2023 • 27 • 1

Print ISSN 2500-0462
Online ISSN 2500-3259

Научный рецензируемый журнал

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.
Периодичность 8 выпусков в год
DOI 10.18699/VJGB-23-01

2023 • 27 • 1

Print ISSN 2500-0462
Online ISSN 2500-3259

Scientific Peer Reviewed Journal

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

Founded in 1997
Published 8 times annually
DOI 10.18699/VJGB-23-01

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

СОДЕРЖАНИЕ • 2023 • 27 • 1

VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2023 • 27 • 1

Вавиловский журнал генетики и селекции. 2023;27(1):5-6
DOI 10.18699/VJGB-23-02

К 40-летию Научно-исследовательского института медицинской генетики Томского национального исследовательского медицинского центра Российской академии наук

Уважаемые читатели!

Предлагаем вашему вниманию выпуск, посвященный 40-летию Научно-исследовательского института медицинской генетики Томского национального исследовательского медицинского центра Российской академии наук. На страницах текущего номера – статьи, подготовленные сотрудниками института и их коллегами по материалам докладов, прозвучавших на XIII научной конференции «Генетика человека и патология», посвященной юбилею института, проведенной в Томске 20–22 ноября 2022 г.

История института началась 6 июля 1982 г. В этот день в Томске состоялось торжественное открытие Отдела медицинской генетики московского Института медицинской генетики Академии медицинских наук СССР. Спустя 5 лет, в 1987 г., Научно-исследовательский институт ме-

Научно-исследовательский институт медицинской генетики Томского НИМЦ РАН.

Генетическая клиника Научно-исследовательского института медицинской генетики Томского НИМЦ РАН.

дицинской генетики был преобразован в самостоятельное научное учреждение и вошел в состав Томского научного центра АМН СССР.

В 1993 г. НИИ медицинской генетики становится Федеральным центром медико-генетической службы Министерства здравоохранения РФ, а в 1994 г. в структуре института открывается Генетическая клиника – первое и единственное в России специализированное медицинское учреждение для пациентов с наследственными заболеваниями. С момента создания и по настоящее время в фокусе внимания института находятся ключевые направления современной медицинской генетики и генетики человека – клиническая генетика, популяционная генетика и геномика, генетика многофакторных болезней, молекулярная цитогенетика, онтогенетика, персонализированная медицина. Важной частью деятельности института стала подготовка научных и медицинских кадров. Институт готовит аспирантов и ординаторов, с 1998 г. работает диссертационный совет по защите диссертаций по специальности «Генетика». В 1989 г. в Сибирском государственном медицинском университете на базе НИИ медицинской генетики был открыт курс медицинской генетики, а в 1999 г. – первая в Сибири кафедра медицинской генетики. В 2016 г. НИИ медицинской генетики вошел в состав Томского национального исследовательского медицинского центра Российской академии наук – крупнейшего в современной России академического медицинского научно-исследовательского учреждения, объединившего научные и клинические базы шести научно-исследовательских институтов. Сегодня НИИ медицинской генетики Томского НИМЦ – динамично развивающийся институт, один из ведущих медико-генетических центров в России. Исследования института направлены на выявление фундаментальных основ наследственной патологии человека, разработку и внедрение технологий диагностики, лечения и профилактики наследственных заболеваний, развитие медицинской генетики в России как важного направления, интегрирующего практически все отрасли современной медицинской науки.

Открывает выпуск аналитический обзор работ, посвященных генетическим аспектам феномена неслучайного сочетания разных болезней, подготовленный Е.Ю. Брагиной и В.П. Пузырёвым. В работе обобщены и систематизированы современные представления о генетических основах синтропных и дистропных заболеваний, генетической архитектуре многофакторных болезней человека – одного из магистральных направлений научных исследований коллектива лаборатории популяционной генетики.

В обзоре Ю.Ю. Коталевской и В.А. Степанова «Молекулярно-генетические основы буллезного эпидермолиза» рассмотрены новые патогенетические механизмы и гены, ответственные за развитие классических и синдромальных форм буллезного эпидермолиза – тяжелого наследственного заболевания, сопровождающегося хрупкостью кожи и мультисистемными поражениями.

В статье Т.В. Никитиной с соавторами «Сравнительная цитогенетика анэмбрионии и неразвивающейся беременности у человека» обобщены результаты многолетнего цитогенетического скрининга эмбриолетальных мутаций, проводимого в лаборатории цитогенетики института. Накопленный уникальный материал позволил выделить специфические хромосомные аномалии в группах эмбрионов, различающихся по степени тяжести нарушений внутриутробного развития.

Развитие технологий полногеномного анализа вывело на новый уровень работы института в области популяционной и эволюционной генетики. В трех статьях, подготовленных под первым авторством В.А. Степановым, В.Н. Харьковым и Н.А. Колесниковым с сотрудниками лаборатории эволюционной генетики, представлены новые данные о структуре и происхождении генофонда ряда коренных сибирских этносов, продемонстрирована информативность идентичных по происхождению блоков генома в установлении генетических связей между популяциями и определении эволюционных механизмов адаптации человека к факторам окружающей среды.

Эпигенетика и молекулярные механизмы регуляции активности генов при патологии эмбрионального развития и многофакторных заболеваниях также находятся в фокусе внимания коллектива. В работе В.В. Деменевой с соавторами приведены данные об особенностях экспрессии генов как с их канонических, так и с альтернативных промоторов ретротранспозона LINE-1 в зависимости от уровня его метилирования в тканях плаценты. В статье И.А. Гончаровой с соавторами прослежена динамика изменений характера метилирования ДНК в клетках печени при прогрессировании фиброза, индуцированного вирусом гепатита С, до гепатоцеллюлярной карциномы.

Завершает выпуск работа наших коллег из Института цитологии и генетики СО РАН (г. Новосибирск), в которой обозначены ключевые требования к пробоподготовке биологического материала для создания Hi-C библиотек ДНК в целях диагностики хромосомных перестроек современными методами 3D геномики.

Редакция «Вавиловского журнала генетики и селекции» поздравляет коллектив НИИ медицинской генетики Томского НИМЦ с юбилеем, желает дальнейшего развития, творческих успехов и новых достижений на благо российской медицинской генетики!

Научные редакторы выпуска:
академик РАН В.А. Степанов
д-р биол. наук, профессор РАН И.Н. Лебедев

Вавиловский журнал генетики и селекции. 2023;27(1):7-17
DOI 10.18699/VJGB-23-03

Original Russian text https://vavilovj-icg.ru/

Genetic outline of the hermeneutics of the diseases connection phenomenon in human

E.Yu. Bragina[1] ✉, V.P. Puzyrev[1, 2]

[1] Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
[2] Siberian State Medical University, Tomsk, Russia
✉ elena.bragina@medgenetics.ru

Abstract. The structure of diseases in humans is heterogeneous, which is manifested by various combinations of diseases, including comorbidities associated with a common pathogenetic mechanism, as well as diseases that rarely manifest together. Recently, there has been a growing interest in studying the patterns of development of not individual diseases, but entire families associated with common pathogenetic mechanisms and common genes involved in their development. Studies of this problem make it possible to isolate an essential genetic component that controls the formation of disease conglomerates in a complex way through functionally interacting modules of individual genes in gene networks. An analytical review of studies on the problems of various aspects of the combination of diseases is the purpose of this study. The review uses the metaphor of a hermeneutic circle to understand the structure of regular relationships between diseases, and provides a conceptual framework related to the study of multiple diseases in an individual. The existing terminology is considered in relation to them, including multimorbidity, polypathies, comorbidity, conglomerates, families, "second diseases", syntropy and others. Here we summarize the key results that are extremely useful, primarily for describing the genetic architecture of diseases of a multifactorial nature. Summaries of the research problem of the disease connection phenomenon allow us to approach the systematization and natural classification of diseases. From practical healthcare perspective, the description of the disease connection phenomenon is crucial for expanding the clinician's interpretive horizon and moving beyond narrow, disease-specific therapeutic decisions.
Key words: diseases connection phenomenon; syntropy; dystropy; comorbidity; hermeneutics.

For citation: Bragina E.Yu., Puzyrev V.P. Genetic outline of the hermeneutics of the diseases connection phenomenon in human. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):7-17. DOI 10.18699/VJGB-23-03

Генетическая канва герменевтики феномена сочетания болезней человека

Е.Ю. Брагина[1] ✉, В.П. Пузырёв[1, 2]

[1] Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
[2] Сибирский государственный медицинский университет Министерства здравоохранения Российской Федерации, Томск, Россия
✉ elena.bragina@medgenetics.ru

Аннотация. Структура заболеваний у человека неоднородна, характеризуется различными вариантами сочетаний болезней, включая сопутствующие патологии, связанные общим патогенетическим механизмом, а также болезни, редко проявляющиеся совместно на фенотипическом уровне. В последнее время отмечается рост интереса к изучению закономерностей развития не отдельных болезней, а целых семейств, связанных общими патогенетическими механизмами и общими генами, вовлеченными в их развитие. В результате установлен существенный генетический компонент, контролирующий образование конгломератов болезней сложным образом, через функционально взаимодействующие модули отдельных генов в генных сетях. Аналитический обзор исследований по проблематике разных аспектов сочетания болезней и является целью настоящей работы. В обзоре использована метафора герменевтического круга для познания структуры закономерных связей между болезнями, приведены концептуальные рамки, связанные с множественностью заболеваний у индивида. Рассмотрена существующая терминология применительно к ним, среди которых мультиморбидность, полипатии, коморбидность, конгломераты, семейства, «вторые болезни», синтропия и другие. Приведены ключевые результаты, чрезвычайно полезные, прежде всего, для описания генетической архитектуры болезней многофакторной природы. Обобщения по проблеме исследования феномена сочетания болезней позволяют приблизиться к систематизации и естественной классификации болезней. С точки зрения практического здравоохранения описание феномена сочетания болезней имеет решающее значение для расширения интерпретационного горизонта клинициста и выхода за пределы узких, ориентированных на конкретную болезнь терапевтических решений.
Ключевые слова: феномен сочетания болезней; синтропия; дистропия; коморбидность; герменевтика.

Introduction

We live in the "Many Worlds in One" (Vilenkin, 2010) and this One World amazes us with the mystery and the universality of the connections of phenomena, the variety of evolutionary and historical events. These events take place both on a cosmic scale and on a planetary scale and the Earthlings (humanity) are the same universality of connections between themselves and the surrounding world. These connections are formed naturally or randomly, they have a long phylogenetic history of 4 billion years and only a hundred-year ontogenetic history of each individual. The structure of "human" connections, which appears in metabolic and morphophysiological variability, forms the basis of medical assessments – the norm or the disease. Since the beginning of the century, a new approach to the study of these issues – the network analysis – has emerged in biology and medicine. The network analysis is an attempt to understand the laws governing all kinds of networks, from the social to the complex gene networks that rule over all cells and traits, determining health or disease (Barabási et al., 2011).

The human genome, as the assemblage of all genes of the *Homo sapiens* species, is in a complex and not fully understood relationship with the environment and society. The peculiarity of such a relationship between genome and phenome is a difference often noted now: the genome is limited (approximately 3 billion base pairs in humans), the phenome is not limited (its limit depends on how far we want to go) (Paigen, Eppig, 2000). A century before the "genomic revolution" took place, in the 1930s, the outstanding Russian geneticist Alexander S. Serebrovskiy, discussing the problem of organic evolution, defined this problem as an "infinite-finite contradiction" in the "unity of an infinite number of traits and a finite number of genes" (Serebrovskiy, 1973).

In such an infinite world with an infinite number of traits, it is always possible (although it is not easy) to observe and identify traits connected to each other, including those related to pathology. In the clinic, this phenomenon forms the basis for diagnosis and healing, and stable combinations of certain disease traits represent an independent subject of research – the phenomenon of connection of diseases or the diseases connection phenomenon (DCP).

In 1970, the American physician and specialist in the field of epidemiology of non-communicable diseases, Alvan R. Feinstein, proposed the term "comorbidity" for combinations of diseases in individuals. Comorbidity means the manifestation of an additional clinical condition that exists or occurs in addition to the index disease under consideration (Feinstein, 1970). Such a clinical condition may be a disease, a pathological syndrome, pregnancy, a long-term "strict" diet, or a complication of drug therapy. Comorbidity is a complex of several diseases (megaforms, conglomerates) that exist simultaneously in individual patients and are observed much more frequently than would be expected in a random distribution.

The popularity of the term "comorbidity" is striking, especially among clinicians: there is the International Research Community on Multimorbidity (IRCMo), the Journal of Multimorbidity and Comorbidity (https://journals.sagepub.com/description/COB) has been published since 2010, and there is an online medical platform for discussing the diagnosis and treatment of patients with comorbid diagnoses (https://nexusacademy.ru/about). The author of the term "comorbidity" is credited with the discovery that "clarified" the interpretation of comorbid pathology (Vertkin, 2015). And yet, there is still a feeling of overestimation of "clarity" in the understanding of the phenomenon and the term. It is similar to the situation described in the novel of the famous Nobel laureate William Faulkner: "They all talked at once, their voices insistent and contradictory and impatient, making of unreality a possibility, then a probability, then an incontrovertible fact, as people will when their desires become words"[1].

And yet, we must agree that the term "comorbidity" has proved especially successful for clinicians. It became an umbrella term for numerous names of combinations of diseases, variants of two or more forms of pathology in patients and, often, in their closest relatives. Sometimes such diseases are called background or concomitant diseases. In general, according to our calculations, the pool of names for such combinations of diseases includes more than 30 terms. Among them: multimorbidity, polypathies, comorbidity, conglomerates, families, "second diseases" and others. Most often there are diseases that have a "common root" (related pathogenesis, trans-syndromal comorbidity), although other combinations of diseases show nothing in common in pathogenesis (trans-nosological comorbidity). Note that specific terminological studies are limited, and as a result we see no consensus (Azaïs et al., 2016; Navickas et al., 2016). However, in the current situation the object of the study is defined, it is "comorbid patient" (Vertkin, 2015). Good quality clinical and epidemiological data is accumulated, which came in time to become the basis for implementation of "omics" approaches to research on the DCP problem. And there is very serious content and a rather serious genetic aspect. This is the subject of this article.

Conceptual toolkit in the genetic study of the DCP

Here we present a set (assemblage) of views (principles, concepts) connected with each other and forming a unified system, that is useful, in our opinion, for understanding (interpretation, explanation) of the DCP. Let us use a metaphor – the "hermeneutic circle" – which describes the mutual agreement between the individual (part) and the whole, like a hermeneutical rule: we must understand the whole in terms of the detail, and the detail in terms of the whole (Gadamer, 2010). If we consider the DCP as a "whole", it would be reasonable to include as "the details" (the components of the hermeneutic circle) the fragments of concepts (doctrines, principles) of outstanding clinical geneticists, such as the Soviet neurologist Sergey N. Davidenkov (1880–1961), the American geneticist Victor A. McKusick (1921–2008), the German pediatrician Meinhard von Pfaundler (1872–1947) and the now living German-American clinician John M. Opitz. All of them are, at the same time, geneticists and, most importantly, practicing physicians who investigated the polymorphism of disease manifestation and the mysterious phenomenon of a combination of several pathologies in one patient.

[1] William Faulkner. The Sound and the Fury. 1929.

Without keeping the chronological order of their publications, we follow the intended logic in presenting the structure of the hermeneutic circle, i. e., those "details" that can be useful in interpreting the DCP as a "whole".

"Lumpers" and "splitters" (McKusick, 1969). In the 1960s, a discussion was opened in the medical genetics community – what is the "nosology" of genetic diseases? Mainly Mendelian diseases were discussed, but also diseases with an inherited predisposition (multifactorial diseases, MFDs). Phenotypically, patients represent a huge clinical diversity, and possibilities of clarifying the etiology of diseases by molecular genetic or cytogenetic methods were limited in those years. So, physicians-researchers were quite free to classify the patients by combining or separating them. However, during the discussion of this problem, an important generalization was proposed and it was the principles of medical genetics: pleiotropism, variability (polymorphism) and genetic heterogeneity (McKusick, 1968). These principles, above all, can be considered stabilizing the semantic context of understanding the DCP. Today's systematists of human pathology also rely on these principles (Biesecker, 1998; Brunner, van Driel, 2004). Moreover, with the advances in genomic medicine, it became possible to describe the genetic architecture of multifactorial diseases, which is understood as the number of genetic polymorphisms that affect the risk of disease, the distribution of their allelic frequencies and their effect sizes, as well as their genetic mode of action (additive, dominant and/or epistatic, pleiotropic) (Wray et al., 2008).

Syndrome as pleiotropy, conditional tropism hypothesis (Davidenkov, 1947; Opitz, Neri, 2013). The word "syndrome" was first used in English in 1541, as noted by (Opitz, Neri, 2013), and is still used to indicate a common cause rather than simply a set of symptoms. The same authors also evaluate another dictionary definition – the syndrome, as a concurrence of manifestations "characterizing a specific disease", a greater-than-chance concurrence of identical or very similar sets of manifestations in two or more individuals suggesting similar pathogenesis, subject to causal verification through the discovery of physical, infectious, toxicological, or genetic factors (Opitz, Neri, 2013).

Today, biochemical and refined molecular/cytogenetic methods identify genetic causes, epigenetic modifications in combined phenotypes or syndromes with high accuracy. The explanation of such combinations, their persistence or "dividing" in descendants, the severity of manifestations of similar combinations, as well as the interpretation of the relationship between multiple variations of the norm or minor anomalies with their advanced forms of pathology was suggested by Sergey N. Davidenkov in the conditional tropism hypothesis (1947). He used the evolutionary-genetic approach to analyze more than one hundred nosological forms of human nervous diseases. The frequency of combined appearance of the diseases of the nervous system in one patient or in one family is explained by conditional tropism: in addition to its own influence on the nervous system development, the pathological property (gene) also has the ability to dramatically enhance the phenotypic expression of other genotype features "moving into the same direction" and including numerous variants. So, for example, a mild excavation of the foot can take the form of a severe Friedreich's deformity.

Associations, syntropies and dystropies, the transitive association hypothesis (Pfaundler, Seht, 1921; Blair et al., 2013). The renowned textbook for the diagnosis of congenital diseases (Jones, 2011) defines associations as combinations of congenital anomalies that have no well-defined etiology and occur together more often than expected by chance alone. Since its inception, the concept of "associations" has engendered feelings of unease and vagueness, as noted (Opitz, Neri, 2013). They agreed on two variants in the definition of the term: coincidental concurrence (simple rencontre or simple juxtaposition) and combination of anomalies (close connection, polytopic defect of a body area). In the 1900s, new designations of essentially the same associations appeared: but the term "multiple abarts" (from the German *abart*, malformation) was proposed for hereditary diseases and congenital malformations, and "syntropy" (*Syntropie* in German) (Pfaundler, Seht, 1921) was proposed for common multifactorial diseases occurring in one patient at the same time. They not only termed the "mutual disposition, attraction" of the two diseases by the term "syntropy"; in addition, on the basis of abundant clinical data and tens of thousands autopsies Pfaundler and Seht recorded another pathological condition opposite to syntropy – "mutual repulsion", incompatibility (incongruity, dissociation) and named it "dystropy" (*Dystropie* in German). At the same time, intermediate, to a certain extent random and "neutral states" also got their name, "neutropy" (*Neutrotropie* in German). According to these researchers, the term "syndrome" can also be regarded as syntropy, because it means a "selective relationship" of its constituent traits. Another property of the unity of pathological conditions is the appearance of at least two diseases in one patient at the same time (synchrony). Thus, syntropy, syndrome, synchrony ("3S") are related concepts and the main factor uniting them is a similar pathogenesis. For example, in relation to atherosclerosis, diabetes and obesity is a "common root" (Stein O., Stein Y., 1995).

In our current definition, syntropy is a natural-species phenomenon of a combination of two or more pathological conditions (nosologies or syndromes) in an individual and his closest relatives, non-random and having an evolutionary genetic basis; it is a part (an extract) of the human phenome, comprised of a landscape of interacting traits and diseases, reflecting continual molecular-genetic causality (Puzyryov, 2002; Puzyrev et al., 2010). The genes involved in the development of syntropies are called syntropic genes. More precisely, syntropic genes are a set of functionally interacting genes localized throughout the genome, coregulated and involved in a metabolic pathway common to a given syntropy (Puzyryov, 2002; Puzyrev et al., 2010). In the case when regulatory relationships lead to the mutual exclusion of certain phenotypes at the clinical level (dystropy), such genes are termed dystropic in relation to the relevant phenotypes. There is some semantic similarity of the concepts of "syntropic and dystropic genes" with the term "core genes", which were discussed in the recently proposed omnigenic model of complex disease (Boyle et al., 2017).

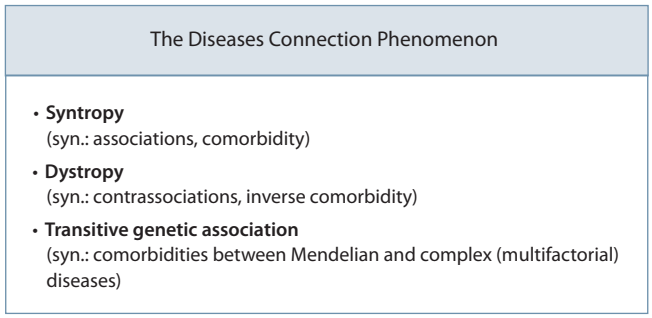

Fig. 1. Classification of disease connection forms in humans.

Finally, let us talk about the transitive genetic association hypothesis. The transitive associations are another form of association from the described above, syntropy (association in the conventional sense and the most common form) and dystropy (dissociation). David R. Blair et al. (2013) hypothesized that statistically significant comorbidities between complex (MFDs) and Mendelian diseases represent a type of genetic association, in which a non-Mendelian phenotype is mapped to the genetic loci that cause the Mendelian disease. In fact, transitive associations are a kind of syntropy, but the phenotype is the result of a combination of complex and Mendelian disease. According to the authors of the hypothesis, such conditions represent about half (54 %) of all comorbid diseases (Blair et al., 2013).

Classification of variants of diseases connection in humans. There is no generally accepted classification of the DCP. Moreover, the tasks of systematization, understanding of the general properties that fix regular connections, in all the variety of such combinations, have not been formulated; the existing attempts to classify such pathological phenomena are still fragmented and conditional. Most often, they are descriptive in nature. This is especially true for the clinical classification of connections designated by the term "comorbidity", and carriers of such pathological features are referred to as "comorbid patients" (Vertkin et al., 2012). Now we can also confirm the attempts to systematize the concept of "syntropy" (Krylov, 2000): by the mechanisms of formation (etiological, pathogenetic, age-related, iatrogenic, random), by the time of occurrence (congenital, delayed, simultaneous, successive) and by clinical significance (inert, interference).

Previously, we (Puzyrev, 2015) proposed the identification of the following forms of diseases connection in individual patients (Fig. 1). The proposed systematization of the DCP forms is also descriptive, but the elements of intrinsic classifications can also be seen in it. This is associated, among other things, with the designation of the key terms of connection characteristics: association and syntropy. There are several subject areas in scientific research (besides medicine), in which the term "syntropy" is used. Viktor B. Vyatkin (2016) designates three fields of science in which the concept of "syntropy" takes an important place, proposing a classification of syntropy (in order of the beginning of their use) into: medical (Pfaundler – von Seht syntropy), biophysical (Fantappiè – Szent-Györgyi – Fuller syntropy), informational (Vyatkin syntropy). In our opinion, these two additional types of syntropy not only have an independent significance, but are also important for the essential understanding of biological processes, including both in general pathology and in the particular pathogenesis of the DCP.

Note that the multiplicity of diseases in an individual is a long-standing problem that had attracted the attention of researchers before the widespread use of the "comorbidity" term. The commonality of the mechanisms of development of non-random pathological connections is reflected in the names of relevant concepts: "the sum of homeostasis diseases" (Dilman, 1968), "diseases of adaptation" (Kaznacheev, 1980), "cardiovascular disease continuum" (Dzau et al., 2006), "metabolic syndrome" (Reaven, 1988). It is important to consider this problem from the genetic perspective, the concepts of diseasome (Goh et al., 2007) and network medicine (Barabási et al., 2011; Kolchanov et al., 2013).

Generalizations on the problem of studying the DCP allow us to approach the intrinsic classifications of the phenomenon. It is important. As Mikhail D. Golubovsky (2006) noted, a good system is an event in science, a conceptual discovery, a new vision of harmony in the chaos of facts. That is why the inclusion of classifications in the hermeneutic circle seems useful.

Actual data on the DCP study

Syntropies (comorbidity)

Syntropy is widespread and more common than we imagine. For example, the 438 common diseases registered in the UK Biobank patient histories (https://www.ukbiobank.ac.uk/) form more than 11,000 possible combinations (Dong et al., 2021). The global nature of the problem has initiated a huge number of studies, mainly of an epidemiological kind. In 2021 alone, the query 'comorbidity' found 34,185 medical and biological articles in the US National Center for Biotechnology Information database (https://pubmed.ncbi.nlm.nih.gov/). Currently, more than 50 million people aged 65 and older – nearly half of Europe's population – have two or more diseases at the same time (Rijken et al., 2018). The number of comorbid patients is predicted to continually increase, affecting up to 68 % of the population by 2035 (Kingston et al., 2018).

Molecular causes of phenotypic connections remain largely unknown, despite active research in this field (Reynolds et al., 2021; Jia et al., 2022; Quick et al., 2022; Shnayder et al., 2022; Wang et al., 2022). Through these studies, it became evident that a significant proportion (46 %) of comorbid conditions is caused by a common component at the level of genes, SNPs, and gene networks interactions (Dong et al., 2021), that in general reflects their pathogenetic relationship. For example, the *HLA-DQB1*, *TLR1*, *WDR36*, *LRRC32*, *IL1RL1*, *GSDMA*, *TSLP*, *IL33*, *SMAD3* genes involved in the pathogenesis of certain allergic diseases are critical for the development of phenotype according to the "atopic march" scenario (Ferreira et al., 2014). Meanwhile, in terms of pathogenesis, seemingly non-obvious connections between diseases are revealed. The existence of many of these connections was not previously even assumed. Varicose veins disease, according to genetic

correlations analysis, is associated with fluid intelligence, prospective memory and educational attainment (Shadrina et al., 2019), and autism is positively correlated with allergic rhinitis and autoimmune disorders (Rzhetsky et al., 2007). A significant addition to the identification of common genes for comorbid conditions is the study of the biological processes in which these genes are involved (Rubio-Perez et al., 2017). The use of such approaches provides a more complete picture of the connections of diseases and common pathogenetic pathways. Knowledge of these connections can be widely applied, including treatment of comorbid patients.

Based on our own research findings on the genetic component of allergic diseases (Freidin et al., 2015) on the one hand, we established the molecular connection of most allergic diseases. On the other hand, with regard to the molecular relationships of allergic diseases with other diseases, we noted their proximity to infectious diseases and a marked distance from autoimmune diseases (Fig. 2).

The *TLR4*, *CAT*, *ANG/RNASE4* genes can make the greatest contribution to the comorbidity of bronchial asthma and hypertension, indicating the importance of inflammation, neovascularization and oxidative stress for the pathogenesis of both diseases (Bragina et al., 2018). The development of bronchial asthma phenotypes in combination with cardiovascular/metabolic disorders is associated with certain genetic variants that affect gene expression, including *CAT*, *TLR4*, *ELF5*, *ABTB2*, *UTP25*, *TRAF3IP3*, *NFKB1*, *LOC105377347*, *C1orf74*, *IRF6* and others, in the target organs of the studied disease profile (Bragina et al., 2022).

Syntropic genes are involved in pathogenesis through complex interactions with other genes, proteins, and environmental factors, which collectively affect the clinical manifestations of comorbidities. In most cases, abnormalities in syntropic genes are localized mainly in non-coding RNAs and intergenic regions functionally associated with the regulation of gene transcription (Dong et al., 2021). In turn, the transcription of syntropic genes depends on epigenetic mechanisms, in particular DNA methylation (Ferreira et al., 2017), which indicates a modifying role of environmental influences on complex phenotype development.

Many syntropic genes are known drug targets for therapy, in particular allergic (*FLG*, *IL13*, *IL1RL1*, *IL6R*, *INPP5D*, *NDFIP1*, *PTGER4*, *TSLP*, *STAT6*) (Ferreira et al., 2017), bronchopulmonary and cardiovascular (*EDNRA*, *ADRB1*, *ADRB2*) diseases (Zolotareva et al., 2019; Dong et al., 2021). More than eight thousand drugs target genes involved in the development of comorbid conditions (Dong et al., 2021). Theoretically, such results not only highlight the important contribution of genes to phenotypic correlations, but also provide an opportunity for drug repurposing to target common genetic components of syntropic diseases.

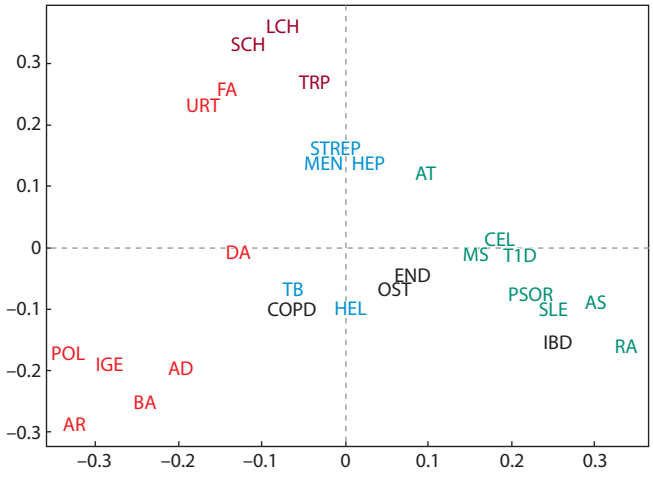

Fig. 2. The results of multidimensional scaling of multifactorial diseases based on the commonality of the genes associated with them (adapted from Freydin et al., 2015).

Abbreviations for diseases: AD – atopic dermatitis, AR – allergic rhinitis, AS – ankylosing spondylitis, AT – autoimmune thyroiditis, BA – bronchial asthma, CEL – celiac disease, COPD – chronic obstructive pulmonary disease, DA – drug allergy, END – endometriosis, FA – food allergy, HEL – Helicobacter infection, HEP – viral hepatitis, IBD – inflammatory bowel disease, IGE – immunoglobulin E level, LCH – leishmaniasis, MEN – meningococcal infection, MS – multiple sclerosis, OST – osteoporosis, POL – pollinosis, PSOR – psoriasis, RA – rheumatoid arthritis, SCH – schistosomiasis, SLE – systemic lupus erythematosus, STREP – streptococcal infection, T1D – type 1 diabetes mellitus, TB – tuberculosis, TRP – trypanosomiasis, URT – urticaria.

Dystropies ("diametrical diseases")

The contrast for syntropy is the diseases that manifest by the phenotypic conflict of one pathological condition with another (dystropy). Dystropy affects various diseases including immune, oncological, neurodegenerative, cardiovascular, autoimmune and others. The spectrum of molecular mechanisms underlying this phenomenon also seems to be very diverse. Research on dystropy focuses on the search for molecular and genetic differences between the diseases. As a result of these studies, differences in the transcription of the same genes in different diseases have been established. Using the example of oncological and neurodegenerative diseases dystropy (Catalá-López et al., 2014), it was revealed that differentially expressed genes are mainly associated with DNA repair, mitochondrial function, stabilization of p53, regulation of angiogenesis, cell cycle, metal ion transport, glucose transport, regulation of apoptotic processes, myeloid leukocyte activation and phagocytosis, mTORC1 and KRAS signaling (Forés-Martos et al., 2021; Pepe et al., 2021). Transcriptional changes in oncogenesis are highly variable; some genes may be activated in some forms of cancer, but suppressed in others, which is probably associated with the features of complex genetic and epigenetic disorders (Zhao et al., 2016). At the same time, common patterns are recorded. In particular, Ibáñez et al. (2014) identified the genes *MT2A*, *MT1X*, *NFKBIA*, *AC009469.1*, *DHRS3*, *CDKN1A*, *TNFRSF1A*, *CRYBG3*, *IL4R*, *MT1M*, *FAM107A*, *ITPKC*, *MID1*, *IL11RA*, *AHNAK*, *KAT2B*, *BCL2*, *PTH1R*, *NFASC* that are simultaneously activated in several disorders of the central nervous system (Alzheimer's disease, Parkinson's disease, schizophrenia) but are suppressed in oncological diseases.

The examples above indicate that phenotypic suppression is mediated by genetic factors. In some cases, potentially "harmful" alleles can be beneficial, creating some kind of trade-off between an increased risk of developing certain diseases and a low risk of developing others. Trade-offs are inevitable,

because the complex integrated functioning of the whole organism needs several interacting parts to work together to perform certain functions. Such integration can lead to a dilemma often called the "cost of complexity" (Wagner et al., 2008), resulting from multiple interacting parts working together to successfully perform a function. Alteration of any part will inevitably negatively affect other features, altering function and reducing overall performance or fitness. Thus, the mechanistic basis for the trade-offs may be focused on pleiotropic genes involved in the biological pathways shared between different traits (Mauro, Ghalambor, 2020). In accordance with this suggestion, the observed divergent nature of the transcription of some genes thought to be important for dystropy can be expected. Diametrical disorders have the intrinsically bidirectional nature of biological processes, whereby expression or activation of genes can be increased or decreased from some optimal value (Crespi, Go, 2015).

Dystropy is significantly formed by drug therapy, because drugs can be connected with the regulation of common molecular processes of phenotypically polar diseases. For example, the use of anticholinesterase agent galantamine and the selective monoamine oxidase inhibitor selegiline in neurodegenerative diseases has anticancer effects (Lazarevic-Pasti et al., 2017; Ryu et al., 2018). Two drugs for breast cancer therapy (exemestane and estradiol) reduce the risk of Alzheimer's disease and other dementias (Branigan et al., 2020; Guglielmotto et al., 2020).

Transitive genetic associations

Genes that can harbor mutations underlying rare and highly penetrant Mendelian diseases affect the development of more common forms of diseases. The effect of mutations can be either a predisposing factor for disease development or *vice versa*, a suppressor of phenotype manifestations. There are various estimates of the involvement of Mendelian genes in the phenotypic expansion of multifactorial pathology. About 300 genes associated with common diseases in genome-wide studies underlie a number of Mendelian diseases (Lupski et al., 2011). By some estimates, the proportion of Mendelian genes in the structure of multifactorial diseases is approximately 23 % (Spataro et al., 2017), but with the growth of genome-wide sequencing data, this amount is likely to increase significantly. In terms of specific pathology, 11 (*ABCG8*, *LCAT*, *APOB*, *APOE*, *LDLR*, *PCSK9*, *CETP*, *LPL*, *LIPC*, *APOA5* and *ABCA1*) out of 30 genes associated with serum lipoprotein concentrations are involved in monogenic disorders of lipid metabolism (Kathiresan et al., 2009). These genes, which are causative variants of both Mendelian disorders and the risk of multifactorial diseases, tend to have higher functional significance and higher expression levels than genes only associated with common diseases. Furthermore, genetic variants in conditionally "Mendelian" genes tend to present higher odds ratios than variants on genes with no link to Mendelian disorders (Spataro et al., 2017).

The idea of a mutational burden materialization in common pathology is not new. The experimental basis for this phenomenon was the publication of Michael S. Brown and Joseph L. Goldstein (Brown, Goldstein, 1986), which showed that patients with heterozygous mutations in the low-density lipoprotein receptor (*LDLR*) gene, along with familial hypercholesterolemia, have coronary atherosclerosis and myocardial infarction. In 2013, David R. Blair (Blair et al., 2013) formulated a hypothesis about the transitivity of rare Mendelian variants into a pathological "allelic continuum" in a wide range of final phenotypic effects from monogenic to complex multifactorial diseases. To date, extensive factual material has been accumulated to support this hypothesis. Carriers of *FLG* gene mutations associated with loss of filaggrin function have an increased risk of developing atopic dermatitis (Sandilands et al., 2007) and bronchial asthma in the context of atopic dermatitis, while at the same time the risk of asthma without atopic dermatitis is reduced (Palmer et al., 2006). This finding suggests that *FLG* gene mutations are an important risk factor for atopy in general, but with different chances for a particular phenotype. Carriers of Gaucher disease mutations, mainly L444P and N370S in the glucocerebrosidase (*GBA*) gene, have an increased risk of Parkinson's disease (Sidransky et al., 2009). Heterozygous carriers of mutations in the cystic fibrosis transmembrane regulator (*CFTR*) gene are predisposed to idiopathic chronic pancreatitis (Weiss et al., 2005) and chronic obstructive pulmonary disease (Divac et al., 2004).

Various approaches are used to gain knowledge about the active contribution of Mendelian disease genes as causative genes for multifactorial diseases. For example, based on the prioritization of data from genome-wide associative studies of various forms of cardiomyopathies, it was found that 70 % of the hypertrophic and 56 % of the dilated cardiomyopathy genes are associated with various Mendelian diseases. This finding suggests that the existing dichotomous classification of diseases – monogenic and multifactorial – has become irrelevant and requires rethinking taking into account new knowledge about the genetic structure of susceptibility (Nazarenko et al., 2022).

The potential of separate gene mutations is evaluated as protective factors in relation to oncological diseases. In particular, activation of apoptosis and autophagy by mutant huntingtin (Gomboeva et al., 2020), as well as the oncotoxic function of CAG repeats (Murmann et al., 2018), the expansion of which causes the Huntington's disease, may prevent the development of most types of cancer in patients with this hereditary disease (Catalá-López et al., 2014). The molecular oncoprotective mechanism of the Laron dwarfism mutation (OMIM #262500) (NM_000163.5(GHR):c.594A>G (p.Glu198=)) in the growth hormone receptor gene is mediated by effects on the activity of genes involved in the control of the cell cycle, mobility, growth and oncogenic transformation (Werner et al., 2020).

Loss of function of individual proteins due to loss-of-function mutations provides specific resistance against some common phenotypes. Protection against type 2 diabetes is associated with carrying a mutation in the zinc transporter type 8 gene (*SLC30A8*) that leads to the synthesis of a truncated protein (Flannick et al., 2014). As a consequence of the resulting deficiency of *SLC30A8* gene function through the mechanism of haploinsufficiency, carriers of mutant alleles have better insulin secretion due to increased glucose

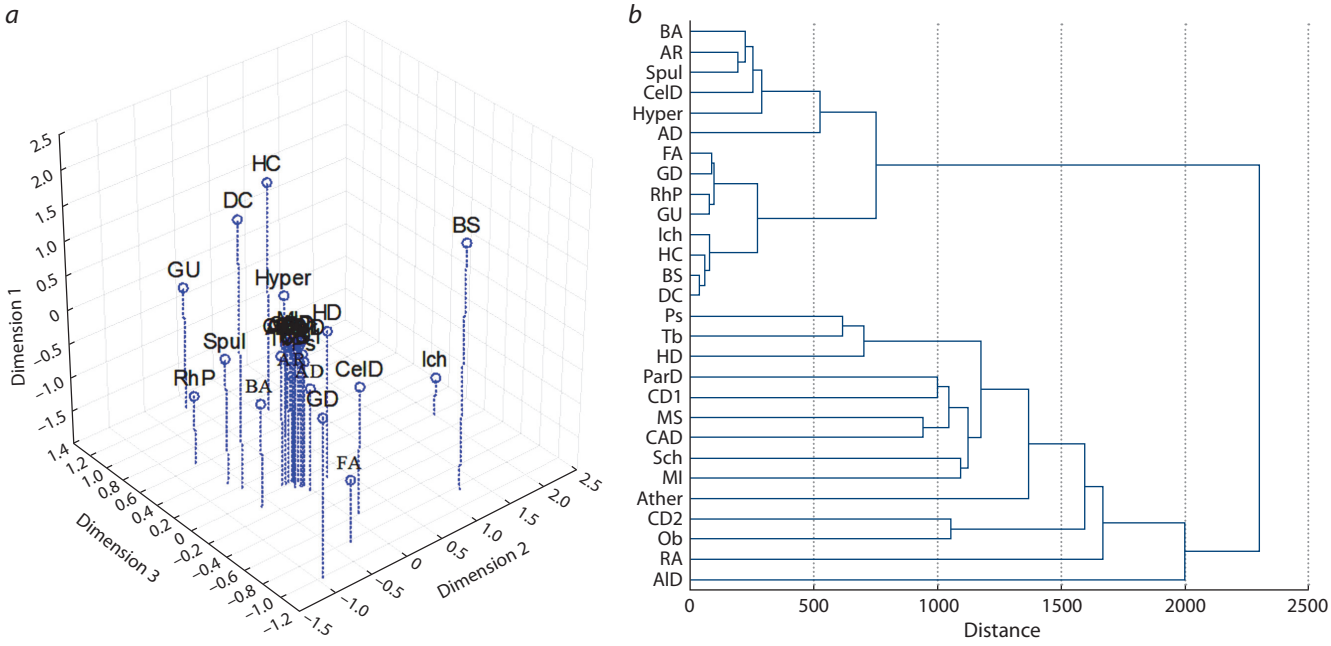

Fig. 3. Modeling of relationships between multifactorial/monogenic diseases by commonality of associated genes, based on: multivariate scaling (*a*) and hierarchical cluster analysis (*b*).

Abbreviations for diseases: AD – eczema (atopic dermatitis), AlD – Alzheimer's disease, AR – allergic rhinitis, Ather – atherosclerosis, BA – atopic bronchial asthma, BS – Brugada syndrome, CAD – coronary artery disease, CD1 – type 1 diabetes, CD2 – type 2 diabetes, CelD – celiac disease, DC – dilated cardiomyopathy, FA – food allergy, GD – Gaucher's disease, GU – gastric ulcer, HC – hypertrophic cardiomyopathy, HD – Huntington's disease, Hyper – arterial hypertension, Ich – ichthyosis, MI – myocardial infarction, MS – multiple sclerosis, Ob – obesity, ParD – Parkinson's disease, Ps – psoriasis, RA – rheumatoid arthritis, RhP – polyposis sinusitis, Sch – schizophrenia, Spul – pulmonary sarcoidosis, Tb – tuberculosis.

sensitivity and proinsulin conversion in the pancreatic beta cells. Another example relates to nonsense mutations (Y142X, C679X, and R46L) in the proprotein convertase subtilisin-kexin type 9 (*PCSK9*) gene underlying familial hypercholesterolemia (OMIM #603776); these mutations result in lower low-density lipoprotein cholesterol level (Cohen et al., 2005). Heterozygous carriers of F508del in the cystic fibrosis transmembrane regulator (*CFTR*) gene, which causes cystic fibrosis, are more resistant to infectious diseases such as cholera, typhoid fever and tuberculosis. Therefore, some authors attribute the high prevalence of cystic fibrosis in the modern human population to the adaptive advantage of mutation carriers (Bosch et al., 2017).

The results of the classification of some multifactorial and Mendelian diseases based on the genes associated with them have identified a large common genetic component of multifactorial diseases (as evidenced by their proximity to the center in Figure 3, *a*). Monogenic diseases are expectedly distant from them, with the exception of Huntington's disease, which is not only close to other neurodegenerative diseases in the degree of gene commonality, but also has molecular similarities with infectious, autoimmune, and cardiometabolic diseases (see Fig. 3, *b*). Overall, in terms of the degree of genetic "commonality" and clustering, most of the diseases studied reflect the generally accepted classification of diseases. However, such modeling has a limitation, since it depends on the extent to which genes are studied, so we should expect a shift in the location of monogenic diseases. At present, the amount of genomic information is rapidly expanding, which brings us closer to filling the gap in the knowledge about disease-associated genes. But even after this gap is filled, a more difficult task remains: to understand the mechanisms of manifestation of the mutation effect and to map the genetic interactions of mutations in different genes, which are combined in a certain way due to structural and molecular interaction (Diss, Lehner, 2018), contributing to phenotypic diversity.

Conclusion

The last decades have been an important milestone for genomic research due to the possibilities of high-throughput technology and the enormous amount of data obtained. It is expected that between 100 million and 2 billion human genomes could be sequenced by 2025, far exceeding growth in other dynamically developing fields that generate Big Data: astronomy, YouTube and Twitter (Stephens et al., 2015). The authors of the aforementioned paper compare genomic research to a "four-headed beast" based on four main demands in genomics throughout the life cycle of the datasets generated by sequencing – acquisition, storage, distribution, and analysis (Stephens et al., 2015). Of these four demands, the greatest effort is required to analyze and comprehend the results obtained, to unravel the complex relationship between genetic variants and phenotypes. This relationship is to a large extent a stochastic process, limited by the genome on the one hand and environmental factors on the other. Consequently, rational ways to comprehend biologically complex objects in the world of Big Data are still relevant.

The results of the study of the diseases connection phenomenon (comorbidity, syntropy/dystropy) accumulated in the scientific literature lead to the necessity and possibility of approaching such a vision of generalization, which was outlined by the outstanding Carl R. Woese in his paper: "...the essence of biology lies not in things as they are, but in things coming into existence" (Woese, Goldenfeld, 2009). In this context, our article attempts to consider the diseases connection phenomenon within the framework of the "hermeneutic circle" metaphor. It is important to note the historical continuity of scientific knowledge on the issue, which was originally based on a holistic view of the development of living organisms, ranging from 'Geoffroyism' (named after Étienne Geoffroy Saint-Hilaire), reflected in the principles of connexion, the unity of elementarity and integrity (Holodkovsky, 1915), to the manifestation of the complex tropism of hereditary factors (Davidenkov, 1947) and the principles of systematization in medical genetics (McKusick, 1968), and finally to the framework of modern concepts of network biology and medicine (Barabási et al., 2011; Kolchanov et al., 2013).

The progress of research on comorbidities has shown the insufficiently comprehensive nature of the existing terminology. For example, in contrast to the term "comorbidity", which has become familiar in medical practice, the genetic discourse of the proximity of concomitant diseases is most fully interpreted by the terms "syntropy" and "dystropy", reflecting the peculiarities of pathogenetic relationships between diseases. The pathogenetic principle of gene involvement in the development of comorbid diseases allowed to classify them as syntropic and dystropic genes (Puzyrev, 2015). Important in this context is the classification of genes on a mechanistic basis into nuclear/core and peripheral genes, that have omnigenic effects on the development of the pathological phenotype through trans- and cis-regulation (Boyle et al., 2017; Liu et al., 2019). It is obvious that, along with nuclear genes, peripheral genes are important objects for MFDs comorbidity studies, because their global activity in specific cell types determines cellular function and disease risk.

The molecular nature of comorbidities, which allows them to be connected in many, often non-fatal and even beneficial combinations, remains difficult to explain due to some "liberties of genome" determined by the dynamic and non-linear nature of the functioning of the system, regulated by feedbacks that can be disrupted in predictable but individual way. The degree of benefit or harm of such combinations of diseases of the conditional "adaptive phenotype" depends on the tradeoffs that are most obvious due to competition for the limited resources of the organism. Probably, vulnerability to some diseases with a relatively low risk of developing others is reduced to the establishment of some "price of complexity", based on the pleiotropic action of genes.

Thus, the diseases connection phenomenon, described in clinical practice for a long time, is of independent interest for fundamental research. The DCP also becomes an additional way to elucidate the etiology and pathogenesis of complex diseases, in the study of which modern methodological and conceptual approaches are involved. On the other hand, the diseases connection phenomenon is important for practical healthcare, since its description is crucial for expanding the clinician's interpretative horizon and moving beyond narrow, disease-specific therapeutic decisions. By expanding our knowledge of the molecular diversity of the human phenome, we can encourage the revision of current disease classifications (Piro, 2012), the identification in such classifications of subtypes with different prognosis for the patient and family members, individual responses to treatment (Manolio, 2013).

References

Azaïs B., Bowis J., Wismar M. Facing the challenge of multimorbidity. *J. Comorb.* 2016;6(1):1-3. DOI 10.15256/joc.2016.6.71.

Barabási A.L., Gulbahce N., Loscalzo J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 2011;12(1):56-68. DOI 10.1038/nrg2918.

Biesecker L.G. Lumping and splitting: molecular biology in the genetics clinic. *Clin. Genet.* 1998;53(1):3-7. DOI 10.1034/j.1399-0004.1998.531530102.x.

Blair D.R., Lyttle C.S., Mortensen J.M., Bearden C.F., Jensen A.B., Khiabanian H., Melamed R., Rabadan R., Bernstam E.V., Brunak S., Jensen L.J., Nicolae D., Shah N.H., Grossman R.L., Cox N.J., White K.P., Rzhetsky A. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell*. 2013;155(1):70-80. DOI 10.1016/j.cell.2013.08.030.

Bosch L., Bosch B., De Boeck K., Nawrot T., Meyts I., Vanneste D., Le Bourlegat C.A., Croda J., da Silva Filho L.V.R.F. Cystic fibrosis carriership and tuberculosis: hints toward an evolutionary selective advantage based on data from the Brazilian territory. *BMC Infect. Dis.* 2017;17(1):340. DOI 10.1186/s12879-017-2448-z.

Boyle E.A., Li Y.I., Pritchard J.K. An expanded view of complex traits: from polygenic to omnigenic. *Cell.* 2017;169(7):1177-1186. DOI 10.1016/j.cell.2017.05.038.

Bragina E.Y., Goncharova I.A., Garaeva A.F., Nemerov E.V., Babovskaya A.A., Karpov A.B., Semenova Y.V., Zhalsanova I.Z., Gomboeva D.E., Saik O.V., Zolotareva O.I., Ivanisenko V.A., Dosenko V.E., Hofestaedt R., Freidin M.B. Molecular relationships between bronchial asthma and hypertension as comorbid diseases. *J. Integr. Bioinform.* 2018;15(4):20180052. DOI 10.1515/jib-2018-0052.

Bragina E.Yu., Goncharova I.A., Zhalsanova I.Z., Nemerov E.V., Nazarenko M.S., Freidin M.B., Puzyrev V.P. Genetic comorbidity of hypertension and bronchial asthma. *Arterial'naya Gipertenziya = Arterial Hypertension*. 2022;28(1):87-95. DOI 10.18705/1607-419X-2022-28-1-87-95. (in Russian)

Branigan G.L., Soto M., Neumayer L., Rodgers K., Brinton R.D. Association between hormone-modulating breast cancer therapies and incidence of neurodegenerative outcomes for women with breast cancer. *JAMA Netw. Open.* 2020;3(3):e201541. DOI 10.1001/jamanetworkopen.2020.1541.

Brown M.S., Goldstein J.L. A receptor-mediated pathway for cholesterol homeostasis. *Science.* 1986;232(4746):34-47. DOI 10.1126/science.3513311.

Brunner H.G., van Driel M.A. From syndrome families to functional genomics. *Nat. Rev. Genet.* 2004;5(7):545-551. DOI 10.1038/nrg1383.

Catalá-López F., Suárez-Pinilla M., Suárez-Pinilla P., Valderas J.M., Gómez-Beneyto M., Martinez S., Balanzá-Martínez V., Climent J., Valencia A., McGrath J., Crespo-Facorro B., Sanchez-Moreno J., Vieta E., Tabarés-Seisdedos R. Inverse and direct cancer comorbidity in people with central nervous system disorders: a meta-analysis of cancer incidence in 577,013 participants of 50 observational studies. *Psychother. Psychosom.* 2014;83(2):89-105. DOI 10.1159/000356498.

Cohen J., Pertsemlidis A., Kotowski I.K., Graham R., Garcia C.K., Hobbs H.H. Low LDL cholesterol in individuals of African descent

resulting from frequent nonsense mutations in *PCSK9*. *Nat. Genet.* 2005;37(2):161-165. DOI 10.1038/ng1509.

Crespi B.J., Go M.C. Diametrical diseases reflect evolutionary-genetic tradeoffs: evidence from psychiatry, neurology, rheumatology, oncology and immunology. *Evol. Med. Public. Health.* 2015;2015(1): 216-253. DOI 10.1093/emph/eov021.

Davidenkov S.N. Evolutionary Genetic Problems in Neuropathology. Leningrad: GIDUV Publ., 1947. (in Russian)

Dilman V.M. Aging, Menopause, Cancer. Moscow, 1968. (in Russian)

Diss G., Lehner B. The genetic landscape of a physical interaction. *eLife.* 2018;7:e32472. DOI 10.7554/eLife.32472.

Divac A., Nikolic A., Mitic-Milikic M., Nagorni-Obradovic L., Petrovic-Stanojevic N., Dopudja-Pantic V., Nadaskic R., Savic A., Radojkovic D. High frequency of the R75Q CFTR variation in patients with chronic obstructive pulmonary disease. *J. Cyst. Fibros.* 2004;3(3):189-191. DOI 10.1016/j.jcf.2004.05.049.

Dong G., Feng J., Sun F., Chen J., Zhao X.M. A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank. *Genome Med*. 2021;13(1):110. DOI 10.1186/s13073-021-00927-6.

Dzau V.J., Antman E.M., Black H.R., Hayes D.L., Manson J.E., Plutzky J., Popma J.J., Stevenson W. The cardiovascular disease continuum validated: clinical evidence of improved patient outcomes. Part I: Pathophysiology and clinical trial evidence (risk factors through stable coronary artery disease). *Circulation.* 2006;114(25):2850-2870. DOI 10.1161/CIRCULATIONAHA.106.655688.

Feinstein A.R. The pre-therapeutic classification of co-morbidity in chronic disease. *J. Chronic Dis.* 1970;23(7):455-468. DOI 10.1016/0021-9681(70)90054-8.

Ferreira M.A., Matheson M.C., Tang C.S., Granell R., Ang W., Hui J., Kiefer A.K., Duffy D.L., Baltic S., Danoy P., Bui M., Price L., Sly P.D., Eriksson N., Madden P.A., Abramson M.J., Holt P.G., Heath A.C., Hunter M., Musk B., Robertson C.F., Le Souëf P., Montgomery G.W., Henderson A.J., Tung J.Y., Dharmage S.C., Brown M.A., James A., Thompson P.J., Pennell C., Martin N.G., Evans D.M., Hinds D.A., Hopper J.L., Australian Asthma Genetics Consortium Collaborators. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *J. Allergy Clin. Immunol.* 2014;133(6):1564-1571. DOI 10.1016/j.jaci.2013.10.030.

Ferreira M.A., Vonk J.M., Baurecht H., Marenholz I., Tian C., Hoffman J.D., Helmer Q., Tillander A., Ullemar V., van Dongen J., … Jorgenson E., Lee Y.A., Boomsma D.I., Almqvist C., Karlsson R., Koppelman G.H., Paternoster L. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* 2017;49(12):1752-1757. DOI 10.1038/ng.3985.

Flannick J., Thorleifsson G., Beer N.L., Jacobs S.B., Grarup N., Burtt N.P., Mahajan A., Fuchsberger C., Atzmon G., Benediktsson R., … Pedersen O., Go-T2D Consortium, T2D-GENES Consortium, Groop L., Cox D.R., Stefansson K., Altshuler D. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* 2014;46(4):357-363. DOI 10.1038/ng.2915.

Forés-Martos J., Boullosa C., Rodrigo-Domínguez D., Sánchez-Valle J., Suay-García B., Climent J., Falcó A., Valencia A., Puig-Butillé J.A., Puig S., Tabarés-Seisdedos R. Transcriptomic and genetic associations between Alzheimer's disease, Parkinson's disease, and cancer. *Cancers* (*Basel*). 2021;13(12):2990. DOI 10.3390/cancers13122990.

Freydin M.B., Ogorodova L.M., Puzyrev V.P. Pathogenetics of Allergic Diseases. Novosibirsk, 2015. (in Russian)

Gadamer G.-G. On the Circle of Understanding. The Relevance of Beauty. Moscow, 2010. (in Russian)

Goh K.I., Cusick M.E., Valle D., Childs B., Vidal M., Barabási A.L. The human disease network. *Proc. Natl. Acad. Sci. USA.* 2007; 104(21):8685-8690. DOI 10.1073/pnas.0701361104.

Golubovsky M.D. Commentary on the Dialogue on Systematics. Nadezhda Mandelstam and Lyubishchev. *Priroda = Nature.* 2006;6: 77-80. (in Russian)

Gomboeva D.E., Bragina E.Y., Nazarenko M.S., Puzyrev V.P. The inverse comorbidity between oncological diseases and Huntington's disease: review of epidemiological and biological evidence. *Russ. J. Genet.* 2020;56(3):269-279. DOI 10.1134/S1022795420030059.

Guglielmotto M., Manassero G., Vasciaveo V., Venezia M., Tabaton M., Tamagno E. Estrogens inhibit amyloid-β-mediated paired helical filament-like conformation of tau through antioxidant activity and miRNA 218 regulation in hTau mice. *J. Alzheimers Dis.* 2020;77(3):1339-1351. DOI 10.3233/JAD-200707.

Holodkovsky N.A. Lamarckism and Geoffreyism. *Priroda = Nature.* 1915;4:533-542. (in Russian)

Ibáñez K., Boullosa C., Tabarés-Seisdedos R., Baudot A., Valencia A. Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS Genet.* 2014;10(2):e1004173. DOI 10.1371/journal.pgen.1004173.

Jia G., Zhong X., Im H.K., Schoettler N., Pividori M., Hogarth D.K., Sperling A.I., White S.R., Naureckas E.T., Lyttle C.S., Terao C., Kamatani Y., Akiyama M., Matsuda K., Kubo M., Cox N.J., Ober C., Rzhetsky A., Solway J. Discerning asthma endotypes through co-morbidity mapping. *Nat. Commun.* 2022;13(1):6712. DOI 10.1038/s41467-022-33628-8.

Jones K.L. Hereditary Syndromes According to David Smith. Atlas-reference book. Moscow: Praktika Publ., 2011. (in Russian)

Kathiresan S., Willer C.J., Peloso G.M., Demissie S., Musunuru K., Schadt E.E., Kaplan L., Bennett D., Li Y., Tanaka T., … Peltonen L., Orho-Melander M., Ordovas J.M., Boehnke M., Abecasis G.R., Mohlke K.L., Cupples L.A. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* 2009;41(1):56-65. DOI 10.1038/ng.291.

Kaznacheev V.P. Modern Aspects of Adaptation. Novosibirsk, 1980. (in Russian)

Kingston A., Robinson L., Booth H., Knapp M., Jagger C., MODEM project. Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. *Age Ageing.* 2018;47(3):374-380. DOI 10.1093/ageing/afx201.

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Lihoschvai V.A. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2013;17(4-2):833-850. (in Russian)

Krylov A.A. To the problem of compatibility of diseases. *Klinicheskaya Meditsyna = Clinical Medicine*. 2000;78(1):56-58. (in Russian)

Lazarevic-Pasti T., Leskovac A., Momic T., Petrovic S., Vasic V. Modulators of acetylcholinesterase activity: from Alzheimer's disease to anti-cancer drugs. *Curr. Med. Chem.* 2017;24(30):3283-3309. DOI 10.2174/0929867324666170705123509.

Liu X., Li Y.I., Pritchard J.K. Trans effects on gene expression can drive omnigenic inheritance. *Cell.* 2019;177(4):1022-1034.e6. DOI 10.1016/j.cell.2019.04.014.

Lupski J.R., Belmont J.W., Boerwinkle E., Gibbs R.A. Clan genomics and the complex architecture of human disease. *Cell.* 2011;147(1): 32-43. DOI 10.1016/j.cell.2011.09.008.

Manolio T.A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* 2013;14(8):549-558. DOI 10.1038/nrg3523.

Mauro A.A., Ghalambor C.K. Trade-offs, pleiotropy, and shared molecular pathways: a unified view of constraints on adaptation. *Integr. Comp. Biol.* 2020;60(2):332-347. DOI 10.1093/icb/icaa056.

McKusick V.A. Some principles of medical genetics. In: Bartalos M. (Ed.) Genetics in Medical Practice. London: Pitman Medical, 1968;43-54.

McKusick V.A. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.* 1969;12(2):298-312. DOI 10.1353/pbm.1969.0039.

Murmann A.E., Gao Q.Q., Putzbach W.E., Patel M., Bartom E.T., Law C.Y., Bridgeman B., Chen S., McMahon K.M., Thaxton C.S., Peter M.E. Small interfering RNAs based on huntingtin trinucleotide repeats are highly toxic to cancer cells. *EMBO Rep.* 2018;19(3):e45336. DOI 10.15252/embr.201745336.

Navickas R., Petric V.K., Feigl A.B., Seychell M. Multimorbidity: what do we know? What should we do? *J. Comorb.* 2016;6(1):4-11. DOI 10.15256/joc.2016.6.72.

Nazarenko M.S., Slepcov A.A., Puzyrev V.P. "Mendelian code" in the genetic structure of complex diseases. *Genetics.* 2022;58(10):1101-1111. DOI 10.31857/S0016675822100058. (in Russian)

Opitz J.M., Neri G. Historical perspective on developmental concepts and terminology. *Am. J. Med. Genet. A.* 2013;161A(11):2711-2725. DOI 10.1002/ajmg.a.36244.

Paigen K., Eppig J.T. A mouse phenome project. *Mamm. Genome.* 2000;11(9):715-717. DOI 10.1007/s003350010152.

Palmer C.N., Irvine A.D., Terron-Kwiatkowski A., Zhao Y., Liao H., Lee S.P., Goudie D.R., Sandilands A., Campbell L.E., Smith F.J., O'Regan G.M., Watson R.M., Cecil J.E., Bale S.J., Compton J.G., DiGiovanna J.J., Fleckman P., Lewis-Jones S., Arseculeratne G., Sergeant A., Munro C.S., El Houate B., McElreavey K., Halkjaer L.B., Bisgaard H., Mukhopadhyay S., McLean W.H. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.* 2006;38(4):441-446. DOI 10.1038/ng1767.

Pepe P., Vatrano S., Cannarella R., Calogero A.E., Marchese G., Ravo M., Fraggetta F., Pepe L., Pennisi M., Romano C., Ferri R., Salemi M. A study of gene expression by RNA-seq in patients with prostate cancer and in patients with Parkinson disease: an example of inverse comorbidity. *Mol. Biol. Rep.* 2021;48(11):7627-7631. DOI 10.1007/s11033-021-06723-0.

Pfaundler M., Seht L.V. Über Syntropie von Krankheitszuständen. *Z. Kinder-Heilk.* 1921;30:100-120. DOI 10.1007/BF02222706.

Piro R.M. Network medicine: linking disorders. *Hum. Genet.* 2012;131(12):1811-1820. DOI 10.1007/s00439-012-1206-y.

Puzyrev V.P. Genetic bases of human comorbidity. *Russ. J. Genet.* 2015;51(4):408-417. DOI 10.1134/S1022795415040092.

Puzyrev V.P., Makeeva O.A., Freidin M.B. Syntropy, genetic testing and personalized medicine. *Per. Med.* 2010;7(4):399-405. DOI 10.2217/pme.10.35.

Puzyryov V.P. Liberties of genome and medical pathogenetics. *Byulleten Sibirskoy Meditsiny = Bulletin of Siberian Medicine.* 2002;1(2):16-29. DOI 10.20538/1682-0363-2002-2-16-29. (in Russian)

Quick C.R., Conway K.P., Swendsen J., Stapp E.K., Cui L., Merikangas K.R. Comorbidity and coaggregation of major depressive disorder and bipolar disorder and cannabis use disorder in a controlled family study. *JAMA Psychiatry.* 2022;79(7):727-735. DOI 10.1001/jamapsychiatry.2022.1338.

Reaven G.M. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes.* 1988;37(12):1595-1607. DOI 10.2337/diab.37.12.1595.

Reynolds R.J., Irvin M.R., Bridges S.L., Kim H., Merriman T.R., Arnett D.K., Singh J.A., Sumpter N.A., Lupi A.S., Vazquez A.I. Genetic correlations between traits associated with hyperuricemia, gout, and comorbidities. *Eur. J. Hum. Genet.* 2021;29(9):1438-1445. DOI 10.1038/s41431-021-00830-z.

Rijken M., Hujala A., van Ginneken E., Melchiorre M.G., Groenewegen P., Schellevis F. Managing multimorbidity: profiles of integrated care approaches targeting people with multiple chronic conditions in Europe. *Health Policy.* 2018;122(1):44-52. DOI 10.1016/j.healthpol.2017.10.002.

Rubio-Perez C., Guney E., Aguilar D., Piñero J., Garcia-Garcia J., Iadarola B., Sanz F., Fernandez-Fuentes N., Furlong L.I., Oliva B. Genetic and functional characterization of disease associations explains comorbidity. *Sci. Rep.* 2017;7(1):6207. DOI 10.1038/s41598-017-04939-4.

Ryu I., Ryu M.J., Han J., Kim S.J., Lee M.J., Ju X., Yoo B.H., Lee Y.L., Jang Y., Song I.C., Chung W., Oh E., Heo J.Y., Kweon G.R. L-Deprenyl exerts cytotoxicity towards acute myeloid leukemia through inhibition of mitochondrial respiration. *Oncol. Rep.* 2018;40(6):3869-3878. DOI 10.3892/or.2018.6753.

Rzhetsky A., Wajngurt D., Park N., Zheng T. Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci. USA.* 2007;104(28):11694-11699. DOI 10.1073/pnas.0704820104.

Sandilands A., Terron-Kwiatkowski A., Hull P.R., O'Regan G.M., Clayton T.H., Watson R.M., Carrick T., Evans A.T., Liao H., Zhao Y., Campbell L.E., Schmuth M., Gruber R., Janecke A.R., Elias P.M., van Steensel M.A., Nagtzaam I., van Geel M., Steijlen P.M., Munro C.S., Bradley D.G., Palmer C.N., Smith F.J., McLean W.H., Irvine A.D. Comprehensive analysis of the gene encoding filaggrin uncovers prevalent and rare mutations in ichthyosis vulgaris and atopic eczema. *Nat. Genet.* 2007;39(5):650-654. DOI 10.1038/ng2020.

Serebrovskiy A.S. Some Problems of Organic Evolution. Moscow, 1973. (in Russian)

Shadrina A.S., Sharapov S.Z., Shashkova T.I., Tsepilov Y.A. Varicose veins of lower extremities: insights from the first large-scale genetic study. *PLoS Genet.* 2019;15(4):e1008110. DOI 10.1371/journal.pgen.1008110.

Shnayder N.A., Novitsky M.A., Neznanov N.G., Limankin O.V., Asadullin A.R., Petrov A.V., Dmitrenko D.V., Narodova E.A., Popenko N.V., Nasyrova R.F. Genetic predisposition to schizophrenia and depressive disorder comorbidity. *Genes* (*Basel*). 2022;13(3):457. DOI 10.3390/genes13030457.

Sidransky E., Nalls M.A., Aasly J.O., Aharon-Peretz J., Annesi G., Barbosa E.R., Bar-Shira A., Berg D., Bras J., Brice A., … Tsuji S., Wittstock M., Wolfsberg T.G., Wu Y.R., Zabetian C.P., Zhao Y., Ziegler S.G. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.* 2009; 361(17):1651-1661. DOI 10.1056/NEJMoa0901281.

Spataro N., Rodríguez J.A., Navarro A., Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum. Mol. Genet.* 2017;26(3):489-500. DOI 10.1093/hmg/ddw405.

Stein O., Stein Y. Smooth muscle cells and atherosclerosis. *Curr. Opin. Lipidol.* 1995;6(5):269-274. DOI 10.1097/00041433-199510000-00005.

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E. Big data: astronomical or genomical? *PLoS Biol.* 2015;13(7):e1002195. DOI 10.1371/journal.pbio.1002195.

Vertkin A.L. Comorbid Patient. Moscow, 2015. (in Russian)

Vertkin A.L., Rumyantsev M.A., Skotnikov A.S. Comorbidity. *Klinicheskaya Meditsyna = Clinical Medicine*. 2012;90(10):4-11. (in Russian)

Vilenkin A. Many Worlds in One. The Search for Other Universes. Moscow: Astrel Publ., 2010. (in Russian)

Vyatkin V.B. About application of the term "syntropy" in scientific research. *Nauchnoye Obozreniye. Referativnyy Zhurnal = Scientific Review. Abstract Journal.* 2016;3:81-84. (in Russian)

Wagner G.P., Kenney-Hunt J.P., Pavlicev M., Peck J.R., Waxman D., Cheverud J.M. Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature.* 2008;452(7186):470-472. DOI 10.1038/nature06756.

Wang M., Tang S., Yang X., Xie X., Luo Y., He S., Li X., Feng X. Identification of key genes and pathways in chronic rhinosinusitis

with nasal polyps and asthma comorbidity using bioinformatics approaches. *Front. Immunol.* 2022;13:941547. DOI 10.3389/fimmu.2022.941547.

Weiss F.U., Simon P., Bogdanova N., Mayerle J., Dworniczak B., Horst J., Lerch M.M. Complete cystic fibrosis transmembrane conductance regulator gene sequencing in patients with idiopathic chronic pancreatitis and controls. *Gut.* 2005;54(10):1456-1460. DOI 10.1136/gut.2005.064808.

Werner H., Sarfstein R., Nagaraj K., Laron Z. Laron syndrome research paves the way for new insights in oncological investigation. *Cells.* 2020;9(11):2446. DOI 10.3390/cells9112446.

Woese C.R., Goldenfeld N. How the microbial world saved evolution from the scylla of molecular biology and the charybdis of the modern synthesis. *Microbiol. Mol. Biol. Rev.* 2009;73(1):14-21. DOI 10.1128/MMBR.00002-09.

Wray N.R., Goddard M.E., Visscher P.M. Prediction of individual genetic risk of complex disease. *Curr. Opin. Genet. Dev.* 2008;18(3):257-263. DOI 10.1016/j.gde.2008.07.006.

Zhao R., Choi B.Y., Lee M.H., Bode A.M., Dong Z. Implications of genetic and epigenetic alterations of *CDKN2A* ($p16^{INK4a}$) in cancer. *EBioMedicine.* 2016;8:30-39. DOI 10.1016/j.ebiom.2016.04.017.

Zolotareva O., Saik O.V., Königs C., Bragina E.Y., Goncharova I.A., Freidin M.B., Dosenko V.E., Ivanisenko V.A., Hofestädt R. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Sci. Rep.* 2019;9(1):16302. DOI 10.1038/s41598-019-52762-w.

ORCID ID

E.Yu. Bragina orcid.org/0000-0002-1103-3073
V.P. Puzyrev orcid.org/0000-0002-2113-4556

Acknowledgements. This study was carried out within the framework of the State Task of the Ministry of Science and Higher Education, No. 122020300041-7.

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

Received October 17, 2022. Revised December 25, 2022. Accepted December 26, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):18-27
DOI 10.18699/VJGB-23-04

Original Russian text https://vavilovj-icg.ru/

Molecular genetic basis of epidermolysis bullosa

Yu.Yu. Kotalevskaya[1, 2] ✉, V.A. Stepanov[3]

[1] Moscow Regional Research and Clinical Institute, Moscow, Russia
[2] Charitable Foundation "BELA. Butterfly Children", Moscow, Russia
[3] Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ kotalevskaya@mail.ru

Abstract. Epidermolysis bullosa (EB) is an inherited disorder of skin fragility, caused by mutations in a large number of genes associated with skin integrity and dermal-epidermal adhesion. Skin fragility is manifested by a decrease in resistance to external mechanical influences, the clinical signs of which are the formation of blisters, erosions and wounds on the skin and mucous membranes. EB is a multisystemic disease and characterized by a wide phenotypic spectrum with extracutaneous complications in severe types, besides the skin and mucous membranes, with high mortality. More than 30 clinical subtypes have been identified, which are grouped into four main types: simplex EB, junctional EB, dystrophic EB and Kindler syndrome. To date, pathogenic variants in 16 different genes are associated with EB and encode proteins that are part of the skin anchoring structures or are signaling proteins. Genetic mutations cause dysfunction of cellular structures, differentiation, proliferation and apoptosis of cells, leading to mechanical instability of the skin. The formation of reduced proteins or decrease in their level leads mainly to functional disorders, forming mild or intermediate severe phenotypes. Absent protein expression is a result of null genetic variants and leads to structural abnormalities, causing a severe clinical phenotype. For most of the genes involved in the pathogenesis of EB, certain relationships have been established between the type and position of genetic variant and the severity of the clinical manifestations of the disease. Establishing an accurate diagnosis depends on the correlation of clinical, genealogical and immunohistological data in combination with molecular genetic testing. In general, the study of clinical, genetic and ultrastructural changes in EB has significantly expanded the understanding of the natural history of the disease and supplemented the data on genotype-phenotype correlations, promotes the search and study of epigenetic and non-genetic disease modifier factors, and also allows developing approaches to radical treatment of the disease. New advances of sequencing technologies have made it possible to describe new phenotypes and study their genetic and molecular mechanisms. This article describes the pathogenetic aspects and genes that cause main and rare syndromic subtypes of EB.
Key words: epidermolysis bullosa; pathogenesis; genotype-phenotype correlations; heterogeneity.

For citation: Kotalevskaya Yu.Yu., Stepanov V.A. Molecular genetic basis of epidermolysis bullosa. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):18-27. DOI 10.18699/VJGB-23-04

Молекулярно-генетические основы буллезного эпидермолиза

Ю.Ю. Коталевская[1, 2] ✉, В.А. Степанов[3]

[1] Московский областной научно-исследовательский клинический институт им. М.Ф. Владимирского, Москва, Россия
[2] Благотворительный фонд «БЭЛА. Дети-бабочки», Москва, Россия
[3] Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ kotalevskaya@mail.ru

Аннотация. Буллезный эпидермолиз (БЭ) – наследственное нарушение, вызывающее хрупкость кожи, обусловленную изменениями генов, отвечающих за целостность кожи и дермо-эпидермальную адгезию. Хрупкость кожи проявляется снижением устойчивости к внешним механическим воздействиям, клинические признаки которой – образование пузырей, эрозий и ран на коже и слизистых оболочках. Для БЭ характерен широкий фенотипический спектр, при тяжелых типах, кроме кожи и слизистых, отмечаются мультисистемность поражения и развитие внекожных осложнений, высокая летальность. Выделено более 30 клинических подтипов БЭ, сгруппированных в четыре основных типа: простой, пограничный, дистрофический БЭ и синдром Киндлера. На сегодняшний день БЭ обусловливают патогенные варианты в 16 различных генах, которые кодируют белки, входящие в состав крепящих структур кожи, и сигнальные белки. Генетические дефекты в этих генах служат причиной нарушения функции клеточных структур, процессов дифференцировки, пролиферации и апоптоза клеток, приводя к механической неустойчивости кожи. Образование укороченных белков или уменьшение их количества обуславливает в основном функциональные нарушения, формируя легкие или среднетяжелые фенотипы. При нулевых генетических вариантах, вследствие которых экспрессия белка утрачивается полностью,

возникают структурные нарушения, влекущие тяжелую клиническую картину. Для большинства вовлеченных в патогенез БЭ генов обнаружены определенные связи между характером и локализацией генетических дефектов с тяжестью клинических проявлений заболевания. Установление точного диагноза зависит от корреляции клинических, генеалогических и иммуногистологических данных в сочетании с молекулярно-генетическим исследованием. В целом изучение клинических, генетических и ультраструктурных изменений при БЭ значительно расширяет понимание естественного течения заболевания и пополняет данные о корреляциях генотип-фенотип, способствует поиску и изучению эпигенетических и негенетических факторов-модификаторов заболевания, а также разработке подходов к радикальному лечению заболевания. Новые возможности технологий секвенирования позволили описать новые фенотипы и изучить их генетические и молекулярные механизмы. В настоящей статье описаны патогенетические аспекты и гены, вызывающие классические и редкие синдромальные подтипы БЭ.
Ключевые слова: буллезный эпидермолиз; патогенез; корреляции генотип-фенотип; гетерогенность.

Introduction

Epidermolysis bullosa (EB) is a group of rare and currently incurable genetically determined hereditary skin diseases. The disease is characterized by fragility of the skin and mucous membranes that occurs with mechanical trauma, seemingly insignificant in terms of shear force, often accompanied by damage to nails, teeth and hair (Pânzaru et al., 2022). The spectrum of characteristic skin manifestations is wide and includes blisters, erosions, wounds that can become chronic, scarring, crusting, milia, skin atrophy, and dyspigmentation. In rare subtypes, it is possible not only to damage the skin, but also muscles, the gastrointestinal tract, kidneys, etc., which is due to the nature of the expression of the defective protein.

The severity of the disease varies from phenotypically mild to severe disabling or lethal variants, which determines the expected prognosis of life expectancy. Severe EB subtypes develop as systemic diseases with secondary multiple organ damage and developmental delay, anemia, affect heart and bones, movement disorders, early susceptibility to skin cancer, and premature death. The treatment of EB is exclusively symptomatic and is aimed at the prevention of mechanical injuries, wound care, treatment of infectious complications and extracutaneous manifestations of the disease. To date, no therapeutic approaches have been able to cure EB patients (Pânzaru et al., 2022).

Epidermolysis bullosa is a demonstrative model of mechanobullous disease, and the study of the underlying mechanisms has made it possible to make significant progress in understanding the fundamentals of the physiology and pathophysiology of the skin. The gained knowledge about EB was reflected in the classification, which was revised several times over the past decade by an international consensus group (Has et al., 2020a). Epidermolysis bullosa is divided into four main types – simplex EB (EBS), junctional EB (JEB), dystrophic BE (DEB) and Kindler's syndrome (KS), which is based on the ultrastructural changes and the level of blisters in the skin and reflects the consequences of genetic defects on the protein function. Epidermolysis bullosa is clinically and genetically very heterogeneous, inherited in an autosomal dominant (AD) or autosomal recessive (AR) pattern of inheritance (Has et al., 2020a). Advances in understanding the pathogenesis of EB contribute to the development of potentially effective protein, cell and gene therapies (Has et al., 2020b).

The epidermal basal layer, basement membrane zone (BMZ) and extracellular matrix are key subregions that take central place in the pathophysiology of EB (Uitto et al., 2017) and genetic changes disturb the structure or function of their proteins (Mariath et al., 2020a). Pathogenic variants in 16 different genes determine the genetic and allelic heterogeneity of EB and the grouping of four main types of EB, including more than 30 clinical subtypes. EB-associated genes encode intracellular, transmembrane or extracellular proteins, mainly structural components of the cytoskeleton (keratin 5 and 14), BMZ (α6β4 integrin, type XVII collagen, laminin-332, type VII collagen, α3 integrin alpha subunit, kindlin-1) or intercellular adhesion proteins (desmoplakin, plakophilin, placoglobin) (see the Table) (Has, Bruckner-Tuderman, 2014). Table presents the key processes of pathogenesis leading to a certain phenotype.

The main EB types

Simplex EB (EBS) is the most common type, accounting for about 70 % of all patients with EB (Has, Fischer, 2019), and includes 14 clinical subtypes according to the latest classification. Simplex EB has a wide range of severity, from mild with blistering of the palms and feet to generalized forms with extracutaneous lesions, sometimes fatal (Fine, 2010). Simplex EB is most often caused by defects in the keratin filaments of basal keratinocytes, has a different genetic basis: it is associated with changes in at least seven genes and represents the greatest clinical diversity.

Most subtypes of EBS are inherited in the AD pattern, although AR inheritance occurs in some regions of the world (Gostyńska et al., 2015; Vahidnezhad et al., 2019). The most common EBS subtypes observed in clinical practice are caused by mutations in the keratin 5 or 14 genes (70–80 % of cases), while according to the literature data, at least 17 % of patients with EBS had mutations *de novo* (Bolling et al., 2011; Wertheim-Tysarowska et al., 2016). In addition, EBS with AD inheritance may be associated with heterozygous variants in the *PLEC* or *KLHL24* genes (Grilletta, 2019; Kiritsi et al., 2021). Rare digenic inheritance caused by mutations in the *KRT5* and *KRT14* genes have also been described in patients with EBS (Sathishkumar et al., 2016).

Keratin 5 and keratin 14 have a similar protein structure consisting of a central α-helical rod domain that is responsible for the polymerization of these proteins to form keratin tonofilaments. The core domain is subdivided into segments 1A, 1B, 2A and 2B by flexible linkers L1, L12 and L2, flanked by variable domains V1 and V2 in both proteins. Also, keratin 5

Classification of epidermolysis bullosa (EB) and main mechanisms of pathogenesis

Subtype	Type	Gene affected	Mechanism
Simplex EB – intraepidermal			
Localized Intermediate	AD	*KRT5*, *KRT14*	Abnormal keratin cytoskeletal network and basal cytolysis
Severe	AD	*KRT5*, *KRT14*	Abnormal keratin cytoskeletal network, clumping of keratin tonofilaments leading to basal cytolysis
With mottled pigmentation	AD	Predominantly *KRT5*, less frequently *KRT14*	Rupture of keratin filaments, basal cytolysis, and additional aggregation of densely packed complex melanosomes in the perinuclear cytoplasm of basal keratinocytes
Migratory circinate	AD	*KRT5*	Keratin 5 elongation due to late termination codon generation leads to T-cell mediated inflammation
Intermediate with cardiomyopathy	AD	*KLHL24*	Pathogenic variants result in a truncated and more stable KLHL24 protein, followed by increased degradation of KRT14
Intermediate with *PLEC* mutations	AD, AR	*PLEC*	Reduced HD due to disruption of the internal plaque to which the keratin cytoskeleton attaches, followed by basal cytolysis
Intermediate with muscular dystrophy	AR	*PLEC*	The cleavage is as close as possible to the BMZ; HD is significantly reduced in size; breaking of the interaction of sarcomeres due to the rodless isoform of plectin inside the Z-disks; defective attachment between assembled desmin filaments triggers the formation of desmin protein aggregates as well as secondary mitochondrial failure
Severe with pyloric atresia	AR	*PLEC*	Absent plectin
EB simplex	AR	*KRT5*, *KRT14*	Absence or significant reduction of bundles of tonofilaments in basal keratinocytes
Localized or intermediate with BP230 deficiency	AR	*DST*	Absence of inner HD plaques, compensatory increase in KRT14 and plectin, which may explain the mild phenotype
Localized or intermediate with exophilin 5 deficiency	AR	*EXPH5*	Disruption of intracellular vesicles transport along actin and tubulin networks; an increase in perinuclear vesicles with abnormal keratin; loss of basal keratinocyte adhesion
Localized with nephropathy (CD151 deficiency)	AR	*CD151*	Pathogenic variants lead to reduced adhesion of keratinocytes mediated by laminin-332-integrin α3ß1 complexes in the epidermis and podocytes
Junctional EB – intralamina lucida			
Severe	AR	*LAMA3*, *LAMB3*, *LAMC2*	Laminin 332 is usually absent; reduced HD; abnormal or absent sub-basal lamina densa; reduction of anchoring filaments
Intermediate	AR	*LAMA3*, *LAMB3*, *LAMC2*, *COL17A*	Reduced laminin-332; absent or reduced collagen of type XVII
With pyloric atresia	AR	*ITGA6*, *ITGB4*	Absent or markedly reduced α6β4 integrin; Pathogenic variants in the *ITGB4* gene leading to partial expression of integrin β4 may cause a milder phenotype
Localized	AR	*LAMA3*, *LAMB3*, *LAMC2*, *COL17A*, *ITGB4*, *ITGA3*	Variable abnormalities and expression levels in defective proteins
Inversa	AR	*LAMA3*, *LAMB3*, *LAMC2*	Reduced expression of laminin-332
Late onset	AR	*COL17A*	Reduced or abnormal expression of type XVII collagen

End of the Table

Subtype	Type	Gene affected	Mechanism
Laryngo-onycho-cutaneous syndrome	AR	*LAMA3*	Abnormally truncated α3A subunit of laminin-332
With interstitial lung disease and nephrotic syndrome	AR	*IGTA3*	Variants with loss of function of the α3 integrin subunit are common; missense variants may cause milder disease and improve survival
Dystrophic EB – sublamina densa			
DDEB, intermediate	AD	*COL7A1*	Reduced or abnormal type VII collagen; usually due to missense mutations causing glycine replacement at the hinge region of the type VII collagen triple helix
DDEB, localized	AD	*COL7A1*	Reduced or abnormal type VII collagen resulting from monoallelic deletions, missense variants, or splice site mutations
DDEB, pruriginosa	AD	*COL7A1*	Pathogenic mechanism is unknown
DDEB, self-improving	AD	*COL7A1*	Intracellular accumulation of unsecreted procollagen VII; retention of type VII collagen in basal keratinocytes; gradual improvement in the formation of type VII collagen and anchoring fibrils for unknown reasons
RDEB, intermediate	AR	*COL7A1*	Combinations of biallelic pathogenic variants in *COL7A1* (missense, nonsense, insertions, deletions, and splice site variants) result in reduced or abnormal production of type VII collagen
RDEB, severe	AR	*COL7A1*	Biallelic null variants in *COL7A1* that result in a markedly reduced or absent type VII collagen and, therefore, in a lack of functional anchoring fibrils
RDEB, inversa	AR	*COL7A1*	It is assumed that specific mutations of arginine and glycine in the triple helix of type VII collagen reduce the thermal stability of the protein, causing clinical manifestations in areas of the body with a higher temperature, incl. on mucous membranes
RDEB, localized	AR	*COL7A1*	Reduced or abnormal type VII collagen
RDEB, pruriginosa	AR	*COL7A1*	As in DDEB, pruriginosa
RDEB, self-improving	AR	*COL7A1*	As in DDEB, self-improving
DEB, severe	AD, AR	*COL7A1*	Pathogenetic mechanisms are unknown, the phenotype occurs in compound heterozygotes for a dominant mutation of glycine in *COL7A1* in one allele and a recessive variant in the second allele, which changes the protein microenvironment in the BMZ area, increasing the severity of clinical manifestations
Kindler syndrome – variable and mixed			
Kindler syndrome	AR	*FERMT1*	Pathogenic variants promote disruption of keratinocyte cytoskeletal networks, abnormal integrin activation, and loss of keratinocyte adhesion to the underlying basement membrane

Note. AD – autosomal dominant type of inheritance; AR – autosomal recessive type of inheritance; BMZ – basement membrane zone; HD – hemidesmosome; DDEB – dominant dystrophic epidermolysis bullosa; RDEB – recessive dystrophic epidermolysis bullosa.

has a conserved H1 and H2 homology domain. The *KRT5* and *KRT14* genes are expressed in the basal keratinocytes of the epidermis, where their protein products combine to form heterodimeric molecules. The K5 and K14 dimers are the main components of the keratinocyte intermediate filament system, which assemble into an intracellular network (Bunick, Milstone, 2017).

Among the pathogenic variants in the *KRT5* and *KRT14* genes predominate dominant missense variants that affect the ability of keratins to interact with their partner. The locations of the pathogenic variant in the functional domains of the *KRT5* or *KRT14* genes are of key importance (Arin et al., 2010). Dominant-negative pathogenic variants are grouped at the beginning of 1A or the end of 2B segments of the helical rod domain of *KRT5* and *KRT14* and are typical of severe generalized EBS, because these domains are highly conserved and are considered critical for filament assembly.

The most common pathogenic variants are: p.Glu477Lys in the *KRT5* gene and p.Arg125Cys, p.Arg125His, p.Asn123Ser in the *KRT14* gene (Bolling et al., 2011; Vahidnezhad et al., 2016). In moderate EBS, pathogenic variants are located in the second part of segments 1A or 2B of the core domain of *KRT5* and *KRT14*. In this subtype, they do not alter the process of keratin elongation during filament assembly, but

impair their function (Has, Bruckner-Tuderman, 2014). In the localized EBS subtype, pathogenic variants are clustered in both *KRT5* and *KRT14*, usually outside the highly conserved core domain boundary motifs, as well as in L12 linkers, in addition, in the *KRT5* gene in the H1 domain, causing structural instability of the filaments (Bardhan et al., 2020). More distinct correlations with the genotype were found in the EBS subtype with spotted pigmentation, which is associated with pathogenic variants in the V1 domain of the *KRT5* gene, so the p.Pro25Leu variant accounts for 90–95 % of mutations in this subtype (Arin et al., 2010).

Severe and moderate EBS with AR inheritance is associated with rare pathogenic biallelic variants in *KRT14* and *KRT5*, which are found in consanguineous families (Vahidnezhad et al., 2016). Homozygous mutations in the *KRT5* gene result in a severe phenotype, extracutaneous manifestations, and early mortality (Has et al., 2006).

The latest revision of the EB classification characterized rare syndromic EBS subtypes associated with mutations in the *PLEC*, *KLHL24*, *DST*, *EXPH5*, and *CD151* genes (see the Table); we will consider them below.

The plectin protein encoded by the *PLEC* gene is a cytoskeletal protein that links the network of intermediate filaments to HD and thus acts as a mediator of the mechanical stability of keratinocytes in the skin (Natsuga, 2015). A large number of alternatively spliced first exons of the plectin gene form multiple protein isoforms and determine different expression in tissues, which ensures clinical diversity and leads to four rare EBS phenotypes.

Pathogenic variants in the *PLEC* gene were mainly found in exons 31 and 32, loss-of-function variants leading to more severe phenotypes such as EBS with pyloric atresia (EBS-AP) and, as a result of null variants of the *PLEC* gene, EBS with muscular dystrophy (EBS-MD), where skeletal muscle fibers lose their structural integrity due to defects in desmin filaments (Natsuga, 2015). Moderate EBS with AR inheritance is caused by a specific homozygous nonsense mutation p.Arg16X in the first exon encoding the plectin 1a isoform, resulting in the absence of only this specific isoform (Gostyńska et al., 2015). Also, in exon 31 of the *PLEC* gene, a dominant amino acid substitution p.Arg2110Trp was described, which leads to a partial loss of protein function and causes HD fragmentation (Kiritsi et al., 2021), which is clinically manifested as moderate EBS.

The KLHL24 protein belongs to a family of highly conserved proteins with BTB/kelch domains; pathogenic variants in the *KLHL24* gene lead to dysregulation of autoubiquitination and change the regulation of degradation of keratin 14 molecules and cause its fragmentation (Dhanoa et al., 2013). In the EBS subtype caused by mutations in the *KLHL24* gene, in all described cases, a heterozygous variant was observed in the start codon, the most common being c.1A-G with a dominant negative effect (Bardhan et al., 2020). Also, 85 % of patients with this subtype of EBS at a young age develop dilated cardiomyopathy caused by KLHL24-mediated degradation of desmin, the main protein of cardiomyocyte intermediate filaments (Grilletta, 2019).

Dystonin (BPAG1) is a member of the plakin protein family (Ganani et al., 2021). The *DST* gene encodes the epithelial BPAG1-e isoform, which is a structural component of internal HD plaques and consists of a helical-helical rod domain and flanking N- and C-termini. The N-terminus of the BPAG1-e protein is involved in its integration into HD and has binding sites for type XVII collagen and β4 integrin, while the C-terminus is the key point of attachment of keratin intermediate filaments (Kumar et al., 2015). Mutations in *BPAG1-e* have been shown to be associated with impaired adhesion of keratinocytes, increased cell migration with reduced expression of β4-integrins on the cell surface (Ganani et al., 2021). Clinically, it leads to a mild phenotype.

The exophilin-5 protein, a RAB27b GTPase effector protein encoded by the *EXPH5* gene, is not a structural component of intermediate filaments, desmosomes, or PD. Although its role is not fully known, it is assumed that it contributes to the regulation of intracellular transport of vesicles, including the control of their formation and movement along the actin and tubulin networks, as well as the secretion of exosomes (Natsuga et al., 2010). Single families are described with homozygous variants in the *EXPH5* gene, leading to a frameshift, as well as in combination with nonsense variants. Mild clinical manifestations have been described.

In the epidermis, the expression of the transmembrane protein CD151 is localized in HD, binding to α6β4 integrin and stabilizing its interaction with laminin-332, and plays a critical role in the formation of the HD complex. CD151 mediates cell adhesion and intracellular vesicular transport of integrins. In the kidneys, it forms complexes with α3β1 and α6β1 integrins and is required for the correct assembly of glomerular and tubular basement membranes (Margadant et al., 2010). A defect in the CD151 protein determines the clinical manifestations in individuals with CD151-associated EBS, including nephropathy with proteinuria (Karamatic Crew et al., 2004).

Junctional EB (JEB) is also a clinically and genetically heterogeneous group of skin fragility disorders, includes nine clinical subtypes, and is a rare type of EB (Has et al., 2020a). JEB subtypes have pathognomonic signs, for example, in severe generalized subtype, granulation tissue is rapidly formed in typical places, and mortality is high (Kiritsi et al., 2011). Phenotypic variability in JEB is extremely wide – from only nail dystrophy to death in the first year of life. Pathogenic variants in seven different genes lead to the development of JEB, all subtypes are inherited in the AR type. Pathogenic variants in the *LAMA3*, *LAMB3*, and *LAMC2* genes encoding the α3, β3, and γ2 chains of laminin-332, as well as in the *COL17A1* gene, encoding type XVII collagen, lead to the most common JEB subtypes (Uitto et al., 2016). Rare JEB phenotypes are associated with deficiency of a6β4 integrin, leading to the development of JEB with pyloric atresia and deficiency of the α3 subunit of a3β1 integrin, causing EBS with respiratory and renal involvement (Kiritsi et al., 2013).

The laminin-332 protein is a heterotrimer consisting of α3, β3, and γ2 chains, which are encoded by the *LAMA3*, *LAMB3*, and *LAMC2* genes, respectively. Together with the extracellular domain of type XVII collagen, they form anchor filaments. The laminin-332 protein binds at its α-chain C-terminus to α3β1 integrins in focal adhesion sites and α6β4 integrins in HD, connecting the surface of basal keratinocytes to the

dermal-epidermal BM (Dogic et al., 1998). In the dermis, the N-terminus of laminin-332 chains bind to type VII collagen, so that anchor filaments and anchor fibrils connect directly (Aumailley et al., 2003). Loss of laminin-332 expression causes extreme skin fragility and excess granulation tissue in generalized severe JEB. In laminin-332-deficient JEB subtypes, the *LAMB3* gene is altered in 70 % of cases. Approximately 9 % of patients with JEB have mutations in the *LAMA3* and *LAMC2* genes, respectively (Varki et al., 2006; Uitto et al., 2016). The most common pathogenic variant is p.R635X, as a "hot" mutation point, which accounts for 45–63 % of all pathogenic alleles of the *LAMB3* gene in generalized severe JEB, resulting in the absence of one of the three proteins that are assembled in laminin-332.

Mild manifestations of EB are caused by missense mutations, splicing site mutations, and deletions with preservation of the reading frame, which, leading to a change in the key positions of protein subunits, affect the ability of laminin α3, β3, and γ2 to assemble into a trimeric molecule, its secondary structure, and its ability to form intracellular anchor fibrils (Kiritsi et al., 2011).

A special phenotype, laryngo-onycho-cutaneous syndrome (LOC syndrome), manifests pathogenic variants that form a stop codon in exon 39, specific for the alpha-3 subunit of the *LAMA3* gene, where three causative variants have been described so far: p.V51fs; p.Gln157Ter; p.Trp16Ter (Wang et al., 2022). Recently, C. Prodinger et al. (2021) reported three new mutations in the *LAMA3* gene outside of exon 39.

Type XVII collagen protein is a homotrimer consisting of three identical subunits, is a transmembrane protein and the main structural component of PD, has both intracellular and extracellular domains. Type XVII collagen acts as a cell surface receptor for extracellular matrix proteins (van den Bergh, Giudice, 2003). The extracellular domain of type XVII collagen is associated with laminin-332; in this regard, it takes part in the creation of anchor filaments, can control cell motility, determines the spatial orientation of laminin-332 and its location in the collagen-IV-containing lamina BM (Tong, Xu, 2004).

This protein also regulates the differentiation of ameloblasts, epithelial cells involved in the formation of tooth enamel (Asaka et al., 2009). Enamel defects, ranging from punctate to generalized hypoplasia, occur in all subtypes of JEB, arising from impaired adhesion of the odontogenic epithelium from which ameloblasts originate (Wright et al., 2015).

Also, type XVII collagen plays a central role in regulating the proliferation of the interfollicular epidermis, participating in the maintenance of hair follicle stem cells, where it controls their aging program, which may explain the irreversible hair loss in people with type XVII collagen deficiency (Matsumura et al., 2016).

Pathogenic variants in the *COL17A1* gene usually result in moderate JEB (Pasmooij et al., 2004), although a few fatal cases have been described with the presence of pathogenic *COL17A1* variants (Murrell et al., 2007). According to D. Kiritsi et al. (2011) 69 % of the *COL17A1* gene variants were nonsense variants, insertions or deletions, 19 % were missense variants, and 12 % were splice site variants. Pathogenic variants leading to exon skipping in the *COL17A1* gene have a mitigating effect on the phenotype, allowing the production of a sufficiently functional protein (Condrat et al., 2019).

In some cases, nonsense mutations can cause mild manifestations of moderate generalized JEB due to alternative splicing mechanisms. It was shown that in patients with a homozygous nonsense mutation p.R795X in exon 33, *COL17A1* mRNA is formed as a result of alternative splicing, which allows the production of a small amount of type XVII collagen.

Integrins are heterodimeric transmembrane receptors consisting of α- and β-subunits that form a functional receptor (Masunaga et al., 2017). In the epidermis, α3β1, α6β4, and α2β1 integrins are the most abundant. The α6β4 integrin binds to laminin-332 and to keratin filaments within the cell, which allows it to coordinate the cellular response and regulate adhesion, migration, and proliferation of keratinocytes. The α6β4 integrin is also involved in the formation of HD integrity and stability and interacts with type XVII collagen, plectin, and dystonin (Has, Nyström, 2015). The group of β1-integrins is associated mainly with the basal surface of keratinocytes and is involved in the formation of focal contacts. The α3β1 integrin is found both on the basal and lateral surfaces of basal keratinocytes, where it can participate in intercellular contacts.

The *ITGA6* gene encodes the α6 subunit, the *ITGB4* gene encodes the β4 subunit of the α6β4 integrin. Pathogenic variants in these genes, leading to premature termination of translation, form a severe phenotype that can be fatal in the neonatal period. Most of the mutations are in the *ITGB4* gene; splicing site variants, small deletions and insertions, amino acid substitutions that lead to a rare subtype, JEB with pyloric atresia, have been described (Masunaga et al., 2017). Studies of genotype and phenotype correlations indicate that variants located in the extracellular domain of *ITGB4* are usually associated with a more severe phenotype compared to those located in the cytoplasmic tail (Mariath et al., 2021). In the *ITGA6* gene, single variants with loss of function in patients from consanguineous families are described, which are clinically manifested by early manifestation and often with a fatal outcome (Schumann et al., 2013; Masunaga et al., 2017).

The *ITGA3* gene encodes the α3 integrin subunit, which is associated with the β1 subunit and forms the α3β1 integrin involved in interactions with extracellular matrix proteins, including laminins. The α3 integrin subunit is expressed in basal keratinocytes, podocytes, tubular epithelial cells, alveolar epithelial cells, and many other tissues (Bardhan et al., 2020).

Several cases of JEB with interstitial lung disease and renal abnormalities have been reported, associated with pathogenic variants in the *ITGA3* gene, the expression of which in different tissues explains the multiple organ damage observed in patients. In addition, the relationship between the α3 integrin subunit and the cell membrane is complex, including post-translational modifications, cleavage, heterodimerization with the β1 integrin subunit, and association with CD151. Amino acid substitutions can interfere with these events and act as null mutations, leading to severe disease (Has et al., 2012); variants that express a residual, truncated, or dysfunctional protein may result in a milder phenotype and improved survival (Liu et al., 2021).

Dystrophic EB (DEB) is divided into two main groups: dominant DEB (DDEB) and recessive DEB (RDEB). Clini-

cal diversity includes 11 subtypes, with all subtypes having cutaneous and extracutaneous manifestations of varying severity. In general, RDEB is more severe than DDEB, ranging from severe skin manifestations with scarring and fibrosis, secondary complications, extracutaneous manifestations, and a high risk of squamous cell carcinoma, to mild skin fragility on the extremities or only nail dystrophy. However, there is a significant phenotypic overlap between AD and AR subtypes, which often makes it clinically difficult to establish the type of inheritance of DEB in a patient, especially if the proband is the only patient in the family.

DEB develops as a result of mutations in only one gene, the *COL7A1* gene, which encodes type VII collagen, the main protein of anchor fibrils that provide BM attachment to the underlying dermis. Pathogenic variants in the *COL7A1* gene lead to a disruption in the production and molecular structure of collagen, causing splitting of the upper layers of the dermis and destruction of anchor fibrils. The nature and location of pathogenic variants are important determinants of the phenotype (Hovnanian et al., 1997), which is determined by the expression and residual function of collagen VII (Mariath et al., 2020).

Type VII collagen is a non-fibrillar collagen synthesized by both epidermal keratinocytes and dermal fibroblasts and is localized in the BM zone below the epithelial layers, representing a homotrimer consisting of three identical α1 polypeptide chains (Uitto et al., 1992). Each α1 polypeptide chain contains a central collagen triple helix domain and terminal non-collagen NC-1 and NC-2 domains (Chung, Uitto, 2010). The triple helical domain consists of a repeating Gly-X-Y sequence interrupted by non-collagenous regions, the largest of which consists of 39 amino acid residues and is known as the "hinge" region.

The NC-1 domain mediates the attachment of anchor fibrils to the basement membrane and islets of collagen IV in the dermis (Bruckner-Tuderman et al., 2013). The NC-2 domain contains conserved cysteines involved in the formation of disulfide bonds, which provide a link between type VII collagen homotrimers. In addition, loops formed by anchor fibrils in the papillary dermis capture and mechanically hold interstitial collagen fibers, which are mainly represented by collagen types I, III and V.

Also, type VII collagen promotes the migration of keratinocytes, which is one of the stages of wound healing, providing their re-epithelialization (Woodley et al., 2008). It has been shown that in DEB the size or number of anchor fibrils is reduced, or they are absent (Uitto, Christiano, 1992), which determines the main mechanism and severity of the development of clinical manifestations. Impaired function of type VII collagen leads to deep skin defects, scarring of the mucous membranes, the formation of milia and fibrosis.

Hundreds of mutations in the *COL7A1* gene associated with DEB are known (Sawamura et al., 2005; Has et al., 2020a). Thus, most cases of DDEB are the result of dominant-negative mutations. Approximately 75 % of DDEB patients have glycine substitution variants in the Gly-X-Y triple helical domain, especially in exons 73, 74, and 75 (Varki et al., 2007). At this hotspot, glycine residue substitutions can lead to greater protein destabilization than glycine residue substitutions within a long, continuous collagen segment, and variants near the hinge region cause protein misfolding and accumulation within cells (Chen et al., 2001). It is also suggested that exon 73 may encode amino acid residues important for the ability of type VII collagen to provide keratinocyte motility (Woodley et al., 2008).

Glycine as well as other amino acid substitutions and splicing variants outside the Gly-X-Y region are also found in DDEB, and intrafamilial phenotypic variability suggests that other factors may influence cell resistance to friction (Koss-Harnes et al., 2002).

Severe generalized RDEB usually results from the absence of a *COL7A1* gene product resulting in null genetic variants on both alleles, about 30 % of which are nonsense stop codon or splicing variants resulting in large deletions, determining disease severity (van den Akker et al., 2011). Many patients with moderate RDEB are compound heterozygous for a premature stop codon and glycine substitution in the collagen domain, another missense variant or variants that disrupt splicing, resulting in destabilization of the triple helix or conformational changes in the protein that affect its functionality (Pânzaru et al., 2022).

This variety of combinations of genetic variants explains the wide range of clinical manifestations. So, for example, p.Gly2049Glu and p.Arg2063Trp variants, adjacent to the "hinge" region, reduce the ability to maintain fibroblast adhesion and lead to a significantly reduced ability to support keratinocyte migration, which slows down the healing of erosions in RDEB patients (Varki et al., 2007). Milder forms of RDEB are often caused by a combination of splicing and missense variants. Glycine substitutions may also occur in RDEB.

Kindler syndrome (KS) is a rare type of EB characterized by skin fragility and acral blistering from birth, development of skin atrophy, photosensitivity, poikiloderma, diffuse palmoplantar hyperkeratosis, and pseudosyndactyly (Lai-Cheong, McGrath, 2022). Morphologically, KS differs from other types of EB in that blistering is variable and can occur at different levels of the dermal-epidermal junction. KS develops as a result of pathogenic variants in the *FERMT1* gene. The disease is inherited according to the AR type.

The *FERMT1* gene encodes the Kindlin-1 protein, which is a multidomain focal adhesion protein. Kindlin-1 is involved in the connection between the actin cytoskeleton and the extracellular matrix through focal adhesion, as well as in integrin-associated signaling pathways (Has et al., 2011). The absence of Kindlin-1 leads to disorganization of keratinocytes as a result of incorrect integrin-mediated cell adhesion and migration (Rognoni et al., 2016). More than 90 pathogenic loss-of-function variants have been registered in the *FERMT1* gene, including: missense, nonsense, and splicing variants; insertions; and Alu-mediated gene rearrangements that result in the absence of the Kindlin-1 protein or the production of a non-functional protein (Lai-Cheong, McGrath, 2022). Environmental factors play an important role in the phenotypic diversity of KS and determine the severity. X. Zhang et al. suggested that homologue 1 of the fermitin family is important for the suppression of UV-induced inflammation and DNA repair (Zhang et al., 2017).

Conclusion

The multisystem manifestations of EB and the involvement of a significant number of proteins that provide mechanical stability of the skin in the pathogenesis are due to its genetic heterogeneity. Pathogenic variants in the genes encoding proteins of the epidermal and dermal anchoring complexes, as well as signal proteins that determine the integrity of the skin, lead to their structural and functional defects. EB is characterized by pronounced clinical variability and, at the same time, similar manifestations in different genotypes. Research and accumulation of the data of the natural history of disease and the genotype-phenotype correlations contribute to understanding the EB pathogenesis and determine the development of approaches for symptomatic and etiopathogenetic, in particular, gene therapy.

References

Arin M.J., Grimberg G., Schumann H., de Almeida H. Jr., Chang Y.-R., Tadini G., Kohlhase J., Krieg T., Bruckner-Tuderman L., Has C. Identification of novel and known *KRT5* and *KRT14* mutations in 53 patients with epidermolysis bullosa simplex: correlation between genotype and phenotype. *Br. J. Dermatol*. 2010;162(6):1365-1369. DOI 10.1111/j.1365-2133.2010.09657.x.

Asaka T., Akiyama M., Domon T., Nishie W., Natsuga K., Fujita Y., Abe R., Kitagawa Y., Shimizu H. Type XVII collagen is a key player in tooth enamel formation. *Am. J. Pathol*. 2009;174(1):91-100. DOI 10.2353/AJPATH.2009.080573.

Aumailley M., el Khal A., Knöss N., Tunggal L. Laminin 5 processing and its integration into the ECM. *Matrix Biol*. 2003;22(1):49-54. DOI 10.1016/S0945-053X(03)00013-1.

Bardhan A., Bruckner-Tuderman L., Chapple I.L.C., Fine J.-D., Harper N., Has C., Magin T.M., Marinkovich M.P., Marshall J.F., McGrath J.A., Mellerio J.E., Polson R., Heagerty A.H. Epidermolysis bullosa. *Nat. Rev. Dis. Primers*. 2020;6(1):78. DOI 10.1038/s41572-020-0210-0.

Bolling M.C., Lemmink H.H., Jansen G.H.L., Jonkman M.F. Mutations in *KRT5* and *KRT14* cause epidermolysis bullosa simplex in 75 % of the patients. *Br. J. Dermatol*. 2011;164(3):637-644. DOI 10.1111/j.1365-2133.2010.10146.x.

Bruckner-Tuderman L., Mcgrath J.A., Robinson E.C., Uitto J. Progress in Epidermolysis bullosa research: summary of DEBRA International Research Conference 2012. *J. Invest. Dermatol*. 2013;133(9):2121-2126. DOI 10.1038/jid.2013.127.

Bunick C.G., Milstone L.M. The X-ray crystal structure of the keratin 1-keratin 10 helix 2B heterodimer reveals molecular surface properties and biochemical insights into human skin disease. *J. Invest. Dermatol.* 2017;137(1):142-150. DOI 10.1016/j.jid.2016.08.018.

Chen M., Keene D.R., Costa F.K., Tahk S.H., Woodley D.T. The carboxyl terminus of type VII collagen mediates antiparallel dimer formation and constitutes a new antigenic epitope for epidermolysis bullosa acquisita autoantibodies. *J. Biol. Chem*. 2001;276(24):21649-21655. DOI 10.1074/JBC.M100180200.

Chung H.J., Uitto J. Type VII collagen: the anchoring fibril protein at fault in dystrophic epidermolysis bullosa. *Dermatol. Clin*. 2010;28(1):93-105. DOI 10.1016/J.DET.2009.10.011.

Condrat I., He Y., Cosgarea R., Has C. Junctional epidermolysis bullosa: allelic heterogeneity and mutation stratification for precision medicine. *Front. Med. (Lausanne)*. 2019;5:363. DOI 10.3389/fmed.2018.00363.

Dhanoa B.S., Cogliati T., Satish A.G., Bruford E.A., Friedman J.S. Update on the Kelch-like (KLHL) gene family. *Hum. Genomics*. 2013;7(1):13. DOI 10.1186/1479-7364-7-13.

Dogic D., Rousselle P., Aumailley M. Cell adhesion to laminin 1 or 5 induces isoform-specific clustering of integrins and other focal adhesion components. *J. Cell Sci*. 1998;111(Pt. 6):793-802. DOI 10.1242/JCS.111.6.793.

Fine J.-D. Inherited epidermolysis bullosa. *Orphanet J. Rare Dis*. 2010;5:12. DOI 10.1186/1750-1172-5-12.

Ganani D., Malovitski K., Sarig O., Gat A., Sprecher E., Samuelov L. Epidermolysis bullosa simplex due to bi-allelic *DST* mutations: Case series and review of the literature. *Pediatr. Dermatol*. 2021;38(2):436-441. DOI 10.1111/pde.14477.

Gostyńska K.B., Nijenhuis M., Lemmink H., Pas H.H., Pasmooij A.M.G., Lang K.K., Castañón M.J., Wiche G., Jonkman M.F. Mutation in exon 1a of *PLEC*, leading to disruption of plectin isoform 1a, causes autosomal-recessive skin-only epidermolysis bullosa simplex. *Hum. Mol. Genet*. 2015;24(11):3155-3162. DOI 10.1093/hmg/ddv066.

Grilletta E.A. Cardiac transplant for epidermolysis bullosa simplex with KLHL24 mutation-associated cardiomyopathy. *JAAD Case Rep*. 2019;5(10):912-914. DOI 10.1016/j.jdcr.2019.08.009.

Has C., Bauer J.W., Bodemer C., Bolling M.C., Bruckner-Tuderman L., Diem A., Fine J.-D., Heagerty A., Hovnanian A., Marinkovich M.P., Martinez A.E., McGrath J.A., Moss C., Murrell D.F., Palisson F., Schwieger-Briel A., Sprecher E., Tamai K., Uitto J., Woodley D.T., Zambruno G., Mellerio J.E. Consensus reclassification of inherited epidermolysis bullosa and other disorders with skin fragility. *Br. J. Dermatol*. 2020a;183(4):614-627. DOI 10.1111/bjd.18921.

Has C., Bruckner-Tuderman L. The genetics of skin fragility. *Annu. Rev. Genomics Hum. Genet*. 2014;15(1):245-268. DOI 10.1146/annurev-genom-090413-025540.

Has C., Castiglia D., del Rio M., Garcia Diez M., Piccinni E., Kiritsi D., Kohlhase J., Itin P., Martin L., Fischer J., Zambruno G., Bruckner-Tuderman L. Kindler syndrome: Extension of FERMT1 mutational spectrum and natural history. *Hum. Mutat*. 2011;32(11):1204-1212. DOI 10.1002/HUMU.21576.

Has C., Chang Y.-R., Volz A., Hoeping D., Kohlhase J., Bruckner-Tuderman L. Novel keratin 14 mutations in patients with severe recessive epidermolysis bullosa simplex. *J. Invest. Dermatol.* 2006;126(8):1912-1914. DOI 10.1038/sj.jid.5700312.

Has C., Fischer J. Inherited epidermolysis bullosa: New diagnostics and new clinical phenotypes. *Exp. Dermatol*. 2019;28(10):1146-1152. DOI 10.1111/exd.13668.

Has C., Nyström A. Epidermal basement membrane in health and disease. *Curr. Top. Membr*. 2015;76:117-170. DOI 10.1016/bs.ctm.2015.05.003.

Has C., South A., Uitto J. Molecular therapeutics in development for epidermolysis bullosa: Update 2020. *Mol. Diagn. Ther*. 2020b;24(3):299-309. DOI 10.1007/s40291-020-00466-7.

Has C., Spartà G., Kiritsi D., Weibel L., Moeller A., Vega-Warner V., Waters A., He Y., Anikster Y., Esser P., Straub B.K., Hausser I., Bockenhauer D., Dekel B., Hildebrandt F., Bruckner-Tuderman L., Laube G.F. Integrin α3 mutations with, lung, and skin disease. *N. Engl. J. Med*. 2012;366(16):1508-1514. DOI 10.1056/NEJMOA1110813.

Hovnanian A., Rochat A., Bodemer C., Petit E., Rivers C.A., Prost C., Fraitag S., Christiano A.M., Uitto J., Lathrop M., Barrandon Y., de Prost Y. Characterization of 18 new mutations in COL7A1 in recessive dystrophic epidermolysis bullosa provides evidence for distinct molecular mechanisms underlying defective anchoring fibril formation. *Am. J. Hum. Genet*. 1997;61(3):599-610. DOI 10.1086/515495.

Karamatic Crew V., Burton N., Kagan A., Green C.A., Levene C., Flinter F., Brady R.L., Daniels G., Anstee D.J. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. *Blood*. 2004;104(8):2217-2223. DOI 10.1182/blood-2004-04-1512.

Kiritsi D., Has C., Bruckner-Tuderman L. Laminin 332 in junctional epidermolysis bullosa. *Cell Adh. Migr*. 2013;7(1):135-141. DOI 10.4161/CAM.22418.

Kiritsi D., Kern J.S., Schumann H., Kohlhase J., Has C., Bruckner-Tuderman L. Molecular mechanisms of phenotypic variability in junctional epidermolysis bullosa. *J. Med. Genet*. 2011;48(7):450-457. DOI 10.1136/JMG.2010.086751.

Kiritsi D., Tsakiris L., Schauer F. Plectin in skin fragility disorders. *Cells*. 2021;10(10):2738. DOI 10.3390/cells10102738.

Koss-Harnes D., Høyheim B., Anton-Lamprecht I., Gjesti A., Jørgensen R.S., Jahnsen F.L., Olaisen B., Wiche G., Gedde-Dahl T. A site-specific plectin mutation causes dominant epidermolysis bullosa simplex Ogna: two identical *de novo* mutations. *J. Invest. Dermatol*. 2002;118(1):87-93. DOI 10.1046/j.0022-202x.2001.01591.x.

Kumar V., Bouameur J.E., Bär J., Rice R.H., Hornig-Do H.T., Roop D.R., Schwarz N., Brodesser S., Thiering S., Leube R.E., Wiesner R.J., Brazel C.B., Heller S., Binder H., Löffler-Wirth H., Seibel P., Magin T.M. A keratin scaffold regulates epidermal barrier formation, mitochondrial lipid composition, and activity. *J. Cell Biol*. 2015; 211(5):1057-1075. DOI 10.1083/JCB.201404147.

Lai-Cheong J.E., McGrath J.A. Kindler syndrome. In: Murrell D. (Ed.). Blistering Diseases: Clinical Features, Pathogenesis, Treatment. Berlin; Heidelberg: Springer, 2022;433-439. DOI 10.1007/978-3-662-45698-9_43.

Liu Y., Yue Z., Wang H., Li M., Wu X., Lin H., Han W., Lan S., Sun L. A novel *ITGA3* homozygous splice mutation in an ILNEB syndrome child with slow progression. *Clin. Chim. Acta*. 2021;523:430-436. DOI 10.1016/J.CCA.2021.10.027.

Margadant C., Charafeddine R.A., Sonnenberg A. Unique and redundant functions of integrins in the epidermis. *FASEB J*. 2010;24(11): 4133-4152. DOI 10.1096/fj.09-151449.

Mariath L.M., Santin J.T., Frantz J.A., Doriqui M.J.R., Schuler-Faccini L., Kiszewski A.E. Genotype-phenotype correlations on epidermolysis bullosa with congenital absence of skin: A comprehensive review. *Clin. Genet*. 2021;99(1):29-41. DOI 10.1111/cge.13792.

Mariath L.M., Santin J.T., Schuler-Faccini L., Kiszewski A.E. Inherited epidermolysis bullosa: update on the clinical and genetic aspects. *An. Bras. Dermatol*. 2020;95(5):551-569. DOI 10.1016/j.abd.2020.05.001.

Masunaga T., Ogawa J., Akiyama M., Nishikawa T., Shimizu H., Ishiko A. Compound heterozygosity for novel splice site mutations of *ITGA6* in lethal junctional epidermolysis bullosa with pyloric atresia. *J. Dermatol*. 2017;44(2):160-166. DOI 10.1111/1346-8138.13575.

Matsumura H., Mohri Y., Thanh Binh N., Morinaga H., Fukuda M., Ito M., Kurata S., Hoeijmakers J., Nishimura E.K. Hair follicle aging is driven by transepidermal elimination of stem cells via COL17A1 proteolysis. *Science.* 2016;351(6273):aad4395. DOI 10.1126/science.aad4395.

Murrell D.F., Pasmooij A.M.G., Pas H.H., Marr P., Klingberg S., Pfendner E., Uitto J., Sadowski S., Collins F., Widmer R., Jonkman M.F. Retrospective diagnosis of fatal BP180-deficient non-Herlitz junctional epidermolysis bullosa suggested by immunofluorescence (IF) antigen-mapping of parental carriers bearing enamel defects. *J. Invest. Dermatol*. 2007;127(7):1772-1775. DOI 10.1038/SJ.JID.5700766.

Natsuga K. Plectin-related skin diseases. *J. Dermatol. Sci*. 2015;77(3): 139-145. DOI 10.1016/j.jdermsci.2014.11.005.

Natsuga K., Nishie W., Shinkuma S., Arita K., Nakamura H., Ohyama M., Osaka H., Kambara T., Hirako Y., Shimizu H. Plectin deficiency leads to both muscular dystrophy and pyloric atresia in epidermolysis bullosa simplex. *Hum. Mutat*. 2010;31(10):E1687-E1698. DOI 10.1002/humu.21330.

Pânzaru M.C., Caba L., Florea L., Braha E.E., Gorduza E.V. Epidermolysis bullosa – a different genetic approach in correlation with genetic heterogeneity. *Diagnostics*. 2022;12(6):1325. DOI 10.3390/diagnostics12061325.

Pasmooij A.M.G., van der Steege G., Pas H.H., Sillevis Smitt J.H., Nijenhuis A.M., Zuiderveen J., Jonkman M.F. Features of epidermolysis bullosa simplex due to mutations in the ectodomain of type XVII collagen. *Br. J. Dermatol*. 2004;151(3):669-674. DOI 10.1111/J.1365-2133.2004.06041.X.

Prodinger C., Chottianchaiwat S., Mellerio J.E., McGrath J.A., Ozoemena L., Liu L., Moore W., Laimer M., Petrof G., Martinez A.E. The natural history of laryngo-onycho-cutaneous syndrome: A case series of six pediatric patients and literature review. *Pediatr. Dermatol*. 2021;38(5):1094-1101. DOI 10.1111/PDE.14790.

Rognoni E., Ruppert R., Fässler R. The kindlin family: functions, signaling properties and implications for human disease. *J. Cell Sci*. 2016;129(1):17-27. DOI 10.1242/JCS.161190.

Sathishkumar D., Orrin E., Terron-Kwiatkowski A., Browne F., Martinez A.E., Mellerio J.E., Ogboli M., Hoey S., Ozoemena L., Liu L., Baty D., McGrath J.A., Moss C. The p.Glu477Lys mutation in keratin 5 is strongly associated with mortality in generalized severe epidermolysis bullosa simplex. *J. Invest. Dermatol.* 2016;136(3):719-721. DOI 10.1016/j.jid.2015.11.024.

Sawamura D., Goto M., Yasukawa K., Sato-Matsumura K., Nakamura H., Ito K., Nakamura H., Tomita Y., Shimizu H. Genetic studies of 20 Japanese families of dystrophic epidermolysis bullosa. *J. Hum. Genet*. 2005;50(10):543-546. DOI 10.1007/S10038-005-0290-4.

Schumann H., Kiritsi D., Pigors M., Hausser I., Kohlhase J., Peters J., Ott H., Hyla-Klekot L., Gacka E., Sieron A.L., Valari M., Bruckner-Tuderman L., Has C. Phenotypic spectrum of epidermolysis bullosa associated with α6β4 integrin mutations. *Br. J. Dermatol*. 2013; 169(1):115-124. DOI 10.1111/bjd.12317.

Tong G., Xu R. The role of collagen XVII in regulating keratinocyte migration. *Lab. Invest*. 2004;84(10):1225-1226. DOI 10.1038/labinvest.3700168.

Uitto J., Bruckner-Tuderman L., Christiano A.M., McGrath J.A., Has C., South A.P., Kopelan B., Robinson E.C. Progress toward treatment and cure of epidermolysis bullosa: Summary of the DEBRA international research symposium EB2015. *J. Invest. Dermatol.* 2016; 136(2):352-358. DOI 10.1016/j.jid.2015.10.050.

Uitto J., Christiano A.M. Molecular genetics of the cutaneous basement membrane zone. Perspectives on epidermolysis bullosa and other blistering skin diseases. *J. Clin. Invest*. 1992;90(3):687-692. DOI 10.1172/JCI115938.

Uitto J., Chung-Honet L.C., Christiano A.M. Molecular biology and pathology of type VII collagen. *Exp. Dermatol*. 1992;1(1):2-11. DOI 10.1111/J.1600-0625.1992.TB00065.X.

Uitto J., Has C., Vahidnezhad H., Youssefian L., Bruckner-Tuderman L. Molecular pathology of the basement membrane zone in heritable blistering diseases: The paradigm of epidermolysis bullosa. *Matrix Biol*. 2017;57-58;76-85. DOI 10.1016/j.matbio.2016.07.009.

Vahidnezhad H., Youssefian L., Saeidian A.H., Mozafari N., Barzegar M., Sotoudeh S., Daneshpazhooh M., Isaian A., Zeinali S., Uitto J. *KRT5* and *KRT14* mutations in epidermolysis bullosa simplex with phenotypic heterogeneity, and evidence of semidominant inheritance in a multiplex family. *J. Invest. Dermatol.* 2016;136(9): 1897-1901. DOI 10.1016/j.jid.2016.05.106.

Vahidnezhad H., Youssefian L., Saeidian A.H., Uitto J. Phenotypic spectrum of epidermolysis bullosa: The paradigm of syndromic versus non-syndromic skin fragility disorders. *J. Invest. Dermatol.* 2019;139(3):522-527. DOI 10.1016/j.jid.2018.10.017.

van den Akker P.C., Jonkman M.F., Rengaw T., Bruckner-Tuderman L., Has C., Bauer J.W., Klausegger A., Zambruno G., Castiglia D., Mellerio J.E., Mcgrath J.A., van Essen A.J., Hofstra R.M.W., Swertz M.A. The international dystrophic epidermolysis bullosa patient registry: an online database of dystrophic epidermolysis bullosa patients and their COL7A1 mutations. *Hum. Mutat*. 2011;32(10): 1100-1107. DOI 10.1002/humu.21551.

van den Bergh F., Giudice G.J. BP180 (type XVII collagen) and its role in cutaneous biology and disease. *Adv. Dermatol*. 2003;19:37-71.

Varki R., Sadowski S., Pfendner E., Uitto J. Epidermolysis bullosa. I. Molecular genetics of the junctional and hemidesmosomal variants. *J. Med. Genet*. 2006;43(8):641-652. DOI 10.1136/JMG.2005.039685.

Varki R., Sadowski S., Uitto J., Pfendner E. Epidermolysis bullosa. II. Type VII collagen mutations and phenotype-genotype correlations in the dystrophic subtypes. *J. Med. Genet*. 2007;44(3):181-192. DOI 10.1136/JMG.2006.045302.

Wang R., Sun L., Habulieti X., Liu J., Guo K., Yang X., Ma D., Zhang X. Novel variants in *LAMA3* and *COL7A1* and recurrent variant in *KRT5* underlying epidermolysis bullosa in five Chinese families. *Front. Med*. 2022;16(5):808-814. DOI 10.1007/S11684-021-0878-X.

Wertheim-Tysarowska K., Ołdak M., Giza A., Kutkowska-Kaźmierczak A., Sota J., Przybylska D., Woźniak K., Śniegórska D., Niepokój K., Sobczyńska-Tomaszewska A., Rygiel A.M., Płoski R., Bal J., Kowalewski C. Novel sporadic and recurrent mutations in *KRT5* and *KRT14* genes in Polish epidermolysis bullosa simplex patients: further insights into epidemiology and genotype-phenotype correlation. *J. Appl. Genet*. 2016;57(2):175-181. DOI 10.1007/s13353-015-0310-9.

Woodley D.T., Hou Y., Martin S., Li W., Chen M. Characterization of molecular mechanisms underlying mutations in dystrophic epidermolysis bullosa using site-directed mutagenesis. *J. Biol. Chem*. 2008;283(26):17838-17845. DOI 10.1074/JBC.M709452200.

Wright J.T., Carrion I.A., Morris C. The molecular basis of hereditary enamel defects in humans. *J. Dent. Res*. 2015;94(1):52-61. DOI 10.1177/0022034514556708.

Zhang X., Luo S., Wu J., Zhang L., Wang W.-hui, Degan S., Erdmann D., Hall R., Zhang J.Y. KIND1 loss sensitizes keratinocytes to UV-induced inflammatory response and DNA damage. *J. Invest. Dermatol*. 2017;137(2):475-483. DOI 10.1016/J.JID.2016.09.023.

ORCID ID

Yu.Yu. Kotalevskaya orcid.org/0000-0001-8405-8223
V.A. Stepanov orcid.org/0000-0002-5166-331X

Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019–2027, agreement No. 075-15-2021-1061, RF 193021X0029).

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

Received October 17, 2022. Revised December 10, 2022. Accepted December 20, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):28-35
DOI 10.18699/VJGB-23-05

Original Russian text https://vavilovj-icg.ru/

Comparative cytogenetics of anembryonic pregnancies and missed abortions in human

T.V. Nikitina ✉, E.A. Sazhenova, E.N. Tolmacheva, N.N. Sukhanova, S.A. Vasilyev, I.N. Lebedev

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ t.nikitina@medgenetics.ru

Abstract. Miscarriage is an important problem in human reproduction, affecting 10–15 % of clinically recognized pregnancies. The cases of embryonic death can be divided into missed abortion (MA), for which the ultrasound sign of the embryo death is the absence of cardiac activity, and anembryonic pregnancy (AP) without an embryo in the gestational sac. The aim of this study was to compare the frequency of chromosomal abnormalities in extraembryonic tissues detected by conventional cytogenetic analysis of spontaneous abortions depending on the presence or absence of an embryo. This is a retrospective study of 1551 spontaneous abortions analyzed using GTG-banding from 1990 to 2022 (266 cases of AP and 1285 cases of MA). A comparative analysis of the frequency of chromosomal abnormalities and the distribution of karyotype frequencies depending on the presence of an embryo in the gestational sac was carried out. Statistical analysis was performed using a chi-square test with a $p < 0.05$ significance level. The total frequency of chromosomal abnormalities in the study was 53.6 % (832/1551). The proportion of abnormal karyotypes in the AP and MA groups did not differ significantly and amounted to 57.1 % (152/266) and 52.9 % (680/1285) for AP and MA, respectively ($p = 0.209$). Sex chromosome aneuploidies and triploidies were significantly less common in the AP group than in the MA group (2.3 % (6/266) *vs* 6.8 % (88/1285), $p = 0.005$ and 4.9 % (13/266) *vs* 8.9 % (114/1285), $p = 0.031$, respectively). Tetraploidies were registered more frequently in AP compared to MA (12.4 % (33/266) *vs*. 8.2 % (106/1285), $p = 0.031$). The sex ratio among abortions with a normal karyotype was 0.54 and 0.74 for AP and MA, respectively. Thus, although the frequencies of some types of chromosomal pathology differ between AP and MA, the total frequency of chromosomal abnormalities in AP is not increased compared to MA, which indicates the need to search for the causes of AP at other levels of the genome organization, including microstructural chromosomal rearrangements, monogenic mutations, imprinting disorders, and epigenetic abnormalities.
Key words: anembryonic pregnancy; missed abortion; miscarriage; karyotype; chromosomal abnormalities; sex chromosomes; triploidy; tetraploidy.

For citation: Nikitina T.V., Sazhenova E.A., Tolmacheva E.N., Sukhanova N.N., Vasilyev S.A., Lebedev I.N. Comparative cytogenetics of anembryonic pregnancies and missed abortions in human. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):28-35. DOI 10.18699/VJGB-23-05

Сравнительная цитогенетика анэмбрионии и неразвивающейся беременности у человека

Т.В. Никитина ✉, Е.А. Саженова, Е.Н. Толмачева, Н.Н. Суханова, С.А. Васильев, И.Н. Лебедев

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ t.nikitina@medgenetics.ru

Аннотация. Невынашивание беременности является серьезной проблемой в репродукции человека, затрагивающей 10–15 % клинически распознаваемых беременностей. Среди случаев эмбриональной гибели можно выделить замершие (неразвивающиеся) беременности (НБ), при которых ультразвуковым признаком гибели эмбриона служит отсутствие сердцебиения, и анэмбрионии (АЭ) – отсутствие эмбриона в полости плодного мешка. Целью данного исследования было сравнение частоты хромосомных аномалий во внезародышевых тканях, выявляемых при стандартном цитогенетическом анализе материала спонтанных абортов, в зависимости от наличия или отсутствия эмбриона. Проведено ретроспективное исследование 1551 спонтанного абортуса, проанализированного с помощью стандартного цитогенетического исследования с 1990 по 2022 г. (266 случаев АЭ и 1285 случаев НБ) в НИИ медицинской генетики Томского НИМЦ. Выполнен сравнительный анализ частоты хромосомных аномалий и распределения частот кариотипов в зависимости от наличия эмбриона в полости плодного мешка. Статистический анализ проводили с использованием критерия хи-квадрат с уровнем значимости $p < 0.05$. Суммарно частота хромосомных аномалий в исследованной выборке составила 53.6 % (832/1551). Доля аномальных кариотипов в группах АЭ и НБ значимо не различалась и составила 57.1 %

(152/266) и 52.9 % (680/1285) для АЭ и НБ соответственно (p = 0.209). При НБ статистически значимо чаще встречались аномалии числа половых хромосом (6.8 % (88/1285) против 2.3 % (6/266), p = 0.005) и триплоидии (8.9 % (114/1285) против 4.9 % (13/266), p = 0.031). В то же время при отсутствии эмбриона статистически значимо чаще регистрировалась тетраплоидия (12.4 % (33/266) против 8.2 % (106/1285), p = 0.031). Соотношение полов (46,XY : 46,XX) среди абортусов с нормальным кариотипом составило 0.54 и 0.74 для АЭ и НБ соответственно. Таким образом, хотя частоты некоторых типов хромосомных аномалий различаются между АЭ и НБ, суммарная частота хромосомных аномалий при АЭ не повышена по сравнению с НБ, что свидетельствует о необходимости поиска причин АЭ на других уровнях организации генома, включая микроструктурные перестройки хромосом, моногенные мутации, нарушения импринтинга и аберрантные эпигенетические модификации.
Ключевые слова: анэмбриония; неразвивающаяся беременность; невынашивание беременности; кариотип; хромосомные аномалии; половые хромосомы; триплоидия; тетраплоидия.

Introduction

Miscarriage is one of the most common issues in human reproduction that results in embryonic or fetal death in 10 to 15 % of all clinically recognized pregnancies (Larsen et al., 2013). Cytogenetic studies reveal chromosomal abnormalities in 50–60 % of first trimester abortions (Menasha et al., 2005; van den Berg et al., 2012; Hardy et al., 2016; Soler et al., 2017; Wang et al., 2020; Wu et al., 2021), and in recent years, there has been an increasing amount of data about the association of miscarriage with copy number variations (CNV), gene mutations, methylation abnormalities and other epigenetic aberrations (Levy et al., 2014; Fu et al., 2018; Fan et al., 2020; Finley et al., 2022). Identification of embryo death causes is necessary to assess the miscarriage risk in subsequent pregnancies; in addition, uncovering a pathogenic factor is important for psychological condition of the couples.

Anembryonic pregnancy is the absence of an embryo in the gestational sac, and it is one of the earliest forms of miscarriage. In anembryonic pregnancy, a blastocyst is implanted into the uterine wall, a gestational sac is formed, but the embryo itself either does not develop initially, or its formation arrests at the earliest stages (no later than the 5th week of gestation), and then only extra-embryonic components of the conceptus continue to proliferate and grow.

As a rule, at around 6 weeks of gestation, the secondary yolk sac and the primary germ layers could be detected within the gestational sac by transvaginal ultrasound, and primitive cardiac tube could be detected during the 7th week. In early pregnancy loss there are several ultrasonography features: the absence of embryonic cardiac activity with a diameter of the gestational sac ≥25 mm, crown–rump length (CRL) ≥7 mm for a period of 6 weeks or more; the absence of an embryo and its cardiac activity 14 days after the detection of a gestational sac without a yolk sac; the absence of an embryo and its cardiac activity 11 days after the detection of a gestational sac with a yolk sac (Doubilet et al., 2013). Thus, ultrasound scanning makes it possible to differentiate two forms of early embryonic death: anembryonic pregnancy (AP) and missed abortion (MA). AP is diagnosed in the absence of an embryo and a secondary yolk sac in the cavity of the gestational sac, for a period of more than 7 weeks (Radzinsky et al., 2015); in addition, ultrasound criteria for AP are a gestational sac more than 13 mm without a yolk sac or more than 18 mm without an embryo. The absence of cardiac activity in the presence of an embryo is a sign of MA.

There are terminological inconsistencies, which make it difficult to compare the results of studies implemented in different centers. The ICD-10 uses the terms ‘blighted ovum’ and ‘missed abortion’, accepted many years ago (Robinson, 1975), which do not quite represent the clinical features found using the ultrasound examination (Farquharson et al., 2005). The European Society of Human Reproduction and Embryology (ESHRE) special group has proposed the terms ‘anembryonic (empty sac) miscarriage’ for a gestational sac ≥8 mm in diameter and without a yolk sac or embryo; ‘yolk sac miscarriage’ for a gestational sac with a yolk sac, but without an embryo; ‘embryonic miscarriage’ with an embryonic CRL of at least 7 mm without cardiac activity (Kolte et al., 2015). Thus, the diagnosis of AP includes both an empty gestational sac (empty sac) and a gestational sac with a yolk sac and without an embryo (yolk sac only).

The estimated frequency of AP among the first trimester pregnancy losses differs: from 16 % in early studies (Robinson, 1975), 22.6 % after IVF (Li et al., 2017), and up to 30–40 % in most studies (Lathi et al., 2007; Cheng et al., 2014; Ouyang et al., 2016; Yoneda et al., 2018). Despite the prevalence of AP, data on the frequency of chromosomal abnormalities in this pathology are contradictory. Intuitively, it seems that such early and pronounced violations, which lead to the developmental arrest of the embryo *per se* at the initial stages of its formation, should be associated with a significantly increased frequency and severity of chromosomal abnormalities. In some studies such association was found (Angiolucci et al., 2011). At the same time, most recent studies demonstrate either the absence of significant differences in the frequency of chromosomal abnormalities between the AP and MA groups (Lathi et al., 2007; Muñoz et al., 2010; Ljunger et al., 2011; Liu et al., 2015), or even a lower frequency of abnormal karyotypes in AP compared to MA (Ginsberg et al., 2001; Cheng et al., 2014; Li et al., 2017; Yoneda et al., 2018; Gu et al., 2021). Therefore, we consider it of current interest to study large samples of AP and MA cases in comparison with the published data. In this work, we studied the frequency and spectrum of chromosomal anomalies detected by cytogenetic analysis of 1551 cases of early miscarriage, depending on the presence or absence of an embryo.

Materials and methods

The object of this study was 1551 spontaneous abortions, karyotyped in the Cytogenetic Laboratory of the Research Institute of Medical Genetics of the Tomsk National Research Medical Center. Products of conception (POC) were obtained from gynecological clinics of Tomsk and Seversk, along with information regarding the patient’s age, woman’s obstetric

and gynecological history, and the number and outcomes of her previous pregnancies. The study was approved by the Biomedical Ethics Committee of the Research Institute of Medical Genetics of the Tomsk National Research Medical Center, Protocol 10, Feb. 15, 2021. Informed consents were obtained from all patients.

In most cases, abortion karyotypes were established using conventional GTG banding after long-term extra-embryonic fibroblast culture (90.9 %, 1410 samples) or direct preparations of the chorionic villi (1.9 %, 29 samples). Conventional comparative genomic hybridization (CGH) (3.5 %, 54 samples) and interphase fluorescence *in situ* hybridization (FISH) with centromere-enumeration probes (3.7 %, 58 samples) were performed in cases where traditional cytogenetic analysis failed. AP were diagnosed by ultrasound examination and included 266 (17.2 %) abortions (with absence of an embryo in the cavity of the gestational sac for more than 7 weeks, a gestational sac more than 13 mm without a yolk sac or more than 18 mm without an embryo). The other 1285 abortions (82.8 %) with an embryo were assigned to the MA group (where an embryonic pole was identified without cardiac activity).

The POC material, usually represented by the fragments of the gestational sac, was delivered to the laboratory in sterile saline, thoroughly washed from blood, and separated from decidual tissues. Methods of embryonic cells culture, chromosome preparations, cytogenetic techniques, FISH and CGH were performed as described previously (Lebedev, Nikitina, 2013).

The calculation of the statistical significance of differences between frequencies was performed using the χ^2 analyses; the normality of the distribution for quantitative indicators was checked using the Kolmogorov–Smirnov test; due to the differences from the normal distribution, comparisons between groups were performed using the nonparametric Mann–Whitney test. A significance level of $p < 0.05$ was applied for all tests. The sex ratio (SR) was calculated as the ratio of karyotypes 46,XY : 46,XX. Recurrent pregnancy loss (RPL) was defined as two or more consecutive miscarriages in a woman's obstetric history.

The study was performed at the Core Medical Genomics Facility of the Tomsk National Research Medical Center (NRMC) of the Russian Academy of Sciences using the resources of the bio-collection "Biobank of the population of Northern Eurasia" of the Research Institute of Medical Genetics, Tomsk NRMC.

Results

Table 1 shows the comparison of the demographic characteristics of the studied groups of abortions. The age of mothers and fathers, the number of woman's pregnancies and spontaneous abortions, and the proportion of couples with RPL in a woman's history did not differ significantly in the samples. However, the gestational age (both by the date of the last menstrual period and by ultrasound examination) in the AP group was significantly less than in the MA group.

In total, abnormal karyotypes were found in 53.6 % (832/1551) of abortions. Table 2 shows the karyotype frequencies. The rates of the different types of chromosomal abnormalities among pregnancy losses with and without an embryo were 52.9 % (680/1285) and 57.1 % (152/266) respectively, and did not differ significantly ($p = 0.209$). We found similar frequencies of chromosomal abnormalities between AP and MA in the autosomal trisomies (27.8 and 22.5 %), autosomal monosomies (1.5 and 0.6 %), structural aberrations (2.3 and 1.0 %) and combined anomalies that include combinations of different types of chromosomal aberrations in one abortion (4.9 and 4.2 %) (see Table 2). At the same time, numerical abnormalities of sex chromosomes in AP were three times less common than in MA (2.3 and 6.8 %, $p = 0.005$). This difference was even more pronounced for monosomy X: 0.8 % (2/266) and 5.0 % (64/1285), $p < 0.001$.

Triploidies occurred significantly less frequently, and tetraploidies occurred significantly more frequently in abortions without embryo in comparison with embryonic miscarriages

Table 1. Comparison of demographic parameters of the AP and MA groups

Parameter	Total	AP	MA	*p*
Maternal age, years	28.6 ± 6.23 (24.0–33.0; 28.0)	28.3 ± 6.14 (24.0–32.0; 28.0)	28.5 ± 6.18 (24.0–33.0; 28.0)	0.491
Paternal age, years	30.9 ± 6.69 (26.0–35.0; 30.0)	30.5 ± 6.52 (26.0–35.0; 30.0)	30.9 ± 6.64 (26.0–35.0; 30.0)	0.405
Gestational age, weeks*	9.4 ± 2.23 (8.0–11.0; 9.1)	9.0 ± 2.13 (7.5–10.1; 9.0)	9.5 ± 2.24 (8.0–11.0; 9.3)	**0.001**
Gestational age in ultrasound, weeks	7.3 ± 1.84 (6.0–8.5; 7.0)	6.6 ± 1.62 (5.5–7.6; 6.5)	7.5 ± 1.84 (6.0–8.5; 7.3)	**<0.001**
No. of pregnancies	2.9 ± 2.20 (1.0–4.0; 2.0)	3.2 ± 2.42 (1.0–4.5; 2.0)	2.9 ± 2.18 (1.0–4.0; 2.0)	0.085
No. of miscarriages	1.6 ± 1.05 (1.0–2.0; 1.0)	1.6 ± 1.13 (1.0–2.0; 1.0)	1.6 ± 1.04 (1.0–2.0; 1.0)	0.633
RPL in anamnesis, %	38.7	35.5	39.4	0.278

Note. Mean ± standard deviation (lower and upper quartile; median); * gestational age calculated by the date of the last menstrual period. Significantly different rates are in bold.

Table 2. Karyotypes of abortions in the AP and MA groups

Karyotypes	Total	AP	MA	*p*
	n = 1551	*n* = 266	*n* = 1285	
46,XX	422 (27.2)	74 (27.8)	348 (27.1)	0.806
46,XY	297 (19.1)	40 (15.0)	257 (20.0)	0.062
Sex chromosome abnormality	94 (6.1)	6 (2.3)	88 (6.8)	**0.005**
Autosomal trisomy	363 (23.4)	74 (27.8)	289 (22.5)	0.062
Triploidy	127 (8.2)	13 (4.9)	114 (8.9)	**0.031**
Tetraploidy	139 (9.0)	33 (12.4)	106 (8.2)	**0.031**
Structural rearrangements	19 (1.2)	6 (2.3)	13 (1.0)	0.094
Autosomal monosomy	12 (0.8)	4 (1.5)	8 (0.6)	0.136
Others	11 (0.7)	3 (1.1)	8 (0.6)	0.372
Combined*	67 (4.3)	13 (4.9)	54 (4.2)	0.618

Note. Percentages are given in parentheses. * Combined – combination of different types of abnormalities. Significantly different rates are in bold.

(4.9 and 8.9 %, $p = 0.031$; 12.4 and 8.2 %, $p = 0.031$ in AP and MA, respectively). Since some of the tetraploid karyotypes in mosaic form may represent cultural artifacts, we reexamined some tetraploid samples using FISH in non-cultured tissues. The frequency of FISH-confirmed tetraploidies showed even more statistically significant differences: 14/266 (5.3 %) at AP *vs* 22/1285 (1.7 %) in MA ($p < 0.001$).

Among the abortions with normal karyotype, the SR was 0.54 for AP and 0.74 for MA; there were no significant differences in the distribution of 46,XX and 46,XY karyotypes ($p = 0.142$).

Discussion

The aim of this study was to compare the frequency of chromosomal abnormalities in pregnancy losses with and without an embryo (MA and AP). In this study, we analyzed a large sample of miscarriages (1551 abortions) and did not find significant differences in the chromosomal abnormality rates between the AP and MA groups (57.1 and 52.9 % respectively, $p = 0.209$). An analysis of previously published comparative studies showed conflicting results regarding the correlation between the karyotype and the presence of an embryo in the gestational sac (Table 3).

As Angiolucci et al. (2011) have reported, the frequency of abnormal karyotypes positively correlated with the diagnosis of AP, however, the authors found this correlation in relation to dead embryos with normal ultrasound signs. We recalculated the data of Angiolucci et al. (2011) using comparison of AP with the total group of abortions with the presence of an embryo, and did not find a statistically significant correlation between the frequency of chromosomal abnormalities and presence/absence of an embryo ($p = 0.381$) (see Table 3). In some studies (Lathi et al., 2007; Muñoz et al., 2010; Ljunger et al., 2011; Liu et al., 2015), no association between the frequency of abnormal karyotypes and the presence or absence of an embryo was found, as in the present work. However, most recent studies of relatively large samples reveal a negative correlation between the absence of an embryo and the chromosomal aberrations rate (Cheng et al., 2014; Li et al., 2017; Yoneda et al., 2018; Gu et al., 2021). This discrepancy may be due to differences in karyotype evaluation methods, sample sizes, and population structure (see Table 3). In addition, some studies were carried out on biased samples, such as women with infertility (Li et al., 2017), or from high-risk groups for aneuploidy (Muñoz et al., 2010), or the average mother age in the sample was more than 35 years (Ginsberg et al., 2001; Muñoz et al., 2010; Angiolucci et al., 2011). Nevertheless, the results of the analysis of the published data, starting from 2001 (since small samples were examined in earlier studies), indicate that the frequency of chromosomal abnormalities in AP is lower than in miscarriage with an embryo (44.4 % (813/1833) *vs* 59.3 % (2701/4558), respectively, $p < 0.001$) (see Table 3). Similar conclusions were obtained in the meta-analysis of Huang et al. (2019) ($p = 0.03$, OR = 0.68, 95 % CI = 0.48–0.97).

A possible explanation may be that specific types of genetic aberrations are critical for different stages of the early embryonic development. This assumption is supported by the different frequency of chromosomal abnormalities for two different pathological phenotypes included in the diagnosis of AP: empty sac and yolk sac only. Thus, in the AP subgroup with an empty sac, the frequency of chromosomal abnormalities was significantly lower than in the AP subgroup with the yolk sac only (Ouyang et al., 2016; Li et al., 2017; Gu et al., 2021) (see Table 3).

The presence of a yolk sac in a gestational sac without an embryo means that the disorders appear after the segregation of the hypoblast and epiblast, which occurs 6–7 days after fertilization. The hypoblast, which gives rise to the endoderm of the yolk sac, continues to develop, while the epiblast,

Table 3. Comparative frequencies of chromosomal abnormalities in AP and MA in various studies

Reference	*N*	Method	Aneuploidy rate in the AP group	Aneuploidy rate in the MA group	*p*
This study	1551	Culture, G-banding	57.1 % (152/266)	52.9 % (680/1285)	0.209
Romanova, 2022	273	Direct, Q-banding	25.7 % (9/35)	67.2 % (160/238)	<0.050
Gu et al., 2021	1102 (887)	CMA and FISH	47.1 % (64/136)	62.1 % (466/751)	0.001
Yoneda et al., 2018	151 (141)	Culture, G-banding	50.8 % (32/53)	77.3 % (68/88)	<0.001
Li et al., 2017	2172 after IVF	CGH and FISH	28.1 % (138/491) empty sac	52.2 % (641/1227)	<0.001
			43.4 % (197/454) yolk sac only		0.002
			35.4 % (335/945) AP total		<0.001
Liu et al., 2015	183	Culture, G-banding	56.1 % (32/57)	61.1 % (77/126)	0.526
Cheng et al., 2014	223	Culture, G-banding	46.3 % (37/80)	61.5 % (88/143)	0.030
Angiolucci et al., 2011	156	G-banding	72.2 % (13/18)	33.8 % (23/68)* 61.6 % (85/138)**	0.006 0.381
Ljunger et al., 2011	259 (239)	Direct, G-banding	54.5 % (48/88)	65.6 % (99/151)	0.092
Muñoz et al., 2010	185	Direct, G-banding	60.5 % (26/43)	67.6 % (96/142)	0.387
Lathi et al., 2007	272	Culture, G-banding	58.2 % (53/91)	68.0 % (123/181)	>0.050
Ginsberg et al., 2001	129	G-banding	57.1 % (12/21)	90.7 % (98/108)	<0.001
Total	6411	–	44.4 % (813/1833)	59.3 % (2701/4558)	<0.001

Note. *N* is the sample size, in brackets is the total number of compared cases of AP and MA; CMA, chromosomal microarray analysis; CGH, comparative genomic hybridization; FISH, fluorescence *in situ* hybridization.
* Relative to abortions with a normal phenotype on ultrasound examination; ** relative to abortions with the presence of an embryo on ultrasound examination.

which gives rise to the three germ layers of the embryo itself (endoderm, ectoderm, and mesoderm), is blocked. An empty gestational sac means that the abnormalities appeared before the separation of the inner cell mass into hypoblast and epiblast, i. e. during implantation (Boss et al., 2018). At such an early stage, the influence of non-genetic factors is unlikely to be significant. The lower frequency of chromosomal abnormalities in such embryos may be due to the fact that at such early stages of development, damage of the activity of genes important in early embryogenesis due to point mutations, CNV, or epigenetic anomalies is more critical than a change in the gene dosage due to aneuploidy.

Considering the predominant contribution of genetic causes (compared to maternal or environmental causes) to a very early arrest of embryonic development, the lower frequency of chromosomal abnormalities in the absence of an embryo in the gestational sac makes it promising to search for genetic aberrations of the sub-chromosomal level and epigenetic anomalies in AP cases (Lebedev et al., 2013). Thus, chromosomal microarray analysis revealed a greater number of CNV in AP compared to MA (299 and 132, respectively), and in AP among pathogenic rearrangements 54.3 % deletions and 45.7 % duplications were found, whereas in MA only duplications were found (Savchenko et al., 2018). Interestingly, the set of genes in CNV also differed: in AP, the genes responsible for basic biological processes, such as migration, cell contacts, and adhesion, were more often affected, while in MA, the genes responsible for morphogenesis were affected.

We found different frequencies of some types of chromosomal abnormalities between abortions with and without an embryo. In our sample, sex chromosome aneuploidies (especially the 45,X) were less common in AP than in MA (see Table 2). Frequency of the 45,X karyotype has been found to be significantly higher in miscarriages with an embryo in most published comparative studies (Minelli et al., 1993; Muñoz et al., 2010; Cheng et al., 2014; Liu et al., 2015; Veropotvelyan, Kodunov, 2015; Li et al., 2017; Ozawa et al., 2019; Gu et al., 2021). These results indicate that monosomy X does not have

a noticeable negative effect on the early development of the embryo *per se*, and such embryos die at later stages, possibly due to a failure of trophoblast function (Ahern et al., 2022).

Triploidy is another type of chromosomal abnormality, which is more common in embryonic miscarriages than in anembryonic ones. Previous studies suggested that the majority of triploidies are the result of fertilization errors leading to either diandry (the presence of two sets of paternal chromosomes) or digyny (the presence of two maternal sets) (Thaker, 2005). Due to the phenomenon of chromosomal imprinting, paternal chromosomes contribute to the preferential proliferation of trophoblast tissues. Perhaps it is diandric triploidy that leads to the AP phenotype, but this assumption needs to be verified.

Common feature for both of the above mentioned types of chromosomal abnormalities is that the mechanism of their origin is not associated with meiotic nondisjunction in oocytes. It is known that cases of X-chromosome monosomy are most often caused by errors in paternal meiosis (Hassold et al., 1988; Segawa et al., 2017), and triploidies are caused mostly by fertilization errors. Therefore, the rate of these types of karyotype abnormalities is increased among abortions from young mothers in comparison with older mothers (Soler et al., 2017; Wang et al., 2020; Gu et al., 2021). Since the mother's age was similar in our AP and MA samples, the higher rate of monosomy X and triploidy in MA in comparison with AP supports the assumption that embryos with these karyotype abnormalities survive better.

We found a higher frequency of tetraploidy in the AP group, which is consistent with the data of (Veropotvelyan, Kodunov, 2015; Ozawa et al., 2019) and implies an unfavorable influence of the tetraploid karyotype, leading to an earlier termination of embryo development.

We found that sex ratio (SR) in abortions with normal karyotype deviates from the expected SR and constitutes 0.74 for MA and 0.54 for AP. Although the differences between the groups did not reach a statistically significant level ($p = 0.142$), they are consistent with the data obtained earlier in our laboratory using significantly smaller samples. In the study (Evdokimova et al., 2000) it was shown that the proportion of 46,XY embryos inversely correlates with the severity of developmental disorders: the SR was 0.77 for spontaneous abortions without significant intrauterine delay of development; 0.60 for MA and 0.31 for AP (compared to 1.10 for control group of induced abortion). One of the reasons for the biased SR may be maternal cell contamination (MCC) of extra-embryonic cell cultures. But since both AP and MA samples were analyzed concomitantly, and the frequency of MCC was low (Nikitina et al., 2005), this equalizes the possible effect of maternal contamination on SR in our study. Interestingly, a large-scale study of SR in early human development (from conception to birth) showed that SR decreases in the first week after conception (due to excess male mortality) and then increases for at least 10–15 weeks (due to excess female mortality) (Orzack et al., 2019). Thus, the excess of female embryo loss in the first trimester of pregnancy probably represents a real phenomenon.

The development of cell-based technologies offers a unique opportunity to study the biological mechanisms that lead to embryogenesis failure. Thus, induced pluripotent stem cells (iPSCs) reproduce the characteristics of embryonic stem cells, including unlimited proliferative capacity and the ability to differentiate into derivatives of three germ layers (pluripotency). It has been shown that iPSCs can be derived from trophoblast tissues not only from embryos with a normal karyotype, but also from embryos with some chromosomal aneuploidies (for example, monosomy X and trisomy 13) (Parveen et al., 2017; Long et al., 2020). If it is possible to reprogram trophoblast cells and obtain iPSC lines from anembryonic cases, this will open up the possibility to study the processes in the derivatives of various germ layers leading to an early developmental arrest of the embryo.

Conclusion

We found that the pattern of chromosomal abnormalities partly differs between AP and MA, and the presence of an embryo is positively correlated with sex chromosome aneuploidy and triploidy, while the absence of an embryo is positively correlated with tetraploidy. At the same time, the total frequency of chromosomal abnormalities in AP and MA did not differ, which indicates the need to search for the causes of AP at other levels of genome organization, including microstructural chromosomal rearrangements, monogenic mutations, imprinting disorders, and other aberrant epigenetic modifications of the genome.

References

Ahern D.T., Bansal P., Faustino I., Kondaveeti Y., Glatt-Deeley H.R., Banda E.C., Pinter S.F. Monosomy X in isogenic human iPSC-derived trophoblast model impacts expression modules preserved in human placenta. *Proc. Natl. Acad. Sci. USA*. 2022;119(40): e2211073119. DOI 10.1073/pnas.2211073119.

Angiolucci M., Murru R., Melis G., Carcassi C., Mais V. Association between different morphological types and abnormal karyotypes in early pregnancy loss. *Ultrasound Obstet. Gynecol.* 2011;37(2):219-225. DOI 10.1002/uog.7681.

Boss A.L., Chamley L.W., James J.L. Placental formation in early pregnancy: how is the centre of the placenta made? *Hum. Reprod. Update*. 2018;24(6):750-760. DOI 10.1093/humupd/dmy030.

Cheng H.H., Ou C.Y., Tsai C.C., Chang S.D., Hsiao P.Y., Lan K.C., Hsu T.Y. Chromosome distribution of early miscarriages with present or absent embryos: female predominance. *J. Assist. Reprod. Genet*. 2014;31(8):1059-1064. DOI 10.1007/s10815-014-0261-9.

Doubilet P.M., Benson C.B., Bourne T., Blaivas M., Society of Radiologists in Ultrasound Multispecialty Panel on Early First Trimester Diagnosis of Miscarriage and Exclusion of a Viable Intrauterine Pregnancy, Barnhart K.T., Benacerraf B.R., Brown D.L., Filly R.A., Fox J.C., Goldstein S.R., Kendall J.L., Lyons E.A., Porter M.B., Pretorius D.H., Timor-Tritsch I.E. Diagnostic criteria for nonviable pregnancy early in the first trimester. *N. Engl. J. Med.* 2013;369(15): 1443-1451. DOI 10.1056/NEJMra1302417.

Evdokimova V.N., Nikitina T.V., Lebedev I.N., Sukhanova N.N., Nazarenko S.A. About the sex ratio in connection with early embryonic mortality in man. *Russ. J. Developmental Biology.* 2000;31(4):251-257.

Fan L., Wu J., Wu Y., Shi X., Xin X., Li S., Zeng W., Deng D., Feng L., Chen S., Xiao J. Analysis of chromosomal copy number in first-trimester pregnancy loss using next-generation sequencing. *Front. Genet.* 2020;11:545856. DOI 10.3389/fgene.2020.545856.

Farquharson R.G., Jauniaux E., Exalto N., ESHRE Special Interest Group for Early Pregnancy (SIGEP). Updated and revised nomenclature for description of early pregnancy events. *Hum. Reprod*. 2005;20(11):3008-3011. DOI 10.1093/humrep/dei167.

Finley J., Hay S., Oldzej J., Meredith M.M., Dzidic N., Slim R., Aradhya S., Hovanes K., Sahoo T. The genomic basis of sporadic and recurrent pregnancy loss: a comprehensive in-depth analysis of 24,900 miscarriages. *Reprod. Biomed. Online.* 2022;45(1):125-134. DOI 10.1016/j.rbmo.2022.03.014.

Fu M., Mu S., Wen C., Jiang S., Li L., Meng Y., Peng H. Whole exome sequencing analysis of products of conception identifies novel mutations associated with missed abortion. *Mol. Med. Rep.* 2018;18(2): 2027-2032. DOI 10.3892/mmr.2018.9201.

Ginsberg N.A., Strom C., Verlinsky Y. Crown-rump lengths in missed miscarriages and trisomy 21. *Ultrasound Obstet. Gynecol.* 2001; 18(5):488-490. DOI 10.1046/j.0960-7692.2001.00571.x.

Gu C., Li K., Li R., Li L., Li X., Dai X., He Y. Chromosomal aneuploidy associated with clinical characteristics of pregnancy loss. *Front. Genet.* 2021;12:667697. DOI 10.3389/fgene.2021.667697.

Hardy K., Hardy P.J., Jacobs P.A., Lewallen K., Hassold T.J. Temporal changes in chromosome abnormalities in human spontaneous abortions: results of 40 years of analysis. *Am. J. Med. Genet. A.* 2016; 170(10):2671-2680. DOI 10.1002/ajmg.a.37795.

Hassold T., Benham F., Leppert M. Cytogenetic and molecular analysis of sex-chromosome monosomy. *Am. J. Hum. Genet*. 1988;42(4): 534-541.

Huang J., Zhu W., Tang J., Saravelos S.H., Poon L.C.Y., Li T.C. Do specific ultrasonography features identified at the time of early pregnancy loss predict fetal chromosomal abnormality? – A systematic review and meta-analysis. *Genes Dis.* 2019;6(2):129-137. DOI 10.1016/j.gendis.2018.10.001.

Kolte A.M., Bernardi L.A., Christiansen O.B., Quenby S., Farquharson R.G., Goddijn M., Stephenson M.D. ESHRE Special Interest Group, Early Pregnancy. Terminology for pregnancy loss prior to viability: a consensus statement from the ESHRE early pregnancy special interest group. *Hum. Reprod.* 2015;30(3):495-498. DOI 10.1093/humrep/deu299.

Larsen E.C., Christiansen O.B., Kolte A.M., Macklon N. New insights into mechanisms behind miscarriage. *BMC Med*. 2013;11:154. DOI 10.1186/1741-7015-11-154.

Lathi R.B., Mark S.D., Westphal L.M., Milki A.A. Cytogenetic testing of anembryonic pregnancies compared to embryonic missed abortions. *J. Assist. Reprod. Genet.* 2007;24(11):521-524. DOI 10.1007/s10815-007-9166-1.

Lebedev I.N., Kashevarova A.A., Skryabin N.A., Nikitina T.V., Lopatkina M.E., Melnikov A.A., Sazhenova E.A., Ivanova T.V., Evtushenko I.D. Array-based comparative genomic hybridization (array-CGH) in analysis of chromosomal aberrations and CNV in blighted ovum pregnancies. *Zhurnal Akusherstva i Zhenskikh Bolezney = Journal of Obstetrics and Women's Diseases*. 2013;62(2):117-125. (in Russian)

Lebedev I.N., Nikitina T.V. Cytogenetics of Human Embryonic Development Disorders (Heredity and Health). Tomsk: Pechatnaya Manufaktura Publ., 2013. (in Russian)

Levy B., Sigurjonsson S., Pettersen B., Maisenbacher M.K., Hall M.P., Demko Z., Lathi R.B., Tao R., Aggarwal V., Rabinowitz M. Genomic imbalance in products of conception: single-nucleotide polymorphism chromosomal microarray analysis. *Obstet. Gynecol*. 2014; 124(2 Pt.1):202-209. DOI 10.1097/aog.0000000000000325.

Li X., Ouyang Y., Yi Y., Tan Y., Lu G. Correlation analysis between ultrasound findings and abnormal karyotypes in the embryos from early pregnancy loss after in vitro fertilization–embryo transfer. *J. Assist. Reprod. Genet.* 2017;34(1):43-50. DOI 10.1007/s10815-016-0821-2.

Liu Y., Liu Y., Chen H., Du T., Tan J., Zhang J. The frequencies of the presence of embryonic pole and cardiac activity in early miscarriages with abnormal karyotypes. *Clin. Exp. Obstet. Gynecol*. 2015; 42(4):490-494.

Ljunger E., Stavreus-Evers A., Cnattingius S., Ekbom A., Lundin C., Annéren G., Sundström-Poromaa I. Ultrasonographic findings in spontaneous miscarriage: relation to euploidy and aneuploidy. *Fertil. Steril*. 2011;95(1):221-224. DOI 10.1016/j.fertnstert.2010.06.018.

Long P., Liu Z., Wu B., Chen J., Sun C., Wang F., Huang Y., Chen H., Li Q., Ma Y. Generation of an induced pluripotent stem cell line from chorionic villi of a Patau syndrome spontaneous abortion. *Stem Cell Res*. 2020;45:101789. DOI 10.1016/j.scr.2020.101789.

Menasha J., Levy B., Hirschhorn K., Kardon N.B. Incidence and spectrum of chromosome abnormalities in spontaneous abortions: new insights from a 12-year study. *Genet. Med.* 2005;7(4):251-263. DOI 10.1097/01.gim.0000160075.96707.04.

Minelli E., Buchi C., Granata P., Meroni E., Righi R., Portentoso P., Giudici A., Ercoli A., Sartor M.G., Rossi A. Cytogenetic findings in echographically defined blighted ovum abortions. *Ann. Genet*. 1993;36(2):107-110.

Muñoz M., Arigita M., Bennasar M., Soler A., Sanchez A., Borrell A. Chromosomal anomaly spectrum in early pregnancy loss in relation to presence or absence of an embryonic pole. *Fertil. Steril*. 2010; 94(7):2564-2568. DOI 10.1016/j.fertnstert.2010.04.011.

Nikitina T.V., Lebedev I.N., Sukhanova N.N., Sazhenova E.A., Nazarenko S.A. A mathematical model for evaluation of maternal cell contamination in cultured cells from spontaneous abortions: significance for cytogenetic analysis of prenatal selection factors. *Fertil. Steril*. 2005;83(4):964-972. DOI 10.1016/j.fertnstert.2004.12.009.

Orzack S.H., Stubblefield J.W., Akmaev V.R., Colls P., Munné S., Scholl T., Steinsaltz D., Zuckerman J.E. The human sex ratio from conception to birth. *Proc. Natl. Acad. Sci. USA*. 2015;112(16): E2102-E2111. DOI 10.1073/pnas.1416546112.

Ouyang Y., Tan Y., Yi Y., Gong F., Lin G., Li X., Lu G. Correlation between chromosomal distribution and embryonic findings on ultrasound in early pregnancy loss after IVF-embryo transfer. *Hum. Reprod*. 2016;31(10):2212-2218. DOI 10.1093/humrep/dew201.

Ozawa N., Ogawa K., Sasaki A., Mitsui M., Wada S., Sago H. Maternal age, history of miscarriage, and embryonic/fetal size are associated with cytogenetic results of spontaneous early miscarriages. *J. Assist. Reprod. Genet.* 2019;36(4):749-757. DOI 10.1007/s10815-019-01415-y.

Parveen S., Panicker M.M., Gupta P.K. Generation of an induced pluripotent stem cell line from chorionic villi of a Turner syndrome spontaneous abortion. *Stem Cell Res.* 2017;19:12-16. DOI 10.1016/j.scr.2016.12.016.

Radzinsky V.E., Makletsova S.A., Aleev I.A., Rudneva O.D., Ryabinkina T.S. Missed miscarriage. Guidelines by MARS Reproductive Health Professional Medical Association. Moscow: StatusPraesens Editorial Office, 2015. (in Russian)

Robinson H.P. The diagnosis of early pregnancy failure by sonar. *Br. J. Obstet. Gynaecol.* 1975;82(11):849-857. DOI 10.1111/j.1471-0528.1975.tb00588.x.

Romanova O.A. Endometrium immunoreactivity in missed miscarriage associated with chromosome aberrations in the chorion. Cand. Med. Sci. Diss. St. Petersburg, 2022. (in Russian)

Savchenko R.R., Kashevarova A.A., Skryabin N.A., Zhigalina D.I., Lopatkina M.E., Nikitina T.V., Vasiliev S.A., Lebedev I.N. Analysis of CNVs in anembryonic pregnancy and missed abortions. *Meditsinskaya Genetika = Medical Genetics.* 2018;17(3):49-54. DOI 10.25557/2073-7998.2018.03.49-54. (in Russian)

Segawa T., Kuroda T., Kato K., Kuroda M., Omi K., Miyauchi O., Watanabe Y., Okubo T., Osada H., Teramoto S. Cytogenetic analysis of the retained products of conception after missed abortion following blastocyst transfer: a retrospective, large-scale, single-centre study. *Reprod. Biomed. Online.* 2017;34(2):203-210. DOI 10.1016/j.rbmo.2016.11.005.

Soler A., Morales C., Mademont-Soler I., Margarit E., Borrell A., Borobio V., Muñoz M., Sánchez A. Overview of chromosome abnormalities in first trimester miscarriages: a series of 1,011 consecutive chorionic villi sample karyotypes. *Cytogenet. Genome Res*. 2017; 152(2):81-89. DOI 10.1159/000477707.

Thaker H.M. The partly molar pregnancy that is not a partial mole. *Pediatr. Dev. Pathol.* 2005;8(2):146-147. DOI 10.1007/s10024-005-2164-3.

van den Berg M.M., van Maarle M.C., van Wely M., Goddijn M. Genetics of early miscarriage. *Biochim. Biophys. Acta.* 2012;1822(12): 1951-1959. DOI 10.1016/j.bbadis.2012.07.001.

Veropotvelyan N.P., Kodunov L.O. The features ratio of chromosomal abnormalities and terms of persistence of gestational sacs among undeveloping pregnancies in cases without embryo (anembrioniya) and with the existing embryo: analysis of 1328 cases. *Zdorov'e Zhenshchiny = Health of Woman*. 2015;5(101):74-80. DOI 10.15574/HW.2015.101.74. (in Russian)

Wang H., Yuan D., Wang S., Luo L., Zhang Y., Ye J., Zhu K. Cytogenetic and genetic investigation of miscarriage cases in Eastern China. *J. Matern. Fetal. Neonatal. Med.* 2020;33(20):3385-3390. DOI 10.1080/14767058.2019.1572738.

Wu X., Su L., Xie X., He D., Chen X., Wang M., Wang L., Zheng L., Xu L. Comprehensive analysis of early pregnancy loss based on cytogenetic findings from a tertiary referral center. *Mol. Cytogenet.* 2021;14(1):56. DOI 10.1186/s13039-021-00577-8.

Yoneda S., Shiozaki A., Yoneda N., Sameshima A., Ito M., Shima T., Nakashima A., Yoshino O., Kigawa M., Takamori R., Shinagawa Y., Saito S. A yolk sac larger than 5 mm suggests an abnormal fetal karyotype, whereas an absent embryo indicates a normal fetal karyotype. *J. Ultrasound Med*. 2018;37(5):1233-1241. DOI 10.1002/jum.14467.

ORCID ID

T.V. Nikitina orcid.org/0000-0002-4230-6855
E.A. Sazhenova orcid.org/0000-0003-3875-3932
E.N. Tolmacheva orcid.org/0000-0002-0716-4302
S.A. Vasilyev orcid.org/0000-0002-5301-070X
I.N. Lebedev orcid.org/0000-0002-0482-8046

Acknowledgements. The study was carried out as part of the state assignment No. 122020300041-7.

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

Received September 30, 2022. Revised November 14, 2022. Accepted November 18, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):36-45
DOI 10.18699/VJGB-23-06

Original Russian text https://vavilovj-icg.ru/

Structure and origin of Tuvan gene pool according to autosome SNP and Y-chromosome haplogroups

V.A. Stepanov, N.A. Kolesnikov, L.V. Valikhova, A.A. Zarubin, I.Yu. Khitrinskaya, V.N. Kharkov

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
Vladimir-kharkov@medgenetics.ru

Abstract. Tuvans are one of the most compactly living peoples of Southern Siberia, settled mainly in the territory of Tuva. The gene pool of the Tuvans is quite isolated, due to endogamy and a very low frequency of interethnic marriages. The structure of the gene pool of the Tuvans and other Siberian populations was studied using a genome-wide panel of autosomal single nucleotide polymorphic markers and Y-chromosome markers. The results of the analysis of the frequencies of autosomal SNPs by various methods, the similarities in the composition of the Y-chromosome haplogroups and YSTR haplotypes show that the gene pool of the Tuvans is very heterogeneous in terms of the composition of genetic components. It includes the ancient autochthonous Yeniseian component, which dominates among the Chulym Turks and Kets, the East Siberian component, which prevails among the Yakuts and Evenks, and the Far Eastern component, the frequency of which is maximum among the Nivkhs and Udeges. Analysis of the composition of IBD-blocks on autosomes shows the maximum genetic relationship of the Tuvans with the Southern Altaians, Khakas and Shors, who were formed during the settlement of the Turkic groups of populations on the territory of the Altai-Sayan region. A very diverse composition of the Tuvan gene pool is shown for various sublines of Y-chromosomal haplogroups, most of which show strong ethnic specificity. Phylogenetic analysis of individual Y-chromosome haplogroups demonstrates the maximum proximity of the gene pool of the Tuvans with the Altaians, Khakas and Shors. Differences in frequencies of Y-chromosome haplogroups between the Todzhans and Tuvans and a change in the frequencies of haplogroups from south to north associated with the East Asian component were found. The majority of the most frequent Y-chromosome haplogroups in the Tuvans demonstrate the founder effect, the formation age of which is fully consistent with the data on their ethnogenesis.
Key words: gene pool; human population; genetic diversity; genetic components; Y-chromosome; Tuvans.

For citation: Stepanov V.A., Kolesnikov N.A., Valikhova L.V., Zarubin A.A., Khitrinskaya I.Yu., Kharkov V.N. Structure and origin of Tuvan gene pool according to autosome SNP and Y-chromosome haplogroups. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):36-45. DOI 10.18699/VJGB-23-06

Структура и происхождение генофонда тувинцев по данным аутосомных SNP и гаплогрупп Y-хромосомы

В.А. Степанов, Н.А. Колесников, Л.В. Валихова, А.А. Зарубин, И.Ю. Хитринская, В.Н. Харьков

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
Vladimir-kharkov@medgenetics.ru

Аннотация. Тувинцы – один из наиболее компактно проживающих народов Южной Сибири, расселенный в основном на территории Тывы. Генофонд тувинцев является достаточно обособленным за счет эндогамии и очень низкой частоты межнациональных браков. Исследована структура генофонда тувинцев и других сибирских популяций по полногеномной панели аутосомных однонуклеотидных полиморфных маркеров и маркерам Y-хромосомы. Результаты анализа частот аутосомных SNP различными методами, сходства по составу гаплогрупп Y-хромосомы и YSTR-гаплотипов показывают, что генофонд тувинцев очень гетерогенен по составу генетических компонентов. Он включает в себя древний автохтонный енисейский компонент, доминирующий у чулымских тюрков и кетов, восточносибирский, преобладающий у якутов и эвенков, и дальневосточный, частота которого максимальна у нивхов и удэгейцев. Анализ состава IBD-блоков на аутосомах демонстрирует максимальное генетическое родство тувинцев с южными алтайцами, хакасами и шорцами, которые формировались при расселении тюркских групп популяций на территории Алтае-Саянского региона. Выявлен очень разнообразный состав генофонда тувинцев по различным сублиниям Y-хромосомных гаплогрупп, большинство из которых показывают сильную этническую специфичность. Филогенетический анализ отдельных Y-хромосомных гаплогрупп демонстрирует максимальную близость генофонда тувинцев с алтайцами, хакасами и шорцами. Внутри тувинского этноса обнаружены значительные различия между выборками из западных, южных и восточных районов Тывы по доле монгольского и енисейского генетических

компонентов. Генетическое разнообразие тувинцев по Y-хромосомным гаплогруппам и максимально разнородный состав генетических компонентов свидетельствуют о самом высоком разнообразии тувинского генофонда по сравнению со всеми коренными народами Сибири. Обнаружены различия по частотам гаплогрупп Y-хромосомы между тоджинцами и тувинцами и изменение частот гаплогрупп с юга на север, связанных с восточноазиатским компонентом. Большинство наиболее частых гаплогрупп Y-хромосомы у тувинцев демонстрирует эффект основателя, возраст формирования которых полностью согласуется с данными об их этногенезе.
Ключевые слова: генофонд; популяции человека; генетическое разнообразие; генетические компоненты; Y-хромосома; тувинцы.

Introduction

From the point of view of studying population and evolutionary genetic processes, analyzing genetic diversity, and reconstructing the genetic history of populations, the gene pool of the indigenous population of Southern Siberia is a unique system. The problems related to the analysis of the composition and ratio of various substrate components among the Siberian peoples, despite the high level of study, have a number of unanswered questions. In this regard, genetics provides the richest opportunities for studying these problems, since the development of new approaches to the analysis of the population gene pool makes it possible to bring ethnogenetic studies to a completely new level. Modern methods used in molecular genetic research and new bioinformatic developments make it possible to reliably identify various ancestral genetic components in the gene pool of various peoples and individuals.

One of the most important problems of ethnology and anthropology of the population of Southern Siberia is the issue of the formation of indigenous ethnic groups, in the solution of which, at present, methods of analyzing genomic data play an important role. The gene pool of the indigenous population of this region was formed due to the long-term and multi-stage mixing of a large number of local gene pools of various tribes of Caucasoid and Mongoloid origin. The indigenous ethnic groups of Southern Siberia are characterized by various anthropological types, complex ethnic and demographic history. The mixture of numerous Turkic, Mongolian, Yeniseian, Samoyed and Ugric groups based on the genetic substrate of the ancient Indo-European tribes and taiga Mongoloids formed as a result a motley picture of the genetic diversity of the population of this region (Gene Pool of the Population of Siberia, 2003).

The processes of merging and assimilation with the participation of various migration flows played an important role in the formation of modern Turkic-speaking populations of Southern Siberia, especially the Tuvans. In the era of the Eneolithic, Bronze and Early Iron Ages, the territory of Tuva was part of the habitat of the ancient Caucasoid population, which later developed the cultures of the Scythian-Siberian world (Alekseev, 1984). The penetration of the Central Asian Mongoloid component into the territory of Southern Siberia dates back to the VII–VI centuries BC. The appearance of the forest, taiga Mongoloid component also dates back to approximately the same time (Kiselev, 1951). Gradually, there was an increase in the Mongoloid component, from the predominance of the Caucasoid in the Scythian time to the formation in the XIII–XIV centuries AD of the modern Central Asian anthropological type of the Tuvans (Debets, 1948).

Tuvans are one of the most compactly living peoples of Russia, settled mainly in the territory of Tuva. In Russia, according to the All-Russian Population Census of 2010, the number of the Tuvans is 263,934 people. At the same time, the gene pool of the Tuvans is relatively isolated, due to endogamy and a very low frequency of interethnic marriages (Puzyrev et al., 1999; Kucher et al., 2003). The heterogeneity of the tribal composition of the Tuvans was shown (Potapov, 1969). For some groups of Tuvans, isolation of local populations was noted, caused both by geographical factors and historically, which is especially pronounced for Tuvans-Todzhans from the northeastern mountainous part of Tuva. In recent years, a number of scientific publications have been devoted to the study of the Tuvinian gene pool, which were focused on the study of the general spectrum of mtDNA lines, Y-chromosome haplogroups, and the detailing of individual clades (Stepanov, Puzyrev, 2000; Stepanov et al., 2001, 2006; Derenko et al., 2006; Kharkov et al., 2013; Damba et al., 2018a, b; Agdzhoyan et al., 2021).

The purpose of this study is a comprehensive analysis of the structure of the gene pool of Tuvans and the reconstruction of their origin in comparison with other populations of the indigenous population of Siberia. To address the issues of genetic proximity of Tuvans with other indigenous peoples, genotyping of a wide genomic set of autosomal markers using high-density biochips, as well as an expanded set of SNP and STR-markers of the Y-chromosome was performed in various indigenous peoples of Siberia.

Materials and methods

The material of the study was DNA samples of men with a total number of 419 samples, representing the indigenous population of the Republic of Tuva. Samples were collected in the village Teeli (west of Tuva) ($N = 44$), village Kungurtug (south-east of the Republic) ($N = 48$), village Toora-Khem (north-eastern part of Tuva) ($N = 23$), and the city of Kyzyl ($N = 304$). Samples from Kyzyl were assigned to the corresponding territorial group according to the birthplaces of the donors. The samples were divided into five territorially distant groups: west (Barun-Khemchigsky, Bai-Taiginsky, Dzun-Khemchigsky, Sut-Kholsky, Mongun-Taiginsky districts) ($N = 169$), center (Chaa-Kholsky, Tandynsky, Kaa-Khemsky, Kyzyl, Ulug-Khem, Chedi-Khol, Piy-Khem, Tes-Khem, Ovyur, Erzin districts) ($N = 179$), east ($N = 71$), including the northeast (Todzhinsky district) ($N = 23$) and southeast (Tere-Kholsky district) ($N = 48$).

The sampling of primary biological material (venous blood) from donors was carried out in compliance with the procedure of written informed consent for the study. For each donor, a

questionnaire was compiled with a brief pedigree, indicating ethnicity and places of birth of ancestors. An individual was assigned to a given ethnic group based on his own ethnic identity, his parents and place of birth.

For the analysis of Y-chromosome haplogroups and haplotypes of Tuvans, all 419 male DNA samples were used. For genotyping on chips, unrelated accessions from the village of Teeli of Bai-Taiga kozhuun (N = 28) were selected. Other populations of the indigenous population of Siberia are represented by: Chulyms, Khakas-Sagays, Khakas-Kachins, Southern Altaians, Kets, Khanty, Tomsk Tatars, Buryats, Yakuts, Evenks, Nivkhs, Udeges, as well as Kalmyks, Dungans and Kirghiz.

Genome-wide genotyping data were obtained using Infinium Multi-Ethnic Global-8 (Illumina) microarrays for SNP genotyping, including over 1.7 million markers. The material was deposited in the bioresource collection "Biobank of the Population of Northern Eurasia". For comparative analysis, we used genotype data for 1677114 autosomal SNPs (Illumina Multi-Ethnic Global-8 biochip) of 917 samples and genotyping data for more than 3000 Y-chromosomal SNPs and 36 YSTRs from more than 1600 male samples representing the indigenous population of Siberia and neighboring regions. More than 30 population samples have been characterized, which are described in detail in our previous works (Kolesnikov et al., 2021, 2022). The NGSadmix method (Scotte, 2013) and the ADMIXTURE program (Alexander et al., 2009, 2011) were used to analyze the component composition and amount of impurities in individuals and populations, and a comparative analysis of autosomal SNP data and haplogroups and haplotypes of Y-chromosomes.

Autosomal SNP genotype array clustering and quality control were performed using a protocol developed by (Guo et al., 2014) using GenomeStudio (Ilumina. GenomeStudio, genotyping module v2.0.3), a software package that Illumina has developed for various genomic analyses. For filtering, normalizing and calculating standard genomic statistics and indicators, the standard set of programs, including vcftools, bcftools, and plink, proved to be optimal. To analyze linkage blocks identical in origin, the Refined IBD algorithm (Browning B.L., Browning S.R., 2013) was used, which shows more accurate results compared to the algorithms built into plink. The genotypes were preliminarily phased using the Beagle 5.1 software (Browning S.R., Browning B.L., 2007). To compare the populations, the sums of the average lengths of blocks identical in origin (IBD segments – identical by descent) were obtained between pairs of individuals.

To study the composition and structure of Y-chromosome haplogroups, two systems of genetic markers were included in the study: diallelic loci represented by SNPs and polyallelic highly variable microsatellites (YSTRs). With the help of 156 SNP markers, the belonging of the samples to different haplogroups was determined. The classification of haplogroups is given in accordance with the data of the International Society for Genetic Genealogy (website www.isogg.org). Analysis of STR haplotypes within haplogroups was performed 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 an ABI Prism 3730 genetic analyzer. Genotyping of SNP markers was performed using PCR and subsequent analysis of DNA fragments using RFLP analysis. Experimental studies were carried out on the basis of the Center for Collective Use of Research Equipment "Medical Genomics" (Research Institute of Medical Genetics of the Tomsk National Research Medical Center). The construction of median networks of Y-chromosome haplotypes was carried out using Network v.10.2.0.0 (Fluxus Technology Ltd; www.fluxus-engineering.com) using the Bandelt median network method (Bandelt, 1999). The generation age of the observed diversity of haplotypes in haplogroups was estimated using the ASD method (Zhivotovsky, 2004), based on the mean square differences in the number of repeats between all markers.

Results and discussion

Genotyping of a large array of SNPs makes it possible to study in great detail the patterns of haplotype diversity that mark various substrate and superstrate layers of the population gene pool, the degree of miscegenation with the alien population at various levels – from individual to generic and ethnic, to conduct a detailed analysis of the demographic history of various populations and analyze the molecular phylogenetic and phylogeographic structure of Y-chromosome haplogroups. This makes it possible to more accurately reconstruct the genetic and demographic events that occurred in the past. The use of modern bioinformatic approaches on a wide array of SNPs and a detailed phylogeny of uniparental lines makes it possible to more accurately reconstruct the formation of the Tuvan gene pool.

After processing the data on the results of a microchip study to filter the progenotyped samples and carry out further calculations, a search was first made among the Tuvans of mestizos using the NGSadmix program. The NGSadmix method, when launched on the data array that we formed, showed that all progenotyped samples of the Tuvans do not have crossbreeding, which is fully consistent with the data of the DNA donor questionnaire. The obtained data on the frequencies of SNPs in the studied population samples were used to elucidate the genetic relationships between different ethnic groups. The ADMIXTURE algorithm was used to reduce the dimension and identify the genetic components.

Component composition of the gene pool of the Tuvans

To identify individual genetic components in the gene pool of the studied populations, the ADMIXTURE program was used, which makes it possible to identify the mixed composition of a set of individuals based on genotype data and, thereby, to make assumptions about the origin of the population. Modeling using ADMIXTURE has recently become one of the main methods of analysis in the study of the gene pools of modern and ancient human populations, allowing you to analyze the same data at different hierarchical levels.

Tuvans, in comparison with most Siberian populations, show a very diverse composition of genetic components.

Their distribution is most clearly manifested at K = 12. For almost all Siberian populations, the complete dominance of one genetic component, characteristic of individual samples or closely related indigenous peoples, is shown. In addition to the Tuvans, a rather heterogeneous component composition was also found among the Khakas-Kachians. The spectrum of genetic components of the Kachins almost completely coincides with the Tuvans, but differs in their proportions.

Altai component. With the maximum frequency in the Tuvans (53 %), the genetic component that dominates in the Southern Altaians (up to 90 %) is represented. Taking into account the fact that the analyzed sample of the Tuvans represents the westernmost region bordering the Republic of Gorny Altai, this is quite natural. It is presented with sufficient frequency among the Kyrgyz (9.8 %) and Khakas-Kachins (7.6 %), related to the Southern Altaians. Probably, this genetic component is associated with the influence of Turkic speakers in the formation of modern South Siberian peoples. Previously, the proximity of the Altaians and Tuvans was shown by analyzing the allele frequencies of the *ZFX* gene (Khitrinskaya et al., 2010), X-linked STR markers (Vagaitseva et al., 2014), enzymes and blood proteins (Spitsyn et al., 1984), frequencies blood groups of the ABO system and according to their anthropological parameters (Bogdanova, 1978a, b; Alekseev, 1984; Alekseeva, 1984).

East Siberian component. The second most common among the Tuvans is the East Siberian genetic component (21 %), which is dominant among the Yakuts (94 %), Evenks from Yakutia (93 %) and Transbaikalia (62 %). This corresponds to the linguistic data on the South Siberian origin of the ancestors of modern Yakuts. It is 30 % among the Buryats, 12 % among the Kachins, and 4 % among the Southern Altaians. The distribution of this genetic component is consistent with the classification of racial types. Tuvans, Tofalars, Yakuts and Dolgans are carriers of the traits of the North Asian minor race – one of the subdivisions of the continental branch of the great Mongoloid race. Two moderately different types are distinguished in the composition of the North Asian Mongoloids – Baikal and Central Asian. The first type is typical primarily for the Tungus-Manchurian peoples, the second – for the Turkic and Mongolian peoples (Turkic Peoples of East Siberia, 2008).

East Asian component. In third place in the Tuvans (11 %) is the dominant component of the Dungans (91 %), Buryats (63 %) and Kalmyks (54 %). It manifests itself most clearly at K = 12. It makes up a larger proportion among the Kirghiz (49 %), Kazakhs (46 %), Uzbeks (43 %), Khakas-Kachins (41 %) and Tomsk Tatars (24 %) and has a small share among the Kachins (4 %) and Southern Altaians (4 %). It is this genetic component that reflects the contribution of the latest groups of immigrants from the territory of Mongolia to the gene pool of the population of Southern Siberia. Almost all other studied populations of Siberia and the Far East – the Yakuts, Shors, Khakas-Sagays and Chulyms demonstrate the almost complete absence of this component. It was not found among the Evenks, Khanty, Kets, Chulyms, Chukchi, Koryaks and Nivkhs. The general picture of the distribution of this genetic component is in good agreement with anthropological and ethnographic data on the influence of the Mongol expansion on the ethnogenesis of the studied ethnic groups.

Yeniseian component. The largest share of this component is characteristic of the Chulym Turks (94 %) and Kets (65 %). In the Kets, its proportion is lower due to miscegenation and the detection by the NGSAdmix method of a recent Caucasoid admixture in many samples. Among Tuvans, its frequency is 6.9 %, and among Kachins, 20 %. The results obtained are in good agreement with the data of ethnology, anthropology and linguistics on the contribution of the Yeniseian component to the formation of various peoples of the Altai-Sayan region and the historical areas of the Yeniseian languages.

Far Eastern component. The last genetic component in the Tuvans present with a significant frequency (4.9 %) prevails in the Nivkhs (96 %) and Udege (56 %). It is present with a low frequency among the Trans-Baikal Evenks (11 %), Buryats (10 %), Kalmyks (8 %) and Dungans (6 %). Probably, its presence reflects the contribution of the taiga Mongoloids, who in ancient times settled westward from Primorye and Transbaikalia.

It can be assumed that the Samoyed component can also be present in the Tuvinian gene pool, however, its determination requires an analysis of the population groups in which it is dominant (Nenets, Enets, Nganasans and Selkups).

Identical in origin clutch blocks. As a result of bioinformatics processing of genotyping data from high-density biochips of various Siberian populations, an analysis was made of the coincidence of DNA fragments common in origin between populations and individuals. A segment with identical nucleotide sequences is IBD in two or more individuals if they inherit it from a common ancestor without recombination, that is, in these individuals the segment has a common origin. The expected length of an IBD segment depends on the number of generations since the last common ancestor. One of the applications of the analysis of genome regions of common origin is the quantitative assessment of the degree of relationship between individuals, which can also supplement information on the genetic relationships of populations (Gusev et al., 2011).

Samples from the sample of the Tuvans showed the maximum match in IBD blocks among themselves (10.07 %), then with a sample of the Southern Altaians (1.62 %), Evenks (0.81 %), Yakuts (0.77 %), Chulyms (0.70 %), Khakas-Sagays (0.66 %), Khakas-Kachins (0.64 %), Buryats (0.58 %), Kalmyks (0.57 %), Udeges (0.39 %) and Khanty (0.38 %). The degree of overlap of IBD blocks between the Tuvans and other population samples is consistent with the results of ADMIXTURE on the distribution of allele frequencies and common genetic components in these populations. The FROH inbreeding coefficient was also calculated for all individuals by homozygosity blocks (ROH). For the Tuvans, its value (0.0151) is much lower than for the Chulyms (0.0292), Kazyms (0.0280) and Russkinskaya Khanty (0.0266), Kets (0.0259) and Khakas-Sagays from the foothill Tashtyp region. Almost equal to the Tuvans in terms of FROH value are the samples of the Southern Altaians (0.0168) and the Khakas-Kachins of the Shirinsky district (0.0146). This indicates the absence of a significant role of inbreeding in the formation of the gene pool of modern Tuvan populations.

Haplogroups of the Y-chromosome

For the most frequent Asian haplogroups of the Y-chromosome in the Tuvans, additional terminal SNPs were genotyped, which made it possible to more accurately separate the samples into individual specific sublines. The frequencies of occurrence are indicated only for them (see the Table). The frequencies of other rather rare haplogroups represented by separate samples, indicated in an earlier article (Kharkov et al., 2013), are not given here, since additional SNPs were not selected for them.

The most frequent Y-chromosome haplogroup in the Tuvans is N1a2b1-B169, which makes up 24 % of the total array of male samples. It is divided into three sublines that differ in terminal SNPs and haplotype clusters. Its variant N1a2b1b2b1 (B178, PF3415, Z35147, Z35149, Z35152) is present with the maximum frequency among the Tuvans. In addition to the Tuvans, two samples of the Southern Altaians belong to it. According to the YFull website, this line was also found in one man from Kyrgyzstan and two from China. The haplotypes of this lineage have a stellar phylogeny, indicating a strong founder effect (Fig. 1).

The age of this line among the Tuvans according to YSTR is 1442 years (SD = 368 years). Its presence among the Altaians, Kirghiz and Chinese in the form of single samples is possibly associated with the inclusion of individual men of Tuvan origin in their composition. This line among the Tuvans represents a common genetic substrate for them, which is unequivocally connected with the heritage of the Samoyed population of the territory of Southern Siberia. The presence of different ethnospecific variants of the N1a2b1 haplogroup among the Tuvans, Khakas, and Shors indicates a significant genetic differentiation between them. This confirms the absence of migrations of carriers of this haplogroup and gene exchange over the past few hundred years. The main factor in its spread on the territory of Tuva was the genetic isolation of local Samoyedic groups and the intensive increase in their population. Four samples of Tuvans belong to a very rare parallel line N1a2b1b2a1~ (B228, Z35125, Z35127, Z35128). It was previously found in Mongols (Illimäe et al., 2016). The third Tuvan subline (xB175, Z35117, Z35118) includes 10 samples.

The second most frequent among the Tuvans is the haplogroup N1a1 (19 %), which is divided into three branches. In the total sample, its frequency is inferior to N1a2b1 by only 5 %, covering slightly less than 30 % of samples in the west of Tuva. The first line N1a1a2~ (B187 xB449) in the total sample of the Tuvans has a frequency of 6.4 %. In the eastern regions – Todzhiinsky and Tere-Kholsky, this haplogroup was not found. This variant is very ethnospecific and is not found in other populations. The sister line parallel to it (N1a1a2 ~ B499) with a relatively recent divergence from the Tuvan line is also characteristic of the Khakas-Sagais and Shors. It dominates in frequency in the Khakas seoks Khyi and Khobyi. Among the Shors, this haplogroup includes all men of the seok Kyi and Kobyi (Kharkov, 2020). On the median networks, the haplotypes of these lines in the Tuvans, Khakas, and Shors form three clusters that do not intersect with the Tuvans, except for one sample (Fig. 2).

At the same time, the haplotypes of the Khakas-Sagays of the Tashtyp district, bordering Shoria, are very close to the Shors and demonstrate a strong recent founder effect. The total age of the Tuvan haplotype cluster was 1863 years (SD = 294 years). This shows the long-standing division of these lineages between the Tuvans, Khakas, and Shors, and rather strong founder effects for individual seoks of Khakas and Shors. This subline has a very limited geographic range. Most likely,

Frequencies of occurrence of the main Y-chromosome haplogroups among the Tuvans

Haplogroup	% (*N*)
N1a2b1b2b1 (B178, PF3415, Z35147, Z35149, Z35152)	23.9 (100)
N1a2b1b2a1~ (B228, Z35125, Z35127, Z35128)	0.9 (4)
N1a2b1b (B169 xB175, Z35117, Z35118)	2.4 (10)
N1a1a2~ (B187 xB449)	6.4 (27)
N1a1a1a1a3a2 (B219 xB199)	11.9 (50)
N1a1a1a (L708, L839 xL392)	0.5 (2)
Q1b1a3b1a~ (B30/YP1691, YP1693, YP1694)	12.9 (54)
C2b1a1a1a1 (F3850)	1.4 (6)
C2b1a1a2a (F1756)	1.4 (6)
C2b1c (M504)	2.8 (12)
C2b1b1 (M77)	10.5 (44)
R1a1a1b2e1~ (YP1505, YP1507, YP1508, YP1509)	4.1 (17)
R1a1a1b2a2a (Z2123)	1.4 (6)
R1a1a1b2a2a3b1a1~ (YP1542-1556)	1.2 (5)
R1a1a1b2 (Y43109)	6.4 (27)

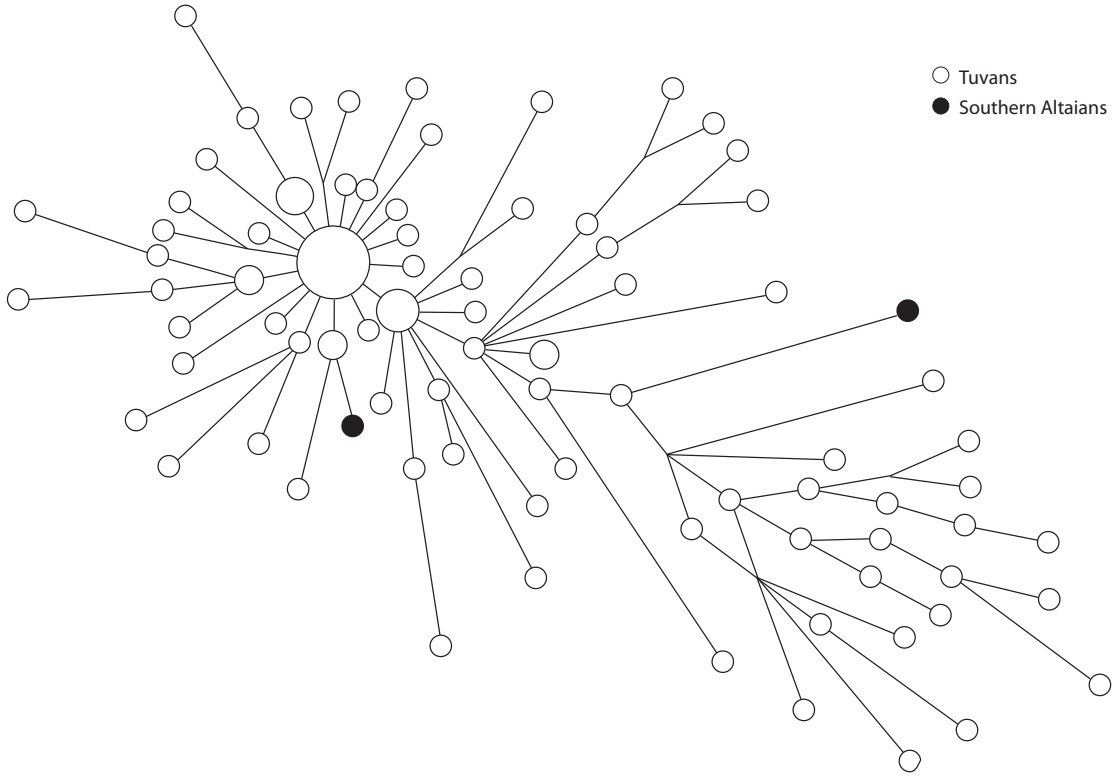

Fig. 1. Median network of YSTR haplotypes of the N1a2b1b2b1 haplogroup in Tuvans and Southern Altaians.

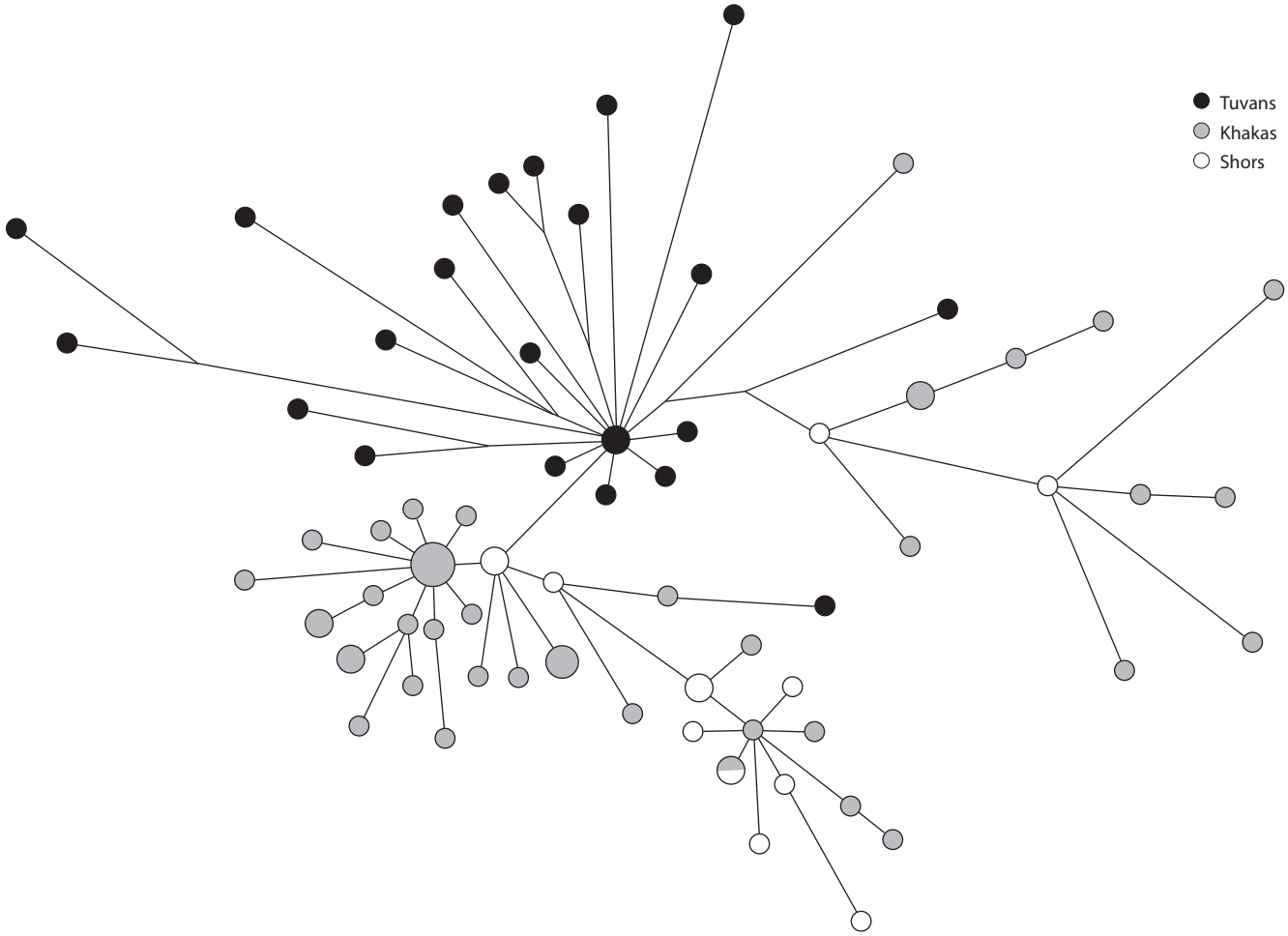

Fig. 2. Median network of YSTR haplotypes of the N1a1a2~ haplogroup in Tuvans, Khakas, and Shors.

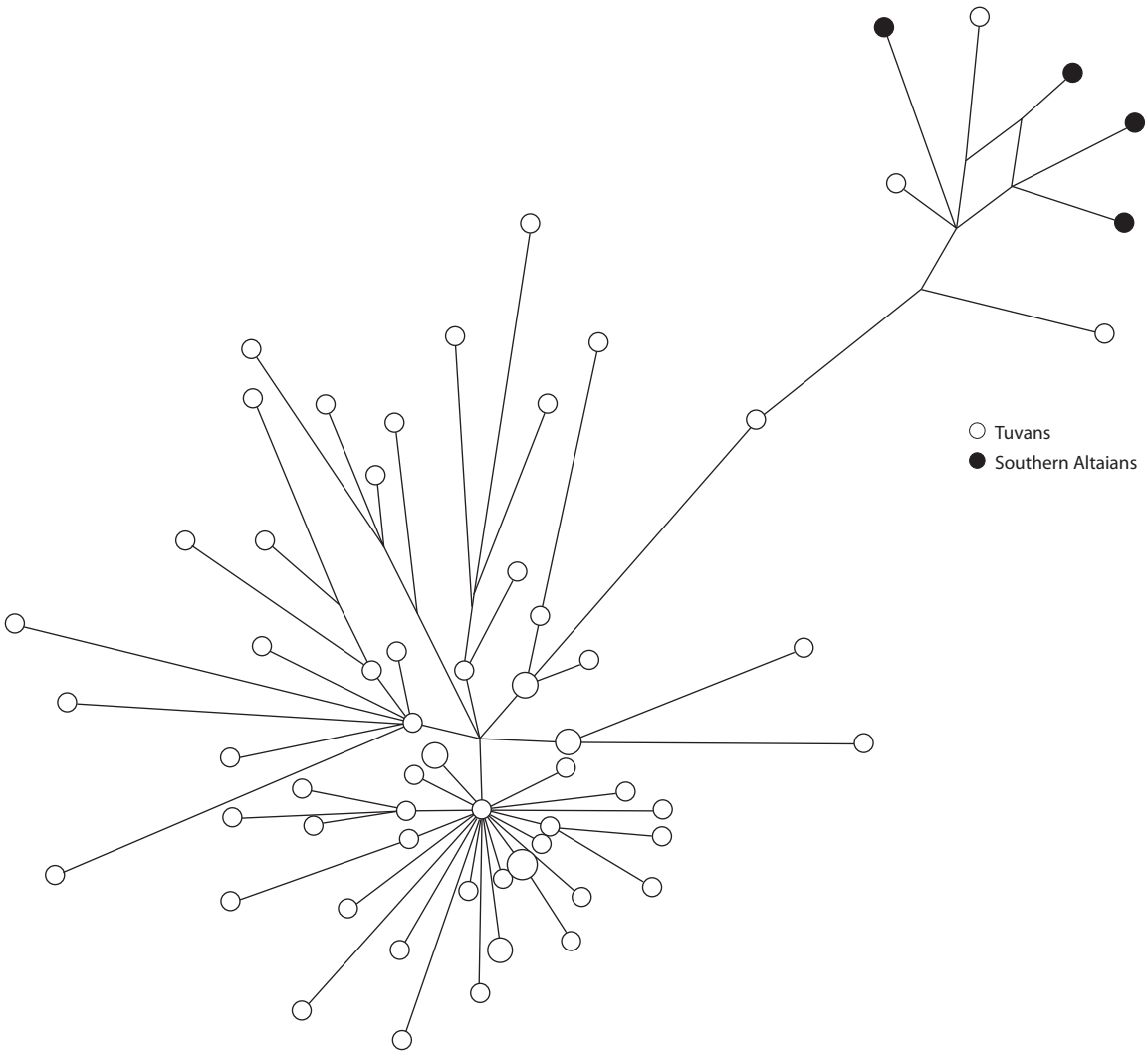

Fig. 3. Median network of YSTR haplotypes of haplogroup Q1b1a3b1a~ – B30 in Tuvans and Southern Altaians.

the initial place of its distribution was the territory of Tuva, from where it spread to Gornaya Shoria and then to Khakassia. It separated from the main stem of N1a1 very early and, like many other rare Y-chromosomal lineages, was preserved with a sufficiently high frequency only in relatively isolated mountain populations. Its separation from the main stem of the N1a1a haplogroup occurred approximately 10,700 years ago (YFull). Due to its population specificity and isolation, the relationship of this variant with the Samoyedic, Ugric, or other genetic components is ambiguous.

The second subline N1a1a among the Tuvans is N1a1a1a1a3a2 (B219 xB199). It is represented in all districts and has a frequency of 11.9 %. It also includes three samples of the Altaians. It was not found among the Khakas, Shors and Chulyms. The line N1a1a1a1a3a2c2-B199, which is closely related to it, dominates in the Eastern Buryats and is represented by a rather high frequency in the Western Buryats. The appearance of these lines is unambiguously connected with the settlement of Mongolian ethnic groups in Tuva, Buryatia and Altai. The age of this line according to haplotypes among the Tuvans was 1500 years (SD = 304 years). The spread of this Y-chromosome lineage occurred a little later than the carriers of the Tuva-Shor-Khakas branch N1a1a2~.

Only two specimens of Tuvan-Todzhans belong to a very rare lineage N1a1a1a (L708, L839 xL392). In terms of haplotypes, it is very close to the Yakut-Evenki haplogroup N1a1a1a1a4a1a1, but is not mutated in its terminal SNPs (M1979, M1984, M1988, M1991). The presence of this Y-chromosome variant in the Todzhans is consistent with the distribution of the East Siberian genetic and overlap in IBD blocks with the Yakuts and Evenks. This lineage also includes four samples of Khakas-Sagay men from the Askizsky district with haplotypes close to those of Tuva.

Haplogroup Q1b1a3b1a~ (B30/YP1691, YP1693, YP1694) occupies 13 % of the total sample of the Tuvans. Its maximum frequency falls on the eastern samples of the Todzhans and Tuvans of the Tere-Kholsky kozhuun (25 %). Four samples of Southern Altaians also belong to this lineage (Fig. 3).

The descending gradient of this haplogroup from east to west was shown on the territory of Tuva earlier (Kharkov et al., 2013; Damba et al., 2018b; Agdzhoyan et al., 2021). The highest frequency of the haplogroup Q1b1a3b1a~ for Tuvans

in Todzha is apparently a consequence of their relative genetic isolation and the preservation of a greater proportion of the local autochthonous Yeniseian genetic component. The age of this line according to haplotypes among the Tuvans was 2187 years (SD = 446 years). The distribution frequency of the haplogroup Q1b1a3b1a~ and its related lines Q1b1a3b1a2-B33 and Q1b1a3b4-B31 in the populations of the indigenous peoples of Southern and Western Siberia reflects the contribution to their gene pools of local aboriginal population groups belonging to the Yeniseian language family, which are quite ancient in origin. Analysis of the Y-chromosomal sublines Q1b1a3b shows that the original center of origin and settlement of its carriers is the territory of modern Tuva.

Different populations with a share of the Mongolian genetic component have different haplogroups and sublines, the origin of which is associated with the settlement of various ethnic groups and migration events of different times. Among the Tuvans, the result of the Mongolian contribution, in addition to N1a1a1a1a3a2, is the haplogroups of the clades C2b1, O2 and O3. All of them are very close to the variants presented with a high frequency among the Mongols, Buryats and Kalmyks. The share of C2b1c (M504) and C2b1a1a1a1 (F3850) is the highest in the southeastern sample (15 %). Lineage C2b1a1a1a1 (F3850) was found only in the southern and southeastern regions. The more frequent line C2b1b1 (M77) shows a clinal decrease in frequency from southeast to west. The same is true for haplogroups O2 and O3. In the gene pool of almost all the populations studied, in which the Mongolian genetic component is not detected by autosomal SNP, the indicated Y-chromosome haplogroups are also absent. Phylogenetic analysis of Y-chromosomal sublines and haplotypes shows that the center of origin and distribution of the carriers of the Mongolian component is the territory of Central Asia.

These haplogroups among the Tuvans are a legacy of the genetic contribution of late Mongoloid migrants, reflecting the contribution of the Xiongnu and Mongolian settlers to the territory of Tuva. Thus, genetic data confirm that the penetration of Mongolian nomads into the territory of Tuva came from the south, gradually spreading to the northern regions, and, accordingly, the mongolization of the population of Tuva was most pronounced precisely in the southern regions. This coincides with the data of paleoanthropology (Alekseev, 1984) and anthropology of the modern population (Bogdanova, 1978a). The data of linguistics characterizing the southeastern dialect as formed as a result of the significant influence of the Mongolian language also completely coincide with the distribution of this component and the frequencies of haplogroups that we obtained.

The haplogroup R1a1a (12 %) among the Tuvans includes seven different lines. Six Tuvan men belong to the R1a1a1b2a2a (Z2123) lineage. Three Altaians and two Kirghiz also belong to it. Five more Tuvans belong to the line R1a1a1b2a2a3b1a1~ (YP1542, YP1556) close to it. It dominates in frequency among the southern Altaians and Teleuts. Twenty-seven Tuvans had the R1a1a1b2 (Y43109) line, which was divided into three variants differing in haplotypes. Sixteen Tuvans and five Southern Altaians belong to one variant. In the second, there are four Tuvans, Khakas from the Turan and Khyzyl Khaya seoks, and almost all Shors from the Tartkyn, Shor-Kyzai and Kara-Shor seoks. In the third, there are seven Tuvans, Khakas from various seoks of the Beltir and Biryusin, and Shors of the seoks of the Cheley and Chediber. This confirms the data that some groups of Tuvans who roamed in the Minusinsk Basin and were later called the “Beltyr” were completely assimilated by local tribes, constituting one of the components of the formation of the ethnos of modern Khakas.

The haplogroup R1a1a1b2e1~ (YP1509) among the Tuvans is also divided by haplotypes into two lines. Nine samples of Tuvans of the first variant are very similar in haplotypes to this variant among the Khakas of the Kharga seoks and the Shors of the Karga and Cheli seoks and one Teleut. Eight samples belong to another specific variant common among the Telengits and Northern Altaians.

A very large diversity of the R1a1a haplogroup was shown among the indigenous population of the Altai-Sayan region. Its various sublines split long ago and show no star-like haplotype phylogeny, reduced diversity, or traces of the founder effect. This indicates a significant size of the effective size of the populations of the ancient Caucasoids and Turks, who introduced these components into the gene pool of modern Tuvans, Khakas and Shors. Founder effects with significant demographic growth were found only in the Southern Altaians, Kirghiz and Teleuts in the haplogroup R1a1a1b2a2a3b1a1~. The distribution of various discovered sublines of the haplogroup R1a1a in the territory of Tuva, Altai, Khakassia and Shoria is most likely associated with the Turks and the Yeniseian Kyrgyz.

Of the other haplogroups among the Tuvans, eight more are single samples (D, E, I1, I2a, J1, J2a, J2a1 and R1b). Most likely, their presence is partly due to the recent miscegenation and earlier dispersal of the Central Asian populations. The results of the study of the detailed phylogeny of Y-chromosome haplogroups made it possible to more accurately analyze the component composition of the Tuvan gene pool. This is a more accurate addition to the analysis of autosomal markers, which makes it possible to reconstruct in detail the formation of their gene pool. This information is also important for describing the similarities and differences between the compared groups, as well as the processes of their ethnogenesis. Various Y-chromosome haplogroups in the Tuvan gene pool demonstrate their genetic affinity with the Altaians, Khakas, Shors, Buryats, Mongols, Evenks, Kets, Chulym Turks, and Teleuts. This allows us to characterize in more detail the gene pool of the indigenous South Siberian population and the genetic relationships and continuity of populations living in this territory.

Conclusion

Thus, in the present study, a detailed study of the gene pool of Tuvans was carried out based on the data of high-density biochips and a wide range of SNPs of the non-recombining part of the Y-chromosome. A very heterogeneous composition of the gene pool of the Tuvans and Khakas was found, both in autosomal SNPs and in various sublines of Y-chromosomal haplogroups. The maximum closeness of the gene pool of the Tuvans with the Altaians, Khakas and Shors is shown. Analy-

sis of IBD blocks and individual rare variants of male lines demonstrates traces of more ancient connections with the ancient aboriginal population of this region and the populations of Eastern Siberia and the Far East. Within the Tuvan ethnos, significant differences were found between samples from the western, southern, and eastern regions of Tuva in terms of the proportion of the Mongolian and Yeniseian genetic component. The genetic diversity of the Tuvans in Y-chromosomal haplogroups and the most heterogeneous composition of genetic components indicate the highest diversity of the Tuvan gene pool, compared to all other indigenous peoples of Siberia.

In the future, we plan to analyze in more detail the structure of the gene pools of the South and West Siberian populations by adding population samples of the Samoyedic peoples – the Nenets and Selkups.

References

Agdzhoyan A.T., Damba L.D., Gurianov V.M., Zaporozhchenko V.V., Balanovsky O.P. Phylogenetic analysis of the South Siberian Q-YP1102 haplogroup based on the data on Y-SNP and Y-STR markers in Tuvans and surrounding populations. *Russ. J. Genet*. 2021;57:1398-1407. DOI 10.1134/S1022795421120024.

Alekseev V.P. Brief account of the paleoanthropology of Tuva in connection with historical issues. In: Anthropoecological Research in Tuva. Moscow: Nauka Publ., 1984;6-75. (in Russian)

Alekseeva T.I. Anthropological features of modern Tuvans. Cephalometry and cephaloscopy. In: Anthropoecological Research in Tuva. Moscow: Nauka Publ., 1984;75-114. (in Russian)

Alexander D.H., Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 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. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 1999;16(1):37-48. DOI 10.1093/oxfordjournals.molbev.a026036.

Bogdanova V.I. Anthropological study of modern Tuvans in 1972–1976. In: Field Studies of the Institute of Ethnography in 1976. Moscow: Nauka Publ., 1978a;187-198. (in Russian)

Bogdanova V.I. Some issues of the origins of the anthropological composition of present-day Tuvan people. *Sovetskaya Etnografiya = Soviet Ethnography.* 1978b;6:46-58. (in Russian)

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.

Damba L.D., Aiyzhy E.V., Mongush B.B.O., Zhabagin M.K., Yusupov Yu.M., Sabitov Zh.M., Agdzhoyan A.T., Markina N.V., Dorzhu Ch.M., Balanovskaya E.V., Balanovsky O.P. Complex approach to the clan structure of Tuvans by the example of Mongush and Oorzhak clans. *Vestnik Tuvinskogo Gosudarstvennogo Universiteta. № 2. Estestvennye i Sel'skokhozyajstvennye Nauki = Bulletin of Tuva State University. No. 2. Natural and Agricultural Sciences*. 2018a;37(2):37-44. (in Russian)

Damba L.D., Balanovskaya E.V., Zhabagin M.K., Yusupov Yu.M., Bogunov Yu.V., Sabitov Zh.M., Agdzhoyan A.T., Korotkova N.A., Lavryashina M.B., Mongush B.B., Kavai-ool U.N., Balanovsky O.P. Estimating the impact of Mongol expansion on gene pool of Tuvans. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018b;22(5):611-619. DOI 10.18699/VJ18.402. (in Russian)

Debets G.F. Paleoanthropology of the USSR. Moscow; Leningrad: Publishing House of the USSR Academy of Sciences, 1948. (in Russian)

Derenko M., Malyarchuk B., Denisova G., Wozniak M., Dambueva I., Dorzhu C., Luzina F., Miścicka-Sliwka D., Zakharov I. Contrasting patterns of Y-chromosome variation in South Siberian population from Baikal and Altai-Sayan regions. *Hum. Genet*. 2006;118(5): 591-604. DOI 10.1007/s00439-005-0076-y.

Gene Pool of the Population of Siberia. Novosibirsk: Publ. House of the Institute of Archeology and Ethnography SB RAS, 2003. (in Russian)

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.

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.

Ilumäe A.-M., Reidla M., Chukhryaeva M., Järve M., Post H., Karmin M., Saag L., Agdzhoyan A., Kushniarevich A., Litvinov S., Ekomasova N., Tambets K., Metspalu E., Khusainova R., Yunusbayev B., Khusnutdinova E.K., Osipova L.P., Fedorova S., Utevska O., Koshel S., Balanovska E., Behar D.M., Balanovsky O., Kivisild T., Underhill P.A., Villems R., Rootsi S. Human Y chromosome haplogroup N: A non-trivial time-resolved phylogeography that cuts across language families. *Am. J. Hum. Genet.* 2016;99(1): 163-173. DOI 10.1016/j.ajhg.2016.05.025.

Kharkov V.N., Khamina K.V., Medvedeva O.F., Simonova K.V., Khitrinskaya I.Yu., Stepanov V.A. Gene-pool structure of Tuvinians inferred from Y-chromosome marker data. *Russ. J. Genet*. 2013; 49(12):1236-1244. DOI 10.1134/S102279541312003X.

Kharkov V.N., Novikova L.M., Shtygasheva O.V., Luzina F.A., Khitrinskaya I.Yu., Volkov V.G., Stepanov V.A. Gene pool of Khakass and Shors for Y chromosome markers: common components and tribal genetic structure. *Russ. J. Genet*. 2020;56(7):849-855. DOI 10.1134/S1022795420070078.

Khitrinskaya I.Yu., Khar'kov V.N., Stepanov V.A. Genetic diversity of the chromosome X in aboriginal Siberian populations: The structure of linkage disequilibrium and haplotype phylogeography of the *ZFX* locus. *Mol. Biol.* 2010;44(5):709-719. DOI 10.1134/S0026893310050055.

Kiselev S.V. History of South Siberia. Moscow: Publishing House of the USSR Academy of Sciences, 1951. (in Russian)

Kolesnikov N.A., Kharkov V.N., Zarubin A.A., Radzhabov M.O., Voevoda M.I., Gubina M.A., Khusnutdinova E.K., Litvinov S.S., Ekomasova N.V., Shtygasheva O.V., Maksimova N.R., Sukhomyasova A.L., Stepanov V.A. Features of the genomic distribution of runs of homozygosity in the indigenous population of Northern Eurasia at the individual and population levels based on high density SNP analysis. *Russ. J. Genet*. 2021;57(11):1271-1284. DOI 10.1134/S1022795421110053.

Kolesnikov N.A., Kharkov V.N., Zarubin A.A., Voevoda M.I., Gubina M.A., Shtygasheva O.V., Maksimova N.R., Sukhomyasova A.L., Stepanov V.A. Signals of directed selection in the Indigenous populations of Siberia. *Russ. J. Genet*. 2022;58(4):473-477. DOI 10.1134/S102279542204007X.

Kucher A.N., Ondar E.A., Stepanov V.A. Tuvinians: genes, demography, health. Tomsk: Pechatnaya Manufaktura Publ., 2003. (in Russian)

Potapov L.P. Essays on the Folk Life of the Tuvans. Moscow: Nauka Publ., 1969. (in Russian)

Puzyrev V.P., Erdynieva L.S., Kucher A.N. Genetic and Epidemiological Study of the Population of Tuva. Tomsk: STT Publ., 1999. (in Russian)

Skotte L., Korneliussen T.S., Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013;195(3):693-702. DOI 10.1534/genetics.113.154138.

Spitsyn V.A., Boeva S.B., Filippov I.K. Genetic and anthropological study of the indigenous population of the Altai-Sayan highland. In: Anthropo-Ecological Research in Tuva. Moscow: Nauka Publ., 1984;185-194. (in Russian)

Stepanov V.A., Kharkov V.N., Puzyrev V.P. Evolution and phylogeography of human Y-chromosomal lineages. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeding Scientists*. 2006;10(1):57-73. (in Russian)

Stepanov V.A., Khitrinskaya I.Yu., Puzyrev V.P. Genetic differentiation of the Tuva population with respect to the Alu-insertions. *Russ. J. Genet*. 2001;37(4):453-459. DOI 10.1023/A:1016623030663.

Stepanov V.A., Puzyrev V.P. Analysis of the allele frequencies of seven Y-chromosome microsatellite loci in three Tuvinian populations. *Russ. J. Genet*. 2000;36(2):179-185.

Turkic Peoples of East Siberia. Moscow: Nauka Publ., 2008. (in Russian)

Vagaitseva K.V., Kharkov V.N., Cherpinskaya K.V., Khitrinskaya I.Yu., Stepanov V.A. Genetic variability of X-linked STR markers in Siberian populations. *Mol. Biol*. 2015;49(2):267-274. DOI 10.1134/S0026893315020132.

Zhivotovsky L.A., Underhill P.A., Cinnioglu 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. 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.

ORCID ID

V.A. Stepanov orcid.org/0000-0002-5166-331X
N.A. Kolesnikov orcid.org/0000-0001-8855-577X
A.A. Zarubin orcid.org/0000-0001-6568-6339
V.N. Kharkov orcid.org/0000-0002-1679-2212

Acknowledgements. The study was supported by the Russian Science Foundation grant No. 22-64-00060 (https://rscf.ru/project/22-64-00060/).

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

Received October 14, 2022. Revised December 28, 2022. Accepted December 28, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):46-54
DOI 10.18699/VJGB-23-07

Original Russian text https://vavilovj-icg.ru/

Relationship of the gene pool of the Khants with the peoples of Western Siberia, Cis-Urals and the Altai-Sayan Region according to the data on the polymorphism of autosomic locus and the Y-chromosome

V.N. Kharkov ✉, N.A. Kolesnikov, L.V. Valikhova, A.A. Zarubin, M.G. Svarovskaya, A.V. Marusin, I.Yu. Khitrinskaya, V.A. Stepanov

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ vladimir-kharkov@medgenetics.ru

Abstract. Khanty are indigenous Siberian people living on the territory of Western Siberia, mainly on the territory of the Khanty-Mansiysk and Yamalo-Nenets Autonomous Okrugs. The present study is aimed at a comprehensive analysis of the structure of the Khanty gene pool and their comparison with other populations of the indigenous population of Southern and Western Siberia. To address the issues of genetic proximity of the Khanty with other indigenous peoples, we performed genotyping of a wide genomic set of autosomal markers using high-density biochips, as well as an expanded set of SNP and STR markers of the Y-chromosome in various ethnic groups: Khakas, Tuvans, Southern Altaians, Siberian Tatars, Chulyms (Turkic language family) and Kets (Yeniseian language family). The structure of the gene pool of the Khanty and other West Siberian and South Siberian populations was studied using a genome-wide panel of autosomal single nucleotide polymorphic markers and Y-chromosome markers. The results of the analysis of autosomal SNPs frequencies by various methods, the similarities in the composition of the Y-chromosome haplogroups and YSTR haplotypes indicate that the Khanty gene pool is quite specific. When analyzing autosomal SNPs, the Ugrian genetic component completely dominates in both samples (up to 99–100 %). The samples of the Khanty showed the maximum match in IBD blocks with each other, with a sample of the Kets, Chulyms, Tuvans, Tomsk Tatars, Khakas, Kachins, and Southern Altaians. The degree of coincidence of IBD blocks between the Khanty, Kets, and Tomsk Tatars is consistent with the results of the distribution of allele frequencies and common genetic components in these populations. According to the composition of the Y-chromosome haplogroups, the two samples of the Khanty differ significantly from each other. A detailed phylogenetic analysis of various Y-chromosome haplogroups made it possible to describe and clarify the differences in the phylogeny and structure of individual ethnospecific sublines, to determine their relationship, traces of population expansion in the Khanty gene pool. Variants of different haplogroups of the Y-chromosome in the Khanty, Khakas and Tuvans go back to their common ancestral lines. The results of a comparative analysis of male samples indicate a close genetic relationship between the Khanty and Nenets, Komi, Udmurts and Kets. The specificity of haplotypes, the discovery of various terminal SNPs confirms that the Khanty did not come into contact with other ethnic groups for a long time, except for the Nenets, which included many Khanty clans.
Key words: gene pool; human population; genetic diversity; genetic components; Y-chromosome; Khanty.

For citation: Kharkov V.N., Kolesnikov N.A., Valikhova L.V., Zarubin A.A., Svarovskaya M.G., Marusin A.V., Khitrinskaya I.Yu., Stepanov V.A. Relationship of the gene pool of the Khants with the peoples of Western Siberia, Cis-Urals and the Altai-Sayan Region according to the data on the polymorphism of autosomic locus and the Y-chromosome. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):46-54. DOI 10.18699/VJGB-23-07

Связь генофонда хантов с народами Западной Сибири, Предуралья и Алтая-Саян по данным о полиморфизме аутосомных локусов и Y-хромосомы

В.Н. Харьков ✉, Н.А. Колесников, Л.В. Валихова, А.А. Зарубин, М.Г. Сваровская, А.В. Марусин, И.Ю. Хитринская, В.А. Степанов

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ vladimir-kharkov@medgenetics.ru

Аннотация. Ханты – коренной сибирский народ, проживающий на территории Западной Сибири, в основном на территории Ханты-Мансийского и Ямало-Ненецкого автономных округов. Настоящее исследование направлено на комплексный анализ структуры генофонда хантов и их сравнение с другими популяциями коренного

населения Южной и Западной Сибири. Для решения вопросов генетической близости хантов с другими коренными народами выполнено генотипирование широкого геномного набора аутосомных маркеров с помощью высокоплотных биочипов, а также расширенного набора SNP- и STR-маркеров Y-хромосомы у различных этнических групп: хакасов, тувинцев, южных алтайцев, сибирских татар, чулымцев (тюркская языковая семья) и кетов (енисейская языковая семья). Результаты анализа частот аутосомных SNP различными методами, сходства по составу гаплогрупп Y-хромосомы и YSTR-гаплотипов свидетельствуют, что генофонд хантов достаточно специфичен. При анализе аутосомных SNP в обеих выборках полностью доминирует угорский генетический компонент (до 99–100 %). Выборки хантов показали максимальное совпадение по IBD-блокам между собой, с выборкой кетов, чулымцев, тувинцев, томских татар, хакасов-качинцев и южных алтайцев. Степень совпадения IBD-блоков между хантами, кетами и томскими татарами согласуется с результатами распределения в этих популяциях частот аллелей и общих генетических компонентов. По составу гаплогрупп Y-хромосомы две выборки хантов значительно различаются между собой. Детальный филогенетический анализ различных гаплогрупп Y-хромосомы позволил описать и уточнить различия в филогении и структуре отдельных этноспецифичных сублиний, определить их родство, следы экспансии численности в генофонде хантов. Варианты разных гаплогрупп Y-хромосомы у хантов, хакасов и тувинцев восходят к общим для них предковым линиям. Результаты сравнительного анализа образцов мужчин также свидетельствуют о близком генетическом родстве между хантами и ненцами, коми, удмуртами и кетами. Специфичность гаплотипов, обнаружение различных терминальных SNP подтверждают, что ханты достаточно долго не имели контактов с другими этносами, кроме ненцев, в состав которых вошло много хантыйских родов.

Ключевые слова: генофонд; популяции человека; генетическое разнообразие; генетические компоненты; Y-хромосома; ханты.

Introduction

The study of the structure of the gene pools of populations of various Siberian regions is one of the priority areas of modern human genetics and helps to reveal in detail some of the issues related to their ethnogenesis.

The Khanty are an indigenous people living on the territory of Western Siberia, mainly on the territory of the Khanty-Mansiysk and Yamalo-Nenets Autonomous Okrugs, as well as the Tyumen Region. Small groups of Khanty live in the north of the Tomsk Region and in the Komi Republic. According to the All-Russian census of 2010, the number of Khanty was 30,943 people, of which 61.6 % lived in the Khanty-Mansi Autonomous Okrug and 30.7 %, in the Yamalo-Nenets Autonomous Okrug. The Khanty have three large ethnographic groups that coincide with the groups of their language dialects – northern, southern and eastern, and the southern (Irtysh) Khanty were Turkified and became part of the Siberian Tatars, having mixed with them, and were also assimilated by Russian settlers (Peoples of West Siberia…, 2005).

Khanty populations are of considerable interest for population genetic studies, both due to the relatively poor knowledge with the involvement of modern genomic technologies, and due to the specificity of the gene pools of their individual groups that developed under conditions of long-term genetic isolation.

The settlement of the Khanty in antiquity was very wide – from the lower reaches of the Ob in the north to the Baraba steppes in the south and from the Yenisey in the East to the Trans-Urals, including the rivers Northern Sosva and Lyapin, as well as part of the rivers Pelym and Konda in the west. Since the 19th century, the Mansi began to move beyond the Urals from the Kama and Ural regions, being pressed by the Komi-Zyryans and Russians. From an earlier time, part of the southern Mansi also left to the north in connection with the creation in the XIV–XV centuries of the Tyumen and Siberian khanates – the states of the Siberian Tatars, and later (XVI–XVII centuries) with the development of Siberia by the Russians. In the XVII–XVIII centuries, the Mansi already lived on Pelym and Konda. Part of the Khanty also moved from the western regions to the east and north (to the Ob from its left tributaries), which is recorded by the statistical data of the archives. Their place was taken by the Mansi. So, by the end of the XIX century, there was no Ostyak population left on the rivers Northern Sosva and Lyapin: they either moved to the Ob or merged with the newcomers (The Peoples of Russia, 1994).

In the north, the Khanty came into contact with the Nenets, some of them were assimilated by them, which is confirmed by ethnographic data, as well as our study of the tribal structure of the Gydan Nenets according to Y-chromosome markers (Kharkov et al., 2021). The migration of the Khanty to the north and east continued into the 20th century. By the 20th century, the southern Khanty were almost completely assimilated by the Tatars and Russians.

Historically, the Khanty population was not homogeneous either in language or culture. Some scientists divide the Khanty language into two large groups – western and eastern, while others still subdivide the western dialects into southern and northern. In anthropological terms, the Khanty are the most characteristic representatives of the Ural type, which also includes the Mansi, Selkups, Nenets, Baraba Tatars, Shors, Northern Altaians and Khakas. The closest relatives of the Khanty in origin, language and culture are the Mansi (Brook, 1986).

The purpose of this study is a comprehensive analysis of the structure of the Khanty gene pool and the reconstruction of their origin in comparison with other populations of the indigenous population of Southern and Western Siberia. To address the issues of genetic proximity of the Khanty with other indigenous peoples, genotyping of a wide genomic set of autosomal markers using high-density biochips, as well as an expanded set of SNP and STR-markers of the Y-chromosome was performed in various ethnic groups: Khakas, Tuvans, Southern Altaians, Siberian Tatars, Chulyms (Turkic language family) and Kets (Yeniseian language family).

Materials and methods

The material of the study was DNA samples of men and women from two populations of the Khanty in the village of Russkinskaya, Surgut district and the village of Kazym, Beloyarsky district of the Khanty-Mansi Autonomous Okrug. The sampling of primary biological material (venous blood) from donors was carried out in compliance with the procedure of written informed consent for the study. For each donor, a questionnaire was compiled with a brief pedigree, indicating ethnicity and places of birth of ancestors. An individual was assigned to a given ethnic group based on their own ethnic identity, their parents and place of birth.

For the analysis of Y-chromosome haplogroups and haplotypes of the Khanty, 120 DNA samples of men from the village of Russkinskaya ($N = 64$) and the village of Kazym ($N = 54$) of the Khanty-Mansi Autonomous Okrug were used. For genotyping on high-density microchips, unrelated samples from the village of Kazym ($N = 30$) and the village of Russkinskaya ($N = 26$) were selected. Other populations of the indigenous population of Siberia are represented by: Chulyms ($N = 22$), Khakas (Sagays of the Tashtyp district, $N = 29$ and Kachins of the Shirinsky district, $N = 26$), Southern Altaians (village of Beshpeltir of the Chemal district, $N = 24$ and Kulada village, Ongudaysky district, $N = 25$), Kets (Kellogg village, Turukhansky district, Krasnoyarsk Territory, $N = 15$), Tomsk Tatars (Chernaya Rechka village, Eushta village and Takhtamyshevo village, Tomsky district, $N = 20$), Tuvinians (Teeli village of Bai-Taiginsky kozhuun, $N = 28$).

Genome-wide genotyping data were obtained using Infinium Multi-Ethnic Global-8 (Illumina) microarrays for SNP genotyping, including over 1.7 million markers. The material was deposited in the bioresource collection "Biobank of the Population of Northern Eurasia".

Autosomal SNP (single nucleotide polymorphism) genotype array clustering and quality control were performed using a protocol developed by (Guo et all., 2014) using GenomeStudio (Ilumina. GenomeStudio) (genotyping module v2.0.3), a software package that Illumina developed for various genomic analyses. For filtering, normalizing and calculating standard genomic statistics and indicators, the standard set of programs, including vcftools, bcftools, and plink, proved to be optimal.

To analyze linkage blocks identical in origin, the Refined IBD algorithm (Browning B.L., Browning S.R., 2013) was used, which shows more accurate results compared to the algorithms built into plink. The genotypes were preliminarily phased using the Beagle 5.1 software (Browning S.R., Browning B.L., 2007). To compare the populations, the sums of the average lengths of blocks identical in origin (IBD segments – identical by descent) were obtained between pairs of individuals.

The tSNE method was used to analyze genetic relationships between populations. The NGSadmix method (Scotte et al., 2013) and the ADMIXTURE program (Alexander et al., 2009; Alexander, Lange, 2011) were used to analyze the component composition and the amount of impurities in individuals and populations.

To study the composition and structure of Y-chromosome haplogroups, two systems of genetic markers were included in the study: diallelic locuses represented by SNPs and polyallelic highly variable microsatellites (YSTRs). With the help of 138 SNP markers, the belonging of the samples to different haplogroups was determined. The classification of haplogroups is given in accordance with the data of the International Society for Genetic Genealogy (website www.isogg.org).

Analysis of STR haplotypes within haplogroups was performed 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 an ABI Prism 3730 genetic analyzer. Genotyping of SNP markers was performed using PCR and subsequent analysis of DNA fragments using RFLP analysis.

Experimental studies were carried out on the basis of the Center for the Collective Use of Research Equipment "Medical Genomics" (Research Institute of Medical Genetics of the Tomsk National Research Medical Center). The construction of median networks of Y-chromosome haplotypes was carried out using the Network v.10.2.0.0 (Fluxus Technology Ltd; www.fluxus-engineering.com) using the Bandelt median network method (Bandelt et al., 1999). The generation age of the observed diversity of haplotypes in haplogroups was estimated using the ASD method (Zhivotovsky et al., 2004) based on the mean square differences in the number of repeats between all markers.

Results and discussion

The large array of data on autosomal SNPs obtained as a result of genotyping of high-density microarrays in samples of the Khanty and other indigenous Siberian peoples makes it possible to characterize the gene pool of the studied samples in the most detailed way using various methods. Genotyping of an extended set of specific Y-chromosome SNPs from various haplogroups makes it possible to describe the molecular-phylogenetic and phylogeographic structure of individual Y-chromosome haplogroups much more accurately.

After processing the data on the results of a microarray study to filter the progenotyped samples and perform further calculations, a search was carried out among the mestizo Khanty using the NGSadmix program. The algorithm of this program makes it possible to determine the ratio of ancestral components from NGS data with a relatively shallow coverage depth. The calculation principle is similar to other programs such as FRAPPE and ADMIXTURE, but NGSadmix, unlike them, works effectively when there is statistical uncertainty in individual genotypes. The NGSadmix method, when run on the data array we formed, showed that almost all Khanty samples do not have crossbreeding, which is fully consistent with the data from the DNA donor questionnaire. Crossbreeding with Russians (up to 30 %) was found only for one man from the village of Russkinskaya. His belonging to the European Y-chromosomal lineage R1b1a1b-L407 confirms the miscegenation on the paternal side. This sample was excluded from further calculations.

The obtained data on the frequencies of SNPs in the studied samples were used to elucidate the genetic relationships between the population samples included in the work. For dimensionality reduction, spatial analysis, and identification

of genetic components, we settled on two algorithms: tSNE and ADMIXTURE. The tSNE method makes it possible to more clearly divide the data array into separate ethnospecific groups of samples compared to the PCA method.

Genetic relationships of the Khanty with other populations of Western and Southern Siberia

When analyzing the data array on the frequencies of autosomal SNPs using the tSNE method at the level of individual samples (Fig. 1). It is shown that the two samples of the Khanty are very close, while the samples of the Kazym and Russkinskaya Khanty do not intersect on the graph and are separated from each other.

The Khanty are characterized by specific features of the gene pool and do not cluster with other populations. Compared with subethnic groups of the Khakas and Southern Altaians from different settlements, more geographically distant samples of the Khanty demonstrate a much greater genetic closeness. The samples of the Kets and Tomsk Tatars are closest to the Khanty. The genetic distances between the Khanty and the populations of Southern Siberia are much greater. Samples that are ethnically and geographically close to each other are located quite close in the Fig. 1, but each sample is included in a separate ethnospecific cluster. The exception is only a few single samples of the Khakas.

Component composition of the gene pool of populations. Modern methods used in genomic studies and new bioinformatic approaches make it possible to reliably identify ancestral genetic components of different origins in the gene pool of various populations and individuals. To identify individual genetic components in the gene pool of the studied populations, the ADMIXTURE program was used, which makes it possible to identify the mixed composition of a set of individuals based on genotype data and, thereby, to make assumptions about the origin of the population.

Modeling using ADMIXTURE has recently become one of the main methods of analysis in the study of the gene pools of modern and ancient human populations, allowing you to analyze the same data at different hierarchical levels. When the number of ancestral components is set to more than two, in most of the studied populations, a genetic component specific to the Khanty is revealed, which is most clearly manifested in the analyzed array of population samples at K = 8, which can be interpreted as the "Ugric" genetic layer in the gene pool of modern populations. The Khanty are characterized by the dominance of this component, which is their genetic basis (up to 99–100 % at the level of most individuals). A significant proportion of this component is also found in the Kets (up to 45–50 % in some individuals) and Tomsk Tatars (up to 5–9 %). Previously, it was shown that this component also occupies a significant share in the gene pool of the populations of the Volga-Ural region – the Bashkirs (up to 25 %), Maris (up to 20 %), Komi, Udmurts and Chuvashs (up to 15 %). It is present with less frequency in almost all South Siberian samples, among the Tuvans, Chulyms, Altaians, and Khakas of Sagays (from 5 to 10 %) (Kharkov et al., 2020).

The dominance of the Ugric component in all Khanty samples, starting from K = 3, and the almost complete absence of other genetic components in their genomes at the individual and population level, indicates that their ancestral populations

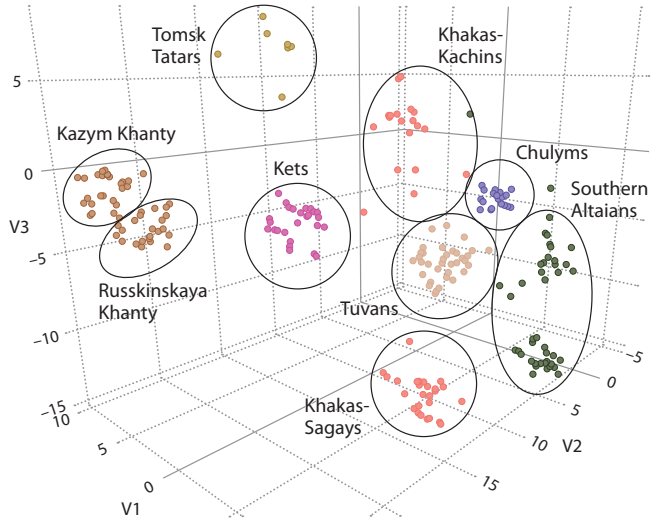

Fig. 1. Differentiation of the genomes of the population of Southern and Western Siberia by three components of tSNE.

were in genetic isolation for a very long time. This suggests that the ancient Ugric population of the modern territory of the Khanty settlement did not mix with other ethnic groups and confirms the absence of other groups of migrants from the territory of Southern Siberia and the steppe zone.

The result obtained shows that the overall picture of the distribution of the components is in good agreement with the geographical location of the studied populations, binding to a specific region, anthropological and linguistic differences. This information makes it possible to more accurately judge the similarities and differences between the compared populations, the composition of ancestral components, as well as the process of formation of their gene pool.

Identical in origin clutch blocks. As a result of bioinformatics processing of genotyping data from high-density biochips of various Siberian populations, an analysis was made of the coincidence of DNA fragments common in origin between populations and individuals. A segment with identical nucleotide sequences is IBD in two or more individuals if they inherit it from a common ancestor without recombination, that is, in these individuals the segment has a common origin. The expected length of an IBD segment depends on the number of generations since the last common ancestor. One of the applications of the analysis of genome regions of common origin is the quantitative assessment of the degree of relationship between individuals, which can also supplement information on the genetic relationships of populations (Gusev et al., 2012).

The samples of the Khanty showed the maximum match in IBD blocks with each other (6 %), then with a sample of the Kets (1.45 %), Chulyms (0.71 %), Tuvans (0.35 %), Tomsk Tatars (0.33 %), Khakas Kachins (0.32 %), and Southern Altaians (0.28 %). At the same time, among the Khanty, a greater coincidence of IBD blocks is observed in Russkinskaya (23.5 %), compared with Kazym (18.1 %).

The degree of overlap of IBD blocks between the Khanty, Kets, and Tomsk Tatars is consistent with the results of tSNE and ADMIXTURE in terms of the distribution of allele fre-

quencies and common genetic components in these populations. At the same time, in Khanty population from Russkinskaya, who have the largest sum of average lengths of IBD segments between pairs of individuals, the greatest contribution is made by IBD longer than 10 cm (42–46 %), which indicates a strong recent inbreeding within the population. To confirm this, the FROH inbreeding coefficient was calculated for all individuals for the three classes of homozygosity blocks (ROH). For the West Siberian populations, the Chulym population (0.0292), the Kazym (0.0280) and Russkinskaya Khanty (0.0266) and Kets (0.0259) populations, which are close in value, have the maximum values. Among the South Siberian populations, including the Altaians, Tomsk Tatars, Tuvans and Khakas, the maximum value was also found for the sample of Khakas-Sagays from the foothill Tashtyp region (0.0318), twice as high as the Khakas-Kachins of the plain Shirinsky region. The minimum value is typical for the Tomsk Tatars (0.0071).

There is a high correlation for FROH > 1.5 with the sum of mean IBD segment lengths (IBD > 1.5 cM) between pairs of individuals within Siberian populations ($r = 0.9246$, $p < 5.612e\text{-}09$). To calculate the Spearman correlation coefficient, cor.test was used in the R program. The ratio of the sum of the average lengths of IBD segments (IBD > 1.5 cM) between pairs of individuals to the coefficient of genomic inbreeding (FROH > 1.5) in the Russkinsskaya Khanty is higher than in Kazym Khanty. These indicators of genomic inbreeding and distribution of IBD lengths within Khanty populations are in good agreement with their territorial isolation and confirm the absence of recent gene flows between populations for hundreds of years.

Haplogroups of the Y-chromosome. As a result of the analysis of the frequency of occurrence of the used SNP markers in the studied samples of the Khanty, eight haplogroups of the Y-chromosome were identified. According to the composition and frequencies of haplogroups, the samples of Russian and Kazym Khanty men are very different from each other. Only two haplogroups are present in both samples (see the Table).

Thirty-nine samples belong to the N1a2b1b1 subline in the Russkinskaya Khanty, and only three in the Kazym ones. Terminal for this line, the Khanty have SNPs Y68212, Y70717, Y70315, Y70327. This Khanty subline is close to the N1a2b1b1 variants in the Chulyms (VL65, Z35095, Z35099, Z35102) and Khakas-Kachins (Z35093, Z35097, Z35103) (Valikhova et al., 2022).

The haplogroup N1a2b1b1 among the Khanty is ethnospecific and does not coincide in terminal SNPs and haplotypes with the dominant among the Nenets N1a2b1b1a~ (B171, B170, Z35091, Z35092) (Kharkov et al., 2021).

A feature of the ethnic composition of the majority of the South Siberian peoples is the presence of clans (seoks), where kinship is counted along the male line. Such a generic structure is typical for the Shors, Khakas, northern and southern Altaians, and Teleuts. All other samples of men from various West and South Siberian populations (the Enets, Khakas-Sagays, Shors, Chelkans and Tuvans, as well as the Khakas seoks formerly part of the Beltirs and Biryusins, assimilated in the late 19th and early 20th centuries) belong to others sublines of haplogroup N1a2b. The median network of haplotypes (Fig. 2) demonstrates a stellate phylogeny in the Khanty with a recent founder effect and a predominance of the ancestral haplotype in frequency.

The specific cluster of Khanty haplotypes is equidistant from all seoks of the Khakas-Kachins. The age of this cluster among the Khanty was 858 years (SD = 338 years), which is approximately one and a half to two times higher than the age of the clusters of the Kachin seoks Khaskha – 487 years (SD = 153 years), Yzyr – 501 years (SD = 203 years), Sokhkhy – 585 years (SD = 215 years) (Kharkov et al., 2020) and Chulym Turks 667 years (SD = 194 years). Thus, the Khanty in this haplogroup have a direct genetic connection with the Kachins, Chulyms and Nenets, whose ancestral lines diverged quite a long time ago and reflect their connection with the peoples of the Samoyedic language group.

The second haplogroup N1a2b2a1 (VL97, L1419, Y3185, Y3188, Y3189, Y3190, Y111190) is common for two Khanty samples (previously designated as the European N1b-E lineage). This subline was found among the Bashkirs, Kazan Tatars, Komi, Mari, Karelians, Vepsians, Finns and Russians (https://www.yfull.com/). Phylogenetically closest to the Khanty along this line are the Komi samples. Ethnospecific branches of the Khanty and Komi unite SNPs Y65017 and Y89655, not found in other populations. The Khanty and Komi have the least ancient common ancestor for this haplogroup, compared to other European populations.

Frequency of occurrence of Y-chromosome haplogroups in the Khanty

Haplogroup	Village of Russkinskaya % (*N*)	Village of Kazym % (*N*)
N1a2b1b1 (Y68212, Y70717, Y70315, Y70327)	60.9 (39)	5.5 (3)
N1a2b2a1 (VL97, L1419, Y3185, Y3188, Y3189, Y3190, Y111190)	9.4 (6)	16.6 (9)
N1a2b2b1~ (Z35076)	–	5.5 (3)
N1a1a1a1a2a1c1~ (Y13850, Y13852)	–	24.1 (13)
N1a2b1b1b1~ (B172, Z35108)	–	9.2 (5)
Q1b1a3b1a2~ (Z35974 xB32, B33, Z35993)	–	38.9 (21)
Q1a2b~ (M25, L716, YP1674, YP1676)	4.7 (3)	–
R1a1a1b2 (Y43850, S7280, FGC687, FGC38304)	25.0 (16)	–

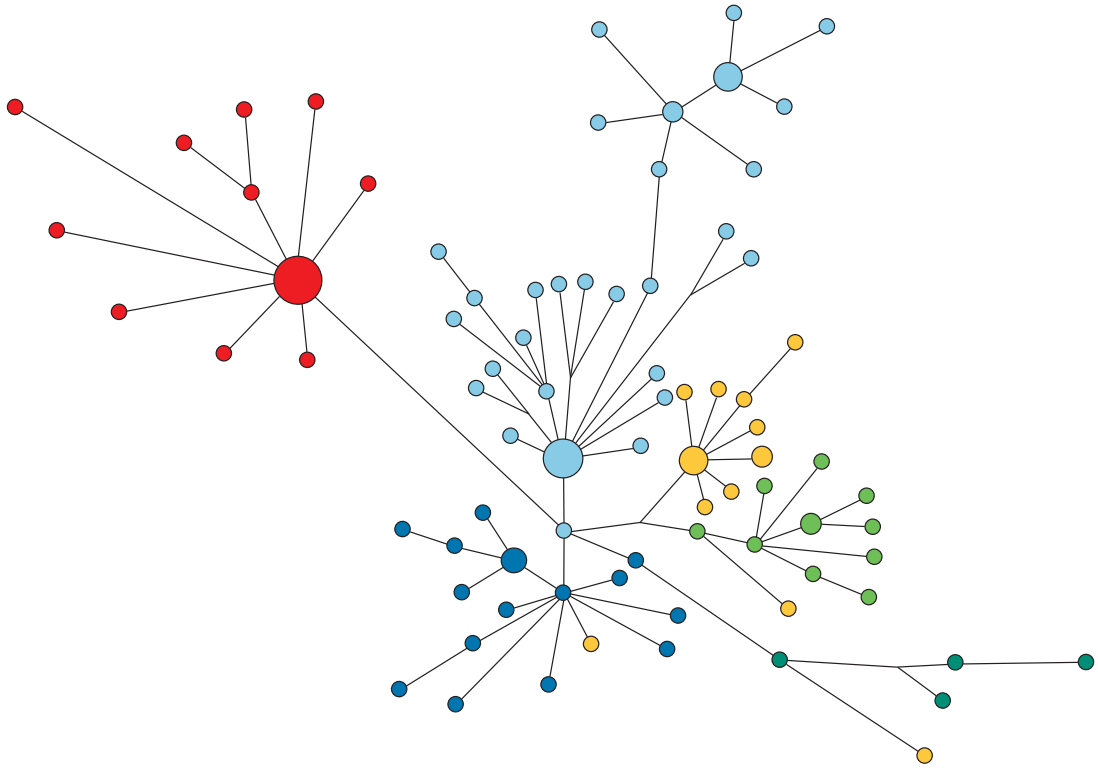

Fig. 2. The median network of YSTR haplotypes of the N1a2b1b1 haplogroup in the Khanty, Chulyms and Khakas-Kachins.
The Khanty are marked in light blue, the Chulyms are in red, the Khakas of the Sokhkhy seok are in blue, the Khakas of the Yzyr seok are in green, and the yellow are Khakases seok Hhaskha, dark green – seok Purut.

According to the YFull website, this branch split from the ancestral line about 2800 years ago. Theoretically, there are two options for the appearance of this haplogroup among the modern Khanty and Komi: 1) inheritance from a common ancient ancestral group of Ugric tribes; 2) the recent mixing of Khanty with ethnic Komi migrants to Siberia. However, the results of the analysis of genomic data using NGSadmix, ADMIXTURE, IBD blocks and differences in terminal SNPs of Y-chromosomes do not confirm the second variant. The YSTR haplotypes of this line in the Khanty and Komi also differ by several mutations. Previously, V.N. Pimenoff et al. suggested in their work that when the Ob-Ugric Khanty and Mansi went to the western slopes of the Ural Mountains and to the north-west of Siberia, a unique association N1b-A and N1b-E was formed (Pimenoff et al., 2008). This combination of N1b sublines in the Khanty and Mansi suggests a recent confluence of the western and eastern lineages in North Western Siberia. Our new data do not contradict this version.

All other haplogroups are represented only in individual samples of the Khanty. The haplogroup N1a2b2b1~ (Z35076) includes three samples of the Kazym Khanty. The lineage N1a2b2b1~ (B528, Y24382, Z35076, Z35077) closest to it is also common among the Komi. The Udmurts, Tatars, Chuvashs and Bashkirs have its more modern line (B226). The YSTR haplotypes of this haplogroup in the Komi and Udmurts are closer to each other than to the Khanty samples. The presence among the Khanty and Komi of two haplogroups, N1a2b2a1 and N1a2b2b1~, with ethnospecific terminal SNPs and different haplotypes indicates their inheritance from fairly ancient common ancestors, most likely part of the early Ugric population of these territories.

Thirteen samples of the Kazym Khanty belong to the haplogroup N1a1a1a1a2a1c1~ (Y13850, Y13852). Seven of them have the surname Pyak, which is Nenets in origin, referring to the Forest Nenets. All seven of these samples have very close haplotypes and are descendants of a relatively recent common Nenets ancestor. In the questionnaires of these men, who consider themselves Khanty, Nenets ancestors were indicated on the paternal line with different depths. The remaining six men of this haplogroup differ in haplotypes from the Pyak genus.

In our study of the Taz Nenets (Kharkov et al., 2021), it was found that all men representing the Khanty origin of the Salinder, Lar and Tibichi clans completely belong to this haplogroup. Representatives of these genera formed in the XVIII–XIX centuries in the lower reaches of the Ob as a result of the development of the Nenets large-herd reindeer husbandry and the involvement of part of the northern Khanty in it (Kvashnin, 2003). All haplotypes of the Kazym Khanty of this haplogroup differ significantly from the haplotypes of the Taz Nenets.

The other five samples of the Kazym Khanty belong to the haplogroup N1a2b1b1b1~ (B172, Z35108). All previously surveyed Nenets men from the Vanuito phratry belonging to the Vanuito, Puiko and Yaungat clans, and the Purungui clan of Khanty origin, belong to it. Four samples of the Khanty differ in haplotypes from the Nenets, but one almost completely coincides with them. Such a division into haplotypes specific to the Khanty and close to the Nenets coincides with the data

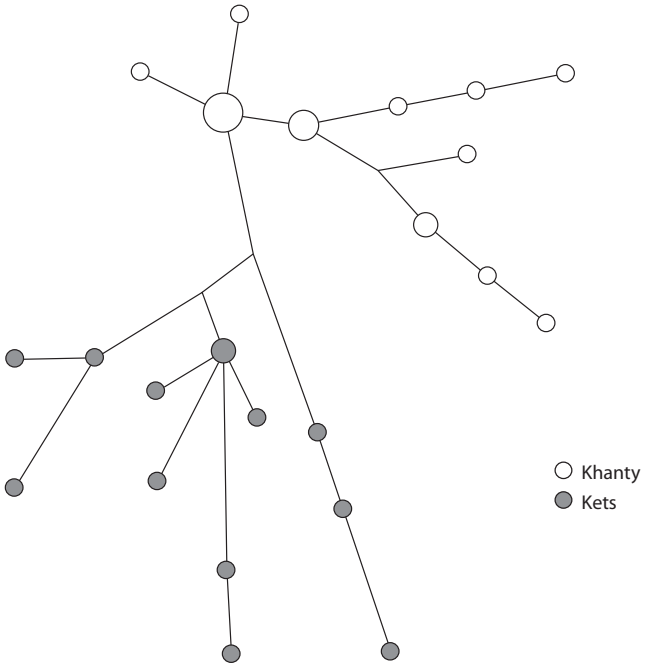

Fig. 3. Median network of YSTR haplotypes of haplogroup Q1b1a3b1a2~ in Khanty and Kets.

on the haplogroup N1a1a1a1a2a1c1~. It is obvious that the gene pool of the Kazym Khanty includes precisely the variants of these haplogroups of Khanty origin, but relatively recently marriages were also made with the Forest Nenets. The absence of these haplogroups in the Russkinskaya Khanty is in good agreement with the data on the distribution of IBD blocks and the coefficient of genomic inbreeding.

The distribution of various haplogroups of the N clade of the Y-chromosome in the studied populations is in good agreement with the frequency of the Ugric genetic component. Phylogenetic analysis of Y-chromosomal sublines and haplotypes of various haplogroups of the N clade shows that the center of origin and distribution of the carriers of the Ugric component in Southern, Western Siberia and Eastern Europe is the territory of modern Altai and Sayan Mountains. The obtained results are well comparable with the data of ethnology, anthropology and linguistics on the contribution of the Uralic component to the formation of various peoples of the Altai-Sayan and the historical areas of Ugric and other languages of the Uralic language family.

Almost 40 % of men from Kazym belong to the haplogroup Q1b1a3b1a2~ (Z35974 xB32, B33, Z35993). The lineage Q1b1a3b1a2~ (B33, Z35991) specific to the Kets population is closest to it. In addition to the Kets, this variant also prevails among the Selkups from the Tomsk Region and the Krasnoyarsk Region. A more distant line Q1b1a3b1a~ (B30, YP1693 xZ35991) is common in Tuvan populations, with a maximum frequency in the eastern mountainous regions of Tuva (up to 25 %). Khanty samples show a specific haplotype spectrum with a recent founder effect that is not observed in the Kets (Fig. 3).

The distribution of these sublines in the populations of the Khanty, Kets, and Tuvans is in good agreement with the shares of matches in IBD blocks between them, the tSNE plot, and the distribution of the Ugric genetic component in these populations over the autosomal part of their gene pool. The presence of this lineage among the Khanty is not due to recent borrowing from other aboriginal populations (Kets and Selkups), but to the fact that it was already part of the settling ancestral groups.

Three men from the village of Russkinskaya have a completely different haplogroup of the Q clade – Q1a2b~ (M25, L716, YP1674, YP1676). This is a very rare haplogroup not found in other Siberian populations. It is presented with the maximum frequency among the ethnic Turkmens of Karakalpakstan, Iran and Afghanistan (Grugni et al., 2012; Skhalyakho et al., 2016). In most other ethnic groups, its frequency is very low. Khanty haplotypes are quite different from other populations. Most likely, the presence of this line among them is not a consequence of recent miscegenation, but is a legacy of the Ugric groups that migrated from southern Siberia and the Urals to the north.

The last haplogroup, which includes 16 Khanty men from the village of Russkinskaya, is R1a1a1b2-Y43850. The haplotypes of all samples are quite close, which indicates a recent founder effect (Fig. 4).

Khanty-specific terminal SNPs are S7280, FGC687, and FGC38304. The R1a1a1b2-Y43850 variants closest to this lineage are represented with a high frequency in the Khakas and Shors, and less frequently in the Tuvans and Northern Altaians. According to YFull, this haplogroup is approximately 3800 years old. All of these patterns belong to four different lineages that split a long time ago. The age of the haplotype cluster in the Khanty was 933 years (SD = 336 years), which is approximately one and a half times less than the age of the South Siberian lines. The Khakas seok Piltir is 1469 years old (SD = 342 years) (Y39884 xY43109). The lineage of this haplogroup (Y62155.2) specific for the Biryusa Khakas seoks of Turan, Khyzyl Khaya and Shor seoks of Tartkyn, Shor-Kyzai and Kara-Shor has approximately the same age – 1315 years (SD = 227 years). The branch with a wider distribution in the Sayan-Altai populations (Y43109) is even older – 1566 years (SD = 350 years). The difference in SNP and STR among the Khakas, Shors, Tuvans, and Northern Altaians is greater than with the Khanty.

A strong heterogeneity of the studied samples of the Khanty in terms of the composition and frequencies of various haplogroups is shown. The phylogeny of various lineages of two haplogroups, N1a2b1b1 and R1a1a1b2-Y43850, indicates their South Siberian origin in the Khanty gene pool. The territory of the Sayan and Altai was the primary focus of the generation of diversity and the expansion of the number of ancestral groups of carriers of these haplogroups in Siberia. It is most likely that the distribution of most Y-chromosome haplogroups among the Khanty occurred during the initial settlement of the Ugric tribes to the north and west.

It is necessary to take into account the fact that the range of modern Khanty is located to the north of the territory of their ancestors. The West Siberian and Volga-Ural regions were the place of secondary generation of diversity, but not the formation of the N1a2 haplogroup itself. At the moment, there is no final opinion regarding the place of formation of the ethnoi of the Uralic language family, but numerous data, including the results of studies of the phylogeny and phylo-

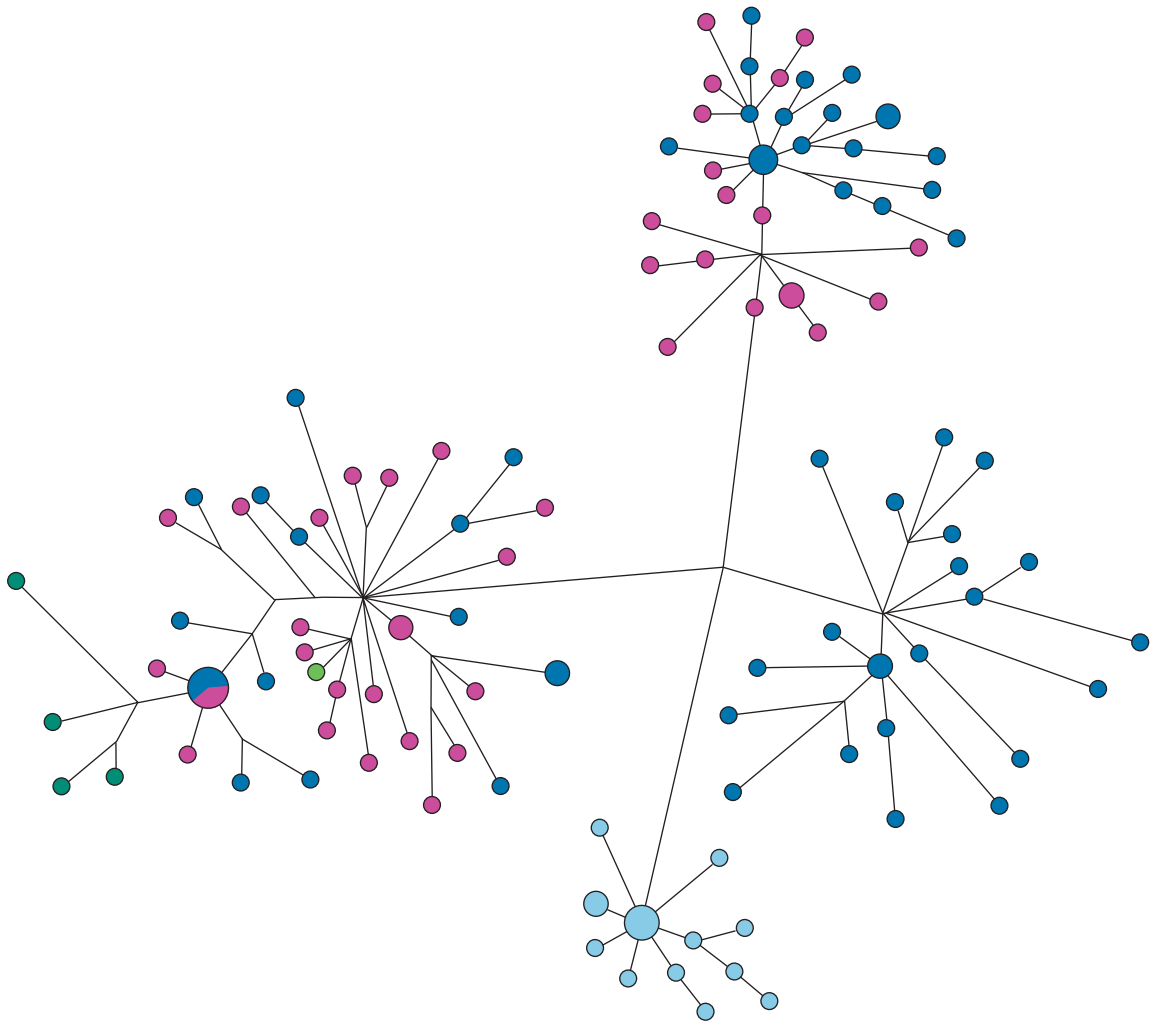

Fig. 4. Median network of YSTR haplotypes of haplogroup R1a1a1b2-Y43850 in the Khanty, Khakas, Shors, Tuvans and Altaians. Khanty are in light blue, Khakas are in blue, Shors are in crimson, Tuvans are in dark green, Altaians are in green.

geography of clade N haplogroups, point to Southern Siberia. Linguistic paleontology points to the Proto-Ural ecological area as a territory limited in the west by the Ural Range, in the north by approximately the Arctic Circle, in the east by the area of the lower reaches of the Angara and Podkamennaya Tunguska and the middle reaches of the Yenisey, in the south by approximately the modern southern border of the West Siberian taiga from the northern foothills of the Sayan and Altai to the lower reaches of the Tobol and the Middle Urals inclusively (Napolskikh, 2018).

Conclusion

Thus, the gene pool of the two Khanty populations is a heterogeneous set of Y-chromosome haplogroups, but very similar in autosomal markers. The expanded composition of terminal SNPs for the identified haplogroups made it possible to describe in detail and clarify the differences in the phylogeny and structure of individual ethnospecific sublines, to determine their relationship, and traces of population expansion in the Khanty gene pool. The results of a comparative analysis of male samples indicate a close genetic relationship of the Khanty with the Altai-Sayan Khakas and Tuvans, as well as with the Nenets, Komi, Udmurts and Kets. The specificity of haplotypes and the detection of various terminal SNPs indicate that the Khanty did not come into contact with other ethnic groups for a long time. The only exception is the Nenets, which included many Khanty clans. For the northern population of the Kazym Khanty, Y-chromosomal lines show a small contribution of the Forest Nenets.

The results obtained do not contradict the generally accepted versions of the Khanty ethnogenesis, but allow us to take a fresh look at this process. The main factor in the formation of the Khanty gene pool was their territorial genetic isolation and later mixing with the newcomer Samoyed population, which, when switching to tundra reindeer husbandry, led to a strong demographic growth of their clans as part of the Nenets. The relatively low genetic diversity in autosomal SNPs and the rather high level of inbreeding in the Khanty confirm this. New information about the structure of the Khanty gene pool is an important addition to the existing anthropological, archaeological, ethnological and linguistic data on their formation and kinship with other peoples.

References

Alexander D.H., Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 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.094052.109.

Bandelt H.J., Forster P., Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 1999;16(1):37-48. DOI 10.1093/oxfordjournals.molbev.a026036.

Brook S.I. The World Population. Ethnodemographic Reference Book. Moscow: Nauka Publ., 1986. (in Russian)

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.

Grugni V., Battaglia V., Hooshiar Kashani B., Parolo S., Al-Zahery N., Achilli A., Olivieri A., Gandini F., Houshmand M., Sanati M.H., Torroni A., Semino O. Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. *PLoS One.* 2012;7(7):e41252. DOI 10.1371/journal.pone.0041252.

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.

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.

Kharkov V.N., Novikova L.M., Shtygasheva O.V., Luzina F.A., Khitrinskaya I.Y., Volkov V.G., Stepanov V.A. Gene pool of Khakass and Shors for Y chromosome markers: common components and tribal genetic structure. *Russ. J. Genet.* 2020;56(7):849-855. DOI 10.1134/S1022795420070078.

Kharkov V.N., Valikhova L.V., Yakovleva E.L., Serebrova V.N., Kolesnikov N.A., Petelina T.I., Khitrinskaya I.Yu., Stepanov V.A. Reconstruction of the origin of the Gydan Nenets based on genetic analysis of their tribal structure using a new set of YSTR markers. *Russ. J. Genet*. 2021;57(12):1414-1423. DOI 10.1134/S1022795421120061.

Kvashin Yu.N. Gydan Nenets: the History of the Formation of the Modern Generic Structure (18-20th Centuries). Moscow: IEA RAS Publ., 2003. (in Russian)

Napolskikh V.V. Essays on Ethnic History. Kazan: Khalikov Institute of Archeology, 2018. (in Russian)

Peoples of West Siberia: Khanty. Mansi. Selkups. Nenets. Enets. Nganasans. Kets. Moscow: Nauka Publ., 2005. (in Russian)

Pimenoff V.N., Comas D., Palo J.U., Vershubsky G., Kozlov A., Sajantila A. Northwestern Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. *Eur. J. Hum. Genet.* 2008;16(10):1254-1264. DOI 10.1038/ejhg.2008.101.

Skhalyakho R.A., Zhabagin M.K., Yusupov Yu.M., Agdzhoyan A.T., Sabitov Zh.M., Gurianov V.M., Balaganskaya O.A., Dalimova D.A., Davletchurin D.Kh., Turdikulova Sh.U., Chukhryaeva M.Ch., Asilgujin R.R., Akilzhanova A.R., Balanovsky O.P., Balanovska E.V. Gene pool of Turkmens from Karakalpakstan in their Central Asian context (Y-chromosome polymorphism). *Vestnik Moskovskogo Universiteta. Seriya 23. Antropologiya = Moscow University Anthropology Bulletin*. 2016;3:86-96. (in Russian)

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.

The Peoples of Russia: Encyclopedia. Moscow: Bol'shaya Rossiyskaya Entsyklopedia 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.

Zhivotovsky L.A., Underhill P.A., Cinnioglu 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. 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.

ORCID ID

V.N. Kharkov orcid.org/0000-0002-1679-2212
N.A. Kolesnikov orcid.org/0000-0001-8855-577X
V.A. Stepanov orcid.org/0000-0002-5166-331X

Acknowledgements. The study was supported by the Russian Science Foundation grant No. 22-64-00060 (https://rscf.ru/project/22-64-00060/).

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

Received October 14, 2022. Revised November 23, 2022. Accepted November 27, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):55-62
DOI 10.18699/VJGB-23-08

Original Russian text https://vavilovj-icg.ru/

Blocks identical by descent in the genomes of the indigenous population of Siberia demonstrate genetic links between populations

N.A. Kolesnikov ✉, V.N. Kharkov, K.V. Vagaitseva, A.A. Zarubin, V.A. Stepanov

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ nikita.kolesnikov@medgenetics.ru

Abstract. The gene pool of the indigenous population of Siberia is a unique system for studying population and evolutionary genetic processes, analyzing genetic diversity, and reconstructing the genetic history of populations. High ethnic diversity is a feature of Siberia, as one of the regions of the peripheral settlement of modern human. The vast expanses of this region and the small number of aboriginal populations contributed to the formation of significant territorial and genetic subdivision. About 40 indigenous peoples are settled on the territory of the Siberian historical and ethnographic province. Within the framework of this work, a large-scale population study of the gene pool of the indigenous peoples of Siberia was carried out for the first time at the level of high-density biochips. This makes it possible to fill in a significant gap in the genogeographic picture of the Eurasian population. For this, DNA fragments were analyzed, which had been inherited without recombination by each pair of individuals from their recent common ancestor, that is, segments (blocks) identical by descent (IBD). The distribution of IBD blocks in the populations of Siberia is in good agreement with the geographical proximity of the populations and their linguistic affiliation. Among the Siberian populations, the Chukchi, Koryaks, and Nivkhs form a separate cluster from the main Siberian group, with the Chukchi and Koryaks being more closely related. Separate subclusters of Evenks and Yakuts, Kets and Chulyms are formed within the Siberian cluster. Analysis of SNPs that fell into more IBD segments of the analyzed populations made it possible to compile a list of 5358 genes. According to the calculation results, biological processes enriched with these genes are associated with the detection of a chemical stimulus involved in the sensory perception of smell. Enriched for the genes found, molecular pathways are associated with the metabolism of linoleic, arachidonic, tyrosic acids and by olfactory transduction. At the same time, an analysis of the literature data showed that some of the selected genes, which were found in a larger number of IBD blocks in several populations at once, can play a role in genetic adaptation to environmental factors.
Key words: IBD; human populations; Siberian populations.

For citation: Kolesnikov N.A., Kharkov V.N., Vagaitseva K.V., Zarubin A.A., Stepanov V.A. Blocks identical by descent in the genomes of the indigenous population of Siberia demonstrate genetic links between populations. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):55-62. DOI 10.18699/VJGB-23-08

Идентичные по происхождению блоки в геномах коренного населения Сибири демонстрируют генетические связи между популяциями

Н.А. Колесников ✉, В.Н. Харьков, К.В. Вагайцева, А.А. Зарубин, В.А. Степанов

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ nikita.kolesnikov@medgenetics.ru

Аннотация. Генофонд коренного населения Сибири представляет собой уникальную систему с точки зрения исследования популяционно- и эволюционно-генетических процессов, анализа генетического разнообразия и реконструкции генетической истории популяций. Высокое этническое разнообразие является особенностью Сибири как одного из регионов периферийного расселения современного человека. Огромные пространства этого региона и малочисленность аборигенного населения способствовали формированию значительной территориальной и генетической подразделенности. На территории сибирской историко-этнографической провинции расселены около 40 коренных народностей. Проведено масштабное популяционное исследование генофонда коренных народов Сибири на уровне высокоплотного ДНК-микрочипа Infinium Multi-Ethnic Global-8, позволяющее заполнить существенный пробел в геногеографической картине населения Евразии. Для этого были отобраны и проанализированы фрагменты ДНК, унаследованные без рекомбинации каждой парой индивидов от их недавнего общего предка, т. е. сегменты (блоки), идентичные по происхождению (IBD). Распределение блоков IBD в популяциях Сибири хорошо согласуется с географической близостью популяций и их языковой принадлежностью. Чукчи, коряки и нивхи среди сибирских популяций формируют отдельный от основной группы Сибири кластер, причем чукчи и коряки являются более

близкородственными. Образуются отдельные субкластеры эвенков и якутов, кетов и чулымцев, тувинцев и алтайцев внутри сибирского кластера. Анализ SNP, которые попадали в большее количество IBD-сегментов анализируемых популяций, позволил составить список из 5358 генов. По результатам расчета, обогащенные этими генами биологические процессы связаны с обнаружением химического раздражителя, участвующего в сенсорном восприятии запаха. Обогащенные найденными генами молекулярные пути связаны с метаболизмом линолевой, арахидоновой, тирозиновой кислот и путем обонятельной трансдукции. При этом анализ литературных данных показал, что некоторые из отобранных генов, которые встречались в большем количестве блоков IBD сразу в нескольких популяциях, могут играть роль в адаптации человека к факторам окружающей среды.
Ключевые слова: IBD; популяции человека; сибирские популяции.

Introduction

Genetic and demographic processes in populations, population fluctuations, cross-breeding events, migrations and natural selection affect the structure of genetic diversity in the genomes of individuals and populations as a whole. In particular, genetic and demographic processes lead to the formation of linkage blocks of common origin (identity by descent, IBD). A segment having identical nucleotide sequences is IBD in two or more individuals if they have inherited it from a common ancestor without recombination, that is, in these individuals the segment has a common origin. The expected length of an IBD block depends on the number of generations that have passed since the segment appeared in the last common ancestor (Browning S.R., Browning B.L., 2010; Palamara et al., 2012).

IBD segments can be used to reveal the demographic history of populations, including bottleneck effects and gene flows in populations (Gusev et al., 2012). Recent studies have shown differences in IBD distribution between African, Asian, and European populations, as well as IBD segments shared with ancient genomes such as those of Neanderthals and Denisovans (Hochreiter, 2013).

Close relatives have rather long DNA fragments identical with each other and, accordingly, in most chromosomes there are blocks of considerable length identical by descent (> 66.7 cM), as a result of which the expected length of the total IBD is about 1700 cM. Cousins and second cousins are expected to have multiple regions (more than 2.5 expected segments) due to the presence of recent ancestors determining their relationship. For first cousins, each IBD is expected to have a total length greater than 62 cM and for second cousins, 25 cM. Distant cousins that are fourth cousins or more distant are very likely to carry one or more regions from their nearest common ancestor. Such couples include the vast majority of people in a particular population and are usually referred to as "unrelated" because the proportion and number of IBD across the genome is expected to be relatively small between them (Gusev et al., 2012).

IBD segments can also help in the detection of natural selection signals in the human genome. Searching for regions with an excess of IBD segments allows the identification of genomic regions in the human genome that are under very recent and strong selection, since selection generally increases the number of IBD segments among individuals in a population (Albrechtsen et al., 2010; Han, Abney, 2011).

Regarding the populations of the indigenous ethnic groups of Siberia, it has been suggested that large-scale dispersal and mixing of populations probably may explain the unusually high proportion of IBD between populations (Pugach et al., 2016).

The purpose of this study was to analyze the structure of the gene pool of populations of the indigenous population of Siberia, based on the identification of linkage blocks identical by descent, and their intra- and interpopulation distribution.

Materials and methods

Genome-wide genotyping data were generated using Infinium Multi-Ethnic Global-8 microarrays (Illumina) with over 1.7 million markers. Samples with more than 5 % missing positions, as well as SNPs with more than 10 % missing genotypes, were excluded from the analysis. The data were preliminarily filtered by the minimum rare allele frequency (MAF, minor allele frequency > 0.01). As a result, 886,889 autosomal SNPs were included in the final data set.

Populations of the indigenous peoples of Siberia ($N = 477$) are represented by Altaians (B – the village of Beshpeltir, Chemalsky district, $N = 24$ and K – the village of Kulada, Ongudaysky district, $N = 25$), Buryats (A – the village of Aginskoye, Aginsky district, $N = 23$ and K – the village of Kurumkan, Kurumkansky district, $N = 28$), Kalmyks ($N = 29$), Kets ($N = 15$), Koryaks ($N = 20$), Chukchi ($N = 25$). The Koryak material was collected in the Koryak Autonomous District of the Kamchatka Region. A population sample of the Chukchi whose blood samples were collected in the villages of Lorino, Sireniki, Yanarykot and Novoe Chaplino of the Chukotka Autonomous Okrug belongs to the coastal group, Nivkhs ($N = 13$), Tatars (T – Tomsk, $N = 20$), Tuvans ($N = 28$), Udeges ($N = 15$), Khantami (K – the village of Kazym, Beloyarsk district, $N = 30$ and R – the village of Russkinskaya, Surgut district $N = 26$), Khakases (T – Sagais of the Tashtyp district $N = 29$ and S – Kachins of the Shirinsky district $N = 26$), Chulyms ($N = 22$), Evenks (Z – Transbaikalian (Chara village of the Kalarsky district, Moklakan village and Tupik village of the Tungiro-Olyokma region) $N = 25$ and Y – Yakut Evenks ($N = 28$) and Yakuts ($N = 26$).

The material was deposited in the bioresource collection "Biobank of the Population of Northern Eurasia". The characteristics of the studied populations are presented in Table 1.

Phasing of genotypes was carried out using Beagle 4.1 software (Browning S.R., Browning B.L., 2007). The Refined IBD algorithm, refined-ibd.16May19.ad5.jar version (Browning B.L., Browning S.R., 2013), was used to analyze genome blocks identical by descent. To compare the populations, the sums of the average lengths of the IBD segments between pairs of individuals were obtained for the following length ranges: 1.5–1.999 cM, 2–3.999, 4–7.999, 8–15.999 and >16 cM (for convenience, these ranges are referred to further in the text as 1.5–2 cM, 2–4, 4–8 and 8–16 cM). A heat map with a dendrogram based on the logarithm of the sum of the average

Table 1. Characteristics of the studied population samples of the indigenous population of Siberia

Population	Location	Sample size	Language affiliation (family/group)	Anthropological type
Altaians (B)	Chemalsky District	24	Altaic/Turkic	Mongoloid (Central Asian)
Altaians (K)	Ongudaysky District	25		
Buryats (A)	Aginsky District	23	Altai/Mongolian	
Buryats (K)	Kurumkansky District	28		
Chukchi	Chukotka Autonomous Okrug	25	Chukchi-Kamchatka languages	Mongoloid (Arctic)
Chulyms	Tomsk Oblast	22	Altaic/Turkic	Uralic, mongoloid (South Siberian)
Evenks (Y)	Republic of Sakha (Yakutia)	28	Altaic/Tungus-Manchu languages	Mongoloid (Baikal)
Evenks (Z)	Zabaykalsky Krai	25		
Kalmyks	Republic of Kalmykia	29	Altai/Mongolian	Mongoloid (Central Asian)
Kets	Krasnoyarsk Krai	15	Yenisei	Yenisei
Khakas (S)	Kachintsy, Altaysky District	26	Altaic/Turkic	Uralic, mongoloid (South Siberian)
Khakas (T)	Sagays, Tashtypsky District	29		
Khanty (K)	Beloyarsky District	30	Ural/Ob-Ugric	Ural
Khanty (R)	Surgutsky District	26		
Koryaks	Kamchatka Krai	20	Chukchi-Kamchatka languages	Mongoloid (Arctic)
Nivkhs	Sakhalin Oblast	13	Paleoasian language	Mongoloid (Sakhalino-Amur)
Tatars (T)	Tomsk Tatars	20	Altaic/Turkic	Uralic, mongoloid (South Siberian)
Tuvans	Tyva Republic	28		Mongoloid (Central Asian)
Udege	Primorsky Krai	15	Altaic/Tungus-Manchu languages	Mongoloid (Baikal)
Yakuts	Republic of Sakha (Yakutia)	26	Altaic/Turkic	Mongoloid (Central Asian)

IBD segment lengths between pairs of individuals was built using the heatmap.2 package in the R software environment.

We also identified SNPs that fell into a larger number of IBD segments of the analyzed populations (the frequency of SNPs in IBD was higher than the 99th quantile of the frequency distribution), determined the belonging of these SNPs to genes, and assessed the biological significance of the resulting list of these genes. For this analysis, we used the WebGestalt web resource (WEB-based Gene SeT AnaLysis Toolkit); in particular, the analysis of KEGG paths and gene ontologies (Gene Ontology) was conducted using the ORA (over-representation analysis) method.

Results and discussion

For a more detailed analysis of the genetic relationship of the Siberian populations and to find out to what extent their genetic structure can be explained by recent local migrations, we isolated and analyzed DNA fragments that were inherited without recombination by each pair of individuals from their recent common ancestor, that is, segments (blocks) identical by descent (IBD).

The populations inhabiting the territory of Siberia are characterized by a unique genetic and demographic history, which is reflected, among other things, in the distribution of IBD blocks both within the populations and between them. We calculated the sums of the average lengths of the IBD segments between pairs of individuals and, based on their results, built a heat map with a dendrogram based on the logarithm of the sum of the average lengths of the IBD segments (Fig. 1).

The number of common segments among representatives of different populations is consistent with the geography of their residence, since the peoples living nearby can be influenced by common genetic and demographic processes. Analysis of the heat map demonstrates the clustering of the populations of the Siberian group, linking peoples by place of origin. Among the Siberian populations, the Chukchi, Koryaks and Nivkhs form a separate cluster from the main group of Siberian populations, with the Chukchi and Koryaks being more closely related. Separate subclusters of Evenks and Yakuts, Kets and Chulyms, Tuvans and Altaians are formed within the Siberian cluster.

With the gradation of the IBD segments with different mean length, the trend generally remains, but some differences appear that more accurately characterize the recent admixture between the peoples. For longer IBDs, the clusters are more in line with the current geographic location of the populations, reflecting the recent exchange of common regions. In three length ranges with IBD sizes (1.52–2, 2–4, and 4–8 cM), populations are better divided into closely geographically located pairs: Koryak-Chukchi, Yakut-Evenki.

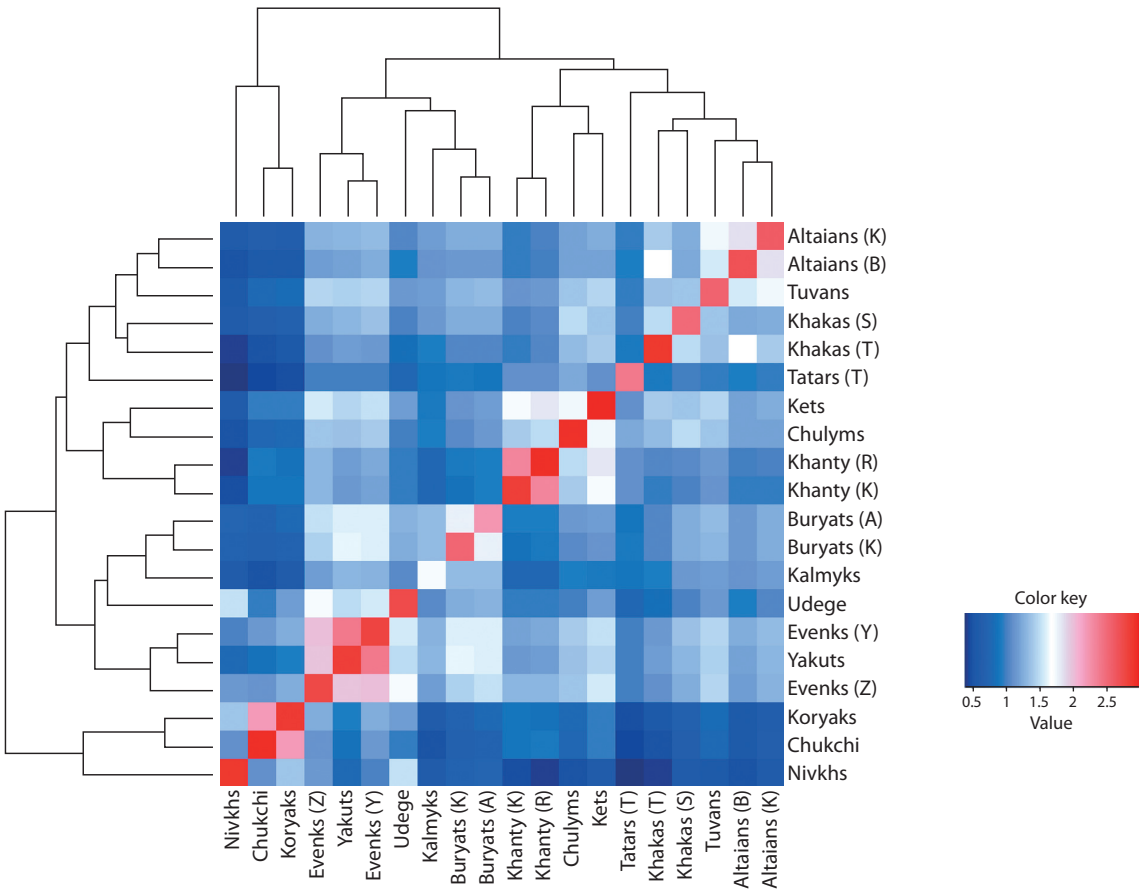

Fig. 1. Heatmap with dendrogram based on the logarithm of the sum of mean IBD segment lengths (>1.5 cM) between pairs of individuals.

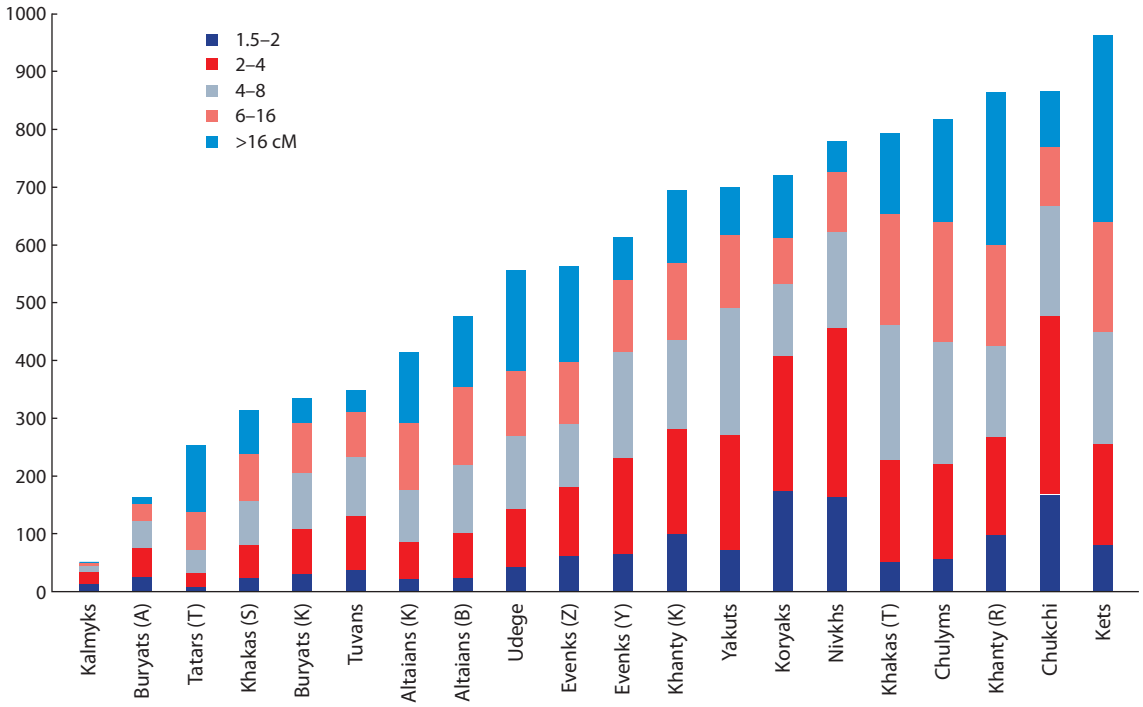

Fig. 2. Diagram of the sum of the average lengths of IBD segments between pairs of individuals in the studied populations for IBDs of different sizes (1.5–2, 2–4, 4–8, 8–16, >16 cM).

The Kets almost equally share IBD blocks with the Chulyms (18.7–27.2–7.7 cM) and Khanty (23.4–24.4–4.8 cM for Khanty (K) and 25.9–30.1–7.9 cM for the Khanty (R)), while among the Khanty, a greater value of common IBD blocks is observed in the Russian Khanty, which corresponds to their closer geographical location compared to the Khanty of the Beloyarsky district.

For the Khakas, who are more distant from the Kets, the values of the total IBD blocks are almost two to three times lower than for the Khanty (9.8–10.4–2.5 cM for the Khakas (S) and 9.1–10.6–3.8 cM for Khakas (T)). Despite the fact that the Evenks, Tuvans and Yakuts are even more remote, for the Tuvans (12.3–12.8–2.1 cM) and the Yakuts (12.8–12.5–2.1 cM), there are similar values of total IBD blocks, for the Trans-Baikal Evenks (16.4–16.9–3.0 cM), it is worth noting that the values of the IBD blocks are greater than with the Yakut Evenks (14.7–14.0–2.6 cM).

With few exceptions, common IBD segments between Siberian populations are better explained by the geographic proximity of the populations rather than by their linguistic affiliation. For example, the Yakut Evenks living in the territory of Yakutia have more IBDs in common with the Yakuts (252.7 cM) than with the Transbaikal Evenks (102.5 cM). At the same time, the sum of the average lengths of IBD segments between pairs of individuals between two populations of the Evenks is comparable to the sum between the Transbaikalian Evenks and Yakuts (95.2 cM).

In terms of intrapopulation IBD analysis, in general, individuals from populations of the Far North and Far East (Koryaks, Chukchi, Nivkhs) share more IBDs with specimens from the same group than individuals from the South Siberian populations such as the Altaians and Tuvans. At the same time, in the Chukchi, Koryaks, and Nivkhs, short IBD fragments of 1.5–4 cM (55–57–59 %) make the largest contribution, which may indicate a bottleneck in the past during migrations to the north and northeast and/or a strong isolation from other populations inhabiting the territory of Siberia. At the same time, in the Chulyms, Khanty (R), inhabiting the central part of Siberia and having the largest sum of average lengths of IBD segments between pairs of individuals, the largest contribution is made by IBDs longer than 8 cM (47–51 %), which indicates a strong recent inbreeding within the population.

The most genetically heterogeneous Siberian populations, which have minimal values for the genomic inbreeding coefficient (F_{ROH}) (Kolesnikov et al., 2021), also have minimal values for the sum of the average lengths of IBD segments. This is most pronounced in the populations of Kalmyks, Agin Buryats, and Tomsk Tatars (Fig. 2).

In Buryat populations, there are no significant differences in the average total length between pairs of individuals with other populations, but there is a significant difference in the distribution of IBD within populations. Thus, the Buryats (K) have a significantly larger average total length between pairs of individuals within the population (335.4 cM), compared with the Buryats (A) (163.5 cM), largely due to medium and long IBD. The Agin Buryats have a much higher proportion of short IBDs (16–30–28–19–7 %) than the Kurumkan Buryats (9–24–29–25–13 %). Despite the fact that for the Buryats, who have a large total population, with their numbers almost doubling from 237 thousand in 1926 to 461 thousand in 2010, this difference between populations can be explained by a sharp increase in the population of the village of Aginskoye from 451 people in 1908 up to 4556 people in 1939 and 15 thousand in 2010, and the stability of the population of the village of Kurumkan, with a population of 5617 people in 1979 and 5465 in 2010, located in one of the remote regions of Buryatia.

A similar dynamics is observed in the populations of the Chukchi (20–36–22–12–11 %), Koryaks (24–33–17–11–15 %), Nivkhs (21–38–21–13–7 %) and Kalmyks (29–39–22–7–3 %), followed by a sharper decrease in the proportion of long IBD. The Chukchi, Nivkhs and Koryaks, which have a small population (up to 13 thousand), are characterized by the absence of sharp fluctuations in the number of populations over the past hundred years with an increase of 29–9–6 %, respectively, which could also contribute to a reduction in the total long IBD segments subject to a small number of closely related marriages within the population.

A total of 189,314 SNPs were obtained for 20 populations, falling into the largest number of IBD segments of the analyzed populations (the frequency of SNPs falling into IBD is higher than the 99th quantile of the frequency distribution). Of these, 88,530 SNPs are located in intergenic regions, the rest are located in the region of 5358 genes. Table 2 shows the genes that were shown in four or more populations.

From the list of 5358 genes most frequently found in IBD blocks, 1694 were annotated using the KEGG database according to WebGestalt. As a result, analysis taking into account the Benjamini–Hochberg correction (FDR = 0.05) revealed molecular KEGG pathways enriched in these genes: the linoleic acid pathway hsa00591 (FDR = 0.0051) including 17 genes (*CYP2C8*, *CYP2C9*, *PLA2G1B*, *PLB1*, *CYP1A2*, *CYP2C19*, *CYP3A4*, *PLA2G10*, *PLA2G2A*, *PLA2G2C*, *PLA2G2D*, *PLA2G2E*, *PLA2G2F*, *PLA2G4A*, *PLA2G4C*, *PLA2G5*, *PLA2G6*), the arachidonic acid pathway hsa00590 (FDR = 0.0240) including 27 genes (*CYP2C8*, *CYP2C9*, *PLA2G1B*, *PLB1*, *ALOX12*, *ALOX12B*, *ALOX15B*, *ALOX5*, *CYP2B6*, *CYP2C19*, *GPX1*, *GPX3*, *GPX7*, *PLA2G10*, *PLA2G2A*, *PLA2G2C*, *PLA2G2D*, *PLA2G2E*, *PLA2G2F*, *PLA2G4A*, *PLA2G4C*, *PLA2G5*, *PLA2G6*, *PTGIS*, *PTGS1*, *PTGS2*, *TBXAS1*), tyrosine metabolism pathway hsa00350 (FDR = 0.0240) including 18 genes (*ADH1A*, *ADH1B*, *ADH1C*, *ADH4*, *ADH5*, *ADH6*, *ADH7*, *ALDH3B1*, *ALDH3B2*, *AOC2*, *AOC3*, *DDC*, *GOT1*, *HPD*, *IL4I1*, *PNMT*, *TYR*, *TYRP1*), olfactory transduction pathway hsa04740 (FDR = 4.55E-08) including 159 genes.

The metabolic conversion of polyunsaturated fatty acids (PUFAs) such as linoleic acid into biologically active long chain PUFAs (> 20 carbons, LC-PUFAs) such as arachidonic acid is essential for proper metabolism. LC-PUFAs and their metabolites are important structural and signaling components for numerous biological systems, including brain development and function, innate immunity, and energy homeostasis (Marszalek, Lodish, 2005; Calder, 2013). There are also food sources of preformed LC-PUFAs in eggs and some meats containing arachidonic acid (Horrocks, Yeo, 1999; Howe et al., 2006; Chilton et al., 2014). The patterns found in the distribution of IBD containing genes involved in fatty acid

Table 2. Genes falling into regions with the maximum number of IBD blocks in more than four populations

Gene	Altaians (B)	Altaians (K)	Buryats (A)	Buryats (K)	Chulyms	Evenks (Y)	Evenks (Z)	Kalmyks	Kets	Khakas (S)	Khakas (T)	Khanty (K)	Khanty (R)	Koryaks	Nivkhs	Tatars (T)	Tuvans	Udege	Yakuts
AAGAB	+	+	+											+			+		
GSE1						+							+	+	+				+
IQCH	+	+	+											+			+		
IQCH-AS1	+	+	+											+			+		
SMAD3	+	+	+											+			+		
ANGPTL1	+						+					+	+						
RALGPS2	+						+					+	+						
ARFGAP1							+	+							+			+	
C15orf61		+	+											+			+		
C16orf74						+			+				+		+				
CALML4		+	+											+			+		
CHRNA4							+	+							+			+	
CLN6		+	+											+			+		
COL20A1							+	+							+			+	
FEM1B		+	+											+			+		
FLJ16779							+	+							+			+	
ITGA11		+	+											+			+		
LINC02206		+	+											+			+		
LOC100130587							+	+							+			+	
LOC101929076		+	+											+			+		
LOC102723493		+	+											+			+		
MAP2K5		+	+											+			+		
MIR99AHG							+	+						+	+				
NKAIN4							+	+							+			+	
PIAS1		+	+											+			+		
RALGPS2	+						+					+	+						
SKOR1		+	+											+			+		

metabolism may indicate recent directional selection associated with adaptation to dietary habits in cold climates or reflect the influence of a Western diet (Chilton et al., 2014).

Positive selection in genes that affect the level of LC-PUFAs, as well as the metabolic efficiency by which LC-PUFAs are formed in populations of the Pygmies on Flores Island (Tucci et al., 2018), Greenland Inuit (Fumagalli et al., 2015) and Native Americans (Amorim et al., 2017; Harris et al., 2019), is also thought to be associated with dietary habits in cold climates, although the exact selection pressure is unknown (Fumagalli et al., 2015). An example is the similarity of genotypes in the genes of the olfactory system, which may play a role in the formation or maintenance of social bonds between individuals within a population (Christakis, Fowler, 2014). For such genotypes, higher rates of positive selection have been found (Fu et al., 2012).

The analysis of gene ontologies (according to WebGestalt, 3511 genes turned out to be annotated according to the gene ontology database) showed nine statistically significant biological processes (taking into account the Benjamini–Hochberg correction (FDR = 0.05)) associated with the detection of a chemical stimulus involved in the sensory perception of smell (GO:0050911, GO:0007608, GO:0050907, GO:0050906, GO:0051606, GO:0009593, GO:0007600, GO:0007606, GO:0050877).

An analysis of the literature also revealed that a number of genes that fall into IBD blocks can play a significant role in the formation of oncology and influence the treatment. For

example, the *AAGAB* gene is included in the IBD blocks in five populations. *AAGAB* consists of 10 exons encoding a 315 amino acid (aa) protein, the AAGAB (alpha and gamma adaptin binding) protein. *AAGAB* is widely expressed and interacts with the gamma-adaptin and alpha-adaptin adapter protein complexes, AP1 and AP2. It is involved in membrane transport and plays a role in endocytosis and protein sorting. Heterozygous *AAGAB* mutations cause pitted palmoplantar keratoderma type 1 (PPKP1), a skin disease characterized by punctate hyperkeratosis of the palms and soles (Kiritsi et al., 2013). *AAGAB* is also a promising biomarker for chemotherapy response and outcome during breast cancer treatment. However, the exact role of *AAGAB* in the development of breast cancer is currently unclear and potentially requires further study (Bownes et al., 2019).

Another gene, *GSE1*, which falls into IBD blocks in five populations, may function as an oncogene in breast, stomach, and prostate cancer, and may also be important in the treatment of patients with prostate cancer (Bamodu et al., 2021).

IQCH-AS1 encoding antisense RNA IQCH 1 (IQCH-AS1) correlates with survival and diagnosis of cancer patients, but its role in the development of thyroid cancer and doxorubicin chemosensitivity remains unclear (Fei et al., 2022).

The role of *SMAD3* in the regulation of genes important for cell development, such as differentiation, growth and death, implies that changing its activity or suppressing its activity can lead to the formation or development of cancer.

Also, some of the genes presented in Table 1 may play a role in human adaptation to environmental factors. For example, the *IQCH* gene may play a regulatory role in spermatogenesis (Yin et al., 2005) and is also associated with adult growth in Mongols (Kimura et al., 2008).

The *CHRNA4* gene encodes for the α4β2 subcomponent of nicotinic receptors in the human brain. Individuals with certain *CHRNA4* genotypes have been shown to be better at tracking and identifying multiple objects in visual search tasks (Espeseth et al., 2010). Polymorphisms in the *CHRNA4* gene also seem to contribute to personality development by affecting the degree of developmental sensitivity to both normal and adverse environmental conditions (Grazioplene et al., 2013).

The *COL20A1* gene encoding type XX alpha 1 collagen is noted in a number of genes with non-synonymous changes with a high frequency in modern humans compared to archaic hominids, which probably contributed to the development of unique human traits and is an interesting object for study (Kuhlwilm, Boeckx, 2019).

Conclusion

Thus, as a result of our study, new information was obtained on the structure and composition of the gene pools of the indigenous peoples of Siberia, their genetic relationships and genetic and demographic processes based on an analysis of the distribution of linkage blocks identical in origin. The results obtained demonstrate the clustering of Siberian populations, linking peoples by place of origin, demonstrating a common origin and a high degree of kinship. The populations inhabiting the territory of Siberia are characterized by a unique genetic and demographic history, which is reflected in the distribution of IBD blocks both within the population and between them. The analysis of IBD blocks significantly complements the study of the formation and interaction of ethnic groups, but does not provide unambiguous answers for populations developing under conditions of complex ethnogenesis. With few exceptions, the overall IBD in Siberia is better explained by the geographical proximity of the populations rather than by their linguistic affiliation.

Analysis of SNPs that fell into more IBD segments of the analyzed populations made it possible to compile a list of 5358 genes. According to the results of calculations, biological processes enriched in these genes are associated with the detection of a chemical stimulus involved in the sensory perception of odor. Enriched with the found genes, molecular pathways are associated with fatty acid metabolism and olfactory transduction. At the same time, an analysis of the literature data showed that some of the selected genes, which were found in a larger number of IBD blocks in several populations at once, can play a role in human adaptation to environmental factors and are promising targets for further study.

References

Albrechtsen A., Moltke I., Nielsen R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics.* 2010; 186(1):295-308. DOI 10.1534/genetics.110.113977.

Amorim C.E.G., Nunes K., Meyer D., Comas D., Bortolini M.C., Salzano F.M., Hünemeier T. Genetic signature of natural selection in first Americans. *Proc. Natl. Acad. Sci. USA.* 2017;114(9):2195-2199. DOI 10.1073/pnas.1620541114.

Bamodu O.A., Wang Y.H., Ho C.H., Hu S.W., Lin C.D., Tzou K.Y., Wu W.L., Chen K.C., Wu C.C. Genetic suppressor element 1 (GSE1) promotes the oncogenic and recurrent phenotypes of castration-resistant prostate cancer by targeting tumor-associated calcium signal transducer 2 (TACSTD2). *Cancers (Basel).* 2021;13(16):3959. DOI 10.3390/cancers13163959.

Bownes R.J., Turnbull A.K., Martinez-Perez C., Cameron D.A., Sims A.H., Oikonomidou O. On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res.* 2019;21(1):73. DOI 10.1186/s13058-019-1159-3.

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.

Browning S.R., Browning B.L. High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet*. 2010;86(4): 526-539. DOI 10.1016/j.ajhg.2010.02.021.

Calder P.C. Long chain fatty acids and gene expression in inflammation and immunity. *Curr. Opin. Clin. Nutr. Metab. Care*. 2013;16(4): 425-433. DOI 10.1097/MCO.0b013e3283620616.

Chilton F.H., Murphy R.C., Wilson B.A., Sergeant S., Ainsworth H., Seeds M.C., Mathias R.A. Diet-gene interactions and PUFA metabolism: A potential contributor to health disparities and human diseases. *Nutrients.* 2014;6(5):1993-2022. DOI 10.3390/nu6051993.

Christakis N.A., Fowler J.H. Friendship and natural selection. *Proc. Natl. Acad. Sci. USA.* 2014;111(Suppl. 3):10796-10801. DOI 10.1073/pnas.1400825111.

Espeseth T., Sneve M.H., Rootwelt H., Laeng B. Nicotinic receptor gene *CHRNA4* interacts with processing load in attention. *PLoS One.* 2010;5(12):e14407. DOI 10.1371/journal.pone.0014407.

Fei Y., Li Y., Chen F. LncRNA-IQCH-AS1 sensitizes thyroid cancer cells to doxorubicin *via* modulating the miR-196a-5p/PPP2R1B signalling pathway. *J. Chemother.* 2022;1-9. DOI 10.1080/1120009X.2022.2082348.

Fu F., Nowak M.A., Christakis N.A., Fowler J.H. The evolution of homophily. *Sci. Rep*. 2012;2:845. DOI 10.1038/srep00845.

Fumagalli M., Moltke I., Grarup N., Racimo F., Bjerregaard P., Jørgensen M.E., Korneliussen T.S., Gerbault P., Skotte L., Linneberg A., Christensen C., Brandslund I., Jørgensen T., Huerta-Sánchez E., Schmidt E.B., Pedersen O., Hansen T., Albrechtsen A., Nielsen R. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*. 2015;349(6254):1343-1347. DOI 10.1126/science.aab2319.

Grazioplene R.G., Deyoung C.G., Rogosch F.A., Cicchetti D. A novel differential susceptibility gene: *CHRNA4* and moderation of the effect of maltreatment on child personality. *J. Child Psychol. Psychiatry*. 2013;54(8):872-880. DOI 10.1111/jcpp.12031.

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.

Han L., Abney M. Identity by descent estimation with dense genome-wide genotype data. *Genet. Epidemiol*. 2011;35(6):557-567. DOI 10.1002/gepi.20606.

Harris D.N., Ruczinski I., Yanek L.R., Becker L.C., Becker D.M., Guio H., Cui T., Chilton F.H., Mathias R.A., O'Connor T.D. Evolution of hominin polyunsaturated fatty acid metabolism: from Africa to the New World. *Genome Biol. Evol.* 2019;11(5):1417-1430. DOI 10.1093/gbe/evz071.

Hochreiter S. HapFABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Res.* 2013;41(22):e202. DOI 10.1093/nar/gkt1013.

Horrocks L.A., Yeo Y.K. Health benefits of docosahexaenoic acid (DHA). *Pharmacol. Res.* 1999;40(3):211-225. DOI 10.1006/phrs.1999.0495.

Howe P., Meyer B., Record S., Baghurst K. Dietary intake of long-chain ω-3 polyunsaturated fatty acids: contribution of meat sources. *Nutrition.* 2006;22(1):47-53. DOI 10.1016/j.nut.2005.05.009.

Kimura T., Kobayashi T., Munkhbat B., Oyungerel G., Bilegtsaikhan T., Anar D., Jambaldorj J., Munkhsaikhan S., Munkhtuvshin N., Hayashi H., Oka A., Inoue I., Inoko H. Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33-q23 in Mongolians. *Hum. Genet*. 2008;123(6):655-660. DOI 10.1007/s00439-008-0512-x.

Kiritsi D., Chmel N., Arnold A.W., Jakob T., Bruckner-Tuderman L., Has C. Novel and recurrent *AAGAB* mutations: clinical variability and molecular consequences. *J. Invest. Dermatol.* 2013;133(10):2483-2486. DOI 10.1038/jid.2013.171.

Kolesnikov N.A., Kharkov V.N., Zarubin A.A., Radzhabov M.O., Voevoda M.I., Gubina M.A., Khusnutdinova E.K., Litvinov S.S., Ekomasova N.V., Shtygasheva O.V., Maksimova N.R., Sukhomyasova A.L., Stepanov V.A. Features of the genomic distribution of runs of homozygosity in the indigenous population of Northern Eurasia at the individual and population levels based on high density SNP analysis. *Russ. J. Genet.* 2021;57(11):1271-1284. DOI 10.1134/S1022795421110053.

Kuhlwilm M., Boeckx C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* 2019;9(1):8463. DOI 10.1038/s41598-019-44877-x.

Marszalek J.R., Lodish H.F. Docosahexaenoic acid, fatty acid-interacting proteins, and neuronal function: breastmilk and fish are good for you. *Annu. Rev. Cell Dev. Biol.* 2005;21:633-657. DOI 10.1146/annurev.cellbio.21.122303.120624.

Palamara P.F., Lencz T., Darvasi A., Pe'er I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet*. 2012;91(5):809-822. DOI 10.1016/j.ajhg.2012.08.030.

Pugach I., Matveev R., Spitsyn V., Makarov S., Novgorodov I., Osakovsky V., Stoneking M., Pakendorf B. The complex admixture history and recent southern origins of Siberian populations. *Mol. Biol. Evol.* 2016;33(7):1777-1795. DOI 10.1093/molbev/msw055.

Tucci S., Vohr S.H., McCoy R.C., Vernot B., Robinson M.R., Barbieri C., Nelson B.J., Fu W., Purnomo G.A., Sudoyo H., Eichler E.E., Barbujani G., Visscher P.M., Akey J.M., Green R.E. Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. *Science.* 2018;361(6401):511-516. DOI 10.1126/science.aar8486.

Yin L.L., Li J.M., Zhou Z.M., Sha J.H. Identification of a novel testis-specific gene and its potential roles in testis development/spermatogenesis. *Asian J. Androl*. 2005;7(2):127-137. DOI 10.1111/j.1745-7262.2005.00041.x.

ORCID ID

N.A. Kolesnikov orcid.org/0000-0001-8855-577X
V.N. Kharkov orcid.org/0000-0002-1679-2212
K.V. Vagaitseva orcid.org/0000-0003-4877-9749
A.A. Zarubin orcid.org/0000-0001-6568-6339
V.A. Stepanov orcid.org/0000-0002-5166-331X

Acknowledgements. The study was supported by the Russian Science Foundation, grant No. 22-64-00060, (https://rscf.ru/project/22-64-00060/).

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

Received October 21, 2022. Revised January 17, 2023. Accepted January 24, 2023.

Вавиловский журнал генетики и селекции. 2023;27(1):63-71
DOI 10.18699/VJGB-23-09

Original Russian text https://vavilovj-icg.ru/

Expression of the *NUP153* and *YWHAB* genes from their canonical promoters and alternative promoters of the LINE-1 retrotransposon in the placenta of the first trimester of pregnancy

V.V. Demeneva[1] ✉, E.N. Tolmacheva[1], T.V. Nikitina[1], E.A. Sazhenova[1], S.Yu. Yuriev[2], A.Sh. Makhmutkhodzhaev[2], A.S. Zuev[1], S.A. Filatova[3], A.E. Dmitriev[3], Ya.A. Darkova[3], L.P. Nazarenko[1], I.N. Lebedev[1, 2], S.A. Vasilyev[1, 3]

[1] Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
[2] Siberian State Medical University, Tomsk, Russia
[3] National Research Tomsk State University, Tomsk, Russia
✉ deviva2503@gmail.com

Abstract. The placenta has a unique hypomethylated genome. Due to this feature of the placenta, there is a potential possibility of using regulatory elements derived from retroviruses and retrotransposons, which are suppressed by DNA methylation in the adult body. In addition, there is an abnormal increase in the level of methylation of the LINE-1 retrotransposon in the chorionic trophoblast in spontaneous abortions with both normal karyotype and aneuploidy on different chromosomes, which may be associated with impaired gene transcription using LINE-1 regulatory elements. To date, 988 genes that can be expressed from alternative LINE-1 promoters have been identified. Using the STRING tool, genes (*NUP153* and *YWHAB*) were selected, the products of which have significant functional relationships with proteins highly expressed in the placenta and involved in trophoblast differentiation. This study aimed to analyze the expression of the *NUP153* and *YWHAB* genes, highly active in the placenta, from canonical and alternative LINE-1 promoters in the germinal part of the placenta of spontaneous and induced abortions. Gene expression analysis was performed using real-time PCR in chorionic villi and extraembryonic mesoderm of induced abortions ($n = 10$), adult lymphocytes ($n = 10$), spontaneous abortions with normal karyotype ($n = 10$), and with the most frequent aneuploidies in the first trimester of pregnancy (trisomy 16 ($n = 8$) and monosomy X ($n = 6$)). The LINE-1 methylation index was assessed in the chorionic villi of spontaneous abortions using targeted bisulfite massive parallel sequencing. The level of expression of both genes from canonical promoters was higher in blood lymphocytes than in placental tissues ($p < 0.05$). However, the expression level of the *NUP153* gene from the alternative LINE-1 promoter was 17 times higher in chorionic villi and 23 times higher in extraembryonic mesoderm than in lymphocytes ($p < 0.05$). The expression level of *NUP153* and *YWHAB* from canonical promoters was higher in the group of spontaneous abortions with monosomy X compared to all other groups ($p < 0.05$). The LINE-1 methylation index negatively correlated with the level of gene expression from both canonical (*NUP153* – $R = –0.59$, *YWHAB* – $R = –0.52$, $p < 0.05$) and alternative LINE-1 promoters (*NUP153* – $R = –0.46$, *YWHAB* – $R = –0.66$, $p < 0.05$). Thus, the observed increase in the LINE-1 methylation index in the placenta of spontaneous abortions is associated with the level of expression of the *NUP153* and *YWHAB* genes not only from alternative but also from canonical promoters, which can subsequently lead to negative consequences for normal embryogenesis.

Key words: miscarriage; placenta; retrotransposon LINE-1; DNA methylation; *NUP153*; *YWHAB*.

For citation: Demeneva V.V., Tolmacheva E.N., Nikitina T.V., Sazhenova E.A., Yuriev S.Yu., Makhmutkhodzhaev A.Sh., Zuev A.S., Filatova S.A., Dmitriev A.E., Darkova Ya.A., Nazarenko L.P., Lebedev I.N., Vasilyev S.A. Expression of the *NUP153* and *YWHAB* genes from their canonical promoters and alternative promoters of the LINE-1 retrotransposon in the placenta of the first trimester of pregnancy. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):63-71. DOI 10.18699/VJGB-23-09

Экспрессия генов *NUP153* и *YWHAB* с их канонических промоторов и альтернативных промоторов ретротранспозона LINE-1 в плаценте первого триместра беременности

В.В. Деменева[1] ✉, Е.Н. Толмачева[1], Т.В. Никитина[1], Е.А. Саженова[1], С.Ю. Юрьев[2], А.Ш. Махмутходжаев[2], А.С. Зуев[1], С.А. Филатова[3], А.Е. Дмитриев[3], Я.А. Даркова[3], Л.П. Назаренко[1], И.Н. Лебедев[1, 2], С.А. Васильев[1, 3]

[1] Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
[2] Сибирский государственный медицинский университет Министерства здравоохранения Российской Федерации, Томск, Россия
[3] Национальный исследовательский Томский государственный университет, Томск, Россия
✉ deviva2503@gmail.com

Аннотация. Для плаценты характерен уникальный гипометилированный геном. Благодаря этой особенности плаценты в первом триместре беременности наблюдается потенциальная возможность использования регуляторных элементов, полученных от ретровирусов и ретротранспозонов, которые во взрослом организме подавляются метилированием ДНК. Кроме того, у спонтанных абортусов и с нормальным кариотипом, и с анеуплоидиями отмечается аномальное повышение уровня метилирования ретротранспозона LINE-1 в трофобласте хориона, что может быть связано с нарушением транскрипции генов с использованием регуляторных элементов LINE-1. На сегодняшний день идентифицировано 988 генов, способных экспрессироваться с альтернативных промоторов LINE-1. Из них с помощью инструмента STRING были отобраны гены, продукты которых взаимодействуют с белками, экспрессирующимися на высоком уровне в плаценте и участвующими в дифференцировке трофобласта, *NUP153* и *YWHAB*. Целью настоящего исследования являлся анализ экспрессии генов *NUP153* и *YWHAB* с канонических и альтернативных промоторов ретротранспозона LINE-1 в зародышевой части плаценты спонтанных и медицинских абортусов первого триместра беременности. Определение уровня экспрессии генов проводили с помощью ПЦР в режиме реального времени в лимфоцитах взрослых индивидов ($n = 10$), в ворсинах хориона и экстраэмбриональной мезодерме медицинских абортусов ($n = 10$) и спонтанных абортусов с нормальным кариотипом ($n = 10$) и с наиболее частыми анеуплоидиями в I триместре беременности (трисомия 16 ($n = 8$) и моносомия X ($n = 6$)). Индекс метилирования LINE-1 оценивали в ворсинах хориона спонтанных абортусов с помощью таргетного бисульфитного массового параллельного секвенирования. Уровень экспрессии обоих генов с канонических промоторов был выше в лимфоцитах крови, чем в тканях плаценты ($p < 0.05$). Однако уровень экспрессии гена *NUP153* с альтернативного промотора LINE-1 был выше в 17 раз в ворсинах хориона и в 23 раза – в экстраэмбриональной мезодерме по сравнению с лимфоцитами ($p < 0.05$). Между группами спонтанных абортусов с моносомией X и остальными группами были выявлены статистически значимые различия. Уровень экспрессии *NUP153* и *YWHAB* с канонических промоторов был выше в группе спонтанных абортусов с моносомией X по сравнению со всеми другими группами ($p < 0.05$). Индекс метилирования LINE-1 отрицательно коррелировал с уровнем экспрессии генов как с канонических (*NUP153* – $R = –0.59$, *YWHAB* – $R = –0.52$, $p < 0.05$), так и с альтернативных промоторов LINE-1 (*NUP153* – $R = –0.46$, *YWHAB* – $R = –0.66$, $p < 0.05$). Таким образом, наблюдаемое нами повышение индекса метилирования LINE-1 в плаценте спонтанных абортусов связано с уровнем экспрессии генов *NUP153* и *YWHAB* не только с альтернативных, но и с канонических промоторов, что в дальнейшем может приводить к негативным последствиям для нормального эмбриогенеза.
Ключевые слова: невынашивание беременности; плацента; ретротранспозон LINE-1; метилирование ДНК; *NUP153*; *YWHAB*.

Introduction

In humans, reproductive losses are more common in the first trimester of pregnancy than in other periods of embryogenesis. One of the most common causes of early embryonic death is an abnormal number of chromosomes (aneuploidy), which leads to severe developmental anomalies. The formation of aneuploidy with meiotic and mitotic origin corresponds to the waves of epigenetic reprogramming, in particular, genome demethylation in the zygote and at the cleavage stage. Early blastocyst demonstrates less DNA methylation at the latter stage than cells at any other moment of ontogeny (Smith et al., 2012). A rapid wave of *de novo* DNA methylation for the inner cell mass then follows while the trophectoderm remains hypomethylated (Santos et al., 2010).

Throughout pregnancy, the placenta has a unique hypomethylated epigenetic landscape compared to other extraembryonic and embryonic tissues, which may indicate its special functions (Robinson, Price, 2015). Hypomethylation in placental DNA occurs mainly in "partially methylated domains" and is unevenly distributed throughout the genome. "Partially methylated domains" refers to large (> 100 kb) regions of low DNA methylation alternating with regions of higher DNA methylation (Schroeder et al., 2013).

The placenta exhibits reduced DNA methylation of some types of repetitive genome elements (Price et al., 2012). One of them, the LINE-1 retrotransposon (long interspersed nuclear element-1), is the largest, occupying approximately 20 % of the genome, and the most evolutionarily young class of retrotransposons in humans, retaining the ability to transpose (Ostertag et al., 2001). The transcriptional activity of LINE-1 is suppressed by DNA methylation during most periods of ontogeny.

An important feature of LINE-1 that requires attention is its high level of methylation in blood leukocytes, regardless of age and gender, while the level of LINE-1 methylation in other tissues has its tissue-specific differences (Chalitchagorn et al., 2004). It was shown that for the placenta as an independent organ, the level of methylation of retrotransposons doesn't always coincide with the global level of methylation of the entire genome. The level of LINE-1 methylation in the tissues of the placenta of the third trimester of pregnancy significantly decreases compared to the first trimester of pregnancy. At the same time, changes in the DNA methylation level of the entire genome are not found between the first and third trimester placentas (He et al., 2014).

It can be assumed that LINE-1 methylation and activation are transiently regulated during normal placental development. This raises the question of a possible functional role for LINE-1 retrotransposon sequences in placental development. Previously, we found that some spontaneous abortions with normal karyotypes were characterized by epigenetic disorders similar to spontaneous abortions with aneuploidy. In particular, some spontaneous abortions with a normal karyotype had an increased methylation index in the LINE-1 retrotransposon promoter, which was characteristic of groups of spontaneous abortions with trisomy 16 and monosomy X (Vasilyev et al., 2021b).

One of the possible roles of LINE-1 may be the usage of its regulatory sequences to influence the transcription of adjacent genes. This effect becomes feasible because LINE-1 includes

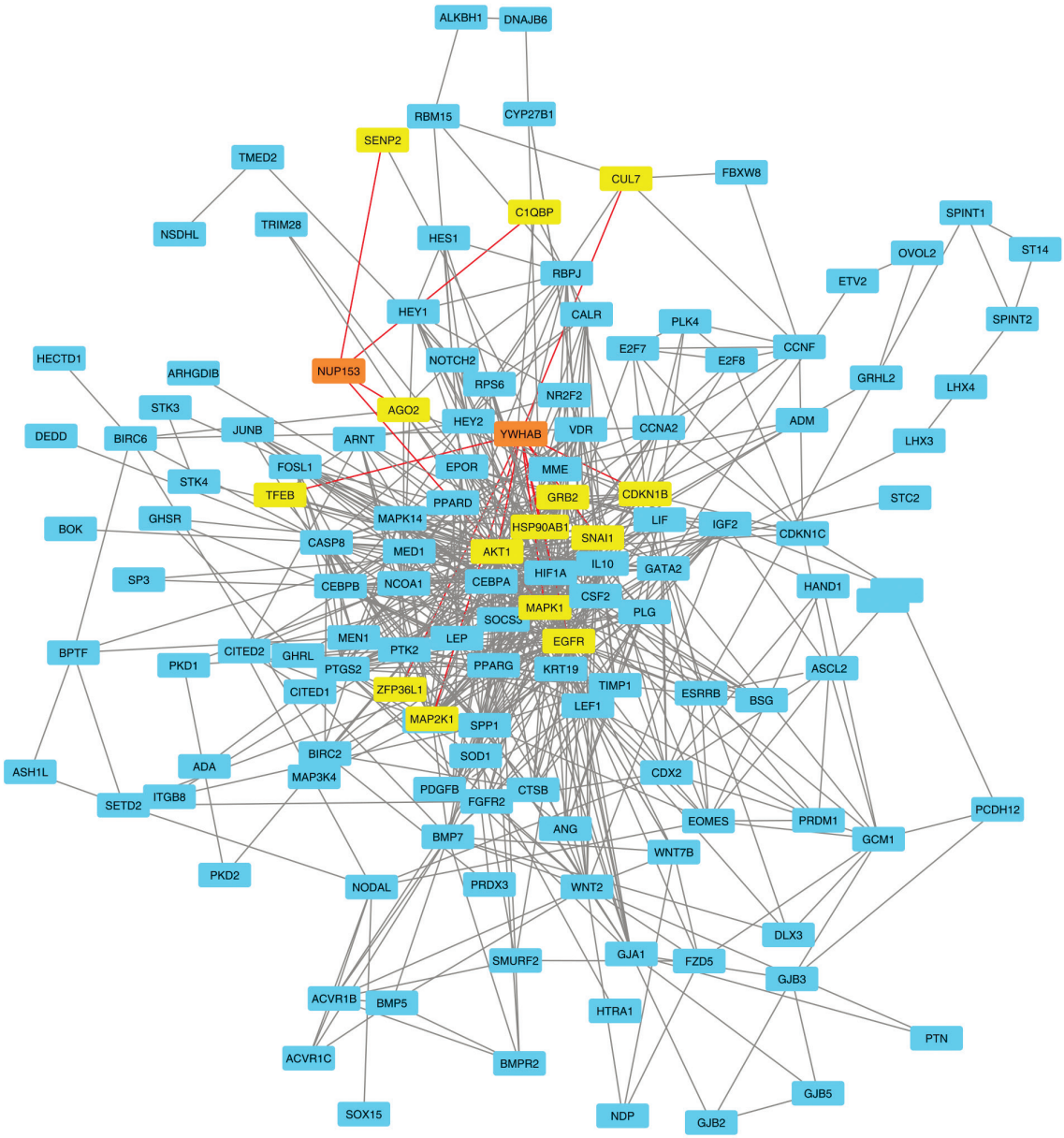

Fig. 1. Functionally significant connections of the proteins involved in the development of the placenta (GO:0061450, GO:0097360, GO:0001890) with the *NUP153* and *YWHAB* genes according to STRING.

Yellow shows proteins that have functional bonds (highlighted in red) with the *NUP153* and *YWHAB* genes (marked in orange) (STRING score > 0.4).

a sense promoter that controls the transcription of the ORF1 and ORF2p proteins required for retrotransposition, and an antisense promoter that controls the transcription of chimeric transcripts, LINE-1 5′-antisense sequences spliced with exons of neighboring genes (Denli et al., 2015). LINE-1 antisense transcripts affect up to 4 % of all human genes, and LINE-1 antisense promoters are actively transcribed in various types of human cells, including embryonic tissues. A total of 988 genes that can be expressed from alternative LINE-1 promoters have been identified so far (Criscione et al., 2016b). It is possible that the expression of multiple genes in extraembryonic tissues may occur predominantly from alternative LINE-1 promoters because LINE-1 promoters are hypomethylated in the placenta.

Using the STRING tool, two genes, *NUP153* and *YWHAB*, were selected among the genes capable of expression from alternative LINE-1 promoters. Their products showed a high level of expression in the placenta and are functionally associated with proteins involved in trophoblast differentiation (according to Gene Ontology: GO:0061450, trophoblast cell migration; GO:0097360, chorionic trophoblast cell proliferation; GO:0001890, placenta development) (Fig. 1). The *NUP153* gene functions as a scaffolding element in the nuclear phase of the nuclear pore complex. It is required for normal nuclear-cytoplasmic transport of proteins and mRNA during somatic cell division (Bilir et al., 2019) and in mouse embryonic stem cells (Souquet et al., 2018). The *YWHAB* gene belongs to the group of genes responsible for signal transduction by binding

to phosphoserine-containing proteins. The protein encoded by the gene interacts with RAF1 and CDC25 phosphatases and may play a role in mitogenic signaling and cell cycle regulatory mechanisms. It was shown that *YWHAB* overexpression stimulates and maintains attachment-independent cell growth in a fibroblast cell line isolated from mouse embryos (Sasaki et al., 2014).

The aim of this study was to analyze the expression of the *NUP153* and *YWHAB* genes from canonical and alternative LINE-1 promoters in the germinal part of the placenta of spontaneous and induced abortions.

Materials and methods

The samples were from chorionic villi and extraembryonic mesoderm of induced abortions (IA) (n = 10, gestational age 8.2 ± 2.3 weeks), spontaneous abortions (SA) with normal karyotype (n = 10, gestational age 7.2 ± 1.4 weeks), trisomy 16 (n = 8, gestational age 6.5 ± 0.8 weeks), and monosomy X (n = 6, gestational age 8.6 ± 0.7 weeks). Samples were taken from the Biobank of Northern Eurasia of the Research Institute of Medical Genetics of the Tomsk National Research Medical Center. The samples were obtained from 2004 to 2021 and stored in liquid nitrogen, before their use for analysis. The study was conducted in compliance with ethical standards by the Helsinki Declaration of the World Medical Association. The study was approved by the Biomedical Ethics Committee of the Research Institute of Medical Genetics of the Tomsk National Research Medical Center (November 9, 2020/No. 7).

A standard cytogenetic analysis was performed on direct preparations of chorionic villi and fibroblast cultures of the extraembryonic mesoderm to determine the karyotype (Lebedev et al., 2004). Karyotyping results for 14 trisomic and monosomic SA samples were confirmed by fluorescence *in situ* hybridization (FISH). Aneuploidy mosaicism was assessed with a lower cutoff of 10 % and an upper cutoff of 90 %.

Centromere-specific DNA probes for X chromosomes were used for the analysis of monosomy X and subtelomeric DNA probes (16q and 16p) were used for the analysis of trisomy 16. The analysis was carried out according to the protocol described elsewhere (Vasilyev et al., 2010). Four samples had a mosaic karyotype with a trisomy level from 10 to 90 %. The remaining 10 spontaneous abortions with a higher proportion of trisomy or monosomy were classified as having pure aneuploidy. The blood lymphocytes of IA parents (5 couples, age 30.8 ± 2.7) were used as a comparison group that were contained in the Lyra reagent (Biolabmix, Russia) before the start of the experiment.

RNA was isolated from chorionic villi and extraembryonic mesoderm by the phenol-chloroform method. All tissues were stored in liquid nitrogen from the moment of obtaining the material of the studied samples to the beginning of RNA isolation. Tissue separation was preliminarily carried out in RNAlater (Invitrogen, USA) to stabilize the RNA in the samples. Each sample was homogenized in a mortar with liquid nitrogen, adding 500 µl of Lyra reagent (Biolabmix, Russia). The lysate was incubated first for 5 min at 55 °C, then for 5 min at room temperature. The lysate was then centrifuged at 12,000 rpm for 10 min to remove undissolved fragments, and the supernatant was transferred into a new tube. A volume of 0.2 ml of chloroform was added per each 1 ml of Lyra reagent, followed by shaking (by hand) for 15 s, followed by incubation of the mixture for 10 min at room temperature, and centrifugation at 10,000 g for 10 min at 4 °C. Next, 0.5 ml of 100 % cold isopropanol was added to the aqueous phase containing RNA per each 1 ml of Lyra reagent, and the mixture was incubated at –20 °C for 10 min, after which the sample was centrifuged at 12,000 g for 10 min at 4 °C.

The precipitate was washed twice with 80 % cold ethanol at 10,000 g for 5 min at 4 °C. The precipitate was then dried for 2 min in a concentrator (Eppendorf, USA) (parameters: 45 °C, V-AL). After this, 40 µl of DEPC water and 1 µl of RiboLock (Thermo, USA) were added to dissolve the precipitate and left for 10 min at room temperature until complete dissolution. All samples were kept on ice to avoid RNA degradation during isolation whereas at the incubation stage all steps were performed at room temperature. All samples were stored at –80 °C after isolation.

The RNA was treated with DNase (Biolabmix) to obtain pure RNA. Further, the OT-M-MuLV-RH kit (Biolabmix) with a random hexaprimer was used for reverse transcription. The reverse transcription reaction mixture included 1.5 µg RNA, 3 µl hexaprimer, 4 µl KCl reaction buffer, 2 µl 0.1 M DDT, 1 µl 10 mM dNTP mix, and 1 µl revertase. Two types of primers were designed for the *NUP153* and *YWHAB* genes: the first for long products that are expressed only from canonical promoters, and the second for short products that are expressed from alternative LINE-1 antisense promoters (see the Table).

The *NUP153* gene includes 22 exons, while the short transcript from the alternative LINE-1 promoter contains only exons 21–22. Primers were designed in exons 16–17 for detecting *NUP153* gene transcripts from the canonical promoter and in exons 21–22 for detecting transcripts from the alternative promoter. Two normal long transcripts with exons 7 or 6 are transcribed from the normal promoter of the *YWHAB* gene. In this regard, primers were designed for each product. For the first transcript with seven exons, primers were designed in exons 1–2. For the second transcript with 6 exons, primers were designed in exons 1–3. Primers of the short product from the alternative LINE-1 promoter of the *YWHAB* gene were designed in exons 4–7 (Fig. 2). The expression from alternative gene promoters was taken to be the difference between the level of gene expression estimated using primers specific to the region downstream the alternative promoter and the level of gene expression estimated using primers annealing upstream in the first exons. This value was used for data analysis and is displayed on the charts. For the *YWHAB* gene, the sum of expression levels of both long transcripts was subtracted from the expression level of the canonical promoter.

The methylation index was assessed in 19 CpG sites of the LINE-1 promoter in chorionic villi of spontaneous abortions using targeted bisulfite massive parallel sequencing. Library preparation and evaluation were carried out according to a previously published protocol (Vasilyev et al., 2021a). Statistical analysis of data was performed using Statistica 10.0 software.

Sequences of oligonucleotide primers for assessing
the level of expression of the *NUP153*, *YWHAB*, and *GAPDH* genes using real-time PCR

Gene	Transcript	Nucleotide sequence
NUP153	Transcript from the canonical promoter (NM_001278209.2, 22 exons)	F 5′-TGTATGTCTGAGAAACCAGGAAGTT-3′
		R 5′-GTAGAGTCTGCCTTATTCTGCACTA-3′
	Shortened transcript from the alternative promoter LINE-1 (2 exons)	F 5′-CAGCATTTACAGTGGGGTCAAAT-3′
		R 5′-CAACACCAATGTGACCTTTATTTCC-3′
YWHAB	Transcript from the canonical promoter (NM_003404, 7 exons)	F 5′-GCTCGGAAGGGTCTTTGTTC-3′
		R 5′-TCTATCCACAGCCGAATGGG-3′
	Transcript from the canonical promoter (NM_139323, 6 exons)	F 5′-GAGTAGTGGGCTTAGGAAGGAAGAG-3′
		R 5′-CTTTTATCCATTGTCATTCCCGTGG-3′
	Shortened transcript from the alternative promoter LINE-1 (4 exons)	F 5′-CTGTAGCCTGGCAAAAACGG-3′
		R 5′-TCCGATGTCCACAGAGTGAGA-3′
GAPDH	Transcript from the canonical promoter (NM_002046.7, 10 exons)	F 5′-GCCAGCCGAGCCACATC-3′
		R 5′-GGCAACAATATCCACTTTACCAGA-3′

Note. F– forward primer; R – reverse primer.

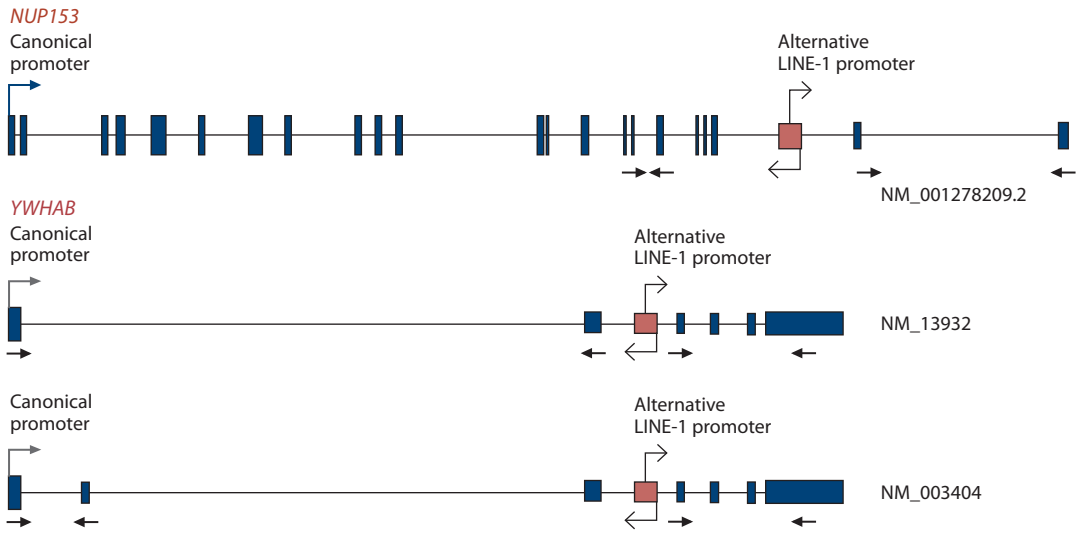

Fig. 2. Scheme of the location of alternative LINE-1 promoters for the *NUP153* and *YWHAB* genes.

The arrows schematically mark the hybridization sites of oligonucleotide primers. The arrow starting at the beginning of the LINE-1 element indicates the direction of transcription from the direct LINE-1 promoter, which is canonical. The second arrow pointing in the opposite direction marks the direction of expression from the alternative antisense LINE-1 promoter, which is also alternative for the studied genes.

Results

The expression level of the *NUP153* gene from the canonical promoter was 12.5 times higher in lymphocytes than in placental tissues ($p = 0.000001$). The expression level of the *YWHAB* gene from the canonical promoter was also on average higher in blood lymphocytes than in placental tissues (by 4.6 times) (transcript NM_13932 ($p = 0.00003$)). The expression level of the NM_003404 transcript of the *YWHAB* gene was highly variable in lymphocytes. However, the expression level of the *NUP153* gene from alternative LINE-1 promoters was statistically significantly higher in extraembryonic tissues compared to lymphocytes of adults (17 times in chorionic villi and 23 times in extraembryonic mesoderm, $p < 0.05$) (Fig. 3). The levels of expression of both genes from canonical promoters were higher in the SA group with monosomy X than in the groups of SA with normal karyotype (Fig. 4).

The level of methylation of the LINE-1 retrotransposon promoter was assessed in the chorionic villi of spontane-

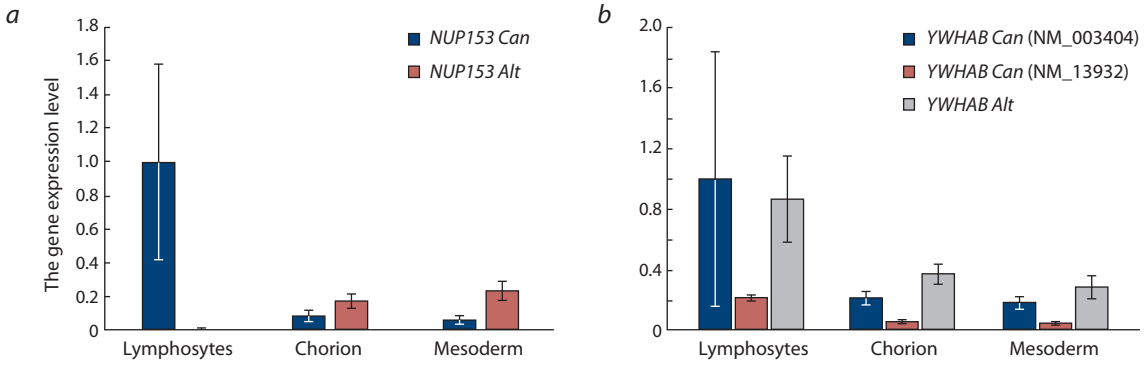

Fig. 3. Comparison of the *NUP153* (*a*) and *YWHAB* (*b*) gene expression levels from canonical promoters and alternative LINE-1 promoters in blood lymphocytes, chorion, and placental mesoderm.

Values are given as fold differences relative to the level of gene expression from the canonical promoter in adult lymphocytes. Expression levels of two different transcripts from the canonical promoter (NM_003404, NM_139323) are shown for the *YWHAB* gene. The reference gene is *GAPDH*. *Can* is the canonical promoter, and *Alt* is the alternative LINE-1 promoter.

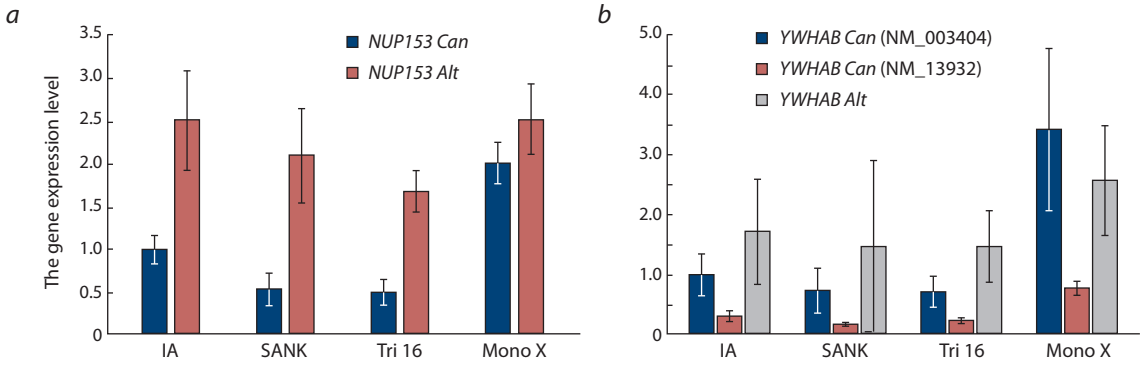

Fig. 4. Comparison of the expression level of the *NUP153* (*a*) and *YWHAB* (*b*) genes from canonical promoters and alternative LINE-1 promoters between groups of spontaneous abortions and induced abortions.

Values are given as fold differences relative to the level of gene expression of the canonical promoter in the group of induced abortions. SANK – spontaneous abortions with normal karyotype; Tri 16 – spontaneous abortions with trisomy 16; Mono X – spontaneous abortions with monosomy X.

ous abortions with different karyotypes. The average level of LINE-1 methylation in chorionic villi of SA was 41.9± ±5.8 % with trisomy 16, 39.7±3.6 % with monosomy X, and 38.4±3.9 % with normal karyotype. The LINE-1 methylation index negatively correlated with the level of gene expression from both canonical (*NUP153* – $R = –0.59$, $p < 0.003$; *YWHAB* – $R = –0.52$, $p < 0.01$) and alternative LINE-1 promoters (*NUP153* – $R = –0.46$, $p = 0.03$; *YWHAB* – $R = –0.66$, $p = 0.001$) (Fig. 5).

Discussion

In the present work, it was found that the level of expression of the *NUP153* and *YWHAB* genes in the placenta from canonical promoters was lower compared to the adult blood lymphocytes, but the expression of the *NUP153* gene from the alternative LINE-1 promoter was higher in the placenta. This result has supported the hypothesis that in the placenta, the expression of genes from alternative promoters derived from retroviruses and retrotransposons can be activated due to the hypomethylated epigenetic landscape. This assumption is also supported by the enrichment of genes that are tissue-specifically expressed in the placenta among all genes which can be transcribed from alternative LINE-1 promoters (Criscione et al., 2016a).

We have not found significant differences in the level of expression of the *YWHAB* and *NUP153* genes from alternative promoters between groups of spontaneous abortions with different karyotypes and the control group of induced abortions. At the same time, the levels of expression of both genes from canonical promoters were higher in the group of spontaneous abortions with monosomy X. However, it has been found that the level of expression of the studied genes changes in individual spontaneous abortions depending on changes in the level of LINE-1 methylation. The obtained data clearly demonstrated that the expression level of the *NUP153* and *YWHAB* genes from the canonical and alternative LINE-1 promoters correlates with the LINE-1 methylation level: the higher the LINE-1 methylation level, the lower the expression.

There can be several reasons for the relationship between the level of LINE-1 methylation and the expression of the studied genes from both promoters. First, a short transcript from an alternative promoter may be associated with the

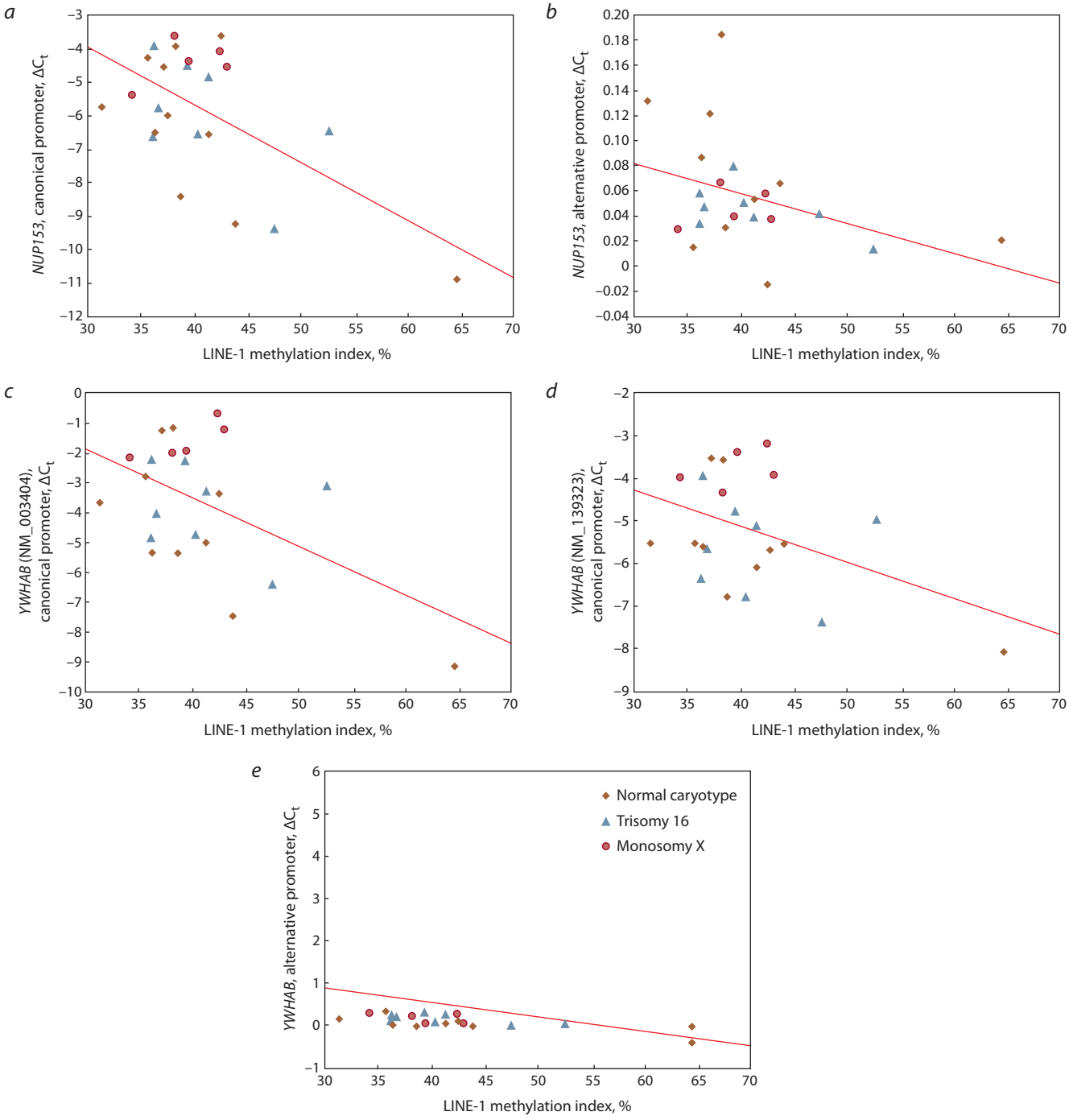

Fig. 5. Correlation of the *NUP153* and *YWHAB* gene expression with the LINE-1 methylation index in the chorionic trophoblast of spontaneous abortions with normal karyotype, trisomy 16, and monosomy X.

Correlations of the LINE-1 methylation index with various transcripts of the *NUP153* and *YWHAB* genes: *a* – *NUP153* gene expression from the canonical promoter; *b* – *NUP153* gene expression from the alternative promoter; *c* – *YWHAB* (NM_003404) gene expression from the canonical promoter; *d* – *YWHAB* (NM_139323) gene expression from the canonical promoter; *e* – *YWHAB* gene expression from the alternative promoter. SANK – spontaneous abortions with normal karyotype; Tri 16 – spontaneous abortions with trisomy 16; Mono X – spontaneous abortions with monosomy X.

activation of gene transcription from the canonical promoter. However, this option seems unlikely, because the expression level of the studied genes from canonical promoters against the background of genome hypomethylation in the placenta was lower than in lymphocytes, which are characterized by a LINE-1 methylation level of more than 70 % (Rosser, An, 2012). This should be the opposite if this hypothesis is correct.

Second, the level of methylation of the LINE-1 retrotransposon may reflect the global level of genome methylation and the level of methylation in the canonical promoter of the studied genes. This variant seems to be more likely, but also doesn't remove the issue of reduced expression of the studied genes in the placenta against the background of a hypomethylated epigenetic landscape compared to adult lymphocytes.

The expression of the studied genes is regulated not only by methylation but also by tissue-specific transcription factors.

It remains unclear whether the *NUP153* and *YWHAB* gene expression both from the canonical and alternative promoter plays a functional role in the placenta or whether these transcripts are by-products of the genome hypomethylation. Potentially, the impaired *NUP153* expression can have a negative impact on the nuclear-cytoplasmic transport of proteins and mRNA, and the abnormal *YWHAB* gene expression can affect the transmission of cell signals.

NUP153 and *YWHAB* gene products have significant functional connections with proteins involved in the differentiation of the trophoblast (see Fig. 1). *NUP153* interacts with the *AGO2*, *SENP2*, *C1QBP*, and *PPARD* genes. A list of significant connections is wider for the *YWHAB* gene – it interacts with the *TFEB*, *CUL7*, *ZFP36L*, *MAP2K1*, *AKT1*, *CDKN1B*, *SNAI1*, *MAPK1*, and *EGFR* genes.

The impaired function of each of these genes has a negative effect on the normal course of embryogenesis. For example, the normal expression of *MAPK1* is necessary for the development of non-embryonic ectoderm during placentogenesis. Its absence can lead to embryo death due to abnormal development and hypovascularization of the placenta (Bissonauth et al., 2006). The *CUL7* gene is actively expressed in the cell lines of the trophoblast. Protein deficiency of the *CUL7* gene is associated with a delay in intrauterine development due to abnormal development of the placenta, which leads to intrauterine hypoxia (Fahlbusch et al., 2012). The deficit can lead to the occurrence of cutaneous or hypodermal hemorrhages, as well as the development of trophoblast with abnormal vascular structure at later stages of gestation (Arai et al., 2003). *CUL7* mutations in the embryo line are associated with the 3-M syndrome, which is characterized by pre- and postnatal growth retardation (Maksimova et al., 2007; Fu et al., 2010).

The *SENP2* gene belongs to the family of ubiquitin-like proteins and is localized in the cell in the nuclear pores and cytoplasm (Talamillo et al., 2020). *SENP2* mutations impair cell cycle progression during trophoblast development in mice: deletion of *SENP2* impairs the p53/Mdm2 pathway, affecting trophoblast progenitor cells and their maturation (Chiu et al., 2008). *SENP2* influences the normal development of cardiomyocytes during further differentiation. Overexpression causes abnormal proliferation of cardiomyocytes with dysregulation of cyclin and cyclin-dependent kinase inhibitors, leading to congenital heart anomalies (Kim et al., 2012). On the other hand, deletions also cause defects in myocardial development due to reduced proliferation (Kang et al., 2010).

It is logical to assume that the existing functional relationships of the *NUP153* and *YWHAB* genes with genes involved in trophoblast differentiation can go both in a negative direction and in a protective one. Pathological changes in the expression of the *NUP153* and *YWHAB* genes can potentially lead to impaired function of other genes, the formation of a pathological embryo phenotype, or even embryonic death.

Conclusion

We have revealed that the *NUP153* and *YWHAB* genes in the placenta tissues are predominantly expressed from alternative LINE-1 promoters located in the introns. Even though the expression from alternative promoters of LINE-1 was higher than with canonical gene promoters for all groups (spontaneous and induced abortions), and there were no significant differences in the level of expression of the *YWHAB* and *NUP153* genes from alternative promotors between groups, we have seen a trend towards the general decrease in expression in spontaneous abortions compared to induced abortions. However, it has been found that the level of expression of the studied genes changes in individual spontaneous abortions, depending on changes in the level of genome methylation. The obtained data demonstrate the relationship between the levels of the *NUP153* and *YWHAB* gene expression from canonical and alternative LINE-1 promoters with LINE-1 methylation levels in extraembryonic tissues of spontaneous abortions.

Thus, an increase in the LINE-1 methylation index in the placenta of spontaneous abortions may be associated with a decrease in gene expression not only from alternative but also from canonical promoters. The revealed features of the relationship between the LINE-1 methylation level with the *NUP153* and *YWHAB* gene expression levels indicate an existing mechanism for self-regulation of normal embryogenesis, disturbance of which can lead to embryo death.

References

Arai T., Kasper J.S., Skaar J.R., Ali S.H., Takahashi C., DeCaprio J.A. Targeted disruption of *p185/Cul7* gene results in abnormal vascular morphogenesis. *Proc. Natl. Acad. Sci. USA*. 2003;100(17):9855-9860. DOI 10.1073/pnas.1733908100.

Bilir Ş., Kojidani T., Mori C., Osakada H., Kobayashi S., Koujin T., Hiraoka Y., Haraguchi T. Roles of Nup133, Nup153 and membrane fenestrations in assembly of the nuclear pore complex at the end of mitosis. *Genes Cells*. 2019;24(5):338-353. DOI 10.1111/gtc.12677.

Bissonauth V., Roy S., Gravel M., Guillemette S., Charron J. Requirement for *Map2k1* (*Mek1*) in extra-embryonic ectoderm during placentogenesis. *Development*. 2006;133(17):3429-3440. DOI 10.1242/dev.02526.

Chalitchagorn K., Shuangshoti S., Hourpai N., Kongruttanachok N., Tangkijvanich P., Thong-ngam D., Voravud N., Sriuranpong V., Mutirangura A. Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene*. 2004;23(54):8841-8846. DOI 10.1038/sj.onc.1208137.

Chiu S.Y., Asai N., Costantini F., Hsu W. SUMO-specific protease 2 is essential for modulating p53-mdm2 in development of trophoblast stem cell niches and lineages. *PLoS Biol.* 2008;6(12):e310. DOI 10.1371/journal.pbio.0060310.

Criscione S.W., Teo Y.V., Neretti N. The chromatin landscape of cellular senescence. *Trends Genet*. 2016a;32(11):751-761. DOI 10.1016/j.tig.2016.09.005.

Criscione S.W., Theodosakis N., Micevic G., Cornish T.B., Burns K.H., Neretti N., Rodić N. Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics.* 2016b;17:463. DOI 10.1186/s12864-016-2800-5.

Denli A.M., Narvaiza I., Kerman B.E., Pena M., Benner C., Marchetto M.C., Diedrich J.K., Aslanian A., Ma J., Moresco J.J., Moore L., Hunter T., Saghatelian A., Gage F.H. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell*. 2015;163(3):583-593. DOI 10.1016/j.cell.2015.09.025.

Fahlbusch F.B., Dawood Y., Hartner A., Menendez-Castro C., Nögel S.C., Tzschoppe A., Schneider H., Strissel P., Beckmann M.W., Schleussner E., Ruebner M., Dörr H.G., Schild R.L., Rascher W., Dötsch J. Cullin 7 and Fbxw 8 expression in trophoblastic cells is regulated via oxygen tension: implications for intrauterine growth restriction? *J. Matern. Fetal Neonatal Med.* 2012;25(11):2209-2215. DOI 10.3109/14767058.2012.684166.

Fu J., Lv X., Lin H., Wu L., Wang R., Zhou Z., Zhang B., Wang Y.L., Tsang B.K., Zhu C., Wang H. Ubiquitin ligase cullin 7 induces epithelial-mesenchymal transition in human choriocarcinoma cells. *J. Biol. Chem.* 2010;285(14):10870-10879. DOI 10.1074/jbc.M109.004200.

He Z.M., Li J., Hwa Y.L., Brost B., Fang Q., Jiang S.W. Transition of LINE-1 DNA methylation status and altered expression in first and third trimester placentas. *PLoS One.* 2014;9(5):96994. DOI 10.1371/journal.pone.0096994.

Kang X., Qi Y., Zuo Y., Wang Q., Zou Y., Schwartz R.J., Cheng J., Yeh E.T.H. SUMO-specific protease 2 is essential for suppression of polycomb group protein-mediated gene silencing during embryonic development. *Mol. Cell.* 2010;38(2):191-201. DOI 10.1016/j.molcel.2010.03.005.

Kim E.Y., Chen L., Ma Y., Yu W., Chang J., Moskowitz I.P., Wang J. Enhanced desumoylation in murine hearts by overexpressed SENP2 leads to congenital heart defects and cardiac dysfunction. *J. Mol. Cell. Cardiol.* 2012;52(3):638-649. DOI 10.1016/j.yjmcc.2011.11.011.

Lebedev I.N., Ostroverkhova N.V., Nikitina T.V., Sukhanova N.N., Nazarenko S.A. Features of chromosomal abnormalities in spontaneous abortion cell culture failures detected by interphase FISH analysis. *Eur. J. Hum. Genet.* 2004;12(7):513-520. DOI 10.1038/sj.ejhg.5201178.

Maksimova N., Hara K., Miyashia A., Nikolaeva I., Shiga A., Nogovicina A., Sukhomyasova A., Argunov V., Shvedova A., Ikeuchi T., Nishizawa M., Kuwano R., Onodera O. Clinical, molecular and histopathological features of short stature syndrome with novel CUL7 mutation in Yakuts: new population isolate in Asia. *J. Med. Genet.* 2007;44(12):772-778. DOI 10.1136/jmg.2007.051979.

Ostertag E.M., Kazazian H.H. Jr. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* 2001;35:501-538. DOI 10.1146/annurev.genet.35.102401.091032.

Price E.M., Cotton A.M., Peñaherrera M.S., McFadden D.E., Kobor M.S., Robinson W. Different measures of "genome-wide" DNA methylation exhibit unique properties in placental and somatic tissues. *Epigenetics.* 2012;7(6):652-663. DOI 10.4161/epi.20221.

Robinson W.P., Price E.M. The human placental methylome. *Cold Spring Harb. Perspect. Med.* 2015;5(5):a023044. DOI 10.1101/cshperspect.a023044.

Rosser J.M., An W. L1 expression and regulation in humans and rodents. *Front. Biosci. (Elite Ed.)*. 2012;4(6):2203-2225. DOI 10.2741/537.

Santos F., Hyslop L., Stojkovic P., Leary C., Murdoch A., Reik W., Stojkovic M., Herbert M., Dean W. Evaluation of epigenetic marks in human embryos derived from IVF and ICSI. *Hum. Reprod*. 2010;25(9):2387-2395. DOI 10.1093/humrep/deq151.

Sasaki Y., Taya Y., Saito K., Fujita K., Aoba T., Fujiwara T. Molecular contribution to cleft palate production in cleft lip mice. *Congenit. Anom. (Kyoto)*. 2014;54(2):94-99. DOI 10.1111/cga.12038.

Schroeder D.I., Blair J.D., Lott P., Yu H.O., Hong D., Crary F., Ashwood P., Walker C., Korf I., Robinson W.P., LaSalle J.M. The human placenta methylome. *Proc. Natl. Acad. Sci. USA.* 2013;110(15):6037-6042. DOI 10.1073/pnas.1215145110.

Smith Z.D., Chan M.M., Mikkelsen T.S., Gu H., Gnirke A., Regev A., Meissner A. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature.* 2012;484(7394):339-344. DOI 10.1038/nature10960.

Souquet B., Freed E., Berto A., Andric V., Audugé N., Reina-San-Martin B., Lacy E., Doye V. Nup133 is required for proper nuclear pore basket assembly and dynamics in embryonic stem cells. *Cell Rep.* 2018;23(8):2443-2454. DOI 10.1016/j.celrep.2018.04.070.

Talamillo A., Barroso-Gomila O., Giordano I., Ajuria L., Grillo M., Mayor U., Barrio R. The role of SUMOylation during development. *Biochem. Soc. Trans.* 2020;48(2):463-478. DOI 10.1042/BST20190390.

Vasilyev S.A., Markov A.V., Vasilyeva O.Yu., Tolmacheva E.N., Zatula L.A., Sharysh D.V., Zhigalina D.I., Demeneva V.V., Lebedev I.N. Method of targeted bisulfite massive parallel sequencing of the human LINE-1 retrotransposon promoter. *MethodsX.* 2021a;8:101445. DOI 10.1016/j.mex.2021.101445.

Vasilyev S.A., Timoshevsky V.A., Lebedev I.N. Cytogenetic mechanisms of aneuploidy in somatic cells of chemonuclear industry professionals with incorporated plutonium-239. *Russ. J. Genet*. 2010;46(11):1381-1385. DOI 10.1134/S1022795410110141.

Vasilyev S.A., Tolmacheva E.N., Vasilyeva O.Y., Markov A.V., Zhigalina D.I., Zatula L.A., Lee V.A., Serdyukova E.S., Sazhenova E.A., Nikitina T.V., Kashevarova A.A., Lebedev I.N. LINE-1 retrotransposon methylation in chorionic villi of first trimester miscarriages with aneuploidy. *J. Assist. Reprod. Genet.* 2021b;38(1):139-149. DOI 10.1007/s10815-020-02003-1.

ORCID ID

V.V. Demeneva orcid.org/0000-0002-5315-4914
E.N. Tolmacheva orcid.org/0000-0002-0716-4302
T.V. Nikitina orcid.org/0000-0002-4230-6855
E.A. Sazhenova orcid.org/0000-0003-3875-3932
S.Yu. Yuriev orcid.org/0000-0002-1343-5471
A.Sh. Makhmutkhodzhaev orcid.org/0000-0002-7541-0317
A.S. Zuev orcid.org/0000-0001-9474-9335
S.A. Filatova orcid.org/0000-0002-9344-0253
A.E. Dmitriev orcid.org/0000-0002-0070-4863
Ya.A. Darkova orcid.org/0000-0002-7117-9250
L.P. Nazarenko orcid.org/0000-0002-1861-433X
I.N. Lebedev orcid.org/0000-0002-0482-8046
S.A. Vasilyev orcid.org/0000-0002-5301-070X

Acknowledgements. This work was conducted as part of the implementation of the topic of state task No. 1022072600037-4.

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

Received October 13, 2022. Revised December 26, 2022. Accepted December 30, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):72-82
DOI 10.18699/VJGB-23-10

Original Russian text https://vavilovj-icg.ru/

Changes in DNA methylation profile in liver tissue during progression of HCV-induced fibrosis to hepatocellular carcinoma

I.A. Goncharova ✉, A.A. Zarubin, N.P. Babushkina, I.A. Koroleva, M.S. Nazarenko

Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ irina.goncharova@medgenetics.ru

Abstract. In this study we compared methylation levels of 27,578 CpG sites between paired samples of the tumor and surrounding liver tissues with various degrees of damage (fibrosis, cirrhosis) in HCV-induced hepatocellular carcinoma (HCC) patients, as well as between tumor and normal tissue in non-viral HCC patients, using GSE73003 and GSE37988 data from GEODataSets (https://www.ncbi.nlm.nih.gov/). A significantly lower number of differentially methylated sites (DMS) were found between HCC of non-viral etiology and normal liver tissue, as well as between HCC and fibrosis (32 and 40), than between HCC and cirrhosis (2450 and 2304, respectively, according to GSE73003 and GSE37988 datasets). As the pathological changes in the tissue surrounding the tumor progress, the ratio of hyper-/hypomethylated DMSs in the tumor decreases. Thus, in tumor tissues compared with normal/fibrosis/cirrhosis of the liver, 75/62.5/47.7 % (GSE73003) and 16 % (GSE37988) of CpG sites are hypermethylated, respectively. Persistent hypermethylation of the *ZNF154* and *ZNF540* genes, as well as *CCL20* hypomethylation, were registered in tumor tissue in relation to both liver fibrosis and liver cirrhosis. Protein products of the *EDG4*, *CCL20*, *GPR109A*, and *GRM8* genes, whose CpG sites are characterized by changes in DNA methylation level in tumor tissue in the setting of cirrhosis and fibrosis, belong to "Signaling by G-protein-coupled receptors (GPCRs)" category. However, changes in the methylation level of the "driver" genes for oncopathology (*APC*, *CDKN2B*, *GSTP1*, *ELF4*, *TERT*, *WT1*) are registered in tumor tissue in the setting of liver cirrhosis but not fibrosis. Among the genes hypermethylated in tumor tissue in the setting of liver cirrhosis, the most represented biological pathways are developmental processes, cell-cell signaling, transcription regulation, Wnt-protein binding. Genes hypomethylated in liver tumor tissue in the setting of liver cirrhosis are related to olfactory signal transduction, neuroactive ligand-receptor interaction, keratinization, immune response, inhibition of serine proteases, and zinc metabolism. The genes hypermethylated in the tumor are located at the 7p15.2 locus in the *HOXA* cluster region, and the hypomethylated CpG sites occupy extended regions of the genome in the gene clusters of olfactory receptors (11p15.4), keratin and keratin-associated proteins (12q13.13, 17q21.2, and 21q22.11), epidermal differentiation complex (1q21.3), and immune system function loci 9p21.3 (*IFNA*, *IFNB1*, *IFNW1* cluster) and 19q13.41–19q13.42 (*KLK*, *SIGLEC*, *LILR*, *KIR* clusters). Among the genes of fibrogenesis or DNA repair, cg14143055 (*ADAMDEC1*) is located in the binding region of the HOX gene family transcription factors (TFs), while cg05921699 (*CD79A*), cg06196379 (*TREM1*) and cg10990993 (*MLH1*) are located in the binding region of the ZNF protein family transcription factor (TF). Thus, the DNA methylation profile in the liver in HCV-induced HCC is unique and differs depending on the degree of surrounding tissue lesion – liver fibrosis or liver cirrhosis.
Key words: DNA methylation; chronic hepatitis C; HCV; liver fibrosis; liver cirrhosis; hepatocellular carcinoma.

For citation: Goncharova I.A., Zarubin A.A., Babushkina N.P., Koroleva I.A., Nazarenko M.S. Changes in DNA methylation profile in liver tissue during progression of HCV-induced fibrosis to hepatocellular carcinoma. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):72-82. DOI 10.18699/VJGB-23-10

Изменение профиля метилирования ДНК в ткани печени при прогрессировании HCV-индуцированного фиброза до гепатоцеллюлярной карциномы

И.А. Гончарова ✉, А.А. Зарубин, Н.П. Бабушкина, Ю.А. Королева, М.С. Назаренко

Научно-исследовательский институт медицинской генетики, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ irina.goncharova@medgenetics.ru

Аннотация. С использованием данных GSE73003 и GSE37988, представленных в базе данных GEODataSets (https://www.ncbi.nlm.nih.gov/), проведен сравнительный анализ уровня метилирования 27 578 CpG-сайтов между парными образцами опухолевой и окружающей опухоль тканями печени различной степени поражения (фиброз, цирроз) у больных HCV-индуцированной гепатоцеллюлярной карциномой (ГЦК), а также между опухолевой и нормальной тканью у больного ГЦК невирусной этиологии. Выявлено значительно меньшее число дифференциально метилированных сайтов между нормальной тканью печени и ГЦК невирусной этиологии, а также между ГЦК и фиброзом (32 и 40), чем между ГЦК и циррозом (2450 и 2304 соответственно по данным

GSE73003 и GSE37988). По мере прогрессирования патологического изменения окружающей опухоль ткани уменьшается соотношение количества гипер-/гипометилированных дифференциально метилированных сайтов в опухоли. Так, в опухолевой ткани по сравнению с нормальной/фиброзом/циррозом печени гиперметилированы 75/62.5/47.7 % (GSE73003) и 16 % (GSE37988) CpG-сайтов соответственно. Стойкое гиперметилирование генов *ZNF154* и *ZNF540*, а также гипометилирование *CCL20* зарегистрировано в опухолевой ткани относительно как фиброза, так и цирроза печени. Белковые продукты генов *EDG4*, *CCL20*, *GPR109A* и *GRM8*, CpG-сайты которых характеризуются изменением уровня метилирования ДНК в опухоли на фоне цирроза и фиброза, принадлежат к категории «передачи сигналов рецепторов, связанных с G-белком». Однако изменение уровня метилирования «драйверных» для онкопатологии генов (*APC*, *CDKN2B*, *GSTP1*, *ELF4*, *TERT*, *WT1*) регистрируется в опухолевой ткани на фоне цирроза печени, но не фиброза. Среди гиперметилированных в опухолевой ткани генов на фоне цирроза печени наиболее представленными биологическими путями являются процессы развития, передачи межклеточных сигналов, регуляции транскрипции, связывания с белками Wnt-пути. Гены, гипометилированные в опухолевой ткани печени на фоне ее цирротического поражения, относятся к передаче обонятельных сигналов, нейроактивному взаимодействию лиганда с рецептором, кератинизации, иммунному ответу, ингибированию сериновых протеаз и метаболизму цинка. Гиперметилированные в опухоли гены локализуются в локусе 7p15.2 в регионе кластера *HOXA*, а гипометилированные CpG-сайты занимают протяженные области генома в кластерах генов обонятельных рецепторов (11p15.4), кератина и кератин-ассоциированных белков (12q13.13, 17q21.2 и 21q22.11), комплекса эпидермальной дифференцировки (1q21.3), а также функционирования иммунной системы – локусы 9p21.3 (кластер *IFNA*, *IFNB1*, *IFNW1*) и 19q13.41–19q13.42 (кластеры *KLK*, *SIGLEC*, *LILR*, *KIR*). Среди генов фиброгенеза или репарации ДНК cg14143055 (*ADAMDEC1*) локализован в регионе связывания транскрипционных факторов семейства HOX, а cg05921699 (*CD79A*), cg06196379 (*TREM1*) и cg10990993 (*MLH1*) расположены в области связывания транскрипционных факторов семейства белков ZNF. Таким образом, профиль метилирования ДНК в печени при HCV-индуцированной ГЦК является уникальным и различается в зависимости от степени поражения окружающей ткани – фиброз или цирроз.

Ключевые слова: метилирование ДНК; ХВГС; фиброз печени; цирроз печени; гепатоцеллюлярная карцинома.

Introduction

Malignant neoplasms of the liver are characterized by an increasing incidence rate worldwide (Philips et al., 2021). The highest morbidity and mortality rates are observed in East Asia and Africa, where the leading cause of hepatocellular carcinoma (HCC) is chronic viral hepatitis B and non-alcoholic fatty liver disease (NAFLD). However, in developed countries one of the main causes of HCC development is considered to be chronic viral hepatitis C (chronic HCV, CHCV); and its prevalence is high in Europe and maximal in Eastern European countries, including Russia (Goossens, Hoshida, 2015; Petruzziello et al., 2016).

The molecular mechanisms of HCC development differ significantly depending on the etiology of the disease. Thus, the hepatitis B virus (HBV) can integrate into the genome of the host hepatocyte, which leads to the direct triggering of carcinogenesis through the activation of protooncogenes and/or suppression of the activity of tumor suppressor genes (Levrero, Zucman-Rossi, 2016). In turn, the hepatitis C virus (HCV), which is an RNA virus, has limited ability to integrate into the genome of the host liver cell and realizes its carcinogenic potential by switching on a multi-stage process that leads through chronic liver inflammation and fibrosis progression to the formation and development of tumor clones. The risk of developing HCC in chronic HCV infection is directly related to the severity of liver fibrosis; it is a rare event in the initial stages of fibrosis and occurs significantly more often in patients with cirrhosis (Khatun et al., 2021).

Among the various factors determining susceptibility to HCV infection and the progression of fibrosis to HCC, the genetic and epigenetic component plays an important role. In particular, genome-wide association studies (GWAS) have identified approximately 140 loci, of which 84 are attributed to known genes, the protein products of which are involved in the response to HCV infection, antiviral therapy, spontaneous viral clearance, and the development of complications to interferon therapy (Kanz et al., 2005).

Genes, including *EXO1*, *VCAN*, *KIT* and *MIR200C*, which are associated with the development of HCV-induced HCC and considered as potential targets for pharmacotherapy, have been identified (Goossens, Hoshida, 2015; Schulze et al., 2015; Chen et al., 2021). In addition, microRNAs determined in liver tissue or serum have been shown to have prognostic value in the development of HCV-induced HCC (Aly et al., 2020; Yan et al., 2021).

There are few experimental studies of liver tissue methylome aberrations in liver pathology depending on etiological causes (Neumann et al., 2012; Hlady et al., 2014). The main data regarding viral etiology are the data obtained by comparative analysis of paired tumor and non-tumorous liver tissues in Asian patients with HCC on the Illumina Infinium Human Methylation BeadChip 27k platform (Shen et al., 2012; Mah et al., 2014; Yamada et al., 2016). A number of studies involve reanalysis of the available DNA methylation findings using additional data, including those obtained on the Illumina Human Methylation 450 BeadChip microarray from The Cancer Genome Atlas (Fan et al., 2018; Meng et al., 2018; Wang Y. et al., 2019; Jiang et al., 2020; Zhao et al., 2021).

A comparison of the lists of differentially methylated CpG sites between the analyzed liver tissues in HCC patients in different studies (Shen et al., 2012; Mah et al., 2014; Yamada et al., 2016) reveals significant similarities. For example, the list of hypermethylated genes in tumor tissue presented in the paper of (Yamada et al., 2016) overlaps by 93 % with the data of another group (Mah et al., 2014). A different picture is observed when comparing the results of reanalysis. Thus, common genes are rarely found in the lists of genes significant for the HCC development presented in various studies (Fan et al., 2018; Meng et al., 2018; Wang Y. et al., 2019; Jiang et al., 2020). This can be explained by the different criteria

chosen for the reanalysis of the primary data provided in the GEO repository (Edgar et al., 2002; Barrett et al., 2013). At the same time, none of the mentioned studies took into account the etiology of HCC, and the analyzed group included both carriers of HBV or HCV and patients without viruses or their combinations.

The contribution of DNA methylation to the development of HCV- and HBV-induced HCC has been reviewed in meta-analyses including studies of targeted methylation of genes associated with liver diseases (Zhang et al., 2019, 2022). The genes hypermethylated in liver tumor tissues in HCC of various viral etiologies have been identified. However, these genes are largely common, which does not provide a complete picture of the patterns of the DNA methylation profile in the influence of hepatitis B and C viruses.

Our research team has been working on the genetic aspects of CHCV. As a result, we established the associations of polymorphisms in fibrogenesis genes and DNA repair genes with pathology and pathogenetically significant features, including stages of liver fibrosis (Goncharova et al., 2020). It is possible that there are features of the DNA methylation profile in liver tissue in the setting of fibrosis and cirrhosis induced by HCV and causing HCC.

Thus, the aim of this study was to identify changes in the DNA methylation profile, including the regions of genes involved in fibrogenesis or DNA repair, in liver tissue during the progression of HCV infection from liver fibrosis to HCC using re-analysis of primary data stored in the GEO repository.

Materials and methods

Data from several studies analyzing the profile of DNA methylation in the liver of Asian patients with HCC caused by viral hepatitis B and C on the Illumina Infinium Human Methylation BeadChip 27k platform are available in the GEO database (Table 1). For Caucasians, there is no data available on DNA methylation in HCC in the GEO repository.

From the GSE73003 and GSE37988 datasets, we selected for analysis the patients diagnosed with CHCV by the presence of a hepatitis C virus total antibody (HCVab+) and the absence of a viral hepatitis B surface antigen (HBsAg–). From the GSE73003 dataset, we chose patients with HCV-induced HCC, in which non-tumor liver tissue was characterized by various stages of fibrotic lesion: liver fibrosis in the setting of CHCV (n = 3) and liver cirrhosis (n = 8). In addition, the study included one patient with HCC of unknown etiology, who was HCVab and HBsAg negative, in which the surrounding liver tissue was defined as normal (HCC_normal tissue/normal tissue).

From the GSE37988 array, patients with HCV-induced HCC, in which non-tumor liver tissue was at the stage of cirrhosis (n = 6), were included in the analysis. In the present work, we did not differentiate the tissues and did not use histological sections, but relied only on the data presented in GSE37988 and GSE73003 GEODataSets (https://www.ncbi.nlm.nih.gov/).

As the GSE57956 dataset does not provide information on the etiology of the pathology, in particular hepatitis B and C viral infection, the tissue samples were not included in the present study.

In addition to the 27,578 CpG sites presented on the Illumina Infinium Human Methylation BeadChip 27k methylation array, the methylation status of fibrogenesis genes and DNA repair genes was analyzed separately. We chose genes associated with CHCV, liver fibrosis stages, the rate of fibrosis progression to liver cirrhosis and comorbid pathologies of CHCV, according to our previous studies (Goncharova et al., 2020).

Statistical data analysis was performed using lumi, limma packages in the R software environment (Bioconductor). The correction for multiple comparisons was performed using the Benjamini–Hochberg (FDR) method.

The methylation index β, which represents the ratio of the intensity of fluorescence signals of methylated alleles to the

Table 1. General characterization of studies related to the analysis of the DNA methylation profile in the liver in patients with HCC caused by viral hepatitis B and C using the Illumina Infinium Human Methylation BeadChip 27k

GEO accession number	Population	Number of patients with HCC, liver tissue	Findings	References
GSE37988	Taiwan	n = 62, paired tumor/non-tumor tissues	684 CpG sites were hypermethylated and 1640 were hypomethylated in the tumor compared to non-tumor tissues ($\Delta\beta \geq 0.20$, FDR ≤ 0.05). Hypermethylation in the tumor was confirmed for the *CDKL2*, *STEAP4*, *HIST1H3G*, *CDKN2A* and *ZNF154* genes	Shen et al., 2012
GSE57956	Singapore	n = 59, paired tumor/non-tumor tissues	2037 CpG sites were hypermethylated and 2379 were hypomethylated in the tumor compared to non-tumor tissues ($\Delta\beta > 0.10$, FDR < 0.05). Hypermethylation in the tumor was confirmed for the *SPDY1*, *TSPYL5*, *PKDREJ*, *ZNF154*, *TUBB6*, *CYB5R2* and *SH3YL1* genes, and hypomethylation was confirmed for the *CYB11B1* and *SPRR3* genes	Mah et al., 2014
GSE73003	Japan	n = 20, paired tumor/non-tumor tissues	875 CpG sites were hypermethylated and 1795 were hypomethylated in tumor compared to non-tumor tissues ($\Delta\beta > 0.15$, FDR < 0.01). Hypermethylation in tumor was confirmed for the *APC*, *CDKN2A*, *GSTP1*, *AKR1B1*, *GRASP*, *MAP9*, *NXPE3*, *RSPH9*, *SPINT2*, *STEAP4* and *ZNF154* genes	Yamada et al., 2016

sum of fluorescence signals of methylated and unmethylated alleles, was used as a parameter of DNA methylation level. The methylation index β varies from 0 (unmethylated state) to 1 (complete methylation of all CpG sites at a given position). CpG site was considered as differentially methylated if it had a difference in the average methylation level between the groups of samples with FDR < 0.05 and $|\Delta\beta| \geq 0.2$, which exceeds the microarray measurement error and complements the statistical significance of the differences by a biologically valid criterion.

Functional annotation of protein products of genes containing differentially methylated CpG sites (DMS) was performed using Web-based GEne SeT AnaLysis Toolkit programs with Weighted set cover (Liao et al., 2019) and Metascape (Zhou et al., 2019) category reductions. The categories of the genes described in terms of biological processes and molecular functions correspond to the Gene Ontology (GO) database classifier, in terms of signaling and metabolic pathways correspond to KEGG and Reactome, in terms of drug targets correspond to DrugBank, and in terms of chromosomal localization correspond to Chromosomal Location.

Additionally, we performed the genomic annotation of DMSs in fibrogenesis genes and DNA repair genes in the hepatocellular carcinoma cell line HepG2 using the UCSC Genome Browser (Kent et al., 2002). This annotation allowed us to characterize CpG sites that localize in gene promoters, open chromatin regions accessible to RNA polymerase II or transcription factor (TF) binding sites, and thereby possibly affect changes in gene expression.

Results and discussion

Identification of DMSs and their genes between tumor and non-tumor liver tissues (normal without hepatitis C and B viruses, fibrosis and cirrhosis in the setting of CHCV) in patients with HCC

A comparative analysis of the methylation level of 27,578 CpG sites between paired samples characterized as HCC surrounded by normal tissue and normal liver tissue in a patient without hepatitis C and B viruses (GSE73003) revealed 32 DMSs, among which 24 CpG sites (21 genes) were hypermethylated and 8 CpG sites (7 genes) were hypomethylated in tumor versus normal tissue (Fig. 1, *a*). Two CpG sites were identified in the *RBM4*, *SOX9* and *SPAG8* genes (hypermethylated in tumor tissue), as well as in *ACTA2* (hypomethylated in tumor tissue).

Twenty CpG sites with the greatest differences in methylation levels between tumor and normal liver tissues are presented in Suppl. Material 1[1]. Most of them are located in the region of CpG islands (16 sites or 80 %). Among them are the CpG sites in the *RBM4*, *TRIP12*, *BFSP1*, *FBP1*, *SGCE* and *PTPN4* genes, which have previously been associated with the development of HCC (see Suppl. Material 1).

Forty differentially methylated sites were identified in HCC surrounded by fibrotic tissue versus fibrosis in CHCV (GSE73003) (see Fig. 1, *b*). In the liver tumor tissue, 25 CpG sites (24 genes) were hypermethylated compared to the fibrotic tissue and 15 CpG sites (15 genes) were hypomethylated. Significant changes in methylation levels during oncotransformation of fibrotic liver tissue were shown for the CpG sites of the *ZNF154*, *DNM3*, *DLEC1*, *LYPD3*, *DDX49*, *NEFH*, *CCL20* and *NNMT* genes, which were previously associated with HCC development (see Suppl. Material 1). Moreover, the most significant hypermethylation in the tumor versus fibrosis was detected for two CpG sites located in the region of the CpG island in the 1st exon of the *ZNF154* gene ($\Delta\beta = 0.593–0.596$, FDR < 0.01).

Of all the differentially methylated genes (DMGs), only the *CCL20* protein product is a proangiogenic chemokine that is highly upregulated in cells infected with HCV and induces endothelial cell invasion and migration during HCC formation (Benkheil et al., 2018). The cg21643045 site in the *CCL20* gene, located in exon 1, was hypomethylated in tumor tissue compared to fibrotic tissue ($\Delta\beta = –0.382$, FDR = 0.0235).

A comparison of DNA methylation level between paired samples of liver tissues (tumor and cirrhosis) in CHCV (GSE73003) revealed 2450 DMSs (see Fig. 1, *c*). In tumor-affected liver tissue versus non-tumor tissue, 1168 CpG sites (886 genes) were hypermethylated and 1282 CpG sites (998 genes) were hypomethylated.

Of the twenty CpG sites of genes that showed the most significant changes in methylation level during oncotransformation of liver tissue affected by cirrhosis, the *GRM8*, *DNM3*, *DLEC1*, *ZNF154*, *WNK2*, *MFAP5*, *FOXD3*, *NEFH*, *MTNR1B*, *CCL20* and *RAB31* genes were associated with HCC development (see Suppl. Material 1). Moreover, cg21790626 in the *ZNF154* gene and cg21643045 in the *CCL20* gene were hyper- and hypomethylated, respectively, in tumor tissue versus cirrhotic tissue ($\Delta\beta = 0.598$, FDR = 3.10×10^{-7} and $\Delta\beta = –0.459$, FDR = 1.43×10^{-6}).

A comparative analysis of the methylation level of 27,578 CpG sites between paired samples of tumor and non-tumor liver tissue in the setting of HCV-induced liver cirrhosis (GSE37988) revealed 2304 DMSs (see Fig. 1, *d*). In the liver tumor tissue versus cirrhotic tissue, 386 CpG sites (305 genes) were hypermethylated and 1936 CpG sites (1483 genes) were hypomethylated.

The genes and CpG sites that showed the most significant changes in the methylation level during oncotransformation of liver tissue affected by cirrhosis according to GSE37988 are presented in Suppl. Material 1. Among them, the *MAGEA3*, *APC*, *AKT3*, *MMP26* and *WFDC1* genes are associated with HCC according to previous studies (see Suppl. Material 1). In contrast to the GSE73003, a smaller proportion of CpG sites (7 out of 20, or 35 %) were located in the region of CpG islands. Moreover, only two of them, cg16970232 and cg24332422 in the *APC* gene, were hypermethylated in tumor tissue versus cirrhotic tissue ($\Delta\beta = 0.730$, FDR = 1.0×10^{-4} and $\Delta\beta = 0.581$, FDR = 1.2×10^{-4}).

Characterization of common DMGs between tumor and non-tumor liver tissues (normal without hepatitis C and B viruses, fibrosis and cirrhosis in the setting of CHCV) in patients with HCC

A comparison of the lists of genes containing DMSs between tumor and non-tumor tissues in patients with HCC, depending on the degree of tumor-adjacent liver tissue damage, revealed that the *ZNF154*, *DNM3*, *FLJ21159*, *DLEC1*, *CCDC37*, *NEFH*, *CCL20* and *KRTAP11-1* genes are among the top ones with

[1] Supplementary Materials 1 and 2 are available in the online version of the paper: http://vavilov.elpub.ru/jour/manager/files/Suppl_Goncharova_Engl_27_1.pdf

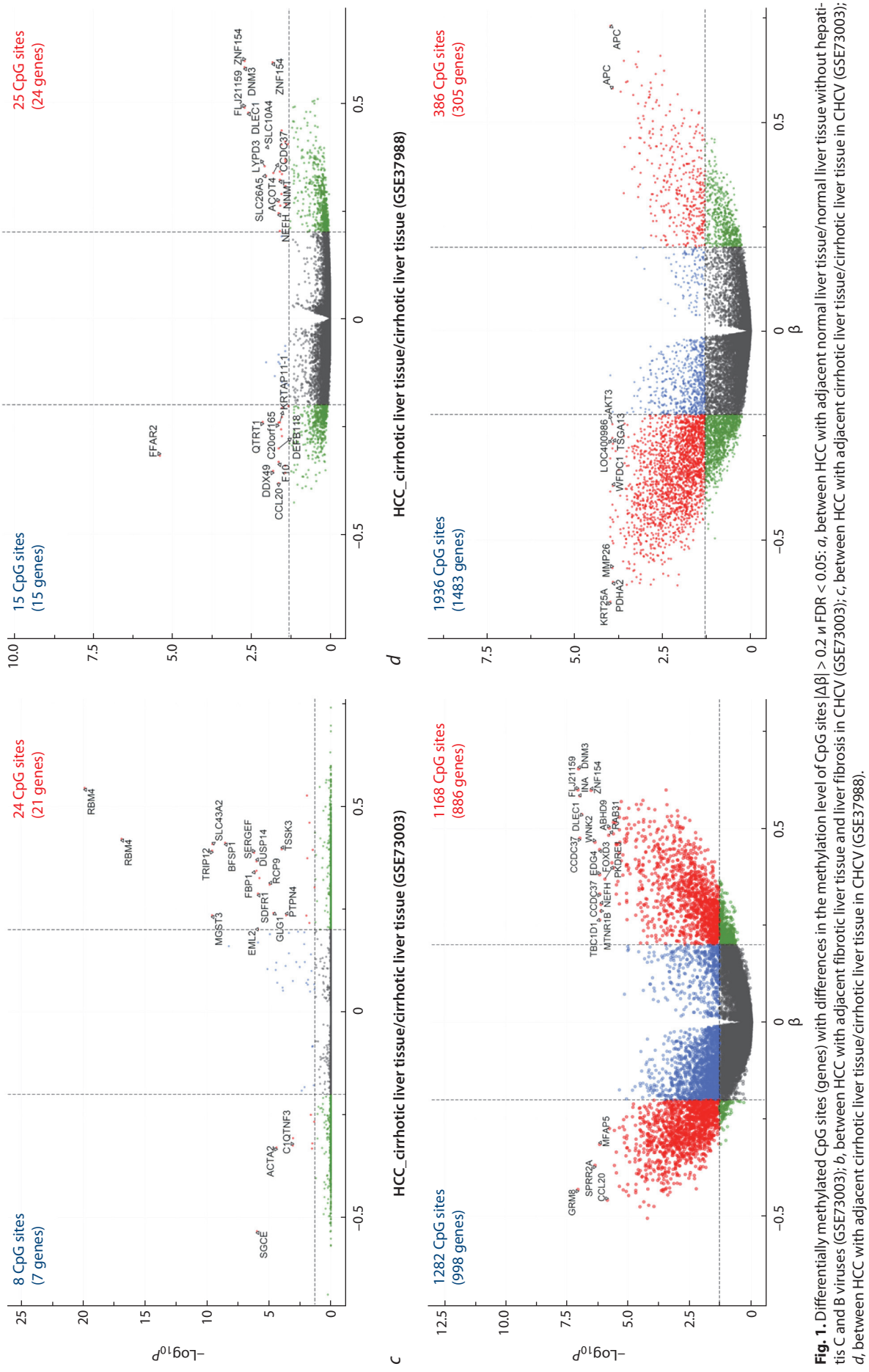

Fig. 1. Differentially methylated CpG sites (genes) with differences in the methylation level of CpG sites $|\Delta\beta| > 0.2$ и FDR < 0.05: *a*, between HCC with adjacent normal liver tissue/normal liver tissue without hepatitis C and B viruses (GSE73003); *b*, between HCC with adjacent fibrotic liver tissue and liver fibrosis in CHCV (GSE73003); *c*, between HCC with adjacent cirrhotic liver tissue/cirrhotic liver tissue in CHCV (GSE73003); *d*, between HCC with adjacent cirrhotic liver tissue/cirrhotic liver tissue in CHCV (GSE37988).

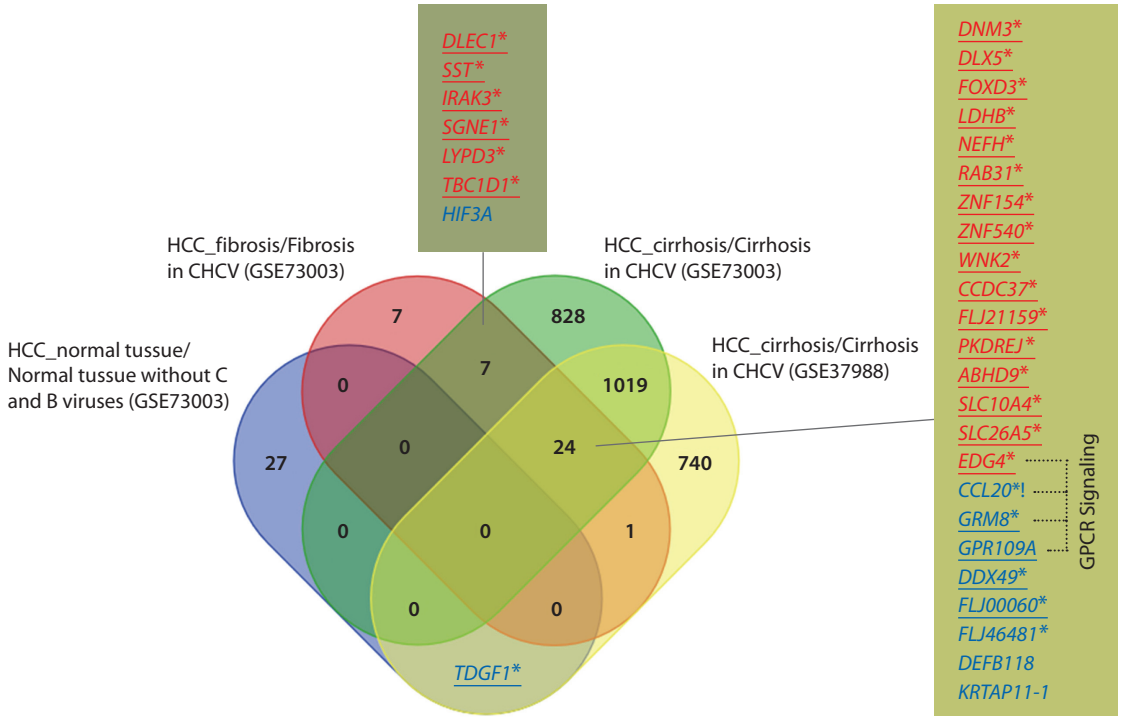

Fig. 2. Venn diagram showing the number of total DMGs between the tumor and adjacent liver tissue of different lesion degrees (normal without C and B viruses, fibrosis and cirrhosis in CHCV) in patients with HCC.

Blue/red – hypo-/hypermethylated genes in tumor tissue versus non-tumor tissue; underlined – location of DMSs in the region of CpG island; */! – gene involved in HCC/HCC in CHCV.

maximum differences in the methylation level of CpG sites between the tissues (see Suppl. Material 1).

The differentially methylated genes between tumor tissues and liver fibrosis/cirrhosis (GSE73003) are characterized by the presence of seven common genes, six of which are hypermethylated in the tumor regardless of the degree of surrounding tissue damage (Fig. 2). Five of the seven DMSs in common genes are located within CpG islands. The *DLEC1*, *SST*, *IRAK3*, *SGNE1*, *LYPD3* and *TBC1D1* genes have previously been shown to be associated with the development of HCC, and the *DLEC1*, *IRAK3* and *SGNE1* genes were hypermethylated in the tumor (Qiu et al., 2008; Kuo et al., 2015; Meng et al., 2018), which is consistent with the results of the present study.

There are 24 DMGs common to tumors in the presence of fibrosis and cirrhosis from the two datasets (GSE73003 and GSE37988). Among them, 16 DMGs (66.7 %) are hypermethylated and located in the CpG island region (see Fig. 2). An association with HCC development has previously been shown for 21 genes: *DNM3* was downregulated and *FOXD3*, *LDHB*, *NEFH*, *ZNF154*, *FLJ21159*, *PKDREJ*, *ABHD9* and *WNK2* were hypermethylated in tumor tissue (Shen et al., 2012; Revill et al., 2013; Liu Z. et al., 2016; Meng et al., 2018; Miller et al., 2021). The *CCDC37*, *CCL20*, *DNM3*, *ZNF154* and *ZNF540* genes overlap with the list of twenty DMGs in HCC regardless of etiology (Shen et al., 2012).

Among the eight genes hypomethylated in the tumor, for *CCL20*, *DDX49* and *GRM8*, the increased expression in blood serum and/or tumor tissue in patients with HCC was previously demonstrated, including upregulation of *CCL20* in HCC in the setting of CHCV (Benkheil et al., 2018; Dai et al., 2021; Gao et al., 2022).

We performed a functional annotation of 24 common DMGs between the tumor and the adjacent liver tissue of various degrees of damage using the Metascape resource (see Fig. 2). It showed the association of hypermethylated (*EDG4*) and hypomethylated genes (*CCL20*, *GPR109A* and *GRM8*) with processes of signaling by G-protein-coupled receptors (R-HSA-372790). Moreover, the expression of the *GRM8* gene in tumor tissue negatively correlates with the survival of patients with HCC, and its methylation level is included in the panel of genes important for disease prediction (Gao et al., 2022). It is thought that GPCRs play the role of oncomodulators, the aberrant expression of which alters various normal signaling pathways in the cells, disrupting angiogenesis, invasion, migration, metastasis, and immune response in HCC initiation and progression, which makes them attractive molecular therapeutic targets (Peng et al., 2018).

The present study revealed hypermethylation of the CpG sites of the *ZNF154* and *ZNF540* genes encoding zinc finger proteins in liver tumor tissue compared to fibrosis and cirrhosis (see Fig. 2). Some proteins of this category are included in the signature of prognostic markers of survival of patients with HBV-induced HCC and are the top hypermethylated genes in HCC of various etiologies (Shen et al., 2012; Wang X. et al., 2021). An analysis of the expression of these genes in the liver showed that in HCC of various etiologies, transcription repression of many zinc finger proteins ZNF is observed (Gonçalves et al., 2022). It is likely that in HCV-induced HCC, the zinc finger protein genes, in particular *ZNF154* and *ZNF540*, can

be promising early markers of oncotransformation, beginning with fibrosis, and not only in the setting of liver cirrhosis.

None of the genes from the list of 24 common DMGs between tumor and adjacent liver tissues in fibrosis and cirrhosis in the setting of CHCV from the two data sets (GSE73003 and GSE37988) were included in the list of known molecular "drivers" of malignancies, including HCC (Hlady et al., 2014; Bailey et al., 2018; Cai et al., 2020; Molina-Sánchez et al., 2020; Zhang et al., 2022). However, such genes are found among DMGs between tumor and cirrhotic tissues. In particular, CpG sites within the CpG islands of the promoters of the *APC*, *CDKN2B*, *GSTP1*, *ELF4* и and *TERT* genes were hypermethylated in tumor tissue, and various CpG sites of the *WT1* gene were characterized by multidirectional changes in their methylation levels.

Functional annotation of DMGs between tumor and non-tumor liver tissues (normal without hepatitis C and B viruses, fibrosis and cirrhosis in the setting of CHCV) in patients with HCC

In terms of the most represented biological pathways and basic molecular functions, the genes harboring hypo- and hypermethylated CpG sites in tumor tissue, compared to cirrhotic tissue in patients with HCC in the setting of CHCV, are similar between the GSE73003 and GSE37988 datasets (Suppl. Material 2). Thus, for genes the CpG sites of which are hypermethylated in tumor tissue, biological processes related to development (GO:0007399, GO:0009790, GO:0048468, FDR < 2.2×10^{-16}) and cell-cell signaling (GO:0007267, FDR < 2.2×10^{-16}, see Suppl. Material 2) are most represented. These results are partially consistent with (Shen et al., 2012) data, where developmental processes are distinguished among the most significant in HCC of various etiologies. Genes containing hypermethylated CpG sites in HCV-induced HCC are similar in molecular functions to genes identified in HCC of various etiologies (Shen et al., 2012) and include transcription regulation and DNA binding (GO:0003700; GO:0140110, GO:0003677, FDR < 0.0002), as well as Wnt-protein binding (GO:0017147, FDR = 1.3×10^{-4}).

The hypermethylated genes are located on chromosome 7 (7p15.2) in the region of the *HOXA* cluster (FDR = 2.3×10^{-5}, see Suppl. Material 2). Previously, identification of DNA methylation signature in liver tissue in HCC showed that 39 out of 214 CpG sites were associated with altered gene expression. This includes genes located in the chr7:27144326–27145664 region in close proximity to homeobox transcription factors (*HOXA6*, *HOXA3*, *HOXA5*, *HOXA7* and *HOXA4*) that are involved in oncogenesis, cell proliferation and migration (Gonçalves et al., 2022).

Hypomethylated genes in HCV-induced HCC are mainly related to the following biological processes: immune and defense responses (GO:0006955, GO:0006952, FDR < 2.2×10^{-16}); G protein-coupled receptor signaling pathway (GO:0007186, FDR < 6.0×10^{-10}); epithelial cell differentiation (GO:0030855, FDR < 2.2×10^{-16}, see Suppl. Material 2), which is partially consistent with the data obtained for HCC of various etiologies (Shen et al., 2012).

According to the molecular functions of hypomethylated genes in HCV-induced HCC of various etiologies (Shen et al., 2012), on the one hand, similarities are revealed with respect to several categories, such as binding to receptors of various antigens, and on the other hand, the activity of peptidase inhibitors, including serine-type peptidase, is noted only in HCV-induced carcinoma (see Suppl. Material 2). The serine protease inhibitor secreted by liver tumor cells (SPINK1 or LC-SPIK) is now known to be a protein that significantly increases in the blood serum of individuals with HCC of viral etiology (Lu et al., 2020).

According to the KEGG and Reactome databases, the most significant molecular pathways for genes hypomethylated in the tumor were olfactory transduction (hsa04740, FDR < < 2.2×10^{-16}), cytokine-cytokine receptor interaction (hsa04060, FDR < 7.8×10^{-7}) and neuroactive ligand-receptor interaction (hsa04080, FDR < 0.0007); signaling by G protein-coupled receptors (GPCRs) (R-HSA-372790, FDR < 3.5×10^{-6}), keratinization (R-HSA-6805567, FDR < 2.2×10^{-16}) and immune system (R-HSA-168256, FDR < 3.4×10^{-6}, see Suppl. Material 2). Apparently, this is due to the fact that DNA hypomethylation in the tumor spreads over extended genome regions in the gene clusters of olfactory receptors (11p15.4), keratin and keratin-associated proteins (12q13.13, 17q21.2 and 21q22.11), epidermal differentiation complex (1q21.3), as well as immune system functioning – loci 9p21.3 (*IFNA*, *IFNB1*, *IFNW1* cluster) and 19q13.41–19q13.42 (*LILR*, *KIR*, *KLK*, *ZNF*, *SIGLEC* clusters, see Suppl. Material 2).

Disruption of epigenetic regulation of the immune system is a common feature in cancers of various localizations (Berglund et al., 2021). Olfactory transduction and neuroactive ligand-receptor interaction are part of the G protein-coupled receptor signaling pathway, the enrichment of which is also common in malignant neoplasms (Wei et al., 2012). Ectopic expression of olfactory receptor genes, associated with epigenetic mechanisms among others, seems to provide invasiveness and metastasis of tumor cells in the late stages of malignancy (Fessahaye et al., 2021). Disruption of the keratinization process is a less frequently reported event in tumor. For it, an association with the DNA hydroxymethylation level in head and neck cancer depending on the carriage of the human papillomavirus is shown (Liu S. et al., 2020), as well as the enrichment of hypomethylated genes in breast cancer (Holm et al., 2016).

The DrugBank database indicates that hypomethylated genes in HCV-induced HCC are involved in zinc metabolism (DB01593); and zinc supplementation may be recommended to reduce the risk of HCC after HCV eradication with direct-acting antiviral agents (Hosui et al., 2021). It is possible that there is an association between zinc deficiency and hypomethylation of DNA in individual genes (Azimi et al., 2022).

The profile of methylation of fibrogenesis genes and DNA repair genes

Ten differentially methylated CpG sites were identified among the genes the protein products of which are involved in the processes of fibrogenesis or DNA repair from the category of genes previously shown to be associated with liver diseases in the studies of our research group (Table 2). The cg03876618 site of the *IGFBP7* gene and the cg14323109 site of the *KDR* gene located in the CpG island regions were hypermethylated in the tumor compared to the surrounding cirrhotic tissue. The CpG sites of the *ADAMDEC1*, *CD79A*, *MMP3* and

Table 2. DMSs of genes involved in fibrogenesis and DNA repair between tumor and cirrhotic liver tissue in patients with HCC

CpG site	Gene	Distance to TSS/location on CpG island	HCC/Cirrhosis			
			β ± SD (GSE73003)	FDR	β ± SD (GSE37988)	FDR
cg14143055	*ADAMDEC1*	1374/no	0.46 ± 0.15/0.70 ± 0.04	0.0106	0.34 ± 0.16/0.73 ± 0.05	0.0038
cg05921699	*CD79A*	477/no	0.52 ± 0.17/0.74 ± 0.03	0.0136	0.45 ± 0.20/0.74 ± 0.05	0.0469
cg16466334	*MMP3*	16/no	0.35 ± 0.16/0.66 ± 0.02	0.0029	0.37 ± 0.22/0.71 ± 0.05	0.0300
cg06196379	*TREM1*	428/no	0.21 ± 0.07/0.44 ± 0.02	0.0004	0.13 ± 0.12/0.36 ± 0.05	0.0139
cg03876618	***IGFBP7***	505/yes	0.55 ± 0.22/0.19 ± 0.02	0.0011	–	–
cg14323109	***KDR***	181/yes	0.34 ± 0.20/0.08 ± 0.02	0.0124	–	–
cg10990993	*MLH1*	1347/no	0.19 ± 0.04/0.42 ± 0.09	0.0036	–	–
cg01053621	*APOA2*	573/no	–	–	0.18 ± 0.12/0.47 ± 0.09	0.0119
cg06531741	*HTR3B*	139/no	–	–	0.45 ± 0.24/0.79 ± 0.03	0.0437
cg03017475	*TAS2R38*	852/no	–	–	0.25 ± 0.11/0.55 ± 0.07	0.0001

Note. TSS – transcription start site; β – methylation level; SD – standard deviation. Bold highlights DMSs/DMGs hypermethylated in tumor tissue compared to liver cirrhosis. The lines indicate that CpG sites are not DMSs between tumor and cirrhotic liver tissues.

TREM1 genes were differentially methylated according to the GSE37988 and GSE73003 datasets (see Table 2).

Genomic annotation using the UCSC Genome Browser showed that in the hepatocellular carcinoma cell line HepG2, the active promoter contains cg01053621 (*APOA2*) and cg10990993 (*MLH1*); and cg03876618 (*IGFBP7*), cg14323109 (*KDR*), cg16466334 (*MMP3*), cg06196379 (*TREM1*) и cg01053621 (*APOA2*) are localized in the RNA polymerase II subunit A binding regions.

The cg14143055 site of the *ADAMDEC1* gene, hypomethylated in tumor tissue, is localized in the binding region of HOX family transcription factors, which play an important role in oncogenesis of various tumors, including HCC (Gonçalves et al., 2022). At the same time, HCV infection and virus core protein expression trigger HOX gene activation (Kasai et al., 2021), which may be one of the factors in the development of HCV-induced HCC.

CpG sites hypomethylated in tumor tissue in HCV-induced HCC – cg05921699 (*CD79A*), cg06196379 (*TREM1*), cg10990993 (*MLH1*) – are located in the binding region of the TFs, representing the zinc finger protein (ZNF) family. ZNF, in addition to regulation of transcription, induce protein-protein interactions, post-transcriptional regulation, lipid metabolism, immune responses, and affect the development of many forms of cancer, including HCC (Li et al., 2022).

In conclusion, it should be noted that our study has a limitation due to the small size of samples of patients with normal and fibrotic tissues surrounding the tumor in CHCV, since in most cases HCC develops in the setting of cirrhotic tissue and other cases are observed much less frequently. The study did not consider the intratumoral and cellular heterogeneity of tissues, which is closely related to the DNA methylation profile (Hlady et al., 2017). Taking into account the fact that the focus of the study was to analyze the DNA methylation profile in the liver in HCV-induced HCC, it is difficult to unambiguously identify CpG sites specific to this pathology. A methodological limitation is the impossibility of distinguishing between DNA methylation and DNA hydroxymethylation, since it undergoes bisulfite modification before hybridization on a methylation profiling microarray.

Conclusion

A comparative analysis of the DNA methylation profile in the liver of patients with HCC between tumor and non-tumor tissues with various degrees of lesion (normal tissue, HCV-induced fibrosis, HCV-induced cirrhosis) showed a significantly lower number of DMSs between HCC and normal tissue without hepatitis C and B viruses/liver fibrosis in CHCV (32 and 40) than between HCC and liver cirrhosis in the setting of HCV in the GSE73003 and GSE37988 datasets (2450 and 2304, respectively).

Based on the fact that the severity of fibrosis correlates with liver function and cirrhosis is the main risk factor for HCC development (Roehlen et al., 2020), we can expect normal and fibrotic liver tissue to be maximally distant from HCC by their epigenetic profile and, as fibrosis progresses to cirrhosis, the number of DMSs between the tumor and the surrounding tissues will decrease. Nevertheless, we see the opposite pattern: the more severe the lesion of the liver tissue surrounding the tumor, the greater the differences in DNA methylation levels observed between them. It is possible that normal liver tissue or tissue with minimal fibrotic lesion helps to restrain the functional imbalance of tumor genome, causing minimal differences in the DNA methylation profile between these tissues. This assumption is indirectly confirmed by the fact that changes in the methylation level of the "driver" genes for HCC are registered in the setting of cirrhosis, but not fibrosis.

As the pathological changes in the liver tissue surrounding the tumor progress, the ratio of hyper-/hypomethylated DMSs in the tumor decreases. Thus, in patients with HCC, 24 CpG sites, or 75 % of all DMSs, are hypermethylated in tumor tissue compared to normal tissue. Compared to liver

tissue affected by fibrosis in the setting of CHCV, 25 out of 40 DMSs, or 62.5 %, are hypermethylated in tumor tissue. When the liver tissue surrounding the tumor is cirrhotic, the number of hypermethylated CpG sites in tumor tissue versus the comparison group is 47.7 and 16 % (GSE73003 and GSE37988, respectively). Previous studies have also revealed the predominance of hypomethylated CpG sites in extended genome regions, including those in the region of genes and intergenic regions, in HCC tumor tissue versus the surrounding cirrhotic liver tissue (Shen et al., 2012; Hlady et al., 2014; Yamada et al., 2016; Yan et al., 2021). The present study shows for the first time that in patients with HCC the tumor in the setting of unaffected liver tissue and with liver fibrosis in CHCV is characterized by a greater proportion of hypermethylated CpG sites, while the number of hypomethylated sites increases in tumor tissue in cirrhosis.

The studies of the profile of gene methylation in the liver in HCC focus on hypermethylated genes, including genes of the ZNF and HOX families, among which the search for markers significant for disease development is performed. At the same time, a comparative analysis showed that in HCV-induced HCC, a greater number of hypermethylated CpG sites were observed in tumor tissue only compared to the surrounding tissue with features of fibrosis. In the case when the tissue surrounding the tumor represents liver cirrhosis, most of the loci in the tumor tissue are hypomethylated, which appears to be a late event that occurs during the transition from the fibrotic damage of liver tissue to malignant transformation.

In this regard, in HCV-induced HCC, attention should also be paid to hypomethylated loci, which, as shown in this study, belong to GPCR proteins (*CCL20*, *GPR109A* and *GRM8*), localized in the binding sites of such TFs as HOX (*ADAMDEC1*), ZNF (*CD79A*, *MLH1*) or in the region of serine protease inhibitor genes, one of which – SPINK1 – is currently considered as a marker capable of detecting HCC of viral etiology at an early stage. In addition, in our work, hypomethylated DMSs were localized in genes associated with zinc metabolism, which is known to play a role in the pathogenesis of many diseases, including HCC.

Thus, the functional state and lesion degree of the tissue surrounding the tumor must be taken into account in studies evaluating the DNA methylation profile in the liver in HCC, since the DMGs spectrum differs significantly between tumor/non-tumor tissue pairs, depending on whether it is relatively normal or with features of fibrosis or cirrhosis. To identify prognostic markers of HCC, including liquid biopsies, the etiology of the disease should be considered, since the spectrum of DMSs and DMGs of HCV-induced HCC only partially overlaps with those identified in the analysis of this pathology of other nature.

References

Aly D.M., Gohar N.A.-H., El-Hady A., Khairy M., Abdullatif M.M. Serum microRNA let-7a-1/let-7d/let-7f and miRNA 143/145 gene expression profiles as potential biomarkers in HCV induced hepatocellular carcinoma. *Asian Pac. J. Cancer Prev*. 2020;21(2):555-562. DOI 10.31557/APJCP.2020.21.2.555.

Azimi Z., Isa M.R., Khan J., Wang S.M., Ismail Z. Association of zinc level with DNA methylation and its consequences: a systematic review. *Heliyon*. 2022;8(10):e10815. DOI 10.1016/j.heliyon.2022.e10815.

Bailey M.H., Tokheim C., Porta-Pardo E., Sengupta S., Bertrand D., Weerasinghe A., Colaprico A., Wendl M.C., Kim J., Reardon B., Ng P.K., Jeong K.J., Cao S., Wang Z., Gao J., Gao Q., Wang F., Liu E.M., Mularoni L., Rubio-Perez C., Nagarajan N., Cortés-Ciriano I., Zhou D.C., Liang W.W., Hess J.M., Yellapantula V.D., Tamborero D., Gonzalez-Perez A., Suphavilai C., Ko J.Y., Khurana E., Park P.J., Van Allen E.M., Liang H., MC3 Working Group, Cancer Genome Atlas Research Network, Lawrence M.S., Godzik A., Lopez-Bigas N., Stuart J., Wheeler D., Getz G., Chen K., Lazar A.J., Mills G.B., Karchin R., Ding L. Comprehensive characterization of cancer driver genes and mutations. *Cell.* 2018;173(2):371-385.e18. DOI 10.1016/j.cell.2018.02.060.

Barrett T., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C.L., Serova N., Davis S., Soboleva A. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41(D1):D991-D995. DOI 10.1093/nar/gks1193.

Benkheil M., Van Haele M., Roskams T., Laporte M., Noppen S., Abbasi K., Delang L., Neyts J., Liekens S. CCL20, a direct-acting pro-angiogenic chemokine induced by hepatitis C virus (HCV): potential role in HCV-related liver cancer. *Exp. Cell Res*. 2018;372(2):168-177. DOI 10.1016/j.yexcr.2018.09.023.

Berglund A., Putney R.M., Hamaidi I., Kim S. Epigenetic dysregulation of immune-related pathways in cancer: bioinformatics tools and visualization. *Exp. Mol. Med*. 2021;53(5):761-771. DOI 10.1038/s12276-021-00612-z.

Cai Y., Tian Y., Wang J., Wei W., Tang Q., Lu L., Luo Z., Li W., Lu Y., Pu J., Yang Z. Identification of driver genes regulating the T-cell – infiltrating levels in hepatocellular carcinoma. *Front. Genet*. 2020;11:560546. DOI 10.3389/fgene.2020.560546.

Chen G., Zhang W., Ben Y. Identification of key regulators of hepatitis C virus-induced hepatocellular carcinoma by integrating whole-genome and transcriptome sequencing data. *Front. Genet*. 2021;12:741608. DOI 10.3389/fgene.2021.741608.

Dai H., Feng J., Nan Z., Wei L., Lin F., Jin R., Zhang S., Wang X., Pan L. Morphine may act via DDX49 to inhibit hepatocellular carcinoma cell growth. *Aging* (*Albany NY*). 2021;13(9):12766-12779. DOI 10.18632/aging.202946.

Edgar R., Domrachev M., Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. DOI 10.1093/nar/30.1.207.

Fan G., Tu Y., Chen C., Sun H., Wan C., Cai X. DNA methylation biomarkers for hepatocellular carcinoma. *Cancer Cell Int*. 2018;18:140. DOI 10.1186/s12935-018-0629-5.

Fessahaye G., Khalid R., Mohamed K.-H., Alsidiq M., Khalid A., Mohamed H., Ibrahim M. Breast cancer and smell: hints from epigenetic and functional alteration of the olfaction. (Preprint). 2021. DOI 10.21203/rs.3.rs-910345/v1.

Gao Y., Liu J., Zhao D., Diao G. A novel prognostic model for identifying the risk of hepatocellular carcinoma based on angiogenesis factors. *Front. Genet.* 2022;13:857215. DOI 10.3389/fgene.2022.857215.

Goncharova I.A., Nazarenko M.S., Beloborodova E.V., Markov A.V., Puzyrev V.P. Genetic structure of a predisposition to the progression of liver fibrosis of various etiologies. *Meditsinskaya Genetika = Medical Genetics*. 2020;19(8):83-84. DOI 10.25557/2073-7998.2020.08.83-84. (in Russian)

Gonçalves E., Gonçalves-Reis M., Pereira-Leal J.B., Cardoso J. DNA methylation fingerprint of hepatocellular carcinoma from tissue and liquid biopsies. *Sci. Rep*. 2022;12(1):11512. DOI 10.1038/s41598-022-15058-0.

Goossens N., Hoshida Y. Hepatitis C virus-induced hepatocellular carcinoma. *Clin. Mol. Hepatol*. 2015;21(2):105-114. DOI 10.3350/cmh.2015.21.2.105.

Hlady R.A., Tiedemann R.L., Puszyk W., Zendejas I., Roberts L.R., Choi J.H., Liu C., Robertson K.D. Epigenetic signatures of alcohol abuse and hepatitis infection during human hepatocarcino-

genesis. *Oncotarget.* 2014;5(19):9425-9443. DOI 10.18632/oncotarget.2444.

Hlady R.A., Zhou D., Puszyk W., Roberts L.R., Liu C., Robertson K.D. Initiation of aberrant DNA methylation patterns and heterogeneity in precancerous lesions of human hepatocellular cancer. *Epigenetics.* 2017;12(3):215-225. DOI 10.1080/15592294.2016.1277297.

Holm K., Staaf J., Lauss M., Aine M., Lindgren D., Bendahl P.-O., Vallon-Christersson J., Barkardottir R.B., Höglund M., Borg Å., Jönsson G., Ringnér M. An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Res*. 2016;18(1):27. DOI 10.1186/s13058-016-0685-5.

Hosui A., Tanimoto T., Okahara T., Ashida M., Ohnishi K., Wakahara Y., Kusumoto Y., Yamaguchi T., Sueyoshi Y., Hirao M., Yamada T., Hiramatsu N. Oral zinc supplementation decreases the risk of HCC development in patients with HCV eradicated by DAA. *Hepatol. Commun*. 2021;5(12):2001-2008. DOI 10.1002/hep4.1782.

Jiang H.Y., Ning G., Wang Y.S., Lv W.-B. 14-CpG-based signature improves the prognosis prediction of hepatocellular carcinoma patients. *BioMed Res. Int.* 2020;2020:9762067. DOI 10.1155/2020/9762067.

Kanz C., Aldebert P., Althorpe N., Baker W., Baldwin A., Bates K., Browne P., van den Broek A., Castro M., Cochrane G., Duggan K., Eberhardt R., Faruque N., Gamble J., Diez F.G., Harte N., Kulikova T., Lin Q., Lombard V., Lopez R., Mancuso R., McHale M., Nardone F., Silventoinen V., Sobhany S., Stoehr P., Tuli M.A., Tzouvara K., Vaughan R., Wu D., Zhu W., Apweiler R. The EMBL nucleotide sequence database. *Nucleic Acids Res.* 2005;33:D29-D33. DOI 10.1093/nar/gki098.

Kasai H., Mochizuki K., Tanaka T., Yamashita A., Matsuura Y., Moriishi K. Induction of HOX genes by hepatitis C virus infection via impairment of histone H2A monoubiquitination. *J. Virol.* 2021;95(6):e01784-20. DOI 10.1128/JVI.01784-20.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996-1006. DOI 10.1101/gr.229102.

Khatun M., Ray R., Ray R.B. Hepatitis C virus associated hepatocellular carcinoma. *Adv. Cancer Res*. 2021;149:103-142. DOI 10.1016/bs.acr.2020.10.003.

Kuo C.C., Shih Y.L., Su H.Y., Yan M.D., Hsieh C.B., Liu C.Y., Huang W.T., Yu M.H., Lin Y.W. Methylation of *IRAK3* is a novel prognostic marker in hepatocellular carcinoma. *World J. Gastroenterol*. 2015;21(13):3960-3969. DOI 10.3748/wjg.v21.i13.3960.

Levrero M., Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma. *J. Hepatol*. 2016;64(1 Suppl.):S84-S101. DOI 10.1016/j.jhep.2016.02.021.

Li X., Han M., Zhang H., Liu F., Pan Y., Zhu J., Liao Z., Chen X., Zhang B. Structures and biological functions of zinc finger proteins and their roles in hepatocellular carcinoma. *Biomark. Res*. 2022;10(1):2. DOI 10.1186/s40364-021-00345-1.

Liao Y., Wang J., Jaehnig E.J., Shi Z., Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47:W199-W205. DOI 10.1093/nar/gkz401.

Liu S., de Medeiros M.C., Fernandez E.M., Zarins K.R., Cavalcante R.G., Qin T., Wolf G.T., Figueroa M.E., D'Silva N.J., Rozek L.S., Sartor M.A. 5-Hydroxymethylation highlights the heterogeneity in keratinization and cell junctions in head and neck cancers. *Clin. Epigenetics.* 2020;12(1):175. DOI 10.1186/s13148-020-00965-8.

Liu Z., Yan H., Zhang J. Blood DNA methylation markers in potentially identified Chinese patients with hepatocellular carcinoma. *Pak. J. Pharm. Sci*. 2016;29(4 Suppl.):1451-1456.

Lu F., Shah P.A., Rao A., Gifford-Hollingsworth C., Chen A., Trey G., Soryal M., Talat A., Aslam A., Nasir B., Choudhry S., Ishtiaq R., Sanoff H., Conteh L.F., Noonan A., Hu K.Q., Schmidt C., Fu M., Civan J., Xiao G., Lau D.T., Lu X. Liver cancer-specific serine protease inhibitor kazal is a potentially novel biomarker for the early detection of hepatocellular carcinoma. *Clin. Transl. Gastroenterol.* 2020;11(12):e00271. DOI 10.14309/ctg.0000000000000271.

Mah W.C., Thurnherr T., Chow P.K., Chung A.Y., Ooi L.L., Toh H.C., Teh B.T., Saunthararajah Y., Lee C.G. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PLoS One*. 2014;9(8):e104158. DOI 10.1371/journal.pone.0104158.

Meng C., Shen X., Jiang W. Potential biomarkers of HCC based on gene expression and DNA methylation profiles. *Oncol. Lett.* 2018;16(3):3183-3192. DOI 10.3892/ol.2018.9020.

Miller B.F., Petrykowska H.M., Elnitski L. Assessing *ZNF154* methylation in patient plasma as a multicancer marker in liquid biopsies from colon, liver, ovarian and pancreatic cancer patients. *Sci. Rep*. 2021;11(1):221. DOI 10.1038/s41598-020-80345-7.

Molina-Sánchez P., Ruiz de Galarreta M., Yao M.A., Lindblad K.E., Bresnahan E., Bitterman E., Martin T.C., Rubenstein T., Nie K., Golas J., Choudhary S., Bárcena-Varela M., Elmas A., Miguela V., Ding Y., Kan Z., Grinspan L.T., Huang K.L., Parsons R.E., Shields D.J., Rollins R.A., Lujambio A. Cooperation between distinct cancer driver genes underlies intertumor heterogeneity in hepatocellular carcinoma. *Gastroenterology*. 2020;159(6):2203-2220.e14. DOI 10.1053/j.gastro.2020.08.015.

Neumann O., Kesselmeier M., Geffers R., Pellegrino R., Radlwimmer B., Hoffmann K., Ehemann V., Schemmer P., Schirmacher P., Bermejo J.L., Longerich T. Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors. *Hepatology*. 2012;56(5):1817-1827. DOI 10.1002/hep.25870.

Peng W.T., Sun W.Y., Li X.R., Sun J.C., Du J.J., Wei W. Emerging roles of G protein-coupled receptors in hepatocellular carcinoma. *Int. J. Mol. Sci*. 2018;19(5):1366. DOI 10.3390/ijms19051366.

Petruzziello A., Marigliano S., Loquercio G., Cozzolino A., Cacciapuoti C. Global epidemiology of hepatitis C virus infection: an up-date of the distribution and circulation of hepatitis C virus genotypes. *World J. Gastroenterol*. 2016;22(34):7824-7840. DOI 10.3748/wjg.v22.i34.7824.

Philips C.A., Rajesh S., Nair D.C., Ahamed R., Abduljaleel J.K., Augustine P. Hepatocellular carcinoma in 2021: an exhaustive update. *Cureus*. 2021;13(11):e19274. DOI 10.7759/cureus.19274.

Qiu G.H., Salto-Tellez M., Ross J.A., Yeo W., Cui Y., Wheelhouse N., Chen G.G., Harrison D., Lai P., Tao Q., Hooi S.C. The tumor suppressor gene *DLEC1* is frequently silenced by DNA methylation in hepatocellular carcinoma and induces G1 arrest in cell cycle. *J. Hepatol*. 2008;48(3):433-441. DOI 10.1016/j.jhep.2007.11.015.

Revill K., Wang T., Lachenmayer A., Kojima K., Harrington A., Li J., Hoshida Y., Llovet J.M., Powers S. Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology*. 2013;145(6):1424-1435.e25. DOI 10.1053/j.gastro.2013.08.055.

Roehlen N., Crouchet E., Baumert T.F. Liver fibrosis: mechanistic concepts and therapeutic perspectives. *Cells.* 2020;9(4):875. DOI 10.3390/cells9040875.

Schulze K., Imbeaud S., Letouzé E., Alexandrov L.B., Calderaro J., Rebouissou S., Couchy G., Meiller C., Shinde J., Soysouvanh F., Calatayud A.L., Pinyol R., Pelletier L., Balabaud C., Laurent A., Blanc J.F., Mazzaferro V., Calvo F., Villanueva A., Nault J.C., Bioulac-Sage P., Stratton M.R., Llovet J.M., Zucman-Rossi J. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* 2015;47(5):505-511. DOI 10.1038/ng.3252.

Shen J., Wang S., Zhang Y.J., Kappil M., Wu H.C., Kibriya M.G., Wang Q., Jasmine F., Ahsan H., Lee P.H., Yu M.W., Chen C.J., Santella R.M. Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology*. 2012;55(6):1799-1808. DOI 10.1002/hep.25569.

Wang X., Xing Z., Xu H., Yang H., Xing T. Development and validation of epithelial mesenchymal transition-related prognostic model for hepatocellular carcinoma. *Aging* (*Albany NY*). 2021;13(10):13822-13845. DOI 10.18632/aging.202976.

Wang Y., Ruan Z., Yu S., Tian T., Liang X., Jing L., Li W., Wang X., Xiang L., Claret F.X., Nan K., Guo H. A four-methylated mRNA

signature-based risk score system predicts survival in patients with hepatocellular carcinoma. *Aging* (*Albany NY*). 2019;11(1):160-173. DOI 10.18632/aging.101738.

Wei P., Tang H., Li D. Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PLoS One*. 2012;7(10):e46887. DOI 10.1371/journal.pone.0046887.

Yamada N., Yasui K., Dohi O., Gen Y., Tomie A., Kitaichi T., Iwai N., Mitsuyoshi H., Sumida Y., Moriguchi M., Yamaguchi K., Nishikawa T., Umemura A., Naito Y., Tanaka S., Arii S., Itoh Y. Genome-wide DNA methylation analysis in hepatocellular carcinoma. *Oncol. Rep*. 2016;35(4):2228-2236. DOI 10.3892/or.2016.4619.

Yan P., Pang P., Hu X., Wang A., Zhang H., Ma Y., Zhang K., Ye Y., Zhou B., Mao J. Specific MiRNAs in naïve T cells associated with hepatitis C virus-induced hepatocellular carcinoma. *J. Cancer*. 2021;12(1):1-9. DOI 10.7150/jca.49594.

Zhang C., Huang C., Sui X., Zhong X., Yang W., Hu X., Li Y. Association between gene methylation and HBV infection in hepatocellular carcinoma: a meta-analysis. *J. Cancer*. 2019;10(25):6457-6465. DOI 10.7150/jca.33005.

Zhang C., Zhang W., Yuan Z., Yang W., Hu X., Duan S., Wei Q. Contribution of DNA methylation to the risk of hepatitis C virus-associated hepatocellular carcinoma: a meta-analysis. *Pathol. Res. Pract*. 2022;238:154136. DOI 10.1016/j.prp.2022.154136.

Zhao P., Malik S., Xing S. Epigenetic mechanisms involved in HCV-induced hepatocellular carcinoma (HCC). *Front. Oncol.* 2021;11:677926. DOI 10.3389/fonc.2021.677926.

Zhou Y., Zhou B., Pache L., Chang M., Khodabakhshi A.H., Tanaseichuk O., Benner C., Chanda S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun*. 2019;10:1523. DOI 10.1038/s41467-019-09234-6.

ORCID ID

I.A. Goncharova orcid.org/0000-0002-9527-7015
A.A. Zarubin orcid.org/0000-0001-6568-6339
N.P. Babushkina orcid.org/0000-0001-6133-8986
I.A. Koroleva orcid.org/0000-0003-1498-6934
M.S. Nazarenko orcid.org/0000-0002-0673-4094

Acknowledgements. The research was carried out as part of the state assignment of the Ministry of Science and Higher Education, No. 122020300041-7.

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

Received October 18, 2022. Revised December 8, 2022. Accepted December 15, 2022.

Вавиловский журнал генетики и селекции. 2023;27(1):83-87
DOI 10.18699/VJGB-23-11

Original Russian text https://vavilovj-icg.ru/

Influence of human peripheral blood samples preprocessing on the quality of Hi-C libraries

M.M. Gridina[1] ✉, E. Vesna[1], M.E. Minzhenkova[2], N.V. Shilova[2], O.P. Ryzhkova[2], L.P. Nazarenko[3], E.O. Belyaeva[3], I.N. Lebedev[3], V.S. Fishman[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Centre for Medical Genetics, Moscow, Russia
[3] Tomsk National Research Medical Center of the Russian Academy of Sciences, Tomsk, Russia
✉ gridinam@gmail.com

Abstract. The genome-wide variant of the chromatin conformation capture technique (Hi-C) is a powerful tool for revealing patterns of genome spatial organization, as well as for understanding the effects of their disturbance on disease development. In addition, Hi-C can be used to detect chromosomal rearrangements, including balanced translocations and inversions. The use of the Hi-C method for the detection of chromosomal rearrangements is becoming more widespread. Modern high-throughput methods of genome analysis can effectively reveal point mutations and unbalanced chromosomal rearrangements. However, their sensitivity for determining translocations and inversions remains rather low. The storage of whole blood samples can affect the amount and integrity of genomic DNA, and it can distort the results of subsequent analyses if the storage was not under proper conditions. The Hi-C method is extremely demanding on the input material. The necessary condition for successfully applying Hi-C and obtaining high-quality data is the preservation of the spatial chromatin organization within the nucleus. The purpose of this study was to determine the optimal storage conditions of blood samples for subsequent Hi-C analysis. We selected 10 different conditions for blood storage and sample processing. For each condition, we prepared and sequenced Hi-C libraries. The quality of the obtained data was compared. As a result of the work, we formulated the requirements for the storage and processing of samples to obtain high-quality Hi-C data. We have established the minimum volume of blood sufficient for conducting Hi-C analysis. In addition, we have identified the most suitable methods for isolation of peripheral blood mononuclear cells and their long-term storage. The main requirement we have formulated is not to freeze whole blood.
Key words: Hi-C; human peripheral blood; blood samples storage.

For citation: Gridina M.M., Vesna E., Minzhenkova M.E., Shilova N.V., Ryzhkova O.P., Nazarenko L.P., Belyaeva E.O., Lebedev I.N., Fishman V.S. Influence of human peripheral blood samples preprocessing on the quality of Hi-C libraries. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(1):83-87. DOI 10.18699/VJGB-23-11

Влияние предварительной обработки образцов периферической крови человека на качество Hi-C библиотек

М.М. Гридина[1] ✉, Э. Весна[1], М.Е. Миньженкова[2], Н.В. Шилова[2], О.П. Рыжкова[2], Л.П. Назаренко[3], Е.О. Беляева[3], И.Н. Лебедев[3], В.С. Фишман[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Медико-генетический научный центр им. академика Н.П. Бочкова, Москва, Россия
[3] Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия
✉ gridinam@gmail.com

Аннотация. Метод захвата конформации хроматина в его полногеномном варианте (Hi-C) – мощный инструмент не только для выявления закономерностей пространственной организации генома, но и для понимания влияния их нарушения на развитие заболеваний. Кроме того, метод может быть использован для детекции хромосомных перестроек, в том числе сбалансированных транслокаций и инверсий. Применение метода Hi-C для поиска хромосомных перестроек получает все более широкое распространение. Это связано с тем, что современные высокопроизводительные методы анализа генома позволяют эффективно детектировать точечные мутации и несбалансированные хромосомные перестройки. Однако чувствительность этих методов для определения сбалансированных транслокаций и инверсий остается достаточно низкой. Хранение образцов цельной крови может влиять на количество и целостность выделяемой из них геномной ДНК, а кроме того, приводить к искажению результатов последующих анализов в том случае, если хранение осуществлялось в ненадлежащих условиях. Метод Hi-C крайне требователен к исходному материалу, так как необходимым условием для его успешного применения и получения качественных данных является сохранение пространственной укладки хроматина внутри ядра. Цель нашего исследования состояла в том, чтобы определить оптимальные условия

хранения крови для проведения последующего анализа Hi-C. Были выбраны 10 различных условий хранения образцов крови и пробоподготовки. Для каждого условия приготовлены Hi-C библиотеки и отсеквенированы, после чего оценивалось качество полученных библиотек. В результате сформулированы требования к хранению и подготовке образцов, необходимые для получения качественных Hi-C данных. Нами установлен минимальный объем образца крови, достаточный для проведения Hi-C анализа. Помимо этого, мы определили способы выделения ядерных элементов крови и их долгосрочного хранения, наиболее подходящие для последующего проведения Hi-C анализа. Основное требование, сформулированное нами, – не замораживать цельную кровь.

Ключевые слова: Hi-C; периферическая кровь человека; хранение образцов крови.

Introduction

The combination of chromatin conformation capture methods with whole genome sequencing led to the development of a simple and efficient Hi-C protocol that allows genome-wide studying of chromatin architecture (Lieberman-Aiden et al., 2009; Rao et al., 2014). In addition to the data concerning the organization and dynamics of chromatin in the nucleus, the Hi-C results showed that the relationship between three-dimensional distance in nuclear space and "nucleotide" distance in genomic coordinates can be described by a power function for all cell types. This means that chromosomal rearrangements have effects not only on the contacts frequency of regions directly located at the points of chromosome breaks, but also change the pattern of three-dimensional contacts of a wide area around the rearrangement boundary (Mozheiko, Fishman, 2019). Chromosomal rearrangements detecting methods based on the analysis of the chromatine three-dimensional organization have recently been proposed (Harewood et al., 2017; Chakraborty, Ay, 2018; Díaz et al., 2018; Fishman et al., 2018; Melo et al., 2020). These methods detect various types of rearrangements, including balanced ones, which are still difficult to detect by other methods (Hakim et al., 2012; Dong et al., 2017). In addition, information about single nucleotide variations can be obtained from Hi-C data (Mozheiko, Fishman, 2019), which is important for medical genetics.

Whole blood is a common biological starting sample for medical genetics. Proper blood samples handling is critical for genome-wide studies. Long-term storage and inadequate storage conditions lead to a decrease in the amount of isolated DNA (Nederhand et al., 2003; Malentacchi et al., 2015; Schröder, Steimer, 2018) and its degradation (Ross et al., 1990; Permenter et al., 2015). A high degree of DNA degradation is a serious problem for subsequent molecular biological analyses (Palmirotta et al., 2011; Malentacchi et al., 2015). For example, an increase in the storage time of a blood sample leads to an overestimation of the level of DNA methylation, which may be due to the different stability of methylated and unmethylated DNA (Schröder, Steimer, 2018).

The key steps of the Hi-C protocol are chromatin fragmentation and ligation. To obtain high-quality datasets, it is necessary that both of these steps take place *in nucleus*, under conditions of maximum preservation of the nucleus integrity. Thus, unlike DNA sequence analysis methods, the Hi-C method imposes additional requirements on the quality of the input material. In this regard, it seems relevant to determine the appropriate storage conditions for blood samples intended for Hi-C analysis.

Materials and methods

Peripheral human blood was collected from the antecubital vein into Vacutainer EDTA Blood Collection Tubes. Blood samples storage conditions and preprocessing steps are specified in the Table and Figure 1.

The isolation of peripheral blood mononuclear cells (PBMC) from 3 ml of whole blood was performed using one of the following methods:

- Red Blood Cell Lysis Buffer (RBCL, BioLegend) was used for lysis of erythrocytes according to the manufacturer's instructions. Then the cells were washed once with phosphate buffer saline (PBS).
- centrifugation 300 g for 10 minutes. Serum, including interphase, was transferred into PBS and centrifuged 300 g for 10 minutes.
- sedimentation method on the density gradient Histopaque-1077 Hybri-max (Sigma) according to the manufacturer's instructions.

Cryopreservation of PBMC was performed in a cell freezing medium: 10 % DMSO, 90 % KSR (Thermo Fisher Scientific). Cells were frozen at –80 °C and stored in liquid nitrogen. After thawing, the cells were washed once with PBS.

Cells were counted and resuspended in PBS at a concentration of 1 million cells/ml. Cell fixation, Hi-C library preparation, and data analysis were performed as described in Gridina et al. (2021) using DNase I (Thermo Fisher Scientific) or S1 nuclease (Thermo Fisher Scientific) for chromatin frag-

Storage and preprocessing conditions

#	A brief description of blood samples storage conditions and preprocessing (the time from the blood collection)
1	Less than 4 hrs; RBCLB
2	Less than 4 hrs; RBCLB; freezing KSR+DMSO
3	Less than 4 hrs; centrifugation
4	Less than 4 hrs; centrifugation; freezing KSR+DMSO
5	24 hrs RT; RBCLB
6	–20 °C 4 days; RBCLB
7	+4 °C 2 days; RBCLB
8	+4 °C 4 days; RBCLB
9	+4 °C 7 days; RBCLB
10	Less than 4 hrs; Histopaque-1077 Hybri-max; freezing KSR+DMSO

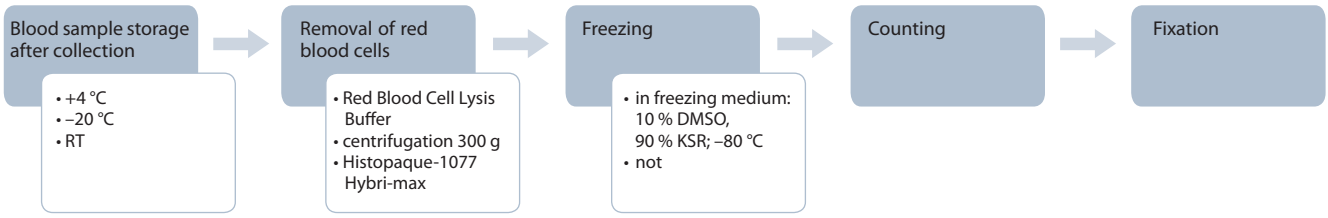

Fig. 1. Blood samples preprocessing.

mentation. HAPA Hyper prep and QIAseq® FX DNA Library Kit (Qiagen) were used for NGS libraries preparation, according to the manufacturer's instructions. The DNA concentration was measured using a Qubit 3.0 fluorimeter (Thermo Fisher Scientific). NGS libraries were sequenced on HiSeq XTen (Illumina) with 150 bp paired reads.

Results and discussion

The first Hi-C step is cells fixation with paraformaldehyde, which is necessary to preserve the native spatial organization of chromatin within the nucleus. Unfortunately, it is not always possible to deliver the sample to the laboratory for fixation on the blood collection day. We decided to systematically estimate the impact of blood storage and preprocessing conditions on the quality of the obtained Hi-C data. Ten conditions were chosen, which included: different methods of PBMC isolation from whole blood; different time and temperature of sample storage; the possibility of freezing PBMC before fixing for long-term storage (see Fig. 1 and the Table).

Although blood sampling is a minimally invasive procedure for biomedical diagnostics, it is clear that there are certain limits on the amount of blood that can be obtained from a patient. Especially if the patient is a small child, or has certain problems with the blood coagulation system. Hi-C analysis requires 1.5–2.5 million cells. Normally, 1 ml of blood contains $(4–11) \times 10^6$ cells. To test each condition, 3 ml of whole blood was taken in two replicates. The PBMC were counted (Fig. 2) after erythrocytes lysis but before cells fixation. A significantly higher number of cells were in the samples processed according to condition #3 (isolation of nuclear elements without RBCL treatment). We did not determine the proportion of living cells during counting. It is possible that dying cells were preserved in samples #3, whereas they were lysed in other cases using RBCL buffer (Brown et al., 2016) or freezing. Supporting the assumption, there were significantly less cells in samples #4 and #10 that were not treated with RBCL but were frozen than in #2.

There were no signs of hemolysis before the start of the isolation of PBMC for all samples except #6, and it was not possible to evaluate this parameter for samples #6. Hemolysis should be avoided, as it is one of the main factors negatively affecting the amount of DNA isolated from blood (Caboux et al., 2012), which may be associated with DNA degradation by nucleases released from degrading cells.

Cell conglomerates were formed in some samples during erythrocyte lysis and subsequent washings. The conglomerates were in both replicates in samples #6, #8, #9 and #10. For these samples, it was not possible to accurately count cells and aliquot them uniformly.

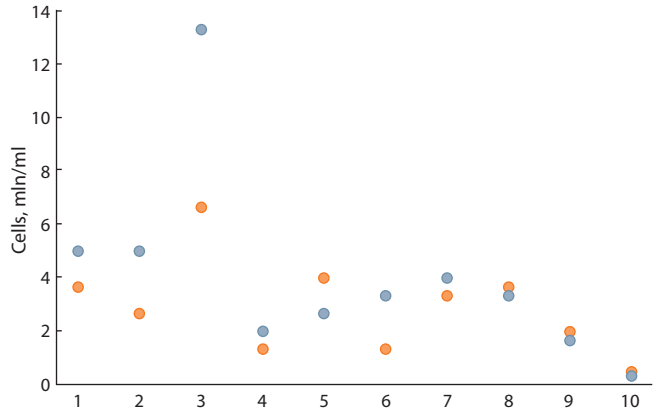

Fig. 2. The PBMC count in 1 ml of blood.

Colors indicate replicates. The horizontal axis represents the storage and preprocessing conditions described in the Table.

2.5 million fixed cells were taken to prepare Hi-C libraries. To assess the quality of Hi-C libraries (Belaghzal et al., 2017), the following controls were made: genomic DNA, DNA after chromatin fragmentation and after ligation. All controls looked accepted (Fig. 3).

We sequenced the Hi-C libraries using paired-end reads with a length of 150 bp, mapped the paired-end reads to the human hg19 genome (GRCh_37) and estimated quality metrics of Hi-C datasets. All libraries had a high proportion of aligned reads (Fig. 4, *a*).

Previously, we have shown (Gridina et al., 2021) that the most important quality metric of Hi-C datasets is the proportion of *cis* interactions (ratio *cis/all* (FF and RR orient)) (see Fig. 4, *b*). It reflects the proportion of Hi-C reads that mapped on the same chromosome among all Hi-C reads. The percentage of *cis* interactions was comparable for all libraries except samples #6 where it was 40.3 and 35.7 %. It means that these Hi-C data are not informative as most fragments ligated randomly. Blood samples #6 were frozen without a cryoprotectant and stored for 4 days at –20 °C. The observed low percentage of *cis* interactions might be due to random DNA strand breaks occurring when cells are frozen without cryoprotectants (Narayanan et al., 2001; Peng et al., 2008; Al-Salmani et al., 2011). On the other hand, this method of freezing leads to ice crystals formation inside the cell and, as a result, to the breaking of cellular and nuclear structures (Mazur, 1984). The release of DNA fragments from the nucleus and their ligation in solution can occur in any way, which leads to the formation of non-informative DNA fragments.

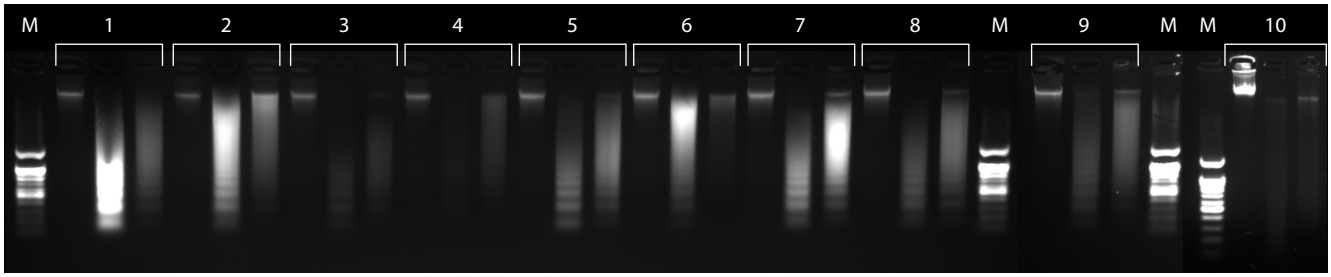

Fig. 3. Chromatin fragmentation and ligation controls in Hi-C experiments.

The numbers represent the storage and preprocessing conditions described in the Table. The order of samples: gDNA, fragmented DNA, ligated DNA. M – DNA ladder 100 bp.

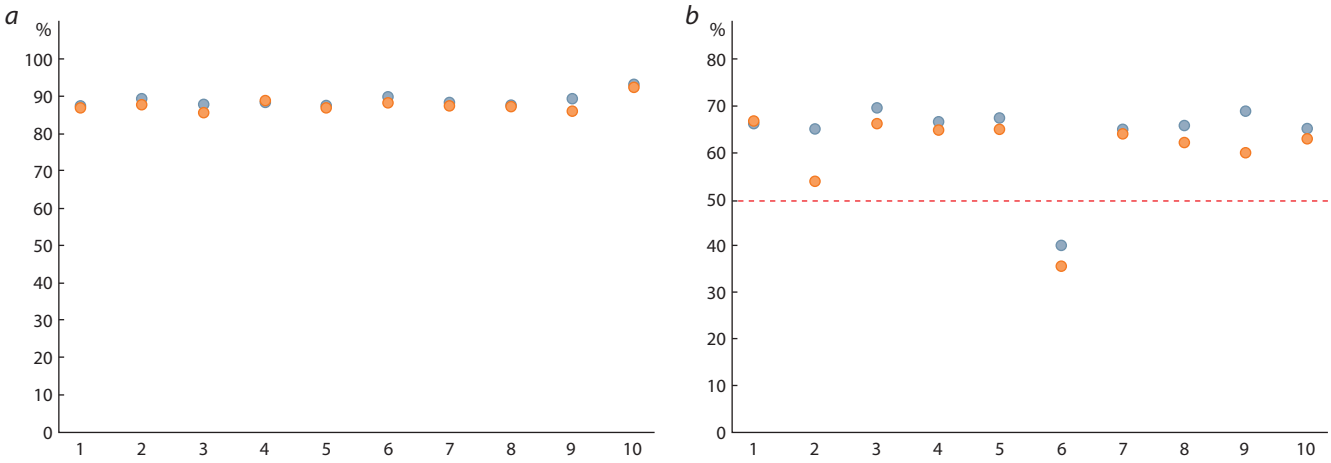

Fig. 4. Quality metrics of Hi-C datasets: *a*, aligned reads; *b*, *cis* interactions.

Colors indicate replicates. The horizontal axes represent the storage and preprocessing conditions described in the Table.

Conclusions

We systematically evaluated various blood samples storage and preprocessing conditions in this work.

As a result, we formulated the following recommendations for the storage and preprocessing of blood samples for Hi-C analysis:

- If it is not possible to deliver the sample on the blood collection day, the samples can be stored at +4 °C for a minimum of 7 days.
- It is better to lyse red blood cells with RBCL buffer before cryopreservation.
- 1–2 ml of whole blood is sufficient (in a person without signs of leukopenia), but if the sample is going to be stored for more than 48 hours, the volume should be increased up to 4–6 ml.
- Never freeze whole blood.

References

Al-Salmani K., Abbas H.H., Schulpen S., Karbaschi M., Abdalla I., Bowman K.J., So K.K., Evans M.D., Jones G.D., Godschalk R.W., Cooke M.S. Simplified method for the collection, storage, and comet assay analysis of DNA damage in whole blood. *Free Radic. Biol. Med.* 2011;51(3):719-725. DOI 10.1016/j.freeradbiomed.2011.05.020.

Belaghzal H., Dekker J., Gibcus J.H. Hi-C 2.0: an optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*. 2017;123:56-65. DOI 10.1016/j.ymeth.2017.04.004.

Brown W.E., Hu J.C., Athanasiou K.A. Ammonium-chloride-potassium lysing buffer treatment of fully differentiated cells increases cell purity and resulting neotissue functional properties. *Tissue Eng. Part C Methods*. 2016;22(9):895-903. DOI 10.1089/ten.tec.2016.0184.

Caboux E., Lallemand C., Ferro G., Hémon B., Mendy M., Biessy C., Sims M., Wareham N., Britten A., Boland A., Hutchinson A., Siddiq A., Vineis P., Riboli E., Romieu I., Rinaldi S., Gunter M.J., Peeters P.H.M., van der Schouw Y.T., Travis R., Bueno-de-Mesquita H.B., Canzian F., Sánchez M.-J., Skeie G., Olsen K.S., Lund E., Bilbao R., Sala N., Barricarte A., Palli D., Navarro C., Panico S., Redondo M.L., Polidoro S., Dossus L., Boutron-Ruault M.C., Clavel-Chapelon F., Trichopoulou A., Trichopoulos D., Lagiou P., Boeing H., Fisher E., Tumino R., Agnoli C., Hainaut P. Sources of pre-analytical variations in yield of DNA extracted from blood samples: analysis of 50,000 DNA samples in EPIC. *PLoS One*. 2012;7(7):e39821. DOI 10.1371/journal.pone.0039821.

Chakraborty A., Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics*. 2018;34(2):338-345. DOI 10.1093/bioinformatics/btx664.

Díaz N., Kruse K., Erdmann T., Staiger A.M., Ott G., Lenz G., Vaquerizas J.M. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun*. 2018;9(1):4938. DOI 10.1038/s41467-018-06961-0.

Dong Z., Wang H., Chen H., Jiang H., Yuan J., Yang Z., Wang W.-J., Xu F., Guo X., Cao Y., Zhu Z., Geng C., Cheung W.C., Kwok Y.K., Yang H., Leung T.Y., Morton C.C., Cheung S.W., Choy K.W. Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the

future of clinical cytogenetics. *Genet. Med*. 2018;20(7):697-707. DOI 10.1038/gim.2017.170. Epub 2017. Nov. 2.

Fishman V.S., Salnikov P.A., Battulin N.R. Interpreting chromosomal rearrangements in the context of 3-dimentional genome organization: a practical guide for medical genetics. *Biochemistry* (*Mosc.*). 2018;83(4):393-401. DOI 10.1134/S0006297918040107.

Gridina M., Mozheiko E., Valeev E., Nazarenko L.P., Lopatkina M.E., Markova Z.G., Yablonskaya M.I., Voinova V.Y., Shilova N.V., Lebedev I.N., Fishman V. A cookbook for DNase Hi-C. *Epigenetics Chromatin*. 2021;14(1):15. DOI 10.1186/s13072-021-00389-5.

Hakim O., Resch W., Yamane A., Klein I., Kieffer-Kwon K.-R., Jankovic M., Oliveira T., Bothmer A., Voss T.C., Ansarah-Sobrinho C., Mathe E., Liang G., Cobell J., Nakahashi H., Robbiani D.F., Nussenzweig A., Hager G.L., Nussenzweig M.C., Casellas R. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature*. 2012;484(7392):69-74. DOI 10.1038/nature10909.

Harewood L., Kishore K., Eldridge M.D., Wingett S., Pearson D., Schoenfelder S., Collins V.P., Fraser P. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol*. 2017; 18(1):125. DOI 10.1186/s13059-017-1253-8.

Lieberman-Aiden E., van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., Dekker J. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293. DOI 10.1126/science.1181369.

Malentacchi F., Ciniselli C.M., Pazzagli M., Verderio P., Barraud L., Hartmann C.C., Pizzamiglio S., Weisbuch S., Wyrich R., Gelmini S. Influence of pre-analytical procedures on genomic DNA integrity in blood samples: the SPIDIA experience. *Clin. Chim. Acta*. 2015;440: 205-210. DOI 10.1016/j.cca.2014.12.004.

Mazur P. Freezing of living cells: mechanisms and implications. *Am. J. Physiol*. 1984;247(3):C125-C142. DOI 10.1152/ajpcell.1984.247.3.C125.

Melo U.S., Schöpflin R., Acuna-Hidalgo R., Mensah M.A., Fischer-Zirnsak B., Holtgrewe M., Klever M.-K., Türkmen S., Heinrich V., Pluym I.D., Matoso E., Bernardo de Sousa S., Louro P., Hülsemann W., Cohen M., Dufke A., Latos-Bieleńska A., Vingron M., Kalscheuer V., Quintero-Rivera F., Spielmann M., Mundlos S. Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. *Am. J. Hum. Genet*. 2020;106(6):872-884. DOI 10.1016/j.ajhg.2020.04.016.

Mozheiko E.A., Fishman V.S. Detection of point mutations and chromosomal translocations based on massive parallel sequencing of enriched 3C libraries. *Russ. J. Genet*. 2019;55(10):1273-1281. DOI 10.1134/S1022795419100089.

Narayanan S., O'Donovan M.R., Duthie S.J. Lysis of whole blood *in vitro* causes DNA strand breaks in human lymphocytes. *Mutagenesis*. 2001;16(6):455-459. DOI 10.1093/mutage/16.6.455.

Nederhand R.J., Droog S., Kluft C., Simoons M.L., De Maat M.P.M., Investigators of the EUROPA trial. Logistics and quality control for DNA sampling in large multicenter studies. *J. Thromb. Haemost*. 2003;1(5):987-991. DOI 10.1046/j.1538-7836.2003.00216.x.

Palmirotta R., Ludovici G., De Marchis M.L., Savonarola A., Leone B., Spila A., De Angelis F., Della Morte D., Ferroni P., Guadagni F. Preanalytical procedures for DNA studies: the experience of the interinstitutional multidisciplinary BioBank (BioBIM). *Biopreserv. Biobank*. 2011;9(1):35-45. DOI 10.1089/bio.2010.0027.

Peng L., Wang S., Yin S., Li C., Li Z., Wang S., Liu Q. Autophosphorylation of H2AX in a cell-specific frozen dependent way. *Cryobiology*. 2008;57(2):175-177. DOI 10.1016/j.cryobiol.2008.06.005.

Permenter J., Ishwar A., Rounsavall A., Smith M., Faske J., Sailey C.J., Alfaro M.P. Quantitative analysis of genomic DNA degradation in whole blood under various storage conditions for molecular diagnostic testing. *Mol. Cell. Probes*. 2015;29(6):449-453. DOI 10.1016/j.mcp.2015.07.002.

Rao S.S.P., Huntle M.H., Durand N.C., Stamenova E.K., Bochkov I.D., Robinson J.T., Sanborn A.L., Machol I., Omer A.D., Lander E.S., Aiden E.L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665-1680. DOI 10.1016/j.cell.2014.11.021.

Ross K.S., Haites N.E., Kelly K.F. Repeated freezing and thawing of peripheral blood and DNA in suspension: effects on DNA yield and integrity. *J. Med. Genet*. 1990;27(9):569-570. DOI 10.1136/jmg.27.9.569.

Schröder C., Steimer W. gDNA extraction yield and methylation status of blood samples are affected by long-term storage conditions. *PLoS One*. 2018;13(2):e0192414. DOI 10.1371/journal.pone.0192414.

ORCID ID

M.M. Gridina orcid.org/0000-0002-7972-5949
E. Vesna orcid.org/0000-0003-3480-3963
M.E. Minzhenkova orcid.org/0000-0001-5458-0408
N.V. Shilova orcid.org/0000-0002-0641-1084
I.N. Lebedev orcid.org/0000-0002-0482-8046
V.S. Fishman orcid.org/0000-0002-5573-3100

Acknowledgements. The study was supported by the Russian Science Foundation, project No. 22-24-00190. Samples from the biocollection "Biobank of the population of Northern Eurasia" of the Research Institute of Medical Genetics, Tomsk NRMC were used in the study.
We thank Alexandra Yan and Artem Nurislamov.

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

Received November 8, 2022. Revised December 29, 2022. Accepted December 30, 2022.

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".

Регистрационное свидетельство ПИ № ФС77-45870 выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 20 июля 2011 г.

«Вавиловский журнал генетики и селекции» включен 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, Российский индекс научного цитирования, ВИНИТИ, базы данных Emerging Sources Citation Index (Web of Science), Zoological Record (Web of Science), Scopus, PubMed Central, Ebsco, DOAJ, Ulrich's Periodicals Directory, Google Scholar, Russian Science Citation Index на платформе Web of Science, каталог научных ресурсов открытого доступа ROAD.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте ИЦиГ СО РАН, https://sites.icgbio.ru/vogis/
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/

✉ e-mail: 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.

Подписано в печать 20.02.2023. Выход в свет 28.02.2023. Формат 60 × 84 $^{1}/_{8}$. Усл. печ. л. 10.23.
Уч.-изд. л. 12.2. Тираж 150 экз. (1-й завод 1–45 экз.) Заказ № 27. Цена свободная.

Отпечатано в Сибирском отделении РАН, Морской проспект, 2, Новосибирск, 630090.