigma, USA). Targeted bisulfite massive parallel sequencing was performed on a MiSeq device (Illumina, USA) using a Micro kit (2x150).

The quality of the reads was evaluated using FastQC v0.11.8, after which the remaining adapter sequences and low-quality reads were trimmed using Trim-Galore. The reads were then mapped to bisulfite-converted target sequences using the bwa-meth tool (v0.2.2) with default parameters. Methylation data in the context of CpG were extracted from the resulting BAM files using the MethylDackel tool. The results were presented as the methylation index, which is the ratio of the number of cytosines to the total number of cytosines and thymines in a specific CpG site. In addition, the average methylation index was calculated along all target sites.

A limitation of the targeted bisulfite massive parallel sequencing used in this study is the impossibility of differentiating hypomethylation of the identified imprinting centers from uniparental disomy of chromosomes 14 and 15, as well as from microdeletions in these regions. Therefore, in case of detection of a decrease in the methylation index in the imprinting centers of the *DLK1* and *MKRN3* genes, real-time PCR was performed for the intergenic locus 14q32.3 and the *NIPAI* gene (15q11.2), respectively, to exclude deletion variants in these regions.

Statistical analysis was performed using the Statistica 10.0 software package (StatSoft, USA). The Mann–Whitney rank test was used to compare the methylation index between groups of samples. The differences were considered statistically significant at $p < 0.05$.

The study was conducted using the equipment of the center for collective use "Medical Genomics" of the Tomsk National Research Medical Center of the Russian Academy of Sciences.

Results

Analysis of the methylation index of 52 CpG dinucleotides of the imprinting center of the *DLK1* gene (*MEG3/DLK1*:IG-DMR) in the control group showed that three loci (DLK1_3_520, DLK1_4_422, and DLK1_5_509) are hypermethylated, which is atypical for imprinted genes

Sequences of the oligonucleotide primers used for analysis of the imprinting centers of the *DLK1* and *MKRN3* gene

Primer name	Primer sequence, 5′ – 3′	Position in genome (GRCh38)	Length of product, bp	Number of CpG dinucleotides
		MEG3/DLK1:IG-DMR		
D1R1	GTTGTTTTGATTTGTTAGGTT	chr100809428–100810087	659	11
D1F1	ACCTTATCCCACATAAAATA			
D1F2	ATTTTTGATTTGTAGTTGGG	chr100809940–100810497	557	13
D1R2	CCTATCTTACTCTTCTTAAAAAAC			
D1F3	GATAAGGTAGGATAAGAAAAGTA	chr100810080–100810720	640	14
D1R3	CCAAAATTCAATAACTCAAACTC			
D1F4	AGGTTTGAGTTTGAGTTATT	chr100810692–100811466	655	22
D1R4	ACAATTTAACAACAACTTTCCTC			
D1F5	GTAGTTTTTAGTAGTGTAGT	chr100811196–100811815	619	13
D1R5	CTACACCTTTTAATCCAAAAA			
Total		2,387 bp		52*
		MKRN3:TSS-DMR		
M3F1	TAGATGGGATAAAAGAAGGTAAT	chr23561945–23562513	568	22
M3R1	CCTCAAAAACTAAAACCTAA			
M3F2	TTTTTTTGTGTTAGGAGATGAGA	chr23562966–23563522	556	2
M3R2	CTTTCTCCTTTTCATAATTACCA			
M3F3	TTAGAAAAGAAGGTATAGTTGAG	chr23562528–23563130	602	2
M3R3	ACCCCATAAAATTCTTAAAC			
Total		1,577 bp		25*

* Taking into account the overlap of the analyzed regions.

(Fig. 1*a*). When superimposing the methylation index of the control group and the group of probands with PP for these CpG dinucleotides, it is also evident that these loci have an increased methylation index and do not differ from the control group (Fig. 1*a*). Therefore, these loci were not included in the present study. In contrast to the *DLK1* imprinting center, only one of the 25 CpG dinucleotides (MKRN3_3_296) of the *MKRN3* gene imprinting center (*MKRN3*:TSS-DMR) in the control group was atypically hypermethylated (Fig. 1*b*). When comparing the methylation of the control group and probands with PP, an increased methylation index was also observed for this locus (Fig. 1*b*). Therefore, this locus was also not included in further analysis.

The average methylation index of the *DLK1* imprinting center in probands with PP was not significantly different from the control group (51.46 ± 9.59 % *vs* 58.44 ± 7.15 %, respectively, $p = 0.058$). Although a slight decrease in this index was noted, it was not statistically significant (Fig. 1*a*). Girls with PP were then divided into three subgroups based on their symptoms: central PP, isolated thelarche, and isolated adrenarche. Pairwise comparisons between these subgroups also failed to reveal significant differences in the methylation level of this gene (Fig. 2*a*): 58.44 ± 7.15 % in the control group, 53.32 ± 8.56 % in the group with central PP, 57.40 ± 9.31 %, with isolated thelarche, and 59.74 ± 2.05 %, with isolated adrenarche ($p > 0.05$). Comparison of the methylation index of each patient with PP and the control group revealed that only one patient with central PP showed a significant decrease in methylation (31.16 % *vs* 58.44 ± 7.15 %, $p < 0.01$). Consequently, in the PP group, there is a decrease in methylation levels in the imprinting control region of the *DLK1* gene in 2.22 % (1/45) of patients, and in girls with central PP, this decrease occurs in 3.84 % (1/26) of cases.

Analysis of the methylation index of the imprinting center of the *MKRN3* gene showed no significant differences between patients with PP and the control group (53.71 ± 10.30 % *vs* 56.92 ± 9.26 %, $p = 0.71$). However, when the group with PP was divided into three subgroups (central PP, isolated thelarche, and isolated adrenarche), there was a significant increase in the methylation index in the adrenarche group compared to the control (72.00 ± 7.84 % *vs* 56.92 ± 9.44 %, $p < 0.005$) (Fig. 2*b*). Comparison of the methylation level of each patient with the control group revealed that three patients with central PP had reduced levels (34.09, 29.01, and 32.08 %, compared to 56.92 ± 9.44 % in the control, $p < 0.01$). Therefore, 6.66 % of patients (3/45) in the PP group and 11.53 % of patients in the central PP group (3/26) had reduced methylation levels of the *MKRN3* imprinting center.

The decrease in the methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes may indicate both uniparental disomy of chromosomes 14 and 15 of maternal origin and deletions of these regions on chromosomes of pa-

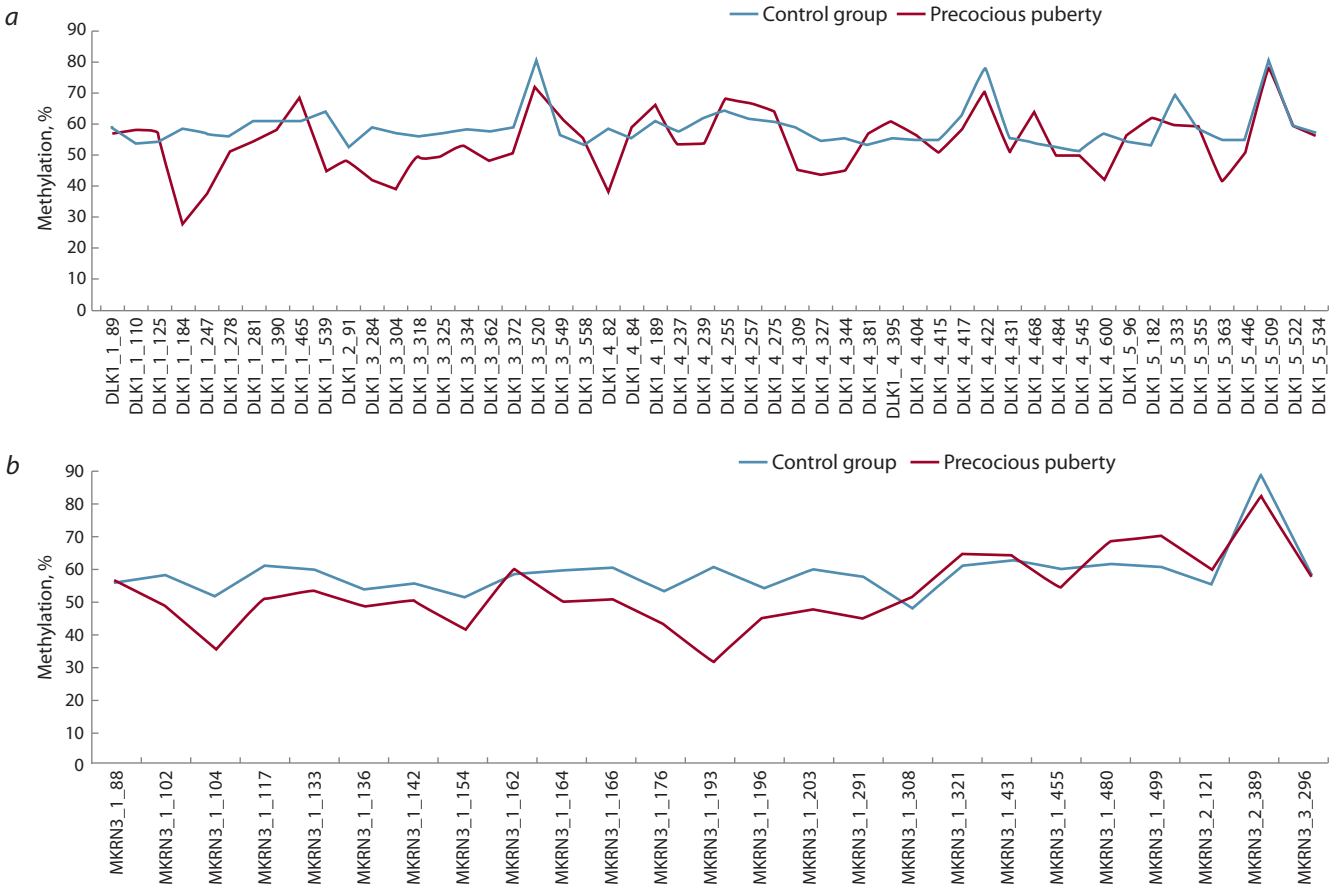


Fig. 1. Methylation index of CpG dinucleotides of the imprinting centers of the *DLK1* (*a*) and *MKRN3* (*b*) genes.

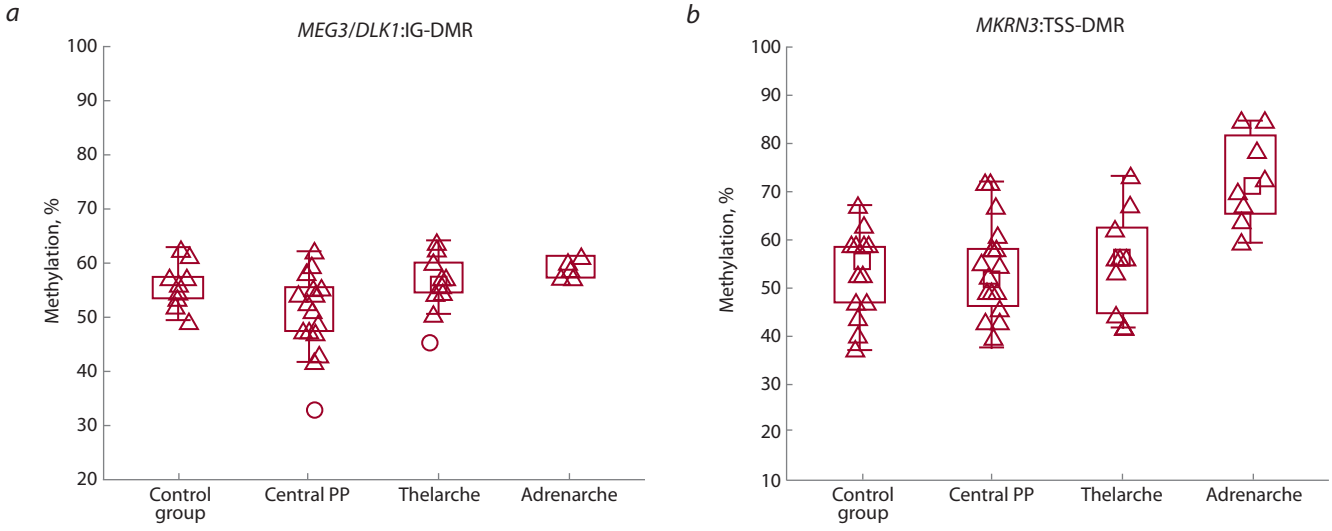


Fig. 2. Pairwise comparison of the mean methylation index of the *DLK1* and *MKRN3* imprinted genes between the groups.

ternal origin. Deletion variants that included the imprinting centers of the *DLK1* and *MKRN3* genes were excluded by real-time PCR. Uniparental disomy for chromosomes 14 and 15 was not excluded. At the same time, uniparental disomy of chromosomes 14 and 15 is a very rare event (1:50,000 people for chromosome 15) (Butler, 2020).

Thus, in the present study, the analysis of the methylation index of the imprinting centers of the *DLK1* and *MKRN3* genes between patients with PP and the control group showed no significant differences between the two groups. However, the group of patients with isolated adrenarche had a significantly increased methylation index of the *MKRN3* imprinting center

(72.00 ± 7.84 % *vs* 56.92 ± 9.44 %, $p < 0.005$). In the group of patients with central PP, 3.8 % of patients had a decreased methylation index of the *DLK1* gene, and 11.5 %, that of the *MKRN3* gene. The presence of epimutations simultaneously in two loci was not detected in any proband.

Discussion

In recent years, there has been an increasing understanding of how epigenetic factors influence the timing of puberty. The current study found a statistically significant reduction in the methylation level of the imprinting centers for the *DLK1* and *MKRN3* genes among a group of girls with central PP. This reduction was observed in 3.84 % of patients for *DLK1*, and 11.53 %, for *MKRN3*. It is worth noting that a decrease in methylation level (hypomethylation) at the *DLK1* imprinting site has also been reported in 8.8 % of individuals with Temple syndrome (Narusawa et al., 2024), a disorder characterized by pre- and postnatal growth retardation, hypotonia, feeding difficulties, delayed motor development, joint laxity, trunk obesity, and facial dysmorphisms (e. g., broad forehead, short nose, micrognathia). In some cases, an early onset of puberty has also been reported (Kagami et al., 2019).

The core clinical features of Temple syndrome are similar to those of Russell–Silver (OMIM 180860) and Prader–Willi (OMIM 176270) syndromes (Kagami et al., 2015; Abi Habib et al., 2019). In the present study, a decrease in the methylation index of the *DLK1* gene imprinting center was found in one proband with PP, who also had obesity, pericallosal lipoma, and hypogenesis of the corpus callosum. In addition to PP and obesity, the clinical picture of the patient did not correspond to the phenotype of Temple syndrome. According to the literature, Canton et al. (2020) showed a decrease in the methylation index of the *DLK1* imprinting center in 2.5 % (3/120) of patients with PP, prenatal and postnatal growth retardation, hypotonia, and feeding difficulties. Further study of these cases revealed the presence of uniparental disomy of chromosome 14 of maternal origin in two of them.

Hypomethylation of the imprinting center of the *MKRN3* gene may be associated with dysfunction of other genes that are normally imprinted and expressed only from the paternal chromosome. These genes include *MAGEL2*, *NECDIN*, *NPAP1*/*C15orf2*, *SNURF-SNRPN*, *MKRN3-AS*/*ZNF127-AS*, *UBE3A-AS*, *IPW*, *SNORD116*, *SNORD115*, *SNORD64*, *SNORD107*, *SNORD108*, *SNORD109A*, and *SNORD109B*.

Uniparental disomy of the maternal chromosome 15 or deletion of the paternal copy of these genes can cause Prader–Willi syndrome. The clinical picture of this syndrome includes muscular hypotonia, obesity, hyperphagia, short stature, hypogonadism, and mental retardation of varying severity. Rarely, this syndrome can be accompanied by PP (Nicoara et al., 2023). Patients with hypomethylation of the imprinting center of the *MKRN3* gene, as identified in this study, had central PP, but without the clinical features of Prader–Willi syndrome. Deletion of the 15q11-q13 region was excluded in these patients. If they had uniparental disomy of this region from the mother, it would certainly lead to the clinical picture of Prader–Willi syndrome. Therefore, it seems that these patients likely only had hypomethylation of the *MKRN3* imprinting center.

An increase in the methylation index of the imprinting center of the *MKRN3* gene has also been shown in the group of probands with isolated adrenarche. Premature adrenarche in children is not caused by premature activation of the hypothalamic-pituitary-gonadal axis, but it is associated with excessive activation of the reticular zone of the adrenal glands, which is the source of dehydroepiandrosterone (DHEA). DHEA is converted into testosterone and dihydrotestosterone in peripheral tissues. During intrauterine development, the adrenal glands start to produce DHEA, which is used to produce placental estriol. A timely increase in androgen levels is necessary for puberty, promoting growth and strengthening of bone tissue, as well as stimulating the process of red blood cell production. Moderate levels of DHEA activate the development of the prefrontal cortex in the brain, providing neuroprotective and neurotrophic effects, and regulate the function of GABA receptors, which are the main inhibitory neurotransmitters in the central nervous system.

Hypermethylation of the *MKRN3* imprinting center leads to increased expression of the gene, resulting in an increased amount of protein produced. MKRN3 then ubiquitinates various factors to suppress the production of gonadotropin-releasing hormone 1 (*GNRH1*). Ubiquitination of the transcriptional repressor MBD3 inhibits its binding to both the *GNRH1* promoter and DNA TET2 demethylase, resulting in epigenetic silencing of *GNRH1* transcription. Additionally, MKRN3 mediates the ubiquitination of poly(A)-binding proteins, such as PABPC1, PABPC3, and PABPC4. This reduces their binding to the poly(A) tail of target mRNA, including *GNRH1* mRNA. This affects the stability and translation of these mRNAs (Li C. et al., 2020; Fanis et al., 2022). The *MKRN3* expression is also found in the adrenal glands, although its role in activating adrenarche remains unknown.

Genome-wide, exome-wide, and methylome-wide association studies on the pathogenetic heterogeneity of PP have shown that genetic variants in the *KISS1R*, *KISS1*, *MKRN3*, and *DLK1* genes occur in approximately 10 % of sporadic cases and 27 % of familial cases. Copy number variations (7q11.23 deletions, Xp22.33 deletions, 1p31.3 duplications) are found in 4 % of patients. Epigenetic abnormalities at the imprinting centers of the *DLK1* and *MKRN3* genes are rare, as reported by Canton et al. (2020). Whole exome sequencing has also identified rare, *de novo* variants that cause loss of gene function in the dominant state. These include the *TNRC6B* (22q13.1), *AREL1* (14q24.3), *PROKR2* (20p12.3), and *LIN28B* (6q16.3) genes, although their roles in the disease process are not fully understood (Shim et al., 2022).

Methylome analysis of this pathology has shown the presence of hypomethylation in the promoter region of the *ZFP57* gene (6p22.1) (Bessa et al., 2018). *ZFP57* contains a KRAB domain, which is a transcriptional repressor, and different genetic variants of this gene have been linked to multi-locus imprinting disturbance (MLID), such as transient neonatal diabetes mellitus (OMIM 601410). This condition is accompanied by hypomethylation of several other imprinted genes, including *PLAGL1*, *GRB10*, and *PEG3* (Monteagudo-Sánchez et al., 2020; Mackay et al., 2024). The expression of the *ZFP57* gene in the hypothalamus of female rhesus macaques has been shown to increase during peripubertal development. This indicates an enhanced repression of downstream target genes of *ZFP57* (Bessa et al., 2018). The increased expression of *ZFP57* has been observed in the hypothalamus of mature

female monkeys. This indicates that this gene may play a role in suppressing the activity of transcription repressors involved in puberty, such as the Polycomb complex (Lomniczi et al., 2015).

Thus, it has been shown that not only genetic variations, but also a disruption of the methylation pattern of the imprinting regions of the *DLK1* and *MKRN3* genes, can be a cause of PP.

Conclusion

Puberty is a multifactorial process with a triggering role of genetic and epigenetic factors. In this study, no significant changes in methylation of the imprinting centers of the *DLK1* and *MKRN3* genes were shown between patients with the clinical picture of PP compared to the control group. Despite this, it was found that a group of patients with isolated adrenarche had an increased methylation index of the imprinting center of the *MKRN3* gene. Among the girls with central PP, 3.8 % had a decreased methylation index for the imprinting centers of the *DLK1* and 11.5 % of the *MKRN3* genes, and the total contribution from methylation disorders in these genes was 15.3 %. No probands had epimutations in both of these loci at the same time.

Therefore, it was demonstrated that not only genetic, but also epigenetic changes may be responsible for PP through methylation disorders in the imprinting regions of the *DLK1* and *MKRN3* genes. However, there are some limitations to the conclusions drawn, and there may be other epigenetic factors that can also influence the formation of PP, such as epimutations in other imprinted genes or mutations in the imprinting centers that control the expression of imprinted genes. Nevertheless, this study shows the epigenetic role of the imprinted genes *DLK1* and *MKRN3* in the development of PP.

References

Abi Habib W., Brioude F., Azzi S., Rossignol S., Linglart A., Sobrier M.-L., Giabicani É., Steunou V., Harbison M.D., Le Bouc Y., Netchine I. Transcriptional profiling at the *DLK1/MEG3* domain explains clinical overlap between imprinting disorders. *Sci Adv*. 2019;5:e9425. doi 10.1126/sciadv.aau9425

Abreu A.P., Toro C.A., Song Y.B., Navarro V.M., Bosch M.A., Eren A., Liang J.N., Carroll R.S., Latronico A.C., Ronnekleiv O.K. *MKRN3* inhibits the reproductive axis through actions in kisspeptin-expressing neurons. *J Clin Invest*. 2020;130(8):4486-4500. doi 10.1172/JCI136564

Alghamdi A. Precocious puberty: types, pathogenesis and updated management. *Cures*. 2023;15(10):e47485. doi 10.7759/cureus.47485

Bessa D.S., Maschietto M., Aylwin C.F., Canton A.P.-M., Brito V.N., Macedo D.B., Cunha-Silva M., … Netchine I., Krepischi A.C.V., Lomniczi A., Ojeda S.R., Latronico A.C. Methylome profiling of healthy and central precocious puberty girls. *Clin Epigenetics*. 2018;10:e146. doi 10.1186/s13148-018-0581-1

Butler M.G. Imprinting disorders in humans: a review. *Curr Opin Pediatr*. 2020;32(6):719-729. doi 10.1097/MOP.0000000000000965

Canton A.P.M., Steunou V., Sobrier M.-L., Montenegro L.R., Bessa D.S., Gomes L.G., Jorge A.A.L., Mendonca B.B., Brito V.N., Netchine I., Latronico A.C. Investigation of imprinting defects in *MKRN3* and *DLK1* in children with idiopathic central precocious puberty through specific DNA methylation analysis. *J Endocr Soc*. 2020;4(1):SUN-090.A426. doi 10.1210/jendso/bvaa046.847

Canton A.P.M., Krepischi A.C.V., Montenegro L.R., Costa S., Rosenberg C., Steunou V., Sobrier M.L., … Jorge A.A.L., Mendonca B.B., Netchine I., Brito V.N., Latronico A.C. Insights from the genetic characterization of central precocious puberty associated with multiple anomalies. *Hum Reprod*. 2021;36(2):506-518. doi 10.1093/humrep/deaa306

Chebotareva Yu.Yu., Petrov Yu.A., Rodina M.A. Some aspects of precocious puberty in preschool-age girls. *Russian Journal of Woman and Child Health*. 2022;5(3):215-222. doi 10.32364/2618-8430-2022-5-3-215-222 (in Russian)

Faienza M.F., Urbano F., Moscogiuri L.A., Chiarito M., De Santis S., Giordano P. Genetic, epigenetic and enviromental influencing factors on the regulation of precocious and delayed puberty. *Front Endocrinol* (*Lausanne*). 2022;13:e1019468. doi 10.3389/fendo.2022.1019468

Fanis P., Morrou M., Tomazou M., Michailidou K., Spyrou G.M., Toumba M., Skordis N., Neocleous V., Phylactou L.A. Methylation status of hypothalamic *Mkrn3* promoter across puberty. *Front Endocrinol* (*Lausanne*). 2022;13:e1075341. doi 10.3389/fendo.2022.1075341

Gomes L.G., Cunha-Silva M., Crespo R.P., Ramos C.O., Montenegro L.R., Canton A., Lees M., … Baracat E.C., Jorge A.A.L., Mendonca B.B., Brito V.N., Latronico A.C. *DLK1* is a novel link between reproduction and metabolism. *J Clin Endocrinol Metab*. 2019;104(6):2112-2120. doi 10.1210/jc.2018-02010

Kagami M., Mizuno S., Matsubara K., Nakabayashi K., Sano S., Fuke T., Fukami M., Ogata T. Epimutations of the IG-DMR and the *MEG3*-DMR at the 14q32.2 imprinted region in two patients with Silver-Russell syndrome-compatible phenotype. *Eur J Hum Genet*. 2015;23(8):1062-1067. doi 10.1038/ejhg.2014.234

Kagami M., Yanagisawa A., Ota M., Matsuoka K., Nakamura A., Matsubara K., Nakabayashi K., Takada S., Fukami M., Ogatac T. Temple syndrome in a patient with variably methylated CpGs at the primary *MEG3/DLK1*:IG-DMR and severely hypomethylated CpGs at the secondary *MEG3*:TSS-DMR. *Clin Epigenetics*. 2019;11(1):e42. doi 10.1186/s13148-019-0640-2

Li C., Lu W., Yang L., Li Z., Zhou X., Guo R., Wang J., … Wang W., Huang X., Li Y., Gao S., Hu R. *MKRN3* regulates the epigenetic switch of mammalian puberty via ubiquitination of *MBD3*. *Natl Sci Rev*. 2020;7(3):671-685. doi 10.1093/nsr/nwaa023

Li C., Han T., Li Q., Zhang M., Guo R., Yang Y., Lu W., … Zhou V., Han Z., Li H., Wang F., Hu R. *MKRN3*-mediated ubiquitination of Poly(A)-binding proteins modulates the stability and translation of *GNRH1* mRNA in mammalian puberty. *Nucleic Acids Res*. 2021;49(7):3796-3813. doi 10.1093/nar/gkab155

Li L.C., Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics*. 2002;18(11):1427-1431. doi 10.1093/bioinformatics/18.11.1427

Lomniczi A., Wright H., Castellano J.M., Matagne V., Toro C.A., Ramaswamy S., Plant T.M., Ojeda S.R. Epigenetic regulation of puberty via zinc finger protein-mediated transcriptional repression. *Nat Commun*. 2015;6:e10195. doi 10.1038/ncomms10195

Macedo D.B., Kaiser U.B. *DLK1*, Notch signaling and the timing of puberty. *Semin Reprod Med*. 2019;37(4):174-181. doi 10.1055/s-0039-3400963

Mackay D.J.G., Gazdagh G., Monk D., Brioude F., Giabicani E., Krzyzewska I.M., Kalish J.M., … Russo S., Tannorella P., Temple K.I., Õunap K., Tümer Z. Multi-locus imprinting disturbance (MLID): interim joint statement for clinical and molecular diagnosis. *Clin Epigenetics*. 2024;16:99. doi 10.1186/s13148-024-01713-y

Micangeli G., Paparella R., Tarani F., Menghi M., Ferraguti G., Carlomagno F., Spaziani M., Pucarelli I., Greco A., Fiore M. Clinical management and therapy of precocious puberty in the Sapienza university pediatrics hospital of Rome, Italy. *Children* (*Basel*). 2023;10(10):e1672. doi 10.3390/children10101672

Monteagudo-Sánchez A., Hernandez M.J.R., Simon C., Burton A., Tenorio J., Lapunzina P., Clark S., … Kelsey G., López-Siguero J.P., de Nanclares G.P., Torres-Padilla M.E., Monk D. The role of *ZFP57* and additional KRAB-zinc finger proteins in the maintenance of human imprinted methylation and multi-locus imprinting disturbances. *Nucleic Acids Res*. 2020;48(20):11394-11407. doi 10.1093/nar/gkaa837

Narusawa H., Ogawa T., Yagasaki H., Nagasaki K., Urakawa T., Saito T., Soneda S., … Naiki Y., Horikawa R., Ogata T., Fukami M., Kagami M. Comprehensive study on central precocious puberty: molecular and clinical analyses in 90 patients. *J Clin Endocrinol Metab*. 2024;26:e666. doi 10.1210/clinem/dgae666

Nicoara D.M., Scutca A.C., Mang N., Juganaru I., Munteanu A.I., Vitan L., Mărginean O. Central precocious puberty in Prader-Willi syndrome: a narrative review. *Front Endocrinol* (*Lausanne*). 2023; 14:e1150323. doi 10.3389/fendo.2023.1150323

Okae H., Chiba H., Hiura H., Hamada H., Sato A., Utsunomiya T., Kikuchi H., Yoshida H., Tanaka A., Suyama M., Arima T. Genome-wide analysis of DNA methylation dynamics during early human development. *PLoS Genet*. 2014;10:e1004868. doi 10.1371/journal.pgen.1004868

Peterkova V.A., Alimova I.L., Bashnina E.B., Bezlepkina O.B., Bolotova N.V., Zubkova N.A., Kalinchenko N.Yu., … Malievskiy O.A., Orlova E.M., Petryaykina E.E., Samsonova L.N., Taranushenko T.E. Clinical guidelines “Precocious puberty”. *Problemy Endocrinologii = Problems of Endocrinology*. 2021;67(5):84-103. doi 10.14341/probl12821 (in Russian)

Roberts S.A., Kaiser U.B. Genetic etiologies of central precocious puberty and the role of imprinted genes. *Eur J Endocrinol*. 2020; 183(4):107-117. doi 10.1530/EJE-20-0103

Sazhenova E.A., Vasilyev S.A., Rychkova L.V., Khramova E.E., Lebedev I.N. Genetics and epigenetics of precocious puberty. *Russ J Genet*. 2023;59(12):1277-1287. doi 10.1134/S1022795423120104

Shim Y.S., Lee H.S., Hwang J.S. Genetic factors in precocious puberty. *Clin Exp Pediatr.* 2022;65(4):172-181. doi 10.3345/cep.2021.00521

Tucci V., Isles A.R., Kelsey G., Ferguson-Smith A.C. Genomic imprinting and physiological processes in mammals. *Cell*. 2019;176(5): 952-965. doi 10.1016/j.cell.2019.01.043

Conflict of interest. The authors declare no conflict of interest.
Received October 15, 2024. Revised December 18, 2024. Accepted December 28, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):448-457

doi 10.18699/vjgb-25-48

Expression analysis of microRNA and lncRNA in visceral adipose tissue of obese and non-obese individuals

A. Bairqdar [1, 3], D.E. Ivanoshchuk [1, 2], O.V. Tuzovskaya [2], N.S. Shirokova [1], E.V. Kashtanova [2], Y.V. Polonskaya [2], Y.I. Ragino [2], E.V. Shakhtshneider [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Internal and Preventive Medicine – Branch of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Department of Genetics, Novosibirsk State University, Novosibirsk, Russia

shakhtshneyderev@bionet.nsc.ru

Abstract. Long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) play important roles in all biological processes, including adipogenesis, lipid metabolism, and insulin response. Analyzing expression patterns of lncRNAs and miRNAs in human visceral fat tissue can enhance our understanding of their roles in metabolic disorders. Our research aims to investigate the expression of lncRNAs (ASMER1, SNHG9, P5549, P19461, and GAS5) and miRNAs (miR-26A, miR-222, miR-221, and miR-155) in visceral adipose tissues of individuals with abdominal obesity (n = 70) compared to their levels in non-obese participants (n = 31), using Real-Time PCR. Among the tested miRNAs, only miR-26A was significantly downregulated in the visceral adipose tissue of obese individuals, with no significant change in the expression of miR-26A in obese people with or without type 2 diabetes. Similarly, of the tested lncRNAs, only GAS5 showed significantly higher expression levels in obese patients with type 2 diabetes (T2D) (n = 10) compared to obese patients without T2D (n = 60). To test possible interactions between the analyzed non-coding RNAs, we used Spearman's bivariate correlation test. GAS5 expression levels showed a weak negative correlation ($p < 0.05$, $r_s = 0.25$) with miR-155 levels in obese patients only. Conversely, a strong positive correlation ($p < 0.01$, $r_s = 0.92$) between SNHG9 and GAS5 was found in the non-obese group, with a weaker correlation in abdominally obese patients ($p < 0.01$, $r_s = 0.67$); additionally, miR-26A and miR-155 levels were moderately correlated in the non-obese group ($p < 0.05$, $r_s = 0.47$) and were found to correlate weakly in obese patients ($p < 0.05$, $r_s = 0.26$). Our results showed that abdominally obese participants demonstrated higher expression levels of miR-26A in visceral adipose tissue and a significantly lower correlation between GAS5 and SNHG9 expression when compared to non-obese subjects.

Key words: long non-coding RNA; microRNA; visceral adipose tissues; abdominal obesity; GAS5; SNHG9; miR-26A; miR-155

For citation: Bairqdar A., Ivanoshchuk D.E., Tuzovskaya O.V., Shirokova N.S., Kashtanova E.V., Polonskaya Y.V., Ragino Y.I., Shakhtshneider E.V. Expression analysis of microRNA and lncRNA in visceral adipose tissue of obese and non-obese individuals. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):448-457. doi 10.18699/vjgb-25-48

Author contributions. Conceptualization, D.I., Y.R. and E.S.; methodology, D.I. and A.B.; software, A.B.; validation, D.I. and E.K.; investigation, A.B., O.T., N.S. and Y.P.; resources, E.S.; data curation, O.T.; writing – original draft preparation, A.B., D.I., and E.S.; writing – review and editing, A.B., D.I. and E.S.; project administration, E.S. All authors have read and agreed to the published version of the manuscript.

Funding. The molecular genetic testing was carried out within the framework of the main topic of state assignment No. FWNR-2025-0006.

Анализ экспрессии микроРНК и днкРНК в висцеральной жировой ткани у лиц с ожирением и без ожирения

А. Бейркдар [1, 3], Д.Е. Иванощук [1, 2], О.В. Тузовская [2], Н.С. Широкова [1], Е.В. Каштанова [2], Я.В. Полонская [2], Ю.И. Рагино [2], Е.В. Шахтшнейдер [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт терапии и профилактической медицины – филиал Федерального исследовательского центра Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

shakhtshneyderev@bionet.nsc.ru

Аннотация. Длинные некодирующие РНК (lncRNA) и микроРНК (miRNA) играют важную роль во всех биологических процессах, включая адипогенез, липидный обмен и инсулиновый ответ. Анализ экспрессии lncRNA и miRNA в висцеральной жировой ткани человека может способствовать более глубокому пониманию их роли в развитии метаболических нарушений. Исследование направлено на изучение уровней экспрессии lncRNA

(ASMER1, SNHG9, P5549, P19461 и GAS5) и miRNA (miR-26A, miR-222, miR-221 и miR-155) в висцеральной жировой ткани у людей с абдоминальным ожирением (n = 70) по сравнению с уровнями их экспрессии у лиц без абдоминального ожирения (n = 31) методом количественной ПЦР в реальном времени. Среди протестированных miRNA уровень экспрессии miR-26A был снижен в висцеральной жировой ткани у людей с ожирением. Среди изученных lncRNA GAS5 показал значительное повышение уровня экспрессии у пациентов с ожирением и сахарным диабетом 2-го типа (СД2) (n = 10) по сравнению с лицами с ожирением без СД2 (n = 60). Уровни экспрессии GAS5 показали слабую отрицательную корреляцию ($p < 0.05$, $r_s = 0.25$) с уровнями miR-155 только у пациентов с ожирением. Положительная корреляция ($p < 0.01$, $r_s = 0.92$) между SNHG9 и GAS5 была обнаружена в группе людей без ожирения, с уменьшением коэффициента корреляции у пациентов с ожирением ($p < 0.01$, $r_s = 0.67$). Кроме того, уровни miR-26A и miR-155 умеренно коррелировали в группе без ожирения ($p < 0.05$, $r_s = 0.47$) и демонстрировали слабую корреляцию у пациентов с ожирением ($p < 0.05$, $r_s = 0.26$). Наши результаты показали, что у пациентов с абдоминальным ожирением наблюдаются повышенные уровни экспрессии miR-26A в висцеральной жировой ткани и слабая корреляция между экспрессией GAS5 и SNHG9 по сравнению с лицами без абдоминального ожирения.

Ключевые слова: длинная некодирующая РНК; микроРНК; висцеральная жировая ткань; абдоминальное ожирение; GAS5; SNHG9; микроРНК-26A; микроРНК-155

Introduction

The prevalence of obesity has been rising over the last four decades and it is currently considered a global pandemic (Valenzuela et al., 2023). Visceral fat accumulation has a stronger association with metabolic disorders compared to subcutaneous fat, despite both primarily consisting of white adipose tissue (Nicklas et al., 2003).

The complexity of white adipose tissue lies in its complex role as an endocrine organ, secreting various hormones, enzymes, growth factors, and multiple non-coding RNAs, all of which regulate numerous metabolic processes in the organism (Lustig et al., 2022). In the last decade, studies have been focusing on long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) as key regulators in the fine-tuning of gene function across various biological processes (Mohanty et al., 2015; Rupaimoole, Slack, 2017). lncRNAs can be transcribed from different regions in the DNA and are typically categorized into three types: intergenic, antisense, and intronic. According to the latest recommendation, lncRNAs are transcripts longer than 500 nucleotides (Mattick et al., 2023), are less evolutionarily conserved than mRNAs, contain fewer exons, are characterized by lower splicing efficiency than mRNAs. lncRNAs are less abundantly expressed than other non-coding RNAs (Statello et al., 2021). Their functions include RNA-chromatin, RNA-RNA, and RNA-protein interactions (Ferrer, Dimitrova, 2024). The functioning mechanism of lncRNA involves the regulation of chromatin structure and transcription, usually through the methylation of enhancer regions (Li W. et al., 2016). lncRNAs interact with mRNAs, modifying their splicing and stability (Morrissy et al., 2011). Additionally, lncRNA can hybridize with mature RNAs at the 5′-region, increasing its translation efficiency (Carrieri et al., 2012; Zucchelli et al., 2015).

Altered expression profiles of lncRNAs have been found in patients with acute myocardial infarction (Zhong et al., 2018), multiple malignancies (Li L. et al., 2021), and, most recently, Alzheimer's disease (Zhang M. et al., 2019; Zhang J.-J. et al., 2021). Studies associating lncRNAs with metabolic disorders have investigated their roles in adipose tissue regulation and development (Tello-Flores et al., 2021; Corral et al., 2022; Sufianov et al., 2023), with accumulating evidence emphasizing their roles in obesity, type 2 diabetes, and other related disorders (Liu et al., 2018; Su et al., 2023).

miRNAs are short in length, approximately 22 nucleotides long, usually derived from longer primary miRNA (pre-miRNA) transcripts. Most miRNAs are processed from a non-coding transcript originating from the intron or non-coding regions, and occasionally from exons of protein-coding genes (Dexheimer, Cochella, 2020).

The direct association of microRNAs with a wide variety of diseases – from various malignancies (Reddy, 2015; Ali Syeda et al., 2020) and autoimmune diseases (Zhang L. et al., 2020), to metabolic disorders including obesity (Veie et al., 2023) and type 2 diabetes (Miao et al., 2018) – is well-established. Currently, miRNAs are being explored as biomarkers in obesity (Gouda et al., 2023) and as therapeutic targets due to their short size and promising *in vitro* (Acharya et al., 2019) and *in vivo* results (Lhamyani et al., 2021).

miR-155, miR-221, and miR-222 were found to be decreased during adipogenic differentiation of human mesenchymal stromal cells, suggesting their role as negative regulators of the process (Skårn et al., 2012). Moreover, miR-26A overexpression in transgenic mice led to reduced visceral fat levels and an improved lipid profile (Zeng H. et al., 2021). miR-222 blood levels were higher in obese subjects after surgeries leading to weight loss, while in the same subjects, miR-221 levels were reduced (Ortega et al., 2013). Circulating levels of miR-155 were higher in obese patients, while miR-26A was downregulated in the same group (Kim et al., 2020).

Non-coding RNAs function as networks that, if disrupted, could lead to multiple disorders including, but not limited to, obesity (Ma et al., 2023); miR-26A is found to be downregulated by the lncRNA GAS5 sponge mechanism in degenerated nucleus pulposus cells (Tan et al., 2021). Additionally, GAS5 is suggested to participate in the pathogenesis of pneumonia by downregulating miR-155 and inhibiting apoptosis (Wang et al., 2021). The miRNA-lncRNA interaction network in obesity has recently gained interest. Guo and Cao (2019) described a potential interaction of lncRNA RP11-552F3.9 with miR-130b and miR-23a in LEP gene regulation and adipogenic differentiation. Additionally, RP11-142A22.4 was identified as a functional site for binding miR-587, making it unavailable and subsequently promoting adipogenesis (Zhang T. et al., 2020). Another example is lncRNA Adi, discovered in 2020, which was found to bind to miR-449a in adipose-derived stem cells, inducing their adipogenic differentiation (Chen et al., 2020).

Previous studies investigating the expression profiles of lncRNAs in obese patients were limited to describing blood levels (Sun et al., 2016; Cabiati et al., 2022; Rasaei et al., 2024) and/or had limited population samples (Tait et al., 2020; Tan et al., 2021). Additionally, in 2016, Sun et al. identified new lncRNAs (P19641, P5549, P21015) that are correlated with obesity, but fell short of identifying the splicing variants expressed in blood samples (Sun et al., 2016). Similarly, Lv et al.'s study describing GAS5 function in visceral fat tissues included only female subjects (Lv et al., 2022).

Despite the attempt by Kim et al. to identify differentially expressed miRNAs in visceral fat tissues, their population consisted only of females and was limited to 20 subjects (Kim et al., 2020). The same limitation was noted in the study by Capobianco et al. (2012).

Due to the limitations of the aforementioned studies, our research aims to first identify the specific splicing variant of P19641, P5549, SNHG9, and ASMER1 in human visceral fat tissue samples. Subsequently, we aim to measure the expression levels of multiple lncRNAs (P19641, P5549, SNHG9, ASMER1, and GAS5) and miRNAs (miR-221, miR-222, miR-155, and miR-26A) in our population.

Materials and methods

Study design. The study protocol was approved by the local Ethics Committee of the Institute of Internal and Preventive Medicine – Branch of the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia (Protocol No. 66 from October 10, 2023) – and was conducted in accordance with the principles of the Declaration of Helsinki of the World Medical Association. Informed consent was obtained from all participants.

A total of 101 accepted individuals underwent a comprehensive clinical examination program. This included the collection of sociodemographic data, administration of a standardized questionnaire on smoking and alcohol use, documentation of chronic disease history, and recording of medication usage; the main medication used is described in Table 1, additional information about administered medications can be found in the supplementary file (Supplementary Material 1)[1]. Additionally, the program involved the Rose cardiological questionnaire, anthropometric measurements (height, body weight, and waist circumference), and blood pressure measurements. Blood serum samples were collected for biochemical assays measuring total cholesterol, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides, and fasting glucose levels.

Lipid levels (cholesterol, triglycerides, and low-density and high-density lipoprotein cholesterols) and glucose concentration were measured using a biochemical analyzer KoneLab 300i (Thermo Fisher Scientific Oy, Finland) with Thermo Fisher Scientific reagents (USA). The values of LDL-C concentration were calculated using the Friedwald formula. The atherogenic coefficient was calculated using the formula: IA = (TC − HDLC)/HDLC. The levels of leptin and adiponectin were determined by multiplex analysis using the Human Adipokine Magnetic Bead Panel 1 kit (EMD Millipore Corporation, Darmstadt, Germany) on a Luminex 20 MAGPIX flow cytometer (Luminex Corporation, Austin, TX, USA).

All 101 subjects scheduled for elective surgery were included in the study based on the main inclusion criteria of having no records of chronic inflammation or symptoms of acute inflammation prior to surgery. The participants were subsequently divided into two groups: 70 participants with abdominal obesity (AO+) and 31 non-obese subjects (AO–) (Table 1). Abdominal obesity was defined as a body mass index (BMI, kg/m^2) of ≥25.0 or a waist circumference (WC) of ≥80 cm in women and ≥94 cm in men, in accordance with the latest medical guidelines (Yumuk et al., 2015; Dedov et al., 2021). Type 2 diabetes (T2D) was diagnosed based on the criteria set forth by the American Diabetes Association. No adjustment for age was applied since the study group was intended to represent only the population aged 45–55. Additionally, no adjustment for sex or medications was applied because the resulting subgroups would be relatively small and could generate statistically unreliable data (Table 1). Individuals exhibiting chronic or acute inflammation on the day of surgery, as well as pregnant women and women on maternity leave, were excluded from the study.

Sample collection and RNA extraction. During the operation, samples of visceral adipose tissue were collected, washed with PBS and subsequently preserved in RNAlater solution (ThermoFisher, USA) at –20 °C until further handling. lncRNA was extracted using the total RNA extraction protocol, microRNA was extracted using the microRNA extraction protocol. Both procedures were performed with the Total and microRNA Extraction Kit (LRU-100-50) (Biolabmix, Russia). All samples were treated with RNase-free DNase (Biolabs, USA).

The concentration of yielded RNA was determined using the BioTek Epoch Analyzer (Agilent Technologies, USA), and the quality was estimated by the absorbance ratios A260/230 and A260/280. Additionally, the quality of RNA preservation was assessed by determining the integrity of 28S, 18S, and 5S bands on 1 % agarose gel electrophoresis.

lncRNA reverse transcription and relative quantification. Reverse transcription of total RNA was performed using the MMuLV Reverse Transcription Kit. A total of 1,000 ng of each sample was used in a 40 μL reaction volume, which included Random Primer6, oligo(dT), and 5X reaction buffer (Kolenda et al., 2021). All cDNA samples were diluted to a total volume of 100 μL.

For P5549, P19641, ASMER1, and SNHG9 lncRNAs, we designed a multiple set of specific primers for each splicing variant listed in the Ensembl database (Ensembl.org) at the start of the study (October 2023). For GAS5 lncRNA, we used a single set of primers specific for the canonical transcript ENST00000702964.1. All primers were designed using the PrimerBlast tool (NCBI, USA). Primer efficiency was determined by the standard curve method with four crossing points. Specificity was assessed by analyzing melting curves from 65 to 95 °C with a 0.5 °C increment per cycle and by determining product length using 5 % acrylamide gel. Conventional PCR for splicing variant identification and qPCR for relative expression analysis were performed using SYBR GREEN I intercalating dye master mix "BioMaster HS-qPCR SYBR Blue (2×)" (Biolabmix, Russia) on the LightCycler 96 system

[1] Supplementary Materials 1 and 2 are available at:
https://vavilovj-icg.ru/download/pict-2025-29/appx15.pdf

Table 1. Clinical characteristics, body composition, and biochemical variables of the study individuals divided into an obese (AO+) and a non-obese group (AO–)

Variable	AO+	AO–	*p*	Population of Western Siberia (Russia)
Number of participants	70	31	N/A[1]	3,132
T2D[2] established diagnosis	10	1	N/A	No data
Age	54 (42–61.3)	48 (35–60)	0.13	56.5 ± 0.10
Males/Females, %	41.40/59.60	70/30	N/A	
BMI[3], kg/m^2	32.3 (29.7–35.5)	22.6 (20.3–24.6)	0.0001	28.1 ± 0.1
Wasit circumference, cm	102.39 (96–109.7)	82 (76–90)	0.0001	93.5 ± 0.2
Blood glucose levels, mmol/L	6.00 (5.4–6.6)	5.8 (5.5–6.5)	0.73	5.7 ± 0.04
Total cholesterol, mmol/L	5.1 (4.4–5.9)	4.9 (3.8–5.7)	0.52	6.33 ± 0.03
Triglycerides, mmol/L	1.6 (1.2–2.1)	1.2 (1.0–1.5)	0.001	1.56 ± 0.02
HDL-C[4], mmol/L	1.3 (1.0–1.6)	1.5 (1.0–1.9)	0.11	1.54 ± 0.01
LDL-C[5], mmol/L	3.0 (2.8–3.6)	2.3 (1.9–3.6)	0.47	3.16 ± 0.02
Undergoing glucose-lowering therapy, %	10.0	3.2	N/A	N/A
Undergoing lipid-lowering therapy, %	15.7	3.2	N/A	N/A

Note. Values are indicated as median and (25–75 % IQR), while reference values are indicated as mean ± standard error of the mean. [1] Not applicable, [2] type 2 diabetes, [3] body mass index, [4] high-density lipoprotein cholesterol, [5] low-density lipoprotein cholesterol.

(Roche, Switzerland). The amplification program was 95 °C for 3 min, followed by a 3-step amplification (95 °C for 15 sec, 63 °C for 15 sec, 72 °C for 20 sec) for a total of 40 cycles, followed by product melting as described above. Sequences of all primers used and standard curves are available in the Supplementary Material 1. Expression of all lncRNAs was normalized to the stable reference gene GAPDH expression levels (Mehta et al., 2010; Ebrahimi et al., 2020).

miRNA reverse transcription and relative quantification. Reverse transcription of microRNA was performed using the stem-loop pulsed reverse transcription protocol with 150 µL of the extracted microRNA (Varkonyi-Gasic et al., 2007). Expression was measured using SYBR GREEN I intercalating dye master mix "BioMaster HS-qPCR SYBR Blue (2×)" (Biolabmix, Russia) on the LightCycler 96 system (Roche, Switzerland). Primers designed for the reverse transcription product included a reverse primer specific for the stem-loop structure and a forward primer specific for the miRNA of interest (Varkonyi-Gasic et al., 2007). Efficiency was determined by the standard curve method with four crossing points. Specificity was assessed by analyzing melting curves from 65 to 95 °C with a 0.5 °C increment per cycle and by determining product length using 5 % acrylamide gel analysis. For miRNA quantification, we designed reverse transcription primers and specific qPCR primers using the online tool sRNAPrimerDB (srnaprimerdb.com). Sequences of all primers used and standard curves are available in the Supplementary Material 1. Expression of all microRNAs was normalized to miRNA-103, given its proven stable expression levels in visceral fat tissue (Ragni et al., 2021).

lncRNA sequence verification by Sanger sequencing. Sanger sequencing was used to verify the PCR amplification product sequence. The PCR product was purified using the Reaction Mixtures DNA Isolation Kit (DR-50, Biolabmix, Russia). Sequencing was performed using the BigDye Terminator v3.1 Kit (ThermoFisher, USA) according to the manufacturer's protocol, and capillary electrophoresis was conducted using the SeqStudio Genetic Analyzer (ThermoFisher, USA). All sequences were verified using the Blastn online alignment tool (Ensembl.org). Alignment options were set to default, and only the first sequence hit was reported (Table 2) with an identity of ~100 % and the highest percentage of coverage.

Statistical analysis. The ΔCt method was used to calculate the relative expression ($2^{-\Delta Ct}$, fold of expression relative to the reference gene by the formula $\Delta Ct = (Ct_{measured\ transcript} - Ct_{reference\ transcript})$ of lncRNA and microRNA (Livak, Schmittgen, 2001). All variables were tested for normal distribution using the Kolmogorov–Smirnov test. The Mann–Whitney test was used for group comparison. A *p*-value of 0.05 was considered significant (*), and $p < 0.01$ was considered highly significant (**). No cutoff for the fold change was assigned. Variables are presented as the median and 25–75 % interquartile range (IQR). Correlation between continuous variables and expression values of both miRNA and lncRNA was tested using Spearmans's bivariate correlation analysis and results were reported as Spearman's rank correlation coefficient and *p*-value for statistical significance. All statistical analyses were performed using SPSS software version 26 for Windows (IBM, USA). All graphs were generated using Prism GraphPad software version 8.2.1 for Windows (Dotmatics, USA).

Results

Subject characteristics

Clinical characteristics of the studied groups are presented in Table 1. A comparative analysis of obese and non-obese individuals revealed differences in triglyceride levels ($p = 0.01$).

Table 2. Splicing variants of the investigated lncRNAs identified to be expressed in visceral fat tissue

lncRNA symbol	Ensembl gene code	Ensembl code of the detected transcript	Blastn aligned sequence
P5549	ENSG00000259200	ENST00000559600.1	ENST00000559600.1
P19461	ENSG00000259719	ENST00000664388.1 ENST00000666463.1	ENST00000664388.1 ENST00000666463.1
SNHG9	ENSG00000255198	ENST00000531523.3	ENST00000531523.3
ASMER1	ENSG00000281903	Not detected	Not detected

Table 3. Expression levels of different measured lncRNAs and miRNAs in visceral fat tissues of obese patients (AO+) and non-obese participants (AO–), as well as in obese individuals with abdominal obesity and type 2 diabetes (AO+&T2D+) and obese participants without type 2 diabetes (AO+&T2D–)

Transcript symbol	AO+	AO–	*p*	AO+&T2D+	AO+&T2D–	*p*
miR-155	0.44 (0.20–0.72)	0.41 (0.27–0.49)	0.93	0.5 (0.19–0.87)	0.4 (0.20–0.72)	0.20
miR-26A	7.29 (4.97–10.2)	10.63 (7.24–14.0)	0.003**	7.0 (5.09–9.80)	7.7 (4.70–10.20)	0.94
lncRNA SNHG9	0.24 (0.18–0.48)	0.27 (0.22–0.50)	0.35	0.44 (0.20–1.40)	0.22 (0.15– 0.48)	0.54
lncRNA GAS5	1.00 (0.50–1.77)	1.23 (0.82–1.73)	0.10	1.34 (1.08–9.06)	0.87 (0.49–1.70)	0.03*

Values of reference biochemical markers for the West Siberian population were described in a survey conducted in 2019 (Semaev et al., 2019).

We aimed, first, to identify the transcribed splicing variant of all investigated lncRNAs. Second, we quantified various lncRNAs and miRNAs for subsequent statistical analysis of correlation with phenotypes of obesity, diabetes, and different biochemical markers. The final section investigates the possible correlation between lncRNA and miRNA expression, which might indicate potential interaction or a common regulatory pathway involved in adipogenesis and the pathology of obesity.

lncRNA expression

Out of the two explored alternative expression forms of lncRNA P5549, we identified ENST00000559600.1 to be solely expressed in our sample pool. ENST00000559600.1 is defined to be the canonical variant in the Ensembl database, has exons 2, 3 and 4 alternatively spliced, while exons 6, 7 and 8 are kept (Table 2). On the other hand, P19461 was found to be expressed in two alternative forms out of the five forms that we tested for. ENST00000664388.1 is the only transcript with exon 10 kept and defined as the canonical form by the Ensembl database. ENST00000666463.1 is the shortest splicing form of P19461 with only one exon (exon 1). We also tested for the expression of two different transcripts of SNHG9 and found the canonical transcript ENST00000531523.3 to be the only expressed form in visceral fat samples; this transcript has exon 2 out of all three exons alternatively spliced. Finally, none of the four investigated ASMER1 splicing variants had detected expression levels (Table 2). Sequence data of all PCR products used for identification of lncRNAs are available in the Supplementary Material 1.

The fluorescence signal of specific PCR products for lncRNA P5549 and P19461 was detected at late cycles and was lower than the detection limits of LightCycler 96, which made further precise relative quantification not possible by a simple qPCR method. Relative quantification showed no significant difference of mean SNHG9 and GAS5 expression levels between obese patients and non-obese subjects. After adjusting for type 2 diabetes diagnosis, GAS5 expression levels were significantly higher (~3 fold of change of the mean) ($p = 0.03$) in obese subjects with T2D compared to obese patients without T2D (Fig. 1). Expression levels of measured lncRNAs are shown in Table 3.

Correlation analysis between lncRNA expression levels and various metabolic parameters, such as BMI, waist circumference, total cholesterol, and lipid profile, did not show any significant correlation in either of the studied groups. The results of correlation tests performed on lncRNAs are provided in the Supplementary Material 2.

microRNA expression

Expression levels of miR-221 and miR-222 were undetectable using the method described before. Expression levels of miR-155 were not significantly changed between the obese and the non-obese group, while expression levels of miR-26A were significantly lower in the obese group (~1.5 fold of change of the mean) ($p < 0.01$) (Fig. 2). After adjustment for type 2 diabetes diagnosis, miR-155 and miR-26A expression levels showed no significant differences between obese patients diagnosed with T2D compared to obese individuals without T2D (Fig. 2). Mean expression levels of miRNAs in the studied groups are described in Table 3.

We also performed a correlation test between miR-26A and miR-155 expression levels and multiple metabolic parameters. In obese subjects, miR-155 expression showed a moderate positive correlation with BMI ($r_s = 0.32$, $p = 0.009$). Interestingly, in non-obese subjects, the opposite pattern of correlation was observed, where miR-155 negatively correlated with both BMI ($r_s = –0.44$, $p = 0.029$) and waist circumference ($r_s = –0.44$, $p = 0.027$). A table of all correlation tests performed on miRNAs is provided in the Supplementary Material 2.

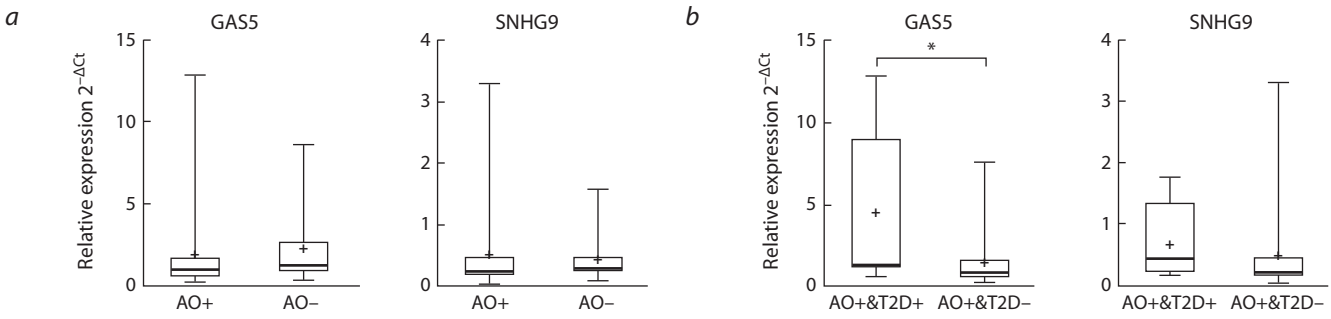


Fig. 1. A box plot representing the relative expression of lncRNA GAS5 and SNHG9.

a, Comparison of lncRNA expression in the obese group (AO+) and the non-obese group (AO–); *b*, comparison of lncRNA expression between obese subjects diagnosed with type 2 diabetes (AO+&T2D+) and obese subjects without an established diagnosis of T2D (AO+&T2D–). The box represents the 25th–75th percentile range, error bars represent the minimum and maximum values, the median is represented as a line inside the box, and the mean, as a + sign. * $p < 0.05$.

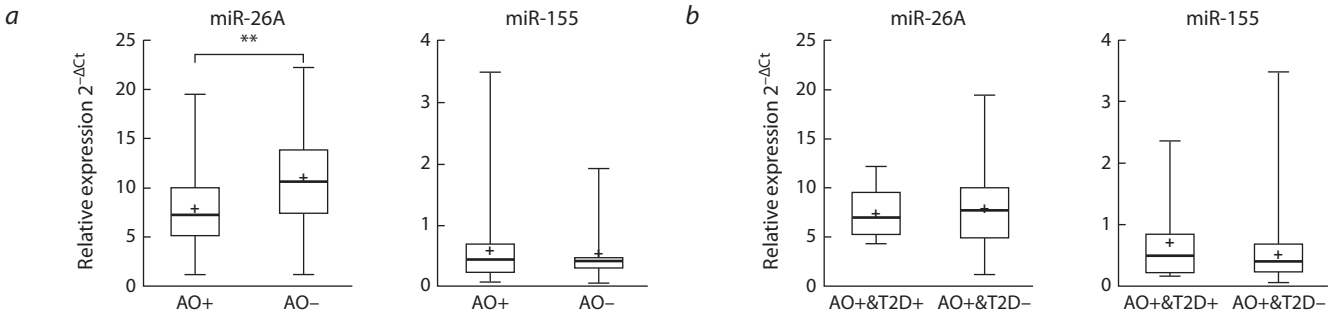


Fig. 2. A box plot representing the relative expression of miR-26A and miR-155.

a, Comparison of miRNAs expression in the obese group (AO+) and the non-obese group (AO–); *b*, comparison of miRNAs expression between obese subjects diagnosed with type 2 diabetes (AO+&T2D+) and obese subjects without an established diagnosis of T2D (AO+&T2D–). The box represents the 25th–75th percentile range, error bars represent the minimum and maximum values, the median is represented as a line inside the box, and the mean, as a + sign. ** $p < 0.01$.

Correlation of lncRNA-lncRNA, miRNA-miRNA and lncRNA-miRNA expression levels

We tested the correlation between miRNA and lncRNA expression levels (Table 4). In obese patients, lncRNA GAS5 showed a weak negative correlation with miR-155 expression levels ($r_S = 0.25$, $p < 0.05$), whereas this correlation was not observed in non-obese participants. Additionally, no correlation was found between GAS5 and miR-26A expression levels in either group. Similarly, SNHG9 did not correlate with any of the investigated miRNAs in either group. Interestingly, the non-obese group showed a very strong correlation ($r_S = 0.92$, $p < 0.0001$) between SNHG9 and GAS5 expression levels. However, the correlation between SNHG9 and GAS5 expression levels in obese subjects was moderate and significant

Table 4. Spearman's correlation between pairs of measured non-coding RNAs in the obese (AO+) and the non-obese group (AO–)

RNA	SNHG9	GAS5	miR-26A	miR-155
		AO+		
SNHG9	1	0.67**	–0.25	–0.20
GAS5	0.67**	1	–0.11	–0.25*
miR-26A	–0.25	–0.11	1	0.26*
miR-155	–0.20	–0.25*	0.26*	1
		AO–		
SNHG9	1	0.92**	–0.18	0.06
GAS5	0.92**	1	–0.29	–0.01
miR-26A	–0.18	–0.29	1	0.47*
miR-155	0.06	–0.01	0.47*	1

Note. The strength of correlation is described by the Spearman's rank correlation coefficient (r_s). * Indicates a significant correlation with 2-tailed *p*-value < 0.05, ** indicates a significant correlation with 2-tailed *p*-value < 0.01.

($r_s = 0.67$, $p < 0.0001$). Similarly, miR-26A and miR-155 displayed a moderate but significant ($r_s = 0.47$, $p < 0.05$) correlation in the non-obese group; the same correlation was reduced and manifested to be weak in obese participants ($r_s = 0.26$, $p < 0.05$). Scatter plots representing bivariate correlations are available in the Supplementary Material 2.

Discussion

The first objective of our study was to identify the specific transcription variant expressed in visceral fat tissue for each of the lncRNAs P5549, SNHG9, P19641, and ASMER1 (Table 2). Using specific primers and Sanger sequencing, we revealed for the first time that lncRNA P5549 was represented by the transcript ENST00000559600.1, and SNHG9 was solely identified as ENST00000531523.3. P19641 was expressed in two different splicing forms: ENST00000664388.1 and ENST00000666463.1, which can be attributed to the inefficient splicing of some lncRNAs (Li L. et al., 2021). The specific splicing variant expressed in visceral adipose tissue suggests a functional significance for the specifically transcribed exons in the interaction of the studied lncRNA with its targets, whether they be DNA, mRNA, or microRNA. Further studies are necessary to identify the exact functioning sequences of the identified lncRNAs, which could facilitate the implementation of these sequences as future therapeutic targets. Additionally, no expression of any of the tested variants for ASMER1 lncRNA was detected. This may be a sign of very low expression levels that are undetectable by the method used for investigation or an indication of low RNA stability.

Next, we aimed to compare the expression levels of lncRNAs SNHG9 and GAS5 between our obese and non-obese groups. Our analysis showed no significant difference in the mean expression levels of either of the analyzed lncRNAs between the two groups. However, GAS5 expression was significantly higher in obese subjects with type 2 diabetes (T2D) compared to obese subjects without T2D (Fig. 1). This contradicts the results from multiple studies showing lower GAS5 expression levels in blood samples collected from T2D patients compared to non-obese subjects (Fawzy et al., 2020; Luo et al., 2020; Ahmadi et al., 2024). GAS5 was found to be induced by proinflammatory mediators (Mameli et al., 2016), which might explain its higher levels in obese patients with T2D, a condition always linked to chronic adipose tissue inflammation (Zatterale et al., 2020). Our results suggest a different expression pattern of GAS5 in visceral fat tissues and might reflect the delicate balance between its function as an adipogenesis inhibitor and a proinflammatory factor. Despite the novelty of our finding, further verification in a larger population is required, in addition to investigating the mechanisms by which GAS5 might regulate fat metabolism.

The expression levels of miR-221 and miR-222 were undetectable in our study, which might be due to tissue-specific expression patterns or methodological differences compared to previous studies (Markovic et al., 2020; Chan et al., 2022). miR-26A was significantly downregulated in obese subjects, whereas miR-155 showed no significant dysregulation between the two groups (Fig. 2). The downregulation of miR-26A in obesity aligns with the findings of Kim et al. (2020) and emphasizes the role of miR-26A in adipogenesis and metabolic regulation described in previous studies (Acharya et al., 2019; Zeng H. et al., 2021). Additionally, miR-155 expression was correlated with current studies that are examining the potential of non-coding RNAs as biomarkers and therapeutic targets (Winkle et al., 2021). GAS5 has been presented as a promising biomarker in breast cancer treatment and prognosis (Grossi et al., 2023). Meanwhile, miR-26A has already been tested as a therapeutic molecule attenuating visceral fat accumulation in a mouse model (Zeng H. et al., 2021). Our results highlight the necessity for novel research investigating GAS5 as a therapeutic target in metabolic disorders. Additionally, we emphasize the need for further evidence of the efficacy of miR-26A as a potential therapeutic molecule in future preclinical and clinical studies.

Our study also explored potential interactions between the studied non-coding RNAs. We found no correlation between SNHG9 expression and miR-26A or miR-155. On the other hand, GAS5 showed a weak negative correlation with miR-155 levels in visceral fat tissue samples of obese individuals; we didn't observe a similar correlation in non-obese subjects (Table 4). Previous studies described a negative correlation between GAS5 and miR-155-5p in regulatory pathways related to inflammation and apoptosis. GAS5 was described to sponge and regulate miR-155, leading to the upregulation of SIRT1 and suppression of the inflammatory response to lipopolysaccharides (Zeng Z. et al., 2023). Additionally, GAS5 and miR-155 were inversely correlated in pneumonia patients, where GAS5 was downregulated and miR-155 was upregulated in plasma samples. GAS5 overexpression decreased miR-155 expression in human bronchial epithelial cells (Wang et al., 2021). This finding requires further confirmation in experimental studies, which may reveal a similar mechanism of regulation in adipose tissue as in inflammatory responses.

GAS5 expression levels showed no correlation with miR-26A in either of the studied groups. This finding does not align with the results of Tan et al. (2021) where GAS5 was responsible for the downregulation of miR-26A. Therefore, additional studies exploring miR-26A interaction network in visceral adipose tissue are required to clarify the exact mechanism by which miR-26A is negatively regulated in obese individuals.

In an attempt to expand our knowledge of the intricate network of non-coding RNAs, we also investigated the correlation between SNHG9 and GAS5 expression levels. Interestingly, we found a very strong, almost linear, correlation between the two lncRNAs in non-obese subjects ($r_s = 0.92$). This correlation was significantly reduced in obese patients ($r_s = 0.67$). Previous studies investigating expression levels of SNHG9 and GAS5 showed inversed correlation in clear cell renal cell carcinoma (Yang et al., 2020) and glioblastoma (Ji et al., 2020), but no common regulation mechanism was suggested. To the best of our knowledge, such correlation between SNHG9 and GAS5 has never been described before in studies of metabolic disorders. Hence, we can only suggest the existence of coordinated regulation between these two lncRNAs in normal metabolic conditions. This coordination could be disrupted in the condition of obesity, leading to a weaker correlation in the expression levels of the two lncRNAs. However,

further *in vitro* and *in vivo* experiments are required to confirm any type of common regulation mechanism.

A moderate correlation between miR-26A and miR-155 (r_s = 0.47) was found in the non-obese group, and this correlation was shown to be weak in obese patients (r_s = 0.26), which partially aligns with the correlation pattern of the above-described lncRNAs. Previous studies showed parallel downregulation of miR-26A and miR-155 in subjects with obesity (Kim et al., 2020) and multiple sclerosis (Mameli et al., 2016), but none of the studies attempted to test for a correlation between the expression levels of the two miRNAs. Such a correlation in expression levels might be evidence of a common regulation mechanism for miR-26A and miR-155. The reduced correlation in obese patients in our results aligns with the previous observation in lncRNA interaction that suggests the existence of a common regulation mechanism for both miRNAs, which might be disrupted in the case of obesity.

Our study included 101 participants divided into 70 obese patients and 31 non-obese individuals, all within the age group of 45–55 years. The study group was not adjusted for sex and types of medications, which is the first of several limitations we acknowledge. However, previous data investigating the expression of different non-coding RNAs in visceral fat tissue are extremely rare, and our chosen targets had not been investigated in this tissue before. Additionally, we aimed to generate first-time data that can prompt further investigation in this direction.

The limitations of our study also include the use of conventional PCR and subsequent Sanger sequencing for the identification of lncRNA splicing variants. While this method is highly specific, it is limited due to the small range of transcripts it can identify. We recommend further studies using transcriptomic methods to identify novel lncRNAs and new forms of expression for existing RNAs. Another limitation of our study was the inability to quantify the scarcely expressed lncRNAs, which might require preamplification or the use of digital PCR. Finally, our study describes correlations and statistical differences that require further confirmation or rejection by experimental *in vitro* and *in vivo* studies. Transcriptomics might also aid in identifying and quantifying various novel transcripts of lncRNAs expressed in visceral fat tissue, deepening our understanding of adipogenesis and fat metabolism.

Conclusions

In conclusion, the precise network of non-coding RNAs is once more shown to be associated with the development of metabolic and various other diseases. Our results indicated a different pattern of microRNA and lncRNA expression in individuals suffering from obesity and T2D. This dissimilarity highlights the important role of the investigated non-coding RNAs in the formation and differentiation of visceral adipose tissue. Additionally, our results displayed a specific miRNA-miRNA and lncRNA-lncRNA correlation pattern in non-obese individuals, which requires further investigation. Understanding these associations can lead to building a better map of the interaction network in the absence of metabolic disorders. This map can serve as a reference for understanding all possible abnormal or alternative pathways regulating the intricate networks in cases of obesity and other metabolic disorders.

References

Acharya A., Berry D.C., Zhang H., Jiang Y., Jones B.T., Hammer R.E., Graff J.M., Mendell J.T. miR-26 suppresses adipocyte progenitor differentiation and fat production by targeting *Fbxl19*. *Genes Dev*. 2019;33(19-20):1367-1380. doi 10.1101/gad.328955.119

Ahmadi S., Boozarpour S., Sabouri H., Ghalandarayeshi S., Babaee N., Lashkarboloki M., Banikarimi S.A. Expression of circulating long non-coding MALAT1 and GAS5 under metformin treatment in type 2 diabetic patients. *Gene Rep*. 2024;35:101905. doi 10.1016/j.genrep.2024.101905

Ali Syeda Z., Langden S.S.S., Munkhzul C., Lee M., Song S.J. Regulatory mechanism of microRNA expression in cancer. *Int J Mol Sci*. 2020;21(5):1723. doi 10.3390/ijms21051723

Cabiati M., Fontanini M., Giacomarra M., Politano G., Randazzo E., Peroni D., Federico G., Del Ry S. Screening and identification of putative long non-coding RNA in childhood obesity: evaluation of their transcriptional levels. *Biomedicines*. 2022;10(3):529. doi 10.3390/biomedicines10030529

Capobianco V., Nardelli C., Ferrigno M., Iaffaldano L., Pilone V., Forestieri P., Zambrano N., Sacchetti L. miRNA and protein expression profiles of visceral adipose tissue reveal miR-141/YWHAG and miR-520e/RAB11A as two potential miRNA/protein target pairs associated with severe obesity. *J Proteome Res*. 2012;11(6):3358-3369. doi 10.1021/pr300152z

Carrieri C., Cimatti L., Biagioli M., Beugnet A., Zucchelli S., Fedele S., Pesce E., Ferrer I., Collavin L., Santoro C., Forrest A.R.R., Carninci P., Biffo S., Stupka E., Gustincich S. Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature*. 2012;491(7424):454-457. doi 10.1038/nature11508

Chan G.C.K., Than W.H., Kwan B.C.H., Lai K.B., Chan R.C.K., Teoh J.Y.C., Ng J.K.C., Chow K.M., Cheng P.M.S., Law M.C., Leung C.B., Li P.K.T., Szeto C.C. Adipose and plasma microRNAs miR-221 and 222 associate with obesity, insulin resistance, and new onset diabetes after peritoneal dialysis. *Nutrients*. 2022;14(22):4889. doi 10.3390/nu14224889

Chen Y., Li K., Zhang X., Chen J., Li M., Liu L. The novel long non-coding RNA lncRNA-Adi regulates adipogenesis. *Stem Cells Transl Med*. 2020;9(9):1053-1067. doi 10.1002/sctm.19-0438

Corral A., Alcala M., Carmen Duran-Ruiz M., Arroba A.I., Ponce-Gonzalez J.G., Todorčević M., Serra D., Calderon-Dominguez M., Herrero L. Role of long non-coding RNAs in adipose tissue metabolism and associated pathologies. *Biochem Pharmacol*. 2022;206:115305. doi 10.1016/j.bcp.2022.115305

Dedov I.I., Shestakova M.V., Melnichenko G.A., Mazurina N.V., Andreeva E.N., Bondarenko I.Z., Gusova Z.R., … Troshina E.A., Khamoshina M.V., Chechelnitskaya S.M., Shestakova E.A., Sheremet'eva E.V. Interdisciplinary clinical practice guidelines "management of obesity and its comorbidities". *Obesity and Metabolism*. 2021;18(1):5-99. doi 10.14341/omet12714 (in Russian)

Dexheimer P.J., Cochella L. MicroRNAs: from mechanism to organism. *Front Cell Dev Biol*. 2020;8:409. doi 10.3389/fcell.2020.00409

Ebrahimi R., Toolabi K., Jannat Ali Pour N., Mohassel Azadi S., Bahiraee A., Zamani-Garmsiri F., Emamgholipour S. Adipose tissue gene expression of long non-coding RNAs; MALAT1, TUG1 in obesity: is it associated with metabolic profile and lipid homeostasis-related genes expression? *Diabetol Metab Syndr*. 2020;12(1):36. doi 10.1186/s13098-020-00544-0

Fawzy M.S., Abdelghany A.A., Toraih E.A., Mohamed A.M. Circulating long noncoding RNAs H19 and GAS5 are associated with type 2 diabetes but not with diabetic retinopathy: a preliminary study. *Bosn J Basic Med Sci*. 2020;20(3):365-371. doi 10.17305/bjbms.2019.4533

Ferrer J., Dimitrova N. Transcription regulation by long non-coding RNAs: mechanisms and disease relevance. *Nat Rev Mol Cell Biol*. 2024;25(5):396-415. doi 10.1038/s41580-023-00694-9

Gouda W., Ahmed A.E., Mageed L., Hassan A.K., Afify M., Hamimy W.I., Ragab H.M., Maksoud N.A.E., Allayeh A.K., Abdel-

maksoud M.D.E. Significant role of some miRNAs as biomarkers for the degree of obesity. *J Genet Eng Biotechnol*. 2023;21(1):109. doi 10.1186/s43141-023-00559-w

Grossi I., Marchina E., De Petro G., Salvi A. The biological role and translational implications of the long non-coding RNA GAS5 in breast cancer. *Cancers* (*Basel*). 2023;15(13):3318. doi 10.3390/cancers15133318

Guo Z., Cao Y. An lncRNA-miRNA-mRNA ceRNA network for adipocyte differentiation from human adipose-derived stem cells. *Mol Med Rep*. 2019;19(5):4271-4287. doi 10.3892/mmr.2019.10067

Ji J., Zhao L., Zhao X., Li Q., An Y., Li L., Li D. Genome-wide DNA methylation regulation analysis of long non-coding RNAs in glioblastoma. *Int J Mol Med*. 2020;46(1):224-238. doi 10.3892/ijmm.2020.4579

Kim N.H., Ahn J., Choi Y.M., Son H.J., Choi W.H., Cho H.J., Yu J.H., Seo J.A., Jang Y.J., Jung C.H., Ha T.Y. Differential circulating and visceral fat microRNA expression of non-obese and obese subjects. *Clin Nutr*. 2020;39(3):910-916. doi 10.1016/j.clnu.2019.03.033

Kolenda T., Ryś M., Guglas K., Teresiak A., Bliźniak R., Mackiewicz J., Lamperska K. Quantification of long non-coding RNAs using qRT-PCR: comparison of different cDNA synthesis methods and RNA stability. *Arch Med Sci*. 2021;17(4):1006-1015. doi 10.5114/aoms.2019.82639

Lhamyani S., Gentile A.-M., Giráldez-Pérez R.M., Feijóo-Cuaresma M., Romero-Zerbo S.Y., Clemente-Postigo M., Zayed H., Oliva-Olivera W., Bermúdez-Silva F.J., Salas J., Gómez C.L., Hmadcha A., Hajji N., Olveira G., Tinahones F.J., El Bekay R. miR-21 mimic blocks obesity in mice: a novel therapeutic option. *Mol Ther Nucleic Acids*. 2021;26:401-416. doi 10.1016/j.omtn.2021.06.019

Li L., Wei H., Zhang Y.W., Zhao S., Che G., Wang Y., Chen L. Differential expression of long non-coding RNAs as diagnostic markers for lung cancer and other malignant tumors. *Aging*. 2021;13(20):23842-23867. doi 10.18632/aging.203523

Li W., Notani D., Rosenfeld M.G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet*. 2016;17(4):207-223. doi 10.1038/nrg.2016.4

Liu Y., Ji Y., Li M., Wang M., Yi X., Yin C., Wang S., Zhang M., Zhao Z., Xiao Y. Integrated analysis of long noncoding RNA and mRNA expression profile in children with obesity by microarray analysis. *Sci Rep*. 2018;8(1):8750. doi 10.1038/s41598-018-27113-w

Livak K.J., Schmittgen T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods*. 2001;25(4):402-408. doi 10.1006/meth.2001.1262

Luo Y., Guo J., Xu P., Gui R. Long non-coding RNA GAS5 maintains insulin secretion by regulating multiple miRNAs in INS-1 832/13 cells. *Front Mol Biosci*. 2020;7:559267. doi 10.3389/fmolb.2020.559267

Lustig R.H., Collier D., Kassotis C., Roepke T.A., Kim M.J., Blanc E., Barouki R., Bansal A., Cave M.C., Chatterjee S., Choudhury M., Gilbertson M., Lagadic-Gossmann D., Howard S., Lind L., Tomlinson C.R., Vondracek J., Heindel J.J. Obesity I: overview and molecular and biochemical mechanisms. *Biochem Pharmacol*. 2022;199:115012. doi 10.1016/j.bcp.2022.115012

Lv Y., Wang F., Sheng Y., Xia F., Jin Y., Ding G., Wang X., Yu J. Estrogen supplementation deteriorates visceral adipose function in aged postmenopausal subjects via Gas5 targeting IGF2BP1. *Exp Gerontol*. 2022;163:111796. doi 10.1016/j.exger.2022.111796

Ma B., Wang S., Wu W., Shan P., Chen Y., Meng J., Xing L., Yun J., Hao L., Wang X., Li S., Guo Y. Mechanisms of circRNA/lncRNA-miRNA interactions and applications in disease and drug research. *Biomed Pharmacother*. 2023;162:114672. doi 10.1016/j.biopha.2023.114672

Mameli G., Arru G., Caggiu E., Niegowska M., Leoni S., Madeddu G., Babudieri S., Sechi G.P., Sechi L.A. Natalizumab therapy modulates miR-155, miR-26A and proinflammatory cytokine expression in MS patients. *PLoS One*. 2016;11(6):e0157153. doi 10.1371/journal.pone.0157153

Markovic J., Sharma A.D., Balakrishnan A. MicroRNA-221: a fine tuner and potential biomarker of chronic liver injury. *Cells*. 2020;9(8):1767. doi 10.3390/cells9081767

Mattick J.S., Amaral P.P., Carninci P., Carpenter S., Chang H.Y., Chen L.-L., Chen R., … Spector D.L., Ulitsky I., Wan Y., Wilusz J.E., Wu M. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol*. 2023;24(6):430-447. doi 10.1038/s41580-022-00566-8

Mehta R., Birerdinc A., Hossain N., Afendy A., Chandhoke V., Younossi Z., Baranova A. Validation of endogenous reference genes for qRT-PCR analysis of human visceral adipose samples. *BMC Mol Biol*. 2010;11(1):39. doi 10.1186/1471-2199-11-39

Miao C., Zhang G., Xie Z., Chang J. MicroRNAs in the pathogenesis of type 2 diabetes: new research progress and future direction. *Can J Physiol Pharmacol*. 2018;96(2):103-112. doi 10.1139/cjpp-2017-0452

Mohanty V., Gokmen-Polar Y., Badve S., Janga S.C. Role of lncRNAs in health and disease – size and shape matter. *Brief Funct Genomics*. 2015;14(2):115-129. doi 10.1093/bfgp/elu034

Morrissy A.S., Griffith M., Marra M.A. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res*. 2011;21(8):1203-1212. doi 10.1101/gr.113431.110

Nicklas B.J., Penninx B.W.J.H., Ryan A.S., Berman D.M., Lynch N.A., Dennis K.E. Visceral adipose tissue cutoffs associated with metabolic risk factors for coronary heart disease in women. *Diabetes Care*. 2003;26(5):1413-1420. doi 10.2337/diacare.26.5.1413

Ortega F.J., Mercader J.M., Catalán V., Moreno-Navarrete J.M., Pueyo N., Sabater M., Gómez-Ambrosi J., Anglada R., Fernández-Formoso J.A., Ricart W., Frühbeck G., Fernández-Real J.M. Targeting the circulating microRNA signature of obesity. *Clin Chem*. 2013;59(5):781-792. doi 10.1373/clinchem.2012.195776

Ragni E., Colombini A., De Luca P., Libonati F., Viganò M., Perucca Orfei C., Zagra L., de Girolamo L. miR-103a-3p and miR-22-5p are reliable reference genes in extracellular vesicles from cartilage, adipose tissue, and bone marrow cells. *Front Bioeng Biotechnol*. 2021;9:632440. doi 10.3389/fbioe.2021.632440

Rasaei N., Gholami F., Samadi M., Shiraseb F., Khadem A., Yekaninejad M.S., Emamgholipour S., Mirzaei K. The interaction between MALAT1 and TUG1 with dietary fatty acid quality indices on visceral adiposity index and body adiposity index. *Sci Rep*. 2024;14(1):12. doi 10.1038/s41598-023-50162-9

Reddy K.B. MicroRNA (miRNA) in cancer. *Cancer Cell Int*. 2015;15(1):38. doi 10.1186/s12935-015-0185-1

Rupaimoole R., Slack F.J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov*. 2017;16(3):203-222. doi 10.1038/nrd.2016.246

Semaev S., Shakhtshneider E., Orlov P., Ivanoshchuk D., Malyutina S., Gafarov V., Ragino Y., Voevoda M. Association of RS708272 (*CETP* gene variant) with lipid profile parameters and the risk of myocardial infarction in the white population of Western Siberia. *Biomolecules*. 2019;9(11):739. doi 10.3390/biom9110739

Skårn M., Namløs H.M., Noordhuis P., Wang M.-Y., Meza-Zepeda L.A., Myklebost O. Adipocyte differentiation of human bone marrow-derived stromal cells is modulated by microRNA-155, microRNA-221, and microRNA-222. *Stem Cells Dev*. 2012;21(6):873-883. doi 10.1089/scd.2010.0503

Statello L., Guo C.-J., Chen L.-L., Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*. 2021;22(2):96-118. doi 10.1038/s41580-020-00315-9

Su X., Huang H., Lai J., Lin S., Huang Y. Long noncoding RNAs as potential diagnostic biomarkers for diabetes mellitus and complications: a systematic review and meta-analysis. *J Diabetes*. 2023;16(2):e13510. doi 10.1111/1753-0407.13510

Sufianov A., Beilerli A., Kudriashov V., Ilyasova T., Liang Y., Mukhamedzyanov A., Bessonova M., Mashkin A., Beylerli O. The role of long non-coding RNAs in the development of adipose cells. *Noncoding RNA Res*. 2023;8(2):255-262. doi 10.1016/j.ncrna.2023.02.009

Sun J., Ruan Y., Wang M., Chen R., Yu N., Sun L., Liu T., Chen H. Differentially expressed circulating lncRNAs and mRNA identified by microarray analysis in obese patients. *Sci Rep*. 2016;6(1):35421. doi 10.1038/srep35421

Tait S., Baldassarre A., Masotti A., Calura E., Martini P., Varì R., Scazzocchio B., Gessani S., Del Cornò M. Integrated transcriptome analysis of human visceral adipocytes unravels dysregulated microRNA-long non-coding RNA-mRNA networks in obesity and colorectal cancer. *Front Oncol*. 2020;10:1089. doi 10.3389/fonc.2020.01089

Tan L., Xie Y., Yuan Y., Hu K. LncRNA GAS5 as miR-26A-5p sponge regulates the PTEN/PI3K/Akt axis and affects extracellular matrix synthesis in degenerative nucleus pulposus cells *in vitro*. *Front Neurol*. 2021;12:653341. doi 10.3389/fneur.2021.653341

Tello-Flores V.A., Beltrán-Anaya F.O., Ramírez-Vargas M.A., Esteban-Casales B.E., Navarro-Tito N., Alarcón-Romero L.D.C., Luciano-Villa C.A., Ramírez M., Del Moral-Hernández Ó., Flores-Alfaro E. Role of long non-coding RNAs and the molecular mechanisms involved in insulin resistance. *Int J Mol Sci*. 2021;22(14):7256. doi 10.3390/ijms22147256

Valenzuela P.L., Carrera-Bastos P., Castillo-García A., Lieberman D.E., Santos-Lozano A., Lucia A. Obesity and the risk of cardiometabolic diseases. *Nat Rev Cardiol*. 2023;20(7):475-494. doi 10.1038/s41569-023-00847-5

Varkonyi-Gasic E., Wu R., Wood M., Walton E.F., Hellens R.P. Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods*. 2007;3(1):12. doi 10.1186/1746-4811-3-12

Veie C.H.B., Nielsen I.M.T., Frisk N.L.S., Dalgaard L.T. Extracellular microRNAs in relation to weight loss – a systematic review and meta-analysis. *Noncoding RNA*. 2023;9(5):53. doi 10.3390/ncrna9050053

Wang X., Guo P., Tian J., Li J., Yan N., Zhao X., Ma Y. LncRNA GAS5 participates in childhood pneumonia by inhibiting cell apoptosis and promoting SHIP-1 expression via downregulating miR-155. *BMC Pulm Med*. 2021;21(1):362. doi 10.1186/s12890-021-01724-y

Winkle M., El-Daly S.M., Fabbri M., Calin G.A. Noncoding RNA therapeutics – challenges and potential solutions. *Nat Rev Drug Discov*. 2021;20(8):629-651. doi 10.1038/s41573-021-00219-z

Yang W., Zhang K., Li L., Ma K., Hong B., Gong Y., Gong K. Discovery and validation of the prognostic value of the lncRNAs encoding snoRNAs in patients with clear cell renal cell carcinoma. *Aging*. 2020;12(5):4424-4444. doi 10.18632/aging.102894

Yumuk V., Tsigos C., Fried M., Schindler K., Busetto L., Micic D., Toplak H. European guidelines for obesity management in adults. *Obes Facts*. 2015;8(6):402-424. doi 10.1159/000442721

Zatterale F., Longo M., Naderi J., Raciti G.A., Desiderio A., Miele C., Beguinot F. Chronic adipose tissue inflammation linking obesity to insulin resistance and type 2 diabetes. *Front Physiol*. 2020;10:1607. doi 10.3389/fphys.2019.01607

Zeng H., Sun W., Ren X., Xia N., Zheng S., Xu H., Tian Y., Fu X., Tian J. AP2-microRNA-26a overexpression reduces visceral fat mass and blood lipids. *Mol Cell Endocrinol*. 2021;528:111217. doi 10.1016/j.mce.2021.111217

Zeng Z., Lan Y., Chen Y., Zuo F., Gong Y., Luo G., Peng Y., Yuan Z. LncRNA GAS5 suppresses inflammatory responses by inhibiting HMGB1 release via miR-155-5p/SIRT1 axis in sepsis. *Eur J Pharmacol*. 2023;942:175520. doi 10.1016/j.ejphar.2023.175520

Zhang J.-J., Ze-Xuan-Zhu, Guang-Min-Xu, Su P., Lei Q., Li W. Comprehensive analysis of differential expression profiles of long noncoding RNAs with associated co-expression and competing endogenous RNA networks in the hippocampus of patients with Alzheimer's disease. *Curr Alzheimer Res*. 2021;18(11):884-899. doi 10.2174/1567205018666211202143449

Zhang L., Wu H., Zhao M., Chang C., Lu Q. Clinical significance of miRNAs in autoimmunity. *J Autoimmun*. 2020;109:102438. doi 10.1016/j.jaut.2020.102438

Zhang M., Zhang Y.-Q., Wei X.-Z., Lee C., Huo D.-S., Wang H., Zhao Z.-Y. Differentially expressed long-chain noncoding RNAs in human neuroblastoma cell line (SH-SY5Y): Alzheimer's disease cell model. *J Toxicol Environ Health A*. 2019;82(19):1052-1060. doi 10.1080/15287394.2019.1687183

Zhang T., Liu H., Mao R., Yang H., Zhang Yuanchuan, Zhang Yu, Guo P., Zhan D., Xiang B., Liu Y. The lncRNA RP11-142A22.4 promotes adipogenesis by sponging miR-587 to modulate Wnt5β expression. *Cell Death Dis*. 2020;11(6):475. doi 10.1038/s41419-020-2550-9

Zhong Z., Hou J., Zhang Q., Li B., Li C., Liu Z., Yang M., Zhong W., Zhao P. Differential expression of circulating long non-coding RNAs in patients with acute myocardial infarction. *Medicine*. 2018;97(51):e13066. doi 10.1097/MD.0000000000013066

Zucchelli S., Cotella D., Takahashi H., Carrieri C., Cimatti L., Fasolo F., Jones M., Sblattero D., Sanges R., Santoro C., Persichetti F., Carninci P., Gustincich S. SINEUPs: a new class of natural and synthetic antisense long non-coding RNAs that activate translation. *RNA Biol*. 2015;12(8):771-779. doi 10.1080/15476286.2015.1060395

Institutional review board statement. The study was conducted in accordance with the Declaration of Helsinki. Study protocol was approved by the local Ethics Committee of the Institute of Internal and Preventive Medicine – Branch of the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia.

Informed consent statement. Informed consent was obtained from all subjects involved in the study.

Data availability statement. The data presented in this study are available on request from the corresponding author.

Conflict of interest. The authors declare no conflict of interest.

Received November 13, 2024. Revised February 15, 2024. Accepted February 17, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):458-466

doi 10.18699/vjgb-25-49

Genomic prediction of plant traits by popular machine learning methods

K.N. Kozlov [1], M.P. Bankin [1], E.A. Semenova[2], M.G. Samsonova [1]

[1] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia
[2] Far Eastern State Agrarian University, Blagoveshchensk, Amur region, Russia
m.g.samsonova@gmail.com

Abstract. A rapid growth of the available body of genomic data has made it possible to obtain extensive results in genomic prediction and identification of associations of SNPs with phenotypic traits. In many cases, to identify new relationships between phenotypes and genotypes, it is preferable to use machine learning, deep learning and artificial intelligence, especially explainable artificial intelligence, capable of recognizing complex patterns. 80 sources were manually selected; while there were no restrictions on the release date, the main attention was paid to the originality of the proposed approach for use in genomic prediction. The article considers models for genomic prediction, convolutional neural networks, explainable artificial intelligence and large language models. Attention is paid to Data Augmentation, Transfer Learning, Dimensionality Reduction methods and hybrid methods. Research in the field of model-specific and model-independent methods for interpretation of model solutions is represented by three main categories: sensing, perturbation, and surrogate model. The considered examples reflect the main modern trends in this area of research. The growing role of large language models, including those based on transformers, for genetic code processing, as well as the development of data augmentation methods, are noted. Among hybrid approaches, the prospect of combining machine learning models and models of plant development based on biophysical and biochemical processes is emphasized. Since the methods of machine learning and artificial intelligence are the focus of attention of both specialists in various applied fields and fundamental scientists, and also cause public resonance, the number of works devoted to these topics is growing explosively.

Key words: genomic prediction; plant phenotype; machine learning; deep learning; artificial intelligence

For citation: Kozlov K.N., Bankin M.P., Semenova E.A., Samsonova M.G. Genomic prediction of plant traits by popular machine learning methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(3):458-466. doi 10.18699/vjgb-25-49

Funding. The research is funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the World-Class Research Center program: Advanced Digital Technologies (agreement No. 075-15-2020-311 dated 04/20/2022).

Геномное прогнозирование признаков растений популярными методами машинного обучения

К.Н. Козлов [1], М.П. Банкин [1], Е.А. Семенова[2], М.Г. Самсонова [1]

[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия
[2] Дальневосточный государственный аграрный университет, Благовещенск, Амурская область, Россия
m.g.samsonova@gmail.com

Аннотация. Быстро накапливающийся массив геномных данных – секвенированных геномов сельскохозяйственных растений – позволил получить обширные результаты по геномному прогнозированию и выявлению ассоциаций однонуклеотидных полиморфизмов с фенотипическими признаками. Во многих случаях для обнаружения новых связей фенотипов с генотипами предпочтительно использовать методы машинного обучения, глубокого обучения и искусственного интеллекта, в особенности объяснимого, способные распознавать сложные закономерности. Вручную было отобрано 80 источников, при этом ограничения по дате выхода не ставилось, основной интерес представляла оригинальность предлагаемого подхода или модификации для применения в задаче геномного прогнозирования. В статье рассмотрены модели для геномного прогнозирования, сверточные нейронные сети, объяснимый искусственный интеллект и большие языковые модели. Уделено внимание подходам к дополнению данных, переносу знаний, методам снижения размерности и гибридным методам. Приведен пример современного способа кодирования больших геномных данных в искусственные изображения, преимуществом которых являются наглядная визуализация и возможность использования известных моделей для извлечения признаков. Исследования в области модель-

но-специфичных и модельно-независимых методов интерпретации решения моделей представлены тремя основными категориями: зондирование, возмущение и суррогатная модель. В рассмотренных примерах отражены основные современные тренды в изучаемой области. Отмечены растущая роль больших языковых моделей, в том числе основанных на трансформерах, для обработки генетического кода, а также разрабатываемые методы аугментации данных. Дополнительным преимуществом применения языковой модели может стать возможность формулировать запросы на близком к естественному языке и получать ответы за относительно короткое время. Среди гибридных подходов выделена перспективность сочетания моделей машинного обучения и моделей развития растений на основе биофизических и биохимических процессов. Поскольку методы машинного обучения и искусственного интеллекта находятся в фокусе внимания как специалистов в различных прикладных областях, так и фундаментальных ученых, а кроме того, вызывают общественный резонанс, количество посвященных этим темам работ имеет взрывной рост.
Ключевые слова: геномное прогнозирование; фенотип растений; машинное обучение; глубокое обучение; искусственный интеллект

Introduction

To this day, a tremendous amount of genomic data has been accumulated and it continues to grow rapidly. These data include the sequenced genomes of agricultural plants such as chickpea, vigna, soybean, wheat, rye, flax etc. (Bragina et al., 2019; Ichihara et al., 2023; Chamorro-Padial et al., 2024; Tang et al., 2024). Many annotations have been obtained, classical methods of genomic prediction and genome-wide association studies have been successfully applied to these data, and SNPs associated with different important phenotypes have been identified (Hayes, 2013).

Many phenotypic traits that selection programs are targeted to are correlated and thus require use of multi-trait models in order to obtain statistically significant predictions. Machine learning methods are suitable for such a class of problems as well as deep learning models and artificial intelligence, explainable AI in particular, which are able to recognize complex patterns in the given data and generalize extracted knowledge.

The papers for the current review were selected based on the originality of the proposed approach or modification for application to the solution of the genomic prediction problem. The search was performed in PubMed (https://pubmed.ncbi.nlm.nih.gov/, accessed on November 7, 2024) using terms “plants genomic prediction machine learning” and dates from the beginning of the year 2010 to the end of the year 2024, which showed exponential growth of the number of manuscripts per year with a small decrease in the growth rate after the year 2021 (Fig. 1).

Eighty sources were selected manually without restrictions on the publication date. The oldest manuscript was published in the year 1988, the majority of works (60 %) were published after the year 2020, and 20 % of the reviewed papers belong to the last two years (Fig. 2).

List of abbreviations

SNP – single nucleotide polymorphism
GP – genomic prediction
GBLUP – genomic best linear unbiased predictor
ML – machine learning
RRBLUP – ridge regression with best linear unbiased predictor
CNN – convolutional neural network
AIO – artificial image object
PCA – principal component analysis
XAI – Explainable Artificial Intelligence
DT – decision trees
RF – random forest
LLM – large language model
GPT – generative pretrained transformer

Genomic prediction

Genomic prediction (GP) aims to predict the phenotype of an organism given single nucleotide polymorphism (SNP) data (Meuwissen et al., 2001). The wide range of genomic prediction methods can be divided into two groups: linear and nonparametric. Linear methods such as BLUP work well for additive traits. They model the phenotype as a function of the contributions of different factors such as individual markers, weather parameters, field conditions, etc. On the other hand, nonparametric machine learning methods such as support vector machines, random forests, and gradient boosting can model nonlinear traits, providing great flexibility to accommodate complex genotype-phenotype associations (Montesinos-López et al., 2021).

Genomic prediction tools based on statistical methods such as genomic best linear unbiased prediction (GBLUP) are widely used in crop breeding. However, these tools are not designed to account for nonlinear relationships in high-dimensional datasets or to handle high-dimensional datasets such as drone images. Machine learning (ML) algorithms have the potential to surpass the prediction accuracy of current tools used to predict phenotypic traits from genomic data due to their ability to autonomously extract features and represent their relationships at multiple levels of abstraction (Danilevicz et al., 2022).

The accuracy of prediction depends on the quality and pre-processing of phenotypic data, the platform used to obtain genomic information, the population breeding scheme, the internal genetic architecture of the trait, the genetic structure of the population, how genotype-environment interactions are treated, and the prediction method (de Los Campos et al., 2013).

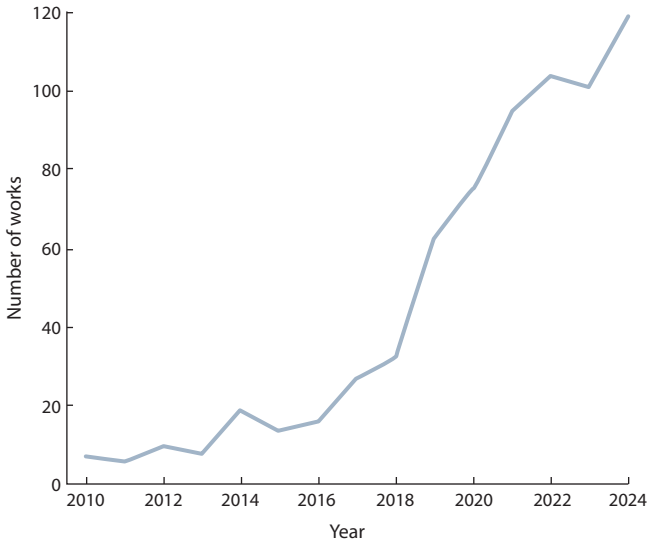


Fig. 1. The growth of the number of works in PubMed.

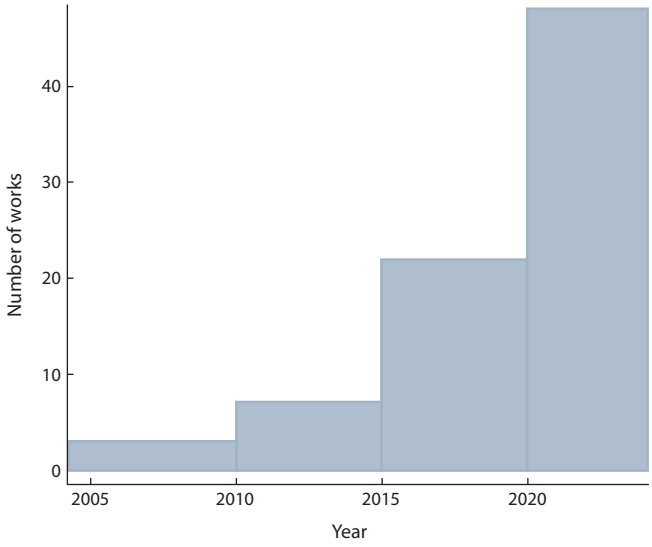


Fig. 2. The distribution of the selected works over the years.

It was reported in (Sandhu et al., 2021) that deep learning models outperformed traditional ridge regression with best linear unbiased prediction (RRBLUP) and Bayesian models under all forecasting scenarios. Machine learning methods were used to increase the statistical power of the models. To apply multi-stage machine learning, a new BioM2 package (Zhang S. et al., 2024) was proposed for the statistical computing system R, which has the ability to apply stratification and aggregation of multivariate data based on biological information to improve the training efficiency of models. In this case, stratification allows one to build subsets of data, for example, training and test samples, by controlling the ratio of the number of objects from different groups, for example, SNPs in genes involved in different processes.

At the same time, aggregation of multivariate data makes it possible to use simpler and more easily interpretable models that can be refined during multi-stage training. An innovative computational framework, PlantMine, which combines feature selection and machine learning methods to efficiently identify key SNPs, was proposed in (Tong et al., 2024), taking critical factors for trait improvement in rice as an example. Various data mining algorithms were applied to the 3,000 Rice Genomes Project dataset. The results highlighted the effectiveness of combining feature selection with machine learning to accurately identify key SNPs, offering a promising avenue to accelerate the breeding of new plant varieties with improved yield and stress tolerance. The overall model performance depended more on the prediction algorithm than the predictor selection method. Among all the models, decision tree-based machine learning methods (random forests and gradient boosting) performed the best, while classical Bayesian methods were prone to overfitting (Sirsat et al., 2022).

Convolutional neural networks and artificial image objects

Among machine learning methods, convolutional neural networks (CNNs) provide the best ability to discover hidden patterns or features from data and are best suited for image analysis (Pook et al., 2020; Montesinos-López et al., 2021). Artificial image objects (AIO) are a new concept for genomic data representation that can be used to encode large genomic data by treating individual genetic variants as pixels (Galli et al., 2022). The advantages of AIOs include convenient, simple visualization, compactness, and the ability to apply a wide range of methods developed for image analysis and classification (Chen X. et al., 2021b), in particular CNNs (Chen X. et al., 2021a). Therefore, AIOs can be used by CNNs for regression and classification tasks (Bavykina et al., 2022).

The algorithm for optimization of data packing in AIO was proposed in (Bazgir et al., 2020). The DeepFeature package proposed in (Sharma et al., 2019, 2021) was developed to transform large-scale experimental data, such as genomic or transcriptomic data, into a form optimal for training a CNN model. The input vector is transformed into a matrix using t-SNE, kernel PCA, PHATE, or UMAP, and the smallest rectangle containing all elements is found using the convex hull algorithm. A rotation is performed to flatten the image, converting Cartesian coordinates into pixel indices.

The application of CNN to AIO processing enables the calculation and visualization of the influence of various factors on the final solution of the model. The work of (Liu et al., 2019) was considered to be the first study to apply the saliency map to identify the most important predictors in soybean. In this study, gaps in the data were treated as a new genotype; as a result, each SNP was encoded with four binary values. The significance value of each geno-

type was calculated as the maximum absolute value of the gradients among these four encoding channels, and the population median was used as a measure of the contribution of the SNP.

The ResNet architecture, widely used in deep learning methods, was adapted for use in genomic selection models in (Xie et al., 2024). Since each locus makes a different contribution to the final phenotype, successive convolutions are more suitable for the genomic selection model than layer pooling. Thus, a deep learning algorithm, ResGS, was proposed that significantly alleviates the problem of degradation, i. e., the decrease in performance with increasing model depth, which can improve the prediction accuracy compared to traditional methods (Wu H. et al., 2024).

Recently, more and more attention has been paid to the internal mechanisms of convolutional neural networks and the reasons why the network makes certain decisions (Wang et al., 2020). Several methods have been proposed, including data permutation and backpropagation approaches (Zhang X., Gao, 2020), gradient-based algorithms (Selvaraju et al., 2020), and class activation maps (Wang et al., 2020). A saliency map represents the spatial regions associated with a particular class in a given image (Simonyan et al., 2014). Class activation maps provide a visual explanation for a single input image (Chattopadhay et al., 2018; Selvaraju et al., 2020), but are sensitive to the model architecture. Gradient-weighted class activation mapping (Grad-CAM) uses the gradients of any target concept fed to the final convolutional layer to create a coarse localization map, which highlights important regions in the image for class prediction (Selvaraju et al., 2017).

Score-CAM, unlike previous class activation mapping-based approaches, removes the dependence on gradients by deriving the weights of each activation map by directly computing the network for instances of the target class, with the final output being a linear combination of the weights and activation maps (Wang et al., 2020). Grad-CAM++ (Chattopadhay et al., 2018), a modification of Grad-CAM (Selvaraju et al., 2020), generalizes CAM to models without global pooling layers. LayerCAM (Jiang et al., 2021) can generate robust class activation maps from a combination of class activation maps from different CNN layers.

Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) aims to overcome the black box problem and provide insight into how AI systems make decisions. Interpretable ML models can explain how they make predictions and identify the factors that influence their results. However, most modern interpretable ML methods were developed for domains such as computer vision, making direct application to bioinformatics problems difficult without customization and domain adaptation.

An interpretable ML model can identify the factors that influence its output (e. g. statistically significant features) and explain the interactions between them (Molnar, 2022). Depending on the level of abstraction, methods can be divided into local and global interpretability methods. While local methods focus on interpreting individual predictions, global ones try to explain the behavior of the entire model in the form of diagrams or lists. Various variants of model-specific and model-independent interpretable ML approaches have been developed, on which an XAI system can be built to improve its local and global interpretability (Wachter et al., 2018), but these methods are most often used to improve visualization (Weber et al., 2023). Linear models, decision trees (DTs), and rule-based systems are less complex and inherently interpretable. However, they are less accurate compared to tree-based ensembles such as random forests (RF) and deep neural networks, resulting in a trade-off between accuracy and interpretability.

Many specific and model-independent interpretable ML methods have been developed (Azodi et al., 2020). All these methods can be divided into three main categories: probing, perturbation, and surrogate models. Examples of probing methods are gradient-based methods such as gradient-weighted class activation mapping (Grad-CAM++) and layered relevance propagation (LRP) (Guidotti et al., 2018). A widely used perturbation-based method is Shapley additive explanations (SHAP). SHAP is based on coalition game theory, i. e., on the average marginal contribution of a feature and the way the payoffs are distributed among its players (Cubitt, 1991).

Since interpretability comes at the cost of a trade-off between accuracy and complexity, studies have proposed training a simple interpretable model to imitate a complex model (Molnar, 2022). A surrogate or simple proxy model is a model interpretation strategy that involves training an initially interpretable model by approximating local black box predictions (Stiglic et al., 2020; Molnar, 2022).

The majority of surrogate model building tools were developed with the aim of improving the interpretability and explainability of black-box ML models covering common problems in computer vision, text mining or structured data, and were based on well-known interpretable ML methods such as LIME (Ribeiro et al., 2016), Model Understanding through Subspace Explanations (MUSE) (Lakkaraju et al., 2019), SHAP (Lundberg, Lee, 2017) (and its variants such as SHAP kernel and SHAP tree), Partial Dependency Graph (PDP), Individual Conditional Expectation (ICE), Permutation Feature Importance (PFI) and Counterfactual Explanations (CE) (Wachter et al., 2018).

Large language models

Recently, the use of large language models (LLM) has become widespread in various fields of science, including decoding genetic text to predict the manifestation of useful traits in plants. LLMs, such as GPT-4, have conquered the world, demonstrating amazing capabilities in natural language proficiency, which immediately prompted researchers to adapt LLMs to a different type of language – the genome, in order to solve complex problems based on

large volumes of accumulated data. The success of LLMs is largely due to the use of transformer-based attention units in the architecture. The use of such architectural solutions allowed the well-known AlphaFold2 neural network (Google DeepMind, 2021) to predict three-dimensional protein structures with unprecedented accuracy. AlphaFold3 (2024), according to the developers, for the first time surpasses physical methods in its prediction of the 3D structure of proteins, as well as the interactions of proteins with each other and with other substances. Profluent's LLM has made it possible to create an artificial protein for genome editing that is comparable in efficiency to the natural one, but has much greater specificity.

The broad implementation of the results of these achievements in production requires a deep understanding of the underlying mechanisms, taking into account complex interactions, accelerating the search for answers to questions arising in practice. In particular, there is a need to shift from identification of SNPs associated with a trait to identification of genes that affect the trait with a greater degree of reliability. In addition, it is necessary to take into account the gene-gene interactions, and to consider not only one trait, but also pairs of related traits. The solution to the described problem is impossible without involving the latest accomplishments in computer science, such as artificial intelligence based on large language models. An additional advantage of using such an approach is the ability to formulate queries in a language close to a natural one and receive answers in a relatively short time.

Research in this area has increased significantly in recent years. For example, a review (Consens et al., 2023) on the application of transformer-like models to genetic data included more than 100 recent papers and noted rapid development in the field. The use of large language models based not only on transformers, but also using the so-called Hyena layer (Poli et al., 2023) to process genomic data was also noted (Nguyen et al., 2023). One interesting approach is the possibility of pre-training such models on genome sequences without using phenotypes.

Currently, the maximum input sequence length among publicly available DNA transformer-based LLMs is limited to only 3×10^4 nucleotides for the GENA-LM architecture. To mitigate this limitation, the performance of a modified recurrent memory transformer (RMT) architecture in the GENA-LM model was studied in (Kuratov et al., 2024) for multiple genomic analysis tasks requiring processing of long DNA sequences. The results obtained in (Kuratov et al., 2024) showed that augmenting GENA-LM with RMT leads to a significant performance improvement.

A new method based on a transformer-like neural network to predict the severity of fusarium and the associated accumulation of the dangerous mycotoxin deoxynivalenol was proposed (Jubair et al., 2021) that used genomic and phenotypic data on the barley. The work showed the superiority of frequency coding of markers and mentioned the high memory requirements of the model when using a large number of markers, which could be reduced using selection by the information criterion.

In the paper (Wu C. et al., 2023), a genomic selection model based on a deep neural network using transformers, convolutional layers, and an additional information module was proposed. The model architecture used encoding of marker positions with trigonometric functions, fast Fourier transform, Gaussian linear activation function (GELU), and included blocks of convolutional network, transformer, and regressor. The model was applied to five datasets, where it outperformed the four methods used for comparison.

An important source of the phenotype prediction accuracy reduction in models based on genomic data is the lack of gene-gene interactions consideration. The work (Cui et al., 2022) proposed an approach for identifying interactions between genes and taking them into account in a deep learning model for phenotype prediction. A layer representing genes as hidden nodes of a sparse network was added to the deep neural network architecture. Importantly, the Shapley values for hidden nodes of the gene layer were used to determine the influence of interactions on the model solution.

Data augmentation

Training large language models requires a large amount of data because there is a large number of unknown parameters. The papers (Jubair et al., 2021) and (Wu C. et al., 2023) consider transformer-like neural network-based models for genomic prediction. In the paper (Jubair et al., 2021), GPTransformer contains two Transformer encoding blocks, uses two nodes for each attention layer, and each Transformer block contains 256 hidden neurons. The output is a vector, which is the input of a feedforward network, which contains one output neuron. The mean squared error (MSE) loss function is used. A dataset of 400 genotypes phenotyped in 3 geographic areas and 2 years, i. e. 2,400 records, was used for training and analysis, and the Pearson correlation coefficient between the model prediction and the data was 0.6, which allowed obtaining significant results.

The GPformer model (Wu C. et al., 2023), based on the transformer-like neural network for predicting phenotype from genotype, was separately trained and tested on the Soybean999, Maize282, Rice469, Wheat599 and Wheat2403 datasets, which have 999, 282, 469, 599 and 2,403 records, respectively. The resulting Pearson correlation coefficient was 0.4–0.8 for different variants.

An additional tool, as in the case of deep learning models for image processing, can be data augmentation, which has recently been studied for deep learning models in the field of bioinformatics. For example, a new approach to augmentation of biological sequence data was proposed in (Ji et al., 2024), in which the chromosome order is changed. This method of generating additional data can be used for training, because the models cannot use the chromosome number as a predictor. In the work (Montesinos-López et

al., 2024) a blending method was considered, which offers a domain-independent approach to augmentation based on the assumption that a linear combination of feature vectors should approximately correspond to a linear combination of their corresponding target values. In (Vilov, Heinig, 2022), data augmentation was successfully used to train a classifier of genomic variants. The approaches based on a generative network (GAN) and a Boltzmann machine (RBM) for compiling synthetic genomes were presented in (Yelmen et al., 2021). In the mentioned works, the authors managed to improve the accuracy and generalization ability of the models, so data augmentation can be used to expand the existing dataset for training the LLM.

A new method was proposed to predict the classification of enhancers into strong and weak using data augmentation and a convolutional neural network ES-ARCNN (Zhang T.-H. et al., 2021). Two data augmentation techniques, such as reverse augmentation and shifting, were used to train ES-ARCNN for previously identified enhancers.

Transfer learning

Transfer learning enables the creation of effective models for a target domain using knowledge from a different but related source domain. In medical research, knowledge transfer can significantly improve the accuracy of disease prediction for data-poor populations with imbalanced data (Gao, Cui, 2022). This approach also has great potential to improve the prediction of complex phenotypic traits, such as plant yield, although it does not work in all cases (Kovalev et al., 2018). Transfer learning is widely used to extract features from images with the models pre-trained on general-purpose datasets and then fine-tuned on a relatively limited, specialized dataset (Kirchler et al., 2022).

To facilitate the application of the Transfer learning approach to phenotype-to-genotype prediction models, an efficient implementation of TrG2P was proposed in (Li et al., 2024). In the developed framework, firstly, convolutional neural networks were trained using genomic data and phenotypic traits with simpler dependencies than a complex target trait, such as yield. Then, the parameters of the convolutional layers of these pre-trained models were transferred to the target trait prediction task, and the fully connected layers were retrained, thus leading to improved prediction accuracy of the resulting model (Li et al., 2024).

Dimensionality reduction methods

The explosive growth of available amounts of data not only brings unprecedented progress in bioinformatics and opportunities to perform predictive modeling (Han, Liu, 2022), but also poses challenges to existing AI methods and tools, such as data heterogeneity, high dimensionality, and volume (Karim et al., 2021). Principal component analysis (PCA) and isometric feature mapping (Isomap) are widely used as dimensionality reduction methods (Fournier, Aloise, 2019). However, the representations obtained by these methods often lose essential properties (Aggarwal, Reddy, 2014), making them less effective against a well-known phenomenon called the curse of dimensionality, especially for high-dimensional datasets (Fournier, Aloise, 2019).

Hybrid methods

With increasing computing power, existing machine learning approaches are frequently combined into complex hybrid models. For example, (Chen C. et al., 2024) considered algorithms that first use BayesR/GWAS to identify a subset of 1,000 markers with moderate to large marginal additive effects, and then use attention networks to make predictions based on these effects and their interactions. Hybrid methods with attention networks yielded the lowest variance in prediction accuracy across all validation datasets and the lowest root mean square error, the criteria usually applied in practical breeding programs. In (Ramzan et al., 2020), a two-step procedure was proposed to solve the problem of detecting a large number of loci with small effects on the phenotype. In the first step, the Wald test statistics values are approximated by cubic splines, and genomic regions with spline's extrema that are higher than expected are considered as quantitative trait loci (QTLs). SNPs in these QTLs are then ranked by their association with the phenotype using a random forest approach. In the work (Nascimento et al., 2024), a Stacking Ensemble Learning (SEL) model was proposed, which combines several models that can potentially predict important traits more accurately than individual ones; the model was applied to the example of coffee breeding in *Coffea arabica*.

A recently proposed direction of research is the combination of machine learning models and crop growth models based on biophysical and biochemical processes (CGM). It has been suggested that such an approach can improve the predictions of integrative traits by decomposing them into simpler intermediate traits with better heritability (Larue et al., 2024). In the study, the combined CGM-GP model outperformed the genomic selection models without CGM integration in the predictive ability, regardless of the regression method used. CGM simulates nonlinear (causal) plant responses to the environment through model parameters (representing genotypic sensitivity to these responses, G×E). Thus, calibrated CGMs for a genotype can be useful for predicting its performance under unknown conditions; on the other hand, it is impossible to predict the performance of unknown genotypes (Larue et al., 2019).

Conclusions

The great variety of machine learning and artificial intelligence methods finds applications in the field of bioinformatics of agricultural plants for such problems as genomic prediction of important phenotypic traits. ML and AI attract close attention of researchers and practitioners from different areas as well as cause resonance in the public, and consequently the number of published manuscripts grows explosively.

The main contemporary trends in the field of ML and AI for GP were included in the review. The examples of the application of common machine learning models and their variants modified for bioinformatics tasks were considered. These examples illustrated the usage of the ML and AI methods alone and in combination with dimensionality reduction and feature selection approaches, the construction of explainable AI solutions and developing hybrid methods. The increasing role of large language models deserves a separate mention, including those based on transformers, and the associated data augmentation methods needed to train models with a huge number of parameters. Transfer learning methods can be used to mitigate the problem of insufficient or imbalanced data.

An important aspect of ML and AI success is data representation, for example, the artificial image objects described in the review make it possible to utilize the powerful and highly efficient apparatus of convolutional neural networks for extraction of characteristic patterns from the data. Such an approach also allows ranking the importance of predictors based on attention maps.

With the rise of the Internet of things, the spread of mobile devices and autonomous robots, a new trend of edge computing started to evolve, seeking solutions to the compactization of models and optimization of algorithms for resource-limited devices. This topic deserves a separate review and was not considered in the current work.

References

Aggarwal C.C., Reddy C.K. (Eds) Data Clustering: Algorithms and Applications. New York: Chapman and Hall/CRC, 2014. doi 10.1201/9781315373515

Azodi C.B., Tang J., Shiu S.-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet*. 2020;36(6):442-455. doi 10.1016/j.tig.2020.03.005

Bavykina M., Kostina N., Lee C.-R., Schafleitner R., Bishop-von Wettberg E., Nuzhdin S.V., Samsonova M., Gursky V., Kozlov K. Modeling of flowering time in *Vigna radiata* with artificial image objects, convolutional neural network and random forest. *Plants*. 2022; 11(23):3327. doi 10.3390/plants11233327

Bazgir O., Zhang R., Dhruba S.R., Rahman R., Ghosh S., Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat Commun*. 2020;11(1):4391. doi 10.1038/s41467-020-18197-y

Bragina M.K., Afonnikov D.A., Salina E.A. Progress in plant genome sequencing: research directions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2019;23(1):38-48. doi 10.18699/VJ19.459 (in Russian)

Chamorro-Padial J., García R., Gil R. A systematic review of open data in agriculture. *Comput Electron Agric.* 2024;219:108775. doi 10.1016/j.compag.2024.108775

Chattopadhay A., Sarkar A., Howlader P., Balasubramanian V.N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA. IEEE, 2018;839-847. doi 10.1109/WACV.2018.00097

Chen C., Bhuiyan S.A., Ross E., Powell O., Dinglasan E., Wei X., Atkin F., Deomano E., Hayes B. Genomic prediction for sugarcane diseases including hybrid Bayesian-machine learning approaches. *Front Plant Sci*. 2024;15:1398903. doi 10.3389/fpls.2024.1398903

Chen X., Chen D.G., Zhao Z., Balko J.M., Chen J. Artificial image objects for classification of breast cancer biomarkers with transcriptome sequencing data and convolutional neural network algorithms. *Breast Cancer Res*. 2021a;23(1):96. doi 10.1186/s13058-021-01474-z

Chen X., Chen D.G., Zhao Z., Zhan J., Ji C., Chen J. Artificial image objects for classification of schizophrenia with GWAS-selected SNVs and convolutional neural network. *Patterns*. 2021b;2(8):100303. doi 10.1016/j.patter.2021.100303

Consens M.E., Dufault C., Wainberg M., Forster D., Karimzadeh M., Goodarzi H., Theis F.J., Moses A., Wang B. To transformers and beyond: large language models for the genome. *arXiv*. 2023. doi 10.48550/arXiv.2311.07621

Cubitt R. The Shapley value: essays in honor of Lloyd S. Shapley. *Econ J*. 1991;101(406):644-646. doi 10.2307/2233574

Cui T., El Mekkaoui K., Reinvall J., Havulinna A.S., Marttinen P., Kaski S. Gene-gene interaction detection with deep learning. *Commun Biol*. 2022;5(1):1238. doi 10.1038/s42003-022-04186-y

Danilevicz M.F., Gill M., Anderson R., Batley J., Bennamoun M., Bayer P.E., Edwards D. Plant genotype to phenotype prediction using machine learning. *Front Genet.* 2022;13:822173. doi 10.3389/fgene.2022.822173

de Los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D., Calus M.P.L. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*. 2013;193(2):327-345. doi 10.1534/genetics.112.143313

Fournier Q., Aloise D. Empirical comparison between autoencoders and traditional dimensionality reduction methods. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy. IEEE, 2019;211-214. doi 10.1109/AIKE.2019.00044

Galli G., Sabadin F., Yassue R.M., Galves C., Carvalho H.F., Crossa J., Montesinos-López O.A., Fritsche-Neto R. Automated machine learning: a case study of genomic "image-based" prediction in maize hybrids. *Front Plant Sci*. 2022;13:845524. doi 10.3389/fpls.2022.845524

Gao Y., Cui Y. Deep transfer learning provides a *Pareto* improvement for multi-ancestral clinico-genomic prediction of diseases. *bioRxiv*. 2022. doi 10.1101/2022.09.22.509055

Guidotti R., Monreale A., Ruggieri S., Pedreschi D., Turini F., Giannotti F. Local rule-based explanations of black box decision systems. *arXiv*. 2018. doi 10.48550/arXiv.1805.10820

Han H., Liu X. The challenges of explainable AI in biomedical data science. *BMC Bioinformatics*. 2022;22(Suppl. 12):443. doi 10.1186/s12859-021-04368-1

Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). In: Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology. Vol. 1019. Totowa, NJ: Humana Press, 2013;149-169. doi 10.1007/978-1-62703-447-0_6

Ichihara H., Yamada M., Kohara M., Hirakawa H., Ghelfi A., Tamura T., Nakaya A., … Komaki A., Fawcett J.A., Sugihara E., Tabata S., Isobe S.N. Plant GARDEN: a portal website for cross-searching between different types of genomic and genetic resources in a wide variety of plant species. *BMC Plant Biol*. 2023:23(1);391. doi 10.1186/s12870-023-04392-8

Ji L., Hou W., Xiong L., Zhou H., Liu C., Li L., Yuan Z. GSCNN: a genomic selection convolutional neural network model based on SNP genotype and physical distance features and data augmentation strategy. *Res Square*. 2024. doi 10.21203/rs.3.rs-3991262/v1

Jiang P.-T., Zhang C.-B., Hou Q., Cheng M.-M., Wei Y. LayerCAM: exploring hierarchical class activation maps for localization. *IEEE Trans Image Process*. 2021;30:5875-5888. doi 10.1109/TIP.2021.3089943

Jubair S., Tucker J.R., Henderson N., Hiebert C.W., Badea A., Domaratzki M., Fernando W.G.D. GPTransformer: a transformer-based deep learning method for predicting Fusarium related traits in barley. *Front Plant Sci.* 2021;12:761402. doi 10.3389/fpls.2021.761402

Karim M.R., Beyan O., Zappa A., Costa I.G., Rebholz-Schuhmann D., Cochez M., Decker S. Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform*. 2021;22(1):393-415. doi 10.1093/bib/bbz170

Kirchler M., Konigorski S., Norden M., Meltendorf C., Kloft M., Schurmann C., Lippert C. transferGWAS: GWAS of images using deep transfer learning. *Bioinformatics*. 2022;38(14):3621-3628. doi 10.1093/bioinformatics/btac369

Kovalev M.S., Igolkina A.A., Samsonova M.G., Nuzhdin S.V. A pipeline for classifying deleterious coding mutations in agricultural plants. *Front Plant Sci*. 2018;9:1734. doi 10.3389/fpls.2018.01734

Kuratov Y., Shmelev A., Fishman V., Kardymon O., Burtsev M. Recurrent memory augmentation of GENA-LM improves performance on long DNA sequence tasks. In: Workshop Machine Learning for Genomics Explorations (MLGenX). 2024. Available: https://openreview.net/pdf?id=K671lCX90x

Lakkaraju H., Kamar E., Caruana R., Leskovec J. Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). New York, NY, USA: Association for Computing Machinery, 2019;131-138. doi 10.1145/3306618.3314229

Larue F., Fumey D., Rouan L., Soulié J.-C., Roques S., Beurier G., Luquet D. Modelling tiller growth and mortality as a sink-driven process using *Ecomeristem*: implications for biomass sorghum ideotyping. *Ann Bot*. 2019;124(4):675-690. doi 10.1093/aob/mcz038

Larue F., Rouan L., Pot D., Rami J.-F., Luquet D., Beurier G. Linking genetic markers and crop model parameters using neural networks to enhance genomic prediction of integrative traits. *Front Plant Sci*. 2024;15:1393965. doi 10.3389/fpls.2024.1393965

Li J., Zhang D., Yang F., Zhang Q., Pan S., Zhao X., Zhang Qi., Han Y., Yang J., Wang K., Zhao C. TrG2P: a transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Commun*. 2024;5(7):100975. doi 10.1016/j.xplc.2024.100975

Liu Y., Wang D., He F., Wang J., Joshi T., Xu D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front Genet*. 2019;10:1091. doi 10.3389/fgene.2019.01091

Lundberg S., Lee S.-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook, NY, USA: Curran Associates Inc., 2017;4768-4777. doi 10.48550/arXiv.1705.07874

Meuwissen T.H., Hayes B.J., Goddard M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819-1829. doi 10.1093/genetics/157.4.1819

Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Independently published, 2022

Montesinos-López O.A., Montesinos-López A., Mosqueda-Gonzalez B.A., Montesinos-López J.C., Crossa J., Ramirez N.L., Singh P., Valladares-Anguiano F.A. A zero altered Poisson random forest model for genomic-enabled prediction. *G3* (*Bethesda*). 2021;11(2):jkaa057. doi 10.1093/g3journal/jkaa057

Montesinos-López O.A., Solis-Camacho M.A., Crespo-Herrera L., Saint Pierre C., Huerta Prado G.I., Ramos-Pulido S., Al-Nowibet K., Fritsche-Neto R., Gerard G., Montesinos-López A., Crossa J. Data augmentation enhances plant-genomic-enabled predictions. *Genes*. 2024;15(3):286. doi 10.3390/genes15030286

Nascimento M., Nascimento A.C.C., Azevedo C.F., de Oliveira A.C.B., Caixeta E.T., Jarquin D. Enhancing genomic prediction with Stacking Ensemble Learning in Arabica Coffee. *Front Plant Sci*. 2024;15:1373318. doi 10.3389/fpls.2024.1373318

Nguyen E., Poli M., Faizi M., Thomas A., Birch-Sykes C., Wornow M., Patel A., Rabideau C., Massaroli S., Bengio Y., Ermon S., Baccus S.A., Ré C. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Red Hook, NY, USA: Curran Associates Inc., 2023;43177-43201. doi 10.48550/arXiv.2306.15794

Poli M., Massaroli S., Nguyen E., Fu D.Y., Dao T., Baccus S., Bengio Y., Ermon S., Ré C. Hyena hierarchy: towards larger convolutional language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML '23). Vol. 202. JMLR.org, 2023;28043-28078. doi 10.48550/arXiv.2302.10866

Pook T., Freudenthal J., Korte A., Simianer H. Using local convolutional neural networks for genomic prediction. *Front Genet*. 2020;11:561497. doi 10.3389/fgene.2020.561497

Ramzan F., Gültas M., Bertram H., Cavero D., Schmitt A.O. Combining random forests and a signal detection method leads to the robust detection of genotype-phenotype associations. *Genes* (*Basel*). 2020;11(8):892. doi 10.3390/genes11080892

Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York, NY, USA: Association for Computing Machinery, 2016;1135-1144. doi 10.1145/2939672.2939778

Sandhu K., Patil S.S., Pumphrey M., Carter A. Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. *Plant Genome*. 2021;14(3):e20119. doi 10.1002/tpg2.20119

Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. IEEE, 2017;618-626. doi 10.1109/ICCV.2017.74

Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359. doi 10.1007/s11263-019-01228-7

Sharma A., Vans E., Shigemizu D., Boroevich K.A., Tsunoda T. DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep*. 2019;9(1): 11399. doi 10.1038/s41598-019-47765-6

Sharma A., Lysenko A., Boroevich K.A., Vans E., Tsunoda T. DeepFeature: feature selection in nonimage data using convolutional neural network. *Brief Bioinform*. 2021;22(6):bbab297. doi 10.1093/bib/bbab297

Simonyan K., Vedaldi A., Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv*. 2014. doi 10.48550/arXiv.1312.6034

Sirsat M.S., Oblessuc P.R., Ramiro R.S. Genomic prediction of wheat grain yield using machine learning. *Agriculture*. 2022;12(9):1406. doi 10.3390/agriculture12091406

Stiglic G., Kocbek P., Fijacko N., Zitnik M., Verbert K., Cilar L. Interpretability of machine learning based prediction models in healthcare. *WIREs Data Min Knowl Discovery*. 2020;10(5):e1379. doi 10.1002/widm.1379

Tang F.H.M., Nguyen T.H., Conchedda G., Casse L., Tubiello F.N., Maggi F. CROPGRIDS: a global geo-referenced dataset of 173 crops. *Sci Data*. 2024;11:413. doi 10.1038/s41597-024-03247-7

Tong K., Chen X., Yan S., Dai L., Liao Y., Li Z., Wang T. PlantMine: a machine-learning framework to detect core SNPs in rice genomics. *Genes*. 2024;15(5):603. doi 10.3390/genes15050603

Vilov S., Heinig M. Neural network approach to somatic SNP calling in WGS samples without a matched control. *bioRxiv*. 2022. doi 10.1101/2022.04.14.488223

Wachter S., Mittelstadt B., Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard J Law Technol*. 2018;31(2):841-887

Wang H., Wang Z., Du M., Yang F., Zhang Z., Ding S., Mardziel P., Hu X. Score-CAM: score-weighted visual explanations for con-

volutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA. IEEE, 2020;111-119. doi 10.1109/CVPRW50498.2020.00020

Weber L., Lapuschkin S., Binder A., Samek W. Beyond explaining: opportunities and challenges of XAI-based model improvement. *Inf Fusion*. 2023;92:154-176. doi 10.1016/j.inffus.2022.11.013

Wu C., Zhang Y., Ying Z., Li L., Wang J., Yu H., Zhang M., Feng X., Wei X., Xu X. A transformer-based genomic prediction method fused with knowledge-guided module. *Brief Bioinform*. 2023;25(1):bbad438. doi 10.1093/bib/bbad438

Wu H., Gao B., Zhang R., Huang Z., Yin Z., Hu X., Yang C.-X., Du Z.-Q. Residual network improves the prediction accuracy of genomic selection. *Anim Genet*. 2024;55(4):599-611. doi 10.1111/age.13445

Xie Z., Xu X., Li L., Wu C., Ma Y., He J., Wei S., Wang J., Feng X. Residual networks without pooling layers improve the accuracy of genomic predictions. *Theor Appl Genet*. 2024;137(6):138. doi 10.1007/s00122-024-04649-2

Yelmen B., Decelle A., Ongaro L., Marnetto D., Tallec C., Montinaro F., Furtlehner C., Pagani L., Jay F. Creating artificial human genomes using generative neural networks. *PLoS Genet*. 2021;17(2):e1009303. doi 10.1371/journal.pgen.1009303

Zhang S., Li P., Wang S., Zhu J., Huang Z., Cai F., Freidel S., Ling F., Schwarz E., Chen J. *BioM2*: biologically informed multi-stage machine learning for phenotype prediction using omics data. *Brief Bioinform*. 2024;25(5):bbae384. doi 10.1093/bib/bbae384

Zhang T.-H., Flores M., Huang Y. ES-ARCNN: predicting enhancer strength by using data augmentation and residual convolutional neural network. *Anal Biochem*. 2021;618:114120. doi 10.1016/j.ab.2021.114120

Zhang X., Gao J. Measuring feature importance of convolutional neural networks. *IEEE Access*. 2020;8:196062-196074. doi 10.1109/ACCESS.2020.3034625

Conflict of interest. The authors declare no conflict of interest.

Received November 26, 2024. Revised January 23, 2025. Accepted January 23, 2025.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(3):467-473

doi 10.18699/vjgb-25-50

Deep learning approach to the estimation of the ratio of reproductive modes in a partially clonal population

T.A. Nikolaeva [1, 2], A.A. Poroshina [1], D.Yu. Sherbakov [1, 2]

[1] Limnological Institute of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia

t.maryanovskaya@alumni.nsu.ru

Abstract. Genetic diversity among biological entities, including populations, species, and communities, serves as a fundamental source of information for understanding their structure and functioning. However, many ecological and evolutionary problems arise from limited and complex datasets, complicating traditional analytical approaches. In this context, our study applies a deep learning-based approach to address a crucial question in evolutionary biology: the balance between sexual and asexual reproduction. Sexual reproduction often disrupts advantageous gene combinations favored by selection, whereas asexual reproduction allows faster proliferation without the need for males, effectively maintaining beneficial genotypes. This research focuses on exploring the coexistence patterns of sexual and asexual reproduction within a single species. We developed a convolutional neural network model specifically designed to analyze the dynamics of populations exhibiting mixed reproductive strategies within changing environments. The model developed here allows one to estimate the ratio of population members who originate from sexual reproduction to the clonal organisms produced by parthenogenetic females. This model assumes the reproductive ratio remains constant over time in populations with dual reproductive strategies and stable population sizes. The approach proposed is suitable for neutral multiallelic marker traits such as microsatellite repeats. Our results demonstrate that the model estimates the ratio of reproductive modes with an accuracy as high as 0.99, effectively handling the complexities posed by small sample sizes. When the training dataset's dimensionality aligns with the actual data, the model converges to the minimum error much faster, highlighting the significance of dataset design in predictive performance. This work contributes to the understanding of reproductive strategy dynamics in evolutionary biology, showcasing the potential of deep learning to enhance genetic data analysis. Our findings pave the way for future research examining the nuances of genetic diversity and reproductive modes in fluctuating ecological contexts, emphasizing the importance of advanced computational methods in evolutionary studies.

Key words: deep learning; convolutional neural network (CNN); Hardy–Weinberg equilibrium; partially clonal population; microsatellites

For citation: Nikolaeva T.A., Poroshina A.A., Sherbakov D.Yu. Deep learning approach to the estimation of the ratio of reproductive modes in a partially clonal population. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025; 29(3):467-473. doi 10.18699/vjgb-25-50

Funding. The study was carried out within the framework of the state budget theme No. 0279-2021-0010 "Genetics of Baikal organism communities: the gene pool structure, conservation strategies".

Acknowledgements. The authors express their gratitude to all sources of funding for the research.

Применение метода глубокого обучения для оценки соотношения репродуктивных режимов в частично клональной популяции

Т.А. Николаева [1, 2], А.А. Порошина [1], Д.Ю. Щербаков [1, 2]

[1] Лимнологический институт Сибирского отделения Российской академии наук, Иркутск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

t.maryanovskaya@alumni.nsu.ru

Аннотация. Генетическое разнообразие биологических объектов, таких как популяции, виды и сообщества, является важнейшим источником информации для понимания их структуры и функционирования. Однако многие экологические и эволюционные проблемы возникают из-за того, что наборы данных содержат относительно небольшое количество выборок, что затрудняет использование традиционных методов анализа. В связи с этим наше исследование предлагает новый подход, основанный на глубоком обучении, для решения одной из самых актуальных задач эволюционной биологии – поиска баланса между половым и бесполым размножением. Половое размножение часто приводит к нарушению выгодных комбинаций генов, которые были отобраны в процессе эволюции. С другой стороны, бесполое размножение позволяет организмам быстрее размножаться без участия самцов, эффективно поддерживая полезные генотипы. Исследование посвящено изучению закономерностей сосуществования полового и бесполого размножения в рамках одного вида. Мы разработали специальную сверточную модель нейронной сети, предназначенную для анализа динамики популяций, которые демонстрируют смешанные репродуктивные

стратегии в изменяющихся условиях. Эта модель позволяет оценить долю потомков репродуктивного размножения, если эта доля остается постоянной в течение достаточного периода времени, в популяциях, состоящих из постоянного числа организмов, с использованием мультиаллельных признаков, таких как микросателлитные повторы. Результаты показали, что модель с точностью 0.99 оценивает соотношение репродуктивных режимов, эффективно справляясь с трудностями, связанными с небольшими выборками. Более того, когда размерность обучающего набора данных соответствует фактическим данным, модель быстрее достигает минимальной ошибки, что подчеркивает важность подбора структуры набора данных для точности предсказаний. Эта работа вносит значительный вклад в понимание динамики репродуктивной стратегии в эволюционной биологии, демонстрируя потенциал глубокого обучения для улучшения анализа генетических данных. Наши результаты открывают двери для будущих исследований, посвященных тонкостям генетического разнообразия и способам размножения в изменчивых экологических условиях, подчеркивая важность современных вычислительных методов в эволюционных исследованиях.
Ключевые слова: глубокое обучение; сверточная нейронная сеть (CNN); равновесие Харди–Вайнберга; частично клональная популяция; микросателлиты

Introduction and motivation

Genetic diversity of biological entities such as populations, species, species communities is the main source of information allowing one to make numerous conclusions about their setup and functioning (Korfmann et al., 2023). Hence, the variety of sampling methods and ways of subsequent experimental data processing have been developed. In contrast to big data applications, where sample sizes typically exceed minimal requirements for robust conclusions, certain problems rely on limited and hard-to-acquire datasets, which complicates processing.

Deep learning has been applied successfully in population genetics in order to study various microevolutionary processes. A recurrent neural network model has been developed to predict recombination maps (Adrion et al., 2020), identify possible cases of positive natural selection (Anders, Korn, 1999; Eğrioğlu et al., 2008) and to estimate the time since the nearest common ancestor (Montinaro et al., 2021). A good predictive effectiveness on simulated data has been shown (Korfmann et al., 2023).

Neural networks were used to elucidate the demographic history of an individual population using genomic data without any preliminary knowledge of the recombination rate (Sanchez et al., 2021). In this study, the authors showed that network architecture is crucial for its performance. A poor design could lead to overfitting and loss of information.

When SNP frequencies were analyzed using MLP (multilayer perception), it led to high prediction errors, since the genomic information was encoded as a simple set of values where the order did not matter, and thus the information provided by the data structure was not used. The MLP configuration has several disadvantages for SNP analysis: (a) the number of estimated network parameters is large, which can lead to an increase in model training time; (b) MLP can extract data geometry only by training, without a guarantee that it will study the spatial structure of the genome. But MLP still works much better than random assumptions or constant prediction (by 32 %) (Sanchez et al., 2021).

In this paper, we apply a deep learning-based approach to one of the most intriguing questions of evolutionary biology: the balance between sexual and asexual reproduction (Schön et al., 2009; Baer, 2020; Otto, 2021; Cohen, Marron, 2023). Sexual reproduction can destroy favorable combinations of genes supported by selection, while the asexual one allows to reproduce twice as fast, since there is no need to produce males for continuous reproduction, and preserve favorable genotypes (Barton, Charlesworth, 1998; Gutiérrez-Valencia et al., 2021).

There are various patterns of coexistence of sexual and asexual reproductive modes in a single species. The sexual and asexual organisms belonging to the same species coexist in the same population, either alternating throughout their life cycle or in spatially or temporarily isolated subpopulations (Tagg et al., 2005; Rossi et al., 2007). Exclusively asexual vertebrates are usually closely related to sexually reproducing species (Janko et al., 2007; Schurko et al., 2009).

Asexual lines (clones) can develop by various mechanisms (spontaneous, contagious or infectious origin, hybridization) from ancestral sexual species (Avise et al., 1992), but the mechanisms of transition may be extremely diverse (Thielsch et al., 2012; Poroshina, Sherbakov, 2023). In order to analyze the exact population processes in organisms able to follow both ways of reproduction, one must be able to estimate the population-wide ratio of reproductive modes. Computer modeling previously allowed us to show that it is possible to do using distortions from equilibrium frequencies of microsatellite alleles (Messer, 2016). Here, we describe the development and testing of a deep learning model designed to study the dynamics of populations with a mixed type of reproduction in a changing environment.

Methods

Experimental data. The experimental data were taken from a published article and represent sets of allele lengths of microsatellites from 44 natural populations of *Daphnia cucullata*, *D. galeata* and *D. longispina* (1715 individuals) expressed in nucleotide pairs (Thielsch et al., 2012). The lengths of microsatellites are converted into matrices reflecting the frequency of occurrence of alleles and analyzed in this form by a neural network.

Simulated data. The training data were generated by a modified version of the Wright–Fisher model (WF), considering a mixed breeding strategy in a population (Messer, 2016). The model describes a population with discrete, nonoverlapping generations. In each generation, the entire population is replaced by the offspring of the previous generation. The parents are selected by random sampling with substitution. In a haploid population of constant size N, the probability that an allele present in i individuals will be present in j individuals in the next generation follows the binomial probability:

$$P_{ij} = \binom{N}{j} (i/N)^j (1 - i/N)^{N-j}, \quad 0 \le i, j \le N. \qquad (1)$$

The transition probabilities P_{ij} determine the Markov process with discrete time in the space of allele frequencies:

$$x(t) = i(t)/N. \quad (2)$$

The expected frequencies of alleles remain constant across generations, whereas the variance for each generation is:

$$Var[x] = x(1 - x)/N. \quad (3)$$

The probability that an allele will eventually become fixed is simply its initial frequency. In particular, the probability of fixing a new mutation present in a single copy is $1/N$ (Ratner, 1972).

Models of genotype distributions resulting in different reproductive modes. If all allele and gene combinations are believed to be of the same adaptive value and the conditions for the WF model are fulfilled, in a sufficiently big population reproducing exclusively sexually the Hardy–Weinberg equilibrium has to be true. In its traditional form, it describes a single locus having two alleles. For this study, we need an expanded model describing equilibrium for multiallelic loci which would be suitable for multiallelic microsatellites markers. Thus, for a gene having m alleles (for microsatellite markers $m > 2$), an array of allele frequencies $P = [p_1, \ldots, p_M]$ and $\sum_{i=1}^{M} p_i = 1$, where M is the number of alleles. The equilibriumprobabilities of diploid genotypes will be:

$$S = P \otimes P. \quad (4)$$

In matrix shape:

$$S = \begin{bmatrix} p_m \\ p_{m-1} \\ \vdots \\ p_1 \end{bmatrix} \otimes [p_1 \; p_2 \; \ldots \; p_m] = \begin{bmatrix} p_m p_1 & p_m p_2 & \cdots & p_m^2 \\ p_{m-1} p_1 & p_{m-1} p_2 & \cdots & p_{m-1} p_m \\ \vdots & \vdots & \ddots & \vdots \\ p_1^2 & p_1 p_2 & \cdots & p_1 p_m \end{bmatrix}. \quad (5)$$

And Hardy–Weinberg equilibrium will be:

$$\sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} = 1. \quad (6)$$

And, according to the WF model, it will hold for generations. In case of asexual reproduction, all ancestors of a given organism will inherit its genotype unless a mutation will transform the ancestral allele into a different one. It is important to note that we assume a fixed number of allowed alleles M, possibly different for each polymorphic locus; therefore, no mutation may increase M and frequencies of alleles will be:

$$A = [p_1, \ldots, p_M] * [p_1, \ldots, p_M]. \quad (7)$$

Assuming that the ratio of organisms resulting from asexual reproduction to the ones resulting from sexual reproduction is α, the genetic setup of a population with two coexisting reproduction strategies will be:

$$AS = \alpha([p_1, \ldots, p_M] * [p_1, \ldots, p_M]) + (1-\alpha) * ([p_1, \ldots, p_M] \otimes [p_1, \ldots, p_M]). \quad (8)$$

Neural network architecture and training. Two sources of noise in real world data have been modeled. Sampling error was mimicked by substituting probabilities of genotypes with their frequencies sampled from a small set of organisms. These frequencies were then converted to probabilities and used for training the network. The resulting values deviate from the expected pattern because of the small sample size.

Possible reasons for additional noise may include misidentification of samples, pipetting mistakes etc. They were simulated by the addition of a random value sampled from a normal distribution with average set to 0 and standard deviation set to 0.05 or any other sufficiently small value.

Neural networks are trained using a matrix of dimension $m \times n$, where m is the number of different alleles of a gene, n is the number of genes, and the element of the matrix a_{ij} is the frequency of occurrence of a combination of the i-th and j-th alleles.

The training set was obtained by repeating simulation of genotype distributions at different α for n genes, for different numbers of alleles M_i for each gene. The allele frequencies were sampled from a uniform distribution and then the genotype frequencies were obtained using (5).

A convolutional neural network (CNN) has been developed. It contains two external and six internal layers, including two convolution layers followed by max-pooling, a flatten layer and two fully connected dense layers (Fig. 1).

The mean absolute error (MAE) was chosen as the loss function. MAE is a measure of errors between paired observations

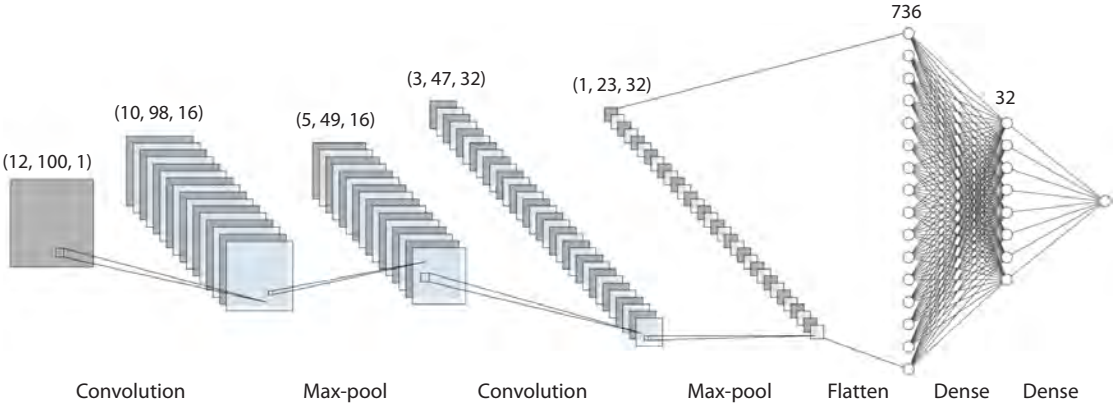


Fig. 1. The structure of the neural network.

A convolutional neural network contains two external and six internal layers: two convolution layers followed by a max-pooling one, a flattening layer and two dense layers.

expressing the same phenomenon. It is calculated as the sum of absolute errors divided by the sample size:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}. \quad (9)$$

The optimization strategy was based on the Adaptive Moment Estimation algorithm (ADAM). It combines both the idea of accumulation of movement and the idea of a weaker update of weights for typical signs. It is one of the most popular adaptive step-size methods (Kingma, Ba, 2014).

Gradient descent (CD) is a method that uses the fixed-point method to zero out the first derivative of the cost function, but it creates difficulties in complex applications.

Estimation of the model's precision. The accuracy of the model was estimated using the coefficient of determination (R^2). The coefficient of determination is the proportion of variance of the dependent variable explained by the dependence model in question, that is, the explanatory variables:

$$R^2 = \frac{\sum_i(\hat{y}_i - \overline{y}_i)^2}{\sum_i(y_i - \overline{y}_i)}. \quad (10)$$

Artificial noise in data. Small data size was modelled by first making a sample of a certain size with genotype quantities (integers) proportional to genotype probabilities calculated as described above and then normalized again to obtain frequencies. Thus, the smaller was the "sample size", the bigger was the distortion. This procedure allowed us to obtain the training set of genetic setups similarly distorted.

Other sources of mistakes include diverse aberrations like misidentification of samples, size calibration errors in the course of fragment analysis, etc. It was modelled by making a vector of random values sampled from normal distribution with the average set to 0 and the standard deviation set to a desired value, and adding this vector to the vector of values delimiting different classes of ratios of individuals resulting from sexual or asexual reproduction.

Results

A deep learning-based method for estimating the ratio of asexual and sexual reproduction in populations capable of switching between these reproductive modes has been developed. In its current form, the method is intended to use multiallelic traits, the most common of which are microsatellite repeats. The method achieved an accuracy value of 0.99. The method of training the neuron network appears to be critically important: our findings reveal that ignoring the variability in allele counts across genes and using uniform genotype matrices significantly reduces model precision. This underscores the importance of accounting for allelic diversity during training. In this regard, for each data set, the model was trained on a simulated data set of a similar dimension to a frequency matrix of the original data.

When the size of the training dataset matches the dimensionality of the actual data, the mean squared error converges to zero more rapidly compared to situations where the training dataset has a larger size (Fig. 2).

With the model architecture chosen, the optimal number of learning epochs turned out to be 15, with the value of the number of epochs, the learning rate equal to 0.01 and the size of the training sample equal to 16, the minimal error value is achieved. With the learning rate of the model equal to 0.1, the error quickly takes a value less than 0.05 and does not rise above this value with the sample sizes of 16 and 32. With a learning rate of 0.1, the result is unstable, and the error value varies from 0.29 to 0.3 and does not drop below even with 50 training epochs (Fig. 3).

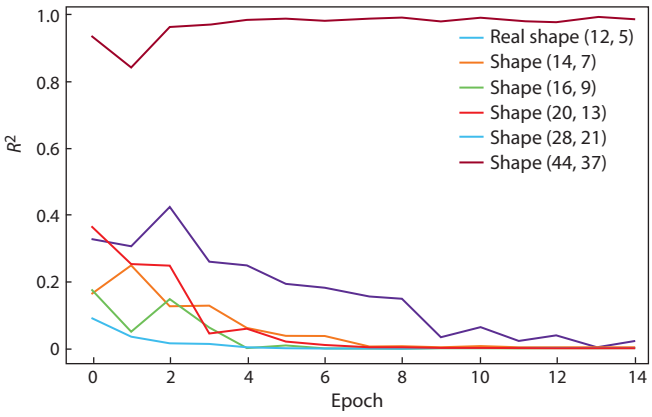


Fig. 2. The dependence of the error value on the learning epoch for different dimensions of a frequency matrix of the training sample.

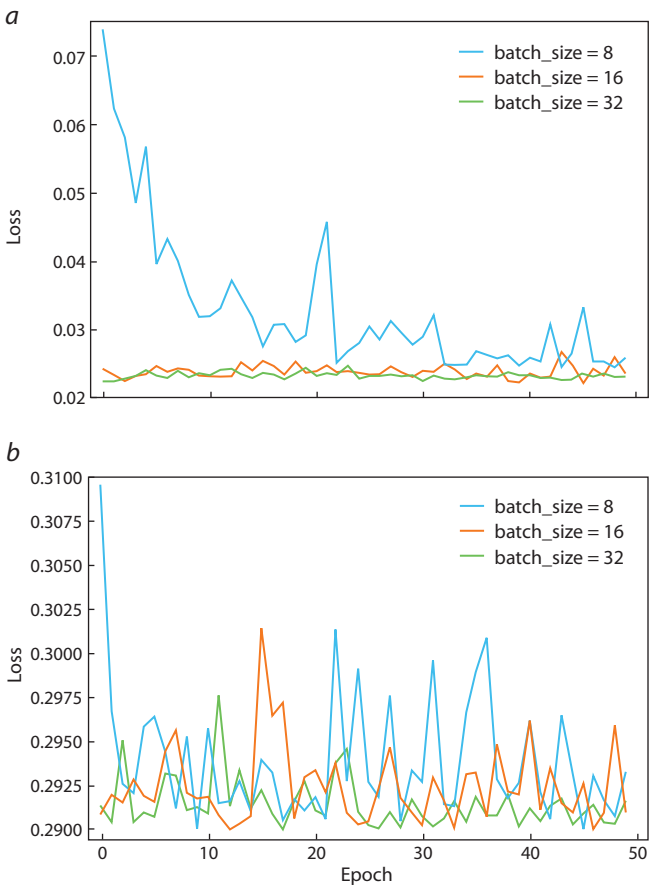


Fig. 3. The dependence of the error value on the learning epoch for different sizes of the training sample and different learning rates.

a – the graph shows the error value depending on the learning epoch, with a model learning rate equal to 0.01; *b* – the graph shows the error value depending on the epoch with a learning rate equal to 0.1.

When noise occurs in the frequency values of the ratio of sexual and asexual reproduction, which may indicate errors occurring during sequencing, the average error values when noise occurs are higher than without noise, but with a standard deviation value of 0.05 differ by no more than 0.01 (Fig. 4). This computational experiment tests the method's resistance to noise caused by sequencing errors.

As the number of individuals in the sample increases, the confidence interval in the early epochs of model learning decreases. When comparing noisy data by sample size and non-noisy data, it can be concluded that the error does not differ much; at the initial stages of model training, the confidence interval is larger, but at the end of model training, both the average and the confidence interval differ slightly (Fig. 5). This computational experiment tests the method's resistance to noise arising from limited sampling for analysis. The model was tested on experimental data, and values. The models obtained as a result of calculations coincided with the experimental data.

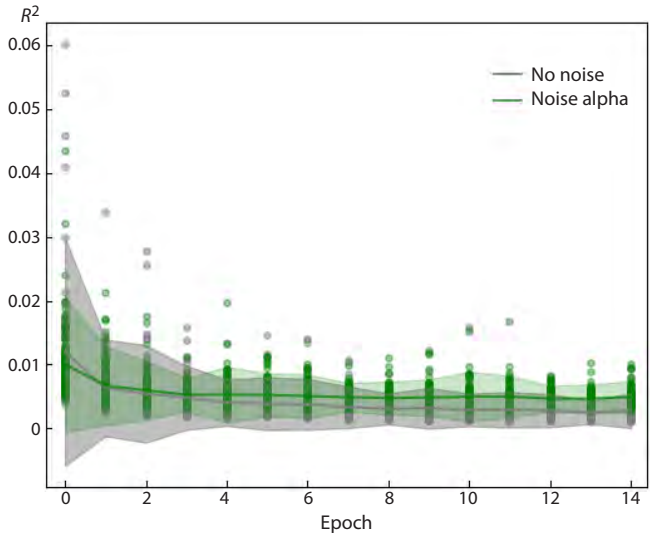


Fig. 4. The effect of noise in the frequency of occurrence of a combination of alleles on the prediction error of the model.

The green color shows the error distribution when training the model on simulated data with artificial noise in the frequencies of occurrence of a combination of alleles having a Gaussian distribution with a standard deviation of 0.05. The grey color shows the error distribution when training the model on simulated data without noise.

Discussion

The model proposed here does not take into account a set of complications quite common in the real-world data. Different loci are often inherited dependently due to topological associations in chromosomes which per se may be of positive selective value and may change the expression level of some genes. These associations may be supported by assorted mechanisms bringing even distant loci physically together. Also, many microsatellites are organized in a more complex way than just a simple repeat of short sequences; in this case, the inheritance of microsatellite alleles may be distorted by non-allelic mutations in the adjacent areas of the genome. These accomplishments become a serious challenge when setting up models, which in turn may cause an unnaturally high level of mistakes in models or the necessity to develop models with many more parameters. This increases the numbers of model parameters' computational time and complexity (Putman, Carbone, 2014).

The advantage of our convolutional neural network compared to traditional approaches is the ability to efficiently extract and process information from multidimensional data

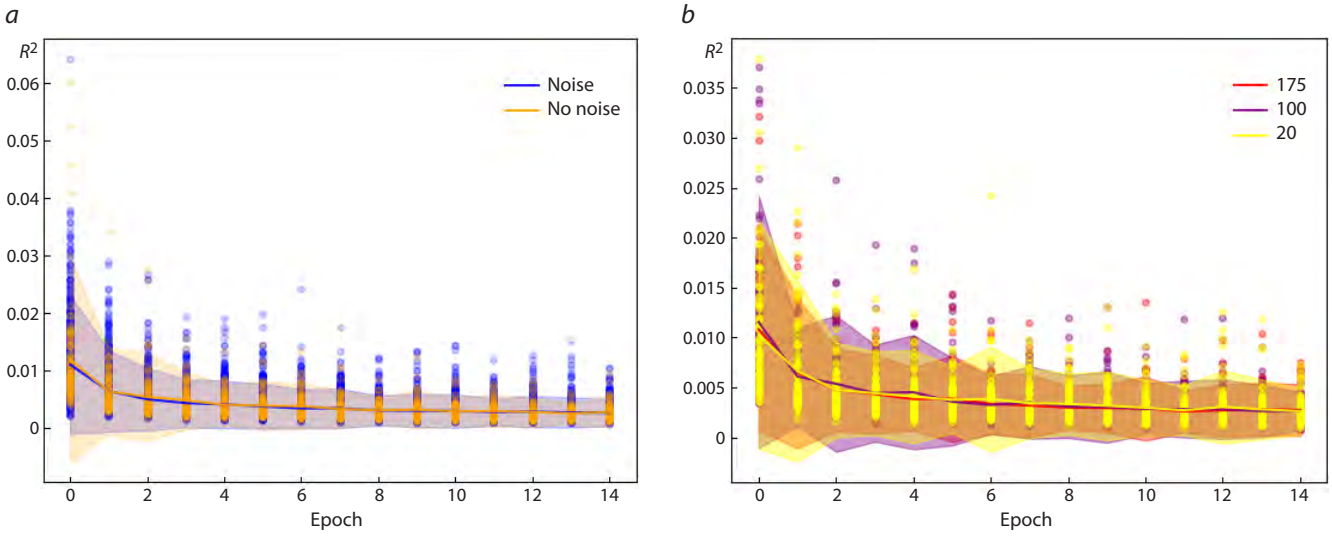


Fig. 5. The dependence of the error value on the epoch of model training with a limited sample.

a – the blue color shows the error distribution when training the model on simulated data containing noise in the form of limiting the sample to the training data and rounding the value of the sum of frequencies in a sample size from 10 to 200, followed by averaging the frequency of occurrence of a combination of alleles. The orange color shows the error distribution when training the model on simulated data without noise. The lines are depicting approximation curves for data with and without noise. The confidence intervals are shown in translucent color; *b* – the red color shows the error distribution for a sample size of 20 out of 200 possible; the purple color shows the error distribution for a sample size of 100 out of 200 possible; the yellow color shows the error distribution for a sample size of 175 out of 200 possible.

structures, which is critically important when analyzing genetic data. In particular, this means an increased ability to recognize complex relationships and mechanisms, which can provide more accurate predictions and a better understanding of genetic interactions. The model proposed is characterized by a high degree of accuracy, it is trained on data, the size of which exactly corresponds to the size of experimentally obtained genetic matrices, thereby minimizing the risk of overfitting, which often occurs when using larger, artifact datasets. This approach made it possible to achieve a significant level of accuracy already at the initial stages of training, which indicates the high efficiency of model training and its ability to quickly adapt to new data.

When training with data that are selected depending on the structure of the actual analyzed data, the model quickly reaches an accuracy of 0.95, after which the overtraining of the model does not happen. Overtraining of a neural network may not occur in some cases, for example, when a linear perceptron is used. In this case, all the minima of the error function are approximately equivalent to the desired point of the global minimum, thus overtraining cannot be achieved.

Noise in the data has a significant impact on neural network prediction results, especially in the cases of analyzing biological data such as DNA sequencing (Kircher, Kelso, 2010). Errors in data acquisition may be due to diverse reasons such as insufficient sample quality: poor quality of DNA or RNA, for example, due to degradation or contamination, can lead to errors in sequencing (Levin et al., 2020), and so can faults in sequencing technology (different sequencing methods have their own limitations and sources of errors). For example, some technologies may have difficulty with repeating sequences or with long DNA fragments (Adiconis et al., 2013). The very PCR process may become a source of noise: when samples are amplified using polymerase chain reaction (PCR), errors can occur, which are then transmitted to the sequencing results (Hsiao, 2019).

Frequency values of the ratio of sexual and asexual reproduction are subject to random deviations. This may make it difficult for the neural network to correctly identify patterns and dependencies. Incorrect or distorted data can cause the model to make incorrect assumptions about the distribution of data, which reduces its generalizing ability.

However, as the results of the present experiment show, with a noise standard deviation of no more than 0.05, which roughly corresponds to the real situation, the difference in predictions is only 0.01. This indicates that the proposed method is sufficiently resistant to frequency noise that occurs during the acquisition of real data.

Modern approaches, such as the use of model ensembles or techniques for estimating the uncertainty of predictions, can also help to effectively deal with noise (Zhou, 2025). This is especially true in biological research, where data may be distorted due to a large variety of reasons. The reasons are not specified in this work, since the noise level in the frequencies, when receiving real data, which are further analyzed, often does not exceed 0.05.

The noise caused by a limited sample size may result in an increase in the prediction error, and its negative impact can be mitigated by using a sufficient sample size. The observed decrease in the confidence interval with an increase in the number of objects in the sample indicates an increase in the accuracy of the model's predictions as more data are accumulated. A comparison of error distributions in noisy and non-noisy data at different stages of training shows that although the confidence interval for noisy data is wider at the initial stages, the error differences become less significant at later stages of training. This may indicate that with an increase in the number of training iterations, the model is able to adapt to noise and adjust its predictions. It is also worth noting that the simulation results agree with experimental data, which confirms the adequacy of the proposed method and its resistance to noise arising from a limited sample size. This opens up the possibility for applying this approach in various fields were working with noisy data is an everyday task, such as genetic research, medical diagnostics, and other scientific fields that otherwise would require the analysis of a larger amounts of complex data.

Conclusion

Application of the described approach has its limits since violations of the equilibrium frequencies of genotypes can arise for a number of reasons not related to reproductive strategy, from genetic drift to sudden demographic changes. Therefore, in each specific case, it is necessary to involve external knowledge regarding the biology of the organisms under study. Further studies of populations with a mixed reproductive strategy and, accordingly, methods for detecting the characteristics of their genetic diversity should take into account, firstly, the inconstancy of the ratios of strategies in a number of generations, and secondly, possible sharp demographic fluctuations. The combinations of these two factors result in unusual patterns of genetic diversity.

References

Adiconis X., Borges-Rivera D., Satija R., DeLuca D.S., Busby M.A., Berlin A.M., Sivachenko A., Thompson D.A., Wysoker A., Fennell T., Gnirke A., Pochet N., Regev A., Levin J.Z. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods*. 2013;10(7):623-629. doi 10.1038/nmeth.2483

Adrion J.R., Galloway J.G., Kern A.D. Predicting the landscape of recombination using deep learning. *Mol Biol Evol*. 2020;37(6):1790-1808. doi 10.1093/molbev/msaa038

Anders U., Korn O. Model selection in neural networks. *Neural Netw*. 1999;12(2):309-323. doi 10.1016/s0893-6080(98)00117-8

Avise J.C., Quattro J.M., Vrijenhoek R.C. Molecular clones within organismal clones. In: Hecht M.K., Wallace B., Macintyre R.J. (Eds) Evolutionary Biology. Vol. 26. Boston, MA: Springer, 1992;225-246. doi 10.1007/978-1-4615-3336-8_6

Baer B. Sexual selection. In: Starr C. (Ed.) Encyclopedia of Social Insects. 2020. Springer, Cham. doi 10.1007/978-3-319-90306-4_104-1

Barton N.H., Charlesworth B. Why sex and recombination? *Science*. 1998;281(5385):1986-1990. doi 10.1126/science.281.5385.1986

Cohen I.R., Marron A. Evolution is driven by natural autoencoding: reframing species, interaction codes, cooperation and sexual reproduction. *Proc Biol Sci*. 2023;290(1994):20222409. doi 10.1098/rspb.2022.2409

Eğrioğlu E., Aladağ Ç.H., Günay S. A new model selection strategy in artificial neural networks. *Appl Math Comput*. 2008;195(2):591-597. doi 10.1016/j.amc.2007.05.005

Gutiérrez-Valencia J., Hughes P.W., Berdan E.L., Slotte T. The genomic architecture and evolutionary fates of supergenes. *Genome Biol Evol*. 2021;13(5):evab057. doi 10.1093/gbe/evab057

Hsiao S.J. Sources of error in molecular diagnostic analyses. In: Dasgupta A., Sepulveda J.L. (Eds) Accurate Results in the Clinical Laboratory. Elsevier, 2019;337-347. doi 10.1016/B978-0-12-813776-5.00021-2

Janko K., Bohlen J., Lamatsch D., Flajšhans M., Epplen J.T., Ráb P., Kotlík P., Šlechtová V. The gynogenetic reproduction of diploid and triploid hybrid spined loaches (*Cobitis*: Teleostei), and their ability to establish successful clonal lineages – on the evolution of polyploidy in asexual vertebrates. *Genetica*. 2007;131(2):185-194. doi 10.1007/s10709-006-9130-5

Kingma D.P., Ba J. Adam: a method for stochastic optimization. *arXiv*. 2014. doi 10.48550/arXiv.1412.6980

Kircher M., Kelso J. High-throughput DNA sequencing – concepts and limitations. *BioEssays*. 2010;32(6):524-536. doi 10.1002/bies.200900181

Korfmann K., Gaggiotti O.E., Fumagalli M. Deep learning in population genetics. *Genome Biol Evol*. 2023;15(2):evad008. doi 10.1093/gbe/evad008

Levin Y., Talsania K., Tran B., Shetty J., Zhao Y., Mehta M. Optimization for sequencing and analysis of degraded FFPE-RNA samples. *J Vis Exp*. 2020;160:e61060. doi 10.3791/61060

Messer P.W. Neutral models of genetic drift and mutation. In: Kliman R.M. (Ed.) Encyclopedia of Evolutionary Biology. Academic Press, Elsevier, 2016;119-123. doi 10.1016/B978-0-12-800049-6.00031-7

Montinaro F., Pankratov V., Yelmen B., Pagani L., Mondal M. Revisiting the out of Africa event with a deep-learning approach. *Am J Hum Genet*. 2021;108(11):2037-2051. doi 10.1016/j.ajhg.2021.09.006

Otto S.P. Selective interference and the evolution of sex. *J Hered*. 2021;112(1):9-18. doi 10.1093/jhered/esaa026

Poroshina A., Sherbakov D. A procedure for modeling genetic diversity distortions in populations of organisms with mixed reproductive strategies. *Mathematics*. 2023;11(13):2985. doi 10.3390/math11132985

Putman A.I., Carbone I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol Evol*. 2014;4(22):4399-4428. doi 10.1002/ece3.1305

Ratner V.A. Principles of Organization and Mechanisms of Molecular Genetic Processes. Novosibirsk: Nauka Publ., 1972 (in Russian)

Rossi V., Gandolfi A., Baraldi F., Bellavere C., Menozzi P. Phylogenetic relationships of coexisting *Heterocypris* (Crustacea, Ostracoda) lineages with different reproductive modes from Lampedusa Island (Italy). *Mol Phylogenet Evol*. 2007;44(3):1273-1283. doi 10.1016/j.ympev.2007.04.013

Sanchez T., Cury J., Charpiat G., Jay F. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. *Mol Ecol Resour*. 2021;21(8):2645-2660. doi 10.1111/1755-0998.13224

Schön I., Martens K., van Dijk P. (Eds) Lost Sex. The Evolutionary Biology of Parthenogenesis. Springer, 2009. doi 10.1007/978-90-481-2770-2

Schurko A.M., Neiman M., Logsdon J.M. Jr. Signs of sex: what we know and how we know it. *Trends Ecol Evol*. 2009;24(4):208-217. doi 10.1016/j.tree.2008.11.010

Tagg N., Doncaster C.P., Innes D.J. Resource competition between genetically varied and genetically uniform populations of *Daphnia pulex* (Leydig): does sexual reproduction confer a short-term ecological advantage? *Biol J Linn Soc*. 2005;85(1):111-123. doi 10.1111/j.1095-8312.2005.00475.x

Thielsch A., Völker E., Kraus R.H.S., Schwenk K. Discrimination of hybrid classes using cross-species amplification of microsatellite loci: methodological challenges and solutions in *Daphnia*. *Mol Ecol Resour*. 2012;12(4):697-705. doi 10.1111/j.1755-0998.2012. 03142.x

Zhou Z.H. Ensemble Methods: Foundations and Algorithms. CRC Press, 2025

Conflict of interest. The authors declare no conflict of interest.
Received November 26, 2024. Revised February 6, 2025. Accepted March 4, 2025.

Прием статей через электронную редакцию на сайте http://vavilov.elpub.ru/index.php/jour
Предварительно нужно зарегистрироваться как автору, затем в правом верхнем углу страницы выбрать «Отправить рукопись». После завершения загрузки материалов обязательно выбрать опцию «Отправить письмо», в этом случае редакция автоматически будет уведомлена о получении новой рукописи.

«Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» до 2011 г. выходил под названием «Информационный вестник ВОГиС»/ "The Herald of Vavilov Society for Geneticists and Breeding Scientists".

Сетевое издание «Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» – реестровая запись СМИ Эл № ФС77-85772, зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 14 августа 2023 г.

Издание включено ВАК Минобрнауки России в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, Russian Science Citation Index, Российский индекс научного цитирования, ВИНИТИ, Web of Science CC, Scopus, PubMed Central, DOAJ, ROAD, Ulrich's Periodicals Directory, Google Scholar.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте https://vavilovj-icg.ru/
и платформе Научной электронной библиотеки, elibrary.ru/title_about.asp?id=32440

англоязычная версия – на сайте vavilov.elpub.ru/index.php/jour
и платформе PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/journals/3805/

✉ email: vavilov_journal@bionet.nsc.ru

Издатель: Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук»,
проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Адрес редакции: проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Секретарь по организационным вопросам С.В. Зубова. Тел.: (383)3634977.
Издание подготовлено информационно-издательским отделом ИЦиГ СО РАН. Тел.: (383)3634963*5218.
Начальник отдела: Т.Ф. Чалкова. Редакторы: В.Д. Ахметова, И.Ю. Ануфриева. Дизайн: А.В. Харкевич.
Компьютерная графика и верстка: Т.Б. Коняхина, О.Н. Савватеева.
Дата публикации 29.05.2025. Формат 60 × 84 $^{1}/_{8}$. Уч.-изд. л. 20.3.