

===== TRANSLATIONS OF PUBLISHED ARTICLES =====

A New Type of Tandem Repeats in the Genome of *Mus Musculus*: GC-Rich Megasatellites

Nafisa Nazipova, Ruslan Tetuev

Institute of Mathematical Problems of Biology RAS, Keldysh's Institute of Applied Mathematics of the Russian Academy of Sciences, Pushchino; Russia

Abstract. With the advance of the third-generation sequencing (TGS) technologies, the technical difficulties in sequencing the repetitive regions of eukaryotic genomes were overcome and, as a result, T2T assemblies of a number of model genomes, including that of the house mouse *Mus musculus*, were completed. This technological breakthrough opened up new possibilities for investigating the structural and functional organization of complex genomes. In our studies, we verified the tandem repeats that were previously identified in the mouse genome and discovered new regions of periodicity, which we called GC-rich megasatellites. This paper describes two megasatellites found in the intergenic regions of chromosome 6 of the mouse genome; each of them has its own characteristics. These extended tandem repeats have an average GC content of 56 % and contain multiple micro- and minisatellite tracks, covering more than a third of each megasatellite copy. In total, there are 83 copies of the first and 71 copies of the second repeat regions, with their average lengths being 4320 and 2440 bp, respectively. A characteristic feature of both tandem repeats is the presence of extended conserved regions alternated with islands of variability (micro- and minisatellite tracks). In the first megasatellite, the islands of variability are formed by the tracks (ACCCC)_n, (AAACG)_n, (AAAC)_n, (GTCT)_n, (GT)_n, (CTTC)_n, (GAGAAG)_n and (TCC)_n; in the second one there are three islands of variability formed by homopurine-homopyrimidine tracks (CT)_n, (CA)_n and (CT)_n. The first megasatellite contains the coding sequence of a proline-rich protein at the 3'-end of each copy; its tandem copies are heterogeneous in the GC-content. Namely, 10% of the pattern at the 5' end and 20% at the 3' end contain more than 65% of GC, whereas the inner 70% of the pattern has ~50% GC. The second megasatellite is more GC-rich than the first one. At the same time, its copies are more uniform in length, GC content, and oligonucleotide-track coverage.

Key words: *third-generation sequencing, Mus musculus genome, chromosome 6, extended tandem repeats, GC-rich megasatellites, microsatellite tracks, minisatellite tracks*

INTRODUCTION

Direct tandem repeats are multiple copies of the same pattern that follow each other, joining in an orderly, head-to-tail manner. The presence of such periodic regions is a common feature of prokaryotic and eukaryotic genomes. The human genome, for example, contains over a million of tandem repeats, many of which are highly polymorphic in regard to the composition and the number of their copies. Tandem repeats can be found in different parts of the genome: intergenic regions, as well as noncoding and coding regions of various genes [1–3]. Moreover, they are often located near or within gene sequences and, thereby, can influence the structure and function of DNA, RNA and proteins – via specific mechanisms, which entail various molecular and cellular consequences [4]. There are data that, in the vicinity of tandem repeats,

the expression of genes can be regulated through changes in the number of the repeat copies [5]. Overall, tandem repeats are highly mutable and can be located in exons, introns and intergenic regions, providing opportunities to modulate gene expression, as well as the structure and function of RNA and proteins [6]. Long tandem repeats are often involved in various disorders, including autism spectrum disorders and cancer. The ways, by which tandem repeats directly or indirectly influence the genome, are discussed in detail in [7].

The presence of repeats greatly complicates reference-free genome assembly. They are the main cause of errors in contigs obtained during *de novo* assembly of the genome [8]. Repeats also introduce ambiguity when the sequenced data are read in the procedure of multiple sequence aligning, this hindering the correction of sequencing errors, identification of single-nucleotide polymorphism, detection of variants and analysis of gene expression [9, 10].

The genome regions occupied by tandem repeats are called satellite DNA. On the basis of motif length, satellite DNA is divided into three classes: microsatellites or short tandem repeats (STR; motif length, 1–5 bp); minisatellites (motif length, 6–79 bp) and satellites (motif length, from 80 to several thousand bp). Microsatellites can be found in exons, introns, 5'- and 3'-untranslated regions of genes (5'-UTR, 3'-UTR) and promoters. Minisatellites are represented, for example, by telomere repeats. The class of satellites contains a great variety of satellite DNA divided into different subgroups: segmental duplications, various centromeric and pericentromeric satellites, macrosatellites (motif length, 100–1000 bp) and megasatellites (motif length, >1000 bp).

All tandem repeats, especially STRs, are characterized by high instability in terms of their length, composition and number of copies. As a rule, the mutation rates in tandem repeats are 10–100,000 times higher than in other parts of the genome [11]. The origin and evolution of repetitive DNA sequences is usually explained by strand mispairing, a mutational process which occurs during DNA replication [12]. Unstable repeats are found in almost 20% of eukaryotic genes and promoters; they determine phenotypic (functional) variability of the cell surface and extracellular proteins and can be involved in the development of pathological conditions.

Functions of tandem repeats

The data obtained in recent years indicate a significant influence of periodic regions on the coding part of the genome and, correspondingly, their active involvement in the regulation of various cellular processes. Tandem repeats are responsible for genetic variability: the number and length of their copies frequently change and, thus, create a genetic diversity, which can contribute to the emergence of new traits. Furthermore, tandem repeats are involved in gene regulation. The repeats located in the regulatory regions can influence gene expression and initiate adaptive changes in response to environmental pressures.

When located within or around specialized regions of chromosomes (e.g., centromeres, telomeres, and subtelomeres), tandem repeats can play a crucial role in maintaining the structural stability of genetic material throughout the cell cycle [13]. During cell cycle, the structure of genetic material is primarily stabilized by a mechanism known as replication fork stabilization. The periodic structure of tandem repeats provides a stable template for DNA polymerase binding and replication initiation. This stability prevents stalling or collapse of replication forks, ensuring accurate and complete DNA replication.

Telomeres, specialized tandem repeats located at the ends of chromosomes, also play a crucial role in maintaining genomic stability. Telomeres protect chromosome ends from degradation, fusion and recognition of DNA breaks. During each round of DNA replication, the basic DNA replication machinery struggles to fully replicate the ends of linear chromosomes. Telomeres are protected by protein complexes, such as shelterin, which includes TRF1, TRF2, POT1 and other proteins, indirectly interacting with telomeres [14]. Owing to this protection, chromosome ends are distinguished from random DNA breaks, which prevents undesirable activation of reparative mechanisms at telomeres. Thus, in combination with the activity of the

enzyme telomerase, telomere tandem repeats ensure the integrity and stability of the genome during successive cell divisions.

Centromeres are chromosomal domains responsible for the accurate transfer of genetic material during cell division. Characterized by highly repetitive DNA regions and associated kinetochore proteins, they are essential for the attachment of microtubules to chromosomes during mitosis [15]. A crucial component of centromeres is an array of tandem repeats, which plays a key role in stabilization of human chromosomes. Variations in these satellites can affect the centromere function [16].

Furthermore, 5' UTRs and 3' UTRs contain various regulatory elements involved in the post-transcriptional gene regulation. They affect mRNA stability, as well as localization and efficiency of translation [17]. UTR-located microsatellites can contribute to gene regulation in many ways; for example, STRs can influence the stability of mRNA molecules. Changes in the STR length can affect folding of UTRs, altering their interactions with RNA-binding proteins, which leads to either degradation or stabilization of mRNA. UTRs can also modulate translation initiation and efficiency. STRs located in 5' UTRs can affect ribosome binding and start codon recognition, changing the rates of translation and protein production. Variations in UTR-located STRs have also been reported to be associated with complex diseases [17].

Tandem repeats can be transcribed into RNA molecules by RNA polymerases [18]. These RNA molecules retain the repetitive nature of the original DNA sequences and can fold into various secondary structures – owing to the intramolecular base pairing within the repeat sequence. RNA molecules derived from tandem repeats can function as non-coding RNAs. They can serve, for example, as scaffolds for the assembly of ribonucleoprotein complexes or regulate gene expression through interactions with RNA-binding proteins or microRNAs. RNAs transcribed from tandem repeats can participate in such regulatory mechanisms as RNA interference, in which RNA pairs with a target mRNA, modulating its stability or translation. RNA molecules synthesized from tandem repeats can also influence cellular processes by sequestering RNA-binding proteins or acting as traps for regulatory factors.

The process of DNA replication often slips in tandem repeats, which leads to the expansion of the repeat region and subsequent gene duplication. As a result, extra gene copies can be created within the same genomic region. Duplicated genes are often subject to variations, such as point mutations or insertions/deletions. Their accumulation over time leads to divergence in the gene sequences and functions. In the process of functional divergence, the duplicated genes can acquire new functions or adopt differential expression patterns. Thus, redundant gene copies can contribute to the expansion and diversity of their gene families, providing evolutionary opportunities for gene innovations and adaptations to environmental or physiological factors. For example, the genes encoding ribosomal RNA (rRNA) are present in the human genome as multiple tandem copies. These rRNA gene repeats facilitate simultaneous synthesis of large amounts of rRNA – to meet the constant need of a cell for the production of ribosomes [19].

Tandem repeats located in the regulatory regions, such as promoters and enhancers, can influence gene expression by affecting the binding of transcription factors. The presence of tandem repeats in these regions can change the three-dimensional structure of chromatin, altering the accessibility of regulatory elements and activation of the transcriptional machinery. Variations in the length and composition of tandem repeats can affect the affinity of transcription factors for binding sites and, thus, determine the levels of gene expression.

Tandem repeats are also vulnerable targets for epigenetic modifications, such as DNA methylation and histone modifications. The length and composition of the tandem-repeat sequence can influence the degree of epigenetic regulation. For example, methylation of a tandem repeat can lead to the formation of repressive chromatin and transcriptional silencing. These epigenetic modifications can have a profound effect on the patterns of gene expression, contributing to the regulation of various cellular processes.

Tandem repeats within exons or introns can influence alternative splicing, a process by which multiple mRNA isoforms are formed from a single gene. Variations in the length of a tandem repeat can influence the process of splicing by either altering the stability of RNA secondary structures or creating binding sites for splicing factors. This can lead to the inclusion or exclusion of specific exons and give rise to diverse protein isoforms with different functions or regulatory properties.

Particular attention to the study of microsatellites has been drawn by the works that elucidated the nature of anticipation, a clinical phenomenon defined as a gradual onset of a genetic disease in the first generation and an earlier manifestation and increased severity of its symptoms in the subsequent generations. Anticipation was first described for myotonic dystrophy; by now, it has been revealed in many monogenic diseases [20]. The genetic mechanism of anticipation is shown to be the expansion of trinucleotide repeats.

When the size of some trinucleotide repeats exceeds a certain threshold, it can lead to genomic instability and pathological consequences. These extended trinucleotide repeats have a tendency to further expand and accumulate in subsequent generations, leading to a dynamic and progressive increase in the repeat length. Extended trinucleotide repeats can disrupt gene function and impair protein production, resulting in alterations in the protein structure and disruption of cellular processes.

The expansion of trinucleotide repeats is responsible for various types of dysfunction in the human body. These dysfunctions include Huntington's disease, fragile X syndrome and several forms of spinocerebellar ataxia. The Lynch syndrome, an autosomal dominant disorder which increases the risk of various types of cancer (colorectal cancer; endometrial adenocarcinoma; tumors of the small intestine, stomach, ureter, renal pelvis, ovaries, brain, and prostate), is also associated with the expansion of trinucleotide repeats. Furthermore, amyotrophic lateral sclerosis, a rapidly progressive and invariably fatal motor-neuron disease, has been identified as one of the neurodegenerative diseases in which microsatellite instability plays a significant role [21].

Recent studies indicate that polymorphic tandem repeats can regulate gene expression in healthy individuals [22]. In the human genome, tandem repeats are frequently found in the genes that control body morphology [23, 24]. Instability of tandem repeats can lead to decreased gene expression, increased disease incidence, and enhanced tumor aggression. The link between tandem-repeat instability and certain pathologies (cancer, autism and neurological disorders) is discussed in [25].

In comparison with syntenic blocks, evolutionary breakpoints in the human genome contain more tandem repeats. The tandem repeats located in the regions of evolutionary breakpoints tend to promote and accelerate the evolution of gene expression; they generate sufficient variability to ensure rapid evolution and adaptation of organisms [26]. Recent studies have shown that STR variations in immune genes can shape the immune response and facilitate adaptation to diverse environmental conditions [27]. The tandem repeats located in the regulatory regions can also contribute to evolutionary adaptation.

GC-Rich Tandem Repeats

As the most dynamic class of genomic variation, tandem repeats have the highest mutation rate among all the genomic elements and, correspondingly, are characterized by a high polymorphism in human populations. Instability of tandem repeats grows with their length, and their expansion in the successive generations is a well-known phenomenon, responsible for at least 50 monogenic human diseases. As many as a third of them are associated with GC-rich trinucleotide repeats [28].

Initially, only trinucleotides were thought to be able to expand in human hereditary diseases, but the “rule of three” was overturned in 1997 [29], when it was discovered that the expansion of the dodecamer $((C)_4G(C)_4GC)_n$ was associated with a monogenic form of epilepsy [30]. Later, expansions of the tetranucleotide repeat $(CCTG)_n$ [31] and the pentanucleotide repeat

(ATTCT)_n [32] were found to be the cause of human neurological disorders. It became evident that it was not only trinucleotide repeats that the expansion was associated with; rather than having a triplet motif, it was about the ability of the sequence to form complex secondary structures and interfere with DNA replication, repair or recombination [33].

Since micro- and minisatellites have a repetitive nature and are characterized by a high level of the GC-content bias, they were assumed to form local secondary structures capable of playing an important role in the mutational process. As shown in the earlier studies of dinucleotide repeats of homopyrimidine-homopurine (TC)_n(GA)_n tracts, they can form triple helices, which are able to block DNA synthesis *in vitro* [34]. Such tracts are scattered all over the human genome [35].

Self-association of oligonucleotides containing homopurine triplet repeats is a known phenomenon described in the literature [33]. Homopurine oligonucleotides can form a few different structures: parallel-stranded duplex [36]; intramolecular antiparallel-stranded duplex [37]; G4-quadruplex, a product of dimerization of the two hairpin duplexes [38].

The abundance of GC-elements in the genome majorly depends on such genomic parameters, as gene density, proteome size, distribution of repetitive elements and methylation patterns. It also strongly correlates with recombination rates and gene expression levels [39]. Thus, the biological significance of genomic GC-elements raises important questions about their evolutionary origin and role.

In comparison with other DNA sequences, the GC-rich ones are more thermally stable. This is because, in contrast to AT, the GC pair has an extra hydrogen bond. There is a hypothesis that the genome may have evolved towards a higher GC content under the conditions of high temperatures [39]. In the light of this hypothesis, the spread of GC-rich isochores over the genomic sequences of warm-blooded vertebrates is considered an adaptation to their body temperature [40]. The hypothesis of thermal adaptation is confirmed, to some degree, by the data that the GC-content of structural RNA in prokaryotes positively correlates with the optimal temperature of their growth [41].

The high-throughput third-generation sequencing technologies have demonstrated how variable the architecture of genomes can be. Some species have small, densely clustered genes with a high AT content, while others have genomes containing GC-rich repeats. Characterization of this variability and identification of organisms exhibiting extreme patterns are crucial for understanding the processes that determine genome function and evolution. In this work, we identify extended regions of mouse genome with GC-bias and analyze them for periodicity.

MATERIALS AND METHODS

In our previous work, we developed a method, which allows one to search for extended fuzzy tandem repeats in DNA sequences by assessing the oscillations of the GC-content curves [42]. The method is based on multiple scanning of a sequence with a sliding window of varying length ($W = 256, 512, 1024 \dots$ bp), which is doubled after each pass up to the maximal expected length of the periodicity pattern. The window of the length W shifts along the sequence with a step d . If the total number of letters G and C in the window sliding along the sequence $S = x_1x_2 \dots x_i \dots x_{i+W-1} \dots x_N$ is defined as F_w^{GC} , then the percentage of the GC-content within the window of the length W at the position x_i will be defined as

$$f_{\%GC}(x_i) = 100\% \cdot F_w^{GC}(x_i) / W, \quad i = 1, \dots, N - W + 1.$$

The shift of the window was taken to be equal to 1. For each window, the percentage of the GC-content was calculated as $f_w^{GC}(x_i)$, and successive values of this characteristic were defined as a GC-profile. The degree of the GC-profile oscillation was quantitatively assessed using a

mathematical apparatus of the generalized spectral-analytical method for analyzing values of a discrete function [43].

Earlier, we used this method to detect tandem repeats in the genome of *Mus musculus* (reference assembly GRCm37.1) [44, 45]. Back then, the mouse genome contained numerous regions of uncertainty, which interrupted the identified periodicities. The latest reference assembly GRCm39 contains approximately 73.5 million bp of unidentified DNA scattered all over the genome [46]. Now, after a T2T (telomere-to-telomere) assembly of the *Mus musculus* genome has been completed (assembly T2T_mhaESC_v1.0, Genbank ID GCA_050437135.1 from May 19, 2025 [46]), we have the opportunity to revise our search for extended tandem repeats based on the oscillations of the nucleotide-content profiles.

The method for revealing tandem repeats on the basis of GC-profile oscillations was optimized for the analysis of large genomes. We applied the method of bitwise shifting [47], which quickly packs a 4-letter alphabetic DNA sequence into a bit string, where each nucleotide is assigned two bits: A $\rightarrow$ 00, C $\rightarrow$ 01, T $\rightarrow$ 10, G $\rightarrow$ 11.

We also developed a method for calculating the GC-content in a window F_w^{GC} shifting along a numerical DNA sequence. The calculations are based on the lookup-table method, which first creates a table of the GC-content of all the possible combinations of four nucleotides ($256 = 4^4$). For example, the 27th position in the table would correspond to the bit combination $00011011_2 = 27_{10}$, (or the ACTG fragment, which has a 50 % GC-content; lookup [27] = 2). In this case, the window is sliding along the sequence with a step $d = 4$.

The combination of the two methods (bitwise packing and table lookup) makes it possible to avoid situations, when we have to compromise between the analysis speed and data compression. On the contrary, the performance tests have shown that it is possible to reduce the memory usage by a factor of 4 and, at the same time, speed up the analysis by 75%.

The degree of oscillation of the discrete GC-percentage function, $f_{\%GC}$, was estimated on the basis of a quick calculation of its second derivative [48]. Any repeats in a segment of the genome of the length L will manifest themselves as quasi-harmonic oscillations of the GC-profile with the same frequency and phase (but the opposite amplitude sign) as oscillations of the 2nd derivative of the function. The proximity of the GC-percentage function to the harmonic function can be estimated by the correlation between $f_{\%GC}$ and its second derivative, where L is the width of the sliding window in which the degree of the GC-profile oscillation is estimated in a given DNA region, starting with x_i :

$$R(x_i) = \frac{\sum_{k=0}^{L-1} (f_{\%GC}(x_i + k) - \bar{f}_{\%GC})(f_{\%GC}''(x_i + k) - \bar{f}_{\%GC}'')}{\sum_{k=0}^{L-1} (f_{\%GC}(x_i + k) - \bar{f}_{\%GC})^2 (f_{\%GC}''(x_i + k) - \bar{f}_{\%GC}'')^2} \approx -1,$$

where $\bar{f}_{\%GC}$ and $\bar{f}_{\%GC}''$ are the mean values of the GC-percentage function and its second derivative.

To automatically determine the boundaries of the area with oscillations, we used the percentage correlation estimate:

$$h = \frac{1 - R(x_i)}{2} \times 100 \, \%.$$

There are many numerical methods for a quick approximate calculation of derivatives of a discrete-argument function. The authors of the method that we used here expanded the function into an incomplete series of orthogonal Chebyshev-Gram polynomials, followed by a quick intraspectral calculation of two derivatives and the correlation between their spectra [49].

Figure 1 shows an example of how this method can be applied for determining the areas of pronounced oscillations of $f_{\%GC}$.

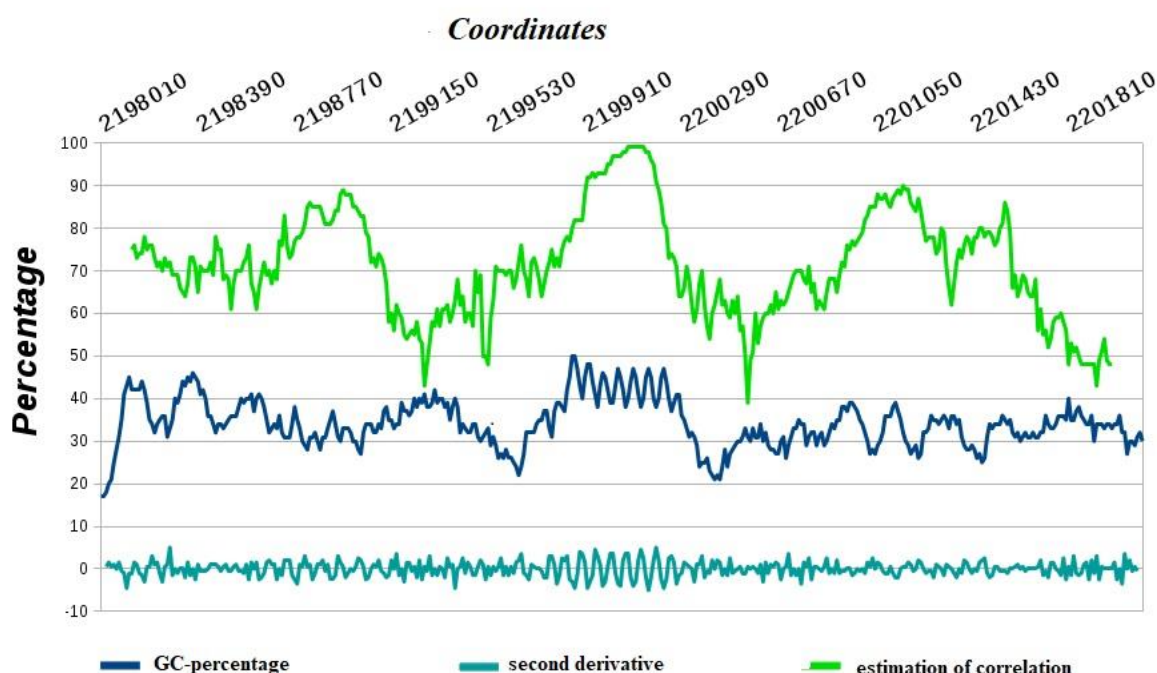


Fig. 1. The percentage GC-content function calculated for a sample DNA sequence (top line) and the estimation of the degree of its oscillations (bottom line).

At the time this study was conducted, the Genbank T2T assembly [46] had not been annotated yet, so we first identified repeats in the T2T assembly and then searched the longest pattern sequence in the reference assembly to annotate the neighborhoods of the identified megasatellites. For this purpose, we used the mouse genome reference assembly GRCm39 (GCF_000001635.27), chromosome 6 (NC_000072.7).

BLASTN [50] was used to refine the coordinates of the megasatellite copies, and multiple alignment was performed with Clustal Ω (Clustal Omega; www.ebi.ac.uk/jdispatcher/msa/clustalo) [51]. Clustal Ω uses seeded guide trees and hidden Markov models to align more than three sequences.

To detect micro- and minisatellite tracks in the copies of megasatellites, we used our own program TrackFinder. Each megasatellite copy was scanned for oligonucleotide tracks, the lengths of which ranged from 1 to 30 bp. In the course of scanning, the program registered the oligonucleotide tandem repeats with the tested length and masked them, starting with the maximal length value. During the next pass through the sequence (with the pattern length reduced by 1), the masked fragments were excluded from the analysis. For the even values of the tested length, possible symmetry of the region was taken into account: e.g., for a sequence of the test length 6, the region ATGATGATGATG was registered not as two copies of the 6-mer ATGATG, but as four copies of the 3-mer ATG. The tracks of mono- and dinucleotides were registered only when their repeat multiplicity was at least 4 and 3, respectively. For oligonucleotides with a length of ≥ 3 bp, detection of two exact copies would be sufficient for registration.

RESULTS

Examination of the T2T *Mus musculus* assembly has revealed several extended fuzzy tandem repeats with the copy lengths greater than 1000 bp on different chromosomes. All of them are characterized by elevated GC-content. In this paper, we shall only analyze repetitive

regions of chromosome 6. In addition to the known repeat, which was found earlier in the GRCm37.1 assembly, the chromosome turned out to have another one, with a pattern twice as long (Fig. 2).

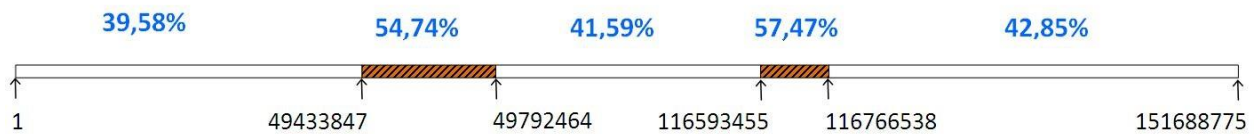


Fig. 2. The average contents of GCs in the two periodic regions of *Mus musculus* chromosome 6 (colored and hatched) with the length of 358937 and 173084 bp are significantly higher than in the remaining parts of the chromosome.

Thus, the analysis of chromosome 6 of the *Mus musculus* T2T-assembly reveals two extended regions with latent periodicity (358937 and 173084 bp in length). Both of them differ from the rest of the chromosome by increased GC-content (Fig. 2).

Megasatellite site 49433847–49792783

The first repetitive region, 358,937 bp in length, contains 83 copies of GC-rich tandem repeats, with the length of copies ranging from 4239 to 4418 bp (average pattern length, 4318 bp). At its 5' and 3' ends, the region contains fragments of its pattern sequence (~5% of the pattern length; 242 and 317 bp, respectively). These pieces of the pattern sequence contain 63.07 and 65.72% of GC pairs. The full-length copies are located at the coordinates 49434088–49792464.

Tandem copies were found to be GC-rich at the ends. With the average GC level of the copies ~55%, the 5' end of the periodic pattern (~450 bp; 10% of the copy length) and the 3' end (900 bp; 20% of the copy length), have their GC percentage level more than 65%. The inner part of the pattern (70% of its length) accounts for ~50% of the GC content of the copies.

The GC-rich 3' end of each of the 83 copies contains a copy of a putative 867-bp gene (Genbank ID NC_000072.7), which in the GRCm39 assembly was called *proline rich protein 36-like*. The 8053-bp *proline rich protein 36 gene* (*Prr36*, NC_000074.7) is located on chromosome 8 and contains seven exons. Our 867-bp gene's BLASTN hits fall on the last exon of *Prr36*.

To study the pathways of the repeat copies evolution resulting from the expansion of short oligonucleotide tracks, we used TrackFinder, which identifies short tandem repeats within each megasatellite copy. The internal repeats identified with TrackFinder, along with their coordinates and multiplicities, are presented in Table S1 (supplementary materials, file [CM115055_1_reps.xlsx](#)). Analyzing these data, one should keep in mind that some differences in the number of copies between mono- and dinucleotide tracks result from the rule of minimal multiplicity (4 and 3 for mono- and dinucleotide tracks, respectively). For example, three copies of the dinucleotide AC (ACACAC) were registered, whereas ACAC tracks were not taken into account. Thus, we avoided registering numerous random duplications – yet, at the same time, lost short repeats at key points of variability. There is no consensus in the literature on the minimal number of copies of mono- and dinucleotide repeats, at which they can be considered tracks.

Table 1 summarizes the results of analysis of the detected megasatellite repeat of chromosome 6 of the *Mus musculus* T2T-assembly (Genbank ID CM115055_1): coordinates and lengths of all its copies; their GC contents, and the length of coverage by micro- and minisatellites in absolute (total number of nucleotides) and relative (percentage of this sum of the copy length) units. Table S1 (supplementary materials; file [CM 115055_1_reps.xlsx](#)) gives coordinates, multiplicities and patterns of the detected micro- and minisatellite tracks in each copy of the tandem repeat. Table rows corresponding to the tracks located in the copies of the

putative protein are highlighted in terracotta. Table 2 shows entry coordinates and other characteristics of the copies of the putative gene.

Table 1. Characteristics of copies of the first megasatellite

#	Coordinates	Copy length	% GC	Length of coverage by short TRs, bp	Coverage length, %
1	49434088-49438425	4338	55.03	1562	36.01
2	49438426-49442762	4337	54.99	1537	35.44
3	49442763-49447094	4332	55.03	1524	35.18
4	49447095-49451432	4338	54.98	1540	35.50
5	49451433-49455825	4393	54.88	1623	36.95
6	49455826-49460178	4353	54.7	1553	35.68
7	49460179-49464537	4359	54.81	1540	35.33
8	49464538-49468955	4418	54.73	1618	36.62
9	49468956-49473371	4416	54.82	1625	36.80
10	49473372-49477705	4334	54.85	1550	35.76
11	49477706-49482029	4324	54.81	1543	35.68
12	49482030-49486346	4317	54.81	1535	35.56
13	49486347-49490718	4372	54.83	1582	36.18
14	49490719-49495018	4300	54.51	1520	35.35
15	49495019-49499370	4352	54.8	1539	35.36
16	49499371-49503668	4298	54.75	1479	34.41
17	49503669-49508073	4405	54.48	1592	36.14
18	49508074-49512440	4367	54.57	1544	35.36
19	49512441-49516796	4356	54.61	1527	35.06
20	49516797-49521184	4388	54.63	1576	35.92
21	49521185-49525507	4323	54.68	1488	34.42
22	49525508-49529799	4292	54.87	1491	34.74
23	49529800-49534117	4318	54.79	1515	35.09
24	49534118-49538483	4366	54.7	1569	35.94
25	49538484-49542852	4369	54.68	1572	35.98
26	49542853-49547147	4295	54.58	1498	34.88
27	49547148-49551486	4339	54.78	1539	35.47
28	49551487-49555772	4286	54.55	1489	34.74
29	49555773-49560045	4273	54.53	1470	34.40
30	49560046-49564363	4318	54.61	1515	35.09
31	49564364-49568667	4304	54.72	1496	34.76
32	49568668-49572925	4258	54.58	1452	34.10
33	49572926-49577251	4326	54.6	1517	35.07
34	49577252-49581564	4313	54.63	1503	34.85
35	49581565-49585852	4288	54.73	1480	34.51
36	49585853-49590153	4301	54.75	1493	34.71
37	49590154-49594417	4264	54.64	1440	33.77
38	49594418-49598721	4304	54.79	1484	34.48
39	49598722-49602980	4259	54.78	1453	34.12
40	49602981-49607233	4253	54.76	1430	33.62
41	49607234-49611536	4303	54.82	1477	34.32
42	49611537-49615876	4340	54.68	1526	35.16
43	49615877-49620211	4335	54.65	1522	35.11
44	49620212-49624527	4316	54.7	1515	35.10
45	49624528-49628832	4305	54.7	1500	34.84
46	49628833-49633088	4256	54.79	1448	34.02
47	49633089-49637407	4319	54.78	1514	35.05
48	49637408-49641754	4347	54.73	1535	35.31
49	49641755-49646073	4319	54.8	1502	34.78
50	49646074-49650373	4300	54.72	1515	35.23
51	49650374-49654659	4286	54.71	1489	34.74

#	Coordinates	Copy length	% GC	Length of coverage by short TRs, bp	Coverage length, %
52	49654660-49658997	4338	54.75	1522	35.09
53	49658998-49663237	4240	54.72	1443	34.03
54	49663238-49667534	4297	54.69	1493	34.75
55	49667535-49671801	4267	54.58	1496	35.06
56	49671802-49676099	4298	54.77	1518	35.32
57	49676100-49680417	4318	54.72	1521	35.22
58	49680418-49684729	4312	54.73	1520	35.25
59	49684730-49689061	4332	54.85	1507	34.79
60	49689062-49693338	4277	54.76	1454	34.00
61	49693339-49697680	4342	54.86	1516	34.91
62	49697681-49701948	4268	54.8	1450	33.97
63	49701949-49706240	4292	54.61	1500	34.95
64	49706241-49710578	4338	54.7	1546	35.64
65	49710579-49714916	4338	54.7	1517	34.97
66	49714917-49719186	4270	54.78	1448	33.91
67	49719187-49723425	4239	54.75	1435	33.85
68	49723426-49727693	4268	54.69	1496	35.05
69	49727694-49731968	4275	54.62	1476	34.53
70	49731969-49736232	4264	54.97	1465	34.36
71	49736233-49740580	4348	54.88	1527	35.12
72	49740581-49744885	4305	54.94	1495	34.73
73	49744886-49749208	4323	54.75	1527	35.32
74	49749209-49753511	4303	54.87	1491	34.65
75	49753512-49757829	4318	54.93	1520	35.20
76	49757830-49762121	4292	54.89	1490	34.72
77	49762122-49766454	4333	54.93	1531	35.33
78	49766455-49770775	4321	54.94	1519	35.15
79	49770776-49775071	4296	54.77	1495	34.80
80	49775072-49779329	4258	54.56	1479	34.73
81	49779330-49783677	4348	54.39	1614	37.12
82	49783678-49788062	4385	54.44	1604	36.58
83	49788063-49792465	4403	54.42	1619	36.77

Table 2. Coordinates of copies of the putative proline-rich protein found within the copies of the large tandem repeat

No.	Coordinates	Copy length	% GC
1	49437542-49438420	879	65.42
2	49441891-49442757	867	65.4
3	49446229-49447089	861	65.39
4	49450564-49451427	864	65.28
5	49454943-49455820	878	65.49
6	49459273-49460173	901	64.93
7	49463641-49464532	892	65.25
8	49468041-49468950	910	65.27
9	49472451-49473366	916	65.28
10	49476824-49477700	877	65.45
11	49481160-49482024	865	65.43
12	49485462-49486341	880	65.45
13	49489819-49490713	895	65.14
14	49494161-49495013	853	65.42
15	49498504-49499365	862	65.66
16	49502811-49503663	853	65.53

No.	Coordinates	Copy length	% GC
17	49507195-49508068	874	65.45
18	49511562-49512435	874	65.45
19	49515915-49516791	877	65.45
20	49520297-49521179	883	65.46
21	49524623-49525502	880	65.45
22	49528936-49529794	859	65.54
23	49533239-49534112	874	65.56
24	49537604-49538478	875	65.6
25	49541968-49542847	880	65.34
26	49546299-49547142	844	65.4
27	49550599-49551481	883	65.57
28	49554936-49555767	832	65.14
29	49559209-49560040	832	65.02
30	49563484-49564358	875	65.49
31	49567792-49568662	871	65.56
32	49572089-49572920	832	65.02
33	49576376-49577246	871	65.67
34	49580689-49581559	871	65.56
35	49584977-49585847	871	65.56
36	49589263-49590148	886	65.58
37	49593575-49594412	838	65.27
38	49597864-49598716	853	65.53
39	49602115-49602975	861	65.62
40	49606373-49607228	856	65.65
41	49610664-49611531	868	65.67
42	49614995-49615871	877	65.11
43	49619333-49620206	874	65.22
44	49623655-49624522	868	65.44
45	49627951-49628827	877	65.45
46	49632222-49633083	862	65.66
47	49636535-49637402	868	65.67
48	49640885-49641749	865	65.66
49	49645207-49646068	862	65.66
50	49649507-49650368	862	65.66
51	49653799-49654654	856	65.65
52	49658128-49658992	865	65.55
53	49662398-49663232	835	65.63
54	49666668-49667529	862	65.66
55	49670962-49671796	835	65.63
56	49675229-49676094	866	65.59
57	49679545-49680412	868	65.67
58	49683854-49684724	871	65.67
59	49688189-49689056	868	65.67
60	49692499-49693333	835	65.63
61	49696799-49697675	877	65.68
62	49701109-49701943	835	65.63
63	49705401-49706235	835	65.63
64	49709703-49710573	871	65.67
65	49714070-49714911	842	65.68
66	49718347-49719181	835	65.63
67	49722586-49723420	835	65.63
68	49726854-49727688	835	65.63
69	49731129-49731963	835	65.63
70	49735351-49736227	877	65.68
71	49739699-49740575	877	65.68
72	49744001-49744880	880	65.68

No.	Coordinates	Copy length	% GC
73	49748333-49749203	871	65.67
74	49752636-49753506	871	65.67
75	49756951-49757824	874	65.68
76	49761252-49762116	865	65.66
77	49765564-49766449	886	65.69
78	49769885-49770770	886	65.69
79	49774193-49775066	874	65.56
80	49778525-49779324	800	65.25
81	49782811-49783672	862	65.78
82	49787196-49788057	862	65.78
83	49791590-49792460	871	65.79

Located in the 3'-neighborhood of the first tandem repeat (coordinates: 49816011–49832827) is *Pdia4* (a gene of *disulfide isomerase-associated protein 4*), which contains a microRNA precursor.

A FASTA file with the sequences of tandem copies of the region 49434088–49792464 ([CM115055_1_reps.fasta](#)), as well as multiple alignment generated by ClustalΩ ([CM115055_1_reps.align](#)) can be found in the supplementary materials.

Megasatellite region 116593455–116766538

The second megasatellite region found on chromosome 6 has already been known to us [44, 45]. The T2T-updated version of the region has a length of 173084 bp and contains 71 pattern copies with lengths ranging from 2333 to 2515 bp (average length, 2438 bp). In the Genbank reference assemblies GRCm37.1 and GRCm39, this GC-rich megasatellite repeat has fewer copies and is located between the *HRH1* and *atg7* genes.

The alignment (file [CM115055_2_reps.align](#)) and Table S2 (file [CM115055_2_reps.xlsx](#)) show how each copy evolved. All the copies contain extended conserved regions and conserved oligonucleotide tracks with the same multiplicity. However, there are three variable homopurine-homopyrimidine tracks (two (CT)_n and one (CA)_n), which are characterized by an active expansion of dinucleotides accompanied by single-nucleotide substitutions. Table 3 shows the coordinates, length and length of coverage by short oligonucleotide tracks (in bp and as a percentage of the copy length) for each copy of the second megasatellite.

The second region is GC-richer than the first: its average GC content is 57.5%. As compared with the long copies of the first megasatellite, the copies of the second one are more uniform in length, GC-composition and coverage by oligonucleotide tracks. A FASTA file with the copy sequences of this megasatellite can be found in the supplementary materials ([CM115055_2_reps.fasta](#)).

Table 3. Characteristics of copies of the second megasatellite

No.	Coordinates	Copy length	GC, %	Length of coverage by short TRs, bp	Coverage, %
1	116593455-116595788	2334	56.34	516	22.11
2	116595789-116598154	2366	56.85	590	24.94
3	116598155-116600539	2385	57.53	598	25.07
4	116600540-116602963	2424	57.55	630	25.99
5	116602964-116605363	2400	57.75	609	25.38
6	116605364-116607765	2402	57.87	623	25.94

No.	Coordinates	Copy length	GC, %	Length of coverage by short TRs, bp	Coverage, %
7	116607766-116610147	2382	57.85	609	25.57
8	116610148-116612553	2406	57.86	634	26.35
9	116612554-116614890	2337	58.19	546	23.36
10	116614891-116617252	2362	57.66	566	23.96
11	116617253-116619590	2338	57.78	550	23.52
12	116619591-116621996	2406	57.61	623	25.89
13	116621997-116624404	2408	57.77	629	26.12
14	116624405-116626834	2430	57.65	611	25.14
15	116626835-116629264	2430	57.57	644	26.50
16	116629265-116631679	2415	57.39	665	27.54
17	116631680-116634113	2434	57.23	691	28.39
18	116634114-116636549	2436	57.27	681	27.96
19	116636550-116638974	2425	57.48	661	27.26
20	116638975-116641397	2423	57.86	572	23.61
21	116641398-116643826	2429	57.68	605	24.91
22	116643827-116646279	2453	57.52	632	25.76
23	116646280-116648701	2422	57.39	563	23.25
24	116648702-116651110	2409	57.2	605	25.11
25	116651111-116653553	2443	57.47	655	26.81
26	116653554-116656000	2447	57.42	636	25.99
27	116656001-116658442	2442	57.45	644	26.37
28	116658443-116660871	2429	56.77	656	27.01
29	116660872-116663265	2394	56.93	578	24.14
30	116663266-116665715	2450	57.31	651	26.57
31	116665716-116668152	2437	57.28	585	24.00
32	116668153-116670593	2441	57.27	619	25.36
33	116670594-116672980	2387	57.31	560	23.46
34	116672981-116675470	2490	56.91	708	28.43
35	116675471-116677911	2441	57.48	629	25.77
36	116677912-116680352	2441	57.56	629	25.77
37	116680353-116682794	2442	57.53	639	26.17
38	116682795-116685135	2341	57.62	565	24.13
39	116685136-116687505	2370	57.89	601	25.36
40	116687506-116689950	2445	57.79	668	27.32
41	116689951-116692391	2441	57.72	666	27.28
42	116692392-116694826	2435	57.74	658	27.02
43	116694827-116697296	2470	57.65	688	27.85
44	116697297-116699801	2505	57.17	699	27.90
45	116699802-116702312	2511	57.23	719	28.63
46	116702313-116704825	2513	57.26	721	28.69
47	116704826-116707340	2515	57.22	723	28.75
48	116707341-116709758	2418	58.02	653	27.01
49	116709759-116712256	2498	57.21	689	27.58
50	116712257-116714708	2452	57.54	668	27.24
51	116714709-116717192	2484	57.53	679	27.33
52	116717193-116719656	2464	57.83	706	28.65
53	116719657-116722110	2454	57.38	662	26.98
54	116722111-116724587	2477	57.61	688	27.78
55	116724588-116727087	2500	57.2	682	27.28
56	116727088-116729545	2458	57.61	698	28.40
57	116729546-116732029	2484	57.17	696	28.02
58	116732030-116734538	2509	57.15	715	28.50
59	116734539-116736972	2434	57.93	669	27.49
60	116736973-116739430	2458	57.49	670	27.26
61	116739431-116741916	2486	57.48	687	27.63

No.	Coordinates	Copy length	GC, %	Length of coverage by short TRs, bp	Coverage, %
62	116741917-116744378	2462	57.84	704	28.59
63	116744379-116746870	2492	57.18	678	27.21
64	116746871-116749334	2464	57.83	706	28.65
65	116749335-116751786	2452	57.38	662	27.00
66	116751787-116754263	2477	57.53	684	27.61
67	116754264-116756762	2499	57.18	681	27.25
68	116756763-116759216	2454	57.54	687	28.00
69	116759217-116761716	2500	57.16	712	28.48
70	116761717-116764163	2447	57.79	681	27.83
71	116764164-116766538	2375	57.05	591	24.88

DISCUSSION

The extended tandem repeats found in the *Mus musculus* genome can be called GC-rich megasatellites. Their key characteristics include localization in the intergenic space, copy lengths greater than 1000 bp, and moderate multiplicity values. Copies of these tandem repeats have extended conserved regions interrupted by micro- and minisatellite tracks of variable length. In the first megasatellite, the islands of variability are formed by the tracks (ACCC)_n, (AAACG)_n, (AAAC)_n, (GTCT)_n, (GT)_n, (CTTC)_n, (GAGAAG)_n and (TCC)_n. The second megasatellite has three islands of variability formed with homopurine-homopyrimidine tracks (CT)_n, (CA)_n, (CT)_n. The conserved regions also contain numerous mono-, di-, and trinucleotide tracks modified with single-nucleotide substitutions.

Apart from detecting long oligonucleotide tracks, the results of which are presented in the supplementary materials (Tables S1 and S2), we counted STRs with the pattern lengths 1–3 bp and multiplicity ≥ 2 for all the copies of the two megasatellites. We found that mono-, di-, and trinucleotide repeats occupy 65.99 % of the first megasatellite and 65.67 % of the second one. Thus, another feature of GC-rich megasatellites is that at least two-thirds of each copy is occupied by STRs. Moreover, the nucleotide content of the detected mono-, di-, and trinucleotide tracks also turned out to be biased: 55.03 % GC and 44.97 % AT for the first megasatellite and 62.16 % GC and 37.84 % AT for the second one.

A frequent occurrence of size variability of micro- and minisatellites has prompted a large number of studies aimed to understand the mechanisms involved. Taking into account the experimental data on the changes of the size of minisatellites in humans during meiosis [52], the researchers initially supposed that minisatellites would expand and contract by homologous recombination, whereas microsatellites would be modified by unrepairable slippage events, taking place between the newly synthesized strand and its template during the S-phase of the DNA synthesis (so-called “replication slippage”) [53]. However, subsequent experiments with both types of the tandem repeats showed that the differences between them were less significant than initially thought. Extensive studies of microsatellites, particularly trinucleotide repeats, revealed that multiple mechanisms were involved in their instability, including replication, meiotic and mitotic homologous recombination and post-replication DNA repair. At the same time, it was shown that minisatellites also evolved by slipping during replication in the S-phase [33].

The mechanisms responsible for the instability of mini- and microsatellites are not fully understood but, at the same time, attract a lot of interest – because in humans, unstable tandem arrays are responsible for a group of genetic diseases known as triplet expansion disorders.

Another intriguing question is how microsatellites evolve. A view, which has recently become generally accepted, is that short microsatellites tend to expand and the longer ones tend to contract [33]. These tendencies are demonstrated by the analysis of tetranucleotide repeat

loci in humans. Correspondingly, a model is created, according to which microsatellites tend to expand until they reach a critical size, above which they tend to contract. This model implies that the size of microsatellites is distributed around this critical threshold value [54]. According to another model, microsatellite evolution is driven by two factors: slippage, which increases the size of repeats and point mutations, which interrupt tandem-repeat sequences, thereby stabilizing them [55, 56]. Our results support both models.

CONCLUSION

The development of third-generation sequencing (TGS) technologies, such as HiFi circular consensus sequencing from PacBio and “ultra-long” (>100,000 bp) sequencing from Oxford Nanopore, made it possible to overcome technical difficulties in sequencing eukaryotic genomes and complete T2T assemblies of a number of them, including the *Mus musculus* genome [46]. This technological breakthrough opened up opportunities for studying the structural and functional organization of complex genomes.

In this work, we have verified tandem repeats previously identified in the intergenic spaces of the *Mus musculus* genome and discovered new intriguing regions of latent periodicity. We suggest to call such regions GC-rich megasatellites. Here, we describe only two of them – with the purpose to demonstrate diversity of these extended tandem repeats, which contain numerous micro- and minisatellite tracks covering more than a third of each copy. It is possible that the biased GC-content of these megasatellites ensures formation of specific secondary structures, facilitating active expansion of microsatellites into these regions. The lengths of the repeat pattern of the two megasatellites are 4320 and 2440, and the numbers of their copies are 83 and 71. The presence of extended conserved regions and a partial protein-coding frame in every copy of the first repeat suggests their origin from a common ancestral region.

One can hypothesize that the mechanism underlying the spread of megasatellite copies is genetic conversion, a process of DNA repair via homologous recombination. This mechanism is activated during reparation of a DNA strand with a double-strand break. When two complementary sequences are paired during homologous recombination, a heteroduplex is formed at the sites of mismatch between the two strands, activating the system of mismatch repair. This system corrects one strand using the other as a template.

Perhaps, there is recombination hotspot near these repeat massives, the presence of which accelerates the repair of double-strand breaks and increases the rates of genetic conversion within this chromosomal region. Because of conversion bias, the regions of recombination hotspot, where conversion events occur particularly frequently, exhibit elevated contents of G and C nucleotides, as compared with their average content in the genome. GC-biased genetic conversion would explain the spatial distribution of GC pairs in the mammalian genomes, the so-called isochoric structure [57].

It should be admitted that the development of advanced *de novo* algorithms for the comprehensive detection of repetitive DNA sequences remains a challenging task due to a high rate of sequencing errors in the TGS reads. The NGS technology yields more accurate but shorter reads, failing to read extended repetitive genomic regions. Experts believe [7] that a hybrid strategy combining short and long reads is currently the most effective way to achieve high-quality sequencing. However, implementing a hybrid strategy would require combining multiple sequencing data, such as NGS reads and TGS reads. Therefore, solving the problem of sequencing errors in TGS reads and making the procedure of repeat detection accurate and complete with an increasing number of high-quality TGS reads will be the subject of future research.

REFERENCES

1. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* 2008. V. 18. P. 1011–1019. doi: [10.1101/gr.070409.107](https://doi.org/10.1101/gr.070409.107)

2. Haubold B., Wiehe T. How repetitive are genomes? *BMC Bioinform.* 2006. V. 7. Article No. 541. doi: [10.1186/1471-2105-7-541](https://doi.org/10.1186/1471-2105-7-541)
3. Yi H., Song H., Hwang J., Kim K., Nierman W.C., Kim H.S. The tandem repeats enabling reversible switching between the two phases of β -lactamase substrate spectrum. *PLoS Genet.* 2014. V. 10. Article No. e1004640. doi: [10.1371/journal.pgen.1004640](https://doi.org/10.1371/journal.pgen.1004640)
4. Bulik-Sullivan B., Finucane H.K., Anttila V., Gusev A., Day F.R., Loh P.R. ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan L., et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 2015. V. 47. P. 1236–1241. doi: [10.1038/ng.3406](https://doi.org/10.1038/ng.3406)
5. O'Dushlaine C.T., Edwards R.J., Park S.D., Shields D.C. Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol.* 2005. V. 6. Article No. R69. doi: [10.1186/gb-2005-6-8-r69](https://doi.org/10.1186/gb-2005-6-8-r69)
6. Hannan A.J. Tandem repeat polymorphisms: Mediators of genetic plasticity modulators of biological diversity and dynamic sources of disease susceptibility. *Adv. Exp. Med. Biol.* 2012. V. 769. P. 1–9.
7. Liao X., Zhu W., Zhou J., Li H., Xu X., Zhang B., Gao X. Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* 2023. V. 6. Article No. 954. doi: [10.1038/s42003-023-05322-y](https://doi.org/10.1038/s42003-023-05322-y)
8. Liao X., Li M., Zou Y., Wu F.-X., Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. *Quant. Biol.* 2019. V. 7. P. 90–109. doi: [10.1007/s40484-019-0166-9](https://doi.org/10.1007/s40484-019-0166-9)
9. Jain C., Rhie A., Hansen N.F., Koren S., Phillippy A.M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods.* 2022. V. 19. No. 6. P. 705–710. doi: [10.1038/s41592-022-01457-8](https://doi.org/10.1038/s41592-022-01457-8)
10. Jakubosky D., D'Antonio M., Bonder M.J., Smail C., Donovan M.K.R., Young Greenwald W.W., Matsui H., i2QTL Consortium, D'Antonio-Chronowska A., Stegle O., et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun.* 2020. V. 11. No. 1. Article No. 2927. doi: [10.1038/s41467-020-16482-4](https://doi.org/10.1038/s41467-020-16482-4)
11. Gymrek M., Willems T., Reich D., Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat. Genet.* 2017. V. 49. No. 10. P. 1495–1501. doi: [10.1038/ng.3952](https://doi.org/10.1038/ng.3952)
12. Castillo-Lizardo M., Henneke G., Viguera E. Replication slippage of the thermophilic DNA polymerases B and D from the *Euryarchaeota* *Pyrococcus abyssi*. *Front. Microbiol.* 2014. V. 5. Article No. 403. doi: [10.3389/fmicb.2014.00403](https://doi.org/10.3389/fmicb.2014.00403)
13. Aguilar M., Prieto P. Telomeres and Subtelomeres Dynamics in the Context of Early Chromosome Interactions During Meiosis and Their Implications in Plant Breeding. *Front. Plant Sci.* 2021. V. 12. Article No. 672489. doi: [10.3389/fpls.2021.672489](https://doi.org/10.3389/fpls.2021.672489)
14. Lim C.J., Cech T.R. Shaping human telomeres: from shelterin and CST complexes to telomeric chromatin organization. *Nat. Rev. Mol. Cell Biol.* 2021. V. 22. P. 283–298. doi: [10.1038/s41580-021-00328-y](https://doi.org/10.1038/s41580-021-00328-y)
15. Lamb J.C., Birchler J.A. The role of DNA sequence in centromere formation. *Genome Biol.* 2003. V. 4. Article No. 214. doi: [10.1186/gb-2003-4-5-214](https://doi.org/10.1186/gb-2003-4-5-214)
16. Miga K.H., Alexandrov I.A. Variation and evolution of human centromeres: a field guide and perspective. *Ann. Rev. Genet.* 2021. V. 55. P. 583–602. doi: [10.1146/annurev-genet-071719-020519](https://doi.org/10.1146/annurev-genet-071719-020519)
17. Ishiura H., Doi K., Mitsui J., Yoshimura J., Matsukawa M.K., Fujiyama A., Toyoshima Y., Kakita A., Takahashi H., Suzuki Y., et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* 2018. V. 50. P. 581–590. doi: [10.1038/s41588-018-0067-2](https://doi.org/10.1038/s41588-018-0067-2)
18. Albertin C.B., Simakov O., Mitros T., Wang Z.Y., Pungor J.R., Edsinger-Gonzales E., Brenner S., Ragsdale C.W., Rokhsar D.S. The octopus genome and the evolution of

- cephalopod neural and morphological novelties. *Nature*. 2015. V. 524. P. 220–224. doi: [10.1038/nature14668](https://doi.org/10.1038/nature14668)
19. Sultanov D., Hochwagen A. Varying strength of selection contributes to the intragenomic diversity of rRNA genes. *Nat. Commun.* 2022. V. 13. Article No. 7245. doi: [10.1038/s41467-022-34989-w](https://doi.org/10.1038/s41467-022-34989-w)
 20. Höweler C.J., Busch H.F.M., Geraedts J.P.M., Niermeijer M.F., Staal A. Anticipation in myotonic dystrophy: fact or fiction? *Brain*. 1989. V. 112. No. 3. P. 779–797. doi: [10.1093/brain/112.3.779](https://doi.org/10.1093/brain/112.3.779)
 21. Taylor J.P., Brown Jr. R.H., Cleveland D.W. Decoding ALS: from genes to mechanism. *Nature*. 2016. V. 539. P. 197–206. doi: [10.1038/nature20413](https://doi.org/10.1038/nature20413)
 22. Fotsing S.F., Margoliash J., Wang C., Saini S., Yanicky R., Shleizer-Burko S., Goren A., Gymrek M. The impact of short tandem repeat variation on gene expression. *Nat. Genet.* 2019. V. 51. No. 11. P. 1652–1659. doi: [10.1038/s41588-019-0521-9](https://doi.org/10.1038/s41588-019-0521-9)
 23. Gemayel R., Vincens M.D., Legendre M., Verstrepen K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 2010. V. 44. P. 445–477. doi: [10.1146/annurev-genet-072610-155046](https://doi.org/10.1146/annurev-genet-072610-155046)
 24. Mukamel R.E., Handsaker R.E., Sherman M.A., Barton A.R., Zheng Y., McCarroll S.A., Loh P.R. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science*. 2021. V. 373. No. 6562. P. 1499–1505. doi: [10.1126/science.abg8289](https://doi.org/10.1126/science.abg8289)
 25. Fueyo R., Judd J., Feschotte C., Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* 2022. V. 23. No. 7. P. 481–497. doi: [10.1038/s41580-022-00457-y](https://doi.org/10.1038/s41580-022-00457-y)
 26. Gemayel R., Cho J., Boeynaems S., Verstrepen K.J. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes (Basel)*. 2012. V. 3. No. 3. P. 461–480. doi: [10.3390/genes3030461](https://doi.org/10.3390/genes3030461)
 27. Shi Y., Niu Y., Zhang P., Luo H., Liu S., Zhang S., Wang J., Li Y., Liu X., Song T., et al. Characterization of genome-wide STR variation in 6487 human genomes. *Nat. Commun.* 2023. V. 14. No. 1. Article No. 2092. doi: [10.1038/s41467-023-37690-8](https://doi.org/10.1038/s41467-023-37690-8)
 28. Schröder C., Horsthemke B., Depienne C. GC-rich repeat expansions: associated disorders and mechanisms. *Med. Genet.* 2022. V. 33. No. 4. P. 325–335. doi: [10.1515/medgen-2021-2099](https://doi.org/10.1515/medgen-2021-2099)
 29. Mandel J.L. Breaking the rule of three. *Nature*. 1997. V. 386. P. 767–768. doi: [10.1038/386767a0](https://doi.org/10.1038/386767a0)
 30. Lalioti M.D., Scott H.S., Buresi C., Rossier C., Bottani A., Morris M.A., Malafosse A., Antonarakis S.E.. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*. 1997. V. 386. P. 847–851. doi: [10.1038/386847a0](https://doi.org/10.1038/386847a0)
 31. Liquori C.L., Ricker K., Moseley M.L., Jacobsen J.F., Kress W., Naylor S.L., Day J.W., Ranum L.P.W. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*. 2001. V. 293. P. 864–867. doi: [10.1126/science.106212](https://doi.org/10.1126/science.106212)
 32. Matsuura T., Fang P., Pearson C.E., Jayakar P., Ashizawa T., Roa B.B., D.L. Nelson. Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? *Am. J. Hum. Genet.* 2006. V. 78. P. 125–129. doi: [10.1086/498654](https://doi.org/10.1086/498654)
 33. Richard G.F., Kerrest A., Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 2008. V. 72. P. 686–727. doi: [10.1128/MMBR.00011-08](https://doi.org/10.1128/MMBR.00011-08)
 34. Baran N., Lapidot A., Manor H. Formation of DNA triplexes accounts for arrests of DNA synthesis at $d(\text{TC})_n$ and $d(\text{GA})_n$ tracts. *Proc. Natl. Acad. Sci. USA*. 1991. V. 88. P. 507–511. doi: [10.1073/pnas.88.2.507](https://doi.org/10.1073/pnas.88.2.507)

35. Arimondo P.B., Barcelo F., Sun J.S., Maurizot J.C., Garestier T., Helene C. Triple helix formation by (G,A)-containing oligonucleotides: asymmetric sequence effect. *Biochemistry*. 1998. V. 37. P. 16627–16635. doi: [10.1021/bi9805588](https://doi.org/10.1021/bi9805588)
36. Rippe K., Fritsch V., Westhof E., Jovin T.M. Alternating d(G-A) sequences form a parallel-stranded DNA homoduplex. *EMBO J*. 1992. V. 11. No. 10. P. 3777–3786. doi: [10.1002/j.1460-2075.1992.tb05463.x](https://doi.org/10.1002/j.1460-2075.1992.tb05463.x)
37. Huertas D., Azorín F. Structural Polymorphism of Homopurine DNA Sequences. d(GGA)_n and d(GGGA)_n Repeats Form Intramolecular Hairpins Stabilized by Different Base-Pairing Interactions. *Biochemistry*. 1996. V. 35. P. 13125–13135. doi: [10.1021/bi961020%2B](https://doi.org/10.1021/bi961020%2B)
38. Shiber M.C., Braswell E.H., Klump H., Fresco J.R. Duplex-tetraplex equilibrium between a hairpin and two interacting hairpins of d(A-G)₁₀ at neutral pH. *Nucleic Acids Res*. 1996. V. 24. No. 24. P. 5004–5012. doi: [10.1093/nar/24.24.5004](https://doi.org/10.1093/nar/24.24.5004)
39. Videvall E. Plasmodium parasites of birds have the most AT-rich genes of eukaryotes. *Microb. Genom*. 2018. V. 4. No. 2. Article No. e000150. doi: [10.1099/mgen.0.000150](https://doi.org/10.1099/mgen.0.000150)
40. Bernardi G., Olofsson B., Filipinski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., Rodier F. The mosaic genome of warm-blooded vertebrates. *Science*. 1985. V. 228. No. 4702. P. 953–958. doi: [10.1126/science.4001930](https://doi.org/10.1126/science.4001930)
41. Hurst L.D., Merchant A.R. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. Biol. Sci*. 2001. V. 268. P. 493–497. doi: [10.1098/rspb.2000.1397](https://doi.org/10.1098/rspb.2000.1397)
42. Tetuev R.K., Nazipova N.N., Pankratov A.N., Dedus F.F. Searching megasatellite tandem repeats in eukaryotic genomes by assessing the oscillations of GC-content curves. *Mathematical biology and bioinformatics*. 2010. Vol. 5. No. 1. P. 30–42. doi: [10.17537/2010.5.30](https://doi.org/10.17537/2010.5.30)
43. Pankratov A.N., Gorchakov M.A., D.A.Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova, Olshevets M.M., Pyatkov M.I., et al. Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences. *Pattern Recognition and Image Analysis*. 2009. V. 19. No. 4. P. 687–692 (in Russian). doi: [10.1134/S105466180904018X](https://doi.org/10.1134/S105466180904018X)
44. Tetuev R.K., Nazipova N.N. Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern. *Repbase Reports*. 2010. V. 10. No. 5. P. 776–776.
45. Tetuev R.K., Dedus F.F., Nazipova N.N. Consensus of repeated region of rat chromosome 4 similar to mouse chromosome 6 repeated region, enclosed in the intergenic region between genes Hrh1 and Atg7. *Repbase Reports*. 2010. V. 10. No. 8. P. 1185–1185.
46. Liu J., Li Q., Hu Y., Yu Y., Zheng K., Li D., Qin L., Yu X. The complete telomere-to-telomere sequence of a mouse genome. *Science*. 2024. V. 386. No. 6726. P. 1141–1146. doi: [10.1126/science.adq8191](https://doi.org/10.1126/science.adq8191)
47. Tetuev R.K., Nazipova N.N. Fast two-bit DNA coding in the TANDIS system for predicting specificity of hybridization of oligonucleotides with their targets. In: *Proceedings of the International Conference “Mathematical Biology and Bioinformatics”*. Ed. V.D. Lakhno. Vol. 10. Pushchino: IMPB RAS, 2024. Article No. e43 (in Russian). doi: [10.17537/icmbb24.59](https://doi.org/10.17537/icmbb24.59)
48. Dedus F.F., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Pankratov A.N., Tetuev R.K. Analytical methods for recognizing repeating structures in the genomes. *Reports of the Academy of Sciences*. 2006. T. 411. No. 5. P. 599 – 602 (in Russian).
49. Tetuev R.K. *Algebra of spectral transformations in data-processing problems*: Thesis for a PhD Degree in Physics and Mathematics. Moscow: Computation Center of the Russian Academy of Sciences, 2007 (in Russian).

50. Zhang Z., Schwartz S., Wagner L., Miller W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 2000. V. 7. No. 1–2. P. 203–214. doi: [10.1089/10665270050081478](https://doi.org/10.1089/10665270050081478)
51. Madeira F., Madhusoodanan N., Lee J., Eusebi A., Niewielska A., Tivey A.R.N., Lopez R., Butcher S. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Research*. 2024. V. 52. No. W1. P. W521–W525. doi: [10.1093/nar/gkae241](https://doi.org/10.1093/nar/gkae241)
52. Jeffreys A.J., Murray J., Neumann R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell*. 1998. V. 2. No. 2. P. 267–273. doi: [10.1016/s1097-2765\(00\)80138-0](https://doi.org/10.1016/s1097-2765(00)80138-0)
53. Strand M., Prolla T.A., Liskay R.M., Petes T.D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*. 1993. V. 365. No. 6443. P. 274–276. doi: [10.1038/365274a0](https://doi.org/10.1038/365274a0)
54. Xu X., Peng M., Fang Z. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 2000. V. 24. No. 4. P. 396–399. doi: [10.1038/74238](https://doi.org/10.1038/74238)
55. Calabrese P.P., Durrett R.T., Aquadro C.F. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics*. 2001. V. 159. No. 2. P. 839–852. doi: [10.1093/genetics/159.2.839](https://doi.org/10.1093/genetics/159.2.839)
56. Santibáñez-Koref M.F., Gangeswaran R., Hancock J.M. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol. Biol. Evol.* 2001. V. 18. No. 11. P. 2119–2123. doi: [10.1093/oxfordjournals.molbev.a003753](https://doi.org/10.1093/oxfordjournals.molbev.a003753)
57. Duret L., Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*. 2009. V. 10. P. 285–311. doi: [10.1146/annurev-genom-082908-150001](https://doi.org/10.1146/annurev-genom-082908-150001)

Received 18.10.2025.

Published 25.10.2025.