2024 • 28 • 6

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-24-64

2024 • 28 • 6

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Published 8 times annually
DOI 10.18699/vjgb-24-64

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

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ • VAVILOV JOURNAL OF GENETICS AND BREEDING

СОДЕРЖАНИЕ • 2024 • 28 • 6

Молекулярная и клеточная биология

Генетика растений

Генетика человека

Медицинская генетика

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

CONTENTS • 2024 • 28 • 6

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):583-591

DOI 10.18699/vjgb-24-65

Effects of the auxin-dependent degradation of the cohesin and condensin complexes on the repair of distant DNA double-strand breaks in mouse embryonic stem cells

A.V. Smirnov, A.S. Ryzhkova, A.M. Yunusova

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
hldn89@gmail.com

Abstract. The SMC protein family, including cohesin and condensin I/II, plays a pivotal role in maintaining the topological structure of chromosomes and influences many cellular processes, notably the repair of double-stranded DNA breaks (DSBs). The cohesin complex impacts DSB repair by spreading γH2AX signal and containing DNA ends in close proximity by loop extrusion. Cohesin supports DNA stability by sister chromatid cohesion during the S/G2 phase, which limits DNA end mobility. Cohesin knockdown was recently shown to stimulate frequencies of genomic deletions produced by distant paired DSBs, but does not affect DNA repair of a single or close DSBs. We examined how auxin-inducible protein degradation of Rad21 (cohesin) or Smc2 (condensins I+II) changes the frequencies of rearrangements between paired distant DSBs in mouse embryonic stem cells (mESCs). We used Cas9 RNP nucleofection to generate deletions and inversions with high efficiency without additional selection. We determined optimal Neon settings and deletion appearance timings. Two strategies for auxin addition were tested (4 independent experiments in total). We examined deletion/inversion frequencies for two regions spanning 3.5 and 3.9 kbp in size. Contrary to expectations, in our setting, Rad21 depletion did not increase deletion/inversion frequencies, not even for the region with an active Ctcf boundary. We actually observed a 12 % decrease in deletions (but not inversions). At the same time, double condensin depletion (Smc2 degron line) demonstrated high biological variability between experiments, complicating the analysis, and requires additional examination in the future. TIDE analysis revealed that editing frequency was consistent (30–50 %) for most experiments with a minor decrease after auxin addition. In the end, we discuss the Neon/ddPCR method for deletion generation and detection in mESCs.

Key words: CRISPR/Cas9; mouse embryonic stem cells; auxin; cohesin; condensin; DNA repair.

For citation: Smirnov A.V., Ryzhkova A.S., Yunusova A.M. Effects of the auxin-dependent degradation of the cohesin and condensin complexes on the repair of distant DNA double-strand breaks in mouse embryonic stem cells. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):583-591. DOI 10.18699/vjgb-24-65

Funding. This work was supported by Russian Science Foundation grant No. 22-74-00084. Cell culture was performed at the Collective Center of ICG SB RAS "Collection of Pluripotent Human and Mammalian Cell Cultures for Biological and Biomedical Research", project number FWNR-2022-0019 (https://ckp.icgen.ru/cells/; http://www.biores.cytogen.ru/brc_cells/collections/ICG_SB_RAS_CELL). Droplet digital PCR was performed using the QX100 equipment (project number FWNR-2022-0015). Sanger DNA sequencing was performed at the Genomics Core Facility (ICBFM SB RAS, Novosibirsk).

Влияние ауксин-зависимой деградации когезина и конденсинов на репарацию двуцепочечных разрывов ДНК в эмбриональных стволовых клетках мыши

А.В. Смирнов, А.С. Рыжкова, А.М. Юнусова

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
hldn89@gmail.com

Аннотация. Семейство SMC-белков, включающее когезин и конденсины I/II, играет ключевую роль в формировании топологической структуры хромосом и косвенно влияет на широкий спектр клеточных процессов, в том числе и на репарацию двуцепочечных разрывов ДНК (DSB). Комплекс когезина регулирует репарацию DSB на нескольких уровнях, например, распространяя сигнал γH2AX и удерживая концы ДНК в непосредственной близости за счет экструзии петель возле разрыва. Когезин также скрепляет сестринские хроматиды во время фазы S/G2, что ограничивает потенциальную подвижность концов ДНК. По имеющимся данным, в фибробластах человека нокдаун когезина стимулирует образование геномных делеций между удаленными DSB (3.2 тыс. п. о.), но не влияет на репарацию одиночных или близких DSB (34 п. о.). Мы решили проверить это наблюдение на эмбриональных стволовых клетках мыши, несущих ауксин-индуцибельный дегрон Rad21 (субъединица когезина) или Smc2 (субъединица конденсинов I+II). Для этого мы использовали нуклеофекцию RNP Cas9 и пары гайдовых

РНК для генерации делеций и инверсий с высокой эффективностью без дополнительной селекции. Мы определили оптимальные условия для эффективной электропорации, включая настройки Neon, а также тайминги появления делеций. Были протестированы две стратегии добавления ауксина (суммарно четыре независимых эксперимента). Были исследованы частоты перестроек в двух сайтах размером около 3.5 и 3.9 тыс. п.о. Вопреки ожиданиям, деплеция Rad21 не увеличивала частоту делеций/инверсий, даже для региона с активной границей Ctcf. Фактически наблюдалось снижение частоты делеций (но не инверсий) на 12 %. Деплеция Smc2 не приводила к заметному увеличению частот делеций/инверсий, возможно, из-за высокой биологической изменчивости между экспериментами. Анализ TIDE показал, что частота редактирования была постоянной для большинства экспериментов (30–50 %), с незначительным снижением после добавления ауксина. В статье также обсуждается применимость метода Neon/ddPCR для создания и детекции делеций в эмбриональных стволовых клетках мыши.

Ключевые слова: CRISPR/Cas9; эмбриональные стволовые клетки мыши; ауксин; когезин; конденсин; репарация ДНК.

Introduction

Properly joining the two ends of a double-strand break (DSB) is crucial for preserving genome integrity. Unprocessed DNA ends can degrade, leading to loss of genetic information. Moreover, because DNA repair occurs in the vast space of the nucleus, incorrect ligation of multiple DSBs can result in chromosomal rearrangements, such as translocations, inversions, deletions, mitotic bridges and even chromothripsis. One-sided breaks that arise during replication are also highly dangerous and must be restrained and connected to the appropriate DNA molecule.

SMC complexes (cohesin, condensin-I, condensin-II) consist of several proteins organized into a ring-shaped structure (Kabirova et al., 2023). They utilize ATP-driven motor activity to shape and organize the genome into topological domains (TADs). Cohesin is an integral part of cellular homeostasis, regulating DNA conformation and topology, thus governing most vital processes from replication and cell division to gene expression and programmed DNA breaks in meiosis or V(D)J recombination. Although many reports have linked cohesin to DNA repair, its exact role therein remains unclear. Cohesin is attracted to DSB foci (Ström et al., 2004; Ünal et al., 2004), but is probably not essential for DNA end-joining *per se* (Gelot et al., 2016). Early cytogenetic and microscopic evidence suggests that cohesin limits DNA end mobility and prevents genomic rearrangements (Wu, Yu, 2012). A study using a genetic reporter showed that cohesin knockdown leads to an increased frequency of deletions when two DSBs are introduced at a distance of 3.2 kbp but does not affect the ligation of closely located breaks (34 bp) (Gelot et al., 2016).

Importantly, these observations were only relevant to the S phase, where cohesin is required for sister chromatid cohesion. Knockdown of cohesin in G1-synchronized cells did not have an effect on deletion frequencies, probably because cohesin molecules physically limit the mobility of the DSB ends to preserve genome integrity only during S phase (Supplementary Material 1)[1]. Generally speaking, cohesin removal does not affect deletion frequencies in G1, because it does not hold fragments together; but in the S phase, the excised fragments and the DSB ends are "stapled" to a sister chromatid (Supplementary Material 1). Cohesin acts on multiple levels to organize DSB repair, including retaining sister chromatids for homologous recombination (HR) and replication fork restart (Wu, Yu, 2012); limiting the mobility of the DSB for a better homology search within confined "repair domains" (Piazza et al., 2021); and amplifying the γH2AX signal by asymmetrically extruding flanking chromatin in the vicinity of Ataxia-telangiectasia mutated (ATM) kinase at the DSB (Arnould et al., 2021).

At the same time, cohesin promotes replication stress by interfering with replication during loop extrusion (Minchell et al., 2020), which complicates the picture even further. Another insight comes from the DIvA (DSB Inducible via AsiSI) U2OS cells. This cell line expresses AsiSI restrictase with attached estrogen receptor ligand-binding domain (Aymard et al., 2014). After induction by 4-hydroxytamoxifen (4OHT), AsiSI translocates into the nucleus and introduces around 100–200 DSBs in annotated genomic loci (Dobbs et al., 2022). Multiple breaks induced with AsiSI tend to cluster together and form special kinds of D-compartments, but cohesin is not required for this process or any other kind of chromatin compartmentalization (Schwarzer et al., 2017; Arnould et al., 2023). Trans-interactions of multiple AsiSI-induced DSBs were also not affected by the Rad21 knockdown, but cohesin was required to reinforce affected TADs locally (Arnould et al., 2023). Thus, the connections between cohesin, chromatin compartmentalization, sister chromatid cohesion, DSB restraining and end joining are highly complex.

The role of condensins, another SMC family of DNA organizers, in DSB repair is still unclear. Their primary function is genome compaction before mitosis and they are mostly not active during interphase, although the complex resides in the nucleus throughout the cell cycle. Recent photobleaching experiments indicated that during interphase condensin II is very efficiently blocked from chromatin by the primary binding partner – the microcephalin protein (Mcph1) (Houlard et al., 2021). Mcph1 plays an important but poorly understood role in DSB repair, such as facilitating HR repair through Rad51 filament stabilization (Wu et al., 2009; Chang et al., 2020). Defects in condensin assembly lead to chromosomal aberrations and sister chromatid interlinks in mitotic chromatin (Wu, Yu, 2012; Baergen et al., 2019). Evidence suggests that yeast condensin cooperates with topoisomerase-II to dissolve DNA knots (Dyson et al., 2021) and condensin II could be directly or indirectly involved in homology-directed repair (Wood et al., 2008).

Does cohesin directly impact the joining of close and distant DSBs? How do TAD features (size, borders, chromatin) influence deletion frequencies? Do condensin complexes play any

[1] Supplementary Materials 1–7 are available at:
https://vavilovj-icg.ru/download/pict-2024-28/appx20.pdf

role in DSB repair? In a series of pilot experiments presented here, we begin to explore some of these glaring questions.

Previously, we obtained and extensively characterized mouse embryonic stem cells (mESCs) with auxin-inducible degron (AID) knock-ins for Rad21 (cohesin) and Smc2 (both condensins I+II). These cells exhibit rapid depletion of the target protein within 1–2 hours of auxin addition (Yunusova et al., 2021). Cas9 activity generates blunt ends at the target sites. Using a pair of gRNA frequently leads to an excision of an intermediate DNA segment, which could lead to deletion or inversion after non-homologous or microhomology-mediated end joining (NHEJ/MMEJ) (Canver et al., 2014;Watry et al., 2020; Li et al., 2021). The cell lines were nucleofected with Cas9 and paired gRNAs, and studied using droplet digital PCR (ddPCR) to detect deletions and inversions in the mESCs population. This approach allowed us to assess the influence of the spatial organization of DNA and chromatin on the joining of two distant DNA ends. Overall, the method demonstrated high efficiency and sensitivity for detecting deletions and inversions. At the same time, the results were somewhat inconsistent, and the method we used might be more challenging than we had anticipated. We discuss its potential and limitations in the following chapters.

Materials and methods

gRNA design and cloning. We selected two genomic regions to induce paired DSBs: the *Ace2* gene locus (ChrX: 162.922.328–162.971.416) and a distinct TAD border that shows strong Ctcf signals in ChIP-seq data for mESCs (Chr5: 49.487.342–49.557.342) (GRCm39). High scoring gRNA sites were chosen using Benchling and Aidit algorithms. The sequences of the optimal gRNAs are listed in the Table. All oligonucleotides used in the study were purchased from DNA-Synthesis (Russia). 100 nt gRNAs were synthesized by the T7 *in vitro* transcription system from a PCR product amplified from a gRNA vector with the T7-primer (overhang 5′-GTTAATACGACTCACTATAG-20nt(gRNA)-3′) and the reverse primer (see the Table) (HiScribe® T7 High Yield RNA Synthesis Kit, E2040S, protocol for short products). After 4 hours at 37 °C, the reaction volume (20 µl) was diluted to 100 µl and treated with 2 µl (4U) of DNaseI (NEB #M0303) in the corresponding buffer. RNA was purified with Monarch® RNA Cleanup Kit (50 µg) (T2040L) and diluted in 30 µl water to achieve concentrations of 2 µg/µl or higher.

mESCs nucleofections. Both mESCs auxin degron cell lines were characterized in our laboratory earlier (Rad21-miniIAA7-eGFP, Smc2-miniIAA7-eGFP) (Yunusova et al., 2021). Cells were cultured on plates coated with a 1 % gelatin solution under 2i conditions (1 µM PD, 3 µM CHIR) in DMEM (Thermo Fisher, USA), supplemented with 7.5 % ES FBS (Gibco, USA), 7.5 % KSR (Gibco), 1 mM L-glutamine (Sigma, USA), NEAA (Gibco), 0.1 mM β-mercaptoethanol, LIF (1000 U/ml, Polygen), and penicillin/streptomycin (100 U/ml each). Upon reaching appropriate confluence (70–80 %), the cells were passaged every two days.

Single nucleofection sample consisted of 5 µl Buffer R with 300 000 cells which were mixed with 5 µl of RNP complex diluted in Buffer R in a 10 µl tip. Nucleofections were carried

Oligonucleotides used in the study

Oligonucleotide	Sequence 5′–3′	Application
gRNA *Ace2* F	cacctgataaagtcagctgt	gRNA sequence
gRNA *Ace2* R2	ataagggcaacgaattgaca	
gRNA Ctcf F	ccttgacaagggcaccatgg	
gRNA Ctcf R2	aagaggctcatcagggactc	
T7 *Ace2*-F F	gttaatacgactcactatagcacctgataaagtcagctgt	T7 *in vitro* transcription
T7 *Ace2*-R2 F	gttaatacgactcactatagataagggcaacgaattgaca	
T7 Ctcf-F F	gttaatacgactcactatagccttgacaagggcaccatgg	
T7 Ctcf-R2 F	gttaatacgactcactatagaaaagcaccgactcggtgcc	
gRNA31 Rev	aaaagcaccgactcggtgcc	
Ace2 F	gcagagtcattattacttccttg	ddPCR for deletions (149 bp) and inversions (154 bp) at the *Ace2* locus
Ace2 R	caacctgggttcagaccctc	
Ace2 Inv R	ggcacaaagagttcatattacttac	
Ace2 Probe	HEX-tacctgcttacaactcagctgagaac-BHQ2	
Ctcf F	ggaggcataataacaactgctc	ddPCR for deletions (205 bp) and inversions (227 bp) at the Ctcf locus
Ctcf R	cagaggttagaacctatgaatcgg	
Ctcf Inv R	ggcacaaagagttcatattacttac	
Ctcf Probe	HEX-agacagagctgatcaagacagcatggt-BHQ2	
Emid1 F	gccaggactgggtagcac	ddPCR for the reference region (79 bp)
Emdi1 R	aggaggctcctgaatttgtgacaag	
Emid1 Probe	FAM-cctgggtcatctgagctgagtcc-BHQ1	
Ace2 TIDE F	gtcatggatgcgctttggat	TIDE PCR (412 bp)
Ace2 TIDE R	aatggagagaatggggcagg	

out at Neon preset condition #10 (Pulse voltage 1000, Pulse width 100, Pulse No. 1). Other tested conditions included #2 (1400, 20, 1), #6 (1100, 30, 1), #7 (1200, 30, 1), #13 (1100, 20, 2), #17 (850, 30, 2). The RNP mix consisted of 0.2 μl of concentrated Cas9-NLS protein (30 pmoles) (Biolabmix, Russia) and 2000 ng of each gRNA (1:2 ratio each). We aimed to set two replicates for the technical experiments (see Fig. 2) and three replicates for deletion/inversion frequencies (DIF) measurements (see Fig. 3). Auxin (500 μM of indole-3-acetic acid (IAA)) was added either 2 hours before nucleofection or right after cell plating after nucleofection, and was kept in the culture medium for the whole period. Target protein degradation was confirmed by microscopic analysis of GFP fluorescence loss (Supplementary Material 2). Cells were collected 24 hours after nucleofection. Genomic DNA was isolated from cells using phenol-chloroform extraction.

ddPCR assays. Droplet digital PCR (ddPCR) was performed using a QX100 system (Bio-Rad, USA) with primers and probes specific for the *Ace2* and Ctcf regions, as well as the reference gene *Emid1* (see the Table). ddPCR reactions were set in 20 μl volumes containing 1× ddPCR Supermix for Probes (no dUTP), 900 nM primers and 250 nM probes, and 50 ng genomic DNA. ddPCR reactions for each sample were performed in duplicates. PCR was conducted according to the following program: 95 °C for 10 min, then 45 cycles of 94 °C for 30 s and 58 °C for 1 min, with a ramp rate of 2 °C per second, and a final step at 98 °C for 10 min. The results were analyzed using QuantaSoft 1.7.4 (Bio-Rad). The resulting number was presented as mean +/− combined SEM. For *Ace2* DIF calculations, the initial ddPCR results were multiplied by two, because the gene is located at the X chromosome (the mESCs DGES-1 line used in the study has male XY origin). Statistical analysis for relative differences between DIF across the experiments was performed with the Student test (control sample frequencies were set as 1).

TIDE sequencing. We PCR-amplified genomic site corresponding to gRNA sites F for *Ace2* (412 bp) from 50 ng of mESCs genomic DNA (samples from Fig. 3) (see the Table). PCR products were purified at 2 % agarose gel and Sanger sequenced using forward primer (reverse primer produced similar estimates in small-scale experiment). Sanger files were compared with wild-type control locus in the TIDE application with mostly default parameters (the start of the alignment window was switched to 91 instead of 100 bp) (http://shinyapps.datacurators.nl/tide/) (Brinkman et al., 2014). Average mutation percentage was calculated for three replicates for each degron.

Results

Implementing ddPCR assay at the mouse *Ace2* locus

First, we set out to optimize the Neon nucleofection parameters for mouse embryonic stem cells (mESCs). The outline of a typical experiment is shown in Fig. 1. Following pilot tests, we estimated the average deletion frequency at multiple sites (based on two replicates) across two genomic regions (Fig. 2*a*). We selected one site (*Ace2* F/R2) for Neon optimization. Initially, we tested the nucleofection parameters for wild-type mESCs on a Neon device across 24 basic settings with an EGFP plasmid (data not shown). From this experiment, we identified six conditions demonstrating higher survival rates and GFP fluorescence (conditions #2, 6, 7, 10, 13, 17, 18). Control mESCs were then nucleofected with Cas9 RNP, and the frequencies of deletions were analyzed by droplet digital PCR (ddPCR) (Fig. 2*b*). We observed a general inverse correlation between cell survival and the efficiency of deletion generation (Fig. 2*b*; Supplementary Material 3). Consequently, we selected condition #10 for further experiments, since high cell mortality is undesirable in our approach. Overall, the detection of deletion alleles with ddPCR proved to be specific, enabling reliable analysis of genomic DNA from the total mESCs population (Supplementary Materials 4, 5).

To optimize auxin addition time points, we conducted a small experiment to evaluate the timing of the appearance of deletions after nucleofection. It is known that in RNP nucleofection experiments mutations accumulate gradually. In our observations with mESCs, a small percentage of deletions (~2 % of the 48-hour level) appeared already in the first 3 hours after nucleofection (Fig. 2*c*). After 24 hours, approximately 63 % of deletions from the 48-hour level were observed. Considering the limited survival of mESCs beyond 24 hours without Rad21 or Smc2, this time frame was selected for subsequent experiments with all auxin degron lines. We also performed additional tests of the RNP stability during pre-incubation at 25 °C, revealing that DPBS buffer could effectively substitute the original Buffer R (Neon) without diminishing efficiency (Fig. 2*d*).

Deletion/inversion frequencies in mESCs degron lines

Using the established protocol, we measured deletion/inversion frequencies (DIF) at two genomic sites in various chromatin contexts. The *Ace2* region was considered a "neutral" region located in the middle of a large TAD and showing no expression in mESCs. Here we focused on the 3495 bp deletion (F-R2) (Fig. 2*a*). We also analyzed DIF at another genomic site – a strong Ctcf boundary (Chr5:49.487.342–49.557.342) (Fig. 2*a*; Supplementary Material 6). To account for biological variability, we analyzed two independent experiments that were set with different cell batches and gRNA preparations (Day A, B). We validated these observations with two alternative auxin treatment strategies (Fig. 1*b*, *c*). In the first strategy, we first nucleofected the cells, then plated them in six wells and added auxin to half of them. This way, degradation starts simultaneously with Cas9 cutting. In the second strategy, auxin was added 2 hours prior to nucleofection to reliably remove all protein complexes (Fig. 1*c*). However, this necessitates separate nucleofections for control and treated samples, introducing additional handling variability. Furthermore, protein depletion prior to nucleofection could potentially increase cellular sensitivity to the procedure, possibly affecting ddPCR outcomes.

Surprisingly, we did not observe a Rad21-dependent DIF increase (Fig. 3*a*). More specifically, we documented a small but reproducible decrease in *Ace2* deletion frequencies in all experiments (mean relative decrease across four experiments: –12.1 %, $p = 0.0423$). Inversion frequencies remained unaffected. It is noteworthy that despite the distinct topological characteristics of the examined regions, there was no visible difference for Rad21-related effects, as the Ctcf region also showed minor and not statistically significant alterations in

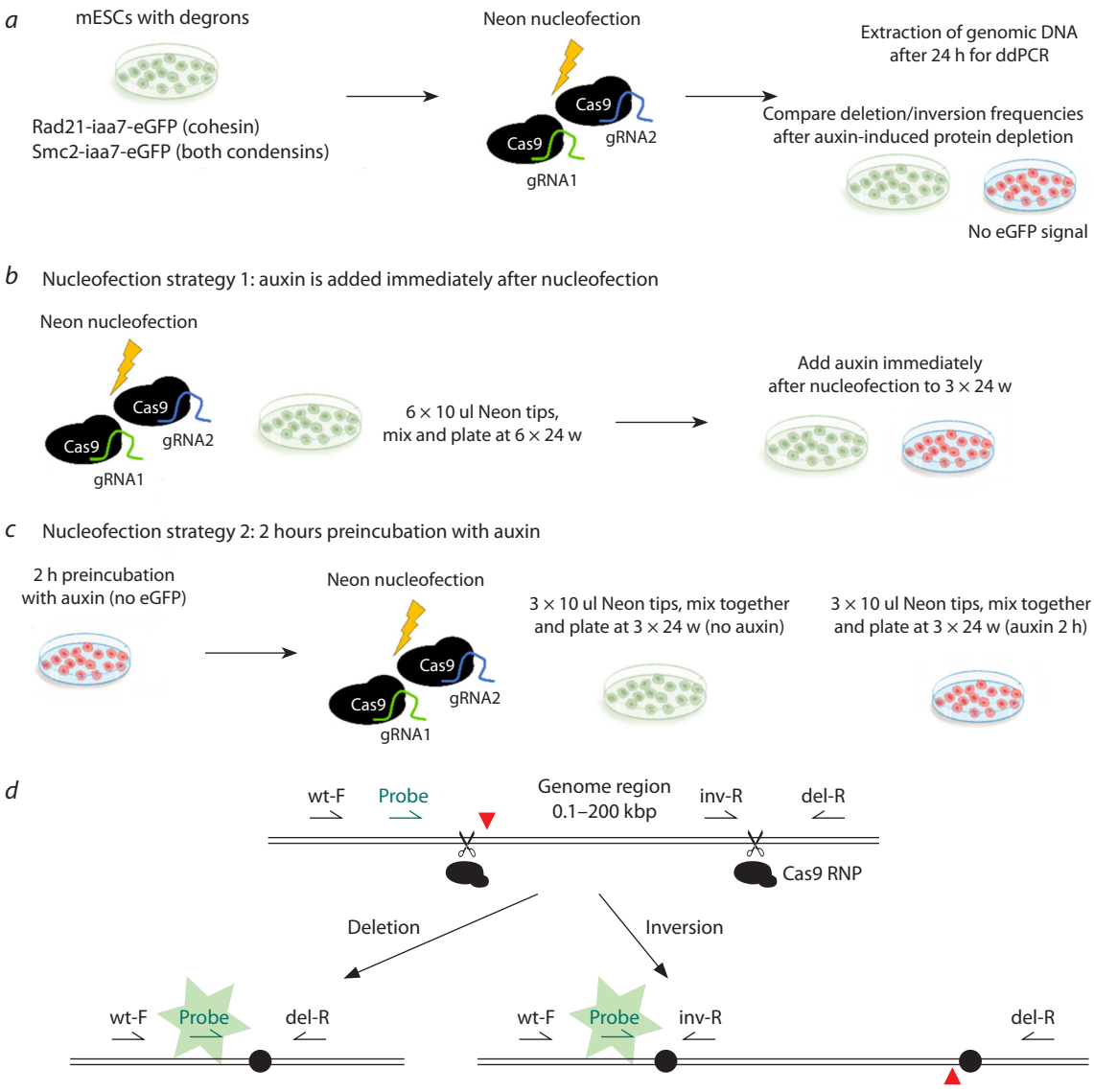


Fig. 1. Experimental approach to study deletion/inversion frequencies in mESCs.

a – mESCs degrons lines were nucleofected with Cas9 and paired gRNAs; After 24 hours, genomic DNA was extracted and analyzed with ddPCR: we measured relative concentrations of the deletion and inversion alleles against the reference gene (*Emid1*). Two different nucleofection strategies with respect to auxin addition were tested; *b* – in the first approach, all cells were mixed together after nucleofection and then split in two sample groups (3+3 × 24 w). Auxin was added immediately after plating to half of the wells; *c* – in the second strategy, cells were preincubated with auxin for 2 hours and then nucleofected independently of control cells. In both cases, auxin was kept in culture medium for the duration of the experiment (24 hours). Degradation of the Rad21 and Smc2 proteins could be tracked by the loss of eGFP fluorescence (Supplementary Material 2); *d* – scheme of the droplet digital PCR modification designed to detect genomic rearrangements (ddXR method) (Watry et al., 2020). Induction of paired DNA breaks could lead to excision of the intermediate fragment, resulting in deletion or inversion. The loss or inversion of the fragment allows to efficiently amplify PCR product, activating probe fluorescence.

deletion frequencies (mean relative decrease across four experiments: –10 %, $p = 0.109$) (Fig. 3*a*).

Conversely, Smc2 depletion showed a trend towards DIF increase in one of the days (Day A, auxin added 2 hours before nucleofection (Fig. 3)), where it reached +33 % (*Ace2* deletions), +75 % (*Ace2* inversions), +61 % (Ctcf deletions), +63 % (Ctcf inversion). This effect, however, was not replicated in the subsequent trial (Day B, auxin added 2 hours before nucleofection) (Fig. 3*b*) upon switching to a different Cas9 batch. Depletion effect on *Ace2* deletions was not statistically significant, nor were changes in *Ace2* inversion frequencies at a significance level of 0.05 (mean relative increase across four experiments: +39 %, $p = 0.088$). These discrepancies could be caused by some unaccounted biological factors, such as varying Cas9 batch efficiencies or differences in cell survival post-nucleofection between experiments (see Discussion). The role of the condensin complexes in distant end joining needs additional examinations in the future.

DSB repair efficiency in degron lines

Our objective was to investigate the impact of SMC protein depletion on DSB repair efficiency, particularly at a single DSB site or at closely positioned pairs of DSBs. Previous research indicated that 34 bp deletions are repaired differently

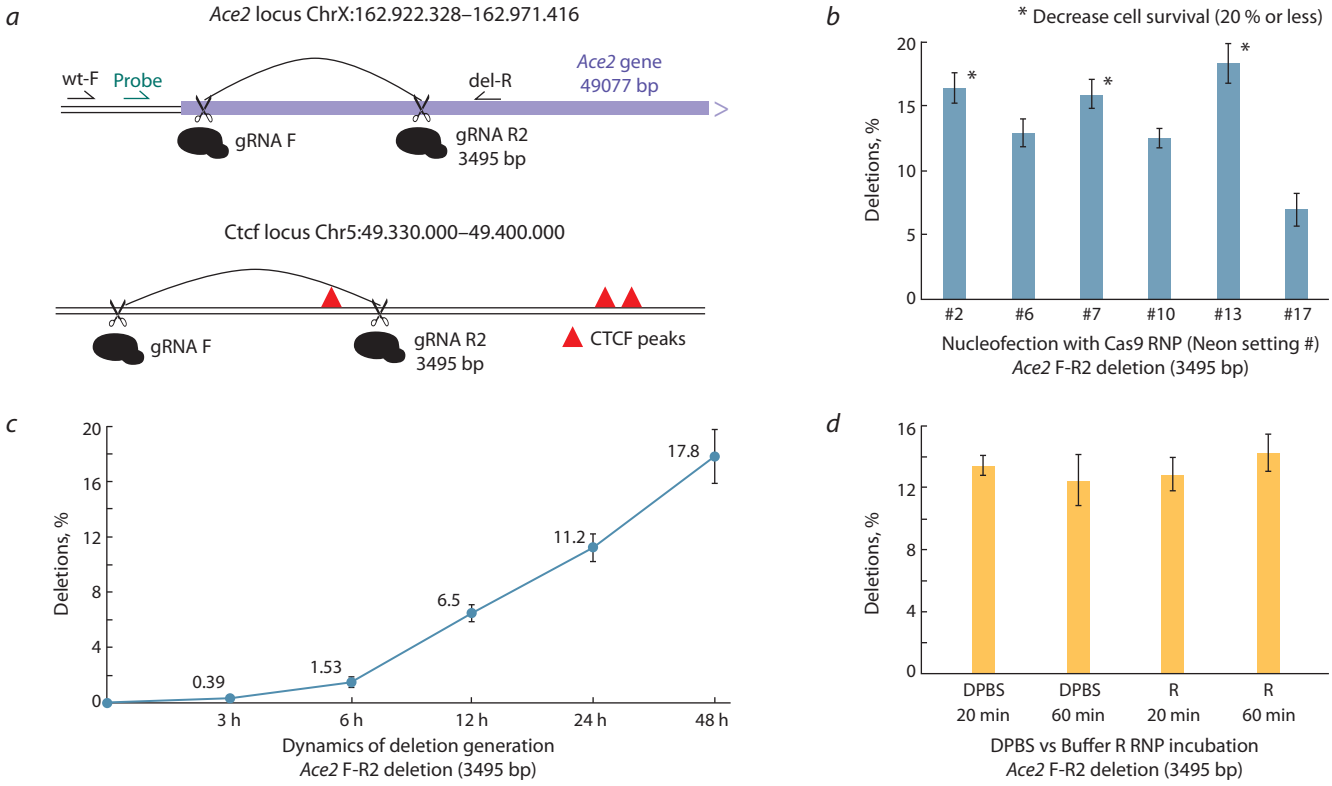


Fig. 2. Optimization of Neon conditions for deletion generation.

a – deletions examined in the study. Primers and the probe for ddPCR are shown for the *Ace2* locus; *b–d* – optimizing mESCs Neon nucleofection conditions with the F-R2 (*Ace2*) gRNA pair.

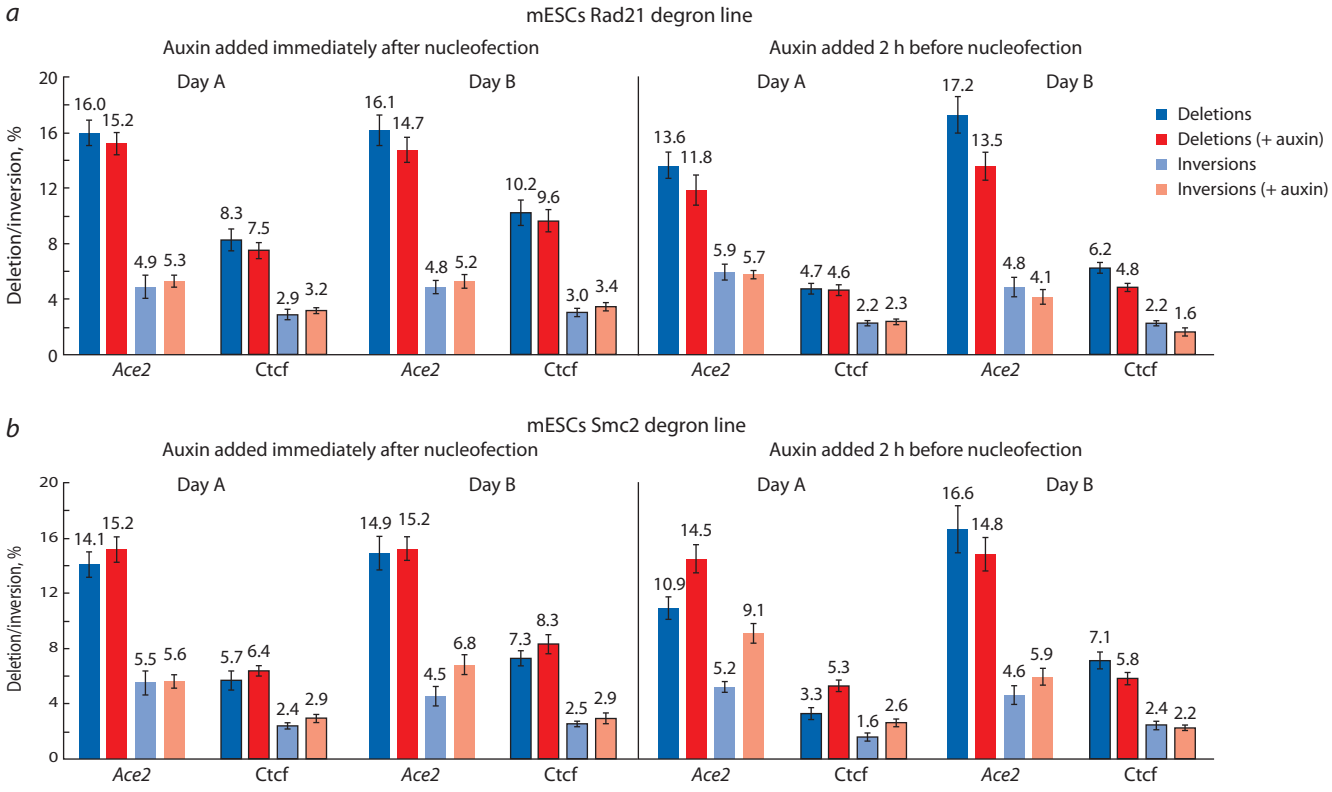


Fig. 3. Deletion/inversion frequencies (DIF) for different genomic regions before and after addition of auxin.

a – DIF for *Ace2* F-R2 and Ctcf F-R2 regions in Rad21 degron line; *b* – DIF for *Ace2* F-R2 and Ctcf F-R2 regions in Smc2 degron line. Data presented as average between three nucleofection replicates and combined SEM. Statistical analysis for mean relative values across four biological experiments is provided in the main text.

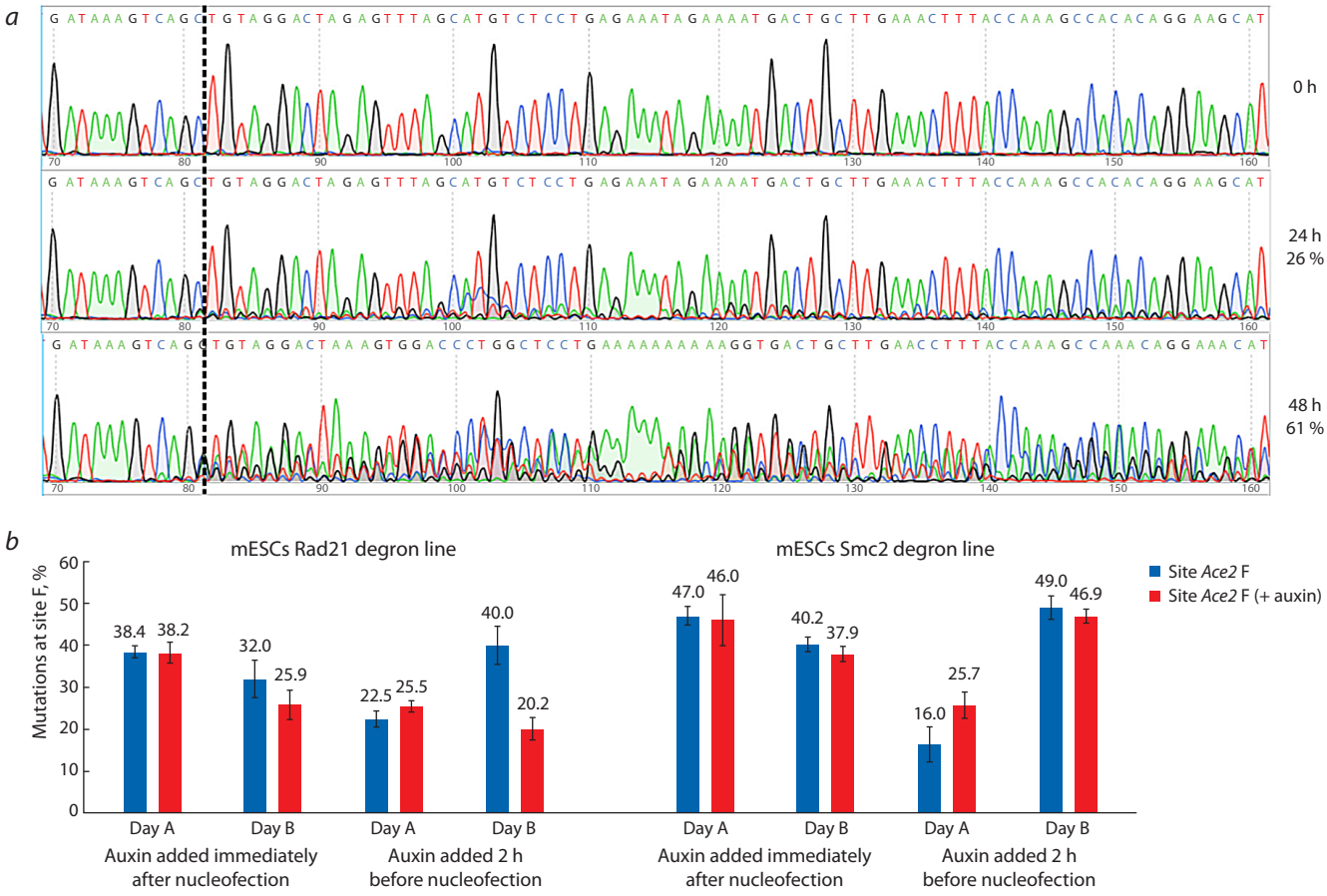


Fig. 4. DSB repair efficiency at a single site *Ace2* F, measured by TIDE.

a – demonstration of Sanger data for the control unedited locus and the mutated locus in the cells from Fig. 2, *d*. Cut site is marked with a dotted line. Editing efficiency measured with TIDE is shown as %; *b* – frequencies of site modifications in various degron lines from Fig. 3. Data shown as average and SEM.

from larger 3200 bp deletions in Rad21-deficient cells (Gelot et al., 2016). One of the drawbacks of the ddPCR method is its inability to detect small deletions, due to interference with the wild-type locus amplification. The authors of the ddXR method recommend digesting genomic DNA to selectively eliminate wild-type genomic loci from ddPCR amplification. With this trick, they were able to amplify deletions as small as 91 bp (Watry et al., 2020).

We managed to apply restriction to a region of 192 bp at the *Ace2* locus (Supplementary Material 7), although attempts to apply it to other short deletions were less successful (data not shown). Given these limitations, the ddPCR method was deemed unsuitable for analyzing 34 bp deletions. Instead, to screen how SMC-protein depletion affects DSB repair at a single end we utilized the Tracking of Indels by DEcomposition (TIDE) method (Brinkman et al., 2014), a straightforward approach based on Sanger sequencing of the break site. This method facilitates the demultiplexing and calculation of Cas9 cut signatures at the break, thereby estimating DSB repair efficiency as a percentage of mutant alleles. Estimating indel mutation signatures at the break site serves not only as an indicator of Cas9 activity but also as a measure of nucleofection efficiency (Fig. 4). We PCR amplified and sequenced regions at the Cas9 target site for the *Ace2* F gRNA (Fig. 2*a*) (the same samples analyzed with ddPCR (Fig. 3)).

For the Smc2 experiments, we did not detect any significant differences in editing efficiency. The slight decrease in efficiency post-auxin treatment was counterbalanced by an increase in DIF (since deletion/inversion events eliminate *Ace2* F sites from PCR amplification in TIDE analysis) (Fig. 4*b*). For Rad21 depletions, a decrease in editing efficiencies was noted, potentially reflecting increased cell vulnerability under high RNP loads in the absence of Rad21. Notably, one experimental condition (Rad21/Smc2, Day A, auxin added 2 h before nucleofection) exhibited a 2-fold reduction in editing efficiencies. In this scenario, auxin addition paradoxically enhanced Cas9 editing for both degron lines (Fig. 4*b*), yet DIF were impacted differently in Rad21 and Smc2 lines (Fig. 3). This suggests that at lower editing efficiencies (RNP load), cells might respond differently to protein depletion. For example, Smc2 depletion could permeabilize cells for nucleofection, possibly due to a cell cycle shift or chromosome decondensation. We plan to perform nucleofections with various RNP concentrations in the future to verify this effect.

Discussion

We have performed a series of experiments with auxin degradation and CRISPR/Cas9-induced DSBs using a collection of mESCs with the SMC degrons. mESCs represent an interesting object for studying DNA repair. For instance,

mESCs mostly rely on HR to preserve genome stability (Choi et al., 2017) and have different end-joining mechanisms based on specialized polymerases (Schimmel et al., 2017). Since mESCs are difficult to edit with lipofection, we adapted a protocol to generate deletions with Neon nucleofections. This method, in conjunction with ddPCR, demonstrated high efficiency and sensitivity in detecting deletions and inversions, with an average modification rate of 60 % for the *Ace2* locus after Neon nucleofection (TIDE at the F site + deletions + insertions) (Fig. 3, 4). This level of editing is notable compared to plasmid transfection outcomes without selection. However, we encountered significant variability in deletion/inversion frequencies (DIF) across experiments, highlighting the influence of numerous biological factors on experimental outcomes.

Cas9, a crucial component in our experiments, can significantly impact DSB repair dynamics. Variations in the Cas9:gRNA ratio can dramatically alter editing outcomes (Chenouard et al., 2023), with repair processes potentially delayed up to 20 hours due to persistent Cas9-DNA binding (Kim et al., 2014; Brinkman et al., 2018). Furthermore, Cas9 retention at break sites can modify blunt ends into 3′-overhang trimmed ends (Stephenson et al., 2018; Jones et al., 2021), necessitating different polymerases for non-homologous end-joining. Variability was also observed between different lots of Cas9-NLS (Biolabmix) even at identical molar concentrations. To account for all these issues, we performed experiments with two strategies of auxin addition and set three nucleofection replicates. We also performed two biological replicates with different mESCs batches, gRNA preps and Neon tips. From our experience, such experiments require very careful examination of the optimal experimental conditions, especially when the gene of interest has strong pleiotropic effects on cell homeostasis.

Our timing analysis indicated that cells accumulate 70 % of deletions within 24 hours, and only 2 % in the first 3 hours, suggesting that auxin could be added within a 0–3 hour window after nucleofection without significantly compromising deletion generation. Furthermore, we confirmed that DPBS incubation does not compromise RNP activity, providing a viable alternative to Buffer R. Notably, immediate post-nucleofection auxin addition exhibited lesser variability compared to a 2-hour pre-incubation strategy (Fig. 3), demonstrating the feasibility of its use in future setups due to its uniform experimental conditions.

Analysis of data on the frequencies of deletions and inversions for various mESCs clones with degrons allowed us to draw the following conclusions. We expected that Rad21 depletion will cause elevated rates of deletions and inversions due to unconstrained movement of the DSB ends, as it was reported by another group. In their report, there was a 30 % increase in the amount of cells with a 3 kb deletion (Gelot et al., 2016) after Rad21 siRNA knockdown. Some other reports using cytogenetic and microscopic analysis also suggested that Rad21 knockdown provokes DNA rearrangements (Wu, Yu, 2012). So far, we have not found any significant stimulatory effects of Rad21 depletion on DIF (Fig. 3). Given that the authors of the initial report (Gelot et al., 2016) worked with a different experimental setting (plasmid transfection with inducible I-SceI, siRNA Rad21 knockdown, SV40-transformed GM639 human WT fibroblasts) and had an alternative detection strategy, our results may reflect differences between the experimental systems. In our setting, the protein was removed almost completely after 2 hours (Yunusova et al., 2021) and Cas9 RNP was active from the beginning (see timings, Fig. 2*c*). Also, mESCs are more sensitive to DNA damage and may react to DSB differently than immortalized fibroblasts (Choi et al., 2017).

It is possible that the absence of Rad21 sensibilizes cells to DNA damage resulting in a decreased opportunity for distant end-joining events to happen, hiding the stimulatory effect. This would lead to a lower amount of TIDE signal, as we see in our data (Fig. 4*b*). However, this does not explain why inversion frequencies are not negatively affected (Fig. 3). In theory, the effect of Rad21 degradation may be more noticeable for extremely distant DSBs, such as a 26 kbp deletion that we plan to analyze in the future (Supplementary Material 6). Correlation between topology and DSB is another long-standing question. In our case, deletion over the Ctcf site at the TAD border was not noticeably affected by cohesin depletion.

Unexpectedly, our findings hint at a significant role of Smc2 depletion in promoting genomic rearrangements, although data variability necessitates further investigation. Condensins are not directly involved in DNA repair, but could affect it via side effects (defects in chromosome segregation, chromatin decondensation in G2/M). Cell cycle is an important determinant of a DSB repair outcome. It is well known that G1 DSBs are repaired with slower kinetics (Arnould et al., 2023). Synchronization of human fibroblasts in the G1 phase showed no end-joining stimulation from Rad21 knockdown (Gelot et al., 2016). We and others analyzed cell cycles in mESCs with Rad21 and Smc2 depletion and found that after 6 hours they accumulate in the G2/M phase (manuscript under preparation). Judging from these data, Rad21 and Smc2 clones have the same cell cycle profile. Thus, cell cycle shift alone would not explain the difference between Rad21 and Smc2 depletion effects. In this study, we could only work with an unsynchronized mESCs population. Synchronization of mESCs is very challenging and imposes additional cell lethality making this approach unsuitable for our goal.

We plan to expand our investigations with the repertoire of deletions at other genomic regions with interesting topological organization. We will also try other improvements, such as NGS sequencing with Unique Molecular Identifiers (UMIs) for Cas9 target sites to account for editing efficiency. In the future, we will also extend our findings to simpler, synchronizable human cell lines such as HAP1 and HCT116, which also harbor Rad21/Smc2 degrons, to further dissect these complex dynamics.

Conclusion

Cohesin facilitates genome stability by limiting DNA movements during replication. By this logic, supported by experimental data, the frequencies of deletion between paired distant breaks will increase after cohesin removal. We could not reproduce these findings in the Rads21 auxin-degron cell line as we did not see an increase in deletion or inversion frequencies. This may reflect differences between experimental systems. Both Rad21 and Smc2 degron studies will require more iterations to account for biological variability.

References

Arnould C., Rocher V., Finoux A.-L., Clouaire T., Li K., Zhou F., Caron P., Mangeot P.E., Ricci E.P., Mourad R., Haber J.E., Noordermeer D., Legube G. Loop extrusion as a mechanism for formation of DNA damage repair foci. *Nature*. 2021;590(7847):660-665. DOI 10.1038/s41586-021-03193-z

Arnould C., Rocher V., Saur F., Bader A.S., Muzzopappa F., Collins S., Lesage E., Le Bozec B., Puget N., Clouaire T., Mangeat T., Mourad R., Ahituv N., Noordermeer D., Erdel F., Bushell M., Marnef A., Legube G. Chromatin compartmentalization regulates the response to DNA damage. *Nature*. 2023;623(7985):183-192. DOI 10.1038/s41586-023-06635-y

Aymard F., Bugler B., Schmidt C.K., Guillou E., Caron P., Briois S., Iacovoni J.S., Daburon V., Miller K.M., Jackson S.P., Legube G. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* 2014;21(4):366-374. DOI 10.1038/nsmb.2796

Baergen A.K., Jeusset L.M., Lichtensztejn Z., McManus K.J. Diminished condensin gene expression drives chromosome instability that may contribute to colorectal cancer pathogenesis. *Cancers* (*Basel*). 2019;11(8):1066. DOI 10.3390/cancers11081066

Brinkman E.K., Chen T., Amendola M., van Steensel B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res*. 2014;42(22):e168. DOI 10.1093/nar/gku936

Brinkman E.K., Chen T., de Haas M., Holland H.A., Akhtar W., van Steensel B. Kinetics and fidelity of the repair of Cas9-induced double-strand DNA breaks. *Mol. Cell*. 2018;70(5):801-813.e6. DOI 10.1016/j.molcel.2018.04.016

Canver M.C., Bauer D.E., Dass A., Yien Y.Y., Chung J., Masuda T., Maeda T., Paw B.H., Orkin S.H. Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.* 2014;289(31):21312-21324. DOI 10.1074/jbc.M114.564625

Chang H.-Y., Lee C.-Y., Lu C.-H., Lee W., Yang H.-L., Yeh H.-Y., Li H.-W., Chi P. Microcephaly family protein MCPH1 stabilizes RAD51 filaments. *Nucleic Acids Res.* 2020;48(16):9135-9146. DOI 10.1093/nar/gkaa636

Chenouard V., Leray I., Tesson L., Remy S., Allan A., Archer D., Caulder A., Fortun A., Bernardeau K., Cherifi Y., Teboul L., David L., Anegon I. Excess of guide RNA reduces knockin efficiency and drastically increases on-target large deletions. *iScience*. 2023;26(4):106399. DOI 10.1016/j.isci.2023.106399

Choi E.-H., Yoon S., Park K.-S., Kim K.P. The homologous recombination machinery orchestrates post-replication DNA repair during self-renewal of mouse embryonic stem cells. *Sci. Rep.* 2017;7(1):11610. DOI 10.1038/s41598-017-11951-1

Dobbs F.M., van Eijk P., Fellows M.D., Loiacono L., Nitsch R., Reed S.H. Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq. *Nat. Commun.* 2022;13(1):3989. DOI 10.1038/s41467-022-31702-9

Dyson S., Segura J., Martínez-García B., Valdés A., Roca J. Condensin minimizes topoisomerase II-mediated entanglements of DNA *in vivo*. *EMBO J.* 2021;40(1):e105393. DOI 10.15252/embj.2020105393

Gelot C., Guirouilh-Barbat J., Le Guen T., Dardillac E., Chailleux C., Canitrot Y., Lopez B.S. The cohesin complex prevents the end joining of distant DNA double-strand ends. *Mol. Cell*. 2016;61(1):15-26. DOI 10.1016/j.molcel.2015.11.002

Houlard M., Cutts E.E., Shamim M.S., Godwin J., Weisz D., Presser Aiden A., Lieberman Aiden E., Schermelleh L., Vannini A., Nasmyth K. MCPH1 inhibits condensin II during interphase by regulating its SMC2-Kleisin interface. *eLife.* 2021;10:e73348. DOI 10.7554/eLife.73348

Jones S.K., Hawkins J.A., Johnson N.V., Jung C., Hu K., Rybarski J.R., Chen J.S., Doudna J.A., Press W.H., Finkelstein I.J. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat. Biotechnol.* 2021;39(1):84-93. DOI 10.1038/s41587-020-0646-5

Kabirova E., Nurislamov A., Shadskiy A., Smirnov A., Popov A., Salnikov P., Battulin N., Fishman V. Function and evolution of the loop extrusion machinery in animals. *Int. J. Mol. Sci.* 2023;24(5):5017. DOI 10.3390/ijms24055017

Kim S., Kim D., Cho S.W., Kim J., Kim J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* 2014;24(6):1012-1019. DOI 10.1101/gr.171322.113

Li D., Sun X., Yu F., Perle M.A., Araten D., Boeke J.D. Application of counter-selectable marker PIGA in engineering designer deletion cell lines and characterization of CRISPR deletion efficiency. *Nucleic Acids Res.* 2021;49(5):2642-2654. DOI 10.1093/nar/gkab035

Minchell N.E., Keszthelyi A., Baxter J. Cohesin causes replicative DNA damage by trapping DNA topological stress. *Mol. Cell*. 2020;78(4):739-751.e8. DOI 10.1016/j.molcel.2020.03.013

Piazza A., Bordelet H., Dumont A., Thierry A., Savocco J., Girard F., Koszul R. Cohesin regulates homology search during recombinational DNA repair. *Nat. Cell Biol.* 2021;23(11):1176-1186. DOI 10.1038/s41556-021-00783-x

Schimmel J., Kool H., van Schendel R., Tijsterman M. Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. *EMBO J.* 2017;36(24):3634-3649. DOI 10.15252/embj.201796948

Schwarzer W., Abdennur N., Goloborodko A., Pekowska A., Fudenberg G., Loe-Mie Y., Fonseca N.A., Huber W., Haering C.H., Mirny L., Spitz F. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*. 2017;551(7678):51-56. DOI 10.1038/nature24281

Stephenson A.A., Raper A.T., Suo Z. Bidirectional degradation of DNA cleavage products catalyzed by CRISPR/Cas9. *J. Am. Chem. Soc.* 2018;140(10):3743-3750. DOI 10.1021/jacs.7b13050

Ström L., Lindroos H.B., Shirahige K., Sjögren C. Postreplicative recruitment of cohesin to double-strand breaks is required for DNA repair. *Mol. Cell*. 2004;16(6):1003-1015. DOI 10.1016/j.molcel.2004.11.026

Ünal E., Arbel-Eden A., Sattler U., Shroff R., Lichten M., Haber J.E., Koshland D. DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol. Cell*. 2004;16(6):991-1002. DOI 10.1016/j.molcel.2004.11.027

Watry H.L., Feliciano C.M., Gjoni K., Takahashi G., Miyaoka Y., Conklin B.R., Judge L.M. Rapid, precise quantification of large DNA excisions and inversions by ddPCR. *Sci. Rep.* 2020;10(1):14896. DOI 10.1038/s41598-020-71742-z

Wood J.L., Liang Y., Li K., Chen J. Microcephalin/MCPH1 associates with the Condensin II complex to function in homologous recombination repair. *J. Biol. Chem.* 2008;283(43):29586-29592. DOI 10.1074/jbc.M804080200

Wu N., Yu H. The Smc complexes in DNA damage response. *Cell Biosci.* 2012;2(1):5. DOI 10.1186/2045-3701-2-5

Wu X., Mondal G., Wang X., Wu J., Yang L., Pankratz V.S., Rowley M., Couch F.J. Microcephalin regulates BRCA2 and Rad51-associated DNA double-strand break repair. *Cancer Res.* 2009;69(13):5531-5536. DOI 10.1158/0008-5472.CAN-08-4834

Yunusova A., Smirnov A., Shnaider T., Lukyanchikova V., Afonnikova S., Battulin N. Evaluation of the OsTIR1 and AtAFB2 AID systems for genome architectural protein degradation in mammalian cells. *Front. Mol. Biosci.* 2021;8:757394. DOI 10.3389/fmolb.2021.757394

Conflict of interest. The authors declare no conflict of interest.
Received March 29, 2024. Revised April 23, 2024. Accepted July 22, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):592-601

DOI 10.18699/vjgb-24-66

Structure and evolution of metapolycentromeres

E.O. Grishko, P.M. Borodin

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

Abstract. Metapolycentromeres consist of multiple sequential domains of centromeric chromatin associated with a centromere-specific variant of histone H3 (CENP-A), functioning collectively as a single centromere. To date, they have been revealed in nine flowering plant, five insect and six vertebrate species. In this paper, we focus on their structure and possible mechanisms of emergence and evolution. The metapolycentromeres may vary in the number of centromeric domains and in their genetic content and epigenetic modifications. However, these variations do not seem to affect their function. The emergence of metapolycentromeres has been attributed to multiple Robertsonian translocations and segmental duplications. Conditions of genomic instability, such as interspecific hybridization and malignant neoplasms, are suggested as triggers for the *de novo* emergence of metapolycentromeres. Addressing the "centromere paradox" – the rapid evolution of centromeric DNA and proteins despite their conserved cellular function – we explore the centromere drive hypothesis as a plausible explanation for the dynamic evolution of centromeres in general, and in particular the emergence of metapolycentromeres and holocentromeres. Apparently, metapolycentromeres are more common across different species than it was believed until recently. Indeed, a systematic review of the available cytogenetic publications allowed us to identify 27 candidate species with metapolycentromeres. The list of the already established and newly revealed candidate species thus spans 27 species of flowering plants and eight species of gymnosperm plants, five species of insects, and seven species of vertebrates. This indicates an erratic phylogenetic distribution of the species with metapolycentromeres and may suggest an independent emergence of the metapolycentromeres in the course of evolution. However, the current catalog of species with identified and likely metapolycentromeres remains too short to draw reliable conclusions about their evolution, particularly in the absence of knowledge about related species without metapolycentromeres for comparative analysis. More studies are necessary to shed light on the mechanisms of metapolycentromere formation and evolution.

Key words: centromere; centromere size; centromere type; metapolycentromeres.

For citation: Grishko E.O., Borodin P.M. Structure and evolution of metapolycentromeres. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):592-601. DOI 10.18699/vjgb-24-66

Funding. This research was funded by the Russian Science Foundation, grant number 23-24-00304.

Acknowledgements. The authors express sincere gratitude to A. Torgasheva and L. Malinovskaya for their assistance in preparing the article and to A. Nurislamov, G. Koksharova and M. Deryuzhenko for the helpful comments.

Структура и эволюция метаполицентромер

Е.О. Гришко, П.М. Бородин

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

Аннотация. Метаполицентромеры состоят из нескольких последовательных доменов центромерного хроматина, связанных со специфичным для центромеры вариантом гистона H3 – CENP-A, которые вместе функционируют как одна центромера. Они были открыты недавно и обнаружены у девяти видов цветковых растений, пяти видов насекомых и шести видов позвоночных животных. В данном обзоре рассматриваются структура метаполицентромер и возможные механизмы их возникновения и эволюции. Метаполицентромеры могут различаться по количеству центромерных доменов, последовательностям ДНК и эпигенетическим модификациям. Однако эти различия, по-видимому, не влияют на их функцию. Появление метаполицентромер объясняют множественными робертсоновскими транслокациями и сегментными дупликациями. В условиях геномной нестабильности (при межвидовой гибридизации и в ходе канцерогенеза) метаполицентромеры могут возникать *de novo*. Гипотеза центромерного драйва представляется убедительным объяснением эволюции центромер в целом и образования метаполицентромер и голоцентромер в частности. По-видимому, метаполицентромеры встречаются чаще, чем принято считать. Систематический обзор доступных цитогенетических публикаций позволил нам дополнительно идентифицировать 27 видов-кандидатов с метаполицентромерами. Таким образом, список уже установленных и вновь найденных видов-кандидатов охватывает 27 видов цветковых и восемь видов голосеменных растений, пять видов насекомых и семь видов позвоночных животных. Виды, включенные

в этот список, спорадически распределены по филогенетическому древу. Это может указывать на независимое эволюционное возникновение метаполицентромер. Однако существующий список видов с идентифицированными и предполагаемыми метаполицентромерами слишком короткий, чтобы сделать надежные выводы об их эволюции, особенно в отсутствие знаний о родственных видах без метаполицентромер для сравнительного анализа. Необходимы дополнительные исследования для того, чтобы пролить свет на механизмы образования и эволюции метаполицентромер.
Ключевые слова: центромера; размер центромеры; тип центромеры; метаполицентромеры.

Four main types of centromeres

The centromere is the region of the chromosome to which spindle filaments attach during mitosis and meiosis. It consists of centromeric DNA and a kinetochore protein complex through which the spindle microtubules attach to the chromosome. Centromeres play a critical role in maintaining chromosome integrity and controlling chromosome segregation during cell division. Disruption of the structure and function of centromeres in mitosis can lead to cell death, and in meiosis, to the formation of unbalanced gametes and sterility. Despite this conserved function, common to all eukaryotes, the centromeres of different organisms can vary significantly in both structure and size (Talbert, Henikoff, 2020). The only epigenetic mark of the centromere, characteristic of the vast majority of species, is the centromere variant of histone H3, the CENP-A protein (Mendiburo et al., 2011).

There are four main types of centromeres: regional centromeres, point centromeres, metapolycentromeres and holocentromeres (Talbert, Henikoff, 2020) (Fig. 1).

Regional centromeres are the most common type of centromere. Cytologically, the regional centromere can be detected as a primary constriction (Flemming, 1882). It is built on centromeric chromatin, marked by CENP-A. Based on centromeric chromatin, the kinetochore is assembled (Cleveland et al., 2003) (Fig. 2). The length of centromeric chromatin varies significantly among different species and can range from several thousand to millions of base pairs (bp) (Haupt et al., 2001; Kanesaki et al., 2015). Usually, centromeric and pericentromeric chromatin consists of highly repeated DNA sequences: satellite DNA or mobile genetic elements. However, centromeres based on non-repeated sequences have also been found (Glöckner, Heidel, 2009; Kanesaki et al., 2015; Talbert et al., 2018). The centromeric sequences of most species consist predominantly of satellite DNA.

Centromeric tandem repeats vary in the number, length, and nucleotide composition of repeating fragments (monomers), but usually have a length of 100–400 bp (Melters et al., 2013). This size ensures DNA coiling around 1–2 nucleosomes. The monomers of the satellite DNA are often A/T rich (Melters et al., 2013), which presumably reduces DNA bending energy and promotes nucleosome folding. The sequences of centromeric repeats can vary even between closely related species (Lee et al., 2005; Talbert et al., 2018). Moreover, even within the same species, centromeres of different chromosomes can consist of either tandem repeats belonging to the same family or completely different repeats (Ahmad et al., 2020; Balzano, Giunta, 2020).

It is known that centromeric repeats are actively transcribed, and the resulting transcripts play an important role in maintaining centromere structure (Talbert, Henikoff, 2018).

Point centromeres are found only in the chromosomes of the budding yeast *Saccharomyces cerevisiae* (Nagpal, Fierz, 2021). They contain only one centromeric nucleosome, the so-called hemisome (heminucleosome), consisting of histones H4, H2A, H2B, and Cse4 (CENP-A homolog) in a single copy (Furuyama, Biggins, 2007; Henikoff et al., 2014). Only one spindle microtubule is attached to the point centromeres (Winey et al., 1995).

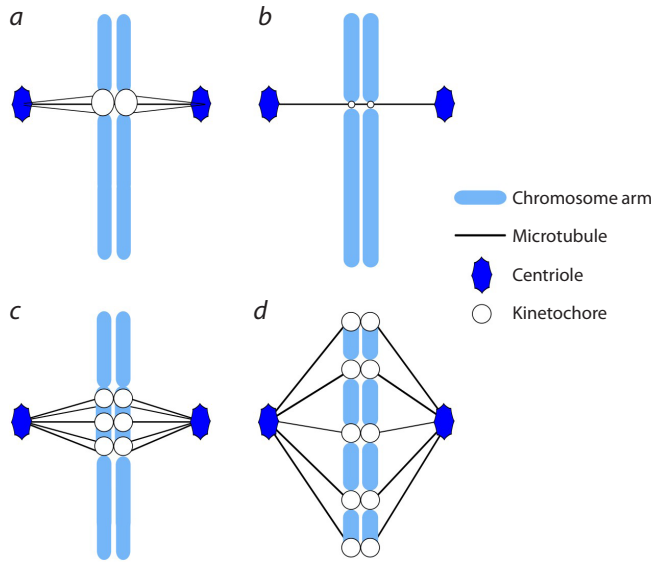


Fig. 1. Four main types of centromeres: regional centromeres (*a*); point centromeres (*b*); metapolycentromeres (*c*), and holocentromeres (*d*).

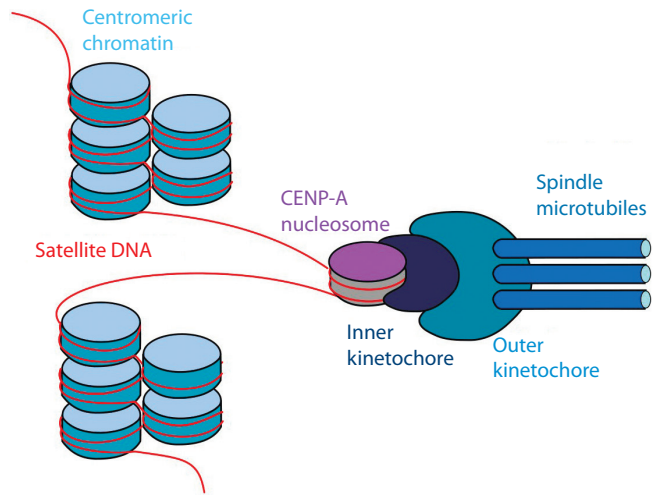


Fig. 2. Regional centromere structure, according to H. Nagpal and B. Fierz (2021), modified.

Holocentromeres do not form a primary constriction since the spindle microtubules have attachment points along the entire length of the chromosome. Some holocentromeres have no centromeric chromatin at all and CENP-A is distributed evenly along the entire length of the chromosome. In other holocentromeres, the centromeric chromatin forms small, equally spaced, repeated clusters along the entire length of the chromosome (Senaratne et al., 2022). Holocentromeres were detected in 700 species of plants and animals with holocentromeres (Melters et al., 2012). More information about this topic can be found in the reviews (Senaratne et al., 2022; Wang et al., 2022; Castellani et al., 2024; Kuo et al., 2024).

Metapolycentromeres consist of several sequential domains of centromeric chromatin associated with CENP-A and functioning as a single centromere. They are considered a transitional type between regional centromeres and holocentromeres (Neumann et al., 2012).

Here we review the structural features and evolution of metapolycentromeres, the most recently discovered and extremely rare type of centromere.

CENP-A as a centromere identifier

The position of the centromere is determined epigenetically, not by a specific DNA sequence, and the centromeric variant of histone H3 is considered the universal epigenetic mark of a functional centromere (Mendiburo et al., 2011). Centromeric histone H3 has several taxon-specific synonyms: CENP-A in animals, CENH3 in plants, CID (centromere identifier) in drosophila, HCP-3 in nematodes, Cnp1 in fission yeast, and Cse4 in budding yeast. In this article, for convenience, we will use the term CENP-A, as it is the most commonly used. CENP-A or its homologues are found in the centromeres of all eukaryotic species studied, with very rare exceptions including some species of lepidopterans and hemipterans, trypanosomes, and fungi (Drinnenberg et al., 2014; van Hooff et al., 2017; Navarro-Mendoza et al., 2019; Senaratne et al., 2021). The presence of CENP-A on a chromosomal site is necessary and sufficient for the formation of a functional centromere and for ensuring its inheritance (Mendiburo et al., 2011).

CENP-A, like canonical histone H3, includes two domains: an N-terminal domain and a C-terminal domain. The latter is integrated into the nucleosomal octamer and forms the nucleosome body (Sullivan K.F. et al., 1994). This domain contains the following regions (from the N end to the C end): αN-helix, α1-helix, Loop1, α2-helix, Loop2, α3-helix, and C-terminal disordered tail (Black et al., 2004; Tachiwana et al., 2012). Human CENP-A shows 48 % homology with the canonical histone H3, making it the most distinct histone H3 variant. The N-terminal domain of human CENP-A is much shorter than that of canonical H3, and the amino acid sequence has the least homology to the canonical H3 sequence of all regions of the protein. The C-terminal domain is 68 % identical to the canonical one (Sullivan K.F. et al., 1994).

Typically, histones are highly conserved, but the amino acid composition of CENP-A varies significantly between different species (Maheshwari et al., 2015). The N-terminal domain and loop 1 of the C-terminal domain interact with centromeric DNA and show signs of positive selection in some organisms, for example, in *Drosophila melanogaster* and *Arabidopsis thaliana*. The main part of the C-terminal domain (except loop 1) is typically conserved (Malik, Henikoff, 2001; Talbert et al., 2002; Maheshwari et al., 2015).

Thus, the centromere's position is epigenetically marked by CENP-A, which is essential for centromere function across eukaryotes. It shows significant interspecies variation and adaptive evolution, highlighting its critical role in centromere functionality and inheritance.

Structure and characteristics of metapolycentromeres

Metapolycentromeres are found in a few species (see the Table). The number of chromosomes containing metapolycentromeres differs between species. In some species, all chromosomes contain metapolycentromeres. In others, metapolycentromeres are present on a few chromosomes or on just one, while the remaining chromosomes contain regional centromeres (Huang Y.-C. et al., 2016; Malinovskaya et al., 2022). Moreover, between populations of the ant species *Trachymyrmex holmgreni*, variation in the number of chromosome pairs containing metapolycentromeres was observed, from 1 to all 20 pairs (Cardoso et al., 2018). Metapolycentromeres also vary in size. They may occupy from 5 to 40 % of the chromosome length (Malinovskaya et al., 2022).

On routinely stained preparations of metaphase chromosomes, metapolycentromeres appear as elongated primary constrictions (Fig. 3*a*) (Drpic et al., 2018; Malinovskaya et al., 2022). Immunolocalization of CENP-A provides a more accurate identification of metapolycentromeres. This method of identification has been applied to the metaphase chromosomes of *Pisum sativum*, *P. fulvum*, *Lathyrus* spp., *Tribolium castaneum*, and *Muntiacus muntjak* (Neumann et al., 2012, 2015; Drpic et al., 2018; Gržan et al., 2020). Recently, L.P. Malinovskaya et al. (2022) and E. Grishko et al. (2023) detected metapolycentromeres in five species of songbirds: Gouldian finch, European pied flycatcher, Eurasian bullfinch, domestic canary, and common linnet, using non-specific antibodies to the human centromere (ACA) on preparations of surface-spread synaptonemal complexes (Fig. 3*b*).

In all cases, the signals from centromeric chromatin domains were distributed in a paired bead-like pattern, with anticentromere antibodies always binding to the outer side of the primary constriction. In some cases, in legumes and songbirds, centromeric chromatin domains were fused, forming a linear structure (Neumann et al., 2012, 2015; Malinovskaya et al., 2022; Grishko et al., 2023). In the songbirds, unequal spacing between domains and unequal numbers of domains on homologous chromosomes of the same karyotype were observed (Grishko et al., 2023).

The use of ChIP-seq with antibodies to CENP-A showed that the centromeric chromatin of peas metapolycentromeres consists predominantly of AT-rich satellite DNA 150–400 bp long. A combination of ChIP-seq with long-read sequencing demonstrated that the centromeric chromatin of metapolycentromeres also contains various retrotransposons. At the moment, the sequence of only one metapolycentromere has been established. The metapolycentromere of *P. sativum* chromosome 6 is 81.6 Mb long and includes nine families of satellite DNA. Satellites from three of these families form up to 1 Mb clusters of centromeric chromatin marked by CENP-A. Except for the enrichment with satellite DNA, the

Species with metapolycentromeres

Species	Reference	Species	Reference
Flowering plants		Flowering plants	
*Allium cepa**	Fiskesjö et al., 1981	*Strophanthus divaricates**	Beentje, 1982
*Allium erdelii**	Kollmann, 1970	*Strophanthus sarmentosus**	Beentje, 1982
*Allium neapolitanum**	Badr, Elkington, 1977	Gymnosperm plants	
*Allium qasyunense**	Kollmann, 1970	*Cryptomeria japonica**	Schlarbaum, Tsuchiya, 1984b
*Allium sativum**	Panda et al., 1979	*Cunninghamia lanceolata**	Schlarbaum, Tsuchiya, 1984a
*Allium subhirsutum**	Badr, Elkington, 1977	*Metasequoia glyptostroboides**	Schlarbaum, Tsuchiya, 1984b
*Allium trifoliatus**	Miceli et al., 1984	*Phyllocladus trichomanoides**	Davies et al., 1997
*Allium trioliatum**	Badr, Elkington, 1977	*Sequoiadendron giganteum**	Schlarbaum, Tsuchiya, 1984b
*Arachis viliosa**	Stalker, Dalmacio, 1981	*Taiwania cryptomerioides**	Schlarbaum, Tsuchiya, 1984a
*Chamaelirium luteum**	Tanaka, 2020	*Taxodium distiehum**	Schlarbaum, Tsuchiya, 1984b
*Colchicum ritchii**	Feinbrun, 1958	*Tsuga longibracteata**	Li, 1991
*Colchicum schimperi**	Feinbrun, 1958	Insects	
*Dioscorea deltoidea**	Bhat, Bindroo, 1980	*Mycetomoellerius urichii*	Teixeira et al., 2022
*Filipendula ulmaria**	Baker H.G., Baker I., 1967	*Solenopsis geminate*	Huang Y.-C. et al., 2016
*Filipendula vulgaris**	Baker H.G., Baker I., 1967	*Solenopsis invicta*	Huang Y.-C. et al., 2016
Lathyrus clymenum	Neumann et al., 2015	*Trachymyrmex holmgreni*	Cardoso et al., 2018
Lathyrus latifolius	Neumann et al., 2015	*Tribolium castaneum*	Gržan et al., 2020
Lathyrus niger	Neumann et al., 2015	Vertebrates	
Lathyrus ochrus	Neumann et al., 2015	*Chloebia gouldiae*	Malinovskaya et al., 2022
Lathyrus sativus	Neumann et al., 2015	*Ficedula hypoleuca*	Malinovskaya et al., 2022
Lathyrus sylvestris	Neumann et al., 2015	*Linaria cannabina*	Grishko et al., 2023
Lathyrus vernus	Neumann et al., 2015	*Mesoplodon carlhubbsi**	Kurihara et al., 2017
Pisum fulvum	Neumann et al., 2015	*Muntiacus muntjak*	Comings, Okada, 1971
Pisum sativum	Neumann et al., 2012	*Pyrrhula pyrrhula*	Grishko et al., 2023
*Rutidosis leiolepis**	Young et al., 2002	*Serinus canaria*	Malinovskaya et al., 2022

Note. The newly mined species with potential metapolycentromeres are indicated by asterisks.

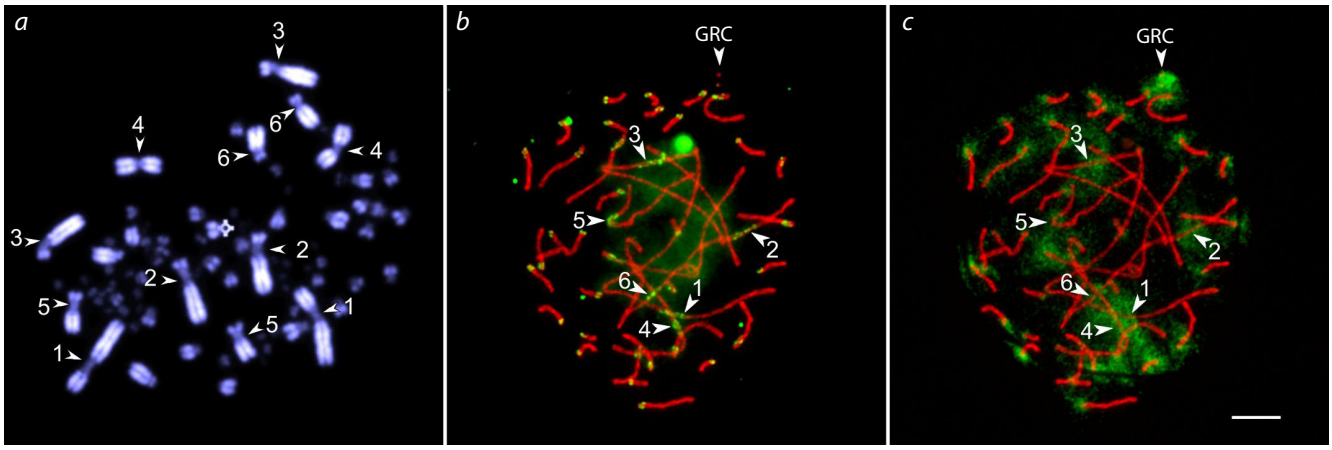


Fig. 3. Mitotic metaphase (*a*) and synaptonemal complexes (*b*, *c*) of the male domestic canary after DAPI staining (*a*) and immunostaining using antibodies against SYCP3, the main protein of the lateral elements of the synaptonemal complex (red), human centromere proteins (green) (*b*) and SYCP3 (red), and H3K9me2/3, histone H3, di- and trimethylated at lysine 9 (green) (*c*).

Numbers indicate macrochromosomes with metapolycentromeres. Arrows indicate extended primary constrictions (*a*) and metapolycentromeres (*b*, *c*). GRC indicates germline restricted chromosome. Bar 5 μm. After L.P. Malinovskaya et al. (2022), modified with permission.

metapolycentromere does not differ from the adjacent regions of the chromosome in DNA methylation patterns, the location of transcriptionally active genes, and retrotransposons (Macas et al., 2023).

E. Grishko et al. (2023) and L.P. Malinovskaya et al. (2022) demonstrated that the metapolycentromeres of the songbirds do not differ from their regional centromeres in the H3K9 methylation patterns (Fig. 3*c*). For example, all macrochromosomes of the domestic canary contain metapolycentromeres, and all of them except the Z chromosome are hypermethylated at H3K9, as well as the regional centromeres of all macrochromosomes except the Z chromosome in several other songbird species studied.

Meanwhile, P. Neumann et al. (2016) revealed a striking similarity between metapolycentromeres and holocentromeres in the patterns of histone modifications H3S10ph, H3S28ph, and H3T3ph distributions in *L. sativus* and *P. sativum* chromosomes. The metapolycentromeres showed a unique pattern of H2AT120ph distribution, significantly different from that of both regional and holocentromeres. The genomes of *Pisum* and *Lathyrus* contain two variants of the CENP-A gene, named CenH3-1 and CenH3-2 (Neumann et al., 2012), the sequences of which show 55 % homology, while corresponding proteins differ in length and amino acid sequence and show 72 % homology (Neumann et al., 2012). Both forms of CENP-A are localized on functional chromatin clusters of metapolycentromeres in these species (Neumann et al., 2015).

Simultaneous immunodetection of CENP-A and tubulin in *P. sativum* revealed colocalization of these proteins in the centromeric region, indicating that each cluster of centromeric chromatin within the metapolycentromere forms a functional kinetochore (Neumann et al., 2012). Regional and metapolycentromeres do not differ in the strength of their suppressive effect on meiotic recombination in the pericentromeric chromosome regions (Grishko et al., 2023).

Thus, metapolycentromeres may vary in the number of centromeric domains and in their genetic content and epigenetic modifications. However, these variations do not seem to affect their function.

Origin of metapolycentromeres

At the moment, several mechanisms for the formation of metapolycentromeres were suggested: multiple Robertsonian translocations in the Indian muntjac (Huang L. et al., 2006), segmental duplications in legumes (Macas et al., 2023), epigenetic charges in the interspecies marsupial hybrids (O'Neill et al., 1998) and expansion of centromeric chromatin and overexpression of the CENP-A protein in the malignant neoplasms (Sullivan L.L. et al., 2011, 2016; Perpelescu et al., 2015).

In the Indian muntjac (*M. muntjak vaginalis*), the metapolycentromere is located on the X chromosome (Drpic et al., 2018). This species has the smallest number of chromosomes among mammals: $2n = 6$ in females and $2n = 7$ in males (Wurster, Benirschke, 1970). The reduction of the chromosome was determined by a fusion of chromosomes in an ancestor with a karyotype of $2n = 70$ (Yang et al., 1997; Chi et al., 2005). The elongated centromere of the X chromosome was suggested to result from several successive Robertsonian translocations (Chang et al., 2001; Huang L. et al., 2006). However, it remains unclear why all autosomes of this species, which also resulted from multiple Robertsonian translocations, have the standard regional centromeres.

In legumes, metapolycentromeres may have arisen through a mechanism associated with the duplication of the centromeric histone H3 gene (Neumann et al., 2015). However, the presence of two CENP-A variants is not a determinant of the presence of metapolycentromeres in *Pisum* and *Lathyrus*. Several plant species have two CENP-A variants but no metapolycentromeres, for example, *A. lyrata* and *Mimulus* spp. (Kawabe et al., 2006; Finseth et al., 2015). Thus, in peas, sequencing of long reads combined with ChIP-seq with antibodies to CENP-A showed the emergence of the newest domain of centromeric chromatin through segmental duplication and subsequent inversion of an existing domain 5.2 Mb long. However, the origin of the remaining domains of centromeric chromatin is unclear (Macas et al., 2023).

Multiple tandem duplications play a major role in the homogenization of centromeric repeat monomers in rice (Ma, Jackson, 2006). They might result from unequal crossing over, gene conversion, duplicate transposition, satellite transposition, and illegitimate recombination (Copenhaver et al., 1999; Ma, Jackson, 2006).

Typically, dicentric and polycentric chromosomes cannot ensure the attachment of unipolar spindle microtubules to their chromatids, which causes chromosome breakage and nondisjunction. Thus, there are mechanisms that select against such a chromosome structure (for example, the elimination of one of the centromeres) (Zhang et al., 2010). However, this does not occur in the case of metapolycentromeres due to the close proximity of the centromeric domains (Neumann et al., 2012). It is known that the distance between two functional centromeres should not exceed 20 Mbp for them to function as one centromere during cell division (Higgins et al., 2005). Apparently, this condition is also satisfied for metapolycentromere domains.

Metapolycentromeres can arise *de novo* from regional centromeres under conditions of genomic instability. Such destabilizing conditions may include interspecific hybridization and malignant neoplasms (Metcalfe et al., 2007; Sullivan L.L. et al., 2011).

The elongated centromeres have been observed in some chromosomes of interspecific hybrids of several marsupial species (kangaroos and wallabies), while the chromosomes of the parental species contained regional centromeres. Interestingly, the elongated centromeres were only present on the maternally derived chromosomes (O'Neill et al., 1998, 2001; Metcalfe et al., 2007; Schroeder-Reiter, Wanner, 2009). This phenomenon was observed in hybrids between the closely related species *Macropus rufogriseus* and *M. agilis*, as well as in those between the phylogenetically distant species *M. eugenii* and *Wallabia bicolor* (O'Neill et al., 1998; Metcalfe et al., 2007). In all these hybrids, the expansion of centromeric chromatin occurred due to an uncontrolled increase in the number of copies of centromeric retrotransposons, and for different hybrids, the families of retrotransposons that facilitated the expansion differed (O'Neill et al., 1998; Metcalfe et al., 2007). Apparently, the changes in the epigenetic context due to hybridization disrupt DNA methylation patterns that normally restrain the activity of centromeric retrotransposons. This, in turn, leads to their repeated copying and the expansion of

the centromeric region (O'Neill et al., 1998). However, it is still not clear why this phenomenon is limited to maternally derived chromosomes.

Expansion of centromeric chromatin also occurs in some human cancer cells (Sullivan L.L. et al., 2011, 2016; Perpelescu et al., 2015). Thus, in cell line GM08148, a rearrangement on chromosome 17 resulted in the centromere entering the euchromatic environment; as a result, CENP-A spread into the short arm and formed an elongated functional centromere on a non-centromeric DNA sequence (Sullivan L.L. et al., 2016). Additionally, overexpression of the CENP-A protein and its chaperone HJURP, along with the disruption of the interaction of the tumor suppressor protein Rb with chromatin in cancer cells, can lead to centromere elongation (Sullivan L.L. et al., 2011; Perpelescu et al., 2015). Altered epigenetic landscapes and uncontrolled proliferation of centromeric sequences may trigger dysregulated expansion of centromeric chromatin.

Metapolycentromere evolution and the centromere drive hypothesis

The conservative centromere function – the attachment of spindle microtubules and subsequent chromosomal segregation – implies strict purifying selection on the components of the centromere: centromere DNA and centromere proteins. However, in reality, we observe a completely opposite picture – both centromeric DNA and centromeric proteins evolve rapidly and often differ significantly even between closely related species. This contradiction is called the "centromere paradox" (Henikoff et al., 2001).

To resolve the centromere paradox, S. Henikoff et al. (2001) suggested the centromere drive hypothesis. This hypothesis suggests that in asymmetric female meiosis, the centromeres segregating in the egg rather than in the polar body ("the strong centromeres") would be favored. However, male meiosis is symmetric. In this case, inequality in centromere strength might lead to chromosome nondisjunction and spermatogenic arrest (Malik, Henikoff, 2001). The resulting conflict might be resolved by a selection for centromeric proteins, which are able to equalize the centromeres and compensate for the fitness costs (Fig. 4). This perpetual tug-of-war between male and female meiosis should result in the rapid evolution of centromeric sequences and proteins (Dawe, Henikoff, 2006).

Selection for "stronger centromeres" in female meiosis might favor variants of the centromeric DNA sequences with enhanced potential to recruit centromeric proteins (in particular CENP-A) and form kinetochores to which more microtubules are attached. This effect may also be enhanced by a selection for an increase in the copy number of such sequences. These processes could cause the occurrence of metapolycentromeres. The suppression of centromeric drive in male meiosis may limit centromere size. Probably, this is why metapolycentromeres are so rare. They have been found in several ant species (Huang Y.-C. et al., 2016; Cardoso et al., 2018) with haploid males. For this reason, there should be no selection for suppression of centromeric drive in male meiosis. This makes Hymenoptera a promising group for the search for new metapolycentromeres.

Thus, the centromere drive hypothesis provides a plausible explanation for the dynamic evolution of centromeres in general and the emergence of metapolycentromeres in particular.

Do metapolycentromeres represent an intermediate stage of evolution between regional centromeres and holocentromeres?

P. Neumann et al. (2012) suggested that metapolycentromeres might represent an intermediate stage of evolution between regional centromeres and holocentromeres. According to this hypothesis, the satellite DNA sequences of the regional centromere, under the influence of centromere drive, might expand so much that they capture the entire chromosome, rendering it holocentric.

During evolution, holocentromeres arose from regional centromeres at least 13 times: four times in plants and nine times in animals (Melters et al., 2012). Despite the common morphological feature (i. e. the absence of a primary constriction for the attachment of spindle filaments), holocentric chromosomes differ from each other in their origin and structure (Melters et al., 2012; Senaratne et al., 2022). Holocentromere centromeric units (chromosomal regions marked with CENP-A) can be based on either satellite or non-repeated

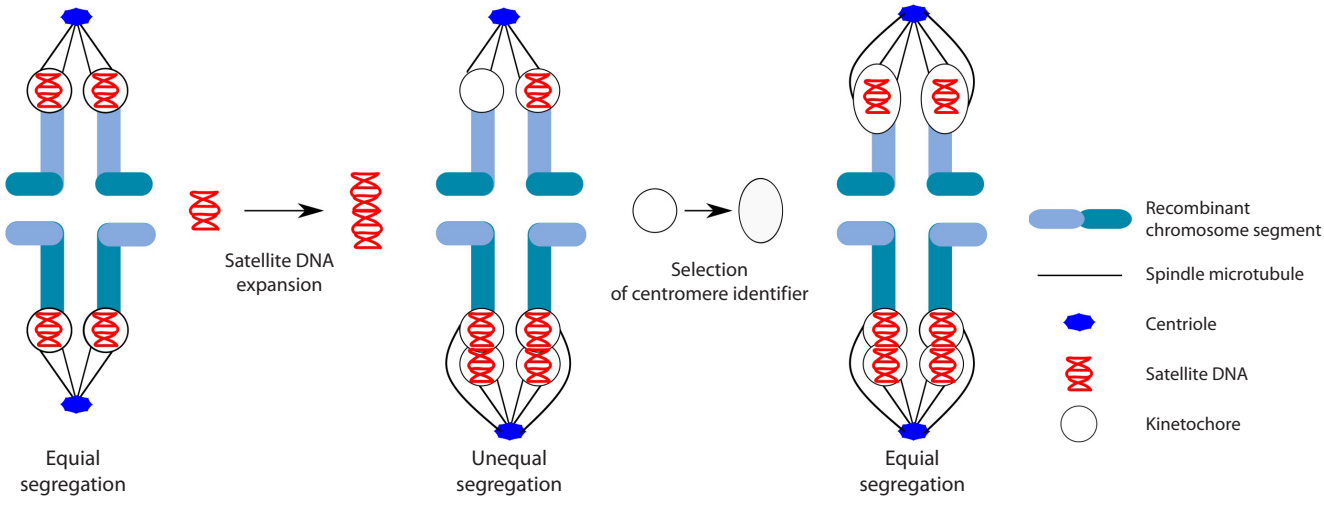


Fig. 4. The model of centromere drive according to S. Henikoff et al. (2001), modified.

DNA sequences (Gassmann et al., 2012; Marques et al., 2015). In turn, satellite holocentromeres are divided into holocentromeres with a large number of small centromeric units and holocentromeres with a small number of large centromeric units (Kuo et al., 2024). Large centromeric units comparable in size to regional centromeres have been discovered in the plants *Chionographis japonica* and *Morus notabilis* (Kuo et al., 2023; Ma et al., 2023). It was suggested that holocentromeres in *C. japonica* formed through multiple misrepaired DNA double-strand breaks associated with the insertion of extra-chromosomal circular DNA (Kuo et al., 2024). These insertions of regions of centromeric chromatin might not occur simultaneously throughout the genome, but evolve from metapolycentromeres.

The genera *Juncus*, *Drosera* and *Cuscuta* include both species with holocentromeres and species with regional centromeres (Pazy, Plitmann, 1994; Shirakawa et al., 2011a, b; Guerra et al., 2019; Neumann et al., 2021; Mata-Sucre et al., 2023). Recently, using ChIP-seq with anti-CENP-A antibodies, it was found that the chromosomes of *J. effusus* bear both regional centromeres and polycentromeres with multiple CENP-A domains (Dias et al., 2024). Such centromeres are similar in structure to metapolycentromeres, but they do not form elongated primary constrictions due to the small number of centromeric domains and their close proximity to each other. The presence of holocentromere and regional centromere species in the genus *Juncus* led to the suggestion that this species represents a transitional form from regional centromeres to holocentromeres. However, not a single "transitive karyotype" containing both metapolycentric and holocentric chromosomes has been discovered.

Even if this hypothesis holds true, it would only explain the origin of holocentricity in a small number of species with holocentric chromosomes, because most holocentric chromosomes do not possess centromere-specific DNA sequences (Talbert, Henikoff, 2020; Senaratne et al., 2021, 2022).

Backward and forward search for metapolycentromeres

We suspect metapolycentromeres are more common than believed. However, finding them is problematic. They can be reliably revealed by immunostaining chromosomes with antibodies to CENP-A or by ChIP-seq with anti-CENP-A antibodies. Metapolycentromeres may also be indirectly detected by the analysis of the copy number of centromeric repeats, by immunostaining for kinetochore proteins, and, in the case of particularly large metapolycentromeres, by routine chromosome staining, which reveals them as elongated primary constrictions. However, indirect methods do not reveal the actual number of functional domains of centromeric chromatin.

The term metapolycentromere was suggested by P. Neumann et al. (2012), and before that date, elongated primary constrictions were not termed metapolycentromeres and often were not mentioned at all. In the backward search for potential metapolycentromeres, we carried out data mining for the cytogenetic articles in the scholar.google.com database (last access: 7th of July 2023) using 18 keywords (Supplementary Material)[1]. We selected all articles written in English that mentioned long primary constrictions in the text or showed them in the micrographs. Table shows the list of already known and newly mined candidate species with metapolycentromeres.

It spans 27 species of flowering and eight species of gymnosperm plants, five species of insects and seven species of vertebrates. It indicates an erratic phylogenetic distribution of the species with metapolycentromeres. This, in turn, may suggest independent evolutionary occurrences of metapolycentromeres. However, the current catalog of species with identified and suspected metapolycentromeres remains too short to draw reliable conclusions about their evolution, particularly in the absence of knowledge about related species without metapolycentromeres for comparative analysis. More studies are necessary to shed light on the mechanisms of metapolycentromere formation and evolution.

[1] Supplementary Material is available at: https://vavilovj-icg.ru/download/pict-2024-28/appx21.pdf

Conclusion

The systematic study of new species with and without metapolycentromeres is important for understanding their evolution. Species with karyotypes containing both regional centromeres and metapolycentromeres are especially interesting. A comparison between the centromeric DNA of metapolycentromeres and regional centromeres may shed light on the mechanisms of metapolycentromere formation.

References

Ahmad S.F., Singchat W., Jehangir M., Suntronpong A., Panthum T., Malaivijitnond S., Srikulnath K. Dark matter of primate genomes: Satellite DNA repeats and their evolutionary dynamics. *Cells*. 2020; 9(12):2714. DOI 10.3390/cells9122714

Badr A., Elkington T.T. Variation of Giemsa C-band and fluorochrome banded karyotypes, and relationships in *Allium* subgen. *Molium*. *Pl. Syst. Evol.* 1977;128(1-2):23-35. DOI 10.1007/BF00985168

Baker H.G., Baker I. The cytotaxonomy of *Filipendula* (Rosaceae) and its implications. *Am. J. Bot.* 1967;54(8):1027-1034. DOI 10.1002/j.1537-2197.1967.tb10729.x

Balzano E., Giunta S. Centromeres under pressure: Evolutionary innovation in conflict with conserved function. *Genes* (*Basel*). 2020; 11(8):912. DOI 10.3390/genes11080912

Beentje H.J. A Monograph on Strophanthus DC. (Apocynaceae). Wageningen, 1982

Bhat B.K., Bindroo B.B. Sex chromosomes in *Dioscorea deltoidea* Wall. *Cytologia* (*Tokyo*). 1980;45(4):739-742. DOI 10.1508/cytologia.45.739

Black B.E., Foltz D.R., Chakravarthy S., Luger K., Woods V.L., Cleveland D.W. Structural determinants for generating centromeric chromatin. *Nature*. 2004;430(6999):578-582. DOI 10.1038/nature02766

Cardoso D.C., Heinze J., Moura M.N., Cristiano M.P. Chromosomal variation among populations of a fungus-farming ant: implications for karyotype evolution and potential restriction to gene flow. *BMC Evol. Biol.* 2018;18(1):146. DOI 10.1186/s12862-018-1247-5

Castellani M., Zhang M., Thangavel G., Mata-Sucre Y., Lux T., Campoy J.A., Marek M., Huettel B., Sun H., Mayer K.F.X., Schneeberger K., Marques A. Meiotic recombination dynamics in plants with repeat-based holocentromeres shed light on the primary drivers of crossover patterning. *Nat. Plants*. 2024;10:423-438. DOI 10.1038/s41477-024-01625-y

Chang S.D., Chao A.S., Lai Y.M., Liu H.Y., Soong Y.K. Interphase FISH-assisted second-trimester termination of a trisomy 21 fetus in an IVF-ET twin pregnancy. A case report. *J. Reprod. Med.* 2001; 46(12):1063-1066

Chi J.X., Huang L., Nie W., Wang J., Su B., Yang F. Defining the orientation of the tandem fusions that occurred during the evolution of Indian muntjac chromosomes by BAC mapping. *Chromosoma*. 2005;114(3):167-172. DOI 10.1007/s00412-005-0004-x

Cleveland D.W., Mao Y., Sullivan K.F. Centromeres and kinetochores. *Cell*. 2003;112(4):407-421. DOI 10.1016/S0092-8674(03)00115-6

Comings D.E., Okada T.A. Fine structure of kinetochore in Indian muntjac. *Exp. Cell Res.* 1971;67(1):97-110. DOI 10.1016/0014-4827(71)90625-2

Copenhaver G.P., Nickel K., Kuromori T., Benito M.I., Kaul S., Lin X., Bevan M., Murphy G., Harris B., Parnell L.D., McCombie W.R., Martienssen R.A., Marra M., Preuss D. Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science.* 1999; 286(5449):2468-2474. DOI 10.1126/science.286.5449.2468

Davies B.J., O'Brien I.E.W., Murray B.G. Karyotypes, chromosome bands and genome size variation in New Zealand endemic gymnosperms. *Plant Syst. Evol.* 1997;208(3-4):169-185. DOI 10.1007/BF00985440

Dawe R.K., Henikoff S. Centromeres put epigenetics in the driver's seat. *Trends Biochem. Sci.* 2006;31(12):662-669. DOI 10.1016/j.tibs.2006.10.004

Dias Y., Mata-Sucre Y., Thangavel G., Costa L., Baez M., Houben A., Marques A., Pedrosa-Harand A. How diverse a monocentric chromosome can be? Repeatome and centromeric organization of *Juncus effusus* (Juncaceae). *Plant J*. 2024;118(6):1832-1847. DOI 10.1111/tpj.16712

Drinnenberg I.A., deYoung D., Henikoff S., Malik H.S. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *eLife*. 2014;3:e03676. DOI 10.7554/eLife.03676

Drpic D., Almeida A.C., Aguiar P., Renda F., Damas J., Lewin H.A., Larkin D.M., Khodjakov A., Maiato H. Chromosome segregation is biased by kinetochore size. *Curr. Biol.* 2018;28(9):1344-1356. DOI 10.1016/j.cub.2018.03.023

Feinbrun N. Chromosome numbers and evolution in the genus *Colchicum*. *Evolution* (*N.Y.*). 1958;12(2):173. DOI 10.2307/2406028

Finseth F.R., Dong Y., Saunders A., Fishman L. Duplication and adaptive evolution of a key centromeric protein in *Mimulus*, a genus with female meiotic drive. *Mol. Biol. Evol.* 2015;32(10):2694-2706. DOI 10.1093/molbev/msv145

Fiskesjö G., Lassen C., Renberg L. Chlorinated phenoxyacetic acids and chlorophenols in the modified *Allium* test. *Chem. Biol. Interact.* 1981;34(3):333-344. DOI 10.1016/0009-2797(81)90105-8

Flemming W. Zellsubstanz, Kern und Zelltheilung. Leipzig: F.C.W. Vogel, 1882. DOI 10.5962/bhl.title.168645

Furuyama S., Biggins S. Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proc. Natl. Acad. Sci. USA.* 2007;104(37):14706-14711. DOI 10.1073/pnas.0706985104

Gassmann R., Rechtsteiner A., Yuen K., Muroyama A., Egelhofer T., Gaydos L., Barron F., Maddox P., Essex A., Monen J., Ercan S., Lieb J.D., Oegema K., Strome S., Desai A. An inverse relationship to germline transcription defines centromeric chromatin in *C. elegans*. *Nature*. 2012;484:534-537. DOI 10.1038/nature10973

Glöckner G., Heidel A.J. Centromere sequence and dynamics in *Dictyostelium discoideum. Nucleic Acids Res.* 2009;37(6):1809-1816. DOI 10.1093/nar/gkp017

Grishko E., Malinovskaya L., Slobodchikova A., Kotelnikov A., Torgasheva A., Borodin P. Cytological analysis of crossover frequency and distribution in male meiosis of cardueline finches (Fringillidae, Aves). *Animals*. 2023;13(23):3624. DOI 10.3390/ani13233624

Gržan T., Despot-Slade E., Meštrović N., Plohl M., Mravinac B. CenH3 distribution reveals extended centromeres in the model beetle *Tribolium castaneum*. *PLoS Genet.* 2020;16(10):e1009115. DOI 10.1371/journal.pgen.1009115

Guerra M., Ribeiro T., Felix L.P. Monocentric chromosomes in *Juncus* (Juncaceae) and implications for the chromosome evolution of the family. *Bot. J. Linn. Soc*. 2019;191(4):475-483. DOI 10.1093/botlinnean/boz065

Haupt W., Fischer T.C., Winderl S., Fransz P., Torres-Ruiz R.A. The CENTROMERE1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin. *Plant J.* 2001;27(4):285-296. DOI 10.1046/j.1365-313x.2001.01087.x

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

Henikoff S., Ramachandran S., Krassovsky K., Bryson T.D., Codomo C.A., Brogaard K., Widom J., Wang J.-P., Henikoff J.G. The budding yeast centromere DNA element II wraps a stable Cse4 hemisome in either orientation *in vivo*. *eLife*. 2014;3:e01861. DOI 10.7554/eLife.01861

Higgins A.W., Gustashaw K.M., Willard H.F. Engineered human dicentric chromosomes show centromere plasticity. *Chromosom. Res.* 2005;13(8):745-762. DOI 10.1007/s10577-005-1009-2

Huang L., Chi J., Nie W., Wang J., Yang F. Phylogenomics of several deer species revealed by comparative chromosome painting with Chinese muntjac paints. *Genetica*. 2006;127(1-3):25-33. DOI 10.1007/s10709-005-2449-5

Huang Y.-C., Lee C.-C., Kao C.-Y., Chang N.-C., Lin C.-C., Shoemaker D., Wang J. Evolution of long centromeres in fire ants. *BMC Evol. Biol.* 2016;16(1):189. DOI 10.1186/s12862-016-0760-7

Kanesaki Y., Imamura S., Matsuzaki M., Tanaka K. Identification of centromere regions in chromosomes of a unicellular red alga, *Cyanidioschyzon merolae*. *FEBS Lett.* 2015;589(11):1219-1224. DOI 10.1016/j.febslet.2015.04.009

Kawabe A., Nasuda S., Charlesworth D. Duplication of centromeric histone H3 (*HTR12*) gene in *Arabidopsis halleri* and *A. lyrata*, plant species with multiple centromeric satellite sequences. *Genetics*. 2006;174(4):2021-2032. DOI 10.1534/genetics.106.063628

Kollmann F. Karyotypes of three *Allium* species of the *erdelii* group. *Caryologia*. 1970;23(4):647-655. DOI 10.1080/00087114.1970.10796400

Kuo Y.T., Câmara A.S., Schubert V., Neumann P., Macas J., Melzer M., Chen J., Fuchs J., Abel S., Klocke E., Huettel B., Himmelbach A., Demidov D., Dunemann F., Mascher M., Ishii T., Marques A., Houben A. Holocentromeres can consist of merely a few megabase-sized satellite arrays. *Nat. Commun.* 2023;14:3502. DOI 10.1038/s41467-023-38922-7

Kuo Y.T., Schubert V., Marques A., Schubert I., Houben A. Centromere diversity: How different repeat-based holocentromeres may have evolved. *BioEssays*. 2024;46(6):2400013. DOI 10.1002/bies.202400013

Kurihara N., Tajima Y., Yamada T.K., Matsuda A., Matsuishi T. Description of the karyotypes of stejneger's beaked whale (*Mesoplodon stejnegeri*) and hubbs' beaked whale (*M. carlhubbsi*). *Genet. Mol. Biol.* 2017;40(4):803-807. DOI 10.1590/1678-4685-gmb-2016-0284

Lee H.R., Zhang W., Langdon T., Jin W., Yan H., Cheng Z., Jiang J. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proc. Natl. Acad. Sci. USA.* 2005;102(33):11793-11798. DOI 10.1073/pnas.0503863102

Li L.C. The karyotype analysis of *Tsuga longibracteata* and its taxonomic significance. *Acta Bot. Yunnan*. 1991;13(3):309-313

Ma B., Wang H., Liu J., Chen L., Xia X., Wei W., Yang Z., Yuan J., Luo Y., He N. The gap-free genome of mulberry elucidates the architecture and evolution of polycentric chromosomes. *Hortic. Res.* 2023;10(7):uhad111. DOI 10.1093/hr/uhad111

Ma J., Jackson S.A. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* 2006;16(2):251-259. DOI 10.1101/gr.4583106

Macas J., Ávila Robledillo L., Kreplak J., Novák P., Koblížková A., Vrbová I., Burstin J., Neumann P. Assembly of the 81.6 Mb centromere of pea chromosome 6 elucidates the structure and evolution of metapolycentric chromosomes. *PLoS Genet.* 2023;19(2):e1010633. DOI 10.1371/journal.pgen.1010633

Maheshwari S., Tan E.H., West A., Franklin F.C.H., Comai L., Chan S.W.L. Naturally occurring differences in CENH3 affect chromosome segregation in zygotic mitosis of hybrids. *PLoS Genet.* 2015;11(1):e1004970. DOI 10.1371/journal.pgen.1004970

Malik H.S., Henikoff S. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. *Genetics*. 2001;157(3):1293-1298. DOI 10.1093/genetics/157.3.1293

Malinovskaya L.P., Slobodchikova A.Y., Grishko E.O., Pristyazhnyuk I.E., Torgasheva A.A., Borodin P.M. Germline-restricted chromosomes and autosomal variants revealed by pachytene karyotyping of 17 avian species. *Cytogenet. Genome Res.* 2022;162(3):148-160. DOI 10.1159/000524681

Marques A., Ribeiro T., Neumann P., Macas J., Novák P., Schubert V., Pellino M., Fuchs J., Ma W., Kuhlmann M., Brandt R., Vanzela A.L.L., Beseda T., Šimková H., Pedrosa-Harand A., Houben A. Holocentromeres in Rhynchospora are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. *Proc. Natl. Acad. Sci. USA*. 2015;112(44):13633-13638. DOI 10.1073/pnas.1512255112

Mata-Sucre Y., Matzenauer W., Castro N., Huettel B., Pedrosa-Harand A., Marques A., Souza G. Repeat-based phylogenomics shed light on unclear relationships in the monocentric genus *Juncus* L. (Juncaceae). *Mol. Phylogenet. Evol*. 2023;189:107930. DOI 10.1016/j.ympev.2023.107930

Melters D.P., Paliulis L.V., Korf I.F., Chan S.W.L. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. *Chromosom. Res.* 2012;20(5):579-593. DOI 10.1007/s10577-012-9292-1

Melters D.P., Bradnam K.R., Young H.A., Telis N., May M.R., Ruby J., Sebra R., Peluso P., Eid J., Rank D., Garcia J., DeRisi J.L., Smith T., Tobias C., Ross-Ibarra J., Korf I., Chan S.W. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* 2013;14(1):R10. DOI 10.1186/gb-2013-14-1-r10

Mendiburo M.J., Padeken J., Fülöp S., Schepers A., Heun P. Drosophila CENH3 is sufficient for centromere formation. *Science.* 2011;334(6056):686-690. DOI 10.1126/science.1206880

Metcalfe C.J., Bulazel K.V., Ferreri G.C., Schroeder-Reiter E., Wanner G., Rens W., Obergfell C., Eldridge M.D.B., O'Neill R.J. Genomic instability within centromeres of interspecific marsupial hybrids. *Genetics*. 2007;177(4):2507-2517. DOI 10.1534/genetics.107.082313

Miceli P., Ficini G., Garbari F. The genus «Allium» L. in Italy. XIII. Morphological, caryological and leaf anatomical study in some C-W Mediterranean triploid populations of «Allium trifoliatum» Cyr. *Webbia*. 1984;38(1):793-803. DOI 10.1080/00837792.1984.10670350

Nagpal H., Fierz B. The elusive structure of centro-chromatin: Molecular order or dynamic heterogenetity. *J. Mol. Biol.* 2021;433(6):166676. DOI 10.1016/j.jmb.2020.10.010

Navarro-Mendoza M.I., Pérez-Arques C., Panchal S., Nicolás F.E., Mondo S.J., Ganguly P., Pangilinan J., Grigoriev I.V., Heitman J., Sanyal K., Garre V. Early diverging fungus *Mucor circinelloides* lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. *Curr. Biol.* 2019;29(22):3791-3802. DOI 10.1016/j.cub.2019.09.024

Neumann P., Navrátilová A., Schroeder-Reiter E., Koblížková A., Steinbauerová V., Chocholová E., Novák P., Wanner G., Macas J. Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet.* 2012;8(6):e1002777. DOI 10.1371/journal.pgen.1002777

Neumann P., Pavlíková Z., Koblížková A., Fuková I., Jedličková V., Novák P., Macas J. Centromeres off the hook: Massive changes in centromere size and structure following duplication of *Cenh3* gene in *Fabeae* species. *Mol. Biol. Evol.* 2015;32(7):1862-1879. DOI 10.1093/molbev/msv070

Neumann P., Schubert V., Fuková I., Manning J.E., Houben A., Macas J. Epigenetic histone marks of extended meta-polycentric centromeres of *Lathyrus* and *Pisum* chromosomes. *Front. Plant Sci.* 2016;7(MAR2016):234. DOI 10.3389/fpls.2016.00234

Neumann P., Oliveira L., Čížková J., Jang T.S., Klemme S., Novák P., Stelmach K., Koblížková A., Doležel J., Macas J. Impact of parasitic lifestyle and different types of centromere organization on chromosome and genome evolution in the plant genus *Cuscuta*. *New Phytologist.* 2021;229(4):2365-2377. DOI 10.1111/nph.17003

O'Neill R.J.W., O'Neill M.J., Marshall Graves J.A. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*. 1998;393(6680):68-72. DOI 10.1038/29985

O'Neill R.J.W., Eldridge M.D.B., Graves J.A.M. Chromosome heterozygosity and *de novo* chromosome rearrangements in mammalian interspecies hybrids. *Mamm. Genome*. 2001;12(3):256-259. DOI 10.1007/s003350010270

Panda B.B., Sahu R.K., Sharma C.B.S.R. Cytogenetic hazards from agricultural chemicals. 2. Selective clastogenesis and spindle inhibition in some plant mitotic systems by the β-exotoxin and the general ineffectiveness of the δ-endotoxin protein of *Bacillus thuringiensis*. *Mutat. Res. Toxicol.* 1979;67(2):161-166. DOI 10.1016/0165-1218(79)90127-7

Pazy B., Plitmann U. Holocentric chromosome behaviour in *Cuscuta* (*Cuscutaceae*). *Plant Syst. Evol*. 1994;191:105-109. DOI 10.1007/BF00985345

Perpelescu M., Hori T., Toyoda A., Misu S., Monma N., Ikeo K., Obuse C., Fujiyama A., Fukagawa T. HJURP is involved in the expansion of centromeric chromatin. *Mol. Biol. Cell*. 2015;26(15):2742-2754. DOI 10.1091/mbc.E15-02-0094

Schlarbaum S.E., Tsuchiya T. The chromosomes of *Cunninghamia konishii*, *C. lanceolata*, and *Taiwania cryptomerioides* (*Taxodiaceae*). *Plant Syst. Evol.* 1984a;145(3-4):169-181. DOI 10.1007/BF00983946

Schlarbaum S.E., Tsuchiya T. Cytotaxonomy and phylogeny in certain species of *Taxodiaceae*. *Plant Syst. Evol.* 1984b;147(1-2):29-54. DOI 10.1007/BF00984578

Schroeder-Reiter E., Wanner G. Chromosome centromeres: Structural and analytical investigations with high resolution scanning electron microscopy in combination with focused ion beam milling. *Cytogenet. Genome Res.* 2009;24(3-4):239-250. DOI 10.1159/000218129

Senaratne A.P., Muller H., Fryer K.A., Kawamoto M., Katsuma S., Drinnenberg I.A. Formation of the CenH3-deficient holocentromere in Lepidoptera avoids active chromatin. *Curr. Biol.* 2021;31(1):173-181.e7. DOI 10.1016/j.cub.2020.09.078

Senaratne A.P., Cortes-Silva N., Drinnenberg I.A. Evolution of holocentric chromosomes: Drivers, diversity, and deterrents. *Semin. Cell Dev. Biol.* 2022;127:90-99. DOI 10.1016/j.semcdb.2022.01.003

Shirakawa J., Hoshi Y., Kondo K. Chromosome differentiation and genome organization in carnivorous plant family Droseraceae. *Chromosome Bot*. 2011a;6(4);111-119. DOI 10.3199/iscb.6.111

Shirakawa J, Katsuya N, Yoshikazu H. A chromosome study of two centromere differentiating *Drosera* species, *D. arcturi* and *D. regia*. *Caryologia*. 2011b;64(4):453-463. DOI 10.1080/00087114.2011.10589813

Stalker H.T., Dalmacio R.D. Chromosomes of *Arachis* species, section *Arachis*. *J. Hered.* 1981;72(6):403-408. DOI 10.1093/oxfordjournals.jhered.a109541

Sullivan K.F., Hechenberger M., Masri K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J. Cell Biol.* 1994;127(3):581-592. DOI 10.1083/jcb.127.3.581

Sullivan L.L., Boivin C.D., Mravinac B., Song I.Y., Sullivan B.A. Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosom. Res.* 2011. DOI 10.1007/s10577-011-9208-5

Sullivan L.L., Maloney K.A., Towers A.J., Gregory S.G., Sullivan B.A. Human centromere repositioning within euchromatin after partial chromosome deletion. *Chromosom. Res.* 2016. DOI 10.1007/s10577-016-9536-6

Tachiwana H., Kagawa W., Kurumizaka H. Comparison between the CENP-A and histone H3 structures in nucleosomes. *Nucleus*. 2012;3(1):6-11. DOI 10.4161/nucl.18372

Talbert P.B., Henikoff S. Transcribing centromeres: Noncoding RNAs and kinetochore assembly. *Trends Genet.* 2018;34(8):587-599. DOI 10.1016/j.tig.2018.05.001

Talbert P.B., Henikoff S. What makes a centromere? *Exp. Cell Res.* 2020;389(2):111895. DOI 10.1016/j.yexcr.2020.111895

Talbert P.B., Masuelli R., Tyagi A.P., Comai L., Henikoff S. Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell*. 2002;14(5):1053-1066. DOI 10.1105/tpc.010425

Talbert P.B., Kasinathan S., Henikoff S. Simple and complex centromeric satellites in *Drosophila* sibling species. *Genetics*. 2018; 208(3):977-990. DOI 10.1534/genetics.117.300620

Tanaka N. Chromosomal traits of *Chamaelirium luteum* (Melanthiaceae) with particular focus on the large heterochromatic centromeres. *Taiwania*. 2020;65(3):286-294. DOI 10.6165/tai.2020.65.286

Teixeira G.A., Barros L.A.C., de Aguiar H.J.A.C., Lopes D.M. Multiple heterochromatin diversification events in the genome of fungus-farming ants: insights from repetitive sequences. *Chromosoma*. 2022;131(1-2):59-75. DOI 10.1007/s00412-022-00770-7

van Hooff J.J., Tromer E., van Wijk L.M., Snel B., Kops G.J. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.* 2017;18(9):1559-1571. DOI 10.15252/embr.201744102

Wang Y., Wu L., Yuen K.W.Y. The roles of transcription, chromatin organisation and chromosomal processes in holocentromere establishment and maintenance. *Semin.Cell Dev. Biol*. 2022;127:79-89. DOI 10.1016/j.semcdb.2022.01.004

Winey M., Mamay C.L., O'Toole E.T., Mastronarde D.N., Giddings T.H., McDonald K.L., McIntosh J.R. Three-dimensional ultrastructural analysis of the *Saccharomyces cerevisiae* mitotic spindle. *J. Cell Biol.* 1995;129(6):1601-1615. DOI 10.1083/jcb.129.6.1601

Wurster D.H., Benirschke K. Indian muntjac, *Muntiacus muntjak*: a deer with a low diploid chromosome number. *Science.* 1970; 168(3937):1364-1366. DOI 10.1126/science.168.3937.1364

Yang F., O'Brien P.C.M., Wienberg J., Neitzel H., Lin C.C., Ferguson-Smith M.A. Chromosomal evolution of the Chinese muntjac (*Muntiacus reevesi*). *Chromosoma*. 1997;106(1):37-43. DOI 10.1007/s004120050222

Young A., Hill J., Murray B., Peakall R. Breeding system, genetic diversity and clonal structure in the sub-alpine forb *Rutidosis leiolepis* F. Muell. (Asteraceae). *Biol. Conserv.* 2002;106(1):71-78. DOI 10.1016/S0006-3207(01)00230-0

Zhang W., Friebe B., Gill B.S., Jiang J. Centromere inactivation and epigenetic modifications of a plant chromosome with three functional centromeres. *Chromosoma*. 2010;119(5):553-563. DOI 10.1007/s00412-010-0278-5

Conflict of interest. The authors declare no conflict of interest.
Received April 16, 2024. Revised July 15, 2024. Accepted July 18, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):602-609

DOI 10.18699/vjgb-24-67

A new leaf pubescence gene, *Hl1*th , introgressed into bread wheat from *Thinopyrum ponticum* and its phenotypic manifestation under homoeologous chromosomal substitutions

A.V. Simonov [1], E.I. Gordeeva [1], M.A. Genaev [1, 2], W. Li [1, 2], I.O. Bulatov[1, 3], T.A. Pshenichnikova [1]

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

Abstract. Blue-grain lines were created on the basis of the spring bread wheat variety Saratovskaya 29 (S29) with chromosome 4B or 4D replaced with chromosome 4Th from *Thinopyrum ponticum*. The leaf pubescence of the two lines differs from S29 and from each other. In this work, we studied the effect of these substitutions on the manifestation of this trait. To quantify pubescence, the LHDetect2 program was used to determine trichome length and number on the leaf fold microphotographs. The key gene *Hl1* on chromosome 4B and another unidentified gene with a weak effect determine the leaf pubescence of the recipient S29. Their interaction leads to the formation of trichomes of up to 300 microns in length. Replacement of both copies of chromosome 4B with two copies of wheatgrass chromosome 4Th modifies leaf pubescence in line S29_4Th(4B) so that the leaf pubescence characteristic of S29 becomes more sparse, and trichomes of up to 600–700 μm in length are formed. Additionally, we described modification of pubescence in the substitution line S29_4Th(4D) where chromosome 4D that does not carry any pubescence gene was replaced. Under this substitution, trichomes of up to 400 μm in length were formed and the average length of trichomes on the underside of the leaf was reduced. The replacement of the *Hl1* gene in the lines was also confirmed by the allelic state of the linked microsatellite marker *Xgwm538*. Thus, as a result of the studies, a new leaf pubescence gene introgressed from *Th. ponticum* into bread wheat was identified. We designated it as *Hl1*th. For the purpose of selection, we propose to use the unlicensed informative microsatellite markers *Xgwm538* and *Xgwm165*, allowing chromosomes 4A, 4B, 4D and 4Th to be distinguished.
Key words: trichome; digital characteristics of pubescence; phenotypic markers; microsatellite markers; interactions of genes.

For citation: Simonov A.V., Gordeeva E.I., Genaev M.A., Li W., Bulatov I.O., Pshenichnikova T.A. A new leaf pubescence gene, *Hl1*th, introgressed into bread wheat from *Thinopyrum ponticum* and its phenotypic manifestation under homoeologous chromosomal substitutions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024; 28(6):602-609. DOI 10.18699/vjgb-24-67

Funding. This work was carried out within the framework of budget project No. FWNR-2022-0017. When processing the data, the computing resources of the "Bioinformatics" Center for Common Use were used with the support of budget project No. FWNR-2022-0020. We express our gratitude to the Center for Collective Use for Plant Reproduction and the Center for Collective Use for Microscopic Analysis at the Institute of Cytology and Genetics SB RAS.

Acknowledgements. The work was supported by the Russian Science Foundation grant No. 23-24-10029 and agreement with the administration of the Novosibirsk region No. R-63.

Новый ген опушения листа *Hl1*th, интрогрессированный в мягкую пшеницу от *Thinopyrum ponticum*, и его фенотипическое проявление при гомеологичных хромосомных замещениях

А.В. Симонов [1], Е.И. Гордеева [1], М.А. Генаев [1, 2], В. Ли [1, 2], И.О. Булатов[1, 3], Т.А. Пшеничникова [1]

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

Аннотация. На основе сорта яровой мягкой пшеницы Саратовская 29 (С29) были созданы голубозерные линии С29_4Th(4B) и С29_4Th(4D) с соответствующим замещением хромосом 4B и 4D хромосомой 4Th от пырея вида *Thinopyrum ponticum*. У этих линий опушение листа отличается от реципиента и различается между собой, в связи

с чем нами проведено исследование эффекта замещений на проявление данного признака. Для количественной оценки опушения была применена программа LHDetect2, определяющая длину и число трихом на микрофотографиях. Опушение листа у сорта С29 определяется главным геном *Hl1* в хромосоме 4B и еще одним геном со слабым эффектом с неизвестной хромосомной локализацией. Их взаимодействие приводит к формированию трихом длиной до 300 мкм. Замещение пары хромосом 4B на пару хромосом 4Th пырея модифицирует опушение листа у линии С29_4Th(4B). Характерное для сорта С29 опушение листа у линии С29_4Th(4B) становится реже, при этом образуются трихомы длиной до 600–700 мкм. Замещение гена *Hl1* на *Hl1*th у линии С29_4Th(4B) также подтверждается аллельным состоянием сцепленного с геном *Hl1* микросателлитного маркера *Xgwm538*. Нами была описана модификация опушения у замещенной линии С29_4Th(4D), где произошло замещение пары хромосом 4D, не содержащей гена опушения. Экспрессирующиеся совместно гены *Hl1* и *Hl1*th у линии С29_4Th(4D) в хромосомах 4B и 4Th соответственно, формируют трихомы длиной более 400 мкм. Однако в таком генотипе снижается средняя длина трихом в сравнении с реципиентом. Таким образом, в результате проведенных исследований идентифицирован новый ген опушения листа, интрогрессированный из вида *Th. ponticum* в мягкую пшеницу, который мы обозначили как *Hl1*th. Для ведения отбора мы предлагаем использовать находящиеся в открытом доступе информативные микросателлитные маркеры *Xgwm538* и *Xgwm165*, позволяющие различать хромосомы 4A, 4B, 4D и 4Th.

Ключевые слова: трихомы; цифровые характеристики опушения; фенотипические маркеры; микросателлитные маркеры; взаимодействие генов.

Introduction

Alien hybridization is widely used in breeding programs to transfer new useful traits into bread wheat (*Triticum aestivum*, AABBDD, $2n = 6x = 42$). For this purpose, both closely related species from the genus *Triticum* L. with similar genomes, such as *Aegilops*, are used, and species from other genera of the family *Poaceae*. Decaploid wheatgrass species *Thinopyrum ponticum* (Podp.) Barkworth & D.R. Dewey ($2n = 10x = 70$, StStStStEeEeEbEbExEx syn. *Agropyron elongatum* Host., *Elytrigia pontica* (Podp.) Holub) belongs to the tertiary gene pool of wheat relatives, and since the mid-20th century it has served as a source of useful genes in wheat breeding (Kroupin et al., 2019). With its tolerance to biotic and abiotic stress factors, *Th. ponticum* has become a donor of effective genes for resistance to various wheat diseases: root rot, leaf, stem and stripe rust, powdery mildew (Li H. et al., 2004; Li H., Wang, 2009; Niu et al., 2014; Wang et al., 2019; Li M. et al., 2021; Yang et al., 2023).

For 30 years, the Institute of Cytology and Genetics SB RAS has been expanding the collection of substituted, isogenic and alloplasmic lines of bread wheat based on the spring variety Saratovskaya 29 (S29) and other varieties. They carry either individual chromosomes, or certain rearrangements in the wheat chromosomes, or the cytoplasm of related species acquired through alien hybridization. Many of these introgressions have been identified using cytological or molecular methods (Leonova et al., 2008; Adonina et al., 2021; Shchukina et al., 2022; Pershina et al., 2023). The 4Th chromosome pair of the species *Th. ponticum* was transferred into the genome of S29 from the winter wheat variety Meropa developed in Bulgaria (Gordeeva et al., 2019). As a result, a substitution line with blue anthocyanin grain color was obtained. It has been established that the *Ba* gene responsible for the blue color of the aleurone layer (Blue aleuron) is located on chromosome 4Th of wheatgrass *Th. ponticum* (Zeven, 1991).

Using GISH analysis, it was shown that the centromeric and pericentromeric regions of chromosome 4Th originate from the E-genome chromosome, and the distal regions of its two arms, from the St-genome chromosome (Zheng et al., 2006). After the selection of hybrid plants in the generation $BC_7F_{2\text{-}3}$, according to the results of cytological and molecular analyses, no recombination was found between the wheat and wheatgrass chromosomes. Therefore, a complete replacement of 4B or 4D chromosome pair with 4Th chromosome pair occurred (Gordeeva et al., 2022). In addition to the blue color of the grain, the changes in the pubescence of leaf blades were visually and tactilely detected in comparison with the recipient in the obtained substitution lines S29_4Th(4B) and S29_4Th(4D) (Gordeeva et al., 2022).

Leaf pubescence is known to be an adaptive trait (Kaur, Kariyat, 2020). Hairiness in rice affects transpiration and drought tolerance, thereby increasing the yield (Hamaoka et al., 2017). A positive effect of this trait on photosynthetic parameters of wheat plants under drought conditions has been shown (Pshenichnikova et al., 2019; Osipova et al., 2020). The pubescence of cereal leaves is presented as outgrowths of epidermal cells – non-secretory trichomes; their length and density varies greatly among the carriers of different genomes (Pshenichnikova et al., 2017). For example, for winter wheat cultivars, leaf pubescence is not typical (our unpublished data), but the phenotypic diversity for this trait among spring wheat cultivars may depend on the region where they were developed (Genaev et al., 2012a).

The occurrence of this trait among cereals corresponds to the "law of homological series", which was formulated in 1920 by N.I. Vavilov (Vavilov, 1935). Among the cereal species, such as rye, barley, rice, and other, more distant species, the accessions may be found with leaf pubescence similar to that found in wheat (Shvachko et al., 2020). The main dominant gene *Hl1* of cv. S29 is located on chromosome 4BL and is responsible for the formation of medium-length trichomes (Maistrenko, 1976; Dobrovolskaya et al., 2007). The non-localized minor gene *Hl3* is also known to form small trichomes and slightly enhances the effect of the *Hl1* gene (Maistrenko, 1976). In the diploid genome of barley (*Hordeum vulgare* L.), the genes for leaf blade and for leaf sheath pubescence were mapped on the long arms of chromosomes 3H and 4H, respectively (Saade et al., 2017; Shvachko et al., 2020). In synthetic hexaploid wheat, the leaf sheath and leaf margin pubescence was associated with *Aegilops tauschii* Coss genome and the responsible gene was found in the long arm of 4D chromosome (Dobrovolskaya et al., 2007; Wan et al., 2015).

The present work is aimed at studying the phenotypic manifestation of a new allele of the *Hl1* gene for leaf pubescence transferred with chromosome 4Th from the species *Thinopyrum ponticum* to the genome of wheat cultivar S29. At the same time, work was carried out to identify the substitution of chromosomes 4B or 4D with chromosome 4Th of wheatgrass using molecular markers. The aim of the work was to study the phenotypical interaction between two genes during the replacement of chromosomes 4B and 4D by quantifying the length and number of trichomes.

Materials and methods

The plant material was represented by the spring recipient cultivar Saratovskaya 29 (S29) and two single chromosome substitution lines S29_4Th(4B) and S29_4Th(4D) (other previously used synonyms, respectively: s:S29_4Th(4B) and s:S29_4Th(4D), Gordeeva et al., 2019, 2022). According to cytological and molecular data (Gordeeva et al., 2019, 2022) the substitution lines are stable.

Analysis of leaf fold image. Microphotographs of transverse folds on the upper and lower sides of a boot leaf were used to determine the number and length of trichomes according to the protocol developed at the Institute of Cytology and Genetics SB RAS (Doroshkov et al., 2009). Images were obtained at the Center for Microscopic Analysis of the Institute of Cytology and Genetics SB RAS on a Carl Zeiss Axioscop 2 plus microscope through a 5x/0.12 lens. The microscope was equipped with an AxoCam HRc digital camera with a TV2/3C 0.63x adapter. The physical size of the field of view during shooting was 2730 × 2163 μm, the resolution of digital photography was 1300 × 1030 pixels. The physical pixel size was 2.1 microns. To obtain digital characteristics of leaf pubescence, the images were analyzed using the computing resources of the Bioinformatics Center for Common Use using the LHDetect2 program developed at the Laboratory of Evolutionary Bioinformatics and Theoretical Genetics of the Institute of Cytology and Genetics SB RAS (Genaev et al., 2012b). The program identifies trichomes, determines their length and produces the result as a sequence of numbers in a text file.

For each of two plants of the same genotype, 12 microphotographs were analyzed, with six folds from the upper and lower sides of the boot leaves. Trichomes formed under the influence of different genes within each class differ greatly in length. Therefore, the length values were presented in logarithmic scale. The calculation of the average length was carried out both in absolute values (microns) and as a decimal logarithm. Additionally, the distribution of trichome lengths and numbers was analyzed.

Statistical processing. The significance of differences between genotypes in length and number of trichomes was assessed using Student's *t*-test, for which MS Excel with the statistical add-in AgCStat was used (Gonchar-Zaykin, Chertov, 2003). The criterion for the significance of differences $p < 0.05$, 0.01 and 0.001 was indicated by one, two and three symbols, respectively: [a] – the difference between the substitution line and the recipient, [b] – the difference between the two substitution lines, * – the difference between the upper and lower sides of the leaf within the genotype. Diagrams of trichome length distribution were constructed in PAST v.3.0 statistical package. The data on the trichomes recognized from a total of six images were used for each genotype.

Genotyping. DNA was isolated from young leaves according to J. Plaschke et al. (1995). Samples diagnostic was made using PCR with microsatellite markers (SSR, simple sequence repeats) developed for chromosomes of the fourth homoeologous group according to recommended amplification programs (Röder et al., 1998). For this purpose, the markers *Xgwm538* and *Xgwm165* were chosen.

The *Xgwm538* marker is located on the long arm of chromosome 4B approximately 2.1 cM proximal to the *Hl1* gene in wheat (Dobrovolskaya et al., 2007). It showed amplification products of 157 bp in size for cv. S29 genome, and 155 bp for cv. Purple Feed (Dobrovolskaya et al., 2007). For cv Chinese Spring, it showed three fragments of 137, 147 and 152 bp, with the last product corresponding to chromosome 4B, and the others amplified from chromosome 4D as shown in null-tetrasomic lines (Brooks et al., 2006). This marker is often used, for example, to map the genes for infection resistance (Sukhwinder-Singh et al., 2003; Brooks et al., 2006; Singh et al., 2012). In our work, we used classical primers of the *Xgwm538* marker (Röder et al., 1998), which show PCR products from chromosomes 4B and 4D. This marker also showed multiple polymorphisms in two species of wheatgrass and wheat-wheatgrass hybrids (Kroupin, 2011). The *Xgwm165* marker is located on the long arms of chromosomes 4B and 4D and on the short arm of chromosome 4A with pericentric inversion (Röder et al., 1998). It is often used to map different genes and QTLs (Pshenichnikova et al., 2012; Salem, Mattar, 2014; Shchukina et al., 2018).

For PCR, a ready-made mixture of BioMaster HS-Taq reagents from BiolabMix LLC was used. PCR products were separated by electrophoresis in 3.5 % agarose gel with the addition of ethidium bromide. For electrophoresis, TBE buffer (Tris-borate-EDTA) and DNA fragment length marker Step50+ (BiolabMix, Novosibirsk, Russia) were used.

Growing conditions. The plant material was grown in a hydroponic greenhouse at the Center for Collective Use of Plant Reproduction of the Institute of Cytology and Genetics SB RAS. Growing conditions: lighting with 600W HPS lamps with adjustable suspension height (up to 45–50,000 lux at the level of the upper leaves) for 12–14 hours at a temperature of 18–20 °C at night and 24–26 °C during the day. Soil substrate: expanded clay, moistened with Knop nutrient solution three times a day.

Results

The pubescence of line S29_4Th(4B) with the replacement of chromosome 4B by chromosome 4Th tactilely distinguished it from the recipient. According to the results of microscopic observations and a detailed study of the leaf pubescence morphology (the method described above), both substitution lines, S29_4Th(4B) and S29_4Th(4D), differed from S29 and from each other. Figure 1 shows microphotographs of leaf folds of the three genotypes, which demonstrate visually distinguished trichomes of different lengths. Digital processing of microphotographs of S29 leaves showed that the average length of trichomes was 64.5 μm on the underside of the leaf and 67.1 μm on the top (see the Table). Their maximum length did not exceed 306 μm. Despite the fact that the longest tri-

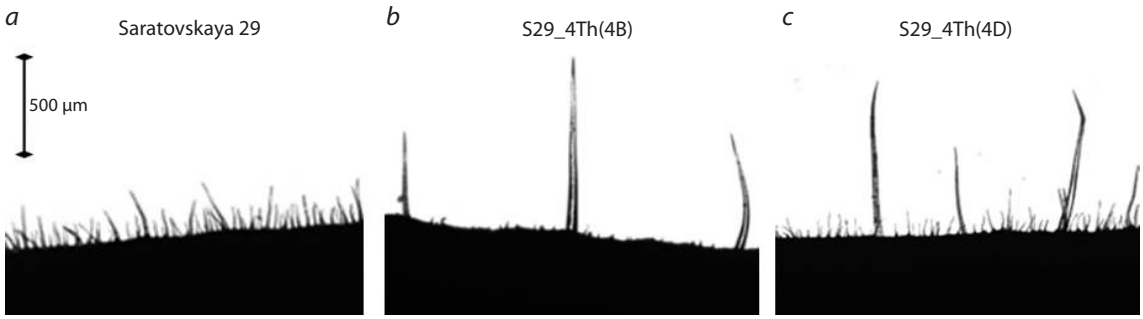


Fig. 1. Effect of substitution of chromosomes 4B and 4D with chromosome 4Th on the leaf pubescence phenotype in S29.
The photographs show trichomes from the folds of the upper part of the leaf in transmitted light.

chomes were formed on the upper side of the leaf, the density of pubescence and the sum of the lengths of all trichomes on the lower side of the leaf were one and a half times higher than on the upper side (see the Table).

The number of trichomes in line S29_4Th(4B) was reduced fivefold in comparison with the recipient (see the Table). On the upper side, single trichomes up to 705.3 µm long were observed, and on the lower side – up to 539.8 µm, which was two times greater than the maximum length of trichomes in S29. Trichomes were more common on the underside of the leaf than on the upper side, and the sum of lengths was twice the sum of the lengths of trichomes from the upper side. The average length of the trichomes on the upper side of the leaf in the line and in S29 did not differ significantly; however, on the lower side of the leaf, the difference in the average lengths of the trichomes was significant. The sum of the lengths of trichomes on both leaf sides in line S29_4Th(4B) was significantly reduced compared to S29.

Small and large trichomes differ in length greatly on the leaf fold of line S29_4Th(4B) (Fig. 1*b*). Figure 2 shows the distribution of trichomes of different lengths according to their number. In line S29_4Th(4B) (red bars), the number of trichomes with a length from 30 to 300 µm is significantly reduced, but a class of trichomes with a length of more than 300 µm has appeared, which is absent in S29. The difference in the average trichome logarithmic lengths of line S29_4Th(4B)

Average morphometric characteristics of leaf pubescence of cv. S29 and substitution lines with introgression from *Th. ponticum*

Genotypes	S29	S29_4Th(4B)	S29_4Th(4D)
Upper side of the leaf			
Average trichome number	32.8 ± 2.7	6.8 ± 1.1^{aaa}	39.8 ± 3.13^{bbb}
Average trichome length, µm	67.1 ± 2.7	71.5 ± 16.5	46.2 ± 1.97^{aaa}
Logarithmic trichome length	1.72 ± 0.02	1.41 ± 0.06^{aaa}	1.56 ± 0.01^{aaabb}
Trichome length limits, µm	8.1–306.6	8.4–705.3	9.5–426.8
Sum of trichome lengths, µm	2204 ± 240	482 ± 104^{aaa}	1840 ± 140^{bbb}
Lower side of the leaf			
Average trichome number	49.8 ± 2.2	10.0 ± 1.07^{aaa}	52.6 ± 4.5^{bbb}
Average trichome length, µm	64.5 ± 1.56	100.8 ± 13.91^{aa}	54.4 ± 2.0^{aaabbb}
Logarithmic trichome length	1.74 ± 0.01	1.59 ± 0.05^{aa}	1.63 ± 0.01^{aaa}
Trichome length limits, µm	11.2–265.3	9.5–539.8	10.4–421.3
Sum of trichome lengths, µm	3215 ± 173	1008 ± 133^{aaa}	2861 ± 239^{bbb}
Significance of differences between the upper and lower sides of the leaf			
Average trichome number	***	*	*
Average trichome length, µm	–	–	**
Sum of trichome lengths, µm	**	**	**

Note. Values marked with superscript "a" are significantly different between the substitution lines and the recipient S29; values marked with superscript "b" are significantly different among the substitution lines; values with superscript asterisks "*" are significantly different between the leaf sides in the same genotype; numbers of superscript symbols indicate significant levels: $p < 0.05^{ab*}$, $p < 0.01^{aabb**}$, $p < 0.001^{aaabbb***}$.

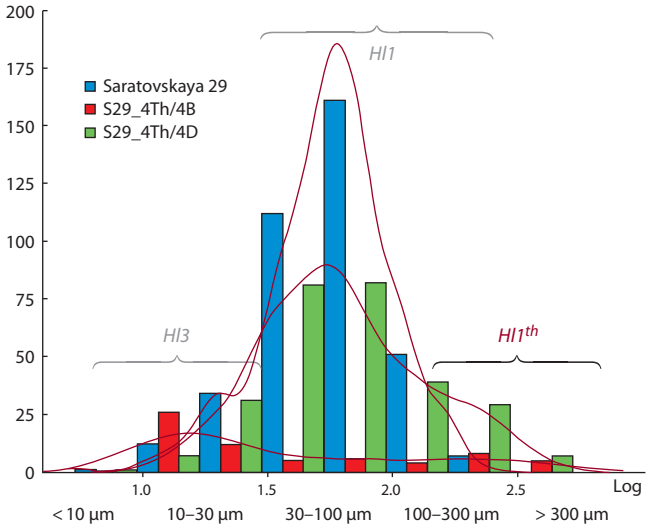


Fig. 2. Distribution of trichome density by length in different genotypes.

The X-axis scale is logarithmic. The Y-axis scale presents the number of trichomes by class from six images with a total width of approximately 13 mm (the width of one image is 2.163 mm). The data is presented for the lower surface of the leaf.

and the recipient was significant on both sides of the leaf (see the Table). The sum of the trichome lengths of the substitution line is also 3–4 times smaller on both sides compared to S29.

A different morphology of trichomes was observed on microphotographs of leaves in line S29_4Th(4D). The dense canopy of pubescence of S29 is preserved, but additional longer trichomes have been formed. Tactilely, they were hardly noticeable against the general background, but in microphotographs they stood out above the main trichome layer typical of S29 (Fig. 1*b*). The maximum trichome length in line S29_4Th(4D) exceeded 400 microns on both sides of the leaf, whereas in S29 it was slightly more than 300 microns on the upper side and more than 200 microns on the lower. In comparison with S29, the total number of trichomes increased insignificantly. At the same time, the difference in this indicator between lines S29_4Th(4D) and S29_4Th(4B) was significant on both sides of the leaf in favor of the first line. The sum of the trichome lengths in line S29_4Th(4D) decreased slightly compared to S29, but it was 3–4 times higher than that of line S29_4Th(4B). Nevertheless, the line with S29_4Th(4D) had a significantly lower average trichome length than the recipient variety with S29. In terms of the average logarithm length of trichomes, line S29_4Th(4D) differed from both S29 and line S29_4Th(4B). In Figure 2, the distribution of trichomes in line S29_4Th(4D) clearly demonstrates a multiple decrease in the number of trichomes of average length and the presence of a class of trichomes longer than in S29.

The microsatellite marker *Xgwm538* amplifying a 157-bp product is closely linked to the *Hl1* gene on chromosome 4B in S29. Using Chinese Spring nulli-tetrasomic lines, it was shown that this marker amplifies a fragment 174 bp in size, specific for chromosome 4B, and two fragments (147 and 137 bp) for chromosome 4D. In S29, only one product less than 150 bp is detected, corresponding to chromosome 4D. The *Xgwm538* marker confirmed in our work the presence of chromosomal substitution in the genome of S29 in both substitution lines (Fig. 3*a*). Line S29_4Th(4B) lacked a fragment larger than 150 bp, which corresponds to the diagnostic fragment for chromosome 4B. On the contrary, line S29_4Th(4D) did not have a fragment smaller than 150 bp, which indicates the presence of chromosome 4D. Thus, the polymorphic marker *Xgwm538* detects wheat chromosomes 4B and 4D of S29 and is not amplified on wheatgrass chromosome 4Th.

The microsatellite marker *Xgwm165* used in our work amplifies fragments on chromosomes 4A, 4B and 4D. We detected bright signals of amplification products of this marker for S29 in an agarose gel (Fig. 3*b*) with sizes of ~200, ~260 bp, as well as a less pronounced signal with a size of ~350 bp.

A PCR fragment about 200 bp was also observed in both substitution lines, which corresponds to chromosome 4A. Line S29_4Th(4B) lacked a 260 bp PCR product, but a 180 bp fragment was detected. A fragment of the same length (180 bp) was detected in line S29_4Th(4D) in combination with PCR products 200 and 260 bp in size as in S29, but there was no signal of 350 bp (Fig. 3*b*). PCR results obtained using *Xgwm165* suggest that 180 bp fragment is synthesized from chromosome 4Th and therefore can be used in determining this chromosomal substitution.

Discussion

The first identified wheat leaf pubescence gene with established chromosomal localization was the *Hl1* gene on chromosome 4B of cv. S29 (Maistrenko, 1976). The replacement of chromosome 4B of this variety with the chromosome of the non-pubescent cultivar Yanetzkis Probat changes the mor-

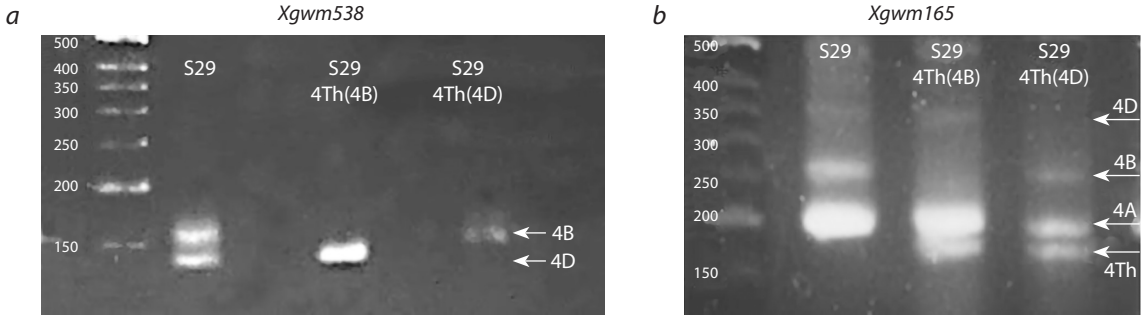


Fig. 3. Electrophoregram of PCR products obtained as a result of DNA amplification of S29 and substitution lines S29_4Th(4B) and S29_4Th(4D) using microsatellite markers *Xgwm538* (*a*) and *Xgwm165* (*b*).

phology of pubescence. This genotype has a noticeably reduced number of trichomes as well as their size (Doroshkov et al., 2016). This phenotype is determined by the presence of gene *Hl3* with a weak effect. In the absence of *Hl1* and *Hl3*, trichomes are practically not formed on the leaves of the S29 isogenic line, as was shown by the development of a glabrous isogenic line of this cultivar (Doroshkov et al., 2016).

Long, sparse trichomes are not typical for leaves of S29. Their appearance in the two substitution lines is apparently determined by a new variant of the pubescence gene transferred from *Th. ponticum*. Our studies indicate that wheatgrass chromosome 4Th carries a new allelic variant of the gene, orthologous to the wheat *Hl1* gene, but with a different phenotypic manifestation. The wheatgrass gene, which replaced wheat gene *Hl1* in line S29_4Th(4B), or was added to it in line S29_4Th(4D), not only forms long trichomes, but also reduces their total number. In accordance with the rules of the Catalog of Gene Symbols for Wheat (McIntosh et al., 2013), we designated the new allele with the symbol *Hl1*Th. Previously, the leaf margin pubescence gene *Hsh* (otherwise *Hs*) was found on chromosome 4H of barley (*Hordeum vulgare* L.) in a region comparable with chromosomes 4B and 4D (Korzun et al., 1999). A QTL associated with leaf margin pubescence was identified on chromosome 4D using the ITMI mapping population (Dobrovolskaya et al., 2007). In this work, we supplemented the homologous series of pubescence genes for the fourth group of chromosomes of cereal plants.

Previously, we suggested that the *Hl1* gene is responsible for the number of trichomes on the leaf surface, that is, for pubescence density (Doroshkov et al., 2014). In the substitution line C29_4Th(4B), the wheatgrass gene *Hl1*th in the absence of *Hl1* stimulated the formation of single long trichomes. When combined in one genotype in line S29_4Th(4D), the *Hl1*th gene apparently has a suppressive effect on *Hl1*, reducing the average length of trichomes and the sum of their lengths. Genes promoting formation of long, rare trichomes on the leaf surface were also localized on other chromosomes of different cultivars. The *Hl2* gene located chromosome on 7B, was found in the Chinese cultivar Hong-mang-mai (Taketa et al., 2002). The *Hl2*^{aesp} gene was introgressed from chromosome 7S of *Aegilops speltoides* Taush. in cv. Rodina (Pshenichnikova et al., 2007). In the species *Triticum timopheevii*, the *Hl*^{tt} gene with a similar phenotypic manifestation was found on chromosome 5A (Simonov et al., 2021).

The substitution lines of S29 were obtained to study the genes regulating anthocyanin biosynthesis. Wheatgrass chromosome 4Th carries the *Ba* gene responsible for the blue color of the grain aleurone layer (Gordeeva et al., 2019), which is also a phenotypic marker of the presence of this chromosome in the genome. However, the grain color is manifested both during the replacement of chromosome 4B and chromosome 4D. The phenotypic effect of introgressed pubescence can serve as a morphological marker for plant selection when obtaining the blue-grained forms with a certain chromosomal substitution in cultivars having the S29-like type of pubescence.

In this work, we studied polymorphism in microsatellite markers that were previously associated with chromosomes of the 4th group and with the *Hl1* gene on chromosome 4B in particular (Dobrovolskaya et al., 2007). Figure 3*a* shows the *Xgwm538* marker, which is located near the *Hl1* gene; in S29, it showed a 157 bp fragment (Dobrovolskaya et al., 2007). This marker can clearly indicate which of the wheat chromosomes, 4B or 4D, is replaced by 4Th from *Th. ponticum.* It was previously noted that the *Xgwm538* marker demonstrates specific fragments for the genomes of *Th. intermedium* and *Th. elongatum* (Kroupin et al., 2011). But in our work, no signals were detected from chromosome 4Th consisting of fragments of the St and E genomes of *Th. ponticum* (Zheng et al., 2006).

Since the wheatgrass chromosome 4Th does not recombine with homeologs, the *Xgwm165* marker was used in this work, amplifying products specific to chromosomes 4A, 4B and 4D (Röder et al., 1998). According to various molecular maps of the GrainGenes database (https://graingenes.org/cgi-bin/GG3/browse.cgi), on chromosome 4B this marker is located proximal to *Xgwm538* at a distance of about 20–30 cM. In the genome of S29, this marker synthesized different PCR products (Fig. 3*b*) for chromosomes of the 4th homoeologous group: for 4A, about 200 bp, for 4B, about 260 bp, and a weak signal of about 350 bp, presumably for chromosome 4D. *Xgwm165* also exhibited a 180 bp fragment for chromosome 4Th. This makes it possible to differentiate wheat plants with different chromosomal substitutions within the fourth group if the parent varieties are not characterized by S29 pubescence.

Trichomes form a special microclimate on the leaf surface; they are able to influence the stability of the surface air layer, changing laminar flows to turbulent ones (Schreuder et al., 2001). Turbulent flows, in turn, contribute to more dynamic gas exchange. Accordingly, changes in the parameters of surface pubescence should affect the parameters of stomatal conductance, the absorption of carbon dioxide and the intensity of moisture evaporation. In the future, it is planned to study these lines on the dynamics of photosynthetic parameters under various growing conditions, in particular, during adaptation to drought.

Conclusion

In our work, for the first time, a new allelic variant of the leaf pubescence gene *Hl1*th transferred from the decaploid species *Thinopyrum ponticum* to bread wheat was discovered and described using digital phenotyping. Its observed phenotypic manifestation against the background of the wheat genome was significantly different from the effect of the wheat gene *Hl1*. In its morphology, it is similar to that of the genes *Hl*^{tt} and *Hl2*^{aesp} localized in chromosomes 5A and 7S of the related cereals *T. timopheevii* and *Ae. speltoides*. The created lines make it possible to compare the adaptive value of similar leaf pubescence morphotypes controlled by different genes within the same model recipient genotype.

References

Adonina I.G., Timonova E.M., Salina E.A. Introgressive hybridization of common wheat: results and prospects. *Russ. J. Genet.* 2021; 57(4):390-407. DOI 10.1134/S1022795421030029

Brooks S.A., See D., Brown-Guedira G. SNP-based improvement of a microsatellite marker associated with Karnal bunt resistance in wheat. *Crop Sci*. 2006;46(4):1467-1470. DOI 10.2135/cropsci2005.05-0065

Dobrovolskaya O.B., Pshenichnikova T.A., Arbuzova V.S., Lohwasser U., Röder M.S., Börner A. Molecular mapping of genes determining hairy leaf character in common wheat with respect to other species of the Triticeae. *Euphytica*. 2007;155:285-293. DOI 10.1007/s10681-006-9329-7

Doroshkov A.V., Arsenina S.I., Pshenichnikova T.A., Afonnikov D.A. The use of computer-based image processing to leaf hairiness analysis in wheat *Triticum aestivum* L. *Informatsionniy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeding Scientists.* 2009;13(1):218-226 (in Russian)

Doroshkov A.V., Afonnikov D.A., Pshenichnikova T.A. Genetic analysis of leaf pubescence in isogenic lines of bread wheat Novosibirskaya 67. *Russ. J. Genet*. 2014;50:153-160. DOI 10.1134/S1022795413120028

Doroshkov A.V., Afonnikov D.A., Dobrovolskaya O.B., Pshenichnikova T.A. Interactions between leaf pubescence genes in bread wheat as assessed by high throughput phenotyping. *Euphytica*. 2016;207:491-500. DOI 10.1007/s10681-015-1520-2

Hamaoka N., Yasui H., Yamagata Y., Inoue Y., Furuya N., Araki T., Ueno O., Yoshimura A. A hairy-leaf gene, BLANKET LEAF, of wild *Oryza nivara* increases photosynthetic water use efficiency in rice. *Rice*. 2017;10(1):20. DOI 10.1186/s12284-017-0158-1

Genaev M.A., Doroshkov A.V., Morozova E.V., Pshenichnikova T.A., Afonnikov D.A. WheatPGE: A system for analysis of relationships among the phenotype, genotype, and environment in wheat. *Russ. J. Genet*. *Appl. Res*. 2012a;2(3):262-269. DOI 10.1134/S2079059712030045

Genaev M.A., Doroshkov A.V., Pshenichnikova T.A., Kolchanov N.A., Afonnikov D.A. Extracting quantitative characteristics of wheat leaf hairiness using image processing technique. *Planta*. 2012b;236:1943-1954. DOI 10.1007/s00425-012-1751-6

Gonchar-Zaykin P.P., Chertov V.G. Nadstroyka k Excel dlya statisticheskoy ocenki i analiza rezul'tatov polevyh i laboratornyh opytov [Elektronnyy resurs]. Available at: URL: http://vniioh.ru/nadstrojka-k-excel-dlya-statisticheskoj-ocenki-i-analiza-rezultatov-polevyx-i-laboratornyx-opytov (Accessed 26.09.2021) (in Russian)

Gordeeva E., Badaeva E., Yudina R., Shchukina L., Shoeva O., Khlestkina E. Marker-assisted development of a blue-grained substitution line carrying the *Thinopyrum ponticum* chromosome 4Th(4D) in the spring bread wheat Saratovskaya 29 background. *Agronomy*. 2019;9:723. DOI 10.3390/agronomy9110723

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;12:2934. DOI 10.3390/agronomy12122934

Kaur J., Kariyat R. Role of trichomes in plant stress biology. In: Núñez-Farfán J., Valverde P. (Eds.). Evolutionary Ecology of Plant-Herbivore Interaction. Springer, 2020;15-35. DOI 10.1007/978-3-030-46012-9_2

Korzun V., Malyshev S., Pickering R.A., Börner A. RFLP mapping of a gene for hairy leaf sheath using a recombinant line from *Hordeum vulgare* L. × *Hordeum bulbosum* L. cross. *Genome.* 1999;42(5):960-963. DOI 10.1139/g99-021

Kroupin P.Yu., Divashuk M.G., Fesenko I.A., Karlov G.I. Adaptation of microsatellite SSR-markers of wheat for the genome analysis of wheatgrass, intermediate wheatgrass, and wheat-wheatgrass hybrids. *Izvestiya Timiryazevskoy Sel'skohozjajstvennoy Akademii = Izvestiya of Timiryazev Agricultural Academy*. 2011;3:49-57 (in Russian)

Kroupin P.Yu., Divashuk M.G., Karlov G.I. Gene resources of perennial wild cereals involved in breeding to improve wheat crop. *Sel'skokhozyaistvennaya Biologiya = Agricultural Biology*. 2019;54(3):409-425. DOI 10.15389/agrobiology.2019.3.409eng

Leonova I.N., Röder M.S., Kalinina N.P., Budashkina E.B. Genetic analysis and localization of loci controlling leaf rust resistance of *Triticum aestivum* × *Triticum timopheevii* introgression lines. *Russ. J. Genet.* 2008;44:1431-1437. DOI 10.1134/S1022795408120077

Li H., Wang X. *Thinopyrum ponticum* and *Th. intermedium*: the promising source of resistance to fungal and viral diseases of wheat. *J. Genet. Genomics.* 2009;36(9):557-565. DOI 10.1016/S1673-8527(08)60147-2

Li H., Conner R.L., Chen Q., Li H., Laroche A., Graf R.J., Kuzyk A.D. The transfer and characterization of resistance to common root rot from *Thinopyrum ponticum* to wheat. *Genome.* 2004;47(1):215-223. DOI 10.1139/g03-095

Li M., Wang Y., Liu X., Li X., Wang H., Bao Y. Molecular cytogenetic identification of a novel wheat – *Thinopyrum ponticum* $1J^S$ (1B) substitution line resistant to powdery mildew and leaf rust. *Front. Plant Sci.* 2021;12:727734. DOI 10.3389/fpls.2021.727734

Maistrenko O.I. Identification and localization of genes controlling the pubescence of the leaf of young soft wheat plants. *Genetika* (*Moscow*). 1976;12(1):5-15 (in Russian)

McIntosh R.A., Devos K.M., Dubcovsky J., Morris C.F., Rogers W.J. Catalogue of Gene Symbols for Wheat. Supplement. 2003. Available at: https://wheat.pw.usda.gov/ggpages/wgc/2003upd.html

Niu Z., Klindworth D.L., Yu G., Friesen T.L., Chao S., Jin Y., Cai X., Ohm J.-B., Rasmussen J.B., Xu S.S. Development and characterization of wheat lines carrying stem rust resistance gene *Sr43* derived from *Thinopyrum ponticum*. *Theor. Appl. Genet.* 2014;127(4):969-980. DOI 10.1007/s00122-014-2272-4

Osipova S.V., Rudikovskii A.V., Permyakov A.V., Rudikovskaya E.G., Permyakova M.D., Verkhoturov V.V., Pshenichnikova T.A. Physiological responses of wheat (*Triticum aestivum* L.) lines with genetically different leaf pubescence. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2020;24(8):813-820. DOI 10.18699/VJ20.678

Pershina L.A., Trubacheeva N.V., Shumny V.K., Badaeva E.D. Development and characterization of a line with substitution of chromosome 4B of wheat *Triticum aestivum* L. on chromosome $4H^{mar}$ of wild barley *Hordeum marinum* ssp. *gussoneanum* (4*x*). *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(6):545-552. DOI 10.18699/VJGB-23-66

Plaschke J., Ganal M.W., Röder M.S. Detection of genetic diversity in closely related bread wheat using microsatellite markers. *Theor. Appl. Genet*. 1995;91(6-7):1001-1007. DOI 10.1007/BF00223912

Pshenichnikova T.A., Lapochkina I.F., Shchukina L.V. The inheritance of morphological and biochemical traits introgressed into common wheat (*Triticum aestivum* L.) from *Aegilops speltoides* Tausch. *Genet. Resour. Crop Evol.* 2007;54(2):287-293. DOI 10.1007/s10722-005-4499-z

Pshenichnikova T.A., Khlestkina E.K., Shchukina L.V., Simonov A.V., Chistyakova A.K., Morozova E.V., Landjeva S., Karceva T., Börner A. Exploitation of Saratovskaya 29 (Janetzkis Probat 4D*7A) substitution and derivate lines for comprehensive phenotyping and molecular mapping of quantitative trait loci (QTL). In: EWAC Newsletter 2012, Proc. of the 15th International EWAC Conference, 7–11 November 2011, Novi Sad, Serbia. 2012;19-22

Pshenichnikova T.A., Doroshkov A.V., Simonov A.V., Afonnikov D.A., Börner A. Diversity of leaf pubescence in bread wheat and relative species. *Genet. Resour. Crop Evol.* 2017;64(7):1761-1773. DOI 10.1007/s10722-016-0471-3

Pshenichnikova T.A., Doroshkov A.V., Osipova S.V., Permyakov A.V., Permyakova M.D., Efimov V.M., Afonnikov D.A. Quantitative characteristics of pubescence in wheat (*Triticum aestivum* L.) are associated with photosynthetic parameters under conditions of normal and limited water supply. *Planta*. 2019;249(3):839-847. DOI 10.1007/s00425-018-3049-9

Röder M.S., Korzun V., Wendehake K., Plaschke J., Tixier M.H., Leroy P., Ganal M.W. A microsatellite map of wheat. *Genetics*. 1998;149(4):2007-2023. DOI 10.1093/genetics/149.4.2007

Saade S., Kutlu B., Draba V., Förster K., Schumann E., Tester M., Pillen K., Maurer A. A donor-specific QTL, exhibiting allelic variation for leaf sheath hairiness in a nested association mapping population, is located on barley chromosome 4H. *PloS One*. 2017;12(12):e0189446. DOI 10.1371/journal.pone.0189446

Salem K.F.M., Mattar M.Z. Identification of microsatellite alleles for salt tolerance at seedling stage in wheat (*Triticum aestivum* L.). *Life Sci. J*. 2014;11(12s):1064-1073

Schreuder M.D.J., Brewer C.A., Heine C. Modelled influences of non-exchanging trichomes on leaf boundary layers and gas exchange. *J. Theor. Biol*. 2001;210:23-32. DOI 10.1006/jtbi.2001.2285

Shchukina L.V., Pshenichnikova T.A., Khlestkina E.K., Misheva S., Kartseva T., Abugalieva A., Börner A. Chromosomal location and mapping of quantitative trait locus determining technological parameters of grain and flour in strong-flour bread wheat cultivar Saratovskaya 29. *Cereal Res. Commun.* 2018;46(4):628-638. DOI 10.1556/ 0806.46.2018.047

Shchukina L.V., Simonov A.V., Demenkova M.A., Klykov A.G., Shamanin V.P., Pozherukova V.E., Lepekhov S.B., Chebatareva M.V., Petin V.A., Börner A., Pshenichnikova T.A. Increased grain protein and gluten contents of bread wheat caused by introgression of a *T. timopheevii* segment into chromosome 2A. *Euphytica.* 2022;218: 170. DOI.10.1007/s10681-022-03121-w

Shvachko N.A., Semilet T.V., Tikhonova N.G. Trichomes in higher plants: homological series in hereditary variability and molecular genetic mechanisms. *Russ. J. Genet*. 2020;56(11):1359-1370. DOI 10.1134/S1022795420110083

Simonov A.V., Smirnova O.G., Genaev M.A., Pshenichnikova T.A. The identification of a new gene for leaf pubescence introgressed into bread wheat from *Triticum timopheevii* Zhuk. and its manifestation in a different genotypic background. *Plant Genet. Resour*. 2021;19(3):238-244. DOI 10.1017/S1479262121000277

Singh A., Pallavi J.K., Gupta P., Prabhu K.V. Identification of microsatellite markers linked to leaf rust resistance gene *Lr25* in wheat. *J. Appl. Genet*. 2012;53(1):19-25. DOI 10.1007/s13353-011-0070-0

Sukhwinder-Singh, Brown-Guedira G.L., Grewal T.S.,·Dhaliwal H.S., Nelson J.C., Singh H., Gill B.S. Mapping of a resistance gene effective against Karnal bunt pathogen of wheat. *Theor. Appl. Genet.* 2003;106(2):287-292. DOI 10.1007/s00122-002-1112-0

Taketa S., Chang C.L., Ishii M., Takeda K. Chromosome arm location of the gene controlling leaf pubescence of a Chinese local wheat cultivar Hong-mang-mai. *Euphytica*. 2002;125(2):141-147. DOI 10.1023/A:1015812907111

Vavilov N.I. Scientific Foundations of Wheat Breeding. Moscow–Leningrad: Selkhozgiz Publ., 1935;70-87 (in Russian)

Wan H., Yang Y., Li J., Zhang Z., Yang W. Mapping a major QTL for hairy leaf sheath introgressed from *Aegilops tauschii* and its association with enhanced grain yield in bread wheat. *Euphytica*. 2015; 205:275-285. DOI 10.1007/s10681-015-1457-5

Wang S., Wang C., Wang Y., Wang Y., Chen C., Ji W. Molecular cytogenetic identification of two wheat – *Thinopyrum ponticum* substitution lines conferring stripe rust resistance. *Mol. Breed.* 2019;39:143. DOI 10.1007/s11032-019-1053-9

Yang G., Deng P., Ji W., Fu S., Li H., Li B., Li Zh., Zheng Q. Physical mapping of a new powdery mildew resistance locus from *Thinopyrum ponticum* chromosome 4AgS. *Front. Plant Sci.* 2023;14:1131205. DOI 10.3389/fpls.2023.1131205

Zeven A.C. Wheats with purple and blue grains: a review. *Euphytica*. 1991;56:243-258

Zheng Q., Li B., Mu S., Zhou H., Li Z. Physical mapping of the blue-grained gene(s) from *Thinopyrum ponticum* 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 December 3, 2023. Revised June 26, 2024. Accepted June 26, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):610-618

DOI 10.18699/vjgb-24-68

The effect of *T. aestivum* chromosomes 1A and 1D on fertility of alloplasmic recombinant (*H. vulgare*)-*T. aestivum* lines depending on cytonuclear compatibility

L.A. Pershina [1, 2], N.V. Trubacheeva [1, 2], V.K. Shumny [1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
pershina@bionet.nsc.ru

Abstract. The effect of *T. aestivum* L. chromosomes 1A and 1D on fertility of recombinant bread wheat allolines of the same origin carrying the cytoplasm of barley *H. vulgare* L. and different levels of cytonuclear compatibility was studied. Alloline L-56 included mainly fully sterile (FS) and partially sterile (PS) plants, alloline L-57 included partially fertile (PF) plants and line L-58 included fertile (F) ones. Analysis of morphobiological traits and pollen painting indicated complete or partial male sterility in plants of allolines L-56 and L-57. To differentiate genotypes with cytonuclear coadaptation and genotypes with cytonuclear incompatibility, PCR analysis of the 18S/5S mitochondrial (mt) repeat was performed. Heteroplasmy (simultaneous presence of barley and wheat mtDNA copies) was found in FS, PS, PF and some F plants, which was associated with a violation of cytonuclear compatibility. Wheat-type homoplasmy (hm) was detected in the majority of the fertile plants, which was associated with cytonuclear coadaptation. The allolines used as maternal genotypes were crossed with wheat-rye substitution lines 1R(1A) and 1R(1D). In F_1, all plants of PF×1R(1A) and PF×1R(1D) combinations were fertile, and in F_2, a segregation close to 3 (fertile) : 1 (sterile) was observed. These results showed for the first time that chromosomes 1A and 1D carry one dominant *Rf* gene, which controls the restoration of male fertility of bread wheat carrying the cytoplasm of *H. vulgare*. All plants of F_1 combinations FS×1R(1A), FS×1R(1D), PS×1R(1A), PS×1R(1D) were sterile, which indicates that a single dose of genes localized on wheat chromosomes 1A or 1D is not enough to restore male fertility in FS and PS plants. All plants of hybrid combinations F(hm)×1R(1A) and F(hm)×1R(1D) in both F_1 and F_2 were fertile, that is, fertility of allolines with cytonuclear coadaptation does not depend on wheat chromosomes 1A and 1D.
Key words: allolines (*H. vulgare*)-*T. aestivum*; chromosomes 1A and 1D; mtDNA; violation of cytonuclear compatibility; cytonuclear coadaptation; *Rf* genes.

For citation: Pershina L.A., Trubacheeva N.V., Shumny V.K. The effect of *T. aestivum* chromosomes 1A and 1D on fertility of alloplasmic recombinant (*H. vulgare*)-*T. aestivum* lines depending on cytonuclear compatibility. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):610-618. DOI 10.18699/vjgb-24-68

Funding. This research was funded by Budget Project FWNR-2022-0017.

Acknowledgements. The authors are grateful to Dr. Shchapova A.I. for the wheat-rye substitution lines provided for the work.

Влияние хромосом 1A и 1D *T. aestivum* на фертильность аллоплазматических рекомбинантных линий (*H. vulgare*)-*T. aestivum* в зависимости от цитоядерной совместимости

Л.А. Першина [1, 2], Н.В. Трубачеева [1, 2], В.К. Шумный [1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
pershina@bionet.nsc.ru

Аннотация. Изучено влияние хромосом 1A и 1D *T. aestivum* L. на фертильность рекомбинантных аллолиний мягкой пшеницы одного происхождения, имеющих цитоплазму ячменя *H. vulgare* L. и разный уровень цитоядерной совместимости. Аллолиния Л-56 включает преимущественно полностью стерильные (ПС) и частично стерильные (ЧС) растения; аллолиния Л-57 – частично фертильные (ЧФ) растения, а линия Л-58 – фер-

тильные (Ф) растения. Результаты анализа морфобиологических признаков и окраски пыльцы указывают на проявление полной или частичной мужской стерильности у растений аллолиний Л-56 и Л-57. Для разделения генотипов с цитоядерной коадаптацией и генотипов, у которых цитоядерная совместимость нарушена, выполнен ПЦР-анализ 18S/5S митохондриального (мт) повтора. Показано, что ПС, ЧС, ЧФ и часть Ф растений характеризуются гетероплазмией (наличием копий мтДНК ячменя и пшеницы), что ассоциировано с нарушением цитоядерной совместимости. У основной части фертильных растений выявлена гомоплазмия (гм) пшеничного типа, что ассоциировано с цитоядерной коадаптацией. Растения аллолиний, использованные в качестве материнских генотипов, были скрещены с пшенично-ржаными замещенными линиями 1R(1A) и 1R(1D). В F_1 все растения комбинаций ЧФ×1R(1A) и ЧФ×1R(1D) были фертильными, а в F_2 наблюдали расщепление, близкое к 3 (фертильные) : 1 (стерильные). Эти результаты впервые показали, что в хромосомах 1A и 1D локализовано по одному доминантному гену *Rf*, контролирующему восстановление мужской фертильности мягкой пшеницы, несущей цитоплазму *H. vulgare*. Все растения F_1 комбинаций ПС×1R(1A), ПС×1R(1D), ЧС×1R(1A), ЧС×1R(1D) стерильные, что указывает на то, что одной дозы генов, локализованных в хромосомах пшеницы 1A или 1D, недостаточно для восстановления мужской фертильности у ПС и ЧС растений. Все растения гибридных комбинаций $Ф_{гм}$×1R(1A) и $Ф_{гм}$×1R(1D) и в F_1 и в F_2 были фертильными, т. е. у аллолиний с цитоядерной коадаптацией нет зависимости проявления фертильности от влияния хромосом пшеницы 1A и 1D.
Ключевые слова: аллолинии (*H. vulgare*)-*T. aestivum*; хромосомы 1A и 1D; мтДНК; нарушение цитоядерной совместимости; цитоядерная коадаптация; гены *Rf*.

Introduction

Alloplasmic lines (allolines) are resulted from repeated crosses of wide F_1 hybrids with a pollen parent. These lines combine the cytoplasm from the maternal species with the nuclear genome from the paternal species (Tsunewaki, 1996). The replacement of cytoplasm affects nuclear-mitochondrial and nuclear-chloroplast interactions (Yang et al., 2008; Crosatti et al., 2013; Soltani et al., 2016) leading to changes in plant development (Badaeva et al., 2006), resistance to stress factors (Buloychik et al., 2002; Talukder et al., 2015; Takenaka et al., 2019), morphological and agronomic traits (Liu C.G. et al., 2002; Atienza et al., 2008; Tao et al., 2011; Klimushina et al., 2013). The most relevant manifestation of cytonuclear conflict is cytoplasmic male sterility (CMS) (Tsunewaki, 1996), which is associated with aberrant mitochondrial genes that negatively affect the development of flower and pollen organs (Yang et al., 2008).

In a number of economically important crops, CMS lines in combination with maintainer and restorer lines carrying male fertility restoration genes (*Rf* – restorer-of-fertility) have been used in hybrid breeding (Islam et al., 2014; Bohra et al., 2016; Gupta et al., 2019). The sources of CMS and restorer genes are a critical tool in this technology. In addition, cytoplasmic substitution results in an increase of cytoplasmic diversity, as has been shown for crops such as rice (Liu Y. et al., 2016), sugar cane (Rafee et al., 2010), and bread wheat (Liu C.G. et al., 2002; Klimushina et al., 2013; Pershina et al., 2018).

In this regard, studying the process of allolines development and the genetic control of fertility restoration is an important task both for identifying new CMS-*Rf* systems for hybrid breeding and for obtaining new genotypes for conventional breeding programs. In bread wheat, male fertility restoration of genotypes carrying the cytoplasm of *T. timopheevii* (Sinha et al., 2013), *H. chilense* (Martin et al., 2010), *Aegilops* species (Tsunewaki, 2015; Hohn, Lukaszewski, 2016) and cultivated barley *H. vulgare* (Pershina et al., 2012; Trubacheeva et al., 2021) has been studied. Most *Rf* genes in bread wheat were located in clusters on chromosomes of the homeologous groups 1, 2 and 6, and the largest number was located on chromosome 1 (Gupta et al., 2019).

In a previous study, we established for the first time that the dominant gene controlling the male fertility restoration of wheat carrying *H. vulgare* cytoplasm was located on the short arm of wheat chromosome 1B (Trubacheeva et al., 2021). In this work, the role of homeologous group 1 for fertility restoration of bread wheat allolines carrying *H. vulgare* cytoplasm continues to be studied. The aim of the work was to study the effect of *T. aestivum* chromosomes 1A and 1D on the male fertility of recombinant wheat allolines carrying cultivated barley cytoplasm depending on the level of their fertility and cytonuclear compatibility. This approach allowed us to identify allolines (*H. vulgare*)-*T. aestivum* as models for studying the localization of the *Rf* genes on chromosomes 1A and 1D.

Materials and methods

Plant material. Three recombinant allolines (*H. vulgare*)-*T. aestivum* derived from individual plants of backcross (BC) generations of a barley-wheat hybrid *H. vulgare* (Nepolegaushii) × *T. aestivum* (Saratovskaya 29), sequentially pollinated with wheat varieties Saratovskaya 29, Mironovskaya 808, Pyrotrix 28, Saratovskaya 29, Pyrotrix 28, were studied (Fig. 1). In previous studies, Saratovskaya 29 was found to be a sterility fixer in backcrossed progenies of barley-wheat hybrids (Pershina et al., 2012), while Mironovskaya 808 and Pyrotrix 28 were identified as male fertility restorers for wheat alloplasmic lines carrying cultivated barley cytoplasm (Pershina et al., 1998, 2012). BC_1–BC_4 generations and the barley-wheat hybrid were male-sterile, but female-fertile, and in BC_5, some 42-chromosomal plants with partially restored male fertility were isolated. Self-pollinated generations F_2BC_5–F_5BC_5 were obtained

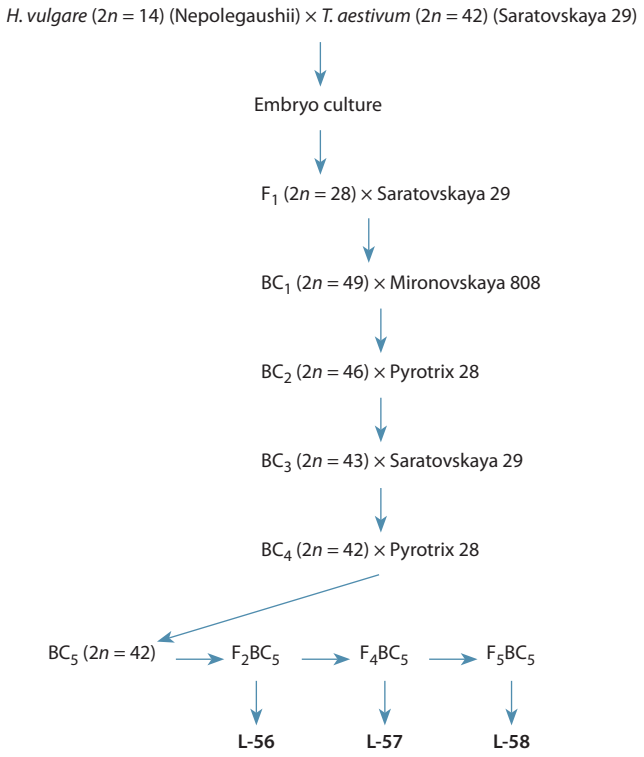


Fig. 1. Production of the alloplasmic recombinant lines (*H. vulgare*)-*T. aestivum* L-56, L-57, L-58.

from these plants, which became the sources of the studied allolines. Alloline L-56 was isolated from F_2BC_5, and allolines L-57 and L-58 were isolated from F_4BC_5 and F_5BC_5, respectively. Beginning from F_3BC_5, plants with the highest level of productivity were used to obtain each subsequent self-pollinated generation.

Methods for studying morphobiological characteristics of alloplasmic recombinant lines. Plants of the lines used were characterized by fertility level: FS – fully sterile (no seeds); PS – partially sterile (1–9 seeds); PF – partially fertile (10–19 seeds); F – fertile (more than 19 seeds per main spike). At least 20 plants of each line grown in a hydroponic greenhouse were evaluated.

Pollen fertility as the main criterion for assessing male fertility/sterility was analyzed in plants with different fertility levels. For this purpose, crushed preparations in Lugol's solution (1 % iodine solution in an aqueous solution of potassium iodide) were prepared on a slide from anthers isolated during flowering from three different flowers of the same spike. Plant height, the number of spikes, main spike length, the number of spikelets per main spike, grain number per main spike and per plant, and 1,000-grain weight were determined for the plants of each alloline. The differences between the average values of the studied traits in alloline L-56 compared with the L-57 line and in alloline L-57 compared with alloline L-58 were statistically evaluated by Student's *t*-test. Data were analyzed using Statistica v.7.0.61.0.

PCR analysis of the 18S/5S mitochondrial (mt) repeat. Specific primers for the 18S/5S repeat were designed based on the mitochondrial genome sequences published earlier (Coulthart et al., 1993). The PCR products were electrophoresed in a 1.5 % agarose gel with 1× TAE buffer and visualized with ethidium bromide. Gel images were captured using the gel documentation system Gel Doc XR+ ("Bio-Rad", USA). Total DNA was isolated from green leaves cut before earing according to a previously published protocol (Current Protocols…, 1987). From one to eight samples from individual genotypes were analyzed. In this part of the work, the control was the barley variety Nepolegaushii as a source of cytoplasm for allolines and the bread wheat variety Pyrotrix 28 as a source of wheat cytoplasm (one of the recurrent genotypes).

Evaluation of the fertility of hybrids between alloplasmic lines and wheat-rye substitution lines 1A(1R) and 1D(1R) in F_1 and F_2. To assess the effect of wheat chromosomes 1A and 1D on the fertility of allolines depending on the level of their cytonuclear compatibility, plants of these lines (as maternal genotypes) were crossed with wheat-rye substitution lines 1A(1R) and 1D(1R) to replace in F_1 one 1A or 1D chromosome of allolines with rye chromosome 1R. The 1A(1R) and 1D(1R) lines used in the work were obtained as a result of substituting wheat Saratovskaya 29 chromosomes with rye chromosome 1R of variety Onokhoyskaya (Shchapova, Kravtzova, 1982). In hybridization, FS and PS plants of alloline L-56, PF plants of alloline L-57 and some F plants of alloline L-58 were used. The spikes of mother plants, as well as F_1 and F_2 plants grown in a hydroponic greenhouse, were bagged before flowering. In individual plants of F_1 and F_2, the seed set in the main spike was assessed. Based on the seed set in the F_2 generation, the individual plants were classified into fertile and sterile groups according to the recommendations of P. Sinha et al. (Sinha et al., 2013): completely sterile plants and plants that set no more than four grains in the main spike were classified as sterile, while those that set five or more grains in the main spike were classified as fertile. Pearson's chi-squared test ($\alpha = 0.05$) was used for the deviation of the observed data from the theoretically expected segregation into fertile and sterile plants in F_2.

Results

Characteristics of the recombinant allolines

Alloline L-56 consisted of partially sterile (60 %) and completely sterile plants (35 %); the frequency of partially fertile plants was 5 % (Table 1).

The majority of plants in alloline L-57 were partially fertile (85 %), the rest were partially sterile (5 %) and fertile (10 %). Alloline L-58 consisted of fertile (92 %) and partially fertile plants (8 %). Figure 2 shows plant spikes with different fertility levels.

In fully sterile plants, stigmas were normally developed, but anthers were absent. In partially sterile and partially

Table 1. Fertility level of recombinant allolines (*H. vulgare*)-*T. aestivum* L-56, L-57, L-58

Allolines	Number of plants studied	Number and frequency* (%) of plants			
		FS (0)	PS (1–9)#	PF (10–19)#	F (>19)#
L-56	20	7 (35 %)	12 (60 %)	1 (5 %)	0
L-57	20	0	1 (5 %)	17 (85 %)	2 (10 %)
L-58	25	0	0	2 (8 %)	23 (92 %)

Note. FS – full sterility; PS – partial sterility; PF – partial fertility; F – fertility. # – grain number per main spike.

fertile plants, anthers were not fully developed compared to fertile plants, and not all pollen grains were stained (Fig. 3).

The comparison of the average values of the studied traits in L-56, represented mainly by sterile and partially sterile plants, compared with L-57, consisting mainly of partially fertile plants, showed that L-56 exceeded L-57 only in terms of the number of spikes per plant. The value of other traits (plant height, length of the main spike, number of spikelets per main spike, grain number per main spike and per plant) in L-56 is significantly lower than in L-57 (Table 2).

In alloline L-57, the values of main spike length, number of spikelets per main spike, grain number per main spike and per plant were significantly lower compared to alloline L-58, represented mainly by fertile plants. Thousand-kernel weight did not differ between the studied allolines.

PCR analysis of 18S/5S mtDNA in recombinant allolines

Heteroplasmy (simultaneous presence of barley and wheat mtDNA copies) was found in all studied plants of alloline L-56, including fully sterile, partially sterile plants and one partially fertile plant (Fig. 4; Table 3). Heteroplasmy was also detected in six partially fertile and two fertile plants of alloline L-57. In alloline L-58, two partially fertile and two fertile plants were found to have heteroplasmy, and six fertile plants had wheat-type homoplasmy. These results were used to divide alloplasmic genotypes into groups with different levels of cytonuclear incompatibility according to the data (Aksyonova et al., 2005; Trubacheeva et al., 2021) (Table 3).

Fig. 2. Plant spikes: 1, 2 – fertile; 3 – partially fertile; 4 – partially sterile; 5 – fully sterile.

In plants with heteroplasmy, cytonuclear compatibility was disrupted, while in plants with homoplasmy it was not.

Analysis of hybrids between recombinant allolines and wheat-rye substitution lines 1R(1A) and 1R(1D)

Individual fully sterile (FS) and partially sterile (PS) plants of alloline L-56 were pollinated with pollen of wheat-rye

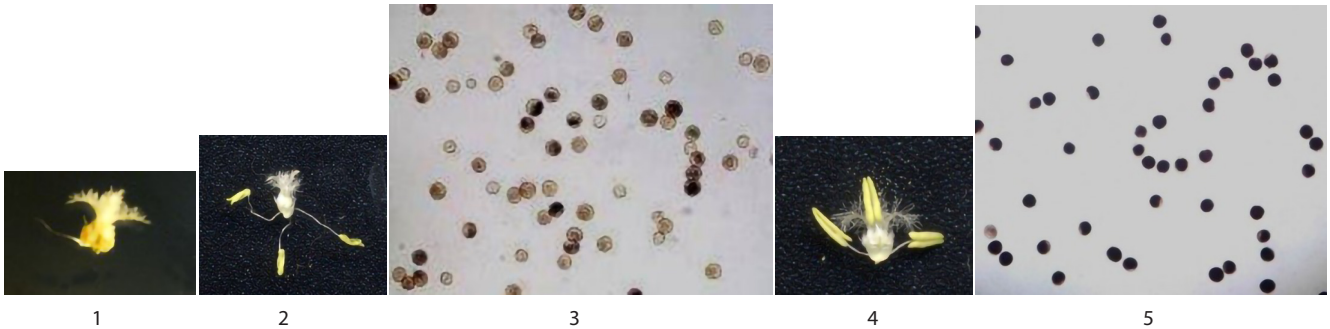


Fig. 3. Stigma (1) of a fully sterile plant; stigma and anthers (2), pollen grains (3) of a partially fertile plant; stigma and anthers (4) and pollen grains (5) of a fertile plant.

Table 2. Agronomic characteristics of recombinant (*H. vulgare*)-*T. aestivum* lines

Traits	L-56	L-57	L-58
Plant height, cm	76.38 ± 2.66$^{(**)}$	87.81 ± 1.97	89.55 ± 2.0
Tiller number	4.78 ± 0.32*	3.92 ± 0.24	3.85 ± 0.22
Main spike length, cm	6.07 ± 0.41$^{(***)}$	8.12 ± 0.34$^{/*/}$	9.14 ± 0.23
Spikelet number per main spike	13.21 ± 0.67$^{(***)}$	16.50 ± 0.25$^{/*/}$	18.15 ± 0.65
Grain number per main spike	3.76 ± 1.98$^{(***)}$	16.35 ± 1.66$^{/***/}$	33.10 ± 2.56
Grain number per plant	15.57 ± 6.35$^{(***)}$	52.50 ± 7.84$^{/***/}$	115.74 ± 13.67
Thousand-kernel weight, g	35.32 ± 1.18	34.25 ± 1.12	35.67 ± 1.31

Note. The difference compared to L-57 is significantly greater at * $p < 0.05$; significantly less at $^{(**)}$ $p < 0.01$ and $^{(***)}$ $p < 0.001$; compared with L-58, significantly less at $^{/*/}$ $p < 0.05$ and $^{/***/}$ $p < 0.001$.

Table 3. Results of the study of the 18S/5S mtDNA repeat in recombinant allolines (*H. vulgare*)-*T. aestivum*

Lines	Fertility level	Number of plants studied	18S/5S mtDNA	Cytonuclear compatibility
L-56	FS	5	B + W	Disrupted
	PS	8	B + W	Disrupted
	PF	1	B + W	Disrupted
L-57	PF	6	B + W	Disrupted
	F	2	B + W	Disrupted
L-58	PF	2	B + W	Disrupted
	F	2	B + W	Disrupted
	F	6	W	Not disrupted
Nepolegaushii	F	2	B	Not disrupted
Pyrotrix 28	F	2	W	Not disrupted

Note. B – barley; W – wheat; Nepolegaushii is a variety of barley; Pyrotrix 28 is a variety of bread wheat.

substitution lines 1R(1A) and 1R(1D). Seeds were set in all combinations of crossing due to female fertility of FS and PS plants. F_1 plants were grown from the set seeds: 18 plants of combination L-56(FS) × 1R(1A), 20 plants of combination L-56(FS) × 1R(1D), 15 plants of combination L-56(PS) × 1R(1A) and 17 plants of combination L-56(PS) × 1R(1D). All F_1 plants of these hybrid combinations did not set seeds from self-pollination (Table 4).

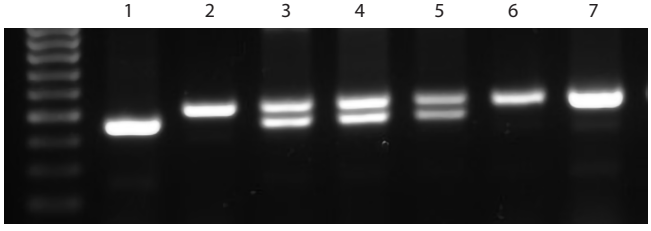


Fig. 4. Agarose gel electrophoresis of PCR products using the 18S/5S mt repeat marker.

1 – barley *H. vulgare* variety Nepolegaushii; 2 – wheat *T. aestivum* variety Pyrotrix 28; 3 – completely sterile L-56 plant; 4 – partially sterile L-56 plant; 5 – partially fertile L-57 plant; 6, 7 – fertile L-58 plants.

The complete sterility of F_1 hybrids heterozygous for wheat chromosomes 1A and 1D indicates that the fertility of partially sterile plants depends on chromosomes 1A and 1D. However, a single dose of the gene localized on each of these chromosomes is not sufficient to restore the male fertility of these plants.

Partially fertile (PF) plants of alloline L-57 were included in hybridization with wheat-rye substitution lines. Fifteen plants were grown from seeds of the hybrid combination L-57(PF) × 1R(1A), and twelve F_1 plants were grown from the combination L-57(PF) × 1R(1D). All F_1 plants were fertile, which indicates that fertility restoration in these allolines is a dominant trait. The analysis of the seed set in the main spike of 74 F_2 plants of the hybrid combination L-57(PF) × 1R(1A) revealed 51 plants that were classified as fertile and 23 plants – as sterile. The observed ratio when

Table 4. Seed setting in F_1 hybrids and segregation for seed setting in F_2 hybrids derived from the crossing of allolines (*H. vulgare*)-*T. aestivum* with wheat-rye substitution lines 1R(1A) and 1R(1D) of variety Saratovskaya 29

Hybrid combination	Generation	Plants			Expected segregation ratio in F_2	χ^2	*p*-value
		Total number	Fertile	Sterile			
L-56(FS) × 1R(1A)	F_1	18	0	18	–		
L-56(FS) × 1R(1D)	F_1	20	0	20	–		
L-56(PS) × 1R(1A)	F_1	15	0	15	–		
L-56(PS) × 1R(1D)	F_1	17	0	17	–		
L-57(PF) × 1R(1A)	F_1	15	15	–			
	F_2	74	51	23	3:1	1.46	0.227
L-57(PF) × 1R(1D)	F_1	12	12	–			
	F_2	61	45	16	3:1	0.05	0.824
L-58($F_{homoplasmy}$) × 1R(1A)	F_1	14	14	–			
	F_2	75	75	–			
L-58($F_{homoplasmy}$) × 1R(1D)	F_1	15	15	–			
	F_2	86	86	–			

segregated into fertile and sterile plants in F_2 fitted well with the theoretically expected segregation ratio of 3 (fertile) : 1 (sterile) with an χ^2 value of 1.46, which is lower than the statistical value of $\chi^2_{05} = 3.84$.

A similar result was obtained for the hybrid combination L-57(PF) × 1R(1D). Out of the 61 F_2 plants studied in this combination, 45 were classified as fertile and 16 were sterile, resulting in a value of $\chi^2 = 0.05$ (Table 4). These results indicated that the fertility of L-57 was dependent on wheat chromosomes 1A and 1D. The ratio of fertile and sterile plants in F_2 of the combinations L-57(PF) × 1R(1A) and L-57(PF) × 1R(1D) showed that male fertility restoration in partially fertile plants of L-57 was controlled by a single dominant gene. One of these genes is localized on chromosome 1A, and the other, on chromosome 1D.

A different result was obtained when crossing fully fertile plants of L-58, in which wheat-type homoplasmy was detected, with wheat-rye substitution lines. All F_1 and F_2 plants of the combinations L-58(F) × 1R(1A) and L-58(F) × 1R(1D) were fertile (Table 4). This means that the fertility of alloline L-58 included in crosses with wheat-rye substitution lines does not depend on bread wheat chromosomes 1A and 1D.

Discussion

There is a strong intergenomic incompatibility between cultivated barley *H. vulgare* and bread wheat *T. aestivum*, which prevents both crossing between them and fertility restoration of hybrids. However, due to the use of methods to overcome incompatibility and the selection of parental genotypes, it was possible to obtain viable barley-wheat F_1 hybrids with female fertility (Pershina et al., 1998). This made it possible to include hybrids in backcrosses with different varieties of bread wheat leading to the elimination of barley chromosomes, the creation of a recombinant wheat nuclear genome and the replacement of a wheat cytoplasm with the cytoplasm of barley in alloplasmic genotypes (Aksyonova et al., 2005; Pershina et al., 2012).

The recombinant allolines (*H. vulgare*)-*T. aestivum* L-56, L-57 or L-58 had the same origin, but differed in morphobiological characteristics and fertility level. The recombinant nuclear genome of these lines was obtained using the varieties of bread wheat Saratovskaya 29, Mironovskaya 808, and Pyrotrix 28. The expression of morphobiological traits in the L-56 line, compared to the L-57 line, represented by partially fertile plants, was suppressed. The L-56 alloline segregated into fully sterile plants and plants with a low fertility level. Apparently, the genome of Saratovskaya 29 prevailed in the nuclear genome of the L-56 line. This variety was a fixer of sterility of bread wheat carrying the cytoplasm of cultivated barley (Pershina et al., 2012).

The absence of anthers in fully sterile plants and incomplete staining of pollen grains in partially fertile plants was caused by CMS, which resulted from disruption of nuclear-mitochondrial interactions (Yang et al., 2008). PCR analysis of the 18S/5S mt repeat in the L-56 and L-57 allolines revealed heteroplasmy, that is, the coexistence of two mtDNA variants, the barley and the wheat type. Heteroplasmy of mtDNA in barley-wheat hybrids and allolines derived from them is a consequence of biparental transmission of mtDNA beginning from F_1 (Aksyonova et al., 2005). This phenomenon has been described for hybrids (*Ae. crassa* × wheat Chinese Spring) (Kawaura et al., 2011)

and allolines (*Ae. longissima*)-*T. turgidum* (Noyszewski et al., 2014). Inheritance of cytoplasmic genomes from both parents, compared with strictly maternal one, results in a greater diversity of mt- and cpDNA variants in hybrids. It has been suggested that biparental inheritance of chloroplasts in angiosperms leads to rescue species with defective plastids (Zhang, Sodmergen, 2010). This mechanism can also reduce the negative impact of cytonuclear incompatibility on the development of F_1 hybrids (Barnard-Kubow et al., 2016). It can be assumed that in the L-56 and L-57 allolines, the presence of wheat copies of mtDNA, along with barley copies, was also a manifestation of neutralization of the cytonuclear conflict between barley cytoplasm and wheat nuclear genome, ensuring the development of viable allolines, albeit with reduced fertility.

When backcrossing hybrids with mtDNA heteroplasmy with the paternal species (wheat), variability was found not only in the nuclear genome, but also in the mitochondrial genome (Aksyonova et al., 2005; Trubacheeva et al., 2012, 2021). When the fertility of the allolines was restored, the number of mtDNA copies of the wheat (paternal) type increased and the original alloplasmic condition appeared to be lost (Aksyonova et al., 2005; Trubacheeva et al., 2021). The same process was observed in the production of wheat allolines carrying the cytoplasm of some *Aegilops* species (Tsukamoto et al., 2000; Hattori et al., 2002). The fully fertile L-58 alloline without CMS was isolated by selecting plants with maximum fertility in the F_5BC_5 generation of the barley-wheat hybrid (Fig. 1). It can be assumed that in L-58, the recombinant nuclear genome without barley chromosomes contains mainly the genomes of the wheat varieties Mironovskaya 808 and Pirotrix 28, which are restorers of fertility in bread wheat with the cytoplasm of cultivated barley (Pershina et al., 2012, 2018). During selection for fertility, as well as during backcrossing, the variability of mtDNA from heteroplasmy to wheat-type homoplasmy correlates with the variability of chloroplast DNA from barley-type homoplasmy to wheat-type homoplasmy (Aksyonova et al., 2005; Trubacheeva et al., 2021).

As follows from the data obtained both in this work and in previously published ones (Aksyonova et al., 2005; Trubacheeva et al., 2012, 2021), heteroplasmy and wheat-type homoplasmy detected in allolines can be used as markers for dividing allolines into groups with cytonuclear incompatibility and cytonuclear coadaptation, since it is not in all cases that the fertility level can be a reliable trait for such a division. For example, in both this and a previously published work (Trubacheeva et al., 2021), mtDNA heteroplasmy was found in some fertile plants, that is, there was a violation of cytonuclear compatibility.

Clear differences between allolines with cytonuclear incompatibility and cytonuclear coadaptation were found when studying the effect of chromosomes 1A and 1D on the fertility of these lines. In allolines L-56 and L-57 with cytonuclear incompatibility, male fertility depends on these wheat chromosomes, but in alloline L-58 without cytonuclear incompatibility, it does not. This can be explained by the fact that in allolines L-56 and L-57 with heteroplasmy, the *Rf* genes located on chromosomes 1A and 1D are necessary to neutralize the sterilizing effect of the cytoplasm. In line L-58 with cytonuclear coadaptation, wheat-type cytoplasm was formed, so the production of male-fertile plants did not depend on the presence of the *Rf* genes on these chromosomes.

We observed similar differences in our previous work (Trubacheeva et al., 2021): the short arm of chromosome 1B affected the fertility of the allolines with cytonuclear incompatibility, but did not affect the fertility of the allolines with cytonuclear compatibility.

Conclusion

To perform this work, among the backcrossed progenies of the barley-wheat hybrid *H. vulgare* × *T. aestivum*, sequentially pollinated with different varieties of bread wheat, three allo-lines of bread wheat with the cytoplasm of cultivated barley were isolated. These allolines of the same origin but differed by fertility and cytonuclear compatibility were used as adequate models to determine the localization of genes controlling the restoration of fertility of bread wheat carrying the cytoplasm of cultivated barley.

Based on the results of segregation in F_2 hybrids obtained from crossing alloline L-57 with wheat-rye substitution lines 1R(1A) and 1R(1D), it was concluded for the first time that chromosomes 1A and 1D carry one dominant *Rf* gene, which controls male fertility restoration of bread wheat with the cytoplasm of cultivated barley. However, a single dose of these genes is not enough to restore the fertility of partially sterile plants. The results of our work supplemented the information on the localization of the *Rf* genes in wheat chromosomes 1A, 1D (this work) and 1BS (Trubacheeva et al., 2021).

An important finding was that the fertility of the line with cytonuclear compatibility did not depend on the chromosomes in which the *Rf* genes were located. This explains the fact that the introgression of alien germplasm into the lines, including the replacement of the short arm of wheat chromosome 1B by the short arm of the rye chromosome 1R, does not violate cytonuclear compatibility, and allolines maintain fertility (Pershina et al., 2018, 2020). Moreover, based on introgression (*H. vulgare*)-*T. aestivum* allolines, DH lines were obtained and used as maternal genotypes to develop commercial high-yielding spring wheat varieties Sigma, Uralosibirskaya 2, Sigma 5 (Belan et al., 2021).

References

Aksyonova E., Sinyavskaya M., Danilenko N., Pershina L., Nakamura C., Davydenko O. Heteroplasmy and paternally oriented shift of the organellar DNA composition in barley-wheat hybrids during backcrosses with wheat parents. *Genome.* 2005;48(5):761-769. DOI 10.1139/g05-049

Atienza S.G., Martin A., Peechioni N., Platani C., Cattivelli L. The nuclear-cytoplasmic interaction controls carotenoid content in wheat. *Euphytica.* 2008;159:325-331. DOI 10.1007/s10681-007-9511-6

Badaeva E.D., Pershina L.A., Bil'danova L.L. Cytogenetic analysis of alloplasmic recombinant lines (*H. vulgare*)-*T. aestivum* unstable in fertility and viability. *Russ. J. Genet*. 2006;42(2):140-149. DOI 10.1134/S1022795406020074

Barnard-Kubow K.B., McCoy M.A., Galloway L.F. Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. *New Phytol*. 2016;213(3):1466-1476. DOI 10.1111/nph.14222

Belan I.A., Rosseeva L.P., Blokhina N.P., Mukhina Y.V., Trubacheeva N.V., Pershina L.A. The use of double haploid lines is an acceleration of breeding process in creating varieties of spring bread wheat. In: Abstracts from the Int. Conf. "Advanced Technologies in Agricultural Production: People, digital, environment (AgroProd 2021)". Omsk, 2021;128-133 (in Russian)

Bohra A., Jha U.C., Adhimoolam P., Bisht D., Sing N.P. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. *Plant Cell Rep*. 2016;35(5):967-993. DOI 10.1007/s00299-016-1949-3

Buloychik A.A., Voluevich E.A., Mikhno A.M. Genome and plasmon effects on expression of the defeated genes of resistance to brown rust in wheat. *Tsitologiya i Genetika = Cytology and Genetics*. 2002; 36(2):11-19 (in Russian)

Coulthart M.B., Spencer D.F., Gray M.W. Comparative analysis of a recombining-repeat-sequence family in the mitochondrial genomes of wheat (*Triticum aestivum* L.) and rye (*Secale cereale* L.). *Curr. Genet.* 1993;23(3):255-264. DOI 10.1007/BF00351504

Crosatti C., Quansah L., Maré C., Giusti L., Roncaglia E., Atienz S.G., Cattivelli L., Fait A. Cytoplasmic genome substitution in wheat affects the nuclear-cytoplasmic cross-talk leading to transcript and metabolite alterations. *BMC Genomics*. 2013;14:868-889. DOI 10.1186/1471-2164-14-868

Current Protocols in Molecular Biology. Greene Publishing Associates. N.Y.: Wiley Interscience, 1987

Gupta P.K., Balyan H.S., Gahlaut V., Saripalli G., Pal B., Basnet B.R., Joshi A.K. Hybrid wheat: past, present and future. *Theor. Appl. Genet*. 2019;132(9):2463-2483. DOI 10.1007/s00122-019-03397-y

Hattori N., Kitagawa K., Takumi S., Nakamura C. Mitochondrial DNA heteroplasmy in wheat, *Aegilops* and their nucleus-cytoplasm hybrids. *Genetics*. 2002;160(4):1619-1630. DOI 10.1093/genetics/160.4.1619

Hohn C.E., Lukaszewski A.J. Engineering the 1BS chromosome arm in wheat to remove the Rf^{multi} locus restoring male fertility in cytoplasms of *Aegilops kotschyi*, *Ae. uniaristata* and *Ae. mutica*. *Theor. Appl. Genet.* 2016;129(9):1769-1774. DOI 10.1007/s00122-016-2738-7

Islam M.S., Studer B., Møller I.M., Asp T. Genetics and biology of cytoplasmic male sterility and its applications in forage and turf grass breeding. *Plant Breed*. 2014;133(3):299-312. DOI 10.1111/pbr.12155

Kawaura K., Saeki A., Masumura T., Morita S., Ogihara Y. Heteroplasmy and expression of mitochondrial genes in alloplasmic and euplasmic wheat. *Genes Genet. Syst.* 2011;86(4):249-255. DOI 10.1266/ggs.86.249

Klimushina M.V., Divashuk M.G., Mokhammed T.A.K., Semenov O.G., Karlov G.I. Analysis of allelic state of genes responsible for baking properties in allocytoplasmic wheat hybrids. *Russ. J. Genet.* 2013; 49(5):530-538. DOI 10.1134/S1022795413050074

Liu C.G., Wu Y.W., Hou H., Zhang C., Zhang Y., McIntosh R.A. Value and utilization of alloplasmic common wheats with *Aegilops crassa* cytoplasm. *Plant Breed*. 2002;121(5):407-410. DOI 10.1046/j.1439-0523.2002.755374.x

Liu Y., Tang L., Xu Q., Ma D., Zha M., Sun J., Chen W. Experimental and genomic evidence for the *indica*-type cytoplasmic effect in *Oryza sativa* L. ssp. *japonica*. *J. Integr. Agric.* 2016;15(10):2183-2191. DOI 10.1016/S2095-3119(15)61190-X

Martin A.C., Atienza S.G., Ramírez M.C., Barro F., Martín A. Molecular and cytological characterization of an extra acrocentric chromosome that restores male fertility of wheat in the msH1 CMS system. *Theor. Appl. Genet.* 2010;121(6):1093-1101. DOI 10.1007/s00122-010-1374-x

Noyszewski A.K., Ghavami F., Alnemer L.M., Soltani A., Gu Y.Q., Huo N., Meinhardt S., Penny M.A., Kianian P.M.A., Kianian S.F. Accelerated evolution of the mitochondrial genome in an alloplasmic line of durum wheat. *BMC Genomics*. 2014;15(1):67. DOI 1471-2164/15/67

Pershina L.A., Numerova O.M., Belova L.I., Devyatkina E.P. Biotechnological and cytogenetic aspects of producing new wheat genotypes using hybrids. *Euphytica*. 1998;100(1-3):239-244. DOI 10.1023/A:1018312408312

Pershina L.A., Devyatkina E.P., Trubacheeva N.V., Kravtsova L.A., Dobrovol'skaya O.B. Characterization of fertility restoration in alloplasmic lines derived from hybridization of self-fertilized of spring of barley–wheat (*Hordeum vulgare* L. × *Triticum aestivum* L.) amphiploid with common wheat varieties Saratovskaya 29 and Pyrotrix 28. *Russ. J. Genet.* 2012;48(12):1184-1190. DOI 10.1134/S1022795412120101

Pershina L.A., Belova L.I., Trubacheeva N.V., Osadchaya T.S., Shumny V.K., Belan I.A., Rosseeva L.P., Nemchenko V.V., Abakumov S.N. Alloplasmic recombinant lines (*H. vulgare*)-*T. aestivum* with 1RS.1BL translocation: initial genotypes for production of common wheat varieties. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018;22(5):544-552. DOI 10.18699/VJ18.393]

Pershina L., Trubacheeva N., Badaeva E., Belan I., Rosseeva L. Study of androgenic plant families of alloplasmic introgression lines (*H. vulgare*)–*T. aestivum* and the use of sister DH lines in breeding. *Plants*. 2020;9(6):764-816. DOI 10.3390/plants9060764

Rafee V.V., Lalitha R., Remadevi A.K., Lekshmi M., Premachandran M.N. Substitution of cytoplasm of sugarcane with that of the wild grass *Erianthus arundinaceus*. *Gregor Mendel Found. J*. 2010; 1(1&2):16-22

Shchapova A.I., Kravtzova L.A. The production of wheat-rye substitution line by using the Giemsa staining technique. *Cereal Res. Commun*. 1982;10(1):33-39

Sinha P., Tomar S.M., Vinod, Singh V.K., Balyan H.S. Genetic analysis and molecular mapping of a new fertility restorer gene *Rf8* for *Triticum timopheevi* cytoplasm in wheat (*Triticum aestivum* L.) using SSR markers. *Genetica.* 2013;141(10-12):431-441. DOI 10.1007/s10709-013-9742-5

Soltani A., Kumar A., Mergoum M., Pirseyedi S.M., Hegstad J.B., Mazaheri M., Kianian S.F. Novel nuclear-cytoplasmic interaction in wheat (*Triticum aestivum*) induces vigorous plants. *Funct. Integr. Genom.* 2016;16(2):171-182. DOI 10.1007/s10142-016-0475-2

Takenaka S., Yamamoto R., Nakamura C. Differential and interactive effects of cytoplasmic substitution and seed aging on submergence stress response in wheat (*Triticum aestivum* L.). *Biotechnol. Biotechnol. Equip*. 2019;33(1):75-85. DOI 10.1080/13102818.2018.1549960

Talukder S.K., Vara Prasad P.V., Todd T., Babar M.A., Poland J., Bowden R., Fritz A. Effect of cytoplasmic diversity on post anthesis heat tolerance in wheat. *Euphytica*. 2015;204:383-394. DOI 10.1007/s10681-014-1350-7

Tao D., Xu P., Zhou J., Deng X., Li J., Deng W., Yang J., Yang G., Li Q., Hu F. Cytoplasm affects grain weight and filled-grain ratio in *indica* rice. *BMC Genet.* 2011;12:53. DOI 10.1186/1471-2156-12-53

Trubacheeva N.V., Kravtsova L.A., Devyatkina E.P., Efremova T.T., Sinyavskaya M.G., Shumny V.K., Pershina L.A. Heteroplasmic and

homoplasmic states of mitochondrial and chloroplast DNA regions in progenies of distant common wheat hybrids of different origins. *Russ. J. Genet. Appl. Res.* 2012;2(6):494-500. DOI 10.1134/S2079059712060147

Trubacheeva N.V., Divashuk M.G., Chernook A.G., Belan I.A., Rosseeva L.P., Pershina L.A. The effect of chromosome arm 1BS on the fertility of alloplasmic recombinant lines in bread wheat with the *Hordeum vulgare* cytoplasm. *Plants.* 2021;10(6):1120. DOI 10.3390/plants10061120

Tsukamoto N., Asakura N., Hattori N., Takumi S., Mori N., Nakamura C. Identification of paternal mitochondrial DNA sequences in the nucleus-cytoplasm hybrid of tetraploidand hexaploid wheat with D and D2 plasmon from *Aegilops* species. *Curr. Genet.* 2000; 38(4):208-217. DOI 10.1007/s002940000153

Tsunewaki K. Plasmon analysis as the counterpart of genome analysis. In: Jauhar P.P. (Ed.) Methods of Genome Analysis in Plants. Boca Raton: CRC Press, 1996;271-299. http://pi.lib.uchicago.edu/1001/cat/bib/2607056

Tsunewaki K. Fine mapping of the first multi-fertility-restoring gene, Rf^{multi}, of wheat for three *Aegilops* plasmons, using 1BS-1RS recombinant lines. *Theor. Appl. Genet.* 2015;128(4):723-732. DOI 10.1007/s00122-015-2467-3

Yang J., Zhang M., Yu J. Mitochondrial retrograde regulation tuning fork in nuclear genes expressions of higher plants. *J. Genet. Genomics*. 2008;35(2):65-71. DOI 10.1016/S1673-8527(08)60010-7

Zhang Q., Sodmergen. Why does biparental plastid inheritance revive in angiosperms? *J. Plant Res.* 2010;123(2):201-206. DOI 10.1007/s10265-009-0291-z

Conflict of interest. The authors declare no conflict of interest.
Received March 25, 2024. Revised May 3, 2024. Accepted May 14, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):619-627

DOI 10.18699/vjgb-24-69

Metabolite concentrations and the expression profiles of the corresponding metabolic pathway genes in eggplant (*Solanum melongena* L.) fruits of contrasting colors

M.A. Filyushin [1], E.A. Dzhos [1, 2], A.V. Shchennikova [1], E.Z. Kochieva [1]

[1] Federal Research Centre "Fundamentals of Biotechnology" of the Russian Academy of Sciences, Moscow, Russia
[2] Federal Scientific Vegetable Center, VNIISSOK village, Moscow region, Russia
michel7753@mail.ru

Abstract. Eggplant (*Solanum melongena* L.) ranks fifth in importance among vegetable crops of the Solanaceae family, in part due to the high antioxidant properties and polyphenol content of the fruit. Along with the popular purple-fruited varieties of *S. melongena*, there are cultivars, the fruits of which are rich in phenolic compounds, but are white-colored due to the lack of anthocyanin biosynthesis. Determination of the amount of anthocyanins and other phenolic compounds, as well as carotenoids and sugars, is included in the assessment of the quality of eggplant fruits of commercial (technical) ripeness. In addition to antioxidant and taste properties, these metabolites are associated with fruit resistance to various stress factors. In this study, a comparative analysis of the content of anthocyanins, carotenoids and soluble sugars (sucrose, glucose, fructose) in the peel and pulp of the fruit of both technical and biological ripeness was carried out in purple-fruited (cv. Vlas) and white-fruited (cv. Snezhny) eggplant accessions of domestic selection. The peel and pulp of biologically ripe fruits of the cvs Vlas and Snezhny were used for comparative transcriptomic analysis. The key genes of the flavonoid and carotenoid metabolism, sucrose hydrolysis, and soluble sugar transport were shown to be differentially expressed between fruit tissues, both within each cultivar and between them. It has been confirmed that the purple color of the peel of the cv. Vlas fruit is due to substantial amounts of anthocyanins. Flavonoid biosynthesis genes showed a significantly lower expression level in the ripe fruit of the cv. Vlas in comparison with the cv. Snezhny. However, in both cultivars, transcripts of anthocyanin biosynthesis genes (*DFR*, *ANS*, *UFGT*) were not detected. Additionally, the purple fruit of the cv. Vlas accumulated more carotenoids and sucrose and less glucose and fructose than the white fruit of the cv. Snezhny. Biochemical data corresponded to the differential expression pattern of the key genes encoding the structural proteins of metabolism and transport of the compounds analyzed.
Key words: eggplant cultivars; *Solanum melongena* L.; carotenoids; anthocyanins; soluble sugars; expression of metabolic pathway genes.

For citation: Filyushin M.A., Dzhos E.A., Shchennikova A.V., Kochieva E.Z. Metabolite concentrations and the expression profiles of the corresponding metabolic pathway genes in eggplant (*Solanum melongena* L.) fruits of contrasting colors. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):619-627. DOI 10.18699/vjgb-24-69

Funding. This research was funded by the Ministry of Science and Higher Education of the Russian Federation in accordance with agreement № 075-15-2022-318 on 20 April 2022 on providing a grant in the form of subsidies from the Federal Budget of the Russian Federation. The grant was provided as state support for the creation and development of a World-class Scientific Center "Agrotechnologies for the Future".

Содержание метаболитов и профиль экспрессии генов соответствующих метаболических путей в контрастных по окраске плодах баклажана (*Solanum melongena* L.)

М.А. Филюшин [1], Е.А. Джос [1, 2], А.В. Щенникова [1], Е.З. Кочиева [1]

[1] Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Россия
[2] Федеральный научный центр овощеводства, пос. ВНИИССОК, Московская область, Россия
michel7753@mail.ru

Аннотация. Баклажан (*Solanum melongena* L.) занимает пятое место по значимости среди овощных культур семейства Пасленовых, в том числе благодаря антиоксидантным свойствам плода за счет высокого содержания различных фенольных соединений. Наряду с популярными фиолетовоплодными сортами *S. melongena* имеются сорта, плоды которых синтезируют фенольные соединения, однако характеризуются белой окраской из-за от-

сутствия биосинтеза антоцианов. Определение количества антоцианов и других фенольных соединений, а также каротиноидов и сахаров входит в оценку качества плодов баклажана коммерческой (технической) спелости. Кроме антиоксидантных и вкусовых качеств, эти метаболиты связаны с устойчивостью плода к различным стрессовым факторам. В данном исследовании проведен сравнительный анализ содержания антоцианов, каротиноидов и растворимых сахаров (сахарозы, глюкозы, фруктозы) в кожице и мякоти плода как технической, так и биологической спелости у фиолетовоплодного (сорт Влас) и белоплодного (сорт Снежный) образцов баклажана отечественной селекции. Кожица и мякоть плода биологической спелости сортов Влас и Снежный были использованы для сравнительного транскриптомного анализа. Показано, что ключевые гены флавоноидного пути, метаболизма каротиноидов, гидролиза сахарозы, а также транспорта растворимых сахаров дифференциально экспрессируются между тканями плода как внутри каждого сорта, так и между сортами. Подтверждена связь фиолетовой окраски кожицы плода сорта Влас с присутствием значительных количеств антоцианов. Определено, что в сравнении с сортом Снежный спелый плод сорта Влас характеризуется существенно более низким уровнем экспрессии генов биосинтеза флавоноидов. Однако у обоих сортов в спелом плоде не выявлены транскрипты генов биосинтеза антоцианов (*DFR*, *ANS*, *UFGT*). Также показано, что в сравнении с белым плодом сорта Снежный фиолетовый плод сорта Влас накапливает больше каротиноидов и сахарозы и меньше глюкозы и фруктозы. Биохимические данные соответствуют профилю дифференциальной экспрессии ключевых генов, кодирующих структурные белки метаболизма и транспорта анализируемых соединений.
Ключевые слова: сорта баклажана; *Solanum melongena* L.; каротиноиды; антоцианы; растворимые сахара; экспрессия генов метаболических путей.

Introduction

Eggplant (*Solanum melongena* L.) is a vegetable crop that ranks fifth in economic importance in the nightshade family (Solanaceae). Despite its heat-loving nature, this crop is grown not only in tropical and subtropical climate zones, but also as a greenhouse crop in regions with cold climate (including the Russian Federation). The most famous are eggplant fruits with peel colored in different shades of purple, which is determined by the content of anthocyanins. The presence of anthocyanins and the fact that the fruit pulp is enriched with phenolic acids indicate powerful antioxidant properties of the eggplant fruit, classifying it as a product with high nutritional/dietary value (Gürbüz et al., 2018; Akhbari et al., 2019; Condurache et al., 2021; Saha et al., 2023).

In addition to purple-fruited varieties, there are also *S. melongena* varieties that produce fruits with white or green peel due to inhibition of anthocyanin biosynthesis (Condurache et al., 2021; Yang et al., 2022; You et al., 2022). The color (white, green, or intermediate shades) is determined by the ratio of two types of plastids in the cells of the fruit – chloroplasts and leucoplasts (Tao et al., 2023). For consumers, white-fruited varieties may be preferable because they lack the bitterness associated with dark-colored fruits due to changes in the content of glycoalkaloids (Lelario et al., 2019; Saha et al., 2023).

Commercial eggplant varieties are characterized by morphological variability, and screening of existing collections for a set of characteristics includes grouping by fruit peel color as the most important trait (Martínez-Ispizua et al., 2021). The assessment of fruit quality focuses on their antioxidant properties (including the determination of phenolic compounds/flavonoids, carotenoids and sugars), and there is wide variation in these terms (Martínez-Ispizua et al., 2021). Purple-fruited varieties, compared to white-fruited varieties, are characterized by greater antioxidant activity and increased content of phenols and carotenoids (both in the peel and in the pulp), and there is little or no difference in the total amount of sugars (Martínez-Ispizua et al., 2021; Colak et al., 2022).

There is no correlation between the content of flavonoids, carotenoids and sugars in eggplant fruits (Martínez-Ispizua et al., 2021). On the other hand, there is indirect evidence for the existence of such a phytohormone-mediated dependence in cherries (Teribia et al., 2016). Namely, there is an inverse correlation between the content of soluble sugars and transzeatin, as well as gibberellin GA4 and anthocyanins; in contrast, abscisic acid (ABA) is positively associated with the amount of anthocyanins and soluble sugars (Teribia et al., 2016). Moreover, the accumulation of anthocyanins is positively correlated with the amount of sugars in the Chinese date *Ziziphus jujube* (Jiang et al., 2020).

All of the antioxidant compounds mentioned, as well as soluble sugars, are closely related to resistance to various stress factors both in the vegetative part of the plant (Keunen et al., 2013; Pérez-Torres et al., 2021; Waadt et al., 2022) and in the fleshy fruit (Shi et al., 2019; Jiang et al., 2020). For example, it has been shown that increased production of phenolic compounds determines the resistance of the eggplant fruit to low temperatures (Shi et al., 2019). Elevated temperature has a positive effect on the content of sugars, anthocyanins, flavonoids and carotenoids in the fruits of the Chinese date *Z. jujube*, but in combination with drought it causes the opposite effect (Jiang et al., 2020).

This work aimed to characterize the fruits of two eggplant varieties, including the determination of the content of anthocyanins, carotenoids and soluble sugars, as well as the expression profile of key genes in the corresponding metabolic pathways. We chose two cultivars of domestic selection that have different fruit colors – white and purple, respectively. A significant difference from similar studies was that we analyzed fruits of not only technical (commercial) ripeness, but also those of biological ripeness.

The following tasks were set: to obtain plant material (fruits of two varieties at the stages of technical and biological ripeness); to determine the content of target metabolites in the peel and pulp of fruits of technical and biological ripeness; to analyze transcriptomes of the peel and pulp of fruits at the stage of biological ripeness, focusing on transcripts of genes of target metabolic pathways; to validate transcriptomic data.

Materials and methods

In a comparative study, we used accessions of two early-ripening eggplant varieties (*S. melongena*), originated by the Federal Scientific Vegetable Center (FSVC, Moscow region) that differed in the color of the ripe fruit. The fruits of cv. Snezhny (ID 9905014, https://gossortrf.ru/registry/) at the stage of technical ripeness have white peel and pulp. The fruits of cv. Vlas (ID 8057522) at technical ripeness have dark purple peel and white flesh. At the stage of biological ripeness, the fruit pulp remains white in both cultivars, and the peel acquires yellowish (cv. Snezhny) or brown (cv. Vlas) shades (Fig. 1).

Plants of the studied varieties were grown (2023) until the fruiting stage in a film greenhouse of the Federal Research Vegetable Center. In August, fruits were collected at technical (commercially mature, CM) and biological (physiologically ripe, PR) ripeness, separated into peel (exocarp) and pulp (mesocarp), grinded in a porcelain mortar in liquid nitrogen and used for biochemical, metabolomic and transcriptomic analyses.

The content of anthocyanins and carotenoids was determined spectrophotometrically in chloroform-methanol extracts according to (Filyushin et al., 2020). Since delphinidin glycosides (93–98 % of the total) dominate among the anthocyanins accumulated in the peel of eggplant fruit (Condurache et al., 2021; Yang et al., 2022), the anthocyanin content was calculated in terms of delphinidin-3-rutinoside.

The content of sugars (glucose, fructose and sucrose) was determined according to metabolome data (unpublished), which were obtained according to (Filyushin et al., 2023a). In short, approximately 0.2 g of finely ground tissue was extracted twice with 200 µl of 80 % methanol. The total extract was evaporated, dissolved in 30 % methanol (at the rate of 50 mg wet weight per 100 µl of the extract) and subjected to mass spectral analysis using ultra-performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UPLC-qTOF-MS/MS) according to the protocol [https://lcms.cz/labrulez-bucket-strapi-h3hsga3/1866243_lcms_148_how_potato_fights_its_enemies_02_2019_ebook_rev_01_9d3990d6c4/1866243-lcms-148-how-potato-fights-its-enemies-02-2019-ebook-rev-01.pdf]. The signal level/100 mg of annotated compounds was used as a relative indicator for sugar content.

Differentially expressed genes (DEGs) encoding proteins involved in sucrose hydrolysis and transport of soluble sugars (invertases and sugar uniporters) were determined from transcriptome data for the peel and pulp of the PR fruit (unpublished). For transcriptomic analysis, preparations of total RNA were isolated (RNeasy Plant Mini Kit, Qiagen, USA) and used for mRNA libraries (NEBNext® mRNA Library Prep Reagent Set for Illumina; New England BioLabs, USA), which were then sequenced (Illumina HiSeq2500; Illumina Inc., USA). Trinity v3.5.13 (https://github.com/trinityrnaseq/trinityrnaseq/wiki) and TransDecoder v5.1.0 (https://github.com/TransDecoder/TransDecoder) were used to assemble and determine coding sequences; CDSs were annotated using NCBI-Blast (https://www.ncbi.nlm.nih.gov/). Relative transcript levels (FPKM; number of fragments per kb transcripts per million mapped fragments) were estimated using RSEM

Fig. 1. Photographs of the fruit of the eggplant cultivars Snezhny (*a*) and Vlas (*b*) in technical (CM; left) and biological (PR; right) ripeness.

The cultivars differ in the color of the fruit peel – white (cv. Snezhny) and purple (cv. Vlas). Scale = 5 cm.

(https://github.com/deweylab/RSEM). To determine DEGs both within varieties (peel *vs.* pulp) and between varieties (peel *vs.* peel; pulp *vs.* pulp), transcriptome data were normalized to the transcript number of the reference gene *GAPDH*.

Structural analysis of DEGs was performed using NCBI-BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and MEGA 7.0 (https://www.megasoftware.net/) using genomic (GCA_000787875.1) (Hirakawa et al., 2014) and transcriptomic (https://www.ncbi.nlm.nih.gov/) *S. melongena* data.

Transcriptomic data were validated using quantitative real-time PCR (qRT-PCR) and a CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories, USA); qRT-PCR program [95 °C – 5 min.; 40 cycles (95 °C – 15 s, 62 °C – 50 s)]. Based on available total RNA preparations, cDNA was synthesized (GoScript™ Reverse Transcription System, Promega, USA) and 3 ng was used in the reaction. The reaction mixture included the "Reaction mixture for qRT-PCR in the presence of SYBR GreenI and ROX" (Sintol LLC, Russia) and gene-specific primers. Reactions were performed in three technical and two biological replicates and normalized to the transcript level of the reference gene *GAPDH* (Zhang et al., 2014).

The obtained biochemical and expression data were statistically processed in GraphPad Prism v.8 (GraphPad Software Inc., USA; https://www.graphpad.com/scientific-software/prism/). To assess the significance of the differences, a *t*-test was used ($p < 0.05$ indicates statistical significance of the differences).

Results

The study was focused on the comparative characteristics of the fruit (technical and biological ripeness) of two eggplant varieties belonging to the same species, *S. melongena*, and differing in the color of the fruit peel. Namely, cv. Snezhny and cv. Vlas with white/yellowish and purple/brown, respectively, colors of the fruit peel at technical/biological ripeness (Fig. 1).

A biochemical analysis of the peel and pulp in the dynamics of fruit ripening showed that the content of anthocyanins corresponds to the color of the analyzed fruit tissues of biological ripeness. In the yellowish peel and white pulp of the fruit of cv. Snezhny, as well as the white pulp of the fruit of cv. Vlas, the amount of anthocyanins showed trace values, while in the purple-brown peel of the fruit of cv. Vlas anthocyanins amount was ~300 times higher (Fig. 2*a*).

The fruits of both varieties, both in technical and biological ripeness, contained traces of carotenoids in the pulp. In the peel, carotenoids accumulated more actively: in cv. Vlas, the amount of carotenoids was ~25 times higher than in cv. Snezhny (Fig. 2*b*).

If the differences in the content of anthocyanins in the analyzed varieties were predictable, then the significant differences in the content of soluble sugars were somewhat unexpected. According to metabolomic profiling of the peel and pulp, it was found that the fruit of cv. Snezhny contains ~2 (peel) and ~5 (pulp) times more hexoses (glucose, fructose), as well as ~2 (peel and pulp) times less sucrose than the fruit of cv. Vlas (Fig. 2*c*).

A comparative analysis of the transcriptomes of the peel and pulp of the fruit of the Snezhny and Vlas varieties identified a number of DEGs, which, as expected (according to the results of biochemical and metabolomic analyses), included genes associated with the metabolism of anthocyanins, carotenoids and sugars (see the Table).

It was found that the key genes of the flavonoid pathway (Zhang et al., 2014; Alappat B., Alappat J., 2020) before anthocyanin synthesis (*CHS1*, *CHS2*, *F3H*) are highly transcribed in the peel of the fruit of cv. Snezhny and are detected in significantly smaller and similar quantities in the pulp (both varieties) and peel (cv. Vlas) (Fig. 3). Considering the branch of the pathway related to anthocyanin synthesis, the level of expression of the first gene of the branch, *DFR*, in the peel and pulp of the fruit of cv. Snezhny is significantly higher than that of cv. Vlas. However, the number of transcripts in the FPKM value for *DFR* is extremely low in all four samples (0.49–3.44), so we cannot speak of a significant difference between varieties, since the level of gene transcripts approaches zero. At the same time, transcripts of subsequent genes of the anthocyanin biosynthesis branch – *ANS* (anthocyanidin synthase) and *UFGT* (UDP-glucosoflavonoid-3-O-glucosyltransferase) – were not included in the list of DEGs and were detected in trace amounts (Fig. 3).

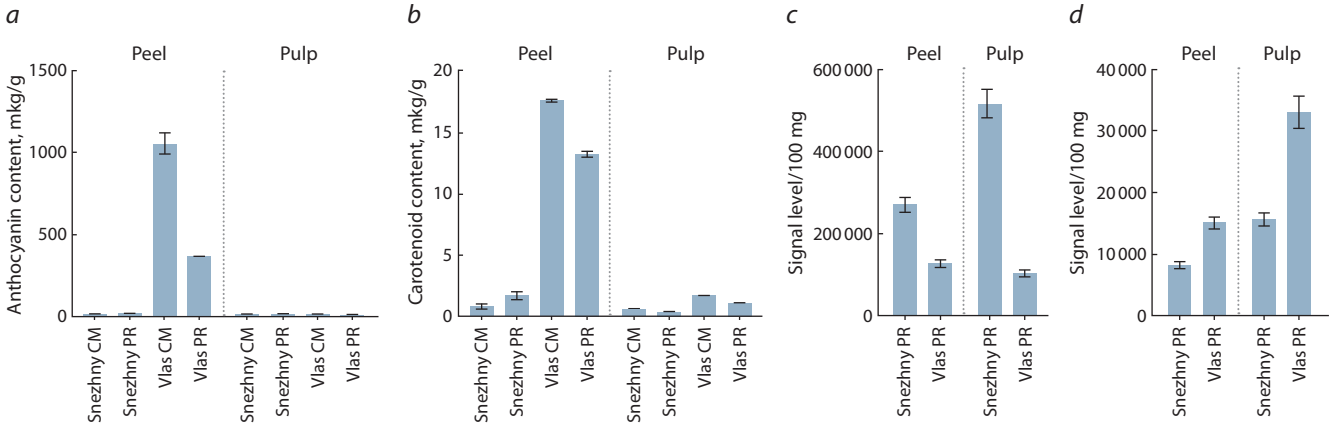


Fig. 2. The content of the sum of anthocyanins (*a*), the sum of carotenoids (*b*), hexoses (total glucose and fructose) (*c*) and sucrose (*d*) in the peel and pulp of the fruit of technical (CM) and biological (PR) ripeness of eggplant cultivars Snezhny and Vlas (*S. melongena*).

The signal level/100 mg of annotated compounds was used as a relative indicator for sugar content obtained from non-targeted metabolomic profiling.

List of DEGs associated with the metabolism of sugars, carotenoids and anthocyanins

Gene	ID, *S. melongena* transcriptome	*S. lycopersicum* homolog, NCBI ID
	Family GH32 (acid invertases; sucrose hydrolysis)	
VINV1	TRINITY_DN2044_c0_g1_i1.p1	acid vacuolar invertase ASK06213.1
CWINV1	TRINITY_DN7423_c0_g1_i23.p1	beta-fructofuranosidase, insoluble isoenzyme CWINV3-like XP_004241885.1
CWINV2	TRINITY_DN29292_c0_g1_i1.p1	beta-fructofuranosidase, insoluble isoenzyme CWINV1 XP_019068732.1
CWINV3	TRINITY_DN3426_c0_g1_i16.p1	cell-wall invertase AAM22409.1
	Family GH100 (neutral/alkaline invertases; sucrose hydrolysis)	
N/AINV1	TRINITY_DN5049_c0_g1_i2.p1	neutral/alkaline invertase 3, chloroplastic, XP_004249987.1
N/AINV2	TRINITY_DN5579_c0_g1_i2.p1	probable alkaline/neutral invertase D, XP_004241837.1
N/AINV3	TRINITY_DN5658_c1_g1_i7.p1	alkaline/neutral invertase A, mitochondrial, XP_004230329.1
N/AINV4	TRINITY_DN6542_c0_g1_i11.p1	alkaline/neutral invertase A, mitochondrial, XP_004230329.1
N/AINV5	TRINITY_DN6803_c0_g1_i6.p1	neutral/alkaline invertase 3, chloroplastic, XP_004249987.1
N/AINV6	TRINITY_DN9045_c1_g1_i1.p1	probable alkaline/neutral invertase D, XP_004238357.1
	Family SWEET (uniporters of soluble sugars)	
SWEET1	TRINITY_DN316_c1_g1_i1.p1	bidirectional sugar transporter SWEET1-like XP_004237723.1
SWEET2	TRINITY_DN2271_c0_g1_i1.p1	bidirectional sugar transporter SWEET1, XP_004242009.1
SWEET3	TRINITY_DN1022_c0_g1_i7.p1	bidirectional sugar transporter N3, XP_019068532.1
SWEET4	TRINITY_DN13252_c0_g1_i6.p1	bidirectional sugar transporter SWEET1-like, XP_004237724.1
SWEET5	TRINITY_DN1022_c0_g1_i4.p1	bidirectional sugar transporter N3, XP_019068532.1
SWEET6	TRINITY_DN10403_c0_g1_i2.p1	bidirectional sugar transporter SWEET2a, XP_004233011.1
	Carotenoid metabolism pathway	
PSY1	TRINITY_DN59246_c0_g1_i2.p1	phytoene synthase 1, NP_001234812.1
PSY2	TRINITY_DN6268_c0_g1_i3.p1	phytoene synthase 2, NP_001234671.1
NCED1	TRINITY_DN3512_c2_g1_i3.p1	9-cis-epoxycarotenoid dioxygenase, NP_001234455.1
	Flavonoid pathway	
CHS1	TRINITY_DN6959_c0_g2_i2.p1	chalcone synthase 1, NP_001234033.2
CHS2	TRINITY_DN6763_c0_g1_i3.p1	chalcone synthase 2, NP_001234036.2
F3H	TRINITY_DN5746_c0_g1_i2.p1	flavanone 3-dioxygenase, NP_001316412.1
DFR	TRINITY_DN49807_c0_g1_i1.p1	dihydroflavonol 4-reductase, NP_001234408.2

Analysis of transcripts of phytoene synthase genes, key isoenzymes of carotenoid metabolism (Rosas-Saavedra, Stange, 2016), showed trace *PSY1* values in the peel (cv. Snezhny) and pulp (both varieties) of the fruit and significant *PSY1* expression level in the fruit peel of cv. Vlas (Fig. 3). Relatively significant numbers of *PSY2* transcripts were found in the fruit peel (both varieties) and pulp (cv. Snezhny). At the same time, the number of *PSY2* transcripts was significantly higher in cv. Snezhny compared to cv. Vlas (Fig. 3). Another DEG associated with carotenoid catabolism, the 9-*cis*-epoxycarotenoid dioxygenase gene (*NCED1*), which catalyzes the synthesis of ABA from xanthophylls of the β,β-branch of the pathway (Rosas-Saavedra, Stange, 2016), was highly transcribed in the fruit peel and pulp of cv. Vlas, while in the fruit of cv. Snezhny, only trace values were detected (Fig. 3).

The list of DEGs associated with irreversible hydrolysis of sucrose and transport of mono- and disaccharides included genes for vacuolar invertase (*VINV1*), cell wall invertases (*CWINV1–3*), neutral/alkaline invertases (*N/AINV1–6*) and sugar uniporters (*SWEET1–6*) (see the Table).

In the peel of the fruit of cv. Snezhny, the highest level of expression was observed for the genes of four invertases (*VINV1*, *CWINV1*, *N/AINV5* and *6*) and three sugar uniporters (*SWEET1*, *3* and *5*); in the fruit pulp – for the genes of four

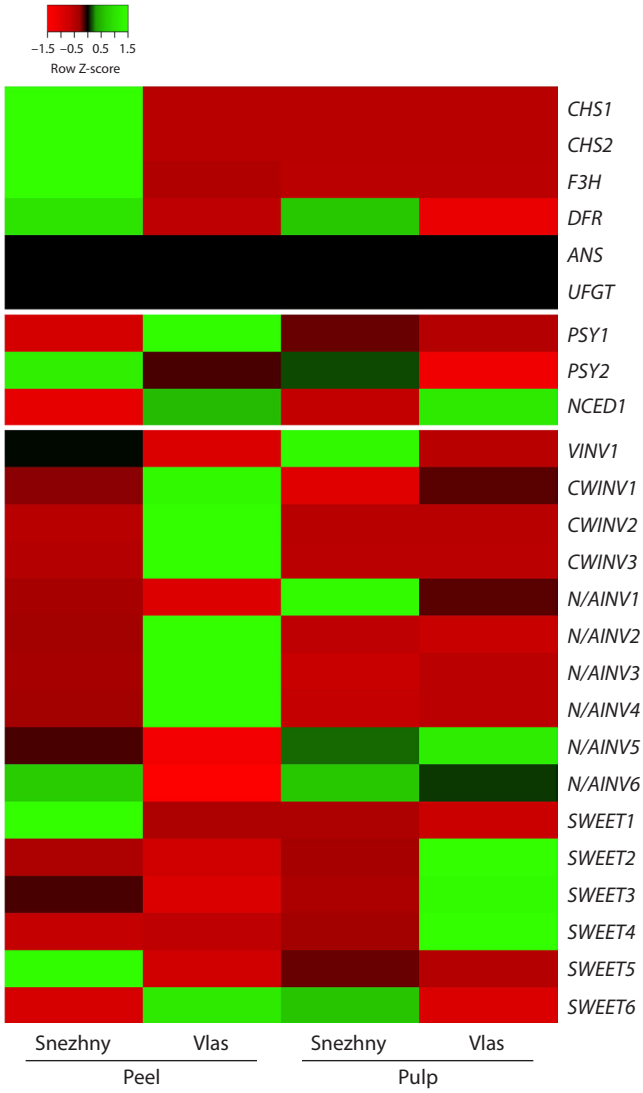


Fig. 3. Heatmap of the expression of DEGs associated with the metabolism of anthocyanins and carotenoids, as well as with the hydrolysis of sucrose and the transport of soluble sugars in the peel and pulp of biologically ripe fruit (PR) of the Snezhny and Vlas cultivars (*S. melongena*).

The heatmap was constructed based on transcriptomic analysis data.

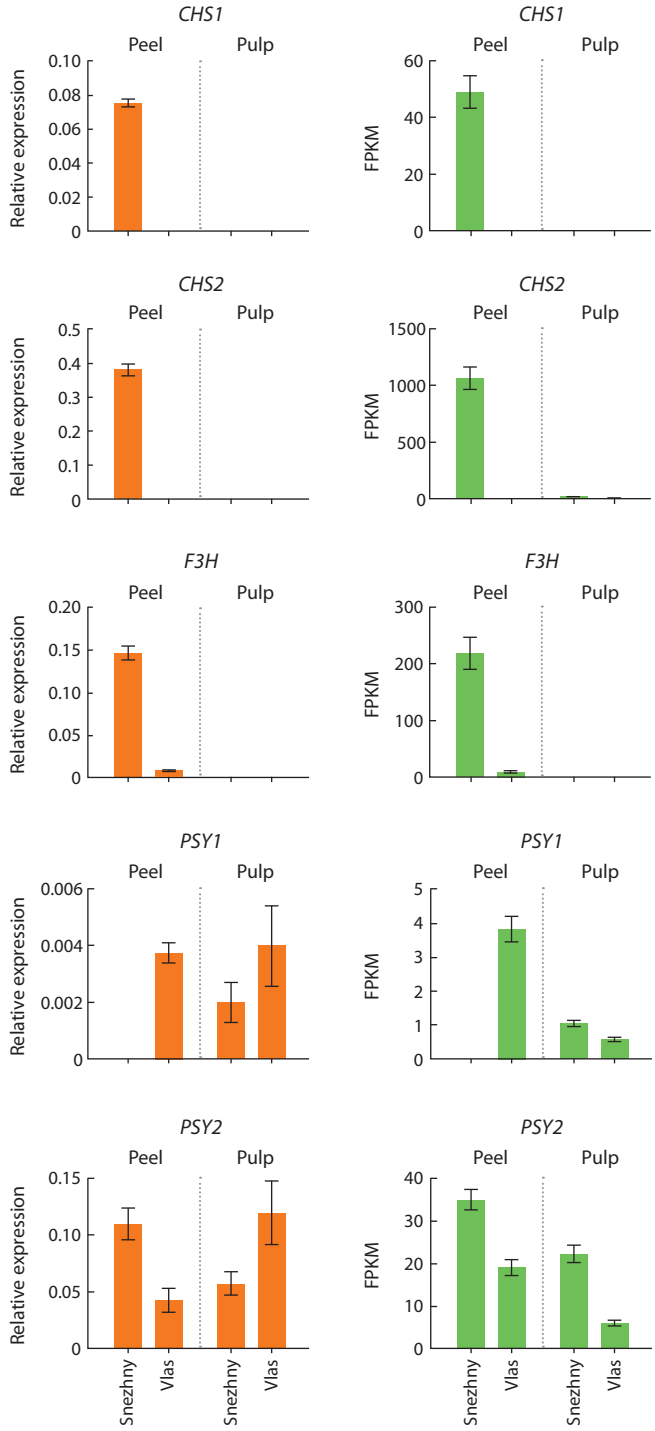


Fig. 4. Relative expression of the *CHS1*, *CHS2*, *F3H*, *PSY1* and *PSY2* genes based on qRT-PCR (left) and transcriptome (right) data.

The absence of the *DFR* and *ANS* transcripts was also confirmed by qRT-PCR; graphs are not shown. The primer sequences for *CHS1*, *CHS2*, *F3H*, *DFR*, *ANS* and the reference gene *GAPDH* were taken from the paper (Filyushin et al., 2023b); for the *PSY1* and *PSY2* genes, from (Kulakova et al., 2023).

invertases (*VINV1*, *N/AINV1*, *5* and *6*) and two sugar uniporters (*SWEET5* and *6*) (Fig. 3).

In general, cv. Vlas differed from cv. Snezhny in higher expression levels and a larger number of DEGs for invertases and sugar uniporters. In the fruit peel of cv. Vlas, the genes of six invertases (*CWINV3*, *CWINV1* and *2*, *N/AINV2–4*) and one sugar uniporter (*SWEET6*) were most highly transcribed, while in the fruit pulp, four invertases (*CWINV1*, *N/AINV1*, *5* and *6*) and three sugar uniporters (*SWEET2–4*) (Fig. 3).

Thus, the expression profile of genes for the metabolism of anthocyanins, carotenoids and sugars varied both within each cultivar (peel *vs*. pulp) and between cultivars (peel *vs*. peel, pulp *vs*. pulp).

Transcriptomic data were validated using qRT-PCR. Namely, in the same fruit tissues, the expression of *CHS1*, *CHS2*, *F3H*, *DFR*, *ANS* (flavonoid pathway), *PSY1* and *PSY2* (carotenogenesis) was determined (Fig. 4). It was shown that the expression pattern of these genes is consistent with transcriptomic data, with the exception of insignificant differences in the ratio of expression levels of the *PSY1* and *PSY2* genes in the fruit pulp between varieties (Fig. 4).

Discussion

The morphological diversity of eggplant cultivars has been the subject of much research, facilitating the optimization of breeding new cultivars with improved characteristics (Martínez-Ispizua et al., 2021). Particular attention is paid to metabolites (content, regulation of synthesis/accumulation) that have antioxidant properties and/or determine the ontogeny/stress resistance and taste of the fruit (Martínez-Ispizua et al., 2021). Nutraceuticals considered mainly include polyphenols, ascorbic acid, carotenoids and, less commonly, glycoalkaloids and sugars (Gürbüz et al., 2018; Akhbari et al., 2019; Condurache et al., 2021; Martínez-Ispizua et al., 2021; Saha et al., 2023).

In this study, accessions of two eggplant *S. melongena* varieties were characterized, which differ in the color of the fruit peel: cv. Snezhny (white color) and cv. Vlas (purple color) (Fig. 1). The characterization included the content of the sum of anthocyanins, the sum of carotenoids and soluble sugars in the peel and pulp of the fruit (CM and PR), accompanied by an analysis of the expression of genes encoding the key stages of metabolism of these compounds in the tissues of the biologically ripe fruit (PR).

Biochemical analysis confirmed that the purple color of the fruit peel of cv. Vlas is due to the presence of anthocyanins (Fig. 2*a*). The significantly higher content of carotenoids in the fruit peel of cv. Vlas in comparison with the pulp, as well as with the fruit of cv. Snezhny (Fig. 2*b*), does not affect the color of the fruit, apparently due to the presence of a large amount of anthocyanins.

In the fruit of cv. Vlas, the content of both pigments decreased significantly during the transition from technical to biological ripeness (Fig. 2*a*, *b*). This may be due to a decrease in the expression of genes for the biosynthesis of these metabolites or to the accelerated catabolism of pigment compounds. A decrease in concentration was also observed for soluble sugars (Fig. 2*c*, *d*). These results correspond to a decrease in the taste and antioxidant characteristics of the fruit at the stage of biological ripeness and explain the commercial use of fruits of technical ripeness.

According to transcriptomic analysis, genes for the metabolism of anthocyanins, carotenoids and sugars are differentially expressed both between fruit tissues within the same variety and between varieties (see the Table). This presumably determines intra- and intervarietal differences in the content of the corresponding compounds in fruit tissues.

In general, the obtained data on the expression of flavonoid pathway genes correspond to the previously shown profile of their expression in eggplant varieties with white and purple peel (Filyushin et al., 2023b). According to these data, between the stages of technical and biological ripeness, significant changes occur in the expression of genes of the flavonoid pathway, resulting in decrease of the content of anthocyanins in the peel of the purple fruit.

An unexpected result was the significantly higher expression of key genes of the flavonoid pathway (up to the anthocyanin branch) in the fruit of cv. Snezhny in comparison with the fruit of cv. Vlas (Fig. 3), which indicates the possibility of more flavonoids (excluding anthocyanins) being synthesized in the fruit of cv. Snezhny. Since the content of carotenoids in the fruit of cv. Snezhny is minimal, and the expression of flavonoid pathway genes is relatively high, it can be assumed that the yellow color of the ripe fruit (PR) of cv. Snezhny (Fig. 1*a*) is associated with the accumulation of flavonoids (colorless or yellow in color). This distinguishes eggplant fruits from the fruits of related species, tomato (*S. lycopersicum*) and pepper (*Capsicum annuum*), the color of which is associated with the accumulation of carotenoids (Filyushin et al., 2020).

In addition, these results are contrary to the few studies comparing the content of phenolic compounds in white and purple eggplant fruits, which indicate a greater accumulation of phenolic compounds in purple fruits (Martínez-Ispizua et al., 2021; Colak et al., 2022). Both studies included the analysis of only one white-fruited variety (Martínez-Ispizua et al., 2021; Colak et al., 2022), as in our case. Thus, white-fruited eggplant varieties can differ significantly from each other in the content of phenolic compounds and, consequently, antioxidant activity.

The shown expression profile of the phytoene synthase genes (*PSY1*, *PSY2*) initiating the biosynthesis of carotenoids (Fig. 3, 4) corresponds to the specificity of each of the two isoenzymes to a certain type of plastid (Rosas-Saavedra, Stange, 2016). Thus, *PSY1*, encoding a chromoplast-specific enzyme, was expressed in trace amounts, while chloroplast-specific *PSY2* corresponded to an order of magnitude more transcripts (Fig. 3, 4). At the same time, a high level of expression of the 9-*cis*-epoxycarotenoid dioxygenase (*NCED1*) gene, which catalyzes the conversion of β,β-branch carotenoids into ABA (Rosas-Saavedra, Stange, 2016), in the fruit of cv. Vlas, and its trace amounts in the fruit of cv. Snezhny (Fig. 3) suggest increased ABA content in the purple-colored fruit. Taking into account the complex functions of ABA (Waadt et al., 2022), this fact may indicate a greater efficiency of development, ripening, and response to stress factors of the purple fruit compared to the white fruit.

ABA content is positively associated with the amount of anthocyanins and soluble sugars (Teribia et al., 2016), although the content of the latter does not correlate with the accumulation of phenolic compounds, as well as carotenoids (Martínez-Ispizua et al., 2021).

The concentration of soluble sugars is regulated, among other things, by hydrolysis (invertases) and transport between tissues (sugar transporters) (Liu et al., 2022; Ren et al., 2022; Filyushin et al., 2023c). The invertase family includes neutral/alkaline (N/AINV) and acidic (vacuolar and cell wall; VINV/CWINV) enzymes that are involved in the regulation of plant ontogeny and stress tolerance (Qian et al., 2016), as well as sugar uniporters of the SWEET family (Fan et al., 2023; Filyushin et al., 2023a).

In comparison with cv. Vlas, the fruits of cv. Snezhny contained more hexoses and less sucrose (Fig. 2), which, at first glance, contradicts the lower activity of invertase genes (Fig. 3). However, these discrepancies may be a consequence of incomplete correspondence of the fruits of the two analyzed varieties in terms of the degree of biological ripeness. Ripe, fleshy fruits are characterized by enlarged cells with large vacuoles that actively accumulate and store sugars (Hedrich et al., 2015). In the peel and pulp of the fruit of cv. Snezhny, the highest level of expression of the only found DEG of vacuolar

invertase, *VINV1*, is detected (Fig. 3), which corresponds to the highest content of hexoses there (Fig. 2) and is probably a sign of complete biological ripeness of the analyzed fruit. At the same time, in the fruit of cv. Vlas, cell wall invertases and neutral/alkaline invertases are highly expressed (Fig. 3), operating in the cytoplasm and chloroplasts (Qian et al., 2016), where hexoses are actively utilized for development processes (Hedrich et al., 2015). That is, the analyzed fruit of a given variety may not have yet reached full ripening and is at an intermediate stage preceding biological ripeness. Also, the observed intervarietal difference in the content of sugars in the fruit may be a consequence of transport regulation of their concentration, including with the help of uniporters of the SWEET family (Filyushin et al., 2023a).

Conclusion

Thus, in this study, a comparative characterization of the ripe fruit of two varieties of eggplant *S. melongena* with white (cv. Snezhny) and purple (cv. Vlas) peel color was carried out using biochemical and transcriptomic analyses. It was shown that the purple color of the fruit of cv. Vlas is associated with the presence of anthocyanins and is accompanied by an increased accumulation of carotenoids and sucrose. This is consistent with the expression profile of genes linked to the key stages of the metabolism of these compounds and the transport of soluble sugars. Compared to cv. Vlas, the fruit of cv. Snezhny is characterized by a large content of hexoses and, possibly, flavonoids.

References

Akhbari M., Hamedi S., Aghamiri Z.S. Optimization of total phenol and anthocyanin extraction from the peels of eggplant (*Solanum melongena* L.) and biological activity of the extracts. *J. Food Measure. Character.* 2019;13:3183-3197. DOI 10.1007/s11694-019-00241-1

Alappat B., Alappat J. Anthocyanin pigments: beyond aesthetics. *Molecules.* 2020;25(23):5500. DOI 10.3390/molecules25235500

Colak N., Kurt-Celebi A., Gruz J., Strnad M., Hayirlioglu-Ayaz S., Choung M.G., Esatbeyoglu T., Ayaz F.A. The phenolics and antioxidant properties of black and purple versus white eggplant cultivars. *Molecules.* 2022;27(8):2410. DOI 10.3390/molecules27082410

Condurache N.N., Croitoru C., Enachi E., Bahrim G.E., Stanciuc N., Rapeanu G. Eggplant peels as a valuable source of anthocyanins: extraction, thermal stability and biological activities. *Plants.* 2021; 10:577. DOI 10.3390/Plants10030577

Fan X.W., Sun J.L., Cai Z., Zhang F., Li Y.Z., Palta J.A. *MeSWEET15a/b* genes play a role in the resistance of cassava (*Manihot esculenta* Crantz) to water and salt stress by modulating sugar distribution. *Plant Physiol. Biochem.* 2023;194:394-405. DOI 10.1016/j.plaphy.2022.11.027

Filyushin M.A., Dzhos E.A., Shchennikova A.V., Kochieva E.Z. Dependence of pepper fruit colour on basic pigments ratio and expression pattern of carotenoid and anthocyanin biosynthesis genes. *Russ. J. Plant Physiol*. 2020;67(6):1054-1062. DOI 10.1134/S1021443720050040

Filyushin M.A., Anisimova O.K., Shchennikova A.V., Kochieva E.Z. Genome-wide identification, expression, and response to *Fusarium* infection of the *SWEET* gene family in garlic (*Allium sativum* L.). *Int. J. Mol. Sci.* 2023a;24(8):7533. DOI 10.3390/ijms24087533

Filyushin M.A., Shchennikova A.V., Kochieva E.Z. Coexpression of structural and regulatory genes of the flavonoid pathway reveals the characteristics of anthocyanin biosynthesis in eggplant organs (*Solanum melongena* L.). *Russ. J. Plant Physiol*. 2023b;70:27. DOI 10.1134/S1021443722603147

Filyushin M.A., Slugina M.A., Shchennikova A.V., Kochieva E.Z. Differential expression of sugar uniporter genes of the *SWEET* family in the regulation of qualitative fruit traits in tomato species (*Solanum* section Lycopersicon). *Russ. J. Plant Physiol*. 2023c; 70(4):70. DOI 10.1134/S102144372360023X

Gürbüz N., Uluişikb S., Frarya A., Fraryc A., Doğanlara S. Health benefits and bioactive compounds of eggplant. *Food Chem.* 2018; 268:602. DOI 10.1016/j.foodchem.2018.06.093

Hedrich R., Sauer N., Neuhaus H.E. Sugar transport across the plant vacuolar membrane: nature and regulation of carrier proteins. *Curr. Opin. Plant Biol.* 2015;25:63-70. DOI 10.1016/j.pbi.2015.04.008

Hirakawa H., Shirasawa K., Miyatake K., Nunome T., Negoro S., Ohyama A., Yamaguchi H., Sato S., Isobe S., Tabata S., Fukuoka H. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. *DNA Res.* 2014;21:649. DOI 10.1093/dnares/dsu027

Jiang W., Li N., Zhang D., Meinhardt L., Cao B., Li Y., Song L. Elevated temperature and drought stress significantly affect fruit quality and activity of anthocyanin-related enzymes in jujube (*Ziziphus jujuba* Mill. cv. 'Lingwuchangzao'). *PLoS One.* 2020;15(11):e0241491. DOI 10.1371/journal.pone.0241491

Keunen E., Peshev D., Vangronsveld J., Van Den Ende W., Cuypers A. Plant sugars are crucial players in the oxidative challenge during abiotic stress: extending the traditional concept. *Plant Cell Environ.* 2013;36(7):1242-1255. DOI 10.1111/pce.12061

Kulakova A.V., Shchennikova A.V., Kochieva E.Z. Expression of carotenoid biosynthesis genes during the long-term cold storage of potato tubers. *Russ. J. Genet*. 2023;59(8):794-807. DOI 10.1134/S1022795423080094

Lelario F., De Maria S., Rivelli A.R., Russo D., Milella L., Bufo S.A., Scrano L. A complete survey of glycoalkaloids using LC-FTICR-MS and IRMPD in a commercial variety and a local landrace of eggplant (*Solanum melongena* L.) and their anticholinesterase and antioxidant activities. *Toxins* (*Basel*). 2019;11(4):230. DOI 10.3390/toxins11040230

Liu Y.H., Song Y.H., Ruan Y.L. Sugar conundrum in plant-pathogen interactions: roles of invertase and sugar transporters depend on pathosystems. *J. Exp. Bot.* 2022;73(7):1910-1925. DOI 10.1093/jxb/erab562

Martínez-Ispizua E., Calatayud Á., Marsal J.I., Mateos-Fernández R., Díez M.J., Soler S., Valcárcel J.V., Martínez-Cuenca M.R. Phenotyping local eggplant varieties: commitment to biodiversity and nutritional quality preservation. *Front. Plant Sci.* 2021;12:696272. DOI 10.3389/fpls.2021.696272

Pérez-Torres I., Castrejón-Téllez V., Soto M.E., Rubio-Ruiz M.E., Manzano-Pech L., Guarner-Lans V. Oxidative stress, plant natural antioxidants, and obesity. *Int. J. Mol. Sci.* 2021;22(4):1786. DOI 10.3390/ijms22041786

Qian W., Yue C., Wang Y., Cao H., Li N., Wang L., Hao X., Wang X., Xiao B., Yang Y. Identification of the invertase gene family (INVs) in tea plant and their expression analysis under abiotic stress. *Plant Cell Rep.* 2016;35(11):2269-2283. DOI 10.1007/s00299-016-2033-8

Ren R., Wan Z., Chen H., Zhang Z. The effect of inter-varietal variation in sugar hydrolysis and transport on sugar content and photosynthesis in *Vitis vinifera* L. leaves. *Plant Physiol. Biochem.* 2022;189: 1-13. DOI 10.1016/j.plaphy.2022.07.031

Rosas-Saavedra C., Stange C. Biosynthesis of carotenoids in plants: enzymes and color. *Subcell. Biochem.* 2016;79:35-69. DOI 10.1007/978-3-319-39126-7_2

Saha P., Singh J., Bhanushree N., Harisha S.M., Tomar B.S., Rathinasabapathi B. Eggplant (*Solanum melongena* L.) nutritional and health promoting phytochemicals. In: Kole C. (Ed.). Compendium of Crop Genome Designing for Nutraceuticals. Singapore: Springer Nature Singapore, 2023;1463-1493. DOI 10.1007/978-981-19-4169-6_53

Shi J., Zuo J., Xu D., Gao L., Wang Q. Effect of low-temperature conditioning combined with methyl jasmonate treatment on the chil-

ling resistance of eggplant (*Solanum melongena* L.) fruit. *J. Food Sci. Technol.* 2019;56(10):4658-4666. DOI 10.1007/s13197-019-03917-0

Tao T., Hu W., Yang Y., Zou M., Zhou S., Tian S., Wang Y. Transcriptomics reveals the molecular mechanisms of flesh colour differences in eggplant (*Solanum melongena*). *BMC Plant Biol.* 2023;23(1):5. DOI 10.1186/s12870-022-04002-z

Teribia N., Tijero V., Munné-Bosch S. Linking hormonal profiles with variations in sugar and anthocyanin contents during the natural development and ripening of sweet cherries. *Nat. Biotechnol.* 2016; 33(6):824-833. DOI 10.1016/j.nbt.2016.07.015

Waadt R., Seller C.A., Hsu P.K., Takahashi Y., Munemasa S., Schroeder J.I. Plant hormone regulation of abiotic stress responses. *Nat. Rev. Mol. Cell Biol.* 2022;23(10):680-694. DOI 10.1038/s41580-022-00479-6

Yang G., Li L., Wei M., Li J., Yang F. SmMYB113 is a key transcription factor responsible for compositional variation of anthocyanin and color diversity among eggplant peels. *Front. Plant Sci.* 2022;13: 843996. DOI 10.3389/fpls.2022.843996

You Q., Li H., Wu J., Li T., Wang Y., Sun G., Li Z., Sun B. Mapping and validation of the epistatic *D* and *P* genes controlling anthocyanin biosynthesis in the peel of eggplant (*Solanum melongena* L.) fruit. *Hortic. Res.* 2022;10(2):uhac268. DOI 10.1093/hr/uhac268

Zhang Y., Hu Z., Chu G., Huang C., Tian S., Zhao Z., Chen G. Anthocyanin accumulation and molecular analysis of anthocyanin biosynthesis-associated genes in eggplant (*Solanum melongena* L.). *J. Agric. Food Chem.* 2014;62:2906. DOI 10.1021/jf404574c

Conflict of interest. The authors declare no conflict of interest.
Received May 13, 2024. Revised June 27, 2024. Accepted June 28, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):628-639

DOI 10.18699/vjgb-24-70

Genotype imputation in human genomic studies

A.A. Berdnikova [1, 2], I.V. Zorkoltseva [1], Y.A. Tsepilov [1], E.E. Elgaeva [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
elizabeth.elgaeva@gmail.com

Abstract. Imputation is a method that supplies missing information about genetic variants that could not be directly genotyped with DNA microarrays or low-coverage sequencing. Imputation plays a critical role in genome-wide association studies (GWAS). It leads to a significant increase in the number of studied variants, which improves the resolution of the method and enhances the comparability of data obtained in different cohorts and/or by using different technologies, which is important for conducting meta-analyses. When performing imputation, genotype information from the study sample, in which only part of the genetic variants are known, is complemented using the standard (reference) sample, which has more complete genotype data (most often the results of whole-genome sequencing). Imputation has become an integral part of human genomic research due to the benefits it provides and the increasing availability of imputation tools and reference sample data. This review focuses on imputation in human genomic research. The first section of the review provides a description of technologies for obtaining information about human genotypes and characteristics of these types of data. The second section describes the imputation methodology, lists the stages of its implementation and the corresponding programs, provides a description of the most popular reference panels and methods for assessing the quality of imputation. The review concludes with examples of the use of imputation in genomic studies of samples from Russia. This review shows the importance of imputation, provides information on how to carry it out, and systematizes the results of its application using Russian samples.

Key words: imputation; genotyping; sequencing; genome-wide association study; human; DNA-microarray.

For citation: Berdnikova A.A., Zorkoltseva I.V., Tsepilov Y.A., Elgaeva E.E. Genotype imputation in human genomic studies. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):628-639. DOI 10.18699/vjgb-24-70

Funding. The work was supported by the Russian Science Foundation grant number 22-15-20037 and the Government of Novosibirsk Region.

Acknowledgements. Authors express gratitude to V.S. Fishman for his recommendations for improvement of the article text.

Импутация генотипов в геномных исследованиях человека

А.А. Бердникова [1, 2], И.В. Зоркольцева [1], Я.А. Цепилов [1], Е.Е. Елгаева [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
elizabeth.elgaeva@gmail.com

Аннотация. Импутация – это метод, позволяющий восстанавливать недостающую информацию о генетических вариантах, которые не удалось генотипировать напрямую с помощью ДНК-микрочипов или секвенирования с низким покрытием. Импутация играет важнейшую роль в полногеномном анализе ассоциаций (genome wide associations study, GWAS). Она приводит к существенному увеличению количества изучаемых вариантов, что повышает разрешающую способность метода и увеличивает сопоставимость данных, полученных в разных когортах и/или с помощью разных технологий, что важно при проведении метаанализов. При ее выполнении информацию о генотипах в исследуемой выборке, у которой известна только часть генетических вариантов, дополняют за счет эталонной (референсной) выборки, имеющей более полные данные о генотипах (чаще всего это результаты полногеномного секвенирования). Импутация стала неотъемлемой частью геномных исследований человека благодаря преимуществам, которые она дает, а также увеличению доступности инструментов для импутации и данных референсных выборок. Обзор посвящен импутации в геномных исследованиях человека. В первом разделе приводятся описание технологий получения информации о генотипах человека и характеристика получаемых типов данных. Во втором разделе представлена методология импутации, перечисляются этапы ее проведения и соответствующие программы, дается опи-

сание наиболее популярных референсных панелей и способов оценки качества импутации. В заключении представлены примеры использования импутации в геномных исследованиях выборок из России. Настоящий обзор показывает важность проведения импутации, дает информацию о том, как ее выполнять, и систематизирует результаты ее применения на примере российских выборок.
Ключевые слова: импутация; генотипирование; секвенирование; полногеномный анализ ассоциаций; человек; ДНК-микрочип.

Technologies for obtaining human genotype data and their features

Human genotype data are a key aspect for many genetic studies. There are several technologies developed to read, analyze and interpret genetic information. The most commonly used methods include Sanger sequencing, next generation sequencing (NGS), and DNA microarrays.

Genotyping using DNA microarrays

A DNA microarray (or simply a "microchip" or "chip", not to be confused with an RNA microarray, which is a different technology) is a small glass or silicon substrate, to which tens of thousands of probes (short single-stranded DNA fragments complementary to certain nucleotide sequences) are attached. These probes are arranged on the chip in such a way that each fragment can be identified by its location (Fig. 1).

During the analysis, fluorescent markers are attached to the studied DNA molecules, which were cut into fragments by restriction endonucleases, and placed on the chip. The target DNA fragments are bound to complementary DNA probes, and all remaining fragments are removed from the chip. Laser beams and computer processing are used to detect the fluorescence of fragments, record the emission (radiation) patterns and subsequently identify the sequences. This method is very fast and allows to simultaneously determine the nucleotide sequence of several DNA fragments (Govindarajan et al., 2012).

An alternative approach to solving the problem of genotyping was implemented by academician A.D. Mirzabekov in domestic developments to create gel microchips (Mirzabekov, 2003). They are a substrate made of glass, plastic or silicone with hemispherical drops of hydrogel fixed on its surface. The distinction of this method is that DNA fragments are immobilized in three-dimensional space, which provides greater sensitivity and capacity of the microchip. This technology has also found its application in RNA analysis, protein and cell biochips.

There are several strategies for identifying single nucleotide polymorphism (SNP) for microarrays (Fig. 2).

Allele distinction by hybridization (Fig. 2*a*). The labeled target DNA hybridizes with probes containing a polymorphic site in the center. Correctly paired oligonucleotides are more stable (have a higher melting temperature) compared to duplexes with a non-complementary base. Therefore, after washing the chip under harsh temperature conditions, only correctly paired chains remain on it. It is common to use multiple fragments for each allele to improve the quality of the signal relative to noise (Wang D.G. et al., 1998).

"Golden Gate" analysis by the Illumina company (Fig. 2*b*). Two allele-specific oligonucleotides, each of which has a 5′ end with different universal primers (P1 and P2) (the primers are labeled with a unique fluorophore for subsequent site discrimination), hybridize in solution with genomic DNA. The third oligonucleotide, in addition to the universal primer (P3), has a tail with a "barcode" sequence complementary to the fragment on the chip. The allele-specific primers extended by a polymerase are ligated to a third oligonucleotide, after which the resulting fragments are amplified using the polymerase chain reaction and hybridized onto the chip. The use of multiple barcodes (one for each locus of interest) allows for analysis of several genomic regions at once (Fan et al., 2003).

Arrayed Primer Extension (APEX, Fig. 2*c*). Here, the chip contains a DNA fragment, the 5′ end of which is fixed to the substrate, and the 3′ end finishes with the nucleotide preceding the SNP being detected. Fragments of genomic DNA are hybridized to the chip, while the desired SNP remains unpaired. During the sequencing reaction, the nucleo-

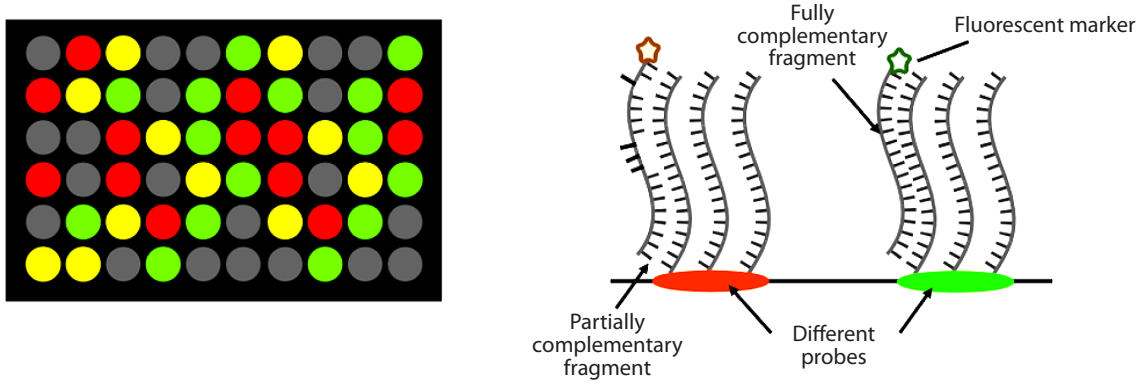


Fig. 1. DNA microarray.

Pseudocolor (red, yellow or green) is determined by the number of molecules bound to the probe and labeled with different dyes. For further explanation of the figure, see the text below.

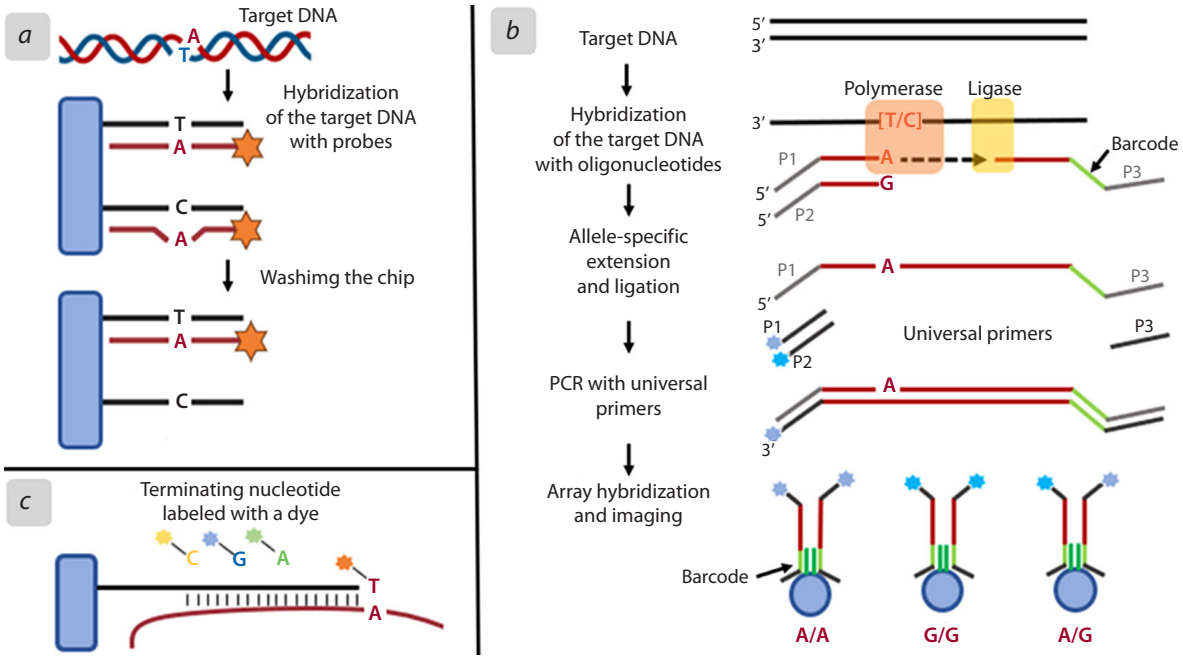


Fig. 2. SNP detection strategies for DNA microarrays.

a – allele-specific hybridization; *b* – "Golden Gate" analysis by the Illumina company; *c* – arrayed primer extension.

tide sequence attached to the substrate is extended by one terminating nucleotide labeled with a dye (Kurg et al., 2000). This nucleotide prevents further growth of the DNA chain, and the color of its dye allows you to determine which of the nucleotides (A, T, G or C) is located at the given position.

One of the main advantages of DNA microarrays is their high throughput capability (Hayat, 2002; Brown et al., 2024). The microarray provides the basis for simultaneous genotyping of thousands of different loci and detection of single nucleotide substitutions. Thus, microarrays are used to analyze large samples in order to genotype frequently occurring genetic variants (with a minor allele frequency in the population > 0.01).

However, there are some limitations in interpreting the results. Microarray data are typically binary (indicating the presence or absence of a specific allele), high-throughput (allowing the analysis of thousands or millions of SNPs), and requiring specialized analysis techniques to extract meaningful information. In this case, we are talking about software (for example, GenomeStudio (Illumina Inc., San Diego, CA, США)), which includes tools for quality control, genotype identification, visualization and data analysis. In addition, microchips can produce both false positive and false negative results. These issues highlight the importance of careful data interpretation and the need to use appropriate statistical methods to control quality and validate results.

Genome sequencing

This chapter describes various sequencing technologies. Around 1976, two methods that could read hundreds of bases in half a day were developed – Sanger and Coulson strand termination and Maxam and Gilbert chemical cleavage (Maxam, Gilbert, 1977; Sanger et al., 1977). In both methods, the analyzed DNA is placed into four test tubes with different compositions of the reaction mixture for a specific type of nitrogenous base (A, T, G or C). Gilbert's method uses DNA, radioactively labeled at one end, and a mixture of enzymes that specifically cut it before a certain type of nucleotide. Sanger sequencing, in contrast, involves primers and dideoxynucleotides that stop chain synthesis when radiolabeled dideoxynucleoside triphosphate (ddNTP; different in each tube) is included. Hence, as a result of implementing either method, labeled DNA fragments of different lengths that end with the same base are formed in each tube. Sequences are separated by length using polyacrylamide plate gel electrophoresis (one lane per base type) at single nucleotide resolution. The image obtained on X-ray film after electrophoresis allows researchers to restore the original sequence. The described methods immediately came into use, and by 1987, automated fluorescent Sanger sequencers could read about 1,000 bases per day (Smith et al., 1986; Connell et al., 1987).

In 2005, next generation sequencing (NGS) technologies were first introduced, which are based on two approaches. The first of these is sequencing by hybridization (SBH). The essence of the method is as follows: first, short sections of DNA are fixed on a glass substrate (DNA chip). Then the fragments to be identified are labeled with fluorophore and applied to the chip for hybridization with the fixed areas. Single-stranded DNA is washed away, and the hybridization pattern is read from the color marks and their brightness. An alternative approach in NGS is sequencing by synthesis (SBS) (Shendure et al., 2017).

As a rule, in technologies that use the SBS technique, pre-fragmented sequences are fixed in a flow cell, where cyclic

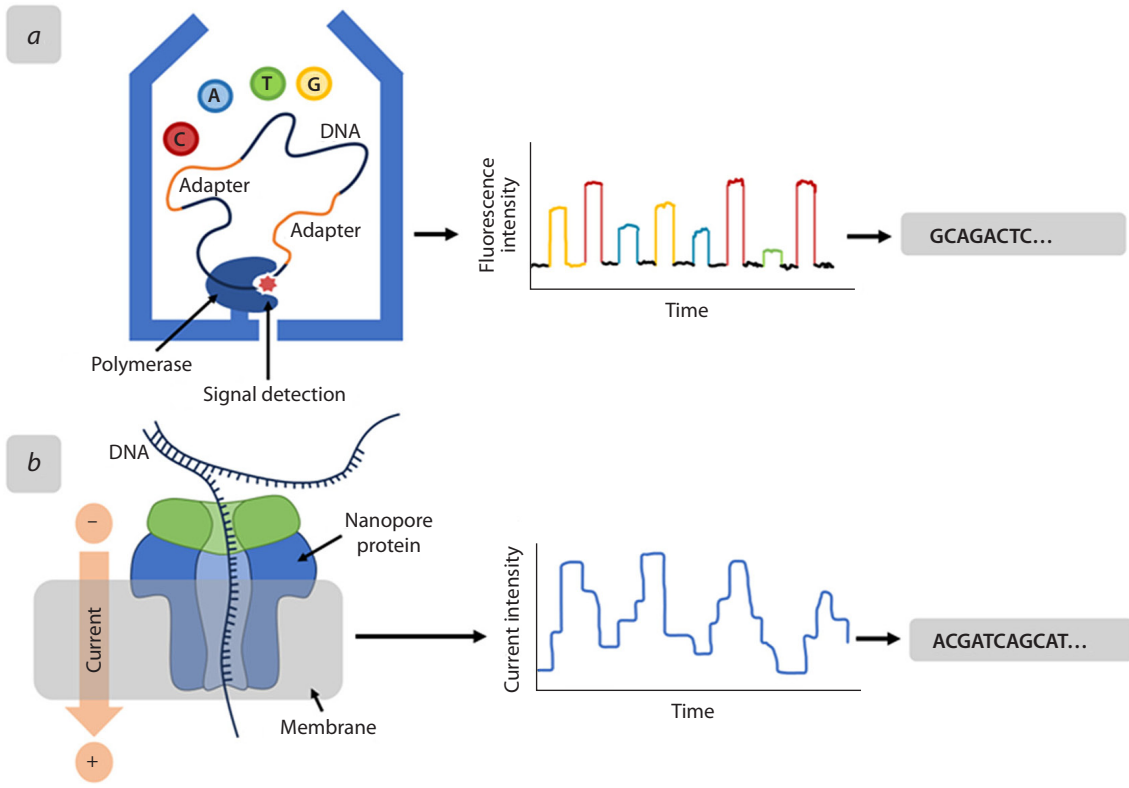


Fig. 3. Third generation sequencing.

а – Pacific Biosciences; *б* – Oxford Nanopore Technology. See the text below for explanation.

synthesis of a new chain occurs. By sequentially adding one of the four deoxynucleotides to the mixture, having removed the previous ones in advance, it is possible to read signals from the cells where the synthesis reaction was successful. Therefore, the output provides information about where which nucleotide is located.

Sequencing technologies with an approach other than NGS were first described in 2008–2009 and named "third generation sequencing" (Check Hayden, 2009). They include two main approaches (Fig. 3).

The first technology, Pacific Biosciences (PacBio) (Rhoads, Au, 2015), is designed to optically monitor DNA synthesis using a polymerase in real time. The structure has a hole less than half the light wavelength that limits fluorescent excitation to a small volume containing only the polymerase and its template (Fig. 3*a*). With such a device, only fluorescently labeled nucleotides included in the growing DNA strand emit signals of sufficient duration to be read. The error rate in this sequencing method is very high (about 10 %), but the errors are distributed randomly. With long reads and tolerance for high GC content and random errors, PacBio provides *de novo* assemblies of unprecedented quality in terms of accuracy and continuity.

The second major third-generation sequencing technology is Oxford Nanopore (ONT) (Deamer et al., 2016). This technique was first proposed in the 1980s. The special chamber where the sequencing process takes place is filled with an electrolytic solution and divided by a two-layer membrane with a nanopore (its dimensions are in the nanorange). Once voltage is applied, the electrolyte ions and the DNA molecule begin to move through the pore. Nucleic acid physically interferes with the migration of ions, which leads to fluctuations in current strength, which allows the nucleotide sequence to be determined (Fig. 3*b*). The main difference from other sequencing technologies is the extreme portability of nanopore devices, which can be as small as a memory stick (USB), as they rely on detecting electronic rather than optical signals.

Comparison of technologies and their application to solve different problems

Most often, Illumina NGS technologies are used for large-scale projects (whole-genome sequencing, transcriptome analysis and epigenetic profiling), but PacBio is more useful for *de novo* assembly, and ONT is more applicable for portable sequencing. The Sanger method is suitable for sequencing short DNA fragments such as individual genes, plasmids or viral genomes.

Also worth mentioning is a sequencing technology competing with Illumina, developed by Complete Genomics and MGI Tech, DNBSEQ-T7 (formerly known as MGISEQ-T7). In DNBSEQ-T7, the clonal amplification process occurs as a rolling circle, i. e., always from the original template, which eliminates the accumulation of DNA polymerase errors (Drmanac et al., 2010). The main advantages of MGI include lower cost compared to Illumina and the ability to process a larger volume of samples in a shorter time. As recent studies

show, the new MGISEQ-2000 sequencer can be used as a full-fledged alternative to Illumina sequencers when conducting whole-genome studies (search for variants, identification of indels), the differences between the two platforms are insignificant (Korostin et al., 2020; Jeon et al., 2021; Feng et al., 2024).

Recently, the effectiveness of using whole-genome sequencing (WGS) for GWAS has been demonstrated (DePristo et al., 2011; Chat et al., 2022). This approach is a promising alternative to genotyping using DNA microarrays, as it allows one to obtain information on a larger proportion of genetic variations, increasing the power of association tests and subsequent fine-mapping analyses (Wang Q.S., Huang, 2022). However, despite the decreasing cost of NGS-based technologies, GWAS mainly use high-throughput and relatively cheap DNA microarrays containing hundreds of thousands to millions of common genetic markers, which make it possible to test almost the entire genome for associations with the trait being studied. SNP genotyping using DNA microarrays can contain up to 5 % errors depending on the manufacturer (Lamy et al., 2006; Yang et al., 2011; Guo et al., 2014). However, existing protocols for quality control of the obtained data can significantly reduce the number of errors (on average by 1.7 %) (Zhao et al., 2018). Thus, microarrays allow fairly accurate genotyping of samples even for species with high heterozygosity (i. e., with greater genetic variation than expected at Hardy–Weinberg equilibrium) (Bourke et al., 2018). Moreover, at the end of 2023, the cost of genotyping a sample on a microchip was an order of magnitude lower than the cost of NGS sequencing, which makes it possible to cover a much larger sample size with the same project budget. Their main disadvantage when conducting GWAS is that they do not allow detection of an association between an SNP and a trait if the genetic variant is not represented on the microarray.

Additional difficulties in using DNA microarrays may arise because the information (such as the location of SNPs on the chromosome) used to design the chip is out of date or differs between manufacturers. The above problems can be solved by imputation of genotyping data (Pasaniuc et al., 2012). This approach allows us to increase the density of coverage for the genetic variants studied (total number of markers) and the proportion of common variants when conducting a meta-analysis (combining data from different studies and/or genotyping platforms) (Li Y. et al., 2009).

A replacement for DNA microarrays could be low-coverage WGS (lcWGS), in which random regions of the genome are sequenced (Chat et al., 2022). Research shows that lcWGS significantly outperforms microarrays in marker density, which also allows for a more thorough assessment of associations with less common variants. Such data also require imputation using haplotypes (e. g., from the 1000 Genomes Project) (Auton et al., 2015). The costs of ultra-low coverage WGS (sequencing depth $\leq 0.5x$) may be comparable to or lower than those of using DNA microarrays, but its potential as an alternative has not yet been sufficiently assessed (Martin et al., 2021).

DNA sequencing and genotyping solve the task of analyzing genetic information in different ways. As such, sequencing allows you to read entire DNA fragments and is therefore applied to identify rare (minor allele frequency < 0.01 %) and *de novo* mutations, and is widely used to study the structure of individual genes or genome regions. Genotyping, on the other hand, is a faster and more cost-effective method for analyzing genetic variation, which is particularly useful for large-scale genomic studies involving thousands or even millions of samples. Thus, if the goal of a study is to comprehensively examine the genetic architecture of a trait or disease, sequencing is likely the best approach. However, if the focus of the study is on common genetic variants, or analysis of the population or kinship structure of the sample, then genotyping is often sufficient and more effective (Gresham et al., 2008).

Imputation of genotyping data

Although sequencing the entire genome of hundreds of thousands of people is not yet feasible, significant progress can be made by identifying only a relatively small number of genetic variants in each person. This type of "incomplete" information is still useful because data on any set of SNPs in a group of people allow inferences to be made about many other unobserved variants in the same people. The approach to accomplish this is called imputation.

Methodology

The imputation procedure includes the following stages: quality control of genotyping data, phasing, imputation itself, and at the final step – quality control of imputed genotypes (Fig. 4).

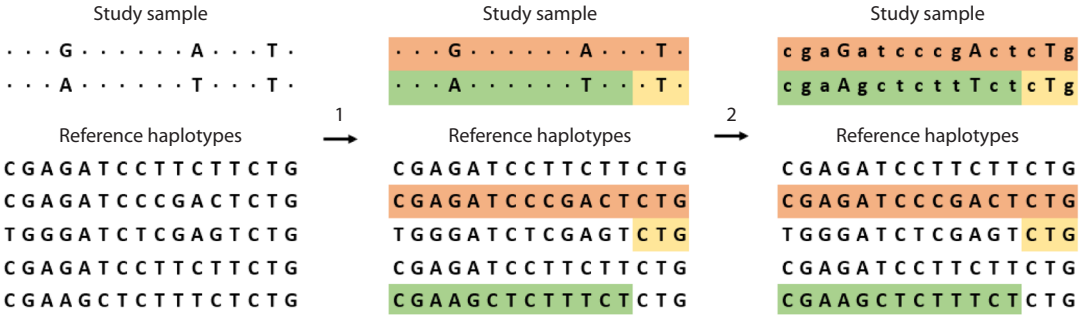


Fig. 4. Imputation of genotyping data.

1 – Phasing; 2 – Imputation itself. See the text below for explanations.

Genetic variants that are located nearby on a chromosome are more likely to be inherited together, which occurs because there are only a few recombinations per chromosome. This principle is called “linkage disequilibrium” (LD). Due to this principle, we observe blocks of haplotypes (haploblocks) – sets of closely spaced genetic variants that were inherited together during evolution.

In imputation, haploblocks are used to identify common short stretches of DNA on chromosomes that individuals in a randomly selected population may have inherited from a common ancestor. By comparing haplotypes in two samples (study and reference) based on a set of common genetic variants, imputation algorithms provide inferences about the genotypes of the studied individuals. Both of these samples must be from the same ethnic group for imputation to produce accurate results (Mills et al., 2020).

Although genotyping data do not contain haplotype information, it can be inferred and reconstructed using stepwise analysis. Phasing is the process of statistically estimating haplotypes. Imputation can be performed on both raw unphased genotyping data and reconstructed mixed haplotypes, although phasing is known to improve imputation accuracy (Anderson et al., 2010). In addition, phasing is often necessary due to the fact that standard imputation algorithms (more about them below) work specifically with haploblocks.

Quality control of genotyping data

An important step in any genomic study is to conduct data quality control. The importance of this step is illustrated by the example of a paper published in Science that was retracted due to insufficient consideration of technical errors in genotyping on an Illumina chip (Marees et al., 2018).

Quality control of DNA microarray genotyping data is divided into two main steps: control at the individual level and control at the marker level. Individual-level control involves removing a sample in the following cases (Anderson et al., 2010):

– there is an observed discrepancy between the phenotype and the genotype (in particular, the phenotypic sex differs from the genetic one);
– the number of heterozygous loci in the genome deviates from the expected value (an overestimation or underestimation of this indicator may indicate sample contamination or inbreeding, respectively);
– the sample contains duplicates, relatives of the first or second degree (similar genotypes will be overrepresented, as a result of which allele frequencies in the population may be displayed unreliably);
– has a different ethnic origin, that is, there is a stratification of the population (the most common approach for identifying such individuals is principal component analysis (PCA) on a kinship matrix).

Data quality control at the level of individual markers also consists of several points that involve the removal of SNPs if:

– minor allele frequency (MAF) < 0.01;
– they are absent from a large part of individuals in the sample;
– they deviate significantly from Hardy–Weinberg equilibrium.

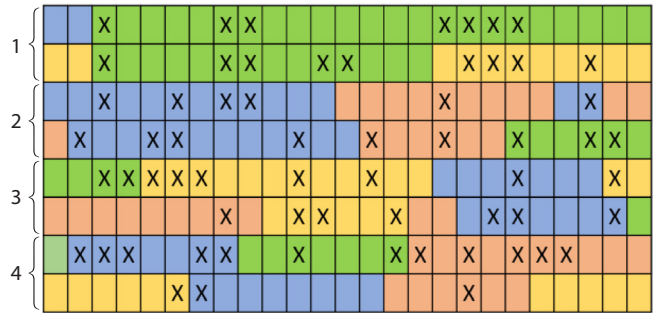


Fig. 5. Visualization of the performance of HMM-based algorithms for four individuals from the reference sample.

Each column is a separate SNP with two alleles (empty and crossed out squares represent different alleles of the same SNP), and each pair of rows represents two copies of DNA (from each parent). Closely related SNPs are grouped by color, and each haplotype is modeled as a mosaic of color combinations (Scheet, Stephens, 2006).

To carry out quality control, a number of publicly available programs are used: PLINK 1.9/PLINK 2 (Purcell et al., 2007; Chang et al., 2015), RICOLI (Lam et al., 2020), SMARTPCA (Price et al., 2006) and FlashPCA (Abraham et al., 2017).

Imputation Tools

Over the past twenty years, several different research groups have developed and published a number of tools for phasing and subsequent imputation, most of which are based on the hidden Markov model (HMM) of Li and Stephens (Li N., Stephens, 2003). This statistical model, first described in 2003, assumes that haplotypes are inherited as haploblocks and that recombination events occur at their boundaries. The model probabilistically reconstructs the studied haplotypes in the form of a mosaic composed of haplotypes from a small reference sample (Fig. 5). It has been shown that methods based on Li and Stephens’ HMM are more accurate and efficient (Weale, 2004) than approaches such as Clark’s algorithm (Clark, 1990) or the EM algorithm (Expectation-Maximization) (Dempster et al., 1977) (Browning S.R., Browning B.L., 2011). Currently, the most commonly used programs implementing Li and Stephens’ HMM are Beagle 5 (Browning B.L. et al., 2021), Eagle2 (Loh et al., 2016) and ShapeIT (Delaneau et al., 2012) for phasing, and also Beagle 5 (Browning B.L. et al., 2018), Impute5 (Rubinacci et al., 2020) and Minimac4 (Das et al., 2016) for imputation. Beagle 5 and ShapeIT2 allow you to perform both of these procedures.

A comparative analysis of current phasing and imputation software showed that, overall, Beagle 5.4 performed slightly better than Impute5 and Minimac4, with a higher concordance rate and high performance even on large data sets (De Marino et al., 2022). However, Minimac4 and Impute5 tend to perform better on rare variants because, unlike Beagle 5.4, which computes clusters of haplotypes and performs calculations based on them, Impute5 and Minimac4 search the entire haplotype space. Minimac4 requires the least amount of memory, but calculations take longer. If memory usage is limited and the loss of accuracy is acceptable, then Minimac4 may be the optimal choice of imputation software.

The above programs can be run from a local server and require reference haplotypes. Nevertheless, most of these large-scale datasets are not publicly available. For this reason, special servers that contain information about different reference panels are most often used for imputation of human data, such as Michigan Imputation Server[1] (Das et al., 2016) and TOPMed Imputation Server[2] (Das et al., 2016). Researchers can upload their datasets there, configure parameters through the web user interface (select tools, reference panels, etc.), perform phasing and genotype imputation on the server, and download the output files.

As disadvantages of this approach, it is worth noting the need to send your data outside the local server (albeit using secure connection protocols) and possible queues. In addition, users are often limited in the choice of programs or reference panels, and cannot combine multiple panels or integrate their own. However, it is possible to bypass these restrictions, for example, using Docker software (Das et al., 2016), and run imputation on your server. The problem with standalone running is a little more complexity due to manual settings, where the user needs to install additional programs for the pipeline and account for library conflicts.

In Supplementary Material 1[3] compares the tools available on the two servers described above.

Reference panels for imputation of human genotyping data

One important issue in genotype imputation is how to select a reference panel that provides high imputation accuracy in the population of interest. As it was shown (Huang, Tseng, 2014), the quality of imputation is affected not only by the size of the panel, but also by the ethnic composition of the reference sample. The most commonly used panels for European populations currently are 1000 Genomes (Sudmant et al., 2015), Haplotype Reference Consortium (HRC) (Haplotype Reference Consortium, 2016) and Trans-Omics for Precision Medicine (TOPMed) (Taliun et al., 2021).

The 1000 Genomes Phase 3 Version 5 reference panel was prepared as part of the 1000 Genomes Project in 2008 (Auton et al., 2015). In total, while using a combination of low-coverage whole-genome sequencing, high-coverage exome sequencing and microarray genotyping, this project was able to characterize 88 million genetic variants (84.7 million SNPs, 3.6 million short insertions/deletions and 60,000 structural variants). This version of the reference panel includes 49 million markers from 2,504 individuals from a mixed population.

The HRC r1.1 2016 reference panel was compiled by the HRC (The Haplotype Reference Consortium) to create a large haplotype reference panel. The HRC panel combines datasets from 20 different studies, most of which were obtained using low-coverage (4–8*x*) whole-genome sequencing and consist of samples of individuals of predominantly European ancestry. The reference panel consists of 64,976 haplotypes of 32 thousand individuals with 39,235,157 SNPs; it does not contain deletions or insertions.

The TOPMed (The Trans-Omics for Precision Medicine) project was initiated in 2010 with the goal of collecting and analyzing whole-genome sequencing data. As of September 2021, TOPMed has approximately 180 thousand participants, predominantly of non-European origin, from more than 85 different studies. A reference panel was created based on the TOPMed database, which includes 286,068,980 SNPs; 5,815,513 insertions and 16,222,592 deletions in the genotypes of 97,256 individuals. These genetic variants are distributed across 22 autosomes and the X chromosome. TOPMed (Version r2) is the first panel that is based solely on deep whole-genome sequencing data and is significantly superior to previously published alternatives.

Although most genetic studies and reference panels focus on samples of individuals of European ancestry, it is worth noting that there are various projects aimed at studying the genetic diversity of other populations. These include ChinaMAP (10,588 samples and 136.7 million SNPs) (Li L. et al., 2021), NARD (1,779 individuals, 40.6 million SNPs) (Yoo et al., 2019), GAsP (1,739 samples, 1 million autosomal SNPs) (Wall et al., 2019), SG10K (4,810 samples, 89.1 million SNPs) (Wu et al., 2019) for samples of people of Asian descent, AFAM (2,269 samples, 45 million SNPs) (O'Connell et al., 2021) and UGR (4778 samples, 2.2 million markers) (Fatumo et al., 2022) for African Americans. The TOPMed panel can also be used to impute non-European samples of individuals of both African and Asian descent.

The ideal solution when selecting a panel for imputation is to combine data from multiple reference samples to construct a combined reference panel. However, different studies tend to use different quality control and variant filtering strategies, which can make pooling results difficult.

Another major issue is restrictions on shared data use. For example, individual-level genotype information in many reference panels is not publicly available; therefore, it may not be possible to directly combine it with sequencing results from other samples. In this regard, the meta-imputation method was proposed (Yu et al., 2022). Instead of combining reference panels, genotypes are first imputed using multiple reference panels separately and then the imputed results are combined into a consistent data set.

Assessment of imputation quality

The quality of genotyping data imputation can be assessed: 1) using standard imputation quality metrics; 2) empirically (for example, conduct a GWAS on the trait of interest and check the reproducibility of association signals known from the literature, or calculate a polygenic estimate of the trait and compare it with real phenotypes).

Imputation quality metrics can also be divided into two large groups (Stahl et al., 2021): 1) those that assess the quality of imputation without using directly genotyped SNPs and are calculated automatically when running the corresponding imputation software, and 2) those that allow the comparison between imputed SNPs and genotypes and are calculated manually.

Quality metrics in the first group are specific to each individual program. For example, for Minimac4 and Beagle 5, the R^2 indicator is estimated (Marchini, Howie, 2010), which

[1] https://imputationserver.sph.umich.edu/index.html#!pages/home
[2] https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home
[3] Supplementary Materials 1–3 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Berd_Engl_28_6.pdf

is calculated differently for each program, while Impute5 calculates the Info parameter (Marchini, Howie, 2010). Because of their specificity, they are not suitable for comparing the quality of data imputed by different methods. This task is successfully accomplished by metrics from the second group, which include: concordation rate (CR), Imputation Quality Score (IQS) (Lin et al., 2010), Hellinger score (Roshyara et al., 2014), squared Euclidean norm score (SEN) (Roshyara et al., 2014) and others. In practice, standard metrics of the first group are most often used.

While conducting imputation, the posterior probabilities of the genotype are estimated. Thus, for biallelic SNPs in an additive model (where the genotype is coded as 0, 1 and 2, and the reference and alternative allele are 0 and 1, respectively), the estimated probability of individual i to have genotype j at a particular locus is denoted as G_j^i (j = 0, 1, 2). This indicator is calculated by appropriate imputation software based on data from the reference and target samples using built-in algorithms (for example, a hidden Markov model, as described above). The dose of the alternative allele is calculated as $D_i = G_1^i + 2G_2^i$.

The R^2 metric is an approximation of the squared correlation between the imputed allele dose and the expected genotype and is calculated as the ratio of the allele dose dispersion and the expected dispersion under Hardy–Weinberg equilibrium

$$\hat{R}_d^2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(D_i - 2\hat{p})^2}{2\hat{p}(1-\hat{p})}, \quad (1)$$

$$\hat{p} = \sum_{i=1}^{N}\frac{D_i}{2N},$$

where N is the number of individuals in the sample; D_i is the dose of the imputed allele for the i-th individual; $\hat{p}$ is the allele frequency estimate.

Many modern algorithms (such as Minimac) carry out imputation on pre-phased genotypes, that is, they work with haplotypes. The formula undergoes slight changes, as the set of genotypes is now described as a pool of $2N$ binary encoded alleles

$$\hat{R}_h^2 = \frac{\frac{1}{2N}\sum_{i=1}^{2N}(H_i - \hat{p})^2}{\hat{p}(1-\hat{p})}, \quad (2)$$

$$\hat{p} = \sum_{i=1}^{2N}\frac{H_i}{2N},$$

where H_i is the probability of the imputed allele in the i-th haplotype (varies from 0 to 1 and is estimated by built-in hidden Markov model algorithms); N is the sample size; $\hat{p}$ is the allele frequency estimate. The derivation of the formulas can be found in Supplementary Materials 2 and 3.

When calculating metrics of the second type, part of the information about genotypes in the sample under study is artificially "masked" (removed from the general data set, while maintaining information about these SNPs). Then the resulting gaps are imputed and compared with real genotypes. For instance, CR represents the proportion of correctly calculated SNPs to all SNPs. The Hellinger exponent is a measure of the distance between two genotype probability distributions and is based on the Bhattacharyya coefficient (Bhattacharyya, 1943), which measures the degree of overlap between two distributions. The SEN metric is the scaled Euclidean distance between the true and imputed dose distributions. Both the Hellinger score and the SEN score are calculated for each individual's distinct SNPs. IQS is based on Cohen's kappa statistic and allows for random co-occurrence between imputed and real SNPs.

As mentioned at the beginning, in addition to the listed metrics, a polygenic score (PGS) of the trait can be used to control the quality of imputation (Choi et al., 2020). It is a measure of an individual's genetic risk for a trait, obtained by summing the quantified effect of many common variants (typically with minor allele frequencies ≥ 1 %) in the genome, each of which may make a small contribution to an individual's genetic risk for that trait or disease. PGS is typically calculated as a weighted sum of a set of genetic variants, usually SNPs, defined as single base pair variations from a reference genome. The resulting score has a distribution close to normal in the general population, with higher scores indicating higher risk.

In general, the equation for calculating a weighted polygenic risk score for an individual is as follows (Collister et al., 2022):

$$\mathrm{PGS}_i = \sum_j^M \hat{\beta} * \mathrm{dosage}_{ij},$$

where M is the number of SNPs in the model; $\hat{\beta}_j$ is an estimate of the effect size of the j-th variant; dosage_{ij} is the genotype encoded 0, 1, 2 for the j-th variant in the genotype of the i-th individual. SNP effect sizes (β) are often obtained from GWAS results.

After calculating the PGS score for a trait, its values are compared with the values of real phenotypes. If there is a significant correlation between these two data sets, we can conclude that the data is of high quality after imputation.

Examples of imputation in genomic studies on Russian samples

Despite the advantages of imputation and phasing described above, there is very little reference to their use in studies of Russian samples. As such, in a 2023 study on depression in a sample of 4,520 individuals from various regions of Russia, imputation was carried out using the HRC and 1000G reference panels using Beagle 5.1 (Pinakhina et al., 2022). Similarly, in a study of the genetic structure of the Western Russian population (sample of 4,145 individuals), the HRC panel was chosen as the panel for imputation; the procedure itself was carried out using Beagle 4.0 and allowed to consider another 10,454,514 imputed genotyped variants in the analysis, in addition to 623,249 genotyped ones (Usoltsev et al., 2023). And in a 2022 study of markers associated with muscle strength and power in 292 Russians (83 of them professional athletes), not only imputation on a 1000G panel, but also phasing using SHAPEIT was carried out (Moreland et al., 2022).

As stated earlier, one of the most important factors for performing high-quality imputation is the correct choice of reference panel. The authors of one work (Kolosov et al., 2022) assessed the reliability of imputation of genotypes of a sample of 230 elderly people from St. Petersburg (501,100 SNP) by such panels as HRC, 1000G, HGDP (Human Genome Diversity Project (Cann et al., 2002) – a reference panel based on 929 people of various ethnic backgrounds). They were able

to increase the total number of variants studied to 37.6, 37.5 and 26.6 million SNPs for each of the panels, respectively, using Beagle 5.1 (the data were pre-phased). In addition, HRC, compared to the other two panels, showed the highest imputation accuracy (IQS and CR metrics).

All of these works use HRC or 1000G as reference panels, but this approach is somewhat outdated and is subject to revision due to the emergence of a larger TOPMed data set, the use of which serves as a kind of gold standard in international studies at the moment. As for the software, various versions of Beagle are used in the reviewed works.

In the mentioned studies on Russian samples, meta-analyses or fine mapping of genes were not carried out; however, as examples from other works show (Barton et al., 2021), thanks to imputation and phasing such analyzes can be done with a significant quality improvement.

Conclusion

From the above, we can conclude that, at the moment, imputation of genotyping data is an integral part of many human genomic studies, in particular GWAS. It provides an increase in the number of SNPs analyzed and makes it possible to combine the results of different studies. Imputation also significantly improves the results of fine mapping, allowing the most accurate identification of specific genetic variants and genes that determine the association of the entire genome region with the trait being studied (Chundru et al., 2019).

It is worth noting that for large-scale studies where sample size and genotyping coverage are important, the combination of DNA microarrays/sequencing with low coverage and further imputation is the most optimal and cheapest data acquisition strategy suitable for most genomic study designs. This combination is used in all major national biobanks, such as UK Biobank (Sudlow et al., 2015), AllOfUs (Ramirez et al., 2022) and others.

Along with the listed advantages, the imputation method has a number of disadvantages and limitations. In particular, reading errors due to low coverage, as well as incorrect selection of parameters for imputation along with an inappropriate reference panel, often lead to low accuracy of the imputed data, which can negatively affect the results of further stages of analysis. It must also be remembered that imputation uses information about haplotypes from the reference sample, so when it becomes outdated, genetic variants that have become frequent in the population relatively recently may be imputed worse (Ali et al., 2022). In addition, a high level of recombination reduces the accuracy of phasing and subsequent imputation of genotypes, and therefore, in some cases, additional recombination analysis is necessary (Weng et al., 2014).

Also, imputation can smooth out genetic differences between individuals in case-control samples (Lau et al., 2024): imputed data may introduce inaccurate genotypes in regions where differences between case and control are expected, and this effect appears regardless of how large and diverse the reference panel is. Finally, when using the method, it is important to remember that what is true for the population as a whole may not always be true for a specific individual.

Currently, there is a wide variety of programs and reference panels for imputation of human genomic data, and, as a consequence, many combinations of them. Due to this, researchers have the opportunity to select the optimal set of imputation tools for the characteristics of the sample and the objectives of a particular study. A review of works on Russian samples showed that the most popular software for imputation is Beagle of various versions, and among reference panels, HRC and 1000G are most often used, which is somewhat different from international practices, where the leader among reference panels is TOPMed.

Greater awareness of the intricacies of imputation and a deliberate approach to the selection of tools will improve the quality of genomic data without increasing the cost of obtaining them, facilitate their integration with the results of other studies, and provide more accurate information about the genetic control of human traits.

References

Abraham G., Qiu Y., Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*. 2017; 33(17):2776-2778. DOI 10.1093/bioinformatics/btx299

Ali A.T., Liebert A., Lau W., Maniatis N., Swallow D.M. The hazards of genotype imputation in chromosomal regions under selection: A case study using the lactase gene region. *Ann. Hum. Genet*. 2022; 86(1):24-33. DOI 10.1111/ahg.12444

Anderson C.A., Pettersson F.H., Clarke G.M., Cardon L.R., Morris A.P., Zondervan K.T. Data quality control in genetic case-control association studies. *Nat. Protoc*. 2010;5(9):1564-1573. DOI 10.1038/nprot.2010.116

Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Abecasis G.R., Bentley D.R., … Min Kang H., 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

Barton A.R., Sherman M.A., Mukamel R.E., Loh P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet*. 2021;53(8):1260-1269. DOI 10.1038/s41588-021-00892-1

Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhyā: Ind. J. Stat*. 1946;7(4):401-406

Bourke P.M., Voorrips R.E., Visser R.G.F., Maliepaard C. Tools for genetic studies in experimental populations of polyploids. *Front. Plant. Sci*. 2018;9:513. DOI 10.3389/fpls.2018.00513

Brown A., Ampratwum P.O., Ray S.D. Microarray analysis. In: Encyclopedia of Toxicology. 4 ed. 2024;6:385-392. DOI 10.1016/B978-0-12-824315-2.00210-4

Browning B.L., Zhou Y., Browning S.R. A One-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet*. 2018;103(3):338-348. DOI 10.1016/j.ajhg.2018.07.015

Browning B.L., Tian X., Zhou Y., Browning S.R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet*. 2021;108(10): 1880-1890. DOI 10.1016/j.ajhg.2021.08.005

Browning S.R., Browning B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet*. 2011;12(10):703-714. DOI 10.1038/nrg3054

Cann H.M., de Toma C., Cazes L., Legrand M.F., Morel V., Piouffre L., Bodmer J., … Zhu S., Weber J.L., Greely H.T., Feldman M.W., Thomas G., Dausset J., Cavalli-Sforza L.L. A human genome diversity cell line panel. *Science*. 2002;296(5566):261-262. DOI 10.1126/science.296.5566.261b

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1):7. DOI 10.1186/s13742-015-0047-8

Chat V., Ferguson R., Morales L., Kirchhoff T. Ultra low-coverage whole-genome sequencing as an alternative to genotyping arrays in

genome-wide association studies. *Front. Genet*. 2022;12:790445. DOI 10.3389/fgene.2021.790445

Check Hayden E. Genome sequencing: the third generation. *Nature*. 2009;457(7231):768-769. DOI 10.1038/news.2009.86

Choi S.W., Mak T.S.-H., O'Reilly P.F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc*. 2020;15(9):2759-2772. DOI 10.1038/s41596-020-0353-1

Chundru V.K., Marioni R.E., Prendergast J.G.D., Vallerga C.L., Lin T., Beveridge A.J., Gratten J., Hume D.A., Deary I.J., Wray N.R., Visscher P.M., McRae A.F. Examining the impact of imputation errors on fine-mapping using DNA methylation QTL as a model trait. *Genetics*. 2019;212(3):577-586. DOI 10.1534/genetics.118.301861

Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol*. 1990;7(2):111-122. DOI 10.1093/oxfordjournals.molbev.a040591

Collister J.A., Liu X., Clifton L. Calculating polygenic risk scores (PRS) in UK biobank: A practical guide for epidemiologists. *Front. Genet*. 2022;13:818574. DOI 10.3389/fgene.2022.818574

Connell C., Fung S., Heiner C., Bridgham J., Chakerian V., Heron E., Jones B., Menchen S., Mordan W., Raff M., Recknor M., Smith L.M., Springer J., Woo S., Hunkapiller M. Automated DNA-sequence analysis. *Biotechniques*. 1987;5:342-348

Das S., Forer L., Schönherr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.-R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R., Fuchsberger C. Next-generation genotype imputation service and methods. *Nat. Genet*. 2016;48(10):1284-1287. DOI 10.1038/ng.3656

De Marino A., Mahmoud A.A., Bose M., Bircan K.O., Terpolovsky A., Bamunusinghe V., Bohn S., Khan U., Novković B., Yazdi P.G. A comparative analysis of current phasing and imputation software. *PLoS One*. 2022;17(10):e0260177. DOI 10.1371/journal.pone.0260177

Deamer D., Akeson M., Branton D. Three decades of nanopore sequencing. *Nat. Biotechnol*. 2016;34(5):518-524. DOI 10.1038/nbt.3423

Delaneau O., Marchini J., Zagury J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods*. 2012;9(2):179-181. DOI 10.1038/nmeth.1785

Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Society*. 1977;39(1):1-38. DOI 10.1111/j.2517-6161.1977.tb01600.x

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*. 2011;43(5):491-498. DOI 10.1038/ng.806

Drmanac R., Sparks A.B., Callow M.J., Halpern A.L., Burns N.L., Kermani B.G., Carnevali P., … Drmanac S., Oliphant A.R., Banyai W.C., Martin B., Ballinger D.G., Church G.M., Reid C.A. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010;327(5961):78-81. DOI 10.1126/science.1181498

Fan J.B., Oliphant A., Shen R., Kermani B.G., Garcia F., Gunderson K.L., Hansen M., … Kruglyak S., Bentley D., Haas J., Rigault P., Zhou L., Stuelpnagel J., Chee M.S. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol*. 2003;68:69-78. DOI 10.1101/sqb.2003.68.69

Fatumo S., Mugisha J., Soremekun O.S., Kalungi A., Mayanja R., Kintu C., Makanga R., Kakande A., Abaasa A., Asiki G., Kalyesubula R., Newton R., Nyirenda M., Sandhu M.S., Kaleebu P. Uganda genome resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genom*. 2022;2(11):100209. DOI 10.1016/j.xgen.2022.100209

Feng Z., Peng F., Xie F., Liu Y., Zhang H., Ma J., Xing J., Guo X. Comparison of capture-based mtDNA sequencing performance between MGI and illumina sequencing platforms in various sample types. *BMC Genomics*. 2024;25(1):41. DOI 10.1186/s12864-023-09938-6

Govindarajan R., Duraiyan J., Kaliyappan K., Palanisamy M. Microarray and its applications. *J. Pharm. Bioallied Sci*. 2012;4(6):310. DOI 10.4103/0975-7406.100283

Gresham D., Dunham M.J., Botstein D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet*. 2008;9(4):291-302. DOI 10.1038/nrg2335

Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc*. 2014;9(11):2643-2662. DOI 10.1038/nprot.2014.174

Hayat M.A. DNA microarrays technology. In: Handbook of Immunohistochemistry and *in situ* Hybridization of Human Carcinomas. 2002;49-55. DOI 10.1016/S1874-5784(04)80015-1

Huang G.-H., Tseng Y.-C. Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proc*. 2014;8(S1):S64. DOI 10.1186/1753-6561-8-S1-S64

Jeon S.A., Park J.L., Park S.-J., Kim J.H., Goh S.-H., Han J.-Y., Kim S.-Y. Comparison between MGI and illumina sequencing platforms for whole genome sequencing. *Genes Genom*. 2021;43(7):713-724. DOI 10.1007/s13258-021-01096-x

Kolosov N., Rezapova V., Rotar O., Loboda A., Freylikhman O., Melnik O., Sergushichev A., Stevens C., Voortman T., Kostareva A., Konradi A., Daly M.J., Artomov M. Genotype imputation and polygenic score estimation in northwestern Russian population. *PLoS One*. 2022;17(6):e0269434. DOI 10.1371/journal.pone.0269434

Korostin D., Kulemin N., Naumov V., Belova V., Kwon D., Gorbachev A. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One*. 2020;15(3):e0230301. DOI 10.1371/journal.pone.0230301

Kurg A., Tõnisson N., Georgiou I., Shumaker J., Tollett J., Metspalu A. Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet. Test*. 2000;4(1):1-7. DOI 10.1089/109065700316408

Lam M., Awasthi S., Watson H.J., Goldstein J., Panagiotaropoulou G., Trubetskoy V., Karlsson R., Frei O., Fan C.-C., De Witte W., Mota N.R., Mullins N., Brügger K., Lee S.H., Wray N.R., Skarabis N., Huang H., Neale B., Daly M.J., Mattheisen M., Walters R., Ripke S. RICOPILI: rapid imputation for COnsortias PIpeLIne. *Bioinformatics*. 2020;36(3):930-933. DOI 10.1093/bioinformatics/btz633

Lamy P., Andersen C.L., Wikman F.P., Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res*. 2006;34(14):e100. DOI 10.1093/nar/gkl475

Lau W., Ali A., Maude H., Andrew T., Swallow D.M., Maniatis N. The hazards of genotype imputation when mapping disease susceptibility variants. *Genome Biol*. 2024;25(1):7. DOI 10.1186/s13059-023-03140-3

Li L., Huang P., Sun X., Wang S., Xu M., Liu S., Feng Z., Zhang Q., Wang X., Zheng X., Dai M., Bi Y., Ning G., Cao Y., Wang W. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res*. 2021;31(12):1308-1310. DOI 10.1038/s41422-021-00564-z

Li N., Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. 2003;165(4):2213-2233. DOI 10.1093/genetics/165.4.2213

Li Y., Willer C., Sanna S., Abecasis G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet*. 2009;10(1):387-406. DOI 10.1146/annurev.genom.9.081307.164242

Lin P., Hartz S.M., Zhang Z., Saccone S.F., Wang J., Tischfield J.A., Edenberg H.J., Kramer J.R., Goate A.M., Bierut L.J., Rice J.P. A new statistic to evaluate imputation reliability. *PLoS One*. 2010;5(3):e9697. DOI 10.1371/journal.pone.0009697

Loh P.-R., Danecek P., Palamara P.F., Fuchsberger C., Reshef Y.A., Finucane H.K., Schoenherr S., Forer L., McCarthy S., Abecasis G.R., Durbin R., L Price A. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet*. 2016;48(11):1443-1448. DOI 10.1038/ng.3679

Marchini J., Howie B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet*. 2010;11(7):499-511. DOI 10.1038/nrg2796

Marees A.T., de Kluiver H., Stringer S., Vorspan F., Curis E., Marie Claire C., Derks E.M. A tutorial on conducting genome wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res*. 2018;27(2). DOI 10.1002/mpr.1608

Martin A.R., Atkinson E.G., Chapman S.B., Stevenson A., Stroud R.E., Abebe T., Akena D., … Ramesar R., Shiferaw W., Stein D.J., Teferra S., van der Merwe C., Zingela Z. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* 2021;108(4):656-668. DOI 10.1016/j.ajhg.2021.03.012

Maxam A.M., Gilbert W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA.* 1977;74(2):560-564. DOI 10.1073/pnas.74.2.560

Mills M.C., Barban N., Tropf F.C. An Introduction to Statistical Genetic Data Analysis. Cambridge, MA: MIT Press, 2020

Mirzabekov A.D. Biochips in the biology and medicine of the XXI century. *Vestnik Rossiyskoj Akademii Nauk = Herald of the Russian Academy of Sciences*. 2003;73(5):412 (in Russian)

Moreland E., Borisov O.V., Semenova E.A., Larin A.K., Andryushchenko O.N., Andryushchenko L.B., Generozov E.V., Williams A.G., Ahmetov I.I. Polygenic profile of elite strength athletes. *J. Strength. Cond. Res*. 2022;36(9):2509-2514. DOI 10.1519/JSC.0000000000003901

O'Connell J., Yun T., Moreno M., Li H., Litterman N., Kolesnikov A., Noblin E., … Wang W., Weldon C.H., Wilton P., Wong C., Auton A., Carroll A., McLean C.Y. A population-specific reference panel for improved genotype imputation in African Americans. *Commun. Biol*. 2021;4(1):1269. DOI 10.1038/s42003-021-02777-9

Pasaniuc B., Rohland N., McLaren P.J., Garimella K., Zaitlen N., Li H., Gupta N., … Haas D.W., Liang L., Sunyaev S., Patterson N., de Bakker P.I.W., Reich D., Price A.L. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet*. 2012;44(6):631-635. DOI 10.1038/ng.2283

Pinakhina D., Yermakovich D., Vergasova E., Kasyanov E., Rukavishnikov G., Rezapova V., Kolosov, … Plotnikov N., Ilinsky V., Neznanov N., Mazo G., Kibitov A., Rakitko A., Artomov M. GWAS of depression in 4,520 individuals from the Russian population highlights the role of MAGI2 (S-SCAM) in the gut-brain axis. *Front. Genet*. 2022;13:972196. DOI 10.3389/fgene.2022.972196

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. 2006;38(8):904-909. DOI 10.1038/ng1847

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 analyses. *Am. J. Hum. Genet*. 2007;81(3):559-575. DOI 10.1086/519795

Ramirez A.H., Sulieman L., Schlueter D.J., Halvorson A., Qian J., Ratsimbazafy F., Loperena R., … Denny J.C., Carroll R.J., Glazer D., Harris P.A., Hripcsak G., Philippakis A., Roden D.M.; All of Us research program. The *All of Us* research program: Data quality, utility, and diversity. *Patterns* (*N Y*). 2022;3(8):100570. DOI 10.1016/j.patter.2022.100570

Rhoads A., Au K.F. PacBio Sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13(5):278-289. DOI 10.1016/j.gpb.2015.08.002

Roshyara N.R., Kirsten H., Horn K., Ahnert P., Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet*. 2014;15(1):88. DOI 10.1186/s12863-014-0088-5

Rubinacci S., Delaneau O., Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet*. 2020;16(11):e1009049. DOI 10.1371/journal.pgen.1009049

Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*. 1977;74(12):5463-5467. DOI 10.1073/pnas.74.12.5463

Scheet P., Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet*. 2006;78(4):629-644. DOI 10.1086/502802

Shendure J., Balasubramanian S., Church G.M., Gilbert W., Rogers J., Schloss J.A., Waterston R.H. DNA sequencing at 40: past, present and future. *Nature*. 2017;550(7676):345-353. DOI 10.1038/nature24286

Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B.H., Hood L.E. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321(6071):674-679. DOI 10.1038/321674a0

Stahl K., Gola D., König I.R. Assessment of imputation quality: comparison of phasing and imputation algorithms in real data. *Front. Genet*. 2021;12:724037. DOI 10.3389/fgene.2021.724037

Sudlow C., Gallacher J., Allen N., Beral V., Burton P., Danesh J., Downey P., Elliott P., Green J., Landray M., Liu B., Matthews P., Ong G., Pell J., Silman A., Young A., Sprosen T., Peakman T., Collins R. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779. DOI 10.1371/journal.pmed.1001779

Sudmant P.H., Rausch T., Gardner E.J., Handsaker R.E., Abyzov A., Huddleston J., Zhang Y., … Gerstein M.B., Bashir A., Stegle O., Devine S.E., Lee C., Eichler E.E., Korbel J.O. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. DOI 10.1038/nature15394

Taliun D., Harris D.N., Kessler M.D., Carlson J., Szpiech Z.A., Torres R., … Cupples L.A., Laurie C.C., Jaquish C.E., Hernandez R.D., O'Connor T.D., Abecasis G.R. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299. DOI 10.1038/s41586-021-03205-y

The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 2016;48:1279-1283. DOI 10.1038/ng.3643

Usoltsev D., Kolosov N., Rotar O., Loboda A., Boyarinova M., Moguchaya E., Kolesova E., … Laiho P., Kostareva A., Konradi A., Shlyakhto E., Palotie A., Daly M.J., Artomov M. Understanding complex trait susceptibilities and ethnical diversity in a sample of 4,145 Russians through analysis of clinical and genetic data. *bioRxiv*. 2023. DOI 10.1101/2023.03.23.534000

Wall J.D., Stawiski E.W., Ratan A., Kim H.L., Kim C., Gupta R., Suryamohan K., … Radha V., Mohan V., Majumder P.P., Seshagiri S., Seo J.-S., Schuster S.C., Peterson A.S. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019;576(7785):106-111. DOI 10.1038/s41586-019-1793-z

Wang D.G., Fan J.-B., Siao C.-J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., Lipshutz R., Chee M., Lander E.S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science.* 1998;280(5366):1077-1082. DOI 10.1126/science.280.5366.1077

Wang Q.S., Huang H. Methods for statistical fine-mapping and their applications to auto-immune diseases. *Semin. Immunopathol*. 2022;44(1):101-113. DOI 10.1007/s00281-021-00902-8

Weale M.E. A survey of current software for haplotype phase inference. *Hum. Genomics*. 2004;1(2):141. DOI 10.1186/1479-7364-1-2-141

Weng Z.-Q., Saatchi M., Schnabel R.D., Taylor J.F., Garrick D.J. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Gen. Select. Evol*. 2014;46(1):34. DOI 10.1186/1297-9686-46-34

Wu D., Dou J., Chai X., Bellis C., Wilm A., Shih C.C., … Wong W.-C., Xie Z., Yeo K.K., Zhang L., Zhai W., Zhao Y. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*. 2019;179(3):736-749.e15. DOI 10.1016/j.cell.2019.09.019

Yang H.-C., Lin H.-C., Kang M., Chen C.-H., Lin C.-W., Li L.-H., Wu J.-Y., Chen Y.-T., Pan W.-H. SAQC: SNP array quality control. *BMC Bioinformatics*. 2011;12(1):100. DOI 10.1186/1471-2105-12-100

Yoo S.-K., Kim C.-U., Kim H.L., Kim S., Shin J.-Y., Kim N., Yang J.S.W., Lo K.-W., Cho B., Matsuda F., Schuster S.C., Kim C., Kim J.-I., Seo J.-S. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med*. 2019;11(1):64. DOI 10.1186/s13073-019-0677-z

Yu K., Das S., LeFaive J., Kwong A., Pleiness J., Forer L., Schönherr S., Fuchsberger C., Smith A.V., Abecasis G.R. Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet*. 2022;109(6):1007-1015. DOI 10.1016/j.ajhg.2022.04.002

Zhao S., Jing W., Samuels D.C., Sheng Q., Shyr Y., Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform*. 2018;19(5):765-775. DOI 10.1093/bib/bbx012

Conflict of interest. The authors declare no conflict of interest.
Received December 12, 2023. Revised April 23, 2024. Accepted July 4, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):640-649

DOI 10.18699/vjgb-24-71

Search for signals of positive selection of circadian rhythm genes *PER1*, *PER2*, *PER3* in different human populations

A.I. Mishina, S.Y. Bakoev, A.Y. Oorzhak, A.A. Keskinov, Sh.Sh. Kabieva, A.V. Korobeinikova, V.S. Yudin, M.M. Bobrova, D.A. Shestakov, V.V. Makarov, L.V. Getmantseva

Centre for Strategic Planning and Management of Biomedical Health Risks of the Federal Medical Biological Agency, Moscow, Russia
arinamishina32@yandex.ru

Abstract. The diversity of geographically distributed human populations shows considerable variation in external and internal traits of individuals. Such differences are largely attributed to genetic adaptation to various environmental influences, which include changes in climatic conditions, variations in sleep and wakefulness, dietary variations, and others. Whole-genome data from individuals of different populations make it possible to determine the specific genetic sites responsible for adaptations and to further understand the genetic structure underlying human adaptive characteristics. In this article, we searched for signals of single nucleotide polymorphisms (SNPs) under selection pressure in people of different populations. To identify selection signals in different population groups, the *PER1*, *PER2* and *PER3* genes that are involved in the coordination of thermogenic functions and regulation of circadian rhythms, which is directly reflected in the adaptive abilities of the organism, were investigated. Data were analyzed using publicly available data from the 1000 Genomes Project for 23 populations. The Extended Haplotype Homozygosity Score statistical method was chosen to search for traces of selection. The comparative analysis performed identified points subject to selection pressure. The SNPs were annotated through the GWAS catalog and manually by analyzing Internet resources. This study suggests that living conditions, climate, and other external factors directly influence the genetic structure of populations and vary across races and geographic locations. In addition, many of the selection variants in the *PER1*, *PER2*, *PER3* genes appear to regulate biological processes that are associated with major modern diseases, including obesity, cancer, metabolic syndrome, bipolar personality disorder, depression, rheumatoid arthritis, diabetes mellitus, lupus erythematosus, stroke and Alzheimer's disease, making them extremely interesting targets for further research aimed at identifying the genetic causes of human disease.
Key words: populations; SNP; adaptation; *PER1*; *PER2*; *PER3*.

For citation: Mishina A.I., Bakoev S.Y., Oorzhak A.Y., Keskinov A.A., Kabieva Sh.Sh., Korobeinikova A.V., Yudin V.S., Bobrova M.M., Shestakov D.A., Makarov V.V., Getmantseva L.V. Search for signals of positive selection of circadian rhythm genes *PER1*, *PER2*, *PER3* in different human populations. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):640-649. DOI 10.18699/vjgb-24-71

Author's contribution. All authors contributed to conceptualization, methodology, investigation, writing original draft preparation, and writing review and editing. All authors have read and agreed to the published version of the manuscript.

Поиск сигналов положительного отбора генов циркадных ритмов *PER1*, *PER2*, *PER3* в различных популяциях людей

А.И. Мишина, С.Ю. Бакоев, А.Ю. Ооржак, А.А. Кескинов, Ш.Ш. Кабиева, А.В. Коробейникова, В.С. Юдин, М.М. Боброва, Д.А. Шестаков, В.В. Макаров, Л.В. Гетманцева

Центр стратегического планирования и управления медико-биологическими рисками здоровью Федерального медико-биологического агентства, Москва, Россия
arinamishina32@yandex.ru

Аннотация. Разнообразие географически распределенных человеческих популяций демонстрирует большую вариацию внешних и внутренних признаков индивидов. Такие различия в значительной степени объясняются генетической адаптацией к различным воздействиям окружающей среды, к которым относят изменения климатических условий, колебания условий сна и бодрствования, вариации рациона и другие. Полногеномные данные, полученные от людей различных популяций, дают возможность идентифицировать конкретные генетические участки, ответственные за эти адаптации, и глубже понимать генетическую

структуру, лежащую в основе адаптивных характеристик человека. В данной работе проведен поиск сигналов однонуклеотидных полиморфизмов (SNP), находящихся под давлением отбора у людей различных популяций. Для выявления сигналов отбора в различных популяционных группах были исследованы гены *PER1*, *PER2* и *PER3*, играющие важнейшую роль в координации термогенных функций и регуляции циркадных ритмов, что напрямую отражается на адаптационных способностях организма. Анализ данных осуществляли на основе общедоступных данных из проекта «1000 геномов» (1000 Genomes Project) по 23 популяциям. Для поиска следов отбора был выбран статистический метод XP-EHH (expanded haplotype homozygosity score). Проведенный сравнительный анализ позволил идентифицировать точки, подверженные давлению отбора. Найденные SNP были аннотированы через каталог GWAS, а также вручную, путем анализа интернет-ресурсов и публикаций. Исследование позволяет сделать вывод о том, что условия проживания, климат и другие внешние факторы напрямую влияют на генетическую структуру популяций и варьируют в зависимости от расы и географического местоположения. Кроме того, многие из вариантов отбора в генах *PER1*, *PER2*, *PER3*, по-видимому, регулируют биологические процессы, связанные с основными современными заболеваниями, включая ожирение, онкологию, метаболический синдром, биполярное расстройство личности, депрессию, ревматоидный артрит, сахарный диабет, красную волчанку, инсульт и болезнь Альцгеймера, что делает их крайне интересными объектами для дальнейших исследований, направленных на идентификацию генетически обусловленных причин заболеваний человека.
Ключевые слова: популяции; SNP; адаптация; *PER1*; *PER2*; *PER3*.

Introduction

Advances in SNP genotyping methods have led to a rapid shift from studies focused on spatially explicit neutral genetic processes to those focused on adaptive genetic processes (Ahrens et al., 2018). One tool for tracking these processes is the search of loci under selection pressure (Carlson et al., 2005). Unique genetic patterns or traces left in genomic regions subjected to selection are called selection signatures (Nielsen, 2005; Jensen et al., 2016; Bakoev et al., 2021). Selection signatures are genomic regions containing DNA sequences functionally involved in the genetic variability of the traits subject to selection (Lopez et al., 2015; Bakoev et al., 2023). Such parts are of interest because of their relevance for tracing evolutionary biology and potential links to genes that control phenotypes in wild and domestic populations (Xu et al., 2015).

Various statistical approaches have been used to identify loci under selection pressure, one of them being extended haplotype homozygosity (EHH) analysis. It should be noted that the word "homozygosity", as part of the term EHH, refers to the probability that two randomly selected chromosomes from a population are identical (at a particular locus or region) (Klassmann, Gautier, 2022). The result interpreted from the theory is that major haplotypes with unusually high EHH and high population frequency indicate the presence of a mutation that became prominent in the human gene pool faster than expected under neutral evolution (Sabeti et al., 2002).

To study the genetic diversity and evolution of human populations, the XP-EHH (Extended Haplotype Homozygosity Score) method is well established to identify potential sites of genetic variation that may be associated with adaptation to different environments and conditions (Voight et al., 2006).

The *PER1*, *PER2* and *PER3* genes are involved in the coordination of circadian rhythms, regulation of the body's adaptive abilities, and are also associated with various diseases (Liberman et al., 2017; Rijo-Ferreira, Takahashi, 2019). For example, a study found a high association of *PER2* gene expression with the adaptation of organisms to low temperatures. S. Chappuis and co-authors (Chappuis et al., 2013) proved that mice with the *Period2* (*PER2*) gene turned off are sensitive to cold because their adaptive thermogenesis system becomes less efficient. Regarding the *PER1* gene, Y. Shi et al. (2021) claim that light adaptation generated by the CRTC1-SIK1 pathway, in which the *PER1* gene is involved, in the suprachiasmatic nucleus provides a robust mechanism that allows the circadian system to maintain homeostasis in the presence of light perturbations. This mechanism appears to be important for rapid adaptation to changing environmental conditions. According to the findings of L. Zhang et al. (2013), a polymorphism in the *PER3* gene is associated with the level of adaptation to shift work schedules and alternating sleep phases in nurses working in shifts.

Thus, genes from the *PER* group are a promising target for finding signals of positive selection in different human populations. In addition, the existence of a link between adaptive abilities, selection signals and major modern diseases is of interest.

Materials and methods

Public data from The 1000 Genomes Project Consortium (1000 Genomes, 2008) representing 23 populations grouped into their respective clusters were used for analysis (see the Table).

Plink 1.9 (Purcell et al., 2007) was used to merge all data. Using bcftools, we removed SNP duplicates and SNPs with identical positions, and normalized all data according to the GRCh38 reference. Start and end positions for the *PER1*, *PER2*, and *PER3* gene regions (GRCh 38 assembly) were obtained from NCBI (National Library of Medicine (USA)).

The XP-EHH (Extended Haplotype Homozygosity Score) method implemented in the selscan program (Szpiech, 2021) was used to identify selection signals. Non-standardized scores were normalized using the "norm" script provided in the selscan program. SNPs with values crit = 1/–1 were considered as genetic variants under selection pressure (outliers) (crit = 1 – ancestral allele under selection pressure, crit = –1 – derived allele).

Populations from the 1000 Genomes Project selected for analysis

Group	*N*	Place of residence/ethnic identity
		Africans (AFR)
ESN	97	Southern Nigeria (Esan in Nigeria)
GWD	113	Western District of The Gambia (Gambian in Western Division – Mandinka)
MSL	83	Sierra Leone (Mende in Sierra Leone)
YRI	108	Ibadan, Nigeria (Yoruba in Ibadan, Nigeria)
LWK	108	Webuye Bungoma County in western Kenya (Luhya in Webuye, Kenya)
		Europeans (EUR)
GBR	89	UK (British from England and Scotland) / UK control population
FIN	96	Finland (Finnish in Finland) / Finns
TSI	107	Tuscany, Italy (Toscani in Italia) / Tuscans
IBS	107	Spain (Iberian Populations in Spain) / Spanish
		Mixed-race Americans (AMR)
CLM	94	Medellin Metropolitan Area, Colombia (Colombian in Medellín, Colombia)
MXL	64	Los Angeles, California, USA (Mexican Ancestry in Los Angeles, CA, USA)
PEL	85	Lima-Callao Metropolitan Area, Peru (Peruvian in Lima, Peru)
PUR	104	Puerto Rico (Puerto Rican in Puerto Rico)
		East Asians (EAS)
GIH	103	Houston metropolitan area, Texas, USA (Gujarati Indians in Houston, Texas, USA)
STU	102	UK (Sri Lankan Tamil in the UK)
ITU	102	UK (Indian Telugu in the UK)
PJL	92	Lahore, Pakistan (Punjabi in Lahore, Pakistan)
BEB	84	Bangladesh (Bengali in Bangladesh)
		South Asians (SAS)
CHS	105	Hunan and Fujian Province of South China (Han Chinese South, China)
CHB	103	Residential area of Beijing Normal University (Han Chinese in Beijing, China)
CDX	93	Xishuangbanna Health School Community in Xishuangbanna, Yunnan, China (Chinese Dai in Xishuangbanna, China)
KHV	96	Ho Chi Minh City, Vietnam (Kinh in Ho Chi Minh City, Vietnam)
JPT	103	Tokyo metropolitan area (Japanese in Tokyo, Japan)

Selection signals were determined by inter-population comparisons, using YRIs from the African cluster as the comparison group. This allowed us to determine the outliers between the Yoruba African population (YRI) from Ibadan and other groups (including the African cluster, namely ESN, GWD, MSL and LWK). In addition, selection choices related to within-cluster variability were also of interest. For this purpose, a comparison group was selected in each cluster and analyzed with other groups in the same cluster. Thus, in the EUR cluster, GBR was taken as the comparison group and accordingly analyzed between GBR&FIN, GBR&IBS and GBR&TSI. In the AMR cluster, the PUR group was taken and analyzed between PUR&CLM, PUR&MXL and PUR&PEL. In the EAS and SAS clusters, CHB and BEB groups were defined, respectively, and analyzed between CHB&CDX, CHB&CHS, CHB&KHV, CHB&JPT and BEB&PJL, BEB&ITU, BEB&STU, BEB&GIH, respectively.

Results and their discussion

Genomics and molecular biology have strongly influenced research on "selection and adaptation" through the identification of the genetic basis of various traits associated with

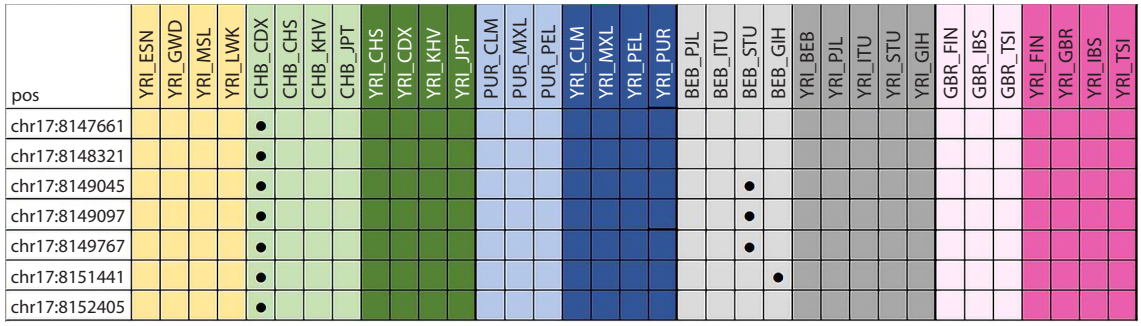


Fig. 1. Genetic variants under selection pressure in the *PER1* gene.
Here and in Fig. 2 and 3 pos – position.

maintenance and health in humans and animals (Hancock et al., 2010; Gintis et al., 2012). Alongside this, the results of the "genetic and genomic revolution" have enabled genome sequencing and provided new tools to measure both past and possibly ongoing adaptations (Zheng et al., 2023).

Human behavior is assumed to be determined by the interaction between nature and societal development (Saravanan et al., 2020). It can be assumed that the features of genetic structure in different human populations, including those associated with the exit of people from Africa, further formed the basis of individual features of human development (Benton et al., 2021). Thus, the *PER1*, *PER2*, and *PER3* genes we considered showed signals of positive selection, some of which were seen in several populations (these variants are mainly localized in the *PER2* gene), while others were found in only one population.

Analysis of the full-genome profiles of the studied populations revealed 110 loci (78 points in the *PER2* gene, 25 in *PER3* and 7 in *PER1*) under selection pressure. When analyzing the *PER1* gene, eight outliers were detected in an intergroup comparison of South Asians living in Beijing (CHB) with South Asians living in Yunan (CDX) (Fig. 1). In addition, four selection pressure sites were also identified in an intergroup comparison of East Asians living in Bangladesh (BEB) with East Asians living in Sri Lanka (STU) and East Asians living in Bangladesh (BEB) with East Asians living in Houston, Texas, Gujarat (GIH) (Fig. 1).

The sites discovered are involved in processes such as: predisposition to the development of major depressive disorder, Parkinson's disease, Alzheimer's disease, alcohol addiction, and breast cancer, as well as longevity (see Supplementary Material)[1].

We would like to pay attention to points under selection pressure in several of the compared groups. Such SNPs were identified in an intergroup comparison of East and South Asians. The points found suggest that similar external factors acted on the compared groups, which had the same effect necessary for the adaptation of the ethnic groups under study. Significant signals at positions chr17:8149097 (predisposition to breast cancer) are worth noting. It is possible that the fixation of alleles in the comparison groups of East Asians from Bangladesh (BEB) and East Asians from Srilanka (STU), and South Asians from Beijing (CHB) and South Asians from Yunnan (CDX) could have occurred due to the prevalence of humid climate in the territories where the ethnicities studied lived. According to some authors, humid climate may be a risk factor in the development of a number of cancers (Maryanaji, 2020; Guo et al., 2021; Pan et al., 2023).

In the analysis of the *PER2* gene, 78 points under selection pressure were identified. Outliers were identified while comparing all of the ethnicities studied with Africans and in single sites in the intergroup comparison of South Asians and Europeans (Fig. 2). Analyzing the points in the compared population groups, it can be concluded that the African population is strongly differentiated from the other ethnicities studied. Annotation of sites under selection pressure in several of the compared groups revealed SNP involvement in the formation of chronotypes, sleep coordination, predisposition to diabetes, stroke, lupus erythematosus and bipolar disorder, intestinal cholesterol absorption, and associations with metabolic phenotype. The associations of SNPs with various diseases and phenotypes in humans are presented in more detail in Supplementary Material.

The presence of the total number of outliers when comparing the population group of Africans and other ethnic groups indicates a long period of influence of certain external factors on all the studied populations. It is interesting to note that all alleles under selection pressure turned out to be derived variants. Since the leading function of the *PER2* gene is the formation of chronotypes, it can be assumed that the finding of the total array of points under selection pressure is also explained by the action of external factors inherent in the area where the studied ethnic groups lived. Such factors include the total number of daylight hours, magnetic field action, climatic peculiarities and others.

Identification of loci under selection pressure in several compared groups between South Asians and Africans, as well as within the groups of South Asians living in Beijing (CHB) and Yunnan (CDX) revealed SNPs responsible for predisposition to the development of a number of gastrointestinal and cardiovascular diseases. At the same time, derived alleles are identified in groups comparing South Asians with Africans,

[1] Supplementary Material is available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Mishina_Engl_28_6.pdf

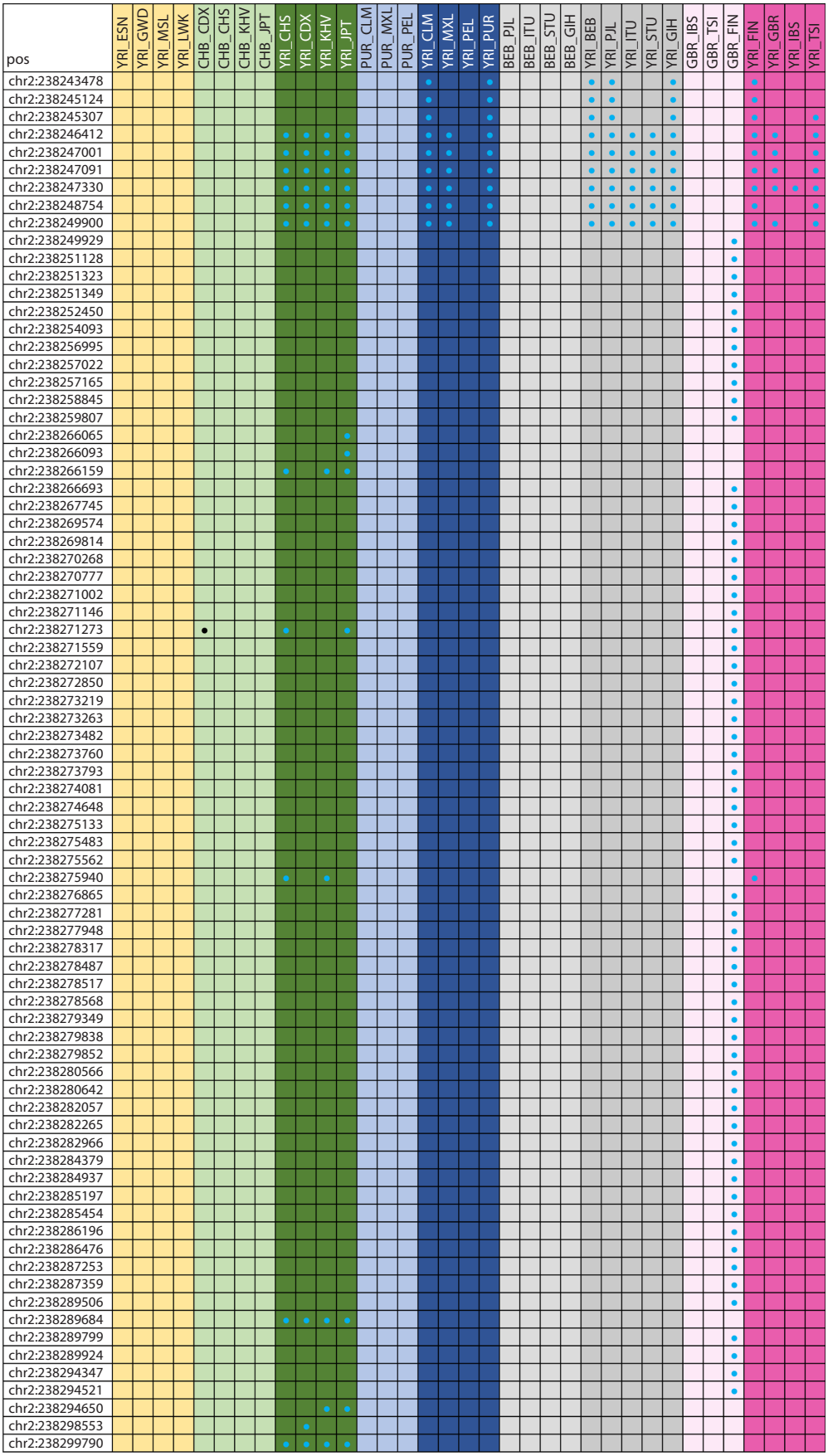


Fig. 2. Genetic variants under selection pressure in the *PER2* gene.

Dots represent variants under selection pressure; black color of the dots means that the ancestral allele was under selection pressure, blue color stands for the derived allele.

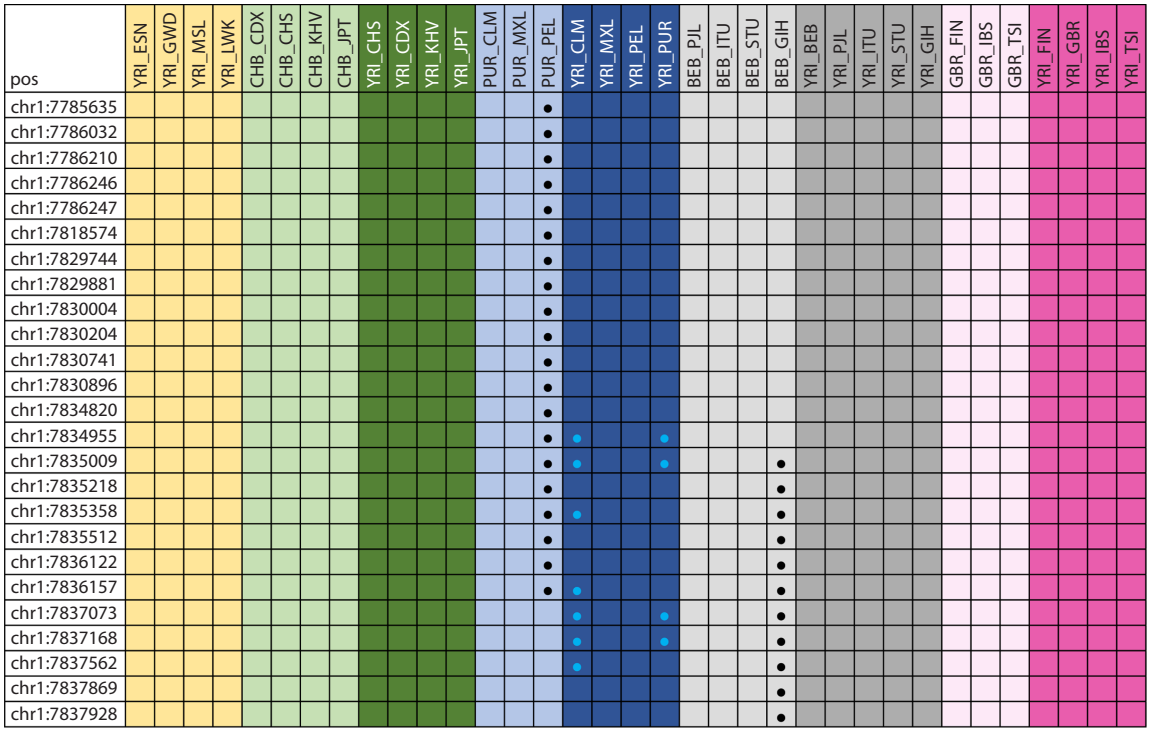


Fig. 3. Genetic variants under selection pressure in the *PER3* gene.

Dots represent variants under selection pressure; black color of the dots means that the ancestral allele was under selection pressure, blue color stands for the derived allele.

and when comparing South Asians within subgroups based on the region of residence, an ancestral allele is detected. Most interesting is the outlier in chr2:238289684 that was found when comparing South Asians to Africans: it is associated with systemic lupus erythematosus, which is caused by disorders such as hormonal imbalance during puberty, stress, and environmental factors, namely sun exposure and viral infections (Quaglia et al., 2021; Kim et al., 2022; Molina et al., 2022).

In our opinion, different levels of viral load as well as authentic climatic conditions may have played a key role in the development of adaptive abilities of these ethnic groups, thus fixing these alleles in the studied populations. The same theory may explain the fixation of loci associated with gastrointestinal diseases. As people migrated from the African continent to other areas, their gastronomic preferences changed, thus modifying the gut microbial ecosystem (Clemente et al., 2015; Syromyatnikov et al., 2022). This suggests that gastrointestinal diseases differed between South Asians and Africans due to differences in the gut microbiome (Donin et al., 2010; Porras et al., 2021).

This study identified 42 points of selection pressure in the *PER3* gene when comparing East Asians living in Bangladesh (BEB) with East Asians living in the Houston, Texas area (GIH), mixed Americans living in Puerto Rico (PUR) with mixed Americans living in Peru (PEL), and when comparing Africans with YRI_CLM Colombians and Africans with YRI_PUR Puerto Ricans (Fig. 3). After annotation, the following associations with the SNPs found were identified: response to the use of lithium medications in the treatment of bipolar disorder, formation of chronotypes of different types, predisposition to depressive disorders, predisposition to metabolic syndrome, likelihood of developing colorectal cancer, and predisposition to obesity. More details of SNP associations with different diseases and phenotypes in humans are presented in Supplementary Material.

The main function of the sites identified by us as being under selection pressure was the formation of the morning-type chronotype. It is worth noting that the internal comparison of the groups of mixed Americans and East Asians identified ancestral alleles, while the comparison of Africans with mixed Americans identified derived alleles. Perhaps the key difference between Africans and mixed Americans is the sleep specificity of these populations. For example, mixed Americans are more likely to have an evening chronotype while Africans have the most frequent morning chronotype (Egan et al., 2017). This may be due to the influence of various external factors such as latitude, longitude, magnetic field action or solar activity.

In addition, the data obtained suggest the influence of external factors on the formation of the studied populations, which, as a result, led to different action of mechanisms of their adaptive abilities. For example, the isolation of the Lima-Callao mixed American (PEL) population from Africans may be due to the remoteness of location of this group of people compared to the other ethnic groups under study. It is reliably known that ethnicities from other parts of Latin America were subjected to more frequent mixing with Europeans compared

to those from Peru (Chacón-Duque et al., 2018). Thus, the identity of the resulting population formed the most isolated genetic cluster.

Discussion

Progressive statistical methods aimed at finding loci under selection pressure have allowed scientists from different countries to conduct studies on this topic. In the authors' works, there are references to individual SNPs that we identified in this study as being under selection pressure. In total, we annotated 35 such sites.

Researchers have done the work of annotating SNPs, finding association with polymorphisms at these sites and correlation with some diseases and physiological features. For example, S.E. Jones et al. analyzed behavioral indicators of circadian rhythms by analyzing whole-genome data in 697,828 residents of the United Kingdom (UK). The study uncovered novel loci associated with the morning-type chronotype. Among these loci, rs58574366 (2:238286196) was identified. Our analyses revealed that this SNP is under selection pressure in comparison groups of Europeans from the United Kingdom (GBR) with Europeans from Finland (FIN). The negative values of the xpehh calculation indices led us to conclude that derived alleles were detected in the two compared samples (Jones et al., 2019).

Another point of interest is rs74508725 (2:238278568). This outlier is found when comparing groups of Europeans from the United Kingdom (GBR) with Europeans from Finland (FIN) and carries negative values, which may indicate differentiation of this site within the studied groups. In the works of G. Kichaev and co-authors (Kichaev et al., 2019), this locus was associated with the phenotype expressed in participants' height.

Locus rs2585399 (17:8151441) was identified by us as being under selection pressure when comparing several groups of people under study at once. These groups include comparisons of East Asians from Bangladesh (BEB) with East Asians from Texas (GIH) and South Asians from Beijing (CHB) with South Asians from Yunan (CDX). An interesting fact is that this SNP was associated with major depressive disorder in the authors' study. Transcriptome association analysis revealed significant associations with NEGR1 expression in the hypothalamus and DRD2 expression in the contiguous nucleus (Levey et al., 2021).

Another selection signal studied previously was rs228654 (1:7837168). However, it is worth noting that comparisons between African and mixed American YRI (Ibadané, Nigeria) to CLM (Medellín, Colombia) and YRI (Ibadané, Nigeria) to PUR (Ruerto Rico) populations revealed negative EHH values, suggesting the presence of a derived allele between the groups. In contrast, positive selection values were found between the groups of East Asians living in Texas (BEB) and East Asians living in Bangladesh (GIH), indicating the presence of an ancestral allele. A group of researchers led by P.R. Jansen (Jansen et al., 2019) analyzed the human genome to gain insights into the pathways, tissues, and cell types involved in the regulation of insomnia. The single nucleotide polymorphism rs228654 was among the loci associated with the development of this disease.

Conclusion

This study suggests that living conditions, climate, and other external factors directly influence the genetic structure of populations and vary by race and geographic location. In addition, many of the selection variants in the *PER1*, *PER2*, *PER3* genes appear to regulate biological processes that are associated with major modern diseases including obesity, cancer, metabolic syndrome, bipolar personality disorder, depression, rheumatoid arthritis, diabetes mellitus, lupus erythematosus, stroke and Alzheimer's disease, making them extremely interesting targets for further research aimed at identifying causal variants of human diseases, including cardiometabolic and psychiatric disorders, as well as cancer.

References

1000 Genomes. [WWW Document]. 2008. URL: https://www.ncbi.nlm.nih.gov/projects/faspftp/1000genomes/ (accessed 9.6.23)

Ahrens C.W., Rymer P.D., Stow A., Bragg J., Dillon S., Umbers K.D.L., Dudaniec R.Y. The search for loci under selection: trends, biases and progress. *Mol. Ecol.* 2018;27(6):1342-1356. DOI 10.1111/mec.14549

Azevedo P.G., Miranda L.R., Nicolau E.S., Alves R.B., Bicalho M.A.C., Couto P.P., Ramos A.V., Souza R.P., Longhi R., Friedman E., Marco L., Bastos-Rodrigues L. Genetic association of the *PERIOD3* (*PER3*) Clock gene with extreme obesity. *Obes. Res. Clin. Pract.* 2021;15(4):334-338. DOI 10.1016/j.orcp.2021.06.006

Bacalini M.G., Palombo F., Garagnani P., Giuliani C., Fiorini C., Caporali L., Stanzani Maserati M., Capellari S., Romagnoli M., De Fanti S., Benussi L., Binetti G., Ghidoni R., Galimberti D., Scarpini E., Arcaro M., Bonanni E., Siciliano G., Maestri M., Guarnieri B.; Italian Multicentric Group on clock genes, actigraphy in AD; Martucci M., Monti D., Carelli V., Franceschi C., La Morgia C., Santoro A. Association of rs3027178 polymorphism in the circadian clock gene *PER1* with susceptibility to Alzheimer's disease and longevity in an Italian population. *GeroScience*. 2022;44(2):881-896. DOI 10.1007/s11357-021-00477-0

Bakoev S., Getmantseva L., Kostyunina O., Bakoev N., Prytkov Y., Usatov A., Tatarinova T.V. Genome-wide analysis of genetic diversity and artificial selection in Large White pigs in Russia. *PeerJ.* 2021;9:e11595. DOI 10.7717/peerj.11595

Bakoev S.Y., Korobeinikova A.V., Mishina A.I., Kabieva S.S., Mitrofanov S.I., Ivashechkin A.A., Akinshina A.I., Snigir E.A., Yudin S.M., Yudin V.S., Getmantseva L.V., Anderzhanova E.A. Genomic signatures of positive selection in human populations of the *OXT*, *OXTR*, *AVP*, *AVPR1A* and *AVR1B* gene variants related to the regulation of psychoemotional response. *Genes* (*Basel*). 2023;14(11): 2053. DOI 10.3390/genes14112053

Baranger D.A.A., Ifrah C., Prather A.A., Carey C.E., Corral-Frías N.S., Drabant Conley E., Hariri A.R., Bogdan R. PER1 rs3027172 genotype interacts with early life stress to predict problematic alcohol use, but not reward-related ventral striatum activity. *Front. Psychol.* 2016;7:464. DOI 10.3389/fpsyg.2016.00464

Benton M.L., Abraham A., LaBella A.L., Abbot P., Rokas A., Capra J.A. The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* 2021;22(5):269-283. DOI 10.1038/s41576-020-00305-9

Biscontin A., Zarantonello L., Russo A., Costa R., Montagnese S. Toward a molecular approach to chronotype assessment. *J. Biol. Rhythms*. 2022;37(3):272-282. DOI 10.1177/07487304221099365

Blomeyer D., Buchmann A.F., Lascorz J., Zimmermann U.S., Esser G., Desrivieres S., Schmidt M.H., Banaschewski T., Schumann G., Laucht M. Association of *PER2* genotype and stressful life events with alcohol drinking in young adults. *PLoS One*. 2013;8(3):e59136. DOI 10.1371/journal.pone.0059136

Bondarenko E.A., Shadrina M.I., Druzhkova T.A., Akzhigitov R.G., Gulyaeva N.V., Gekht A.B., Slominsky P.A. An association study of rs10462021 polymorphism in the clock gene *PERIOD3* and different clinical types of depression. *Mol. Genet. Microbiol. Virol.* 2018; 33(1):26-29. DOI 10.3103/S0891416818010056

Cade B.E. Variation and selection in human circadian clock genes. Doctoral Thesis. University of Surrey, 2010

Carlson C.S., Thomas D.J., Eberle M.A., Swanson J.E., Livingston R.J., Rieder M.J., Nickerson D.A. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 2005;15(11):1553-1565. DOI 10.1101/gr.4326505

Carpen J.D., Archer S.N., Skene D.J., Smits M., von Schantz M. A single-nucleotide polymorphism in the 5′-untranslated region of the *hPER2* gene is associated with diurnal preference. *J. Sleep Res.* 2005;14(3):293-297. DOI 10.1111/j.1365-2869.2005.00471.x

Chacón-Duque J.-C., Adhikari K., Fuentes-Guajardo M., Mendoza-Revilla J., Acuña-Alonzo V., Barquera R., Quinto-Sánchez M., … Poletti G., Gallo C., Bedoya G., Rothhammer F., Balding D., Hellenthal G., Ruiz-Linares A. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* 2018;9(1):5388. DOI 10.1038/s41467-018-07748-z

Chang A.-M., Bjonnes A.C., Aeschbach D., Buxton O.M., Gooley J.J., Anderson C., Van Reen E., Cain S.W., Czeisler C.A., Duffy J.F., Lockley S.W., Shea S.A., Scheer F.A.J.L., Saxena R. Circadian gene variants influence sleep and the sleep electroencephalogram in humans. *Chronobiol. Int.* 2016;33(5):561-573. DOI 10.3109/07420528.2016.1167078

Chang Y.-C., Chiu Y.-F., Liu P.-H., Hee S.W., Chang T.-J., Jiang Y.-D., Lee W.-J., Lee P.-C., Kao H.-Y., Hwang J.-J., Chuang L.-M. Genetic variation in the *NOC* gene is associated with body mass index in chinese subjects. *PLoS One*. 2013;8(7):e69622. DOI 10.1371/journal.pone.0069622

Chappuis S., Ripperger J.A., Schnell A., Rando G., Jud C., Wahli W., Albrecht U. Role of the circadian clock gene *Per2* in adaptation to cold temperature. *Mol. Metab.* 2013;2(3):184-193. DOI 10.1016/j.molmet.2013.05.002

Clemente J.C., Pehrsson E.C., Blaser M.J., Sandhu K., Gao Z., Wang B., Magris M., Hidalgo G., Contreras M., Noya-Alarcón Ó., Lander O., McDonald J., Cox M., Walter J., Oh P.L., Ruiz J.F., Rodriguez S., Shen N., Song S.J., Metcalf J., Knight R., Dantas G., Dominguez-Bello M.G. The microbiome of uncontacted Amerindians. *Sci. Adv.* 2015;1(3):e1500183. DOI 10.1126/sciadv.1500183

Dan Y.-L., Zhao C.-N., Mao Y.-M., Wu Q., He Y.-S., Hu Y.-Q., Xiang K., Yang X.-K., Sam N.B., Wu G.-C., Pan H.-F. Association of PER2 gene single nucleotide polymorphisms with genetic susceptibility to systemic lupus erythematosus. *Lupus*. 2021;30(5): 734-740. DOI 10.1177/0961203321989794

Donin A.S., Nightingale C.M., Owen C.G., Rudnicka A.R., McNamara M.C., Prynne C.J., Stephen A.M., Cook D.G., Whincup P.H. Nutritional composition of the diets of South Asian, black African-Caribbean and white European children in the United Kingdom: The Child Heart and Health Study in England (CHASE). *Br. J. Nutr.* 2010;104(2):276-285. DOI 10.1017/S000711451000070X

Egan K.J., Knutson K.L., Pereira A.C., von Schantz M. The role of race and ethnicity in sleep, circadian rhythms and cardiovascular health. *Sleep Med. Rev.* 2017;33:70-78. DOI 10.1016/j.smrv.2016.05.004

Forbes E.E., Dahl R.E., Almeida J.R.C., Ferrell R.E., Nimgaonkar V.L., Mansour H., Sciarrillo S.R., Holm S.M., Rodriguez E.E., Phillips M.L. *PER2 rs2304672* polymorphism moderates circadian-relevant reward circuitry activity in adolescents. *Biol. Psychiatry*. 2012;71(5):451-457. DOI 10.1016/j.biopsych.2011.10.012

Gafarov V.V., Gagulin I.V., Gromova E.A., Gafarova A.V., Panov D.O. Association of polymorphism rs934945 gene Per2 with sleep disorders in the male population of Novosibirsk 25-44. *Mir Nauki, Kul'tury, Obrazovaniya = The World of Science, Culture and Education*. 2016;5:283-287 (in Russian)

Gintis H., Doebeli M., Flack J. The evolution of human cooperation. *Cliodynamics.* 2012;3(1):172-190. DOI 10.21237/C7CLIO3112928

Gu Z., Wang B., Zhang Y.-B., Ding H., Zhang Y., Yu J., Gu M., Chan P., Cai Y. Association of *ARNTL* and *PER1* genes with Parkinson's disease: a case-control study of Han Chinese. *Sci. Rep.* 2015;5(1): 15891. DOI 10.1038/srep15891

Guo H., Li X., Li W., Wu J., Wang S., Wei J. Climatic modification effects on the association between PM1 and lung cancer incidence in China. *BMC Public Health*. 2021;21(1):880. DOI 10.1186/s12889-021-10912-8

Hancock A.M., Alkorta-Aranburu G., Witonsky D.B., Di Rienzo A. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philos. Trans. R. Soc. B Biol. Sci.* 2010; 365(1552):2459-2468. DOI 10.1098/rstb.2010.0032

Holipah Hinoura T., Kozaka N., Kuroda Y. The correlation between PER3 rs2640908 polymorphism and colorectal Cancer in the Japanese population. *Appl. Cancer Res.* 2019;39(1):3. DOI 10.1186/s41241-019-0072-5

Jansen P.R., Watanabe K., Stringer S., Skene N., Bryois J., Hammerschlag A.R., de Leeuw C.A., Benjamins J.S., Muñoz-Manchado A.B., Nagel M., Savage J.E., Tiemeier H., White T., Tung J.Y., Hinds D.A., Vacic V., Wang X., Sullivan P.F., van der Sluis S., Polderman T.J.C., Smit A.B., Hjerling-Leffler J., Van Someren E.J.W., Posthuma D. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* 2019;51(3):394-403. DOI 10.1038/s41588-018-0333-3

Jensen J.D., Foll M., Bernatchez L. The past, present and future of genomic scans for selection. *Mol. Ecol.* 2016;25(1):1-4. DOI 10.1111/mec.13493

Jones S.E., Lane J.M., Wood A.R., van Hees V.T., Tyrrell J., Beaumont R.N., Jeffries A.R., … Gehrman P.R., Lawlor D.A., Frayling T.M., Rutter M.K., Hinds D.A., Saxena R., Weedon M.N. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nat. Commun.* 2019;10(1):343. DOI 10.1038/s41467-018-08259-7

Kichaev G., Bhatia G., Loh P.-R., Gazal S., Burch K., Freund M.K., Schoech A., Pasaniuc B., Price A.L. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* 2019;104(1):65-75. DOI 10.1016/j.ajhg.2018.11.008

Kim J.-W., Kim H.-A., Suh C.-H., Jung J.-Y. Sex hormones affect the pathogenesis and clinical characteristics of systemic lupus erythematosus. *Front. Med.* 2022;9:906475. DOI 10.3389/fmed.2022.906475

Klassmann A., Gautier M. Detecting selection using extended haplotype homozygosity (EHH)-based statistics in unphased or unpolarized data. *PLoS One*. 2022;17(1):e0262024. DOI 10.1371/journal.pone.0262024

Kripke D.F., Nievergelt C.M., Joo E., Shekhtman T., Kelsoe J.R. Circadian polymorphisms associated with affective disorders. *J. Circadian Rhythms*. 2009;7:2. DOI 10.1186/1740-3391-7-2

Lee H., Nah S.-S., Chang S.-H., Kim H.-K., Kwon J.-T., Lee S., Cho I.-H., Lee S.W., Kim Y.O., Hong S.-J., Kim H.-J. PER2 is downregulated by the LPS-induced inflammatory response in synoviocytes in rheumatoid arthritis and is implicated in disease susceptibility. *Mol. Med. Rep.* 2017;16(1):422-428. DOI 10.3892/mmr.2017.6578

Lesicka M., Jabłońska E., Wieczorek E., Pepłońska B., Gromadzińska J., Seroczyńska B., Kalinowski L., Skokowski J., Reszka E. Circadian gene polymorphisms associated with breast cancer susceptibility. *Int. J. Mol. Sci.* 2019;20(22):5704. DOI 10.3390/ijms20225704

LeVan T.D., Xiao P., Kumar G., Kupzyk K., Qiu F., Klinkebiel D., Eudy J., Cowan K., Berger A.M. Genetic variants in circadian rhythm genes and self-reported sleep quality in women with breast cancer. *J. Circadian Rhythms*. 2019;17(1):184. DOI 10.5334/jcr.184

Levey D.F., Stein M.B., Wendt F.R., Pathak G.A., Zhou H., Aslan M., Quaden R., Harrington K.M., Nuñez Y.Z., Overstreet C., Radhakrishnan K., Sanacora G., McIntosh A.M., Shi J., Shringarpure S.S., Concato J., Polimanti R., Gelernter J. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* 2021;24(7):954-963. DOI 10.1038/s41593-021-00860-2

Levran O., Randesi M., Rotrosen J., Ott J., Adelson M., Kreek M.J. A 3′ UTR SNP rs885863, a *cis*-eQTL for the circadian gene *VIPR2* and lincRNA 689, is associated with opioid addiction. *PLoS One*. 2019;14(11):e0224399. DOI 10.1371/journal.pone.0224399

Liberman A.R., Kwon S.B., Vu H.T., Filipowicz A., Ay A., Ingram K.K. Circadian clock model supports molecular link between *PER3* and human anxiety. *Sci. Rep.* 2017;7(1):9893. DOI 10.1038/s41598-017-07957-4

Lin E., Kuo P.-H., Liu Y.-L., Yang A.C., Kao C.-F., Tsai S.-J. Effects of circadian clock genes and health-related behavior on metabolic syndrome in a Taiwanese population: Evidence from association and interaction analysis. *PLoS One*. 2017;12(3):e0173861. DOI 10.1371/journal.pone.0173861

Lopez M.E., Neira R., Yáñez J.M. Applications in the search for genomic selection signatures in fish. *Front. Genet.* 2015;5:458. DOI 10.3389/fgene.2014.00458

Maryanaji Z. The effect of climatic and geographical factors on breast cancer in Iran. *BMC Res. Notes*. 2020;13(1):519. DOI 10.1186/s13104-020-05368-9

McCarthy M.J., Welsh D.K. Cellular circadian clocks in mood disorders. *J. Biol. Rhythms*. 2012;27(5):339-352. DOI 10.1177/0748730412456367

Melhuish Beaupre L.M., Gonçalves V.F., Zai C.C., Tiwari A.K., Harripaul R.S., Herbert D., Freeman N., Müller D.J., Kennedy J.L. Genome-wide association study of sleep disturbances in depressive disorders. *Mol. Neuropsychiatry*. 2020;5(Suppl. 1):34-43. DOI 10.1159/000505804

Min W., Tang N., Zou Z., Chen Y., Zhang X., Huang Y., Wang J., Zhang Y., Zhou B., Sun X. A panel of rhythm gene polymorphisms is involved in susceptibility to type 2 diabetes mellitus and bipolar disorder. *Ann. Transl. Med.* 2021;9(20):1555. DOI 10.21037/atm-21-4803

Miranda A., Shekhtman T., McCarthy M., DeModena A., Leckband S.G., Kelsoe J.R. Study of 45 candidate genes suggests *CACNG2* may be associated with lithium response in bipolar disorder. *J. Affect. Disord.* 2019;248:175-179. DOI 10.1016/j.jad.2019.01.010

Molina E., Gould N., Lee K., Krimins R., Hardenbergh D., Timlin H. Stress, mindfulness, and systemic lupus erythematosus: An overview and directions for future research. *Lupus*. 2022;31(13):1549-1562. DOI 10.1177/09612033221122980

National Library of Medicine (US) [WWW Document]. URL https://www.ncbi.nlm.nih.gov/(accessed 9.17.23)

Nielsen R. Molecular signatures of natural selection. *Annu. Rev. Genet.* 2005;39(1):197-218. DOI 10.1146/annurev.genet.39.073003.112420

Pan Z., Yu L., Shao M., Ma Y., Cheng Y., Wu Y., Xu S., Zhang C., Zhu J., Pan F., Sun G. The influence of meteorological factors and total malignant tumor health risk in Wuhu city in the context of climate change. *BMC Public Health*. 2023;23(1):346. DOI 10.1186/s12889-023-15200-1

Porras A.M., Shi Q., Zhou H., Callahan R., Montenegro-Bethancourt G., Solomons N., Brito I.L. Geographic differences in gut microbiota composition impact susceptibility to enteric infection. *Cell Rep.* 2021;36(4):109457. DOI 10.1016/j.celrep.2021.109457

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 analyses. *Am. J. Hum. Genet.* 2007;81(3):559-575. DOI 10.1086/519795

Qu F., Qiao Q., Wang N., Ji G., Zhao H., He L., Wang H., Bao G. Genetic polymorphisms in circadian negative feedback regulation genes predict overall survival and response to chemotherapy in gastric cancer patients. *Sci. Rep.* 2016;6(1):22424. DOI 10.1038/srep22424

Quaglia M., Merlotti G., De Andrea M., Borgogna C., Cantaluppi V. Viral infections and systemic lupus erythematosus: new players in an old story. *Viruses*. 2021;13(2):277. DOI 10.3390/v13020277

Rijo-Ferreira F., Takahashi J.S. Genomics of circadian rhythms in health and disease. *Genome Med.* 2019;11(1):82. DOI 10.1186/s13073-019-0704-0

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z.P., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., Ackerman H.C., Campbell S.J., Altshuler D., Cooper R., Kwiatkowski D., Ward R., Lander E.S. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002;419(6909):832-837. DOI 10.1038/nature01140

Sakurada K., Konta T., Takahashi S., Murakami N., Sato H., Murakami R., Watanabe M., Ishizawa K., Ueno Y., Yamashita H., Kayama T. Circadian clock gene polymorphisms and sleep-onset problems in a population-based cohort study: The yamagata study. *Tohoku J. Exp. Med.* 2021;255(4):325-331. DOI 10.1620/tjem.255.325

Saravanan K.A., Panigrahi M., Kumar H., Bhushan B., Dutt T., Mishra B.P. Selection signatures in livestock genome: A review of concepts, approaches and applications. *Livest. Sci.* 2020;241:104257. DOI 10.1016/j.livsci.2020.104257

Schroor M.M., Plat J., Mensink R.P. Relation between single nucleotide polymorphisms in circadian clock relevant genes and cholesterol metabolism. *Mol. Genet. Metab.* 2023;138(4):107561. DOI 10.1016/j.ymgme.2023.107561

Shareefa D. Genetic analysis of bipolar disorder and alcohol use disorder. Doctoral Thesis. University of Cape Town, 2015

Shi Y., Liu Y., Yang L., Yan J. A mathematical model to characterize the role of light adaptation in mammalian circadian clock. *Front. Mol. Biosci.* 2021;8:681696. DOI 10.3389/fmolb.2021.681696

Soria V., Martínez-Amorós È., Escaramís G., Valero J., Pérez-Egea R., García C., Gutiérrez-Zotes A., Puigdemont D., Bayés M., Crespo J.M., Martorell L., Vilella E., Labad A., Vallejo J., Pérez V., Menchón J.M., Estivill X., Gratacòs M., Urretavizcaya M. Differential association of circadian genes with mood disorders: CRY1 and NPAS2 are associated with unipolar major depression and CLOCK and VIP with bipolar disorder. *Neuropsychopharmacology*. 2010;35(6):1279-1289. DOI 10.1038/npp.2009.230

Syromyatnikov M., Nesterova E., Gladkikh M., Smirnova Y., Gryaznova M., Popov V. Characteristics of the gut bacterial composition in people of different nationalities and religions. *Microorganisms*. 2022;10(9):1866. DOI 10.3390/microorganisms10091866

Szpiech Z.A. Selscan 2.0: scanning for sweeps in unphased data. *bioRxiv*. 2021. DOI 10.1101/2021.10.22.465497

Voight B.F., Kudaravalli S., Wen X., Pritchard J.K. A map of recent positive selection in the human genome. *PLoS Biol.* 2006;4(3):e72. DOI 10.1371/journal.pbio.0040072

Wang W.M., Yuan P., Wang J.Y., Ma F., Fan Y., Li Q., Zhang P., Xu B.H. Association of genetic variantions of circadian clock genes and risk

of breast cancer. *Zhonghua Zhong Liu Za Zhi*. 2013;35(3):236-239. DOI 10.3760/cma.j.issn.0253-3766.2013.03.017

Wen M., Jiang X., She H., Han C., Pei Z., Cai Y., Zhang T. The *Per2* polymorphism rs10462023 is associated with the risk of stroke in a Chinese population. *Biol. Rhythm Res.* 2015;46(4):545-551. DOI 10.1080/09291016.2015.1026675

Xu L., Bickhart D.M., Cole J.B., Schroeder S.G., Song J., Tassell C.P., Sonstegard T.S., Liu G.E. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.* 2015;32(3):711-725. DOI 10.1093/molbev/msu333

Zhang L., Ptáček L.J., Fu Y.-H. Diversity of human clock genotypes and consequences. 2013;119:51-81. DOI 10.1016/B978-0-12-396971-2.00003-8

Zheng W., He Y., Guo Y., Yue T., Zhang H., Li J., Zhou B., Zeng X., Li L., Wang B., Cao J., Chen L., Li C., Li H., Cui C., Bai C., Baimakangzhuo, Qi X., Ouzhuluobu, Su B. Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol.* 2023;24(1):73. DOI 10.1186/s13059-023-02912-1

Conflict of interest. The authors declare no conflict of interest.
Received January 31, 2024. Revised June 25, 2024. Accepted June 26, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):650-658

DOI 10.18699/vjgb-24-72

Genetic aspects of lactase deficiency in indigenous populations of Siberia

B.A. Malyarchuk

Institute of Biological Problems of the North of the Far Eastern Branch of the Russian Academy of Sciences, Magadan, Russia
malbor@mail.ru

Abstract. The ability to metabolize lactose in adulthood is associated with the persistence of lactase enzyme activity. In European populations, lactase persistence is determined mainly by the presence of the rs4988235-T variant in the *MCM6* gene, which increases the expression of the *LCT* gene, encoding lactase. The highest rates of lactase persistence are characteristic of Europeans, and the lowest rates are found in East Asian populations. Analysis of published data on the distribution of the hypolactasia-associated variant rs4988235-C in the populations of Central Asia and Siberia showed that the frequency of this variant increases in the northeastern direction. The frequency of this allele is 87 % in Central Asia, 90.6 % in Southern Siberia, and 92.9 % in Northeastern Siberia. Consequently, the ability of the population to metabolize lactose decreases in the same geographical direction. The analysis of paleogenomic data has shown that the higher frequency of the rs4988235-T allele in populations of Central Asia and Southern Siberia is associated with the eastward spread of ancient populations of the Eastern European steppes, starting from the Bronze Age. The results of polymorphism analysis of exons and adjacent introns of the *MCM6* and *LCT* genes in indigenous populations of Siberia indicate the possibility that polymorphic variants may potentially be related to lactose metabolism exist in East Asian populations. In East Asian populations, including Siberian ethnic groups, a ~26.5 thousand nucleotide pairs long region of the *MCM6* gene, including a combination of the rs4988285-A, rs2070069-G, rs3087353-T, and rs2070068-A alleles, was found. The rs4988285 and rs2070069 loci are located in the enhancer region that regulates the activity of the *LCT* gene. Analysis of paleogenomic sequences showed that the genomes of Denisovans and Neanderthals are characterized by the above combination of alleles of the *MCM6* gene. Thus, the haplotype discovered appears to be archaic. It could have been inherited from a common ancestor of modern humans, Neanderthals, and Denisovans, or it could have been acquired by hybridization with Denisovans or Neanderthals. The data obtained indicate a possible functional significance of archaic variants of the *MCM6* gene.

Key words: genetic polymorphism; lactase persistence; *MCM6* gene; LCT gene; human populations; Siberia; archaic variants of polymorphism.

For citation: Malyarchuk B.A. Genetic aspects of lactase deficiency in indigenous populations of Siberia. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):650-658. DOI 10.18699/vjgb-24-72

Генетические аспекты лактазной недостаточности у коренного населения Сибири

Б.А. Малярчук

Институт биологических проблем Севера Дальневосточного отделения Российской академии наук, Магадан, Россия
malbor@mail.ru

Аннотация. Способность метаболизировать лактозу во взрослом состоянии связана с сохранением активности фермента лактазы. В европейских популяциях персистенция лактазы детерминируется главным образом наличием варианта rs4988235-T в гене *MCM6*, который увеличивает экспрессию гена *LCT*, кодирующего лактазу. Наиболее высокие показатели персистенции лактазы характерны для европейцев, а самые низкие – для населения Восточной Азии. Анализ опубликованных данных о распределении варианта rs4988235-C, связанного с гиполактазией, у населения Центральной Азии и Сибири выявил, что частота этого варианта увеличивается в северо-восточном направлении. В Центральной Азии частота этого аллеля составляет 87 %, на юге Сибири – 90.6 % и на северо-востоке Сибири – 92.9 %. Соответственно, в таком же географическом направлении убывает способность населения метаболизировать лактозу. Анализ палеогеномных данных показал, что более высокая частота аллеля rs4988235-T в популяциях Центральной Азии и Южной Сибири связана с распространением на восток древнего населения восточноевропейских степей начиная с эпохи бронзового века. Результаты анализа полиморфизма экзонов и прилегающих к ним интронов генов *MCM6* и *LCT* у коренного населения Сибири свидетельствуют о возможности существования в восточноазиатских популяциях вариантов полиморфизма, потенциально связанных с метаболизмом лактозы. В популяциях Восточной Азии, в том числе в сибирских

этнических группах, обнаружен участок гена *MCM6* длиной ~26.5 тыс. пар нуклеотидов, включающий комбинацию аллелей rs4988285-A, rs2070069-G, rs3087353-T, rs2070068-A. Локусы rs4988285 и rs2070069 находятся в области энхансера, регулирующего активность гена *LCT*. Анализ палеогеномных последовательностей показал, что указанной выше комбинацией аллелей гена *MCM6* характеризуются геномы денисовцев и неандертальцев. Таким образом, обнаруженный гаплотип, по всей видимости, является архаичным. Он мог быть унаследован от общего предка современных людей, неандертальцев и денисовцев, или же был приобретен в результате гибридизации с денисовцами или неандертальцами. Полученные данные свидетельствуют о возможной функциональной значимости архаичных вариантов полиморфизма гена *MCM6*.
Ключевые слова: генетический полиморфизм; персистенция лактазы; ген *MCM6*; ген *LCT*; популяции человека; Сибирь; архаичные варианты полиморфизма.

Introduction

Lactose (milk sugar) is the main disaccharide in the milk of various mammals and its hydrolysis requires the enzyme lactase, encoded by the *LCT* gene, which is predominantly expressed in the small intestine. Lactase activity declines during ontogenesis, which can lead to difficulties digesting lactose in many adults (Ségurel, Bon, 2017). Primary hypolactasia (OMIM: 223100) is characterized by a range of symptoms (bloating, nausea, diarrhoea) after ingestion of milk and dairy products. However, ethnoregional populations around the world have been found to differ in their ability to metabolize lactose (Evershed et al., 2022). It is thought that this ability, or lactase persistence (LP), is inherited. One of the most important genetic polymorphisms that have been linked to LP is the T variant at the rs4988235 locus of the *MCM6* gene, which regulates the expression of the *LCT* gene (Enattah et al., 2002; Olds, Sibley, 2003; Troelsen et al., 2003). Although this genetic variant is about 14,000 nucleotide pairs away from the *LCT* gene (which is why it is often called –13910*T), it is responsible for increasing the enzymatic activity of lactase, which breaks down lactose into glucose and galactose molecules.

It would appear that the lowest LP values are characteristic of East Asian populations, while the highest are found in Europeans (Liebert et al., 2017). This is due to the fact that, according to archaeological data, dairy farming may have emerged in the steppe zone of the North Caucasus and the Black Sea region around 4–5 thousand years ago (kya) (Scott et al., 2022). Paleogenomic data suggests that the frequency of the LP-associated variant rs4988235-T began to increase around 6 kya within the ancestral EHG and CHG genomic components characteristic of Eastern European and Caucasian hunter-gatherers, respectively (Segurel et al., 2020; Irving-Pease et al., 2024). The linkage between the polymorphism variants in the rs4988235 and rs1438307 loci was also revealed, and the increase in frequency of the rs1438307-T allele may have begun much earlier than previously thought, around 12 kya (Irving-Pease et al., 2024). With regard to the rs1438307-T variant, it has been suggested that it may have arisen as a consequence of the adaptation of ancient humans to starvation and exposure to pathogens; this is based on the observation that it is involved in the regulation of the body's energy expenditure and the development of metabolic diseases (Evershed et al., 2022).

Despite the great interest of the genetic and medical communities in the genetic aspects of hypolactasia in human populations, many regions of the world remain poorly studied (Liebert et al., 2017; Anguita-Ruiz et al., 2020). The aim of this paper is to attempt to review the results of studies on the polymorphism of the *LCT* and *MCM6* genes, which are directly related to lactase persistence, in indigenous populations of Siberia, one of the least studied regions.

Distribution of rs4988235 locus polymorphisms in modern and ancient North Asian populations

Genetic and epidemiological studies have indicated that in populations of the European part of Russia, primary hypolactasia is determined predominantly or exclusively by the rs4988235-C allele of the *MCM6* gene (Borinskaya et al., 2006; Kovalenko et al., 2023), and accordingly, LP is defined by the rs4988235-T allele. However, in East Asian populations (including Siberian ones), this relationship is not as clearly evident – some populations (e. g., Buryats and Uyghurs) show very high (at the level of 95 %) frequency of the rs4988235-C variant, which is associated with a reduced prevalence of hypolactasia (Borinskaya et al., 2006; Sokolova et al., 2007). In this context, it has been proposed that the lower frequency of hypolactasia in some ethnic groups of Siberia and Central Asia may be associated with the presence of not only the rs4988235-T variant, but also some other genetic LP markers (Sokolova et al., 2007).

To date, it has been found that in addition to the rs4988235- T allele, several other genetic polymorphism variants that determine the ability to break down lactose are common in African and Middle Eastern ethnic groups, for example they include the rs41525747, rs41380347, rs145946881 and rs182549 loci of the *MCM6* gene (Ingram et al., 2007; Tishkoff et al., 2007). However, data on the association between genetic polymorphisms and LP in East Asian populations is somewhat more conflicting. For instance, in Central Asian populations, a mixed sample of Tajiks and Uzbeks, as well as Kazakhs, showed that the rs4988235-T variant (with frequencies of 10 and 16.5 %, respectively) correlated quite well with the ability to digest lactose (11–30 % in Tajiks and Uzbeks, and 25–32 % in Kazakhs) (Heyer et al., 2011). It would appear that Tibetans, who have a long-standing tradition of consuming yak milk, also digest lactose at a level of around 30 %, but they appear to lack the rs4988235-T and rs182549-T polymorphism variants that are found in neighbouring populations in northern China at a frequency of 3.8 and 6.9 %, respectively (Xu et al., 2010; Peng et al., 2012). Tibetans have been found to have their own spectrum of alleles of the enhancer region of the *MCM6* gene, which may be associated with LP, among which the –13838*A variant appears to predominate with a frequency of about 6.5 % (Peng et al., 2012).

A more detailed study of the genetic adaptation to milk consumption in Central Asian populations, distinguished by their economic patterns, has demonstrated that pastoralists

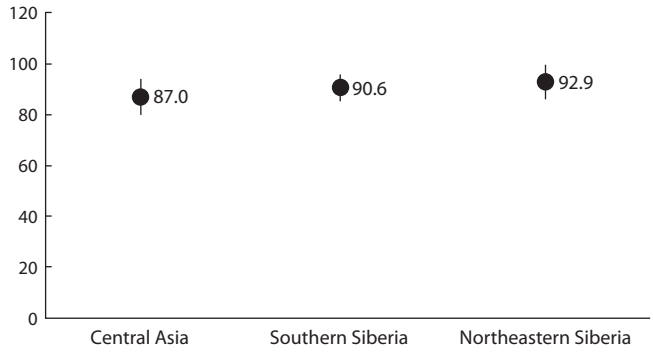
(Kazakhs, Kyrgyz, Karakalpaks, Buryats, Mongolians and Altaians), whose diets rely heavily on dairy products, do not have a higher ability to metabolize lactose than farmers (Turkmens, Tajiks and Uzbeks), who have a higher prevalence of the rs4988235-T variant (Sequrel et al., 2020). The relatively low frequency (~10 %) of this genetic variant is also observed in ethnic groups in southern Siberia (Khakasians, Shorians and Tubalars) who lead a semi-nomadic lifestyle and are engaged in forestry and taiga hunting (Sequrel et al., 2020). The data indicate that the frequency of occurrence of the rs4988235-T variant in Central Asian and Siberian populations is not significantly influenced by economic structure or milk consumption levels.

Table 1 presents the distribution of the rs4988235-C variant in various indigenous populations from Northeast China, Central Asia and Siberia. As can be observed, the frequency of this variant in the populations varies from 70 to 100 %. However, when the samples are divided into three regional groups, there is an increase in the frequency of the rs4988235-C variant from the south to the northeast of Siberia (see the Figure).

The frequency of this allele is 87.0 ± 2.0 % in Central Asia, 90.6 ± 1.7 % in Southern Siberia and 92.9 ± 2.3 % in Northeastern Siberia. Consequently, the population's capacity to metabolize lactose diminishes in a similar geographic direction. It is also noteworthy that the observed differences in the allele frequency of the rs4988235 locus are statistically significant only between the populations of Central Asia and Siberia (both its southern and northeastern parts; $P < 10^{-5}$, Fisher's exact test). Furthermore, the two Siberian populations do not differ from each other ($P = 0.09$). It is important to note that allele frequencies may vary in different samples within the same ethnic group. This is exemplified by the case of the Kazakhs, Buryats, and especially Chukchi (Table 1). In addition to random factors, which are particularly relevant in small sample sizes, admixture with individuals belonging to ethnic groups exhibiting a higher frequency of the rs4988235-T variant may contribute to the observed heterogeneity in allele frequencies.

This was clearly demonstrated in the study of the rs4988235 locus polymorphism in the Nenets, who have been reindeer herders for generations and who practically do not drink milk (Khabarova et al., 2012). It appears that this is primarily due to a high prevalence of lactose intolerance. The frequency of the rs4988235-C variant in the Nenets with all four grandparents of Nenets origin is 92.7 % (the frequency of the rs4988235-CC genotype is 90 %). Concurrently, the prevalence of the rs4988235-C allele among the Nenets with at least one relative of Nenets origin has declined to 73 % (Khabarova et al., 2012). The majority of interethnic marriages with Nenets involve Komi and North Russians, in whom the frequency of the rs4988235-T allele is 35–42 % (Khabarova et al., 2011). A certain decrease in hypolactasia among the populations of the northernmost regions of Europe and Siberia can be attributed to intermarriage contacts with immigrant populations of Eastern European origin, which commenced in the 17th century in conjunction with the expansion of the Russian pioneers and reached its peak during the Soviet period.

The duration and intensity of contacts with Eastern Europeans were evidently greater on the territory of Southwestern Siberia and Central Asia, taking into account the migrations of the populations of the Eastern European steppes during the



Distribution of the rs4988235-C variant in regional groups of Siberia and Central Asia.

The average frequency of the genetic variant (in %) and the limits of the standard deviation of the frequency are shown.

Bronze Age. As previously stated, paleogenomic data indicates that the rs4988235-T mutation emerged approximately 6 kya in the ancient population of the northern Black Sea coast, and later this variant of the polymorphism spread throughout northern Eurasia, from Spain to Kazakhstan (Segurel et al., 2020; Irving-Pease et al., 2024). Furthermore, in Europe, the increased frequency of the rs4988235-T variant, which determines stable lactase activity required for milk digestion in adults, was favored by positive selection due to vitamin D deficiency at high latitudes and the need for increased calcium intake from milk (Kozlov, Vershubskaya, 2017).

A review of paleogenomic data in the AADR database (Allen Ancient DNA Resource, https://reich.hms.harvard.edu/) revealed that the first documented instances of the rs4988235-T allele were observed in ancient European populations, including those in Ukraine (~6 kya), Ireland (5.5 kya), and from 4.5 kya and later in Lithuania, Germany, Czech Republic, Estonia, and Russia. In East Asia, the rs4988235-T allele was first identified in an individual from the Botai archaeological culture in northern Kazakhstan (5.3 kya). In Central Asia (in the ancestors of the Kazakhs, Kyrgyz, Mongolians, Turkmens, Uzbeks, and Tajiks), the frequency of the rs4988235-T variant was 4.2 % between 0.5 and 5.3 kya. According to modern data (Table 1), its value is estimated to be approximately 13 %. It is hypothesized that this mutation was already prevalent in Central Asia with a frequency of approximately 5 % since the Iron Age (Segurel et al., 2020). Consequently, if this variant of genetic polymorphism was subject to strong selective pressure, it should have had sufficient time to reach high frequencies in modern populations. This is estimated by L. Segurel et al. (2020) to be 51 %. However, this was not the case, which leads to the reasonable conclusion that the rs4988235-T variant did not experience significant selective pressure in Central Asian populations, in contrast to Europeans and some populations in Africa and the Middle East (Segurel et al., 2020).

The AADR database indicates that in the ancient populations of Siberia and the Urals (between 0.5 to 10 kya), the rs4988235-T variant was distributed with a frequency of 1.8 %, although only in the westernmost region of this territory. All cases of this allele were registered around 3.1–3.8 kya in representatives of the Karasuk (Southwestern Siberia) and Sintashta (Southern Urals) archaeological cultures. The mean

Table 1. Frequency of the rs4988235-C allele in North Asian populations

Ethnic group	Sample size	Frequency of the rs4988235-C allele	References
Tajiks (Uzbekistan, Tajikistan)	254	0.82	Segurel et al., 2020
Uzbeks (Uzbekistan)	45	0.76	Segurel et al., 2020
Turkmens (Uzbekistan)	50	0.80	Segurel et al., 2020
Kazakhs (China)	94	0.95	Sun et al., 2007
Kazakhs (Uzbekistan)	83	0.83	Heyer et al., 2011
Kazakhs (Uzbekistan)	159	0.79	Segurel et al., 2020
Kazakhs (Kazakhstan)	34	0.88	Sokolova et al., 2007
Altaian Kazakhs	128	0.91	Pilipenko et al., 2016
Kyrgyz (Kyrgyzstan)	201	0.88	Segurel et al., 2020
Karakalpaks (Uzbekistan)	45	0.93	Segurel et al., 2020
Uyghurs (Kazakhstan)	30	0.95	Sokolova et al., 2007
Mongolians (China)	82	0.98	Sun et al., 2007
Mongolians (Mongolia)	32	0.88	Segurel et al., 2020
Manchurians (China)	75	1.0	Sun et al., 2007
Orochen (China)	45	0.99	Sun et al., 2007
Nanai (China)	77	1.0	Sun et al., 2007
Southern Altaians	24	0.85	Cardona et al., 2014
Southern Altaians	62	0.92	Segurel et al., 2020
Northern Altaians	29	0.93	Segurel et al., 2020
Shorians	24	0.94	Cardona et al., 2014
Shorians	29	0.90	Segurel et al., 2020
Khakasians	29	0.86	Segurel et al., 2020
Khakasians	64	0.92	Pilipenko et al., 2016
Buryats	78	0.95	Sokolova et al., 2007
Buryats	24	0.98	Cardona et al., 2014
Buryats	28	0.82	Segurel et al., 2020
Yakuts	22	0.93	Cardona et al., 2014
Yakuts	55	0.95	Liebert et al., 2017
Yakuts	25	0.94	Bersaglieri et al., 2004
Western Evenki	24	0.96	Cardona et al., 2014
Evens	24	0.96	Cardona et al., 2014
Koryaks	25	0.96	Cardona et al., 2014
Chukchi	35	0.94	Borinskaya et al., 2006
Chukchi	14	0.75	Cardona et al., 2014
Eskimo	19	0.97	Cardona et al., 2014

frequency of the rs4988235-T variant in the contemporary indigenous population of Siberia is 0.8 % (Table 1). This suggests that, over the past 3,000 years, the frequency of the allele responsible for the enhancement of lactase enzymatic activity has remained unchanged in Siberian populations, despite changes in dietary habits and increased consumption of dairy products. This evidence indicates that in Siberian populations, the rs4988235-T allele behaves as a neutral variant of genetic polymorphism.

In light of the possibility of additional variants of the enhancer polymorphism of the *MCM6* gene in the East Asian population, it is worth noting that this kind of screening was

Table 2. Frequencies of the rs4988235-T and rs182549-A alleles in Siberian populations (according to Cardona et al., 2014)

Ethnic group	Sample size, *N*	Frequency of the rs4988235-T allele	Frequency of the rs182549-A allele
Southern Altaians	24	0.15	0.15
Shorians	24	0.06	0.06
Buryats	24	0.02	0.02
Yakuts	22	0.07	0.07
Western Evenki	24	0.04	0.04
Evens	24	0.04	0.04
Koryaks	25	0.04	0.04
Chukchi	14	0.25	0.25
Eskimo	19	0.03	0.05

performed for several loci, including rs41525747, rs41380347, rs869051967, rs145946881, and rs182549 (Xu et al., 2010; Liebert et al., 2017; Anguita-Ruiz et al., 2020). The enhancer element of the *LCT* gene was also investigated in two indigenous populations of Southern Siberia – the Altaian Kazakhs and Khakasians (Pilipenko et al., 2016). Nevertheless, the frequencies of alleles potentially associated with LP were generally quite low. The sole exception to this is the rs182549 locus. In some East Asian populations, it has been reported that the rs182549-A allele is more informative than rs4988235-T. This is due to the observed occurrence of the rs182549-A allele in the absence of the rs4988235-T allele (Sun et al., 2007; Mattar et al., 2010; Xu et al., 2010). A similar conclusion has been reached for some populations of African, European, and West Asian origin (Bersaglieri et al., 2004; Coelho et al., 2005; Raz et al., 2013). However, this is at odds with the previous conclusion that a complete linkage disequilibrium exists between the rs4988235-T and rs182549-A alleles (Enattah et al., 2002; Troelsen et al., 2003), which included East Asian populations (Kato et al., 2018).

The data on the frequency of distribution of rs4988235-T and rs182549-A variants in ethnic groups of Siberia (Cardona et al., 2014) also indicate that these alleles are linked. The only exception is the Eskimo group, where the frequency of the rs182549-A allele exceeds that of the rs4988235-T allele (Table 2). Therefore, it is highly unlikely that the rs182549-A allele is responsible for maintaining lactase persistence in the indigenous Siberian population.

Polymorphic variants (including archaic ones) of the *LCT* and *MCM6* genes in indigenous Siberians

The results of the polymorphism analysis of exons and adjacent introns of the *MCM6* and *LCT* genes in the indigenous populations of Siberia clearly indicate the existence of polymorphic variants potentially related to lactose metabolism in East Asian populations. Table 3 presents data on the distribution of *LCT* and *MCM6* gene polymorphisms in 102 individuals from various regions of Siberia. These include the in digenous populations of Northeastern (Eskimo, Chukchi, Koryaks), Central (Evens, Evenki, Yakuts), Southern (Tuvinians, Shorians, Altaians, Buryats), and Western (Kets, Khanty, Mansi, Selkups, and Nenets) Siberia. The data were obtained from a full-genome variability study (Pagani et al., 2016). The *LCT* gene contains 21 polymorphic loci, while the *MCM6* gene has seven. The majority of polymorphic variants identified in indigenous Siberians belong to alleles that are commonly found in both East Asian and European populations. Rare variants characteristic only of the East Asian population were found in the rs201668742, rs144864087, and rs3739021 loci. Similarly, variants characteristic only of Europeans were revealed in the rs34307240 locus.

However, it is noteworthy that a group of polymorphic variants in the rs79023654 locus of the *LCT* gene and the rs4988285, rs2070069, rs3087353, and rs2070068 loci of the *MCM6* gene (highlighted in bold in Table 3) warrant further investigation. The rs79023654, rs4988285, and rs2070069 loci are located in the noncoding region of the genes. The rs3087353 and rs2070068 loci are situated within exons, yet nucleotide substitutions within them fail to result in amino acid substitutions. The alleles listed in Table 3 are linked in both indigenous Siberians and other East Asian populations, including Japanese, Koreans, and Vietnamese (Tables 3 and 4). In individuals from Siberia, all these alleles are known to be present in the rs4988235-CC genotype, which is associated with primary hypolactasia. The rs79023654 locus of the *LCT* gene is located at a distance of ~29.7 thousand nucleotide pairs from the rs4988285 locus of the *MCM6* gene. The polymorphic loci within the *MCM6* gene are located at a distance of ~26.5 thousand nucleotide pairs from each other. Furthermore, the rs4988285 and rs2070069 loci are located in the vicinity of the enhancer that regulates the activity of the *LCT* gene. This suggests the potential functional significance of the identified haplotype's polymorphic variants.

The analysis of the dbSNP data (https://www.ncbi.nlm.nih.gov/snp/) indicated that the rs79023654-A, rs4988285-A, rs2070069-G, and rs3087353-T variants were characteristic of the East Asian population and were observed with a low frequency (approximately 1 %) in the South Asian populations (Table 4). However, the fifth allele from this group, rs2070068-A, was detected with a relatively high frequency (24.7 %) in African populations (Table 4). From this distribution, it can be concluded that the East Asian haplotype rs79023654-A, rs4988285-A, rs2070069-G, and rs3087353-T

Table 3. Polymorphic variants of exons and adjacent introns of the *LCT* and *MCM6* genes and their frequency (in %) in Eurasian populations

Polymorphic variant	Gene	NES (*N* = 25)	CS (*N* = 29)	SS (*N* = 28)	WS (*N* = 20)	EAS	EUR
rs62170085-G	*LCT*	6.0	0	0	2.5	0.26	2.84
rs1042712-C	*LCT*	0	6.9	14.3	10.0	21.7	19.4
rs2278544-G	*LCT*	60.0	51.7	64.3	67.5	43.3	68.6
rs3213890-A	*LCT*	0	6.9	14.3	10.0	20.1	19.5
rs2322659-C	*LCT*	62.0	58.6	64.3	70.0	45.6	66.5
rs2304371-G	*LCT*	0	8.6	21.4	12.5	22.6	23.3
rs3739022-A	*LCT*	36.0	32.8	12.5	20.0	21.5	13.9
rs201668742-T	*LCT*	0	5.2	0	0	0.03	0
rs144864087-C	*LCT*	4.0	6.9	7.1	5.0	1.03	0
rs79023654-A	***LCT***	**4.0**	**10.3**	**10.7**	**7.5**	**16.2**	**0**
rs35093754-C	*LCT*	0	1.7	7.1	2.5	5.04	2.92
rs6719488-T	*LCT*	60.0	48.3	57.1	62.5	39.7	62.6
rs2322812-G	*LCT*	36.0	32.8	12.5	17.5	21.5	13.9
rs2874874-C	*LCT*	36.0	32.8	12.5	17.5	21.5	13.9
rs7579771-A	*LCT*	40.0	51.7	44.6	37.5	60.4	37.4
rs2164210-C	*LCT*	60.0	48.3	55.4	62.5	39.7	62.6
rs60376570-A	*LCT*	36.0	32.8	12.5	17.5	21.5	13.9
rs3816088-C	*LCT*	0	1.7	7.1	2.5	5.04	3.0
rs3754689-T	*LCT*	4.0	19.0	26.8	20.0	37.7	20.2
rs2236783-A	*LCT*	54.0	48.3	51.8	57.5	37.01	62.7
rs34307240-A	*LCT*	2.0	0	1.8	0	0	0.95
rs4988285-A	***MCM6***	**4.0**	**10.3**	**10.7**	**7.5**	**16.2**	**0**
rs3739021-A	*MCM6*	0	1.7	3.6	2.5	0.17	0
rs3087350-T	*MCM6*	0	1.7	7.1	2.5	5.2	3.0
rs2070069-G	***MCM6***	**4.0**	**10.3**	**10.7**	**7.5**	**16.2**	**0**
rs3087353-T	***MCM6***	**4.0**	**10.3**	**12.5**	**7.5**	**15.7**	**0**
rs2070068-A	***MCM6***	**4.0**	**10.3**	**12.5**	**7.5**	**15.8**	**0**
rs1057031-A	*MCM6*	0	8.6	14.3	7.5	21.3	20.5

Note. Designations: NES – Northeastern Siberia; CS – Central Siberia; SS – Southern Siberia; WS – Western Siberia; EAS – East Asia; EUR – Europe. For Siberian populations, frequencies are given according to Pagani et al. (2016), for East Asia and Europe, according to the dbSNP database.

Table 4. Frequency (in %) of variants rs79023654-A, rs4988285-A, rs2070069-G, rs3087353-T and rs2070068-A in world populations

Region/country	rs79023654-A	rs4988285-A	rs2070069-G	rs3087353-T	rs2070068-A
Europe	0	0	0	0	0
Siberia	8.3	8.3	8.3	8.8	8.8
East Asia	16.2	16.2	16.2	15.7	15.8
Japan	13.9	13.9	13.9	13.9	13.9
Vietnam	17.1	17.1	11.2	12.8	13.1
South Korea	18.4	18.5	18.6	18.5	18.5
South Asia	1.1	0.58	0.58	0.58	0.58
Africa	0	0	0	0	24.7

Note. Population frequencies are given according to the dbSNP database; for Siberian populations, according to Pagani et al. (2016).

was formed on the basis of ancestral (African) haplotypes, which were characterized by the rs2070068-A variant. However, an analysis of paleogenomic data (AADR database) revealed that the rs2070068-A variant emerged in Africa at a later point in time than in Eurasia. The earliest documented occurrence of this allele in Africa is associated with the northern region of the continent (in the territory of Morocco) at approximately 14.5 kya. Subsequent cases were identified at approximately 9 kya and later. However, it became evident that in Eurasia, this variant of *MCM6* gene polymorphism was observed in both Denisovans and Neanderthals (individuals who lived between ~40 and 110 kya), as well as in numerous most ancient representatives of *Homo sapiens* in Europe and East Asia (aged between ~34 and 44 kya).

Further analysis of paleogenomic sequence databases (Denisova Variants Track Settings; https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=dhcVcfDenisovaPinky) revealed that the *MCM6* haplotype rs4988285-A, rs2070069-G, rs3087353-T, rs2070068-A was common among Denisovans and Neanderthals. The rs79023654 locus of the *LCT* gene fell into a sequencing region with low coverage. Hence the presence of polymorphism at this locus in Denisovans and Neanderthals remains uncertain.

The results obtained thus suggest that the *MCM6* haplotype detected in the population of East Asia (and, to a much lesser extent, South Asia) is archaic. It is possible that this haplotype was inherited from the common ancestor of *H. sapiens*, Neanderthals, and Denisovans (approximately 600 kya, at the time of the divergence of the ancestor of *H. sapiens* from the ancestors of Neanderthals and Denisovans, as reported by H. Zeberg et al. (2024)). Alternatively, it may have been acquired as a result of hybridization with Neanderthals or Denisovans. Given the distribution of the archaic haplotype in East Asia, it seems more probable that introgression from Denisovans occurred. It has been demonstrated that Neanderthals and Denisovans also exchanged genes – for example, approximately 80–90 kya in southern Siberia (Slon et al., 2018). Consequently, the hypothesis that polymorphic variants were transferred from Denisovans to Neanderthals is also a plausible one.

In recent years, there has been a considerable amount of work done to catalogue archaic variants of genetic polymorphisms that have been identified in the gene pool of modern humans (https://bioinf.eva.mpg.de/catalogbrowser), but the incompleteness of this type of information may be dependent on the extent to which populations have been studied (Zeberg et al., 2024). It seems likely that this database will become much more comprehensive as genomic research continues to expand geographically. There are already some interesting findings of rare ancestral polymorphism variants in widely separated populations – for example, identical alleles of a number of genes in the South African Khoisan and the Philippine Aeta (Zeberg et al., 2024).

There is a great deal of information about the genetic variants that modern humans have inherited from Neanderthals. In particular, there is much to be gained from an understanding of the advantages that humans have gained from admixture, in terms of metabolism, sensory function (especially pain perception), immunity (including SARS-CoV-2), and the expression of some genes (Telis et al., 2020; Zeberg et al., 2020; Pairo-Castineira et al., 2021; Haeggström et al., 2022; Zeberg et al., 2024).

Much less is known about the functional manifestations of Denisovan genetic influence. The main examples of such influence are related to adaptation to high altitude and cold conditions. For instance, a ~33,000 bp fragment of Denisovan DNA has been found in Tibetans that encodes the hypoxia-inducible transcription factor EPAS1, which is involved in adaptation to low oxygen levels (Zhang et al., 2021). In Greenland Eskimos, a ~28,000 bp fragment of Denisovian DNA containing the *WARS* and *TBX15* genes has been identified with high frequency – it is believed that these polymorphic variants may play a role in the adaptation of Arctic indigenous peoples to low temperatures (Racimo et al., 2017). It seems plausible to suggest that the archaic haplotype of the *MCM6* gene found in East Asian populations may be used to implement a specific programme for regulating the enzymatic activity of lactase, which is still relevant today. Further studies are needed to gain a deeper understanding of the specific role of this haplotype in regulating lactase activity. These studies should consider a range of factors, including medical genetics, biochemical and physiological aspects.

Conclusion

Thus, the results of the review of the data on the variability of the *LCT* and *MCM6* genes indicate that from ancient times the indigenous populations of Siberia have been characterized by a low frequency of the rs4988235-T variant, which may contribute to the enhancement of the enzymatic activity of lactase. A certain increase in the frequency of this allele over time in the populations of Central Asia and Southwestern Siberia is associated with the eastward expansion of the ancient populations of the Eastern European steppes starting from the Bronze Age (Heyer et al., 2011; Pilipenko et al., 2016; Segurel et al., 2020). However, it seems that the rs4988235-T variant did not reach high frequencies in Central Asian populations, in contrast to Europe. This may suggest that there is no significant selective pressure on this variant of polymorphism in Central Asian populations (Segurel et al., 2020). It is still unclear why different groups of East Asian populations that traditionally consume dairy products have not developed specific variants of genetic polymorphisms for lactose metabolism. One possible explanation is the hypothesis of cultural adaptation of Central Asian populations, including the development of a culture using bacteria to digest lactose during fermentation, which may have contributed to the establishment of specific microflora in the gut (Segurel et al., 2020).

It is also worth noting that some epigenetic mechanisms (mainly DNA methylation) may also be involved in regulating the expression of lactose metabolism genes (Labrie et al., 2016). It has also been suggested that the type of DNA methylation in the enhancer and promoter regions of the *LCT* gene may be a useful indicator of lactase phenotypes, and it appears that epigenetic modifications may play an important role in the regulation of lactase deficiency (Leseva et al., 2018). Thus, both genetic and epigenetic approaches should be used to investigate the functional significance of polymorphic variants potentially associated with LP, including archaic genetic variants, which the present study has shown to still have some prevalence in human populations.

References

Anguita-Ruiz A., Aguilera C.M., Gil Á. Genetics of lactose intolerance: an updated review and online interactive world maps of phenotype and genotype frequencies. *Nutrients.* 2020;12(9):2689. DOI 10.3390/nu12092689

Bersaglieri T., Sabeti P.C., Patterson N., Vanderploeg T., Schaffner S.F., Drake J.A., Rhodes M., Reich D.E., Hirschhorn J.N. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 2004;74:1111-1120. DOI 10.1086/421051

Borinskaya S.A., Rebrikov D.V., Nefedova V.V., Kofiadi I.A., Sokolova M.V., Kolchina E.V., Kulikova E.A., Chernyshov V.N., Kutsev S.I., Polonikov A.V., Ivanov V.P., Kozlov A.I., Yankovsky N.K. Molecular diagnosis and frequencies of primary hypolactasia in populations of Russia and neighboring countries. *Mol. Biol.* 2006; 40(6):931-935. DOI 10.1134/S0026893306060124

Cardona A., Pagani L., Antao T., Lawson D.J., Eichstaedt C.A., Yngvadottir B., Shwe M.T.T., Wee J., Romero I.G., Raj S., Metspalu M., Villems R., Willerslev E., Tyler-Smith C., Malyarchuk B.A., Derenko M.V., Kivisild T. Genome-wide analysis of cold adaption in indigenous Siberian populations. *PLoS One.* 2014;9:e98076. DOI 10.1371/journal.pone.0098076

Coelho M., Luiselli D., Bertorelle G., Lopes A.I., Seixas S., Destro-Bisol G., Rocha J. Microsatellite variation and evolution of human lactase persistence. *Hum. Genet.* 2005;117(4):329-339. DOI 10.1007/s00439-005-1322-z

Enattah N.S., Sahi T., Savilahti E., Terwilliger J.D., Peltonen L., Järvelä I. Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 2002;30(2):233-237. DOI 10.1038/ng826

Evershed R.P., Davey Smith G., Roffet-Salque M., Timpson A., Diekmann Y., Lyon M.S., Cramp L.J.E., … Tasić N., van Wijk I., Vostrovská I., Vuković J., Wolfram S., Zeeb-Lanz A., Thomas M.G. Dairying, diseases and the evolution of lactase persistence in Europe. *Nature.* 2022;608(7922):336-345. DOI 10.1038/s41586-022-05010-7

Haeggström S., Ingelman-Sundberg M., Pääbo S., Zeberg H. The clinically relevant CYP2C8*3 and CYP2C9*2 haplotype is inherited from Neandertals. *Pharmacogenomics J.* 2022;22(4):247-249. DOI 10.1038/s41397-022-00284-6

Heyer E., Brazier L., Ségurel L., Hegay T., Austerlitz F., Quintana-Murci L., Georges M., Pasquet P., Veuille M. Lactase persistence in Central Asia: phenotype, genotype, and evolution. *Hum. Biol.* 2011; 83(3):379-392. DOI 10.3378/027.083.0304

Ingram C.J., Elamin M.F., Mulcare C.A., Weale M.E., Tarekegn A., Raga T.O., Bekele E., Elamin F.M., Thomas M.G., Bradman N., Swallow D.M. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum. Genet.* 2007;120(6):779-788. DOI 10.1007/s00439-006-0291-1

Irving-Pease E.K., Refoyo-Martínez A., Barrie W., Ingason A., Pearson A., Fischer A., Sjögren K.G., … Korneliussen T., Werge T., Allentoft M.E., Sikora M., Nielsen R., Racimo F., Willerslev E. The selection landscape and genetic legacy of ancient Eurasians. *Nature*. 2024;625(7994):312-320. DOI 10.1038/s41586-023-06705-1

Kato K., Ishida S., Tanaka M., Mitsuyama E., Xiao J.Z., Odamaki T. Association between functional lactase variants and a high abundance of *Bifidobacterium* in the gut of healthy Japanese people. *PLoS One*. 2018;13(10):e0206189. DOI 10.1371/journal.pone.0206189

Khabarova Y., Tornianen S., Tuomisto S., Järvelä I., Karhunen P., Isokoski M., Mattila K. Lactase non-persistent genotype influences milk consumption and gastrointestinal symptoms in Northern Russians. *BMC Gastroenterol.* 2011;11:124. DOI 10.1186/1471-230X-11-124

Khabarova Y., Grigoryeva V., Tuomisto S., Karhunen P.J., Mattila K., Isokoski M. High prevalence of lactase non-persistence among indigenous nomadic Nenets, north-west Russia. *Int. J. Circumpolar Health.* 2012;71(1):1-6. DOI 10.3402/ijch.v71i0.17898

Kovalenko E., Vergasova E., Shoshina O., Popov I., Ilinskaya A., Kim A., Plotnikov N., Barenbaum I., Elmuratov A., Ilinsky V., Volokh O., Rakitko A. Lactase deficiency in Russia: multiethnic genetic study. *Eur. J. Clin. Nutr.* 2023;77(8):803-810. DOI 10.1038/s41430-023-01294-8

Kozlov A.I., Vershubskaya G.G. D-vitamin status and lactase persistence in European populations (review with the elements of meta-analysis. *Vestnik Moskovskogo Universiteta. Seria XXIII. Antropologia = Moscow University Anthropology Bulletin.* 2017;3:68-75 (in Russian)

Labrie V., Buske O.J., Oh E., Jeremian R., Ptak C., Gasiunas G., Maleckas A., Petereit R., Žvirbliene A., Adamonis K., Kriukienė E., Koncevičius K., Gordevičius J., Nair A., Zhang A., Ebrahimi S., Oh G., Šikšnys V., Kupčinskas L., Brudno M., Petronis A. Lactase nonpersistence is directed by DNA-variation-dependent epigenetic aging. *Nat. Struct. Mol. Biol.* 2016;23(6):566-573. DOI 10.1038/nsmb.3227

Leseva M.N., Grand R.J., Klett H., Boerries M., Busch H., Binder A.M., Michels K.B. Differences in DNA methylation and functional expression in lactase persistent and non-persistent individuals. *Sci. Rep.* 2018;8(1):5649. DOI 10.1038/s41598-018-23957-4

Liebert A., López S., Jones B.L., Montalva N., Gerbault P., Lau W., Thomas M.G., Bradman N., Maniatis N., Swallow D.M. World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Hum. Genet*. 2017;136(11-12):1445-1453. DOI 10.1007/s00439-017-1847-y

Mattar R., Monteiro M., Silva J.M., Carrilho F.J. LCT-22018G>A single nucleotide polymorphism is a better predictor of adult-type hypolactasia/lactase persistence in Japanese-Brazilians than LCT-13910C>T. *Clinics* (*Sao Paulo*). 2010;65(12):1399. DOI 10.1590/s1807-59322010001200030

Olds L.C., Sibley E. Lactase persistence DNA variant enhances lactase promoter activity *in vitro*: functional role as a *cis* regulatory element. *Hum. Mol. Genet.* 2003;12(18):2333-2340. DOI 10.1093/hmg/ddg244

Pagani L., Lawson D.J., Jagoda E., Mörseburg A., Eriksson A., Mitt M., Clemente F., … Thomas M.G., Manica A., Nielsen R., Villems R., Willerslev E., Kivisild T., Metspalu M. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016; 538(7624):238-242. DOI 10.1038/nature19792

Pairo-Castineira E., Clohisey S., Klaric L., Bretherick A.D., Rawlik K., Pasko D., Walker S., … Maslove D., Ling L., McAuley D., Montgomery H., Walsh T., Pereira A.C., Renieri A.; GenOMICC Investigators; ISARIC4C Investigators; COVID-19 Human Genetics Initiative; 23andMe Investigators; BRACOVID Investigators; GenCOVID Investigators; Shen X., Ponting C.P., Fawkes A., Tenesa A., Caulfield M., Scott R., Rowan K., Murphy L., Openshaw P.J.M., Semple M.G., Law A., Vitart V., Wilson J.F., Baillie J.K. Genetic mechanisms of critical illness in COVID-19. *Nature*. 2021; 591(7848):92-98. DOI 10.1038/s41586-020-03065-y

Peng M.S., He J.D., Zhu C.L., Wu S.F., Jin J.Q., Zhang Y.P. Lactase persistence may have an independent origin in Tibetan populations from Tibet, China. *J. Hum. Genet.* 2012;57(6):394-397. DOI 10.1038/jhg.2012.41

Pilipenko I.V., Pristyazhnyuk M.S., Kobzev V.F., Voevoda M.I., Pilipenko A.S. Polymorphism of the *LCT* gene regulatory region in Turkic-speaking populations of the Altay-Sayan region (southern Siberia). *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2016;20(6):887-893. DOI 10.18699/VJ16.209 (in Russian)

Racimo F., Gokhman D., Fumagalli M., Ko A., Hansen T., Moltke I., Albrechtsen A., Carmel L., Huerta-Sanchez E., Nielsen R. Archaic adaptive introgression in TBX15/WARS2. *Mol. Biol. Evol.* 2017; 34(3):509-524. DOI 10.1093/molbev/msw283

Raz M., Sharon Y., Yerushalmi B., Birk R. Frequency of LCT-13910C/T and LCT-22018G/A single nucleotide polymorphisms associated with adult-type hypolactasia/lactase persistence among Israelis of different ethnic groups. *Gene.* 2013;519(1):67-70. DOI 10.1016/j.gene.2013.01.049

Scott A., Reinhold S., Hermes T., Kalmykov A.A., Belinskiy A., Buzhilova A., Berezina N., … Krause R., Karapetian M., Stolarczyk E., Krause J., Hansen S., Haak W., Warinner C. Emergence and intensification of dairying in the Caucasus and Eurasian steppes. *Nat. Ecol. Evol*. 2022;6(6):813-822. DOI 10.1038/s41559-022-01701-6

Ségurel L., Bon C. On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 2017;8:297-319. DOI 10.1146/annurev-genom-091416-035340

Segurel L., Guarino-Vignon P., Marchi N., Lafosse S., Laurent R., Bon C., Fabre A., Hegay T., Heyer E. Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biol.* 2020;18(6):e3000742. DOI 10.1371/journal.pbio.3000742

Slon V., Mafessoni F., Vernot B., de Filippo C., Grote S., Viola B., Hajdinjak M., Peyregne S., Nagel S., Brown S., Douka K., Higham T., Kozlikin M.B., Shunkov M.V., Derevianko A.P., Kelso J., Meyer M., Prüfer K., Pääbo S. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature.* 2018;561(7721):113-116. DOI 10.1038/s41586-018-0455-x

Sokolova M.V., Vasilyev E.V., Kozlov A.I., Rebrikov D.V., Senkeeva S.S., Kozhekbaeva Zh.M., Lyundup A.V., Svechnikova N.S., Ogurtsov P.P., Khusnutdinova E.K., Yankovsky N.K., Borinskaya S.A. Polymorphism C/T-13910 of the *LCT* gene regulatory region and lactase deficiency in Eurasian populations. *Ekologicheskaya Genetika = Ecological Genetics.* 2007;5:25-34. DOI 10.17816/ecogen5325-34 (in Russian)

Sun H.M., Qiao Y.D., Chen F., Xu L.D., Bai J., Fu S.B. The lactase gene *–13910T* allele can not predict the lactase-persistence phenotype in north China. *Asia Pac. J. Clin. Nutr.* 2007;16(4):598-601

Telis N., Aguilar R., Harris K. Selection against archaic hominin genetic variation in regulatory regions. *Nat. Ecol. Evol.* 2020;4(11):1558-1566. DOI 10.1038/s41559-020-01284-0

Tishkoff S., Reed F., Ranciaro A., Voight B.F., Babbitt C.C., Silverman J.S., Powell K., Mortensen H.M., Hirbo J.B., Osman M., Ibrahim M., Omar S.A., Lema G., Nyambo T.B., Ghori J., Bumpstead S., Pritchard J.K., Wray G.A., Deloukas P. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet*. 2007;39(1):31-40. DOI 10.1038/ng1946

Troelsen J.T., Olsen J., Møller J., Sjöström H. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*. 2003;125(6):1686-1694. DOI 10.1053/j.gastro.2003.09.031

Xu L., Sun H., Zhang X., Wang J., Sun D., Chen F., Bai J., Fu S. The *–22018A* allele matches the lactase persistence phenotype in northern Chinese populations. *Scand. J. Gastroenterol.* 2010;45(2):168-174. DOI 10.3109/00365520903414176

Zeberg H., Dannemann M., Sahlholm K., Tsuo K., Maricic T., Wiebe V., Hevers W., Robinson H.P.C., Kelso J., Pääbo S. A Neanderthal sodium channel increases pain sensitivity in present-day humans. *Curr. Biol.* 2020;30(17):3465-3469.e4. DOI 10.1016/j.cub.2020.06.045

Zeberg H., Jakobsson M., Pääbo S. The genetic changes that shaped Neandertals, Denisovans, and modern humans. *Cell*. 2024;187(5):1047-1058. DOI 10.1016/j.cell.2023.12.029

Zhang X., Witt K.E., Banuelos M.M., Ko A., Yuan K., Xu S., Nielsen R., Huerta-Sanchez E. The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *Proc. Natl. Acad. Sci. USA*. 2021;118(22):e2020803118. DOI 10.1073/pnas.2020803118

Conflict of interest. The author declares no conflict of interest.

Received April 26, 2024. Revised May 31, 2024. Accepted June 3, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):659-666

DOI 10.18699/vjgb-24-73

Traces of Paleolithic expansion in the Nivkh gene pool based on data on autosomal SNP and Y chromosome polymorphism

V.N. Kharkov [1] ✉, N.A. Kolesnikov [1], L.V. Valikhova[1], A.A. Zarubin [1], A.L. Sukhomyasova [2], I.Yu. Khitrinskaya [1], V.A. Stepanov [1]

[1] Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
[2] M.K. Ammosov North-Eastern Federal University, Yakutsk, Russia
✉ vladimir.kharkov@medgenetics.ru

Abstract. The Nivkhs are a small ethnic group indigenous of the Russian Far East, living in the Khabarovsk Territory and on Sakhalin Island, descending from the ancient inhabitants of these territories. In the Nivkhs, a specific Sakhalin-Amur anthropological type is prevalent. They are quite isolated, due to long isolation from contacts with other peoples. The gene pool of the Nivkhs and other Far Eastern and Siberian populations was characterized using a genome-wide panel of autosomal single-nucleotide polymorphic markers and Y chromosome haplogroups. Bioinformatic processing of frequencies of autosomal SNPs, Y chromosome haplogroups and YSTR haplotypes showed that the Nivkh gene pool is very different from the other populations'. Analysis of the SNP frequencies using the PCA method divided the Far Eastern populations in full accordance with the territories of their residence into the northern group of the Chukchi and Koryaks and the southern group, including the Nivkhs and Udege. The remoteness of the Nivkhs coincides with their geographic localization, with the Nivkhs and Udege demonstrating the greatest kinship. The Nivkhs have a specific component of their gene pool, which is present with much less frequency in the Udege and Transbaikal Evenks. According to the IBD blocks, the genotypes of the Nivkhs show a very small percentage of coincidence with the Udege, Koryaks, Evenks and Chukchi, the value of which is the lowest compared to the IBD blocks among all other Siberian populations. The Nivkh-specific composition of haplogroups and YSTR haplotypes was shown. In the Nivkhs, the C2a1 haplogroup is divided into three sublines, which have a fairly ancient origin and are associated with the ancestors of modern northern Mongoloids. The Nivkh haplogroup O2a1b1a2a-F238 is found among residents of China and Myanmar. The Q1a1a1-M120 line is represented among the Nivkhs, Koryaks, Evenks and Yukaghirs. Phylogenetic analysis of individual Y chromosomal haplogroups demonstrated the closeness of the Nivkh gene pool with the ancient population of the Amur and Okhotsk regions, the Koryaks, the Tungus peoples and the population of Southeast Asia. The Nivkh gene pool confirms the relative smallness of their ancestral groups without mixing with other populations.
Key words: gene pool; human populations; genetic diversity; genetic components; Y chromosome; Nivkhs.

For citation: Kharkov V.N., Kolesnikov N.A., Valikhova L.V., Zarubin A.A., Sukhomyasova A.L., Khitrinskaya I.Yu., Stepanov V.A. Traces of Paleolithic expansion in the Nivkh gene pool based on data on autosomal SNP and Y chromosome polymorphism. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):659-666. DOI 10.18699/vjgb-24-73

Funding. The study was supported by the Russian Science Foundation grant No. 22-64-00060, https://rscf.ru/project/22-64-00060/.

Следы палеолитической экспансии в генофонде нивхов по данным о полиморфизме аутосомных SNP и Y-хромосомы

В.Н. Харьков [1] ✉, Н.А. Колесников [1], Л.В. Валихова[1], А.А. Зарубин [1], А.Л. Сухомясова [2], И.Ю. Хитринская [1], В.А. Степанов [1]

[1] Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
[2] Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Россия
✉ vladimir.kharkov@medgenetics.ru

Аннотация. Нивхи – малочисленный коренной народ Дальнего Востока, проживающий на территории Хабаровского края и острова Сахалин, который относится к потомкам древнего населения этих территорий. У нивхов преобладает специфичный сахалино-амурский антропологический тип. Они являются достаточно обособленными за счет длительной изоляции от контактов с другими народами. Генофонд нивхов охарактеризован по полногеномной панели аутосомных однонуклеотидных полиморфных маркеров и гаплогруппам Y-хромосомы в сравнении с другими дальневосточными и сибирскими популяциями. Биоинформатическая обработка частот

аутосомных SNP, гаплогрупп Y-хромосомы и YSTR-гаплотипов показала, что генофонд нивхов существенно отличается от генофондов других популяций. При анализе частот SNP методом PCA дальневосточные популяции располагаются в полном соответствии с территориями их проживания и делятся на северную группу чукчей и коряков и южную, включающую нивхов и удэгейцев. Удаленность нивхов совпадает с их географической локализацией, при этом нивхи и удэгейцы демонстрируют наибольшее родство. У нивхов выделяется специфичный для них компонент генофонда, который с гораздо меньшей частотой присутствует у удэгейцев и забайкальских эвенков и бурятов-А. По IBD-блокам генотипы нивхов демонстрируют очень небольшую долю совпадения с удэгейцами, коряками, эвенками и чукчами, значение которых является самым низким по сравнению с IBD-блокам между другими сибирскими популяциями. Показан специфичный для нивхов состав гаплогрупп и YSTR-гаплотипов. Гаплогруппа C2a1 у нивхов разделена на три сублинии, которые имеют достаточно древнее происхождение и связаны с предками современных северных монголоидов. Нивхская гаплогруппа O2a1b1a2a-F238 есть у жителей Китая и Мьянмы. Линия Q1a1a1-M120 в исследованных в данной работе выборках представлена у нивхов, коряков, эвенков и юкагиров. Филогенетический анализ отдельных Y-хромосомных гаплогрупп демонстрирует близость генофонда нивхов с коряками и тунгусскими народами, а также родство в меньшей степени с древним населением Приамурья и Приохотья и населением Юго-Восточной Азии. Генофонд нивхов подтверждает относительную малочисленность их предковой группы без смешения с другими популяциями.

Ключевые слова: генофонд; популяции человека; генетическое разнообразие; генетические компоненты; Y-хромосома; нивхи.

Introduction

The Nivkh people are a small ethnic group that lives in the Far Eastern regions of Sakhalin Island and the lower Amur Basin. In 2022, there were about 3,842 Nivkh people. They self-identify as nivkhgu. Neighboring ethnic groups call them Gilyak or Gilyami, and the Russians adopted this name, calling them Gilak. In Tungusic and Manchu languages, the word "Gilyaki" means "people who move with the help of oar-powered boats".

Based on their territory, the Nivkhs can be divided into two groups: the island group (Sakhalin) and the mainland group. In the past, they occupied a much larger territory. On the mainland, their settlement area extended from the Amur River to the Uda Basin, and on Sakhalin, they lived along the western and eastern coastlines and at the mouth of the Poronai River. Nowadays, the Sakhalin Nivkhs live in the northern part of the island and the Tym River basin. In the mainland, the Nivkh people are concentrated in two districts of the Khabarovsk Territory: Nikolaevsky and Ulchsky. They speak a language called Nivkh, which has two dialects: Amur and East Sakhalin. Nivkh is an isolated language, along with Ket and Yukaghir. It was previously classified as part of the Paleoasiatic language family due to its unclear genealogical origins. A strong relationship between Nivkh and the Chukchi-Kamchatkan languages was found in the work of M.D. Fortescue (2011).

The Nivkh people are direct descendants of the ancient population that inhabited Sakhalin and the lower reaches of the Amur River in the past. They are part of the Paleo-Asiatic group of the Mongoloid race, and their anthropological type is similar to that of the Sakhalin-Amur people, which can also be found among the Ulchi people. Together with the Chukchi, Koryaks, and other people of Northeastern Siberia, the Nivkhs belong to the Paleoasian group. There is a theory that the ancestors of the modern Nivkhs, as well as the Eskimo and Native American peoples, were all links in the same ethnic chain that once covered the northwestern coast of the Pacific Ocean. The modern appearance of the Nivkhs has been significantly influenced by their cultural and ethnic interactions with the Tungus-Manchu, Ainu, and Japanese people (The Peoples of Russia, 1994; Sulyandziga et al., 2003; Peoples of North-East Siberia, 2010).

The data obtained from genotyping high-density microarrays for autosomal SNPs in the Nivkhs and other Far Eastern and Siberian indigenous peoples allow us to more accurately describe their gene pool composition, identify common haplotype blocks, and homozygosity patterns compared to limited sets of DNA markers. Genotyping a larger set of specific Y-chromosome SNPs enables a more detailed characterization of the molecular phylogenetic structure of Y-haplogroups. Modern bioinformatics methods for individual genotype analysis allow us to characterize the gene pool of the studied samples as thoroughly as possible using various techniques.

There are a vast number of Single Nucleotide Polymorphisms (SNPs) in the human genome, which makes them an effective tool for analyzing genetic relationships between populations. Modern population genetics has various marker systems, including autosomal and homologous DNA markers that determine the phylogeny of the Y-chromosome and mitochondrial DNA haplogroups.

A specific feature of the mitochondrial gene pools in all Primorye populations is the presence of mtDNA lines belonging to haplogroup Y. The maximum frequencies of these lines were noted in the Sakhalin Nivkh (66.1 %) and Ulchi populations (37.9 %). The frequency of this line is also high in the Ainu (25.5 %), Negidal (21.2 %) (Starikovskaya et al., 2005), Koryak (5.7 %), Even (8.1 %), and Eastern Evenk (8.9 %) populations (Derenko, Malyarchuk, 2010). However, the frequency of this mtDNA line in other Asian populations is significantly lower and decreases as one moves away from the territories where the main carriers of this line reside. The origin of these specific mtDNA lines is associated with the lower reaches of the Amur River and Sakhalin.

The distribution of the Y1a1 mitochondrial DNA (mtDNA) subgroup is limited to the Northeast Asian region. All the mtDNA lines found in the Koryak, Even, Itel'men, Negidal, Nivkh, Orok, and Ainu populations belong exclusively to this subgroup (Horai et al., 1996; Schurr et al., 1999; Bermisheva et al., 2005; Starikovskaia et al., 2005; Derenko, Malyarchuk, 2010). The main area of this mtDNA line and the frequencies of its sublineages correlate well with the distribution of the C2a1 Y-chromosome haplogroup. This is an example of parallel expansions of Y- and mtDNA haplogroups in the same

region. These findings are consistent with previous research on ancient genomes from the Amur River basin, which formed a distinct genetic cluster including both ancient and modern populations from the region (people speaking Tungusic languages and Nivkh) (Wang et al., 2021).

The purpose of this study is to conduct a comprehensive analysis of the genetic structure of the Nivkh population in comparison with other indigenous populations of Siberia and the Far East. In order to address questions about the genetic affinity of the Nivkhs to other indigenous groups, we have performed genotyping for a wide range of autosomal markers on high-density DNA microarrays, as well as for a larger set of SNPs and STR markers on the Y chromosome, in various ethnic groups such as the Udege, Chukchi, Koryak, Yakut, Evenk, Buryat, Tuvinian, Khakass, Southern Altai, Ket, Chulym, and Khant.

Materials and methods

The research material consisted of DNA samples from men and women from the Nivkh population (N = 155) living in the settlements of Nekrasovka and Moskalvo, in the Okhinsky district of the Sakhalin region. Venous blood was collected from donors in accordance with the written informed consent procedure for conducting the study (Protocol No. 10 of the Biomedical Ethics Committee of the Research Institute of Medical Genetics, dated 02/15/2021). For each donor, a questionnaire was completed with a brief family history, indicating ethnicity, ancestral place of birth, and other relevant information. Individuals were assigned to an ethnic group based on their own ethnic identity, their parents' ethnic background, and the place of their birth.

52 DNA samples from the Nivkh population were used to analyze Y-chromosomal haplogroups and haplotypes in men. For high-density genotyping, unrelated Nivkh samples (N = 13) without intermarriage with other ethnic groups were selected. This small number of samples is due to the significant proportion of interethnic marriages among the collected individuals over the past few generations, as well as the relatively small size of the Nivkh population and the presence of close relatives on both the maternal and paternal sides in the samples.

Other populations of the indigenous people of Siberia included in this study are represented by the Udege (N = 15), Koryaks (N = 20), and Chukchi (N = 25). Samples of the Udege were collected from the villages of Krasny Yar and Agzu, in the Pozharsky and Terneysky districts of Primorsky Krai, respectively. The Koryak samples were collected in the Koryak Autonomous Okrug in the Kamchatka region, and the Chukchi samples were collected from various settlements in the Chukotka and Chukotkan Autonomous Okrugs, including the coastal regions of Lorino, Sireniki, Yanaryk, and Novoe Chaplino. Southern Altaians were also included in the study, with samples collected from the Beshpeltir (N = 24) and Kulada (N = 25) villages in the Chemalsky and Ongudaysky districts, respectively. Finally, the Ket samples were collected (N = 15) in the Kellogg settlement of the Turukhansky district in the Krasnoyarsk region; other samples were collected in Tomsk Tatar (Chernaya Rechka, Eushta, and Takhtamyshevo in the Tomsk area, N = 20), Tuvinian (Teeli in Bai-Tayga kozhuun, N = 28), Buryat (Aginskoe in the Aginsky district N = 23 and Kurumkan in the Kurumkan district, N = 28), Khanty (Kazym village in the Beloyarsk district N = 30 and Ruskinskaya in Surgut district N = 26), Khakass (Tashtypsky, N = 29 and Shirinsky N = 26 districts), Chulym (N = 22), Evenk (Zabaikalsky villages Chara, Moklakan, and Tupik, N = 25; Y – Evenks of Yakutia, N = 28), Yakut (Ust-Aldan district village Cheriktey, N = 26) settlements. The material is stored in the bioresources collection "Biobank of the Population of Northern Eurasia".

Genome-wide genotype data were obtained using the Infinium Multi-Ethnic Global 8 microarrays (Illumina), which include over 1.7 million single nucleotide polymorphisms (SNPs). Clustering of the SNP genotype array and quality control were performed using a protocol developed by Y. Guo et al. (2014), using GenomeStudio software (Illumina GenomeStudio, version 2.0.3). A standard set of tools, including vcftools, bcftools, and plink, were used for filtering, normalization, and calculation of standard genomic statistics and metrics.

The Refined IBD algorithm (Browning B.L., Browning S.R., 2013) was used to analyze cluster blocks that are identical in origin. This algorithm produced more accurate results than the algorithms built into plink. The genotypes had been previously phased using the Beagle 5.1 software (Browning S.R., Browning B.L., 2007). To compare populations, we obtained the sums of the average lengths of clusters that were identical in origin between pairs of individuals.

PCA was used to analyze genetic relationships between populations. The NGSadmix technique (Skotte et al., 2013) and ADMIXTURE program (Alexander et al., 2009; Alexander, Lange, 2011) were used to determine the component composition and amount of impurities in individuals and populations.

To study the composition and structure of the Y chromosome, two systems of genetic markers were used in the study: diallelic loci represented by single nucleotide polymorphisms (SNPs) and polyallelic microsatellites with high variability (YSTRs). Using 589 SNPs, men were classified into different haplogroups. Genotyping of SNPs was performed using the polymerase chain reaction (PCR) method and subsequent analysis of DNA fragments through RFLP analysis (restriction fragment length polymorphism). For specific terminal SNPs, a small number of samples were genotyped for individual sub-haplogroups according to their YSTR haplotypes, and the results were obtained through NGS (next-generation sequencing) of the Y chromosome. Haplogroups were designated based on the ISOGG (International Society of Genetic Genealogy) 2019 Y-DNA Haplogroup Tree classification. Analysis of STR haplotypes within haplogroups was carried out using 44 STR markers of the non-recombining part of the Y chromosome (DYS19, 385a, 385b, 388, 389I, 389II, 390, 391, 392, 393, 426, 434, 435, 436, 437, 438 , 439, 442, 444, 445, 448, 449, 456, 458, 460, 461, 481, 504, 505, 518, 525, 531, 533, 537, 552, 570, 576, 635, 643, YCAIIa, YCAIIb, GATA H4.1, Y-GATA-A10, GGAAT1B07). STR markers were genotyped using capillary electrophoresis on ABI Prism 3730 and Nanofor-05 devices.

Experimental studies were conducted at the Center for Collective Use of Scientific Research Equipment "Medical Genomics", which is part of the Research Institute of Medical Genetics at the Tomsk Scientific Center. The median net-

works of Y-chromosome haplotypes were created using the Network v10.2.0.0 software (Fluxus Technology Ltd.; www.fluxus-engineering.com), using the Bandelt median network method (Bandelt et al., 1999). The age of the haplotype diversity observed in haplogroups was estimated using the ASD method (Zhivotovsky et al., 2004), based on the average square difference in the number of repeats between all markers.

Results and discussion

After processing the data based on the results of the microarray study to filter the genotyped samples and carry out further calculations, a search was carried out among the Nivkhs for mestizos using the NGSadmix program. The NGSadmix method, when launched on the data array we generated, showed that all samples of pure Nivkhs do not have crossbreeding with other peoples, which coincides with the results of their survey.

Genetic relationships of the Nivkhs with the peoples of Eastern and Northeastern Siberia

When analyzing data on the frequencies of autosomal SNPs using the PCA method at the level of individual samples (Fig. 1), it is clear that the Nivkhs are closest to the Udege, as well as to the Evenks from Transbaikalia and Yakutia. The Chukchi and Koryaks are very distant from all other populations in the figure, which is consistent with their strong geographic isolation in Northeastern Siberia. It is PC2 that separates them from all analyzed samples, but according to PC1 they are very close to the Nivkhs and Udege. Their strong distance from more southern peoples indicates the presence in the Chukchi-Koryak gene pool of an older, specific genetic component associated with the aboriginal Paleolithic population of the territories where they lived. The Far Eastern samples are divided in full accordance with the territories of their residence into the northern group of the Chukchi and Koryaks and the southern group including the Nivkhs and Udege. The Evenks from Transbaikalia and Yakutia are also close to each other. The Yakuts and Buryats are a little more remote. The distance between the Nivkhs and all other populations in the figure coincides with their geographic location. The Nivkhs, Udege, Chukchi and Koryaks make up the Far Eastern group of populations, with the Nivkhs and Udege showing the greatest kinship.

Almost all samples of individual ethnic groups form specific clusters (Fig. 1), which can partially overlap in this figure, with the exception of the Tomsk Tatars, who have a fairly heterogeneous composition of the gene pool (Valikhova et al., 2022). In the three-component analysis and in the t-SNE plot, all ethnospecific clusters are much more distant from each other. Individual samples from different samples that stand out from these general groups show crossbreeding when analyzed by the NGS-Admix method, which affects their location on the graph.

Component composition of the gene pool of populations

To determine the genetic components in the gene pool of the studied populations, the Admixture program was used, which makes it possible to identify the heterogeneity of the component composition of the genome of individuals based on genotype data and accurately determine their distribution at the level of populations and individual samples. When setting the number of ancestral components to more than four, in most studied populations a genetic component specific to the Nivkhs is revealed, most clearly manifested in the analyzed array of population samples at K = 8, which can be interpreted as the "Sakhalin-Amur" genetic layer in the gene pool of modern populations (Fig. 2).

At K = 8, this component completely dominates among the Nivkhs (0.92) and Udege (0.61), and is found among the Buryats (0.50), Altai-Kizhi (0.34), Khakass-Kachins, Tuvinians (0.30), Altaians of the village of Beshpeltir (0.23), Tomsk Tatars (0.11), Evenks (0.02–0.09), Khakass Sagais (0.07) and Yakuts (0.01). It is possible that this genetic layer is associated with an ancient substrate in these populations.

At K = 10, a more detailed separation occurs (Fig. 3): the Nivkhs have a component specific to them (0.98), highlighted in blue in Figure 3, which is present among the Udege (0.22), and to a small extent among the Transbaikal Evenks (0.05), the Evenks of Yakutia, Khakassians, Tomsk Tatars and Buryats (0.02). The dominance in the frequency of this component in all Nivkh samples confirms that their ancestors had no

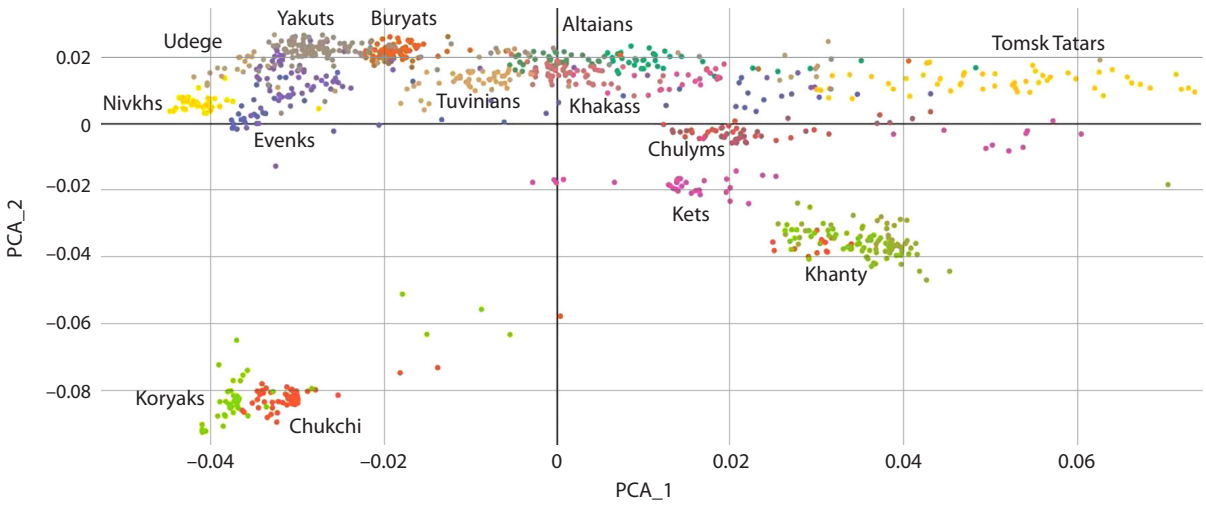


Fig. 1. Differentiation of genomes of the population of the Far East and Siberia according to two PCA components.

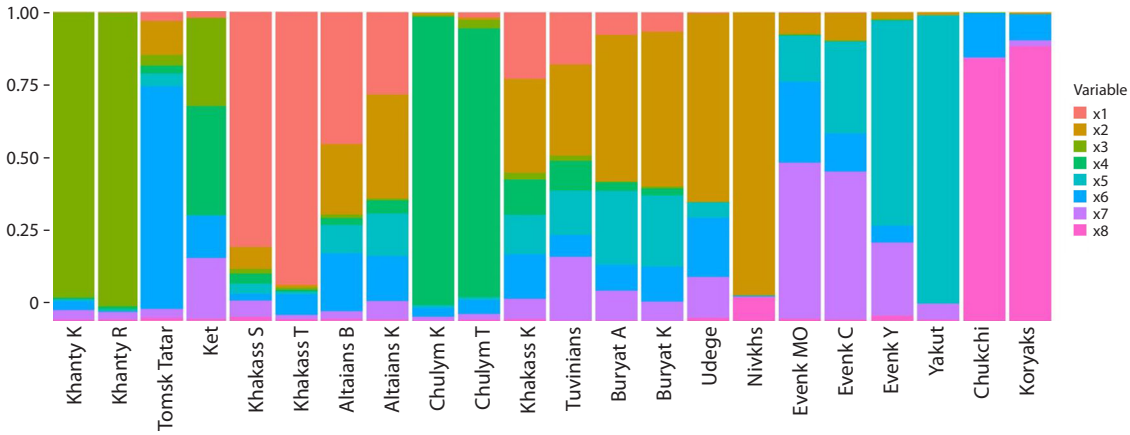


Fig. 2. Ordered picture of Admixture components when ranking Siberian populations from west to east, K = 8.

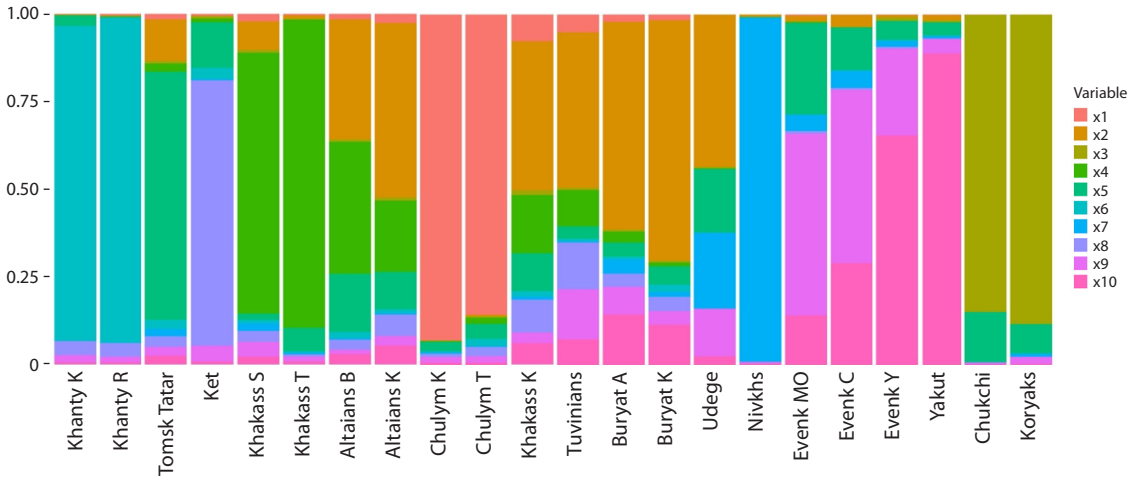


Fig. 3. Ordered picture of Admixture components when ranking Siberian populations from west to east, K = 10.

contact with other peoples for quite a long time and lived in isolation on Sakhalin Island. The data obtained prove that the indigenous population of Sakhalin did not mix with other ethnic groups for a long time.

Blocks identical by descent

Coincidence analysis was carried out at the individual and population levels to assess common ancestry DNA blocks. A fragment that has identical nucleotide sequences in different people is the legacy of their common ancestor. The size of this IBD segment is comparable to the number of generations due to chromosome recombination during the formation of germ cells. The use of information about these genomic regions of common origin at the level of individuals and populations makes it possible to quantify the degree of genetic relatedness between people and provides additional information about the genetic connections of populations (Gusev et al., 2012).

The genotypes of the Nivkhs showed a coincidence in IBD blocks with each other >1.5 cM (11 %), then with the Udege (0.58 %), Koryaks (0.47 %), Evenks (0.28 %) and Chukchi (0.18 %). With other Siberian populations, their share is much lower (Fig. 4). The agreement between the Nivkhs and other studied populations is the lowest compared to other ethnic groups. This confirms their very long isolation and lack of contact with other peoples. The proportion of interpopulation IBD blocks between the Nivkhs, Udege, Koryaks, Chukchi, and Evenks is consistent with the results of PCA and Admixture. Analysis of IBD within the Nivkh, Koryak and Chukchi populations showed that they have more common IBD than people from other samples. At the same time, among the Chukchi (55 %), Koryaks (57 %) and Nivkhs (59 %), the greatest contribution is made by short IBD fragments, which may indicate a “bottleneck” in the past during migrations to the north and northeast or isolation from other populations inhabiting the territory of Siberia.

Genomic inbreeding coefficient

When estimating the genomic inbreeding coefficient for ROH lengths >1.5 Mb, the Nivkhs have a relatively low level of consanguinity (FROH = 0.0268). Among the Koryaks (FROH = 0.0446) and Chukchi (FROH = 0.0431), it is maximum for Siberian populations and is almost twice as high as

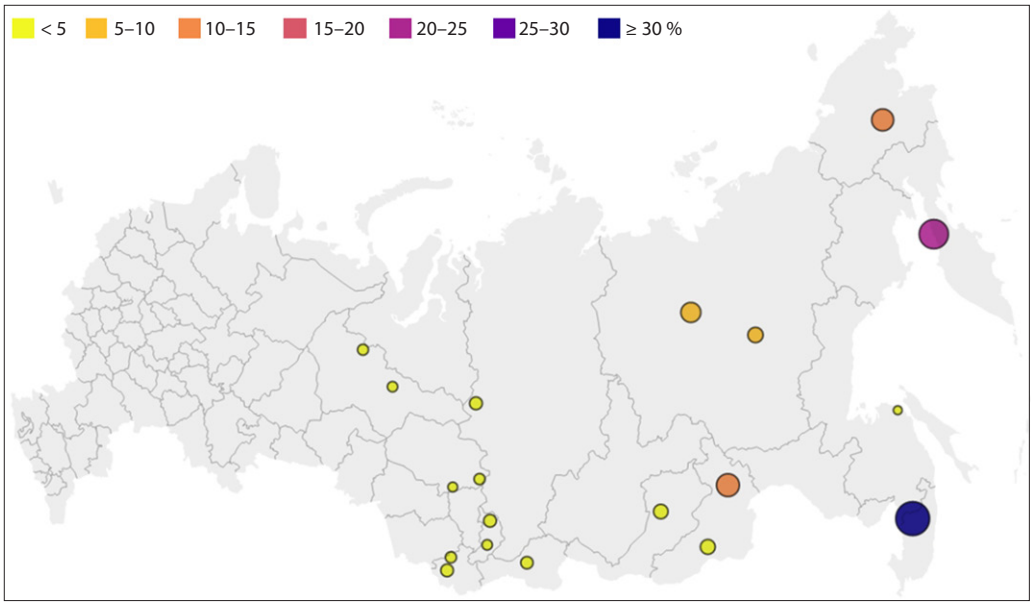


Fig. 4. Sum of segment lengths IBD >1.5 cM between pairs of Nivkh individuals and Siberian populations.

their average value in the territory of Siberia and the Far East. For the Nivkhs, Chukchi and Koryaks, a significant increase in the total length of the average ROH class per individual has been shown compared to other populations. This adds to the comparison with the short ROH class in Siberian populations. The results obtained indicate a relatively small number of ancestral groups of these peoples over many generations and marriage contacts between relatives, as well as a possible "bottleneck" effect. The level of homozygosity in the genomes of representatives of these Far Eastern peoples shows the highest level of inbreeding among all indigenous Siberian peoples. They have long homozygous stretches for all ROH length categories in most samples examined. These results confirm the relatively small size of their ancestral groups over a long period of time and their territorial isolation, which precluded mixing with other populations.

Y-chromosome haplogroups

The results of genotyping of SNP and YSTR markers and determination of Y-chromosome haplogroups in all samples of Nivkh men have been shown to match the data of their questionnaires on the paternal side. All men who are mestizos with Eastern European peoples on their father's side belong to specific European sublines of haplogroups E, I1, N1a1, N1a2 and R1a1. Haplogroups of mestizos with Koreans and Orochons belong to the East Asian variants of the C2, O1 and O2 clades. All other Nivkh samples belong to sublines of three haplogroups specific to them.

The most common haplogroup among the Nivkhs is C2a1 (86 %). With such a high frequency, this Y-chromosomal line has not been recorded in any of the analyzed ethnic groups, and is maximum in purebred male-line Nivkhs compared to other peoples. It is a substrate element of their gene pool, associated with autochthonous population groups of the Okhotsk region.

Of the 37 Nivkh men without paternal crossbreeding, 16 people belong to the C2a1a subline (B90, Z32902, Z32912, Z32919, Z32926, Z32937 (xB93, Z32958)) (see the Table). The age of this lineage was previously determined to be 4,216 years (3,700–4,667) (Liu et al., 2021). This branch forms a special cluster of YSTR haplotypes, characterized by a reduction to ten in the number of tandem repeats in the DYS389I locus, specific for the Nivkhs and Koryaks. The parallel line C2a1a-B93 is also present among the Evenks, Evens, Koryaks, Yukaghirs and Yakuts. Among the Yakut Evenks and Yukaghirs, it is 15–20 %. In a very large sample of Yakuts, only four samples belong to it. One example of this

Frequencies of Y-chromosome haplogroups among the Nivkhs

Haplogroup	% (*N* = 37)
C2a1a1b1a~ – B473, F10085, F13958 (xZ32848, FGC28920, BY186309)	32.4 % (12)
C2a1a – B90, Z32902, Z32912, Z32919, Z32926, Z32937 (xB93, Z32958)	43.2 % (16)
C2a1 – F3447, ACT1932, ACT1942	10.8 % (4)
O2a1b1a2a – F238	5.4 % (2)
Q1a1a1 – M120, F746, Y34108, Y34449	8.1 % (3)

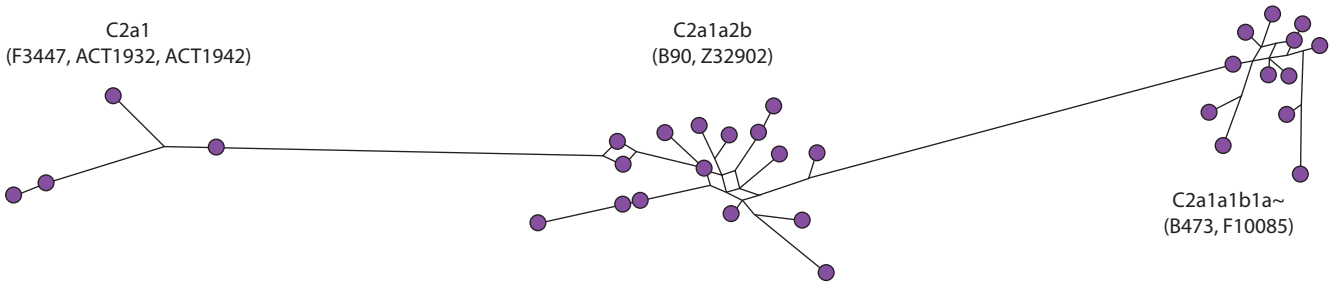


Fig. 5. Median network of YSTR haplotypes of haplogroup C2a1 in the Nivkhs.

branch has also been found among the Transbaikal Evenks. The presence of a specific branch C2a1a2b (B93) in these populations is associated with the ancient indigenous populations of the Amur and Okhotsk regions, which separated from the Asian ancestors from more southern regions a long time ago. According to the research team from Tartu (Karmin et al., 2015), three samples of Koryak men belonging to haplogroup C3c2 have haplotypes almost completely identical to our samples from this line. In two Evenks from Mongolia (Liu et al., 2021) and one from Russia, the C2a1a2b–B90 subline was also discovered (Karmin et al., 2015). This branch is related to C2a1a2b–M86, which previously split with the C2a–M48 branch about 11.6 Kya (Liu et al., 2021). Its spread in Eastern Siberia is associated with the relatively recent migration of Tungus tribes from the Amur region and Manchuria. The Nivkh-specific subline C2a1a2b (xB93) separated from the common ancestor even before the formation of the B93 mutation among the Tungusic peoples.

The second most common line among the Nivkhs is C2a1a1b1a~ F13958 (32.4 %). This line was found in one Kazakh and three Kyrgyz, but in terms of haplotypes they differ significantly from the Nivkhs. According to the YFull website, the age of its common ancestor is 4,300 years (CI: 5,200–3,500). The C2a1 lineage (F3447, ACT1932, ACT1942) includes four Nivkhs. According to the YFull website, the age of its common ancestor is 16,000 years (CI: 17,300–14,800). This line of ancient origin was found in two Chinese people from Liaoning Province, a Korean and a Japanese person.

The large diversity of C2a1 lines among the Nivkhs and their age indicate a very early appearance of this haplogroup in the indicated territory. The spread of this line during the formation of the gene pool of the ancient population of Northeast Asia is associated with the early migrations of Mongoloid tribes. Thus, C2a1 is a marker for the settlement of the ancestors of modern northern continental Mongoloids and their further differentiation in Siberia, as well as the second wave of settlement of America, the representatives of which retained the morphological features of the ancient proto-Mongoloids of Asia.

In general, the Nivkh gene pool, in terms of autosomal SNPs and Y-chromosome haplogroups, on the one hand, occupies an intermediate position between the gene pools of the Koryaks and Udege; on the other hand, it is less diverse in composition and is distinguished by the presence of three specific variants. The highest frequency of haplogroups C2a1a–B90 (xB93) among Siberian populations makes it a unique object for studying the Paleolithic layers of the total Far Eastern gene pool and reconstructing the earliest stages of human settlement of Northeast Asia.

The overall median network of haplotypes of haplogroup C2a1 is very branched, and consists of three clusters of haplotypes that match the genotypes of terminal SNPs for these sublineages (Fig. 5). This corresponds to an estimate of the time of their separation. All three clusters demonstrate the presence of common male ancestors, the descendants of which are all analyzed Nivkh samples.

Thus, the populations that brought haplogroup C2a1 to the territory of the Amur region and Kamchatka apparently migrated north along the Pacific coast. The greatest haplotype diversity of C2a1 in the Far East indicates a significantly earlier appearance of this haplogroup in this territory, in comparison with Southern Siberia. The spread of this line during the formation of the gene pool of the ancient population of North Asia is apparently associated with the migrations of Mongoloid tribes that formed the Central Asian, Baikal and Arctic groups of anthropological types.

Two Nivkhs have haplogroup O2a1b1a2a – F238 (see the Table). It is represented among residents of China and one person from Myanmar. The age of its common ancestor is 7,500 years (CI: 8,600–6,400). Three more Nivkhs belong to the rare line Q1a1a1 – M120, F746, Y34108, Y34449, to which one Koryak, an Evenk from Yakutia and four Yukaghirs belong.

Conclusion

The spread of C2a1 carriers undoubtedly occurred with the assimilation of the more ancient local population. Thus, the Nivkh gene pool is quite specific in the composition of Y-chromosome and mtDNA haplogroups, but very similar in autosomal markers. The results of the analysis of the samples indicate a close genetic relationship of the Nivkhs with the Koryaks, Chukchi, Udege and Evenks. The specificity of the Y-chromosome sublines and YSTR haplotypes proves that the Nivkhs had no contact with other ethnic groups for a long time and lived in relative isolation for many centuries. The results of microarray analysis also confirm this. Data on the Nivkh gene pool complement the results of paleogenetic, linguistic, anthropological and ethnological research areas. According to ethnogenesis, the Nivkhs are Paleo-Asians. It was on their genetic substrate that other Amur peoples were later formed, which is in good agreement with the results of this study of their gene pools.

References

Alexander D.H., Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform*. 2011;12:246. DOI 10.1186/1471-2105-12-246

Alexander D.H., Novembre J., Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. DOI 10.1101/gr.094052

Bandelt H.J., Forster P., Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 1999;16:37-48. DOI 10.1093/oxfordjournals.molbev.a026036

Bermisheva M.A., Kutuev I.A., Spitsyn V.A., Villems R., Batyrova A.Z., Korshunova T.Yu., Khusnutdinova E.K. Analysis of mitochondrial DNA variation in the population of Oroks. *Russ. J. Genet.* 2005;41(1):66-71. DOI 10.1007/pl00022112

Browning B.L., Browning S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013;194(2):459-471. DOI 10.1534/genetics.113.150029

Browning S.R., Browning B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 2007;81(5):1084-1097. DOI 10.1086/521987

Derenko M.V., Malyarchuk B.A. Molecular Phylogeography of the Population of Northern Eurasia Based on Data on Mitochondrial DNA Variability. Magadan, 2010 (in Russian)

Fortescue M.D. The relationship of Nivkh to Chukotko-Kamchatkan revisited. *Lingua.* 2011;121:1359-1376. DOI 10.1016/j.lingua.2011.03.001

Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc*. 2014;9:2643-2662. DOI 10.1038/nprot.2014.174

Gusev A., Palamara P.F., Aponte G., Zhuang Z., Darvasi A., Gregersen P., Pe'er I. The architecture of long-range haplotypes shared within and across populations. *Mol. Biol. Evol*. 2012;29(2):473-486. DOI 10.1093/molbev/msr133

Horai S., Murayama K., Hayasaka K., Matsubayashi S., Hattori Y., Fucharoen G., Harihara S., Park K.S., Omoto K., Pan I.H. mtDNA polymorphism in East Asian populations, with special reference to the peopling of Japan. *Am. J. Hum. Genet.* 1996;59(3):579-590

Karmin M., Saag L., Vicente M., Wilson Sayres M.A., Järve M., Talas U.G., Rootsi S., Ilumäe A.M., Mägi R., Mitt M., … Tyler-Smith K., Underhill P.A., Willerslev E., Nielsen R., Metspalu M., Villems R., Kivisild T. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* 2015;25(4):459-466. DOI 10.1101/gr.186684.114

Liu B.L., Ma P.C. Wang C.Z., Yan S., Yao H.B., Li H.L., Xie Y.M., Meng S.L., Sun J., Cai J.H., Sarengaowa S., Li H., Cheng H.Z., Wei L.H. Paternal origin of Tungusic-speaking populations: insights from the updated phylogenetic tree of Y-chromosome haplogroup C2a-M86. *Am. J. Hum. Biol.* 2021;33(2):e23462. DOI 10.1002/ajhb.23462

Peoples of North-East Siberia. Moscow: Nauka Publ., 2010 (in Russian)

Schurr T., Sukernik R., Starikovskaya Y., Wallace D. Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea–Bering Sea region during the Neolithic. *Am. J. Phys. Anthropol.* 1999;108:1-39. DOI 10.1002/(SICI)1096-8644(199901)108:1<1::AID-AJPA1>3.0.CO;2-1

Skotte L., Korneliussen T., Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013;195(3):693-702. DOI 10.1534/genetics.113.154138

Starikovskaya E.B., Sukernik R.I., Derbeneva O.A., Volodko N.V., Ruiz-Pesini E., Torroni A., Brown M.D., Lott M.T., Hosseini S.H., Huoponen K., Wallace D.C. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann. Hum. Genet.* 2005;69:67-89. DOI 10.1046/j.1529-8817.2003.00127.x

Sulyandziga R.V., Kudryashova D.A., Sulyandziga P.V. Indigenous Peoples of the North, Siberia, and the Far East of the Russian Federation. Review of the current situation. Moscow, 2003 (in Russian)]

The Peoples of Russia: Encyclopedia. Moscow: Bolshaya Rossiyskaya Entsiklopediya Publ., 1994 (in Russian)

Valikhova L.V., Kharkov V.N., Zarubin A.A., Kolesnikov N.A., Svarovskaya M.G., Khitrinskaya I.Yu., Shtygasheva O.V., Volkov V.G., Stepanov V.A. Genetic interrelation of the Chulym Turks with Khakass and Kets according to autosomal SNP data and Y-chromosome haplogroups. *Russ. J. Genet.* 2022;58(10):1228-1234. DOI 10.1134/S1022795422100118

Wang C.C., Yeh H.Y., Popov A.N., Zhang H.Q., Matsumura H., Sirak K., Cheronet O., Kovalev A., Rohland N., Kim A.M., … Schiffels S., Kennett D.J., Jin L., Li H., Krause J., Pinhasi R., Reich D. The genomic formation of human populations in East Asia. *Nature.* 2021;591(7850):413-419. DOI 10.1038/s41586-021-03336-2

Zhivotovsky L.A., Underhill P.A., Cinnioğlu C., Kayser M., Morar B., Kivisild T., Scozzari R., Cruciani F., Destro-Bisol G., Spedini G., Chambers G.K., Herrera R.J., Yong K.K., Gresham D., Tournev I., Feldman M.W., Kalaydjieva L. On the effective mutation rate at Y-chromosome STRs with application to human population divergence time. *Am. J. Hum. Genet.* 2004;74(1):50-61. DOI 10.1086/380911

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

Received April 3, 2024. Revised June 21, 2024. Accepted June 22, 2024.

Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2024;28(6):667-674

DOI 10.18699/vjgb-24-74

Polymorphic variants of the dopamine receptor gene *DRD2* (rs6277, rs1800497) in adolescents with problematic video game use

S.Yu. Tereshchenko, K.V. Afonicheva, I.V. Marchenko, M.V. Shubina, M.V. Smolnikova

Scientific Research Institute of Medical Problems of the North – a separate division of the Federal Research Center "Krasnoyarsk Science Center" of the Siberian Branch of the Russian Academy of Sciences, Krasnoyarsk, Russia
smarinv@yandex.ru

Abstract. Problematic video games use, as a specific form of problematic Internet use, is widespread among adolescents and can have negative effects on their mental and somatic well-being. An increasing incidence of addictive video gaming, as well as the overuse of the Internet, among the young population makes the current study of susceptibility factors, including the genetic component, relevant. There has been a number of investigations related to the involvement of gene variants of the neurotransmitter system in the development of Internet addiction, with the results being different for various ethnic groups. The dopamine type 2 receptor gene (*DRD2*) is one of the candidate genes for susceptibility to video game addiction. The aim of the work was to study polymorphic variants of the dopamine receptor gene *DRD2* (rs6277, rs1800497) in Russian adolescents with problematic use of computer video games. A sampling of 407 adolescents aged 14.1 ± 1.8 years was tested, of which 56 (13.8 %) were identified as having problems with the pathological use of video games use based on the GASA scale results. Boys in the sample proved to be addicted to video games more than girls ($p = 0.041$). As a result of comparing the allele frequency of *DRD2* (rs6277), a tendency to a higher frequency of the minor allele T was revealed in the group of adolescents with problematic video game use compared with adolescents without problematic video game use (i.e. 0.563 and 0.466, respectively, $p = 0.06$). When using the dominant inheritance model, it was revealed that adolescents with problematic use of video games were statistically significantly more likely to carry the T (CT+TT) allele ($p = 0.04$, OR = 2.14, CI = 1.01–4.53). The T allele *DRD2* (rs6277) is associated with low expression of the dopamine receptor D2 and leads to decreasing the density and affinity of extrastriatal dopamine type 2 receptors, which is associated with impaired social communication as well. We suggest that the presence of CT and TT genotypes of rs6277 *DRD2* may be a potential risk factor for developing problematic video game use in adolescents.

Key words: gene polymorphism; dopamine; teenagers; problematic video game use; game addiction; Internet addiction.

For citation: Tereshchenko S.Yu., Afonicheva K.V., Marchenko I.V., Shubina M.V., Smolnikova M.V. Polymorphic variants of the dopamine receptor gene *DRD2* (rs6277, rs1800497) in adolescents with problematic video game use. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(6):667-674. DOI 10.18699/vjgb-24-74

Funding. The study was carried out within the framework of the State Assignment No. 124020100064-6 "Psychosomatic disorders in adolescents of Central Siberia: prevalence, structure, psychological risk factors and neurogenetic predictors".

Полиморфные варианты гена рецептора дофамина *DRD2* (rs6277, rs1800497) у подростков с проблемным использованием компьютерных видеоигр

С.Ю. Терещенко, К.В. Афоничева, И.В. Марченко, М.В. Шубина, М.В. Смольникова

Научно-исследовательский институт медицинских проблем Севера – обособленное подразделение Федерального исследовательского центра «Красноярский научный центр Сибирского отделения Российской академии наук», Красноярск, Россия
smarinv@yandex.ru

Аннотация. Проблемное использование видеоигр как специфическая форма проблемного использования Интернета широко распространено среди подростков и может оказывать негативный эффект на их психическое и соматическое благополучие. Рост зависимости от пользования видеоиграми, как и Интернетом, среди молодого населения делает актуальным изучение факторов подверженности к ним, в том числе генетической составляющей. Существует ряд исследований, посвященных изучению вовлеченности полиморфных вариантов генов системы нейромедиаторов в развитие Интернет-зависимости, результаты которых различаются в разных

этнических группах. Ген рецептора дофамина второго типа *DRD2* является одним из кандидатных генов подверженности к патологической зависимости от использования видеоигр. Целью работы было исследование полиморфных вариантов гена рецептора дофамина *DRD2* (rs6277, rs1800497) у русских подростков с проблемным использованием компьютерных видеоигр. Протестирована выборка из 407 подростков в возрасте 14.1 ± 1.8 года, у 56 (13.8 %) из которых на основании результатов оценки шкалы GASA было выявлено проблемное использование видеоигр. Мальчики в выборке чаще были зависимы от видеоигр, чем девочки (p = 0.041). В результате сравнения частоты аллелей *DRD2* rs6277 обнаружена тенденция к большей частоте минорного аллеля T в группе подростков с проблемным использованием видеоигр по сравнению с подростками без проблемного использования видеоигр (0.563 и 0.466 соответственно, p = 0.06). В доминантной модели наследования у подростков с проблемным использованием видеоигр статистически значимо чаще встречалось носительство аллеля T (CT+TT) (p = 0.04, OR 2.14, CI = 1.01–4.53). Носительство аллеля T *DRD2* rs6277 ассоциировано с низкой экспрессией дофаминового рецептора D2 и приводит к снижению плотности и аффинности экстрастриарных дофаминовых рецепторов второго типа, что сопряжено в том числе с нарушением социальной коммуникации. Мы полагаем, что наличие генотипов CT и TT rs6277 гена *DRD2* может выступать потенциальным фактором риска развития проблемного использования видеоигр у подростков.

Ключевые слова: полиморфизм генов; дофамин; подростки; проблемное использование компьютерных видеоигр; игровая зависимость; интернет-зависимость.

Introduction

Problematic video game use among adolescents is a pressing challenge in modern society and is characterized by excessive passion for video games, leading to negative consequences in various areas of life: social, educational, somatic and psychological (Griffiths et al., 2012; Paulus et al., 2018; Männikkö et al., 2020).

Since video games are currently associated with high Internet use in the vast majority of cases, the problematic use of video games is considered by most experts to be a specific problematic use of the Internet, or Internet addiction. Several synonymous terms can be found in the literature available, essentially describing a single psychological construct: game addiction (ICD-11), Internet gaming disorder (DSM-5), gaming disorder, pathological video gaming, excessive video game use, compulsive gaming, problematic digital gaming, problematic online gaming, problematic video game use (PVGU). These are the terms that are often used interchangeably in scientific publications, however, there may have some semantic aspects depending on the context and the theoretical background of the study. The European Research Group recommends using the term "Problematic Use of the Internet" for generalized Internet addiction and its particular types, i. e. "Problematic Social Media Use" and PVGU (Fineberg et al., 2022). Only one of the many specific types of addictive Internet behavior, namely PVGU, is currently considered to be a mental disorder (Internet Gaming Disorder, DSM-5; American Psychiatric Association, 2013; Gaming Disorder, ICD-11, 2019).

As shown in a systematic review by S. Mihara and S. Higuchi (Mihara, Higuchi, 2017), the prevalence of PVGU varies from 0.7 to 27.5 % and, like generalized Internet addiction, is highly dependent on the questionnaires used and addiction assessment criteria. As with generalized Internet addiction, the prevalence of PVGU shows higher prevalence values in Asian countries with predominantly Mongoloid population compared to other regions (Sussman et al., 2018).

Very few studies have been devoted to finding the genetic basis of Internet addiction as opposed to other types of addictions (e. g. substance abuse or gambling). For example, the first twin study based on an examination of 825 children aged 10–12 years in the Chinese population was conducted in 2014, with the authors being able to estimate the proportion of total variability due to genetic effects, which varied from 58 to 66 % depending on gender (Li M. et al., 2014). Similar results were obtained a little later in the study of Turkish (19–86 %) (Deryakulu, Ursavaş, 2014), Dutch (48 %) (Vink et al., 2016), Australian (41 %) (Long et al., 2016) and German (21–44 %) (Hahn et al., 2017) twin cohorts. Although these data are limited by the sample size and different ethno-geographic conditions, there is likely to be a tendency towards a greater contribution of genetic factors in males. Thus, the presence of a genetic component in developing Internet addiction has been convincingly demonstrated by twin studies using various populations as an example, however, to date, specific genes involved in the mechanisms of such heritability have not been precisely identified.

Therefore, candidate genes are in active study, their polymorphic variants can disrupt the functioning of neurotransmitter systems and cause mental and behavioral disorders. One of them is the dopamine receptor gene *DRD2* (Kim et al., 2022). Dopamine is a hormone responsible for motivation, desire and addiction, functionally associated with the "pleasure centers". Dopaminergic brain neurons form the nigrostriatal, mesolimbic, mesocortical, tuberoinfundibular pathways (Kolotilova et al., 2014). The D2 receptor, classified as inhibitory, is present in high concentrations in the striatum, olfactory tubercle, amygdala, nucleus accumbens, hypothalamus, substantia nigra and ventral tegmental area (Ford, 2014; Arnsten et al., 2015). The human dopamine receptor gene *DRD2* is located on chromosome 11 (q22-q23) and is polymorphic, with different genetic variants altering the availability and expression of the dopamine D2 receptor gene, which affects receptor sensitivity and density (Magistrelli et al., 2021). The rs6277 polymorphism in exon 7 of the *DRD2* gene is a substitution of the amino acid serine for cysteine (Ser311Cys). The homozygous CC genotype of rs6277 *DRD2* causes low sensitivity to dopamine in the striatum (Hänninen et al., 2006). However, outside the striatum (extrastriate area), this genotype has a high affinity to dopamine D2 receptors (Liu et al., 2014; Smith et al., 2017; Della Torre et al., 2018). The dopamine binding potential by D2 receptors in the striatum is higher in carriers of the TT genotype of rs6277 *DRD2*, while the opposite effect is observed in the extrastriate area (Hänninen et al., 2006).

A decrease in DRD2 density in the striatum and environmental effect are known to result in the development of addictions, including alcohol, drugs, computer games (Hill et al., 2008; Bhaskar, Kumar, 2014; Gao et al., 2017; Anokhin et al., 2019; Picci et al., 2022). However, according to the published data, it is debatable which allele (C or T) of rs6277 *DRD2* is associated with addiction to psychoactive substances (Hill et al., 2013). The T allele of rs6277 *DRD2* is shown in some studies to be associated with an increased tendency to pathological addiction to video games (Kim et al., 2022).

However, it is worth noting that genetic factors represent only one aspect of the tendency to addictive behavior, and the influence of the environment and sociocultural factors also play an essential role. Thus, it is known that a stressful environment combined with the T allele of rs6277 *DRD2* causes a decrease in the ability to control craving for computer games (Kim et al., 2022). Individuals with the homozygous TT genotype of rs6277 *DRD2* have been shown to respond better to nicotine replacement therapy than carriers of the C allele (Hill et al., 2008). The C allele variant of rs6277 *DRD2* causes a hypodopaminergic state manifesting as a reduced ability to suppress responses to reward-related stimuli (Machulska et al., 2016; Richter et al., 2017; Rył et al., 2024). Carriers of the homozygous CC genotype of rs6722 *DRD2*, who were abused or experienced traumatic life events in childhood, have been demonstrated to have a high degree of impulsivity and more frequent alcohol consumption in adulthood (Klaus et al., 2021). It is reported that the risk of developing such addiction is higher in adult C allele carriers of rs6277 *DRD2*, whereas in adolescents (11–13 years old), this allelic variant may be protective against the development of dependence on psychoactive substances, as well as predispose to a later onset of alcohol consumption (Picci et al., 2022).

The rs1800497 polymorphism of the *DRD2* gene causes an amino acid substitution of glycine for lysine (Glu713Lys), which leads to a specificity change of dopamine receptor binding. According to some data, this polymorphism is called *DRD2/ANKK1 Taq1A* because it is located within the protein kinase *PKK2* gene (Ankyrin Repeat and Kinase Domain Containing 1 – ANKK1), a protein of the post-receptor intracellular signal transmission system (Gafarov et al., 2019). The rs1800497 *DRD2* polymorphism is also frequently studied in the context of neuropsychiatric disorders and addictions (Volkow et al., 1996; Pohjalainen et al., 1998). The A1 allele (T) carriers were found to have a 30 % decrease in the density of dopamine D2 receptors in the brain striatum, resulting in poor attention and learning ability, an increase in anxiety, and an association with "reward deficiency" and "novelty seeking" syndrome (Klein et al., 2007; Kushnarev, 2022). The presence of the minor T allele of rs1800497 was similarly shown in the work (Pohjalainen et al., 1998) to be associated with a reduced number of dopamine binding sites in the brain. It has been suggested that there is an association between the A1/A1 (TT) and A1/A2 (TC) genotypes of rs1800497 of the *DRD2* gene with "reward deficiency" syndrome (Klein et al., 2007). "Reward deficiency" syndrome causes various mental and behavioral disorders, i. e. nicotine and drug addiction, gambling addiction, ADHD, autism spectrum disorders, eating disorders with compulsive overeating (Pohjalainen et al., 1998). It was found that male carriers of the allelic T variant of rs1800497 are more likely to suffer from addiction to online games (Paik et al., 2017). This allelic variant is also more common in people addicted to playing video games to satisfy their seeking of reward (Werling, Grünblatt, 2022). Thus, people with a low number of dopamine D2 receptors tend to search for extreme ways to enjoy life. Impaired sensitivity of dopamine receptors causes a decrease in people's ability to draw the right conclusions from negative experiences, since dopamine is involved in learning processes and provides the opportunity to effectively learn from mistakes.

The aim of this study was to investigate polymorphic variants of the dopamine receptor gene *DRD2* (rs6277, rs1800497) in adolescents with problematic video game use for possible associations between genetic variants and behavioral aspects of gaming addiction to be identified.

Material and methods

In the present study, psychological and genetic testing of 407 adolescents aged 12–18 years was carried out. All adolescents involved in the study were Russians (verified by both mother and father nationality). Informed consent was obtained from the adolescents or their parents (legal representatives), followed by notification of the voluntary and confidential nature of the study. The study participants were asked to fill out a demographic data questionnaire (gender, age, nationality of mother and father), and a translated version of the Game Addiction Scale for Adolescents (GASA) questionnaire (Lemmens et al., 2009). The GASA questionnaire includes seven questions concerning behavioral disorders in adolescents caused by overuse of Internet games. Each question is assessed on a 5-point scale: "never" (0 points), "rarely" (1 point), "sometimes" (2 points), "often" (3 points), "very often" (4 points). According to the criteria proposed by the authors of the questionnaire (Lemmens et al., 2009), having PVGU was determined (if the teenager answered any four or more of seven questions – "sometimes", "often" or "very often").

After completing the questionnaire, adolescents were asked to provide saliva samples in special containers. Saliva samples were collected using the "Saliva DNA Collection and Preservation Devices" (Cat. No. RU 49080, Norgen Biotek Corp., Canada). DNA was isolated from saliva samples using the DIAtom DNA Prep kit (Isogene Lab, Russia). Genotyping of polymorphic variants rs6277 and rs1800497 *DRD2* was performed using TaqMan technology with probes and primers (DNA Synthesis, Russia) and a reaction mixture (Syntol, Russia) on a Rotor-Gene 6000 device (Qiagen, Germany). The study was approved by the Ethics Committee of FRC KSC SB RAS (Protocol No. 12 dated 12.18.2018).

Statistical analysis was performed using Statistica v.10 software (StatSoft Inc., USA). Differences in categorical data were evaluated using Pearson's χ^2 test with Yates's correction, and the differences in quantitative data were evaluated with Student's *t*-test.

Results

Descriptive statistics of the main variables are presented in Table 1. The mean age of the 407 tested adolescents was 14.1 ± 1.8 years, the ratio of boys/girls = 174 (42.8 %)/233 (57.2 %). PVGU was detected in 56 adolescents (13.8 %) based on the GASA scale assessment results (Table 1). Boys

Table 1. Descriptive statistics of main variables

Parameter	Total	Boys	Girls	*p* (boys–girls)
Age 12–14	241	95 (39.4 %)	146 (60.6 %)	–
Age 15–18	166	79 (47.6 %)	87 (52.4 %)	–
Total number	407	174 (42.8 %)	233 (57.2 %)	–
GASA result (n = 407)				
Gambling addiction scale for adolescents (GASA), score	10.8 ± 6.8	12.2 ± 6.7	9.8 ± 6.9	0.0005 t = 3.5
Problematic Video Game Use (PVGU)	56 (13.8 %)	31 (17.8 %)	25 (10.7 %)	p = 0.041 χ^2 = 4.21, df = 1

Note. Data are presented as n (%) and mean ± standard deviation.

Table 2. Distribution of genotype and allele frequencies of rs6277 for the *DRD2* gene in adolescents with and without PVGU

Genotypes and alleles of rs6277	Without PVGU	With PVGU	χ^2	p	OR	95 % CI
	n = 351	n = 56				
Genotype CC	0.291 (102)	0.161 (9)	4.26	0.12	0.47	0.22–0.99
Genotype CT	0.487 (171)	0.554 (31)			1.31	0.74–2.30
Genotype TT	0.222 (78)	0.285 (16)			1.40	0.74–2.30
Allele C	0.534	0.437	3.62	0.06	0.68	0.45–1.01
Allele T	0.466	0.563			1.47	0.99–2.20

Table 3. Distribution of genotype frequencies of rs1800497 for the *DRD2* gene in adolescents with and without PVGU

Genotypes and alleles of rs1800497	Without PVGU	With PVGU	χ^2	p	OR	95 % CI
	n = 351	n = 56				
Genotype CC	0.638 (224)	0.714 (40)	1.24	0.54	1.42	0.76–2.63
Genotype CT	0.342 (120)	0.268 (15)			0.70	0.37–1.32
Genotype TT	0.020 (7)	0.018 (1)			0.89	0.11–7.40
Allele C	0.809	0.848	0.98	0.32	1.32	0.76–2.28
Allele T	0.191	0.152			0.76	0.44–1.31

had significantly higher mean scores on the game addiction scale than girls. In addition, more PVGU adolescents were detected among boys compared to girls.

The genotype distribution frequency of the polymorphic variants of the *DRD2* gene (rs6277 and rs1800497) in the adolescents studied corresponds to their distribution in Caucasian populations (according to the website ensembl.org). The distribution of genotype frequencies was consistent with the Hardy–Weinberg equilibrium, both for PVGU cases and for the group without PVGU. Thus, the allele frequencies of the selected polymorphic variants in the study population were balanced and, therefore, applicable to association studies.

The distribution of genotype and allele frequencies of the rs6277 and rs1800497 polymorphisms of the *DRD2* gene depending on the PVGU presence and absence is given in Tables 2 and 3, respectively. The genotype frequency distribution of polymorphic variants of rs6277 (*DRD2*) did not differ significantly between the PVGU group and in the group without PVGU (p = 0.12) (Table 2). At the same time, when comparing the frequencies of alleles of *DRD2* rs6277, a clear trend towards a higher frequency of the minor T allele in the group of adolescents with PVGU was found compared to the group without PVGU (p = 0.06). Analysis of polymorphic variants of rs1800497 of the DRD2 gene showed no significant differences in the frequencies of genotypes and alleles between the groups with and without PVGU (Table 3).

Then, we analyzed the distribution of genotype frequencies of the rs6277 polymorphism of the *DRD2* gene using

Table 4. Distribution of genotype frequencies of the rs6277 polymorphism of the *DRD2* gene in adolescents with and without PVGU

Genotypes	Without PVGU	With PVGU	χ^2	*p*	OR	95 % CI
	n = 351	*n* = 56				
CC	0.291	0.161	4.11	0.04	0.47	0.22–0.99
CT+TT	0.709	0.839			2.14	1.01–4.53

the dominant model of inheritance, where heterozygotes and homozygotes for the minor allele of rs6277 of the *DRD2* gene (CT and TT, respectively) were combined (Table 4).

According to the obtained results, in the group of adolescents with PVGU, carriage of the T allele (genotype CT+TT) was statistically significantly more common compared to adolescents without PVGU. Calculation of the odds ratio (OR) demonstrated a significant association between carriage of the T allele and the presence of PVGU in adolescents.

Discussion

The overall PVGU frequency in the studied sample of Russian adolescents was 13.8 %, which is not significantly different from our previously obtained data on the prevalence of computer game addiction, resulting from a large-scale epidemiological project (*n* = 4,514, PVGU prevalence – 10.4 %) (Tereshchenko et al., 2022). Boys in the sample of the present study were more often addicted to video games than girls ($p = 0.041$), which is consistent with the data of the mentioned project and the results of other epidemiological studies using the GASA questionnaire (Mihara, Higuchi, 2017; Tereshchenko et al., 2022). The genotypes and alleles distribution in the studied sample is similar to their frequency in the global population of European descent according to the 1000 Genomes Project and HapMap databases (website: ensemble.org), both for rs6277 and rs1800497. Thus, in terms of the prevalence of the main variables, the population studied is typical enough, and the findings can be successfully extrapolated to other adolescent of European populations.

We have found that the CT and TT genotype carriers of the rs6277 polymorphism of the *DRD2* gene, that is, the T allele carriers, according to the results obtained using the dominant model of inheritance, exhibit signs of PVGU significantly more often than adolescents with the CC genotype.

The T allele carriers of rs6277 of the *DRD2* gene are known to have a lower density and affinity for dopamine D2 receptors in all brain regions (including the prefrontal cortex), excluding the striatum, compared to carriers of the C allele – C/C > C/T > T/T (Hirvonen et al., 2009; Smith et al., 2017). Low DRD2 density in extrastriatal brain region can lead to certain psychophysiological consequences. In particular, the functional effects of the availability of these receptors in extrastriatal regions, including the cortex and thalamus, have been considered in the study devoted to the role of extrastriatal DRD2 (Takahashi et al., 2006). The review includes postmortem examinations as well as *in vivo* studies in humans and animals, considering the role of low functional activity of extrastriatal DRD2 for schizophrenia (Takahashi et al., 2006). Low availability of D2/3 receptors in extrastriatal regions in adult males with socio-communicative deficits in autism has been indicated by C. Murayama et al. to be associated with reduced dopamine receptor density (Murayama et al., 2022).

The T allele carriers of rs6277 of the *DRD2* gene were shown to be less active in suppressing impulsive tendencies to undesirable actions than the C allele carriers (Colzato et al., 2010). In the study by O.H. Della Torre et al., it was found that the T allele carriers of rs6277 of the *DRD2* gene (6–18 years old) were characterized by problems with impulse control, self-control of emotions and volitional personality change (Della Torre et al., 2018). As a theoretical model confirming the genetic data, the authors of the study cite the opinion of G.S. Dichter et al. that a decrease of the dopaminergic activity is associated with learning problems and a lack of self-discipline (Dichter et al., 2012).

Our data on the association between the T allele carriage of rs6277 of the *DRD2* gene and PVGU in adolescents correspond with the study results of E. Kim et al. directly relating the PVGU severity and the T allele carriage in college students ($b = 19.58$, $p = 0.04$) using regression analysis (Kim et al., 2022). Two other studies conducted on samples of adults didn't show such an association (Paik et al., 2017; Rył et al., 2024). The inconsistency of the results obtained can be explained by the differences in age, gender, ethnicity, and number of sampling size of the aforementioned studies. In particular, the influence of the genetic component on addictions may be manifested differently in adolescents and adults. Adolescence is characterized by different time trajectories in developing the limbic system and prefrontal cortex (Casey et al., 2008). Delayed development of the prefrontal cortex compared to the limbic system during adolescence results in weakened cortical inhibition on underlying subcortical structures and increased impulsivity, which contributes to a high risk of developing addictive behavior (He, Crews, 2007).

We believe that the association between the T allele carriage of rs6277 *DRD2* and PVGU in adolescents and students, which Kim et al. (Kim et al., 2022) and our research team have found, provides the theoretical and empirical background. Carrying the T allele of rs6277 leads to a decrease in the density and affinity of extrastriatal dopamine D2 receptors (Hirvonen et al., 2009; Smith et al., 2017) and a peculiar phenomenon of "dopamine desensitization", which is associated with a reduced sensitivity to reward, increased impulsivity, lack of self-discipline (Colzato et al., 2010; Della Torre et al., 2018; Weinstein, Lejoyeux, 2020; Kim et al., 2022), as well as a possible impairment of social communication (Takahashi et al., 2006; Murayama et al., 2022).

Hyporeactivity of the orbitofrontal cortex and decreased dopaminergic function in this brain region are associated with hyposensitivity of the reward system, promoting transgressive behavior, delinquency, and substance abuse (Matthys et al.,

2013). Certain *DRD2* variants were suggested to possibly contribute to the development of a hypodopaminergic state, with partial availability of dopamine receptors determining reduced sensitivity to reward (Alcaro et al., 2021). The latter may lead to the adolescent aiming to receive additional stimulation of the dopaminergic system, which manifests as addictive behavior including an active persistent reward component, such as compulsive use of video games (Weinstein, Lejoyeux, 2020; Kim et al., 2022) or gambling.

Increased impulsivity and impairment of social connection turn out to be the most important predictors of the development of generalized Internet addiction and its specific form – i. e. PVGU. Impulsivity and self-control are associated with a wide range of behavioral characteristics. Empirical studies have shown that people with high self-control are better at controlling their thoughts, regulating their emotions and suppressing their impulses than individuals with low self-control (de Ridder et al., 2012). Low self-control and high impulsivity are closely related to delinquency, crime, antisocial behavior, externalizing behavior, victimization and addictive disorders. One of the psychiatric disorders most associated with Internet addiction has been known to be Attention Deficient Hyperactivity Disorder, characterized by high behavioral impulsivity (Wang et al., 2017). A large number of psychological studies have shown that Internet-addicted behavior is closely associated with low self-control/high impulsivity (Li W. et al., 2016; Li S. et al., 2021; Yu et al., 2021). A meta-analysis of 40 neurophysiological studies of problematic Internet use have shown that, regardless of content, Internet-addicted behavior is characterized by significant impairment in inhibitory control, decision-making and working memory (Ioannidis et al., 2019). A meta-analysis by M. Zhang et al. has demonstrated a common pattern of structural brain changes in chemical and behavioral addictions, i. e. changes in the prefrontal and insular cortex, associated with increased impulsivity (Zhang et al., 2021).

A rare T allele of the rs1800497 *DRD2* polymorphism is also associated with low expression of the dopamine D2 receptor gene in the prefrontal cortex and has been found by S.-H. Paik et al. to be more common among Korean men (19–47 years old) with Internet gaming addiction (Paik et al., 2017). This variant is also more common in Korean young adults (high school students and college students) with PVGU and high reward dependence (Han et al., 2007). However, our study results did not provide any statistically significant differences between different genotypes and alleles of the rs1800497 polymorphism of the *DRD2* gene in groups with and without signs of PVGU. The inconsistency in the analysis results may be due to the ethnic and gender characteristics of the samples, as well as the use of different psychometric tools to verify PVGU. In particular, there are pronounced ethnic differences in the genotype and allele frequencies of the rs1800497 *DRD2* polymorphism in representatives of Caucasian and Mongoloid populations that may have a great impact.

Conclusion

The research results of polymorphic variants of the dopamine receptor gene *DRD2* in adolescents with PVGU allow one to conclude that genetic factors are important for developing this behavioral disorder. The availability of CT and TT genotypes for the polymorphic locus rs6277 of the *DRD2* gene may be a potential risk prediction of developing PGVU in adolescents. Further study of the genetic basis of behavioral disorders will provide personalized approach to the prevention and treatment of game addiction, taking into account the patient's genetic profile.

References

Alcaro A., Brennan A., Conversi D. The SEEKING drive and its fixation: a neuro-psycho-evolutionary approach to the pathology of addiction. *Front. Hum. Neurosci*. 2021;15:635932. DOI 10.3389/fnhum.2021.635932

Anokhin P.K., Veretinskaya A.G., Davidova T.V., Shamakina I.Yu. Dopamine D2 agonists in the treatment of experimental alcoholism. *Patologicheskaya Fiziologiya i Eksperimental'naya Terapiya = Pathological Physiology and Experimental Therapy*. 2019;63(1):33-39. DOI 10.25557/0031-2991.2019.01.33-39 (in Russian)

Arnsten A.F.T., Wang M., Paspalas C.D. Dopamine's actions in primate prefrontal cortex: challenges for treating cognitive disorders. *Pharmacol. Rev*. 2015;67(3):681-696. DOI 10.1124/pr.115.010512

Bhaskar L.V.K.S., Kumar S.A. Polymorphisms in genes encoding dopamine signalling pathway and risk of alcohol dependence: a systematic review. *Acta Neuropsychiatrica*. 2014;26(2):69-80. DOI 10.1017/neu.2013.27

Casey B.J., Jones R.M., Hare T.A. The adolescent brain. *Ann. N.Y. Acad. Sci.* 2008;1124:111-126. DOI 10.1196/annals.1440.010

Colzato L.S., van den Wildenberg W.P.M., Van der Does A.J.W., Hommel B. Genetic markers of striatal dopamine predict individual differences in dysfunctional, but not functional impulsivity. *Neuroscience*. 2010;170(3):782-788. DOI 10.1016/j.neuroscience.2010.07.050

de Ridder D.T.D., Lensvelt-Mulders G., Finkenauer C., Stok F.M., Baumeister R.F. Taking stock of self-control: a meta-analysis of how trait self-control relates to a wide range of behaviors. *Pers. Soc. Psychol. Rev*. 2012;16(1):76-99. DOI 10.1177/1088868311418749

Della Torre O.H., Paes L.A., Henriques T.B., de Mello M.P., Celeri E.H.R.V., Dalgalarrondo P., Guerra-Júnior G., Santos-Júnior A.D. Dopamine D2 receptor gene polymorphisms and externalizing behaviors in children and adolescents. *BMC Med. Genet*. 2018;19(1):65. DOI 10.1186/s12881-018-0586-9

Deryakulu D., Ursavaş Ö.F. Genetic and environmental influences on problematic Internet use: a twin study. *Comput. Hum. Behav.* 2014;39:331-338. DOI 10.1016/j.chb.2014.07.038

Dichter G.S., Damiano C.A., Allen J.A. Reward circuitry dysfunction in psychiatric and neurodevelopmental disorders and genetic syndromes: animal models and clinical findings. *J. Neurodev. Disord*. 2012;4(1):19. DOI 10.1186/1866-1955-4-19

Fineberg N.A., Menchón J.M., Hall N., Dell'Osso B., Brand M., Potenza M.N., Chamberlain S.R., Cirnigliaro G., Lochner C., Billieux J., … Cataldo I., Riva G.M., Yücel M., Flayelle M., Hall T., Griffiths M., Zohar J. Advances in problematic usage of the internet research – a narrative review by experts from the European network for problematic usage of the internet. *Compr. Psychiatry*. 2022;118:152346. DOI 10.1016/j.comppsych.2022.152346

Ford C.P. The role of D2-autoreceptors in regulating dopamine neuron activity and transmission. *Neuroscience*. 2014;282:13-22. DOI 10.1016/j.neuroscience.2014.01.025

Gafarov V.V., Gromova E.A., Panov D.O., Maximov V.N., Gagulin I.V., Gafarova A.V. Association of *DRD2/ANKK1 Taq1A* polymorphism with depression in an open 45–64 year-old male population (international epidemiological HAPIEE and WHO MONICA programs). *Nevrologiya, Neiropsikhiatriya, Psikhosomatika = Neurology, Neuropsychiatry, Psychosomatics*. 2019;11(2):37-41. DOI 10.14412/2074-2711-2019-2-37-41

Gao X., Wang Y., Lang M., Yuan L., Reece A.S., Wang W. Contribution of genetic polymorphisms and haplotypes in *DRD2*, *BDNF*, and

opioid receptors to heroin dependence and endophenotypes among the Han Chinese. *OMICS*. 2017;21(7):404-412. DOI 10.1089/omi.2017.0057

Griffiths M.D., Kuss D.J., King D.L. Video game addiction: past, present and future. *Curr. Psychiatry Rev.* 2012;8(4):308-318. DOI 10.2174/157340012803520414

Hahn E., Reuter M., Spinath F.M., Montag C. Internet addiction and its facets: the role of genetics and the relation to self-directedness. *Addict. Behav*. 2017;65:137-146. DOI 10.1016/j.addbeh.2016.10.018

Han D.H., Lee Y.S., Yang K.C., Kim E.Y., Lyoo I.K., Renshaw P.F. Dopamine genes and reward dependence in adolescents with excessive internet video game play. *J. Addict. Med*. 2007;1(3):133-138. DOI 10.1097/ADM.0b013e31811f465f

Hänninen K., Katila H., Kampman O., Anttila S., Illi A., Rontu R., Mattila K.M., Hietala J., Hurme M., Leinonen E., Lehtimäki T. Association between the C957T polymorphism of the dopamine D2 receptor gene and schizophrenia. *Neurosci. Lett*. 2006;407(3):195-198. DOI 10.1016/j.neulet.2006.08.041

He J., Crews F.T. Neurogenesis decreases during brain maturation from adolescence to adulthood. *Pharmacol. Biochem. Behav*. 2007;86(2):327-333. DOI 10.1016/j.pbb.2006.11.003

Hill S.Y., Hoffman E.K., Zezza N., Thalamuthu A., Weeks D.E., Matthews A.G., Mukhopadhyay I. Dopaminergic mutations: within-family association and linkage in multiplex alcohol dependence families. *Am. J. Med. Genet. B Neuropsychiatr. Genet*. 2008;147B(4):517-526. DOI 10.1002/ajmg.b.30630

Hill S.Y., Lichenstein S., Wang S., Carter H., McDermott M. Caudate volume in offspring at ultra high risk for alcohol dependence: COMT Val158Met, DRD2, externalizing disorders, and working memory. *Adv. J. Mol. Imaging*. 2013;3(4):43-54. DOI 10.4236/ami.2013.34007

Hirvonen M.M., Lumme V., Hirvonen J., Pesonen U., Någren K., Vahlberg T., Scheinin H., Hietala J. C957T polymorphism of the human dopamine D2 receptor gene predicts extrastriatal dopamine receptor availability *in vivo*. *Prog. Neuropsychopharmacol. Biol. Psychiatry*. 2009;33(4):630-636. DOI 10.1016/j.pnpbp.2009.02.021

Ioannidis K., Hook R., Goudriaan A.E., Vlies S., Fineberg N.A., Grant J.E., Chamberlain S.R. Cognitive deficits in problematic internet use: meta-analysis of 40 studies. *Br. J. Psychiatry*. 2019;215(5):639-646. DOI 10.1192/bjp.2019.3

Kim E., Lee D., Do K., Kim J. Interaction effects of DRD2 genetic polymorphism and interpersonal stress on problematic gaming in college students. *Genes* (*Basel*). 2022;13(3):449. DOI 10.3390/genes13030449

Klaus K., Vaht M., Pennington K., Harro J. Interactive effects of *DRD2* rs6277 polymorphism, environment and sex on impulsivity in a population-representative study. *Behav. Brain Res*. 2021;403:113131. DOI 10.1016/j.bbr.2021.113131

Klein T.A., Neumann J., Reuter M., Hennig J., von Cramon D.Y., Ullsperger M. Genetically determined differences in learning from errors. *Science*. 2007;318(5856):1642-1645. DOI 10.1126/science.1145044

Kolotilova O.I., Koreniuk I.I., Khusainov D.R., Cheretaev I.V. Dopaminergic brain system. *Vestnik Bryanskogo Gosudarstvennogo Universiteta = The Bryansk State University Herald*. 2014;4:97-106 (in Russian)

Kushnarev A.P. The study of polymorphism of the dopamine receptor type 4 (DRD4) and dopamine transporter (DAT) genes in individuals with antisocial behavior and representatives of extreme professions. *Meditsina. Sotsiologiya. Filosofiya. Prikladnyye Issledovaniya = Medicine. Sociology. Philosophy. Applied Research.* 2022;3:40-47 (in Russian)

Lemmens J.S., Valkenburg P.M., Peter J. Development and validation of a game addiction scale for adolescents. *Media Psychology*. 2009;12(1):77-95. DOI 10.1080/15213260802669458

Li M., Chen J., Li N., Li X. A twin study of problematic internet use: its heritability and genetic association with effortful control. *Twin Res. Hum. Genet*. 2014;17(4):279-287. DOI 10.1017/thg.2014.32

Li S., Ren P., Chiu M.M., Wang C., Lei H. The relationship between self-control and internet addiction among students: a meta-analysis. *Front. Psychol*. 2021;12:735755. DOI 10.3389/fpsyg.2021.735755

Li W., Zhang W., Xiao L., Nie J. The association of Internet addiction symptoms with impulsiveness, loneliness, novelty seeking and behavioral inhibition system among adults with attention-deficit/hyperactivity disorder (ADHD). *Psychiatry Res*. 2016;243:357-364. DOI 10.1016/j.psychres.2016.02.020

Liu L., Fan D., Ding N., Hu Y., Cai G., Wang L., Xin L., Xia Q., Li X., Xu S., Xu J., Yang X., Zou Y., Pan F. The relationship between DRD2 gene polymorphisms (C957T and C939T) and schizophrenia: a meta-analysis. *Neurosci. Lett*. 2014;583:43-48. DOI 10.1016/j.neulet.2014.09.024

Long E.C., Verhulst B., Neale M.C., Lind P.A., Hickie I.B., Martin N.G., Gillespie N.A. The genetic and environmental contributions to Internet use and associations with psychopathology: a twin study. *Twin Res. Hum. Genet*. 2016;19(1):1-9. DOI 10.1017/thg.2015.91

Machulska A., Zlomuzica A., Rinck M., Assion H.-J., Margraf J. Approach bias modification in inpatient psychiatric smokers. *J. Psychiatric Res.* 2016;76:44-51. DOI 10.1016/j.jpsychires.2015.11.015

Magistrelli L., Ferrari M., Furgiuele A., Milner A.V., Contaldi E., Comi C., Cosentino M., Marino F. Polymorphisms of dopamine receptor genes and Parkinson's disease: clinical relevance and future perspectives. *Int. J. Mol. Sci*. 2021;22(7):3781. DOI 10.3390/ijms22073781

Männikkö N., Ruotsalainen H., Miettunen J., Pontes H.M., Kääriäinen M. Problematic gaming behaviour and health-related outcomes: a systematic review and meta-analysis. *J. Health. Psychol*. 2020;25(1):67-81. DOI 10.1177/1359105317740414

Matthys W., Vanderschuren L.J.M.J., Schutter D.J.L.G. The neurobiology of oppositional defiant disorder and conduct disorder: altered functioning in three mental domains. *Dev. Psychopathol*. 2013;25(1):193-207. DOI 10.1017/S0954579412000272

Mihara S., Higuchi S. Cross-sectional and longitudinal epidemiological studies of Internet gaming disorder: a systematic review of the literature. *Psychiatry Clin. Neurosci*. 2017;71(7):425-444. DOI 10.1111/pcn.12532

Murayama C., Iwabuchi T., Kato Y., Yokokura M., Harada T., Goto T., Tamayama T., Kameno Y., Wakuda T., Kuwabara H., Senju A., Nishizawa S., Ouchi Y., Yamasue H. Extrastriatal dopamine D2/3 receptor binding, functional connectivity, and autism socio-communicational deficits: a PET and fMRI study. *Mol. Psychiatry*. 2022;27(4):2106-2113. DOI 10.1038/s41380-022-01464-3

Paik S.-H., Choi M.R., Kwak S.M., Bang S.H., Chun J.-W., Kim J.-Y., Choi J., Cho H., Jeong J.-E., Kim D.-J. An association study of *Taq1A ANKK1* and *C957T* and *–141C DRD2* polymorphisms in adults with internet gaming disorder: a pilot study. *Ann. Gen. Psychiatry*. 2017;16:45. DOI 10.1186/s12991-017-0168-9

Paulus F.W., Ohmann S., von Gontard A., Popow C. Internet gaming disorder in children and adolescents: a systematic review. *Dev. Med. Child Neurol*. 2018;60(7):645-659. DOI 10.1111/dmcn.13754

Picci G., Fishbein D.H., VanMeter J.W., Rose E.J. Effects of OPRM1 and DRD2 on brain structure in drug-naïve adolescents: genetic and neural vulnerabilities to substance use. *Psychopharmacology* (*Berl.*). 2022;239(1):141-152. DOI 10.1007/s00213-021-06030-3

Pohjalainen T., Rinne J.O., Någren K., Lehikoinen P., Anttila K., Syvälahti E.K., Hietala J. The A1 allele of the human D2 dopamine receptor gene predicts low D2 receptor availability in healthy volunteers. *Mol. Psychiatry*. 1998;3(3):256-260. DOI 10.1038/sj.mp.4000350

Richter A., Barman A., Wüstenberg T., Soch J., Schanze D., Deibele A., Behnisch G., Assmann A., Klein M., Zenker M., Seidenbecher C., Schott B.H. Behavioral and neural manifestations of reward memory in carriers of low-expressing versus high-expressing genetic variants of the dopamine D2 receptor. *Front. Psychol*. 2017;8:654. DOI 10.3389/fpsyg.2017.00654

Rył A., Tomska N., Jakubowska A., Ogrodniczak A., Palma J., Rotter I. Genetic aspects of problematic and risky Internet use in young men – analysis of ANKK1, DRD2 and NTRK3 gene polymorphism. *Genes* (*Basel*). 2024;15(2):169. DOI 10.3390/genes15020169

Smith C.T., Dang L.C., Buckholtz J.W., Tetreault A.M., Cowan R.L., Kessler R.M., Zald D.H. The impact of common dopamine D2 receptor gene polymorphisms on D2/3 receptor availability: C957T as a key determinant in putamen and ventral striatum. *Transl. Psychiatry*. 2017;7(4):e1091-e1091. DOI 10.1038/tp.2017.45

Sussman C.J., Harper J.M., Stahl J.L., Weigle P. Internet and video game addictions: diagnosis, epidemiology, and neurobiology. *Child Adolesc. Psychiatr. Clin. N. Am*. 2018;27(2):307-326. DOI 10.1016/ j.chc.2017.11.015

Takahashi H., Higuchi M., Suhara T. The role of extrastriatal dopamine D2 receptors in schizophrenia. *Biol. Psychiatry*. 2006;59(10):919-928. DOI 10.1016/j.biopsych.2006.01.022

Tereshchenko S., Kasparov E., Semenova N., Shubina M., Gorbacheva N., Novitckii I., Moskalenko O., Lapteva L. Generalized and specific problematic internet use in Central Siberia adolescents: a school-based study of prevalence, age-sex depending content structure, and comorbidity with psychosocial problems. *Int. J. Environ. Res. Public Health*. 2022;19(13):7593. DOI 10.3390/ijerph 19137593

Vink J.M., van Beijsterveldt T.C.E.M., Huppertz C., Bartels M., Boomsma D.I. Heritability of compulsive Internet use in adolescents. *Addict. Biol*. 2016;21(2):460-468. DOI 10.1111/adb.12218

Volkow N.D., Wang G.J., Fowler J.S., Logan J., Hitzemann R., Ding Y.S., Pappas N., Shea C., Piscani K. Decreases in dopamine receptors but not in dopamine transporters in alcoholics. *Alcohol Clin. Exp. Res*. 1996;20(9):1594-1598. DOI 10.1111/j.1530-0277.1996.tb05936.x

Wang B.-Q., Yao N.-Q., Zhou X., Liu J., Lv Z.-T. The association between attention deficit/hyperactivity disorder and internet addiction: a systematic review and meta-analysis. *BMC Psychiatry*. 2017; 17(1):260. DOI 10.1186/s12888-017-1408-x

Weinstein A., Lejoyeux M. Neurobiological mechanisms underlying internet gaming disorder. *Dialogues Clin. Neurosci*. 2020;22(2): 113-126. DOI 10.31887/DCNS.2020.22.2/aweinstein

Werling A.M., Grünblatt E. A review of the genetic basis of problematic Internet use. *Curr. Opin. Behav. Sci*. 2022;46:101149. DOI 10.1016/j.cobeha.2022.101149

Yu Y., Mo P.K.-H., Zhang J., Li J., Lau J.T.-F. Impulsivity, self-control, interpersonal influences, and maladaptive cognitions as factors of internet gaming disorder among adolescents in China: cross-sectional mediation study. *J. Med. Internet. Res*. 2021;23(10):e26810. DOI 10.2196/26810

Zhang M., Gao X., Yang Z., Wen M., Huang H., Zheng R., Wang W., Wei Y., Cheng J., Han S., Zhang Y. Shared gray matter alterations in subtypes of addiction: a voxel-wise meta-analysis. *Psychopharmacology* (*Berl.*). 2021;238(9):2365-2379. DOI 10.1007/s00213-021-05920-w

Conflict of interest. The authors declare no conflict of interest.
Received April 6, 2024. Revised June 30, 2024. Accepted July 15, 2024.

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» – 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 на платформе Web of Science, 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 30.09.2024. Формат 60 × 84 1/8. Уч.-изд. л. 13.3.