Вавиловский журнал генетики и селекции. 2021;25(1):18-29
DOI 10.18699/VJ21.003

Original Russian text www.bionet.nsc.ru/vogis/

Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms

E.V. Ignatieva[1, 2] ✉, E.A. Matrosova[1, 2]

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

Abstract. Whole genome and whole exome sequencing technologies play a very important role in the studies of the genetic aspects of the pathogenesis of various diseases. The ample use of genome-wide and exome-wide association study methodology (GWAS and EWAS) made it possible to identify a large number of genetic variants associated with diseases. This information is accumulated in the databases like GWAS central, GWAS catalog, OMIM, ClinVar, etc. Most of the variants identified by the GWAS technique are located in the noncoding regions of the human genome. According to the ENCODE project, the fraction of regions in the human genome potentially involved in transcriptional control is many times greater than the fraction of coding regions. Thus, genetic variation in noncoding regions of the genome can increase the susceptibility to diseases by disrupting various regulatory elements (promoters, enhancers, silencers, insulator regions, etc.). However, identification of the mechanisms of influence of pathogenic genetic variants on the diseases risk is difficult due to a wide variety of regulatory elements. The present review focuses on the molecular genetic mechanisms by which pathogenic genetic variants affect gene expression. At the same time, attention is concentrated on the transcriptional level of regulation as an initial step in the expression of any gene. A triggering event mediating the effect of a pathogenic genetic variant on the level of gene expression can be, for example, a change in the functional activity of transcription factor binding sites (TFBSs) or DNA methylation change, which, in turn, affects the functional activity of promoters or enhancers. Dissecting the regulatory roles of polymorphic loci have been impossible without close integration of modern experimental approaches with computer analysis of a growing wealth of genetic and biological data obtained using omics technologies. The review provides a brief description of a number of the most well-known public genomic information resources containing data obtained using omics technologies, including (1) resources that accumulate data on the chromatin states and the regions of transcription factor binding derived from ChIP-seq experiments; (2) resources containing data on genomic loci, for which allele-specific transcription factor binding was revealed based on ChIP-seq technology; (3) resources containing *in silico* predicted data on the potential impact of genetic variants on the transcription factor binding sites.
Key words: transcription regulation; genetic variability; pathogenic genetic variants; transcription regulatory regions; transcription factor binding sites (TFBSs); genomic databases.

For citation: Ignatieva E.V., Matrosova E.A. Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and genomic resources for dissecting these mechanisms *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):18-29. DOI 10.18699/VJ21.003

Геномная изменчивость в регуляторных районах генов, ассоциированная с заболеваниями человека: механизмы влияния на транскрипцию генов и полногеномные информационные ресурсы, обеспечивающие исследование этих механизмов

Е.В. Игнатьева[1, 2] ✉, Е.А. Матросова[1, 2]

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

Аннотация. Полногеномные и полноэкзомные технологии секвенирования играют важную роль в исследованиях генетических аспектов патогенеза различных заболеваний. Широкое применение методов полногеномного и полноэкзомного анализа ассоциаций позволяет идентифицировать множество вариантов геномной изменчивости (ГИ), ассоциированных с заболеваниями. Эта информация накапливается в базах данных GWAS central, GWAS catalog, OMIM, ClinVar и др. Большинство вариантов, идентифицированных методикой полногеномного анализа ассоциаций, располагается в некодирующих областях генома человека. По данным проекта ENCODE, доля участков в геноме человека, потенциально задействованных в регуляции транскрипции, во много раз превышает долю кодирующих областей. Таким образом, геномная изменчивость в некодирующих областях генома может повышать предрасположенность к заболеваниям, нарушая функционирование различных регуляторных элементов (промоторов, эн-

© Ignatieva E.V., Matrosova E.A., 2021
This work is licensed under a Creative Commons Attribution 4.0 License

хансеров, участков, определяющих 3D структуру хроматина и т. д.). Однако идентификация механизмов влияния патогенных вариантов ГИ на риск развития заболеваний затруднена ввиду большого разнообразия регуляторных элементов. В обзоре рассмотрены молекулярно-генетические механизмы влияния патогенных вариантов ГИ на экспрессию генов. При этом внимание сосредоточено на транскрипционном уровне регуляции как ключевой стадии, запускающей последовательность этапов экспрессии любого гена. Пусковым событием, опосредующим влияние патогенного варианта ГИ на уровень экспрессии гена, может быть, например, изменение функциональной активности сайтов связывания транскрипционных факторов или уровня метилирования ДНК, что, в свою очередь, отражается на функциональной активности промоторов или энхансеров. Выявление регуляторных эффектов полиморфных локусов невозможно без тесной интеграции современных экспериментальных подходов с компьютерным анализом больших массивов генетических данных, получаемых на основе омиксных технологий. В обзоре кратко описаны наиболее известные открытые полногеномные информационные ресурсы, содержащие данные, полученные на основе омиксных технологий, в том числе: ресурсы, накапливающие сведения о состоянии хроматина и участках его связывания с транскрипционными факторами, выявленными с помощью технологии ChIP-seq; ресурсы по геномным локусам, для которых на основе данных ChIP-seq выявлено аллель-специфичное связывание с транскрипционными факторами; а также ресурсы, содержащие предсказанные *in silico* данные о потенциальном влиянии геномной изменчивости на сайты связывания транскрипционных факторов.
Ключевые слова: регуляция транскрипции; геномная изменчивость; патогенные геномные варианты; районы, регулирующие транскрипцию; сайты связывания транскрипционных факторов; геномные базы данных.

Introduction

At present, largely due to the widespread use of the technology of genome-wide and exome-wide association study (GWAS and EWAS), a large number of polymorphisms associated with diseases have been identified. For example, GWAS central (https://www.gwascentral.org/) contains information on more than 70 million associations between ~3.2 million SNPs and 1451 diseases or phenotypic characteristics (Beck et al., 2020). Experimental datasets of comparable volume have been accumulated in a number of other databases on genotype-phenotype associations (GWAS catalog, OMIM, ClinVar, HGMD, PheGenI, EGA, GAD, dbGaP).

Currently, a large amount of experimental data has been obtained about the disease-associated genetic variants (GVs), but the molecular mechanisms underlying these associations are extremely poorly understood. This is due to the fact that only a relatively small proportion of pathogenic GVs is located in the coding regions of the human genome, changes in the nucleotide sequence of which disrupt the structure and function of proteins. A huge mass of polymorphic loci associated with diseases is located in non-coding regions of the genome (introns, 5′- and 3′-flanking regions of genes, intergenic regions). For example, according to GWAS data, ~90 % of the total number of variants associated with diseases are located in noncoding regions of the human genome (Maurano et al., 2012; Farh et al., 2015).

It is known that non-coding regions of the genome contain regions that perform a wide range of regulatory functions: promoter regions, enhancers, negative regulatory elements, nuclear matrix attachment regions, regions that determine the structure of topologically associating domains (TADs), and other features of 3D organization of genome (Mathelier et al., 2015; Meddens et al., 2019; Ibrahim, Mundlos, 2020). The proportion of regions in the human genome potentially involved in the transcriptional regulation is extremely high. According to the ENCODE project, the chromatin regions corresponding to the peaks of transcription factor (TF) binding identified by the ChIP-seq occupy ~8.1 % of the total genomic DNA (ENCODE Project Consortium, 2012), which is significantly higher than the proportion of coding regions of the human genome (~1.2 %). Considering that not all known TFs and not all cell lines were studied in the ENCODE project, an obviously larger fraction of genomic DNA is involved in the interaction with TFs. The total length of human genome regions with enhancer-associated chromatin features also significantly exceeds the total size of the coding regions: for example, in only one cell type studied (H1-ES), enhancer regions occupy ~3.2 % (Roadmap Epigenomics Consortium et al., 2015).

Studies aimed at identifying the mechanisms of the influence of pathogenic GVs on the predisposition to diseases are carried out very actively, which is reflected in a number of review publications (Mathelier et al., 2015; Merkulov et al., 2018; Smith et al., 2018; Wang et al., 2019; Vohra et al., 2020). The most discussed effect of pathogenic GVs is a change in the binding activity of TFBSs (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Mathelier et al., 2015; Gorbacheva et al., 2018). It has also been shown that polymorphic loci can be associated with alteration of DNA methylation patterns (Howard et al., 2014; Kumar D. et al., 2017; Rahbar et al., 2018; Schmitz et al., 2019) and modifications of histone proteins (Kilpinen et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019), with structural change in chromatin loops (Visser et al., 2015; Zhang et al., 2018) and, as one of the manifestations of this process, with changes in the TADs structure (Cong et al., 2019; Mei et al., 2019). Examples of such effects will be discussed below (Table 1).

The effects of genetic variants on the functional activity of transcription factor binding sites

The key role in the transcriptional regulation is played by transcription factors – proteins that can specifically bind to DNA of the regulatory regions of genes and to initiate the transcription complexes formation. The human genome contains more than 1500 genes encoding TFs (Wingender et al., 2013). TF binding sites, as a rule, have a length of 10–25 nucleotides (Levitsky et al., 2014; Kulakovskiy et al., 2018).

Nucleotide substitutions, as well as short insertions/deletions at polymorphic loci, can disrupt TFBSs or create them *de novo* (see Table 1), and this, in turn, can have both negative and positive effects on the level of gene transcription (Chen L. et al., 2013; Gorbacheva et al., 2018). Such GVs (and the cor-

Table 1. Examples of polymorphic loci associated with pathologies and mechanisms of their action on the gene expression level

Disease or pathology	Polymorphic locus	Localization	Mechanism	Reference
Atopic asthma	rs928413 A→G	*IL33* promoter region	rs928413(G) allele creates a binding site for the transcription factor CREB1 leading to increased expression level of *IL33*	Gorbacheva et al., 2018
Obesity	rs1421085 T→C	Intron of the *FTO* gene which contains the regulatory region of the *IRX3* and *IRX5* genes (the distance between rs1421085 and TSSs of *IRX3* and *IRX5* is ~517,000 and ~1,164,000 bases)	rs1421085(C) variant disrupts a conserved motif for the ARID5B repressor, which leads to derepression of a potent preadipocyte enhancer and a doubling of *IRX3* and *IRX5* expression	Claussnitzer et al., 2015
Pancreatic cancer	rs2001389 A→G	The boundary of TAD located on chromosome 10	The allele G of rs2001389 weakens the CTCF binding activity of DNA, eliminating TAD boundary and altering 3D chromatin structure, and it is related to the lower expression of a putative antioncogene *MFSD13A*	Mei et al., 2019
Disturbances of lipid metabolism	rs174537 G→T	An enhancer region of the FADS cluster	Individuals that have rs174537(T) allele exhibited a higher level of DNA methylation at CpG sites located within regulatory region of FADS cluster, which led to a decrease in transcriptional activity of *FADS1* and *FADS2*	Howard et al., 2014
Atopic dermatitis	rs612529 T→C	*VSTM1* promoter region	The rs612529(T) allele facilitates binding of the transcription factor PU.1, that acts as docking site for DNA demethylases. In carriers of pathogenic variant C, the interaction of PU.1 with DNA is disrupted, as a result, the methylation level of the *VSTM1* promoter is elevated, and this is accompanied by a downregulation of *VSTM1* expression	Kumar D. et al., 2017
Fragile X syndrome	CGG repeat expansion. Healthy individuals harbor between 5 and 55 copies of the CGG repeats, affected patients harbor more than 100 copies	The 5′-untranslated region of *FMR1* gene	CGG repeat expansion disrupts the structure of TAD, that includes *FMR1*. In individuals with mutation-length CGG triplet repeats, the 5′-boundary region of TAD is ablated (this region is hypermethylated and its CTCF occupancy is lost). As a result, one subTAD dissolves. *FMR1*, which is normally associated with the downstream TAD, shifts to the upstream TAD. In this case, *FMR1* promoter is hypermethylated, and *FMR1* expression is downregulated	Sun et al., 2018
Rheumatoid arthritis and type-2 diabetes mellitus	rs7873784 G→C	The 3′-untranslated region of *TLR4* gene	rs7873784(C) allele creates a binding site for transcription factor PU.1, a known regulator of *TLR4* expression. Functional PU.1 binding site augments the enhancer activity of *TLR4* 3′-UTR that leads to increased *TLR4* expression	Korneev et al., 2020
Breast cancer	rs4321755 C→T	Enhancer region of *MRPS30* and *RP11-53O19.1* genes	The risk allele rs4321755(T) creates a GATA3 binding motif within an enhancer, resulting in stronger binding of GATA3 and chromatin accessibility, thereby activating interaction between the enhancer and *MRPS30/RP11-53O19.1* divergent promoter and increasing the expression of *MRPS30* and *RP11-53O19.1* genes	Zhang et al., 2018

responding polymorphic loci) that affect the transcriptional activity of genes are usually called regulatory variants (Kumar S. et al., 2017; Guo, Wang, 2018; Merkulov et al., 2018).

Pathological (that is, associated with a disease) can be both an allelic variant of the DNA sequence containing a disrupted TFBS (Lewinsky et al., 2005; Chen L. et al., 2013; Claussnitzer et al., 2015; Kumar D. et al., 2017; Mei et al., 2019) and an allelic variant, leading to creation of TFBS *de novo* (Gorbacheva et al., 2018; Zhang et al., 2018; Korneev et al., 2020) (see Table 1).

Pathological GVs, affecting the binding activity of TFBSs, can be located not only in promoter regions, but also in regulatory regions located at considerable distance from transcription start sites (TSSs) of genes: enhancers (Lewinsky et al., 2005; Zhang et al., 2018; Meddens et al., 2019), regulatory regions with repressive function (Claussnitzer et al., 2015),

and TAD boundary regions (Mei et al., 2019) (see Table 1). For example, the rs1421085 T→C substitution associated with obesity impairs the functioning of the negative regulatory region controlling expression of the *IRX3* and *IRX5* genes (Claussnitzer et al., 2015). The rs1421085 locus is located in the intron of the *FTO* gene (Fig. 1) at a considerable distance from the transcription start sites of *IRX3* and *IRX5* (~520,000 and ~1,164,000 bases). Normally, the DNA region containing allele T interacts with a repressor factor ARID5B, leading to a decrease in transcriptional activity of *IRX3* and *IRX5* genes. In carriers of the mutant variant of the DNA sequence (allele C), the binding site of the ARID5B repressor factor is disrupted, which causes an excessively high expression of the *IRX3* and *IRX5* genes and activates adipogenesis (Claussnitzer et al., 2015).

Occasionally a nucleotide substitution at a polymorphic locus disrupts the TFBS and this, in turn, affects the functional activity of the TAD (see Table 1). This effect was found in the case of A→G (rs2001389), associated with the risk of pancreatic cancer (Fig. 2). The rs2001389 locus is located in the region that determines the structure of chromatin loops within the TAD. This TAD contains 91 genes and is formed by spatially adjacent chromatin regions (Mei et al., 2019). The DNA region containing the risk allele G is characterized by a reduced ability to interact with CTCF, which in this case acts as a structural protein of chromatin. Normally, CTCF binding ensures the functioning of one of the regions that determines the structure of chromatin loops within the considered TAD. The pathogenic allele G alters the activity of CTCF binding motif within TAD boundary disrupting the stability of corresponding 3D structure of chromatin. As a result, the expression of the genes within this TAD is impaired. In this case, the greatest decrease in *MFSD13A* expression is observed.

The effects of genetic variability on DNA methylation and gene transcriptional activity

DNA methylation doesn't change the nucleotide sequence and is the addition of a methyl group to the fifth carbon atom of cytosine (Angeloni, Bogdanovic, 2019). An increase in the level of DNA methylation, as a rule, leads to a long-term inactivation of the expression of genes lying in the methylated region, since, according to the generally accepted concept, methylation of a DNA region facilitates recruiting protein complexes, containing histone deacetylase (HDAC) (Jones et al., 1998; Nan et al., 1998). DNA methylation can also decrease the ability of some TFs to interact with DNA: it is known that CTCF factors and factors from the ETS family have such sensitivity to methylation (Wang et al., 2019). In contrast, another transcription factor, ZFP57, binds only to methylated DNA (Quenneville et al., 2011). Thus, cytosine methylation can activate different mechanisms of gene transcription regulation, and not always an increase in the methylation level of the regulatory DNA region is associated with a decrease in the expression of the corresponding gene (Izzi et al., 2016; Wang et al., 2019).

Genetic variability affects significantly the methylation of DNA regions that have regulatory potential. Thus, a genome-wide analysis of the methylation patterns of DNA collected from 24 subjects from Norfolk Island genetic isolate (Benton

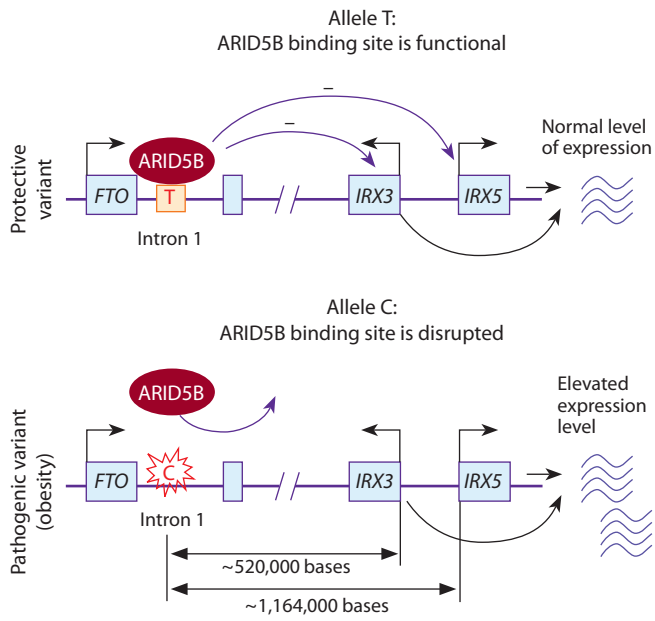

Fig. 1. Disruption of the binding site caused by T→C substitution (rs1421085) weakens ARID5B repressor binding to the regulatory region of the *IRX3* and *IRX5* genes. As a result, the level of expression of *IRX3* and *IRX5* is increased.

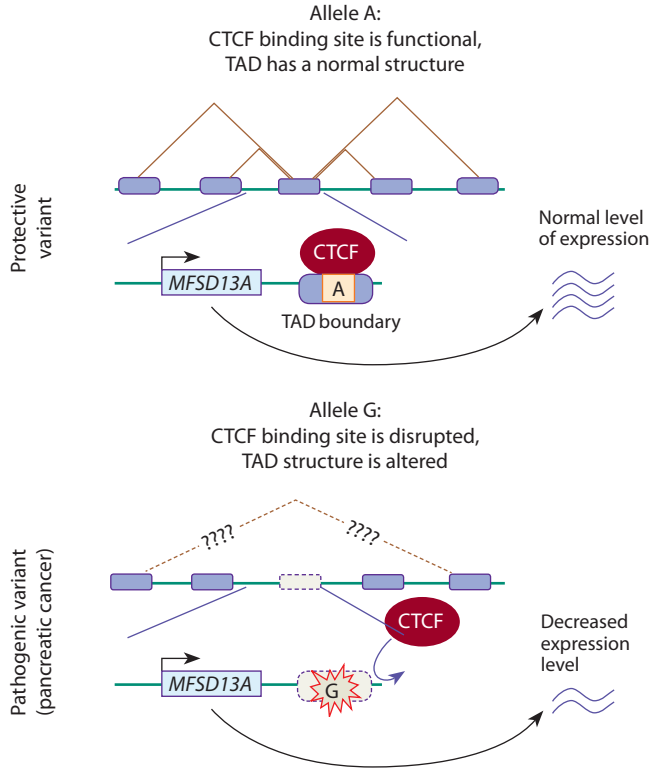

Fig. 2. Disruption of the CTCF binding site caused by the nucleotide substitution (rs2001389) eliminates one of the boundary regions that determine the TAD structure. As a result, the tumor suppressor gene *MFSD13A* expression is downregulated.

The contacts between chromatin regions within the TAD are shown with brown lines. Interrogation points in the bottom figure indicate the lack of data on the new structure of TAD.

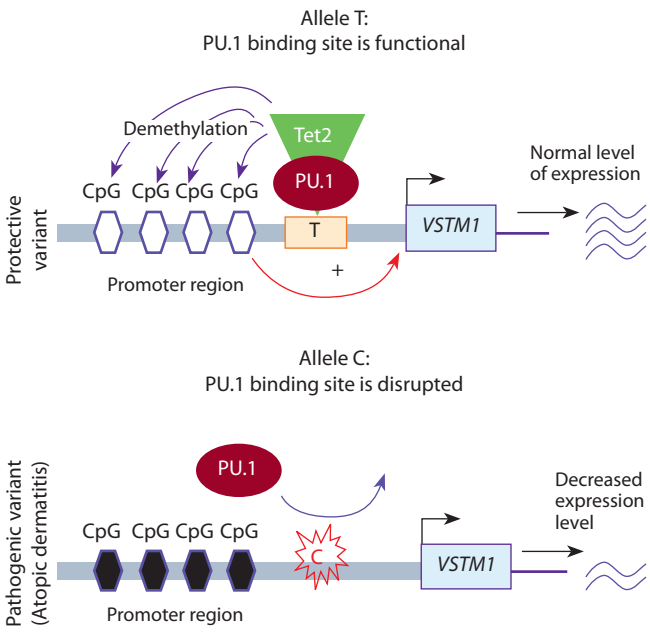

Fig. 3. Disruption of the PU.1 binding site caused by the T→C (rs612529) nucleotide substitution reduces the activity of demethylases (for example, Tet2) that maintain the *VSTM1* promoter region in an active state, and therefore *VSTM1* expression is downregulated.

et al., 2019), identified 12,761 regions containing at least two CpG dinucleotides and having an allele-specific methylation level. In most cases (98 %), regions with allele-specific methylation level are co-localized with single nucleotide variants presented in dbSNP (Benton et al., 2019).

This study (Benton et al., 2019) also analyzed the location of allele-specific methylation regions relative to the set of polymorphic loci associated with human diseases extracted from the GWAS catalog database. It turned out that polymorphic loci associated with diseases overlap with regions of allele-specific methylation twice more often than it would be expected by chance. This means that the change in methylation levels due to genetic variability is one of the factors that increase the risk of disease.

As an example, consider the rs174537 (G→T) polymorphic locus located in the enhancer of the *FADS1* and *FADS2* genes encoding fatty acid desaturases 1 and 2. The T variant of the rs174537 locus is associated with an increased risk of pathological disturbances of lipid metabolism (see Table 1). It was shown that individuals that have rs174537(T) allele had a higher methylation level of the regulatory region of the *FADS1* and *FADS2* genes in human liver (Howard et al., 2014), which led to the suppression of the transcriptional activity of *FADS1* and *FADS2*.

Occasionally, in one of the allelic variants, DNA demethylation occurs, initiated by TF binding to DNA (see Table 1). For example, such a mechanism was revealed for rs612529 T→C. This locus is located in the promoter region of the *VSTM1* (Fig. 3). The low expression of *VSTM1* in monocytes provokes the development of atopic dermatitis. In this cell type, the promoter region containing the protective variant T interacts with the transcription factor PU.1 more actively than the other one containing variant C. PU.1 initiates DNA demethylation by recruiting DNA demethylases (for example, Tet2). As a result, carriers of the T allele have completely demethylated *VSTM1* promoter, and *VSTM1* expression is activated. In carriers of pathogenic variant C, the interaction of PU.1 with DNA is disrupted, as a result, methylation level of the *VSTM1* promoter is elevated, and this is accompanied by a decrease in *VSTM1* expression (Kumar D. et al., 2017).

The effects of the genetic variability on the chromatin states and chromatin spatial organization

Pathogenic GVs may impair the chromatin state (Kilpinen et al., 2013). There are cases when the presence of a pathogenic GV was accompanied by a change in the patterns of histone modification and the appearance (or disappearance) of DNase I hypersensitive sites (McVicker et al., 2013; Visser et al., 2015; Zhang et al., 2018; Cong et al., 2019). In these cases, allele-specific contacts between promoters and enhancers were identified, the number of which correlated with the activity of the enhancer regions.

There are also known cases when structural variations of the genome (insertions, deletions, duplications, inversions, translocations longer than 50 nucleotides) lead to a change in the spatial organization of chromatin, thereby disrupting the expression of genes associated with pathological processes (Sun et al., 2018; Ibrahim, Mundlos, 2020). For example, the expansion of CGG trinucleotide repeats in the 5′-untranslated region (5′-UTR) of the *FMR1* gene, associated with the fragile X syndrome, disrupts the structure of TAD, that includes *FMR1* (Fig. 4, see Table 1). Normally, *FMR1* is very close to the 5′-boundary region of TAD (in Fig. 4, this is TAD1). The DNA region corresponding to this 5′-boundary is hypomethylated and is occupied by CTCF. In individuals with mutation-length CGG triplet repeats (more than 100), this boundary is ablated (this region is hypermethylated and its CTCF occupancy is lost). As a result, TAD1 dissolves and the boundary of the other TAD (in Fig. 4 it is designated as TAD2) shifts to the 3′-region of *FMR1*. Therefore, *FMR1* is within the TAD2, which normally does not contain this gene. In this case, *FMR1* promoter is hypermethylated, and *FMR1* expression is inactivated (Park et al., 2015; Sun et al., 2018).

To study molecular-genetic mechanisms of the effect of genome variability on the 3D chromatin structure, it is necessary to reconstruct the spatial genome organization. The following basic levels of the 3D genome organization have been identified: (1) regulatory DNA loops that bring together promoters and enhancers; (2) topologically associating domains (TADs), within which DNA regions have more contacts with each other than with neighboring domains; (3) A and B compartments corresponding to transcriptionally active and condensed chromatin; and finally (4) chromosome territories (Fishman et al., 2018; Hansen et al., 2018). Disruption of 3D contacts between promoters and enhancers within the TAD, caused, for example, by chromosomal rearrangements, can significantly affect the transcriptional activity of a gene, increasing risk of diseases (Lupiáñez et al., 2015).

The Institute of Cytology and Genetics SB RAS has developed an experimental computer approach for prediction

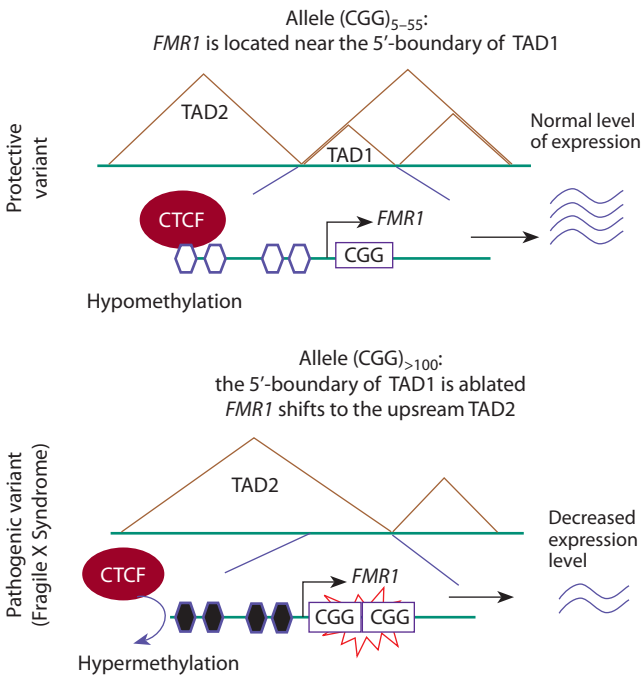

Fig. 4. With an increase in the number of CGG triplet repeats in the 5′-untranslated region of the *FMR1* gene, the DNA region corresponding to the TAD1 boundary region is hypermethylated. This leads to impaired binding of CTCF factors and disrupts a barrier function of the boundary region.

The brown lines show the contacts between chromatin loops within TADs.

physical contacts between promoters and enhancers within the 3D chromatin structure (Fishman et al., 2018; Belokopytova et al., 2020; Belokopytova, Fishman, 2021). The approach is based on the following information: (1) cell type; (2) cell-specific localization of enhancers in the linear genome (from the ENCODE database); (3) transcriptional activity of promoters (from RNA-seq experiments); (4) boundaries of chromatin loop extrusion (based on ChIP-seq mapping of CTCF occupancy in a definite cell type); (5) orientation of CTCF binding motifs (based on motif prediction pipeline); (6) A or B chromatin compartment (according to Hi-C experiments). Analysis of these data using the original 3DPredictor program (Belokopytova et al., 2020), developed on the basis of machine learning algorithms, allows to predict the frequencies of physical contacts between promoters and enhancers in the 3D genome structure with an accuracy that exceeds the accuracy of other known prediction methods.

The 3DPredictor was used to analyze the 3D genome structure in homozygous *DelB/DelB* mice that have a deletion of the 1.5 Mb genomic region containing *Epha4*. This deletion is accompanied by the appearance of additional contacts between *Pax3* gene and *Epha4* enhancer region, altering *Pax3* expression and leading to brachydactyly. Mice with the *DelB/DelB* genotype are a genetic model of human pathology accompanied by limb malformations (Lupiáñez et al., 2015). Testing 3DPredictor on this model has demonstrated the high efficiency of the program: in homozygous *DelB/DelB* mice, ectopic contacts between the *Pax3* gene and *Epha4* enhancers cluster were predicted (Belokopytova et al., 2020), and these predictions were in good agreement with the experimental data.

Genetic variability: combined analysis of heterogeneous big biological and genetic data

As noted above, many polymorphic loci associated with diseases are located at a considerable distance from the coding regions of genes (ENCODE Project Consortium, 2012; Maurano et al., 2012). Additional studies are needed to identify the molecular-genetic mechanisms of the influence of such GVs on the predisposition to diseases. The purpose of such studies is to clarify the regulatory role of GVs. A typical example is the work (Zhang et al., 2018), which made it possible to find a functionally active regulatory variant rs4321755 associated with the risk of breast cancer. The rs4321755 locus is located in a distant enhancer that regulates the expression of the *MRPS30* and *RP11-53O19.1* genes (see Table 1). It turned out that in the presence of the pathogenic variant rs4321755(T), a new GATA3 binding site is created. The transcription factor GATA3 increases the functional activity of the enhancer, this leads to the formation of more contacts between the enhancer and the divergent promoter of the *MRPS30* and *RP11-53O19.1* genes, and increased expression level of these genes. To identify this functionally significant regulatory variant, the authors developed an integrated experimental computer method based on a combined analysis of heterogeneous big biological and genetic data, including: (1) data on allele-specific expression obtained from RNA-seq in combination with data on haplotypes; (2) expression quantitative trait loci (eQTL); (3) genomic distribution of DNAse I hypersensitive sites; (4) localization of ChIP-seq peaks from ENCODE and GEO databases; (5) localization of regulatory motives predicted by computer programs. Similar scenarios for integrated experimental computer research have been implemented in the other studies (Chen C.-Y. et al., 2014; Claussnitzer et al., 2015; Zhao et al., 2019; Li et al., 2020).

This kind of research became possible due to (1) the development of modern high-throughput experimental approaches that allow producing data of different types on a genome-wide scale (parallel high-throughput sequencing, ChIP-seq, 3C, Hi-C, ChIA-PET techniques, DNAse I footprinting, bisulfite sequencing, etc.); (2) development of public information resources accumulating such experimental data. Table 2 provides a brief description of information resources containing genomic data obtained on the basis of omics technologies and used to study the mechanisms by which GVs alter the level of transcription. These resources present (1) the human genome annotation (GENCODE); (2) genome variability in human populations (HapMap, 1000 Genomes Project, IGSR, dbSNP); (3) GVs associated with diseases (GWAS central, GWAS catalog, ClinVar, HGMD, OMIM, etc.); (4) modifications of the chromatin (ENCODE, NIH Roadmap Epigenomics Mapping Consortium); (5) expression quantitative trait loci (GTEx project, eQTL databases, exSNP, etc.); (6) profiling of transcription factor binding events by ChIP-seq (Cistrome Data Browser, GTRD, ReMap); (7) allele-specific binding of TFs, identified using ChIP-seq data in combination with the data on the genotypes of the studied cells (AlleleDB,

Table 2. Information resources on genomic data obtained on the basis of the modern high-performance experimental methods

Information resource	URL	Description
	The human genome annotation	
GENCODE*	https://www.gencodegenes.org/	Reference quality human gene annotations created by merging the results of manual and computational gene annotation methods
	Genetic diversity in human populations	
HapMap (Haplotype Map)	https://www.genome.gov/10001688/international-hapmap-project ftp://ftp.ncbi.nlm.nih.gov/hapmap/	A map of haplotype blocks of the human genome and the specific SNPs that identify the haplotypes (tag SNPs)
1000 Genomes Project (1KGP)	https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/	Genetic variants (single nucleotide polymorphisms, insertions/deletions, structural variants) and genotypes identified in individuals from 26 populations
International Genome Sample Resource (IGSR)	https://www.internationalgenome.org	Combining 1000 Genomes Project data with the other large datasets generated on 1000 Genomes samples by projects such as GEUVADIS, who generated RNA-Seq data on the 1000 Genomes European samples and the YRI population, and ENCODE, who have carried out extensive assays on the NA12878 cell line
dbSNP	https://www.ncbi.nlm.nih.gov/snp/	Human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with population frequency, publication, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations. The human data in dbSNP include submissions from the SNP Consortium, variations mined from genome sequence as part of the human genome project, and individual lab contributions of variations in specific genes, mRNAs, ESTs, or genomic regions
	Disease-associated genetic variants	
GWAS central (Genome-wide association studies central)	https://www.gwascentral.org/	Allele and genotype frequency data, genetic association significance findings. GWAS central gathers datasets from public domain projects, and also encourage direct data submission from the community
GWAS catalog (Genome-wide association studies catalog)	https://www.ebi.ac.uk/gwas/home	Data on associations between polymorphic loci and phenotypic traits extracted from the published GWA studies
OMIM (Online Mendelian Inheritance in Man)	https://www.ncbi.nlm.nih.gov/omim	A compendium of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. OMIM is based on the peer-reviewed biomedical literature
ClinVar (Clinical Variations)	https://www.ncbi.nlm.nih.gov/clinvar/	A public archive of reports of the relationships among human variations and phenotypes
HGMD (The Human Gene Mutation Database)	http://www.hgmd.cf.ac.uk/ac/index.php	All published gene lesions responsible for human inherited disease
PheGenI (The Phenotype-Genotype Integrator)	https://www.ncbi.nlm.nih.gov/gap/phegeni	Phenotype-oriented resource that merges GWAS catalog data with several other databases (Gene, dbGaP, OMIM, eQTL and dbSNP)
EGA (The European Genome-phenome Archive)	https://ega-archive.org/	Data on the relationship between genotypes and phenotypes obtained by various experimental methods (GWAS, exome sequencing, whole-genome sequencing, single-cell sequencing, genotyping)
dbGaP (The database of Genotypes and Phenotypes)	https://www.ncbi.nlm.nih.gov/gap/	Data and results from studies that have investigated the interaction of genotype and phenotype in humans. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits
	Chromatin modifications and chromatin states	
ENCODE (The Encyclopedia of DNA Elements)	http://genome.ucsc.edu/ENCODE/ https://www.encodeproject.org/	Genome-wide profiles of histone modifications, genome-wide DNA methylation profiles, regions of TF binding derived from ChIP-seq experiments, interactions between genomic loci, genomic distribution of DNAse I hypersensitive sites, expression data for more than 300 cell types
NIH Roadmap Epigenomics Mapping Consortium	http://www.roadmapepigenomics.org/	Human epigenomic data (DNA methylation profiles, histone modifications, chromatin accessibility, etc.). Annotation of the human genome in accordance with the classifications of chromatin states (15, 18, 25-state models)

End of Table 2

Information resource	URL	Description
	Expression quantitative trait loci (eQTL)	
Genotype-Tissue Expression (GTEx) project	https://www.gtexportal.org/home/	Expression and eQTL data in 54 human cell types with a healthy phenotype
eQTL databases	https://www.hsph.harvard.edu/liming-liang/software/eqtl/	Expression quantitative trait loci derived from lymphoblastoid cell lines
exSNP	http://www.exsnp.org/	eQTL data from six cell types (LCLs, B cells, monocytes, brain, liver, and skin) integrated with SNPs in disease risk loci from GWA studies of seven common human diseases
eQTL Catalogue	https://www.ebi.ac.uk/eqtl/	Cis-eQTLs and splicing QTLs from all available public studies on human (including GTEx project data)
eQTL Browser	http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/	eQTLs identified in recent studies in multiple tissues
	Profiling of transcription factor binding events by ChIP-seq	
Cistrome Data Browser	http://cistrome.org/db/#/	The ChIP-seq, DNase-seq and ATAC-seq data: (1) genomic regions interacting with TFs, (2) DNase I hypersensitive sites, (3) the binding locations of modified histone proteins. The data has been assigned statuses according to six quality control criteria
Gene Transcription Regulation Database (GTRD)	https://gtrd.biouml.org/#!	A collection of ChIP-seq experiments aimed at finding TF binding sites in the human and mouse genomes
ReMap (Global map of regulatory elements)	http://remap.univ-amu.fr/	A collection of ChIP-seq, ChIP-exo, DAP-seq experiments from public resources (GEO, ENCODE, ENA). Chromatin regions in contact with TFs, transcriptional coactivators, and chromatin remodeling factors
	Allele-specific binding of TFs, identified using ChIP-seq data in combination with the data on the genotypes of the studied cells	
AlleleDB	http://alleledb.gersteinlab.org/	Genomic annotation of cis-regulatory SNVs associated with allele-specific binding and expression derived from RNA-seq and ChIP-seq data of 383 individuals from the 1000 Genomes Project
AlleleSeq	http://alleleseq.gersteinlab.org/	Allele-specific binding of six TFs (cFos, cMyc, JunD, Max, NfκB, CTCF) identified using variation data for NA12878 from the 1000 Genomes Project as well as matched, deeply sequenced RNA-Seq and ChIP-Seq data sets generated for this purpose
	The effects of genetic variants on TFBSs predicted *in silico* by computer programs	
HaploReg	https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php	Annotation of polymorphic loci within haplotype blocks that were defined using LD information from the 1000 Genomes Project. Annotation includes: (1) chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, (2) sequence conservation across mammals, (3) the effect of GVs on regulatory motifs, (4) the effect of GVs on expression from eQTL studies
SNP2TFBS	http://ccg.vital-it.ch/snp2tfbs/	Genetic variants from 1000 Genomes Project, which, according to *in silico* predictions, affect the similarity of TFBSs with weight matrices
rSNPBase	http://rsnp3.psych.ac.cn/index.do	SNP-related regulatory elements (TF binding regions, TADs, mature miRNA regions, predicted miRNA target sites, etc.), SNP-related regulatory element-target gene pairs, SNP-based regulatory networks
rVarBase	http://rv.psych.ac.cn/	Annotation of polymorphic loci (including copy number variations). Annotation includes (1) chromatin state, (2) related regulatory element (CpG islands, matched TF binding sites, miRNA target sites, etc.), (3) target genes
	Information resources integrating or accumulating diverse types of data	
UCSC Genome Browser	https://genome.ucsc.edu/	Data is integrated based on a graphical interface that allows visualizing genome sequences along with a large number of annotations and features (positions of transcripts, GC percent, chromatin states, histone marks, contacts between chromatin regions, expression, genetic variability, etc.). Data can be retrieved in text format via special Table Browser program
Ensembl Genome Browser	https://www.ensembl.org/index.html	Data is integrated based on a graphical interface that allows visualizing genome sequences along with a large number of annotations and features (positions of transcripts, GC percent, chromatin states, genetic variability, etc.). Tables of Ensembl data can be downloaded via the highly customizable BioMart data mining tool
GEO (Gene Expression Omnibus)	https://www.ncbi.nlm.nih.gov/gds	The largest public repository that archives and freely distributes comprehensive sets of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community

* GENCODE reference gene annotations for the human and mouse genomes are also available through the UCSC Genome Browser (https://genome.ucsc.edu/) and the Ensembl genome browser (https://www.ensembl.org/index.html).

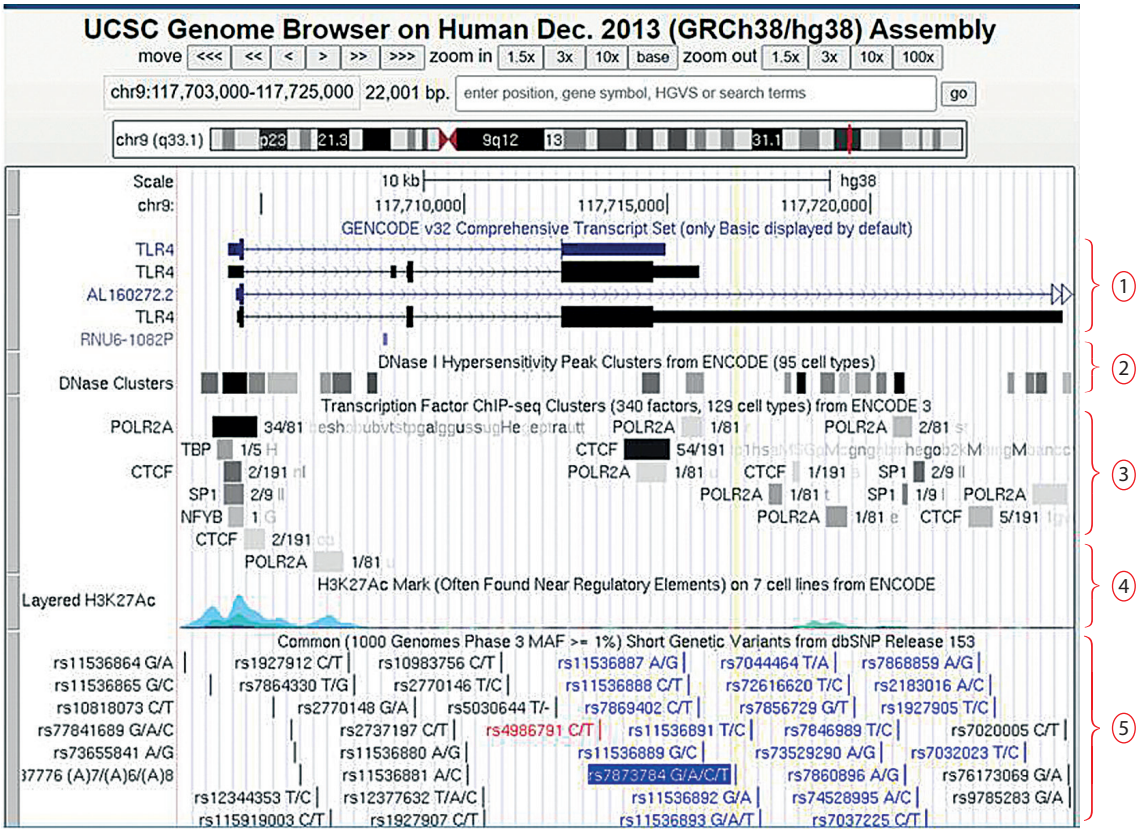

Fig. 5. The view of the human genomic region (chromosomal coordinates chr9: 117,703,000–117,725,000) displayed by the Genome Browser of the University of California, Santa Cruz, USA (UCSC Genome Browser, https://genome.ucsc.edu/).

(1) transcripts of the *TLR4* gene, displayed according to the GENCODE v32 release; (2) DNase I hypersensitivity peak clusters derived from assays in 95 cell types (as a part of the ENCODE project); (3) transcription factor binding derived from a large collection of ChIP-seq experiments performed by the ENCODE project; (4) levels of enrichment of the H3K27Ac histone mark across the genome as determined by a ChIP-seq assay on 7 cell lines from ENCODE (H3K27Ac is the acetylation of lysine 27 of the H3 histone protein, and it is often found near regulatory elements); (5) short genetic variants from dbSNP release 153. The yellow vertical line marks the position the SNP rs7873784 located in the 3′-UTR of *TLR4* gene and associated with development of rheumatoid arthritis and type 2 diabetes (see Table 1). According to (Korneev et al., 2020), the G⟶C substitution at the rs7873784 locus creates PU.1 binding site, that increases the activity of the enhancer located in the 3′-UTR of the *TLR4* gene.

AlleleSeq); (8) the effects of genetic variability on TFBSs predicted *in silico* by computer programs (HaploReg, SNP2TFBS, rSNPBase, rVarBase).

A separate category of information resources includes: (1) the genome browser of the University of California, Santa Cruz, USA (UCSC Genome Browser, https://genome.ucsc.edu/) and (2) the genome browser of the Ensembl database which is a joint research project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute (Ensembl Genome Browser, https://www.ensembl.org/index.html). These genome browsers integrate data on genome sequences and its features obtained by different research groups using a wide range of experimental methods (Lee et al., 2020; Yates et al., 2020). The websites of these browsers provide access to the primary DNA sequences and genome annotations for many organisms (including vertebrates and several other model species). Browser's graphical interfaces allow to obtain scalable maps of genomic regions and to visualize interactively a large number of annotations and features (for example, positions of transcripts, positions of GVs, chromatin regions interacting with TFs detected by ChIP-seq experiments, data on genome-wide mapping of DNase I hypersensitive sites, etc.) (Fig. 5).

The websites of the UCSC Genome Browser and Ensembl Genome Browser provide access to software tools for extraction data as text files: UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) and BioMart data mining tool (https://www.ensembl.org/info/data/biomart/index.html).

Information resources on allele-specific binding of transcription factors and on the effects of genetic variants on TFBSs predicted *in silico*

As noted above, the influence of pathogenic GVs on gene expression is often mediated through a change in the functional activity of TFBSs. In this regard, information resources that include whole genome data on allele-specific binding of TFs, identified based on the ChIP-seq method, can be extremely useful. A range of approaches have been developed to identify allele-specific binding of TFs (Rozowsky et al., 2011; Reddy et al., 2012; Waszak et al., 2014; Younesy et

al., 2014). These approaches are based on the analysis of the ChIP-seq data in combination with the sequencing data, which allow to find heterozygous loci within a single genome and to phase genotypes of the studied cells. Thus, for each type of cells examined, its own set of genomic loci interacting with a specific transcription factor in an allele-specific manner can be identified. For example, in (Cavalli et al., 2016a), the ChIP-seq data for 55 TFs in the HepG2 cells and 57 TFs in the HeLa-S3 cells were analyzed. In HepG2 cells, 3001 genomic loci with allele-specific signals were found, and 712 loci were found in HeLa-S3 cells. The authors note the pronounced tissue-specific nature of allele-specific TF binding: of the entire set of identified loci, only 34 were found in both cell lines (Cavalli et al., 2016a).

The data on allele-specific binding of TFs are collected in the following information resources: AlleleDB (http://alleledb.gersteinlab.org/) (Chen J. et al., 2016), AlleleSeq (http://alleleseq.gersteinlab.org/) (Rozowsky et al., 2011) (see Table 2), as well as in the supplemental files to publications (Cavalli et al., 2016a, b, 2019; Shi et al., 2016).

Studies aimed at identifying allele-specific TF binding made it possible to estimate the number of genetic variants that affect the binding of a particular transcription factor to DNA in a particular cell type. The average number of such events registered for a single transcription factor can range from 19 to 37 for cells with a normal karyotype (GM12878, H1-hESC) and from 12 to 55 for cancer cell lines (SK-N-SH, K562) (Cavalli et al., 2016a, b).

When generating hypotheses on the mechanisms that mediate the effect of GVs on disease risk, one can also use the data on the effects of genetic variants on the functional activity of TFBSs predicted *in silico.* Such information is accumulated in specialized databases: HaploReg (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward, Kellis, 2012), SNP2TFBS (http://ccg.vital-it.ch/snp2tfbs/) (Kumar S. et al., 2017), rSNPBase (http://rsnp3.psych.ac.cn/index.do) (Guo, Wang, 2018), rVarBase (http://rv.psych.ac.cn) (see Table 2).

Conclusion

A significant proportion of pathogenic genetic variants associated with diseases are located in non-coding regions of the human genome. Such genetic variants can with a high degree of probability disrupt functional activity of regulatory regions that control the transcriptional activity of genes. The examples of the mechanisms of influence of pathogenic genetic variants on gene expression considered in this review confirm this possibility. The studies that have made it possible to identify these mechanisms are complex and are based on the analysis of big heterogeneous genetic data. The online omics data resources provide ample opportunities for such research. Further development of experimental techniques and bioinformatics methods for analyzing the data obtained with the help of this techniques, as well as an increase in the set of investigated cell types, will significantly expand these possibilities.

References

Angeloni A., Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem*. 2019;63(6):707-715. DOI 10.1042/EBC20190030.

Beck T., Shorter T., Brookes A.J. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res*. 2020;48(D1):D933-D940. DOI 10.1093/nar/gkz895.

Belokopytova P., Fishman V. Predicting genome architecture: challenges and solutions. *Front. Genet*. 2021. DOI 10.3389/fgene.2020.617202.

Belokopytova P.S., Nuriddinov M.A., Mozheiko E.A., Fishman D., Fishman V. Quantitative prediction of enhancer-promoter interactions. *Genome Res*. 2020;30(1):72-84. DOI 10.1101/gr.249367.119.

Benton M.C., Lea R.A., Macartney-Coxson D., Sutherland H.G., White N., Kennedy D., Mengersen K., Haupt L.M., Griffiths L.R. Genome-wide allele-specific methylation is enriched at gene regulatory regions in a multi-generation pedigree from the Norfolk Island isolate. *Epigenetics Chromatin*. 2019;12(1):60. DOI 10.1186/s13072-019-0304-7.

Cavalli M., Baltzer N., Umer H.M., Grau J., Lemnian I., Pan G., Wallerman O., Spalinskas R., Sahlén P., Grosse I., Komorowski J., Wadelius C. Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci. Rep.* 2019;9(1):2695. DOI 10.1038/s41598-019-39633-0.

Cavalli M., Pan G., Nord H., Wallén Arzt E., Wallerman O., Wadelius C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics.* 2016a;107(6):248-254. DOI 10.1016/j.ygeno.2016.04.006.

Cavalli M., Pan G., Nord H., Wallerman O., Wallén Arzt E., Berggren O., Elvers I., Eloranta M.L., Rönnblom L., Lindblad Toh K., Wadelius C. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* 2016b;135(5):485-497. DOI 10.1007/s00439-016-1654-x.

Chen C.-Y., Chang I.-S., Hsiung C.A., Wasserman W.W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics*. 2014;7:34. DOI 10.1186/1755-8794-7-34.

Chen J., Rozowsky J., Galeev T.R., Harmanci A., Kitchen R., Bedford J., Abyzov A., Kong Y., Regan L., Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.* 2016;18(7):11101. DOI 10.1038/ncomms11101.

Chen L., Liang Y., Qiu J., Zhang L., Chen X., Luo X., Jiang J. Significance of rs1271572 in the estrogen receptor beta gene promoter and its correlation with breast cancer in a southwestern Chinese population. *J. Biomed. Sci*. 2013;20:32. DOI 10.1186/1423-0127-20-32.

Claussnitzer M., Dankel S.N., Kim K.-H., Quon G., Meuleman W., Haugen C., Glunk V., Sousa I.S., Beaudry J.L., Puviindran V., Abdennur N.A., Liu J., Svensson P.-A., Hsu Y.-H., Drucker D.J., Mellgren G., Hui C.-Ch., Hauner H., Kellis M. *FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med*. 2015;373:895-907. DOI 10.1056/NEJMoa1502214.

Cong Z., Li Q., Yang Y., Guo X., Cui L., You T. The SNP of rs6854845 suppresses transcription via the DNA looping structure alteration of super-enhancer in colon cells. *Biochem. Biophys. Res*. 2019;514:734-741. DOI 10.1016/j.bbrc.2019.04.190.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. DOI 10.1038/nature11247.

Farh K.K.-H., Marson A., Zhu J., Kleinewietfeld M., Housley W.J., Beik S., Shoresh N., Whitton H., Ryan R.J.H., Shishkin A.A., Hatan M., Carrasco-Alfonso M.J., Mayer D., Luckey C.J., Patsopoulos N.A., De Jager P.L., Kuchroo V.K., Epstein C.B., Daly M.J., Hafler D.A., Bernstein B.E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343. DOI 10.1038/nature13835.

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*. 2018;83(4):393-401. DOI 10.1134/S0006297918040107.

Gorbacheva A.M., Korneev K.V., Kuprash D.V., Mitkin N.A. The risk G allele of the single-nucleotide polymorphism rs928413 creates a

CREB1-binding site that activates *IL33* promoter in lung epithelial cells. *Int. J. Mol. Sci*. 2018;19(10):2911. DOI 10.3390/ijms1910 2911.

Guo L., Wang J. rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res*. 2018;46(D1):D1111-D1116. DOI 10.1093/nar/gkx1101.

Hansen A.S., Cattoglio C., Darzacq X., Tjian R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus.* 2018; 9(1):20-32. DOI 10.1080/19491034.2017.1389365.

Howard T.D., Mathias R.A., Seeds M.C., Herrington D.M., Hixson J.E., Shimmin L.C., Hawkins G.A., Sellers M., Ainsworth H.C., Sergeant S., Miller L.R., Chilton F.H. DNA methylation in an enhancer region of the FADS cluster is associated with FADS activity in human liver. *PLoS One*. 2014;9(5):e97510. DOI 10.1371/journal.pone.0097510.

Ibrahim D.M., Mundlos S. Three-dimensional chromatin in disease: what holds us together and what drives us apart? *Curr. Opin. Cell Biol*. 2020;64:1-9. DOI 10.1016/j.ceb.2020.01.003.

Izzi B., Pistoni M., Cludts K., Akkor P., Lambrechts D., Verfaillie C., Verhamme P., Freson K., Hoylaerts M.F. Allele-specific DNA methylation reinforces *PEAR1* enhancer activity. *Blood.* 2016;128: 1003-1012. DOI 10.1182/blood-2015-11-682153.

Jones P.L., Veenstra G.J., Wade P.A., Vermaak D., Kass S.U., Landsberger N., Strouboulis J., Wolffe A.P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet*. 1998;19:187-191. DOI 10.1038/561.

Kilpinen H., Waszak S.M., Gschwind A.R., Raghav S.K., Witwicki R.M., Orioli A., Migliavacca E., Wiederkehr M., Gutierrez-Arcelus M., Panousis N., Yurovsky A., Lappalainen T., Romano-Palumbo L., Planchon A., Bielser D., Bryois J., Padioleau I., Udin G., Thurnheer S., Hacker D., Core L.J., Lis J.T., Hernandez N., Reymond A., Deplancke B., Dermitzakis E.T. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342:744-747. DOI 10.1126/science.1242463.

Korneev K.V., Sviriaeva E.N., Mitkin N.A., Gorbacheva A.M., Uvarova A.N., Ustiugova A.S., Polanovsky O.L., Kulakovskiy I.V., Afanasyeva M.A., Schwartz A.M., Kuprash D.V. Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression. *Biochim. Biophys. Acta Mol. Basis Dis.* 2020;1866(3):165626. DOI 10.1016/j.bbadis.2019.165626.

Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.

Kumar D., Puan K.J., Andiappan A.K., Lee B., Westerlaken G.H., Haase D., Melchiotti R., Li Z., Yusof N., Lum J., Koh G., Foo S., Yeong J., Alves A.C., Pekkanen J., Sun L.D., Irwanto A., Fairfax B.P., Naranbhai V., Common J.E., Tang M., Chuang C.K., Jarvelin M.R., Knight J.C., Zhang X., Chew F.T., Prabhakar S., Jianjun L., Wang Y., Zolezzi F., Poidinger M., Lane E.B., Meyaard L., Rötzschke O. A functional SNP associated with atopic dermatitis controls cell type-specific methylation of the *VSTM1* gene locus. *Genome Med*. 2017;9(1):18. DOI 10.1186/s13073-017-0404-6.

Kumar S., Ambrosini G., Bucher P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res*. 2017;45(D1):D139-D144. DOI 10.1093/nar/gkw1064.

Lee C.M., Barber G.P., Casper J., Clawson H., Diekhans M., Gonzalez J.N., Hinrichs A.S., Lee B.T., Nassar L.R., Powell C.C., Raney B.J., Rosenbloom K.R., Schmelter D., Speir M.L., Zweig A.S., Haussler D., Haeussler M., Kuhn R.M., Kent W.J. UCSC Genome Browser enters 20th year. *Nucleic Acids Res*. 2020;48(D1):D756-D761. DOI 10.1093/nar/gkz1012.

Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.

Lewinsky R.H., Jensen T.G.K., Møller J., Stensballe A., Olsen J., Troelsen J.T. T_{-13910} DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity *in vitro*. *Hum. Mol. Genet*. 2005;14(24):3945-3953. DOI 10.1093/hmg/ddi418.

Li S., Li Y., Li X., Liu J., Huo Y., Wang J., Liu Z., Li M., Luo X.-J. Regulatory mechanisms of major depressive disorder risk variants. *Mol. Psychiatry*. 2020;25(9):1926-1945. DOI 10.1038/s41380-020-0715-7.

Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Klopocki E., Horn D., Kayserili H., Opitz J.M., Laxova R., Santos-Simarro F., Gilbert-Dussardier B., Wittler L., Borschiwer M., Haas S.A., Osterwalder M., Franke M., Timmermann B., Hecht J., Spielmann M., Visel A., Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell.* 2015;161(5):1012-1025. DOI 10.1016/j.cell.2015.04.004.

Mathelier A., Shi W., Wasserman W.W. Identification of altered cis-regulatory elements in human disease. *Trends Genet*. 2015;31(2): 67-76. DOI 10.1016/j.tig.2014.12.003.

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., Sandstrom R., Qu H., Brody J., Shafer A., Neri F., Lee K., Kutyavin T., Stehling-Sun S., Johnson A.K., Canfield T.K., Giste E., Diegel M., Bates D., Hansen R.S., Neph S., Sabo P.J., Heimfeld S., Raubitschek A., Ziegler S., Cotsapas C., Sotoodehnia N., Glass I., Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. DOI 10.1126/science.1222794.

McVicker G., van de Geijn B., Degner J.F., Cain C.E., Banovich N.E., Raj A., Lewellen N., Myrthil M., Gilad Y., Pritchard J.K. Identification of genetic variants that affect histone modifications in human cells. *Science.* 2013;342:747-749. DOI 10.1126/science.1242429.

Meddens C., van der List A.C.J., Nieuwenhuis E.E.S., Mokry M. Non-coding DNA in IBD: from sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut*. 2019;68(5):928-941. DOI 10.1136/gutjnl-2018-317516.

Mei S., Ke J., Tian J., Ying P., Yang N., Wang X., Zou D., Peng X., Yang Y., Zhu Y., Gong Y., Zhong R., Chang J., Miao X. A functional variant in the boundary of a topological association domain is associated with pancreatic cancer risk. *Mol. Carcinog*. 2019;58(10): 1855-1862. DOI 10.1002/mc.23077.

Merkulov V.M., Leberfarb E.Y., Merkulova T.I. Regulatory SNPs and their widespread effects on the transcriptome. *J. Biosci*. 2018;43(5): 1069-1075. DOI 10.1007/s12038-018-9817-7.

Nan X., Ng H.H., Johnson C.A., Laherty C.D., Turner B.M., Eisenman R.N., Bird A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*. 1998;393:386-389. DOI 10.1038/30764.

Park C.-Y., Halevy T., Lee D.R., Sung J.J., Lee J.S., Yanuka O., Benvenisty N., Kim D.-W. Reversion of *FMR1* methylation and silencing by editing the triplet repeats in fragile X iPSC-derived neurons. *Cell. Rep*. 2015;13(2):234-241. DOI 10.1016/j.celrep.2015.08.084.

Quenneville S., Verde G., Corsinotti A., Kapopoulou A., Jakobsson J., Offner S., Baglivo I., Pedone P.V., Grimaldi G., Riccio A., Trono D. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell*. 2011;44(3):361-372. DOI 10.1016/j.molcel.2011.08.032.

Rahbar E., Waits C.M.K., Kirby E.H., Jr., Miller L.R., Ainsworth H.C., Cui T., Sergeant S., Howard T.D., Langefeld C.D., Chilton F.H. Allele-specific methylation in the *FADS* genomic region in DNA from human saliva, CD4+ cells, and total leukocytes. *Clin. Epigenetics*. 2018;10:46. DOI 10.1186/s13148-018-0480-5.

Reddy T.E., Gertz J., Pauli F., Kucera K.S., Varley K.E., Newberry K.M., Marinov G.K., Mortazavi A., Williams B.A., Song L., Crawford G.E., Wold B., Willard H.F., Myers R.M. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res*. 2012;22(5):860-869. DOI 10.1101/gr.131201.111.

Roadmap Epigenomics Consortium, Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M.J., … Hirst M., Meissner A., Milosavljevic A., Ren B., Stamatoyannopoulos J.A., Wang T., Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015; 518(7539):317-330. DOI 10.1038/nature14248.

Rozowsky J., Abyzov A., Wang J., Alves P., Raha D., Harmanci A., Leng J., Bjornson R., Kong Y., Kitabayashi N., Bhardwaj N., Rubin M., Snyder M., Gerstein M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol*. 2011;7:522. DOI 10.1038/msb.2011.54.

Schmitz R.J., Lewis Z.A., Goll M.G. DNA methylation: shared and divergent features across eukaryotes. *Trends Genet*. 2019;35(11): 818-827. DOI 10.1016/j.tig.2019.07.007.

Shi W., Fornes O., Mathelier A., Wasserman W.W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* 2016;44(21):10106-10116. DOI 10.1093/nar/gkw691.

Smith A.J.P., Deloukas P., Munroe P.B. Emerging applications of genome-editing technology to examine functionality of GWAS-associated variants for complex traits. *Physiol. Genomics*. 2018;50(7): 510-522. DOI 10.1152/physiolgenomics.00028.2018.

Sun J.H., Zhou L., Emerson D.J., Phyo S.A., Titus K.R., Gong W., Gilgenast T.G., Beagan J.A., Davidson B.L., Tassone F., Phillips-Cremins J.E. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell*. 2018;175(1):224-238. DOI 10.1016/j.cell.2018.08.005.

Visser M., Palstra R.J., Kayser M. Allele-specific transcriptional regulation of *IRF4* in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the *IRF4* promoter. *Hum. Mol. Genet*. 2015;24(9):2649-2661. DOI 10.1093/hmg/ddv029.

Vohra M., Sharma A.R., Prabhu B.N., Rai P.S. SNPs in sites for DNA methylation, transcription factor binding, and miRNA targets leading to allele-specific gene expression and contributing to complex disease risk: a systematic review. *Public Health Genomics*. 2020;23: 1-16. DOI 10.1159/000510253.

Wang H., Lou D., Wang Z. Crosstalk of genetic variants, allele-specific DNA methylation, and environmental factors for complex disease risk. *Front. Genet*. 2019;9:695. DOI 10.3389/fgene.2018.00695.

Ward L.D., Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930-D934. DOI 10.1093/nar/gkr917.

Waszak S.M., Kilpinen H., Gschwind A.R., Orioli A., Raghav S.K., Witwicki R.M., Migliavacca E., Yurovsky A., Lappalainen T., Hernandez N., Reymond A., Dermitzakis E.T., Deplancke B. Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics*. 2014;30(2):165-171. DOI 10.1093/bioinformatics/btt667.

Wingender E., Schoeps T., Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res*. 2013;41(D1):D165-D170. DOI 10.1093/nar/gks1123.

Yates A.D., Achuthan P., Akanni W., Allen J., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Azov A.G., Bennett R., Bhai J., … Perry E., Ruffier M., Trevanion S.J., Cunningham F., Howe K.L., Zerbino D.R., Flicek P. Ensembl 2020. *Nucleic Acids Res*. 2020; 48(D1):D682-D688. DOI 10.1093/nar/gkz966.

Younesy H., Möller T., Heravi-Moussavi A., Cheng J.B., Costello J.F., Lorincz M.C., Karimi M.M., Jones S.J.M. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics*. 2014;30(8): 1172-1174. DOI 10.1093/bioinformatics/btt744.

Zhang Y., Manjunath M., Zhang S., Chasman D., Roy S., Song J.S. Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. *Cancer Res*. 2018;78(7):1579-1591. DOI 10.1158/0008-5472.CAN-17-3486.

Zhao T., Hu Y., Zang T., Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet*. 2019;10:1021. DOI 10.3389/fgene.2019.01021.

ORCID ID
E.V. Ignatieva orcid.org/0000-0002-8588-6511

Acknowledgements. The study was supported from the funds of the budget project No. 0259-2021-0009.
Conflict of interest. The authors declare no conflict of interest.
Received December 28, 2020. Revised January 18, 2021. Accepted January 18, 2021.