Вавиловский журнал генетики и селекции. 2023;27(6):662-675
DOI 10.18699/VJGB-23-77

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

Human-genome single nucleotide polymorphisms affecting transcription factor binding and their role in pathogenesis

E.V. Antontseva ✉, A.O. Degtyareva, E.E. Korbolina, I.S. Damarov, T.I. Merkulova

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

Abstract. Single nucleotide polymorphisms (SNPs) are the most common type of variation in the human genome. The vast majority of SNPs identified in the human genome do not have any effect on the phenotype; however, some can lead to changes in the function of a gene or the level of its expression. Most SNPs associated with certain traits or pathologies are mapped to regulatory regions of the genome and affect gene expression by changing transcription factor binding sites. In recent decades, substantial effort has been invested in searching for such regulatory SNPs (rSNPs) and understanding the mechanisms by which they lead to phenotypic differences, primarily to individual differences in susceptibility to diseases and in sensitivity to drugs. The development of the NGS (next-generation sequencing) technology has contributed not only to the identification of a huge number of SNPs and to the search for their association (genome-wide association studies, GWASs) with certain diseases or phenotypic manifestations, but also to the development of more productive approaches to their functional annotation. It should be noted that the presence of an association does not allow one to identify a functional, truly disease-associated DNA sequence variant among multiple marker SNPs that are detected due to linkage disequilibrium. Moreover, determination of associations of genetic variants with a disease does not provide information about the functionality of these variants, which is necessary to elucidate the molecular mechanisms of the development of pathology and to design effective methods for its treatment and prevention. In this regard, the functional analysis of SNPs annotated in the GWAS catalog, both at the genome-wide level and at the level of individual SNPs, became especially relevant in recent years. A genome-wide search for potential rSNPs is possible without any prior knowledge of their association with a trait. Thus, mapping expression quantitative trait loci (eQTLs) makes it possible to identify an SNP for which – among transcriptomes of homozygotes and heterozygotes for its various alleles – there are differences in the expression level of certain genes, which can be located at various distances from the SNP. To predict rSNPs, approaches based on searches for allele-specific events in RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, MPRA, and other data are also used. Nonetheless, for a more complete functional annotation of such rSNPs, it is necessary to establish their association with a trait, in particular, with a predisposition to a certain pathology or sensitivity to drugs. Thus, approaches to finding SNPs important for the development of a trait can be categorized into two groups: (1) starting from data on an association of SNPs with a certain trait, (2) starting from the determination of allele-specific changes at the molecular level (in a transcriptome or regulome). Only comprehensive use of strategically different approaches can considerably enrich our knowledge about the role of genetic determinants in the molecular mechanisms of trait formation, including predisposition to multifactorial diseases.
Key words: regulatory single-nucleotide polymorphism; transcription factor-binding sites; gene expression; genome-wide studies.

For citation: Antontseva E.V., Degtyareva A.O., Korbolina E.E., Damarov I.S., Merkulova T.I. Human-genome single nucleotide polymorphisms affecting transcription factor binding and their role in pathogenesis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(6):662-675. DOI 10.18699/VJGB-23-77

Однонуклеотидные замены в геноме человека, влияющие на связывание факторов транскрипции, и их роль в развитии патологий

Е.В. Антонцева ✉, А.О. Дегтярева, Е.Е. Корболина, И.С. Дамаров, Т.И. Меркулова

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

Аннотация. Однонуклеотидные замены, также называемые однонуклеотидными полиморфизмами (single nucleotide polymorphism, SNP), – это наиболее распространенный тип вариаций генома человека. Подавляющая часть выявленных в геноме человека SNP не оказывает какого-либо воздействия на молекулярный фенотип, однако некоторые способны приводить к изменению функции гена или уровня его экспрессии. В то же время большинство SNP, ассоциированных с некими признаками или патологиями, картируются в регуляторных областях генома, изменяют потенциальные сайты связывания транскрипционных факторов и, соответственно, могут влияют на экспрессию генов. В последние десятилетия значительные усилия были направлены на поиск таких регуляторных SNP (rSNP), а так-

же на понимание механизмов, посредством которых они приводят к фенотипическим различиям, в первую очередь к разной предрасположенности к заболеваниям и индивидуальной чувствительности к лекарственным препаратам. Развитие технологии NGS (next generation sequencing) способствовало не только выявлению огромного количества SNP и поиску их ассоциации (genome wide association studies, GWAS) с некими заболеваниями или фенотипическими проявлениями, но и развитию более производительных подходов для их функциональной аннотации. Стоит отметить, что наличие ассоциации не позволяет выделить функциональный, действительно связанный с болезнью вариант последовательности ДНК из множества маркерных, которые выявляются за счет неравновесия по сцеплению. Более того, установление ассоциаций генетических вариантов с заболеванием не дает сведений о функциональности этих вариантов, что необходимо для выяснения молекулярных механизмов развития патологии и разработки эффективных методов ее лечения и профилактики. В связи с этим функциональный анализ SNP, аннотированных в GWAS каталоге, как на полногеномном уровне, так и на уровне отдельных SNP в последние годы стал особенно актуальным. В настоящее время активно развивается полногеномный поиск потенциально регуляторных SNP без каких-либо предварительных знаний об их ассоциации с признаком. Так, картирование локусов количественных признаков экспрессии (eQTL, expression quantitative trait loci) позволяет выявить SNP, для которого в транскриптомах гомозигот по разным его аллелям, а также гетерозигот наблюдаются различия в уровне экспрессии неких генов, причем как близко расположенных, так и на значительном удалении. Для предсказания регуляторных SNP используют также подходы, основанные на поиске аллель-специфических событий в данных RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, MPRA и т. д. Однако для более полной характеристики таких rSNP необходимо устанавливать их ассоциацию с признаком, в частности с предрасположенностью к некой патологии или с чувствительностью к лекарственным препаратам. Таким образом, именно комплексное использование двух основанных на противоположных принципах подходов к поиску значимых для развития признака (патологии) SNP: с одной стороны, исходящего из данных по ассоциации SNP с неким признаком, а с другой стороны, идущего от определения аллель-специфичных изменений на молекулярном уровне (в транскриптоме или регуломе) – существенно обогащает картину наших знаний о роли генетических детерминант в молекулярных механизмах формирования признаков, включая предрасположенность к многофакторным заболеваниям.

Ключевые слова: регуляторный однонуклеотидный полиморфизм; сайты связывания транскрипционных факторов; экспрессия генов; полногеномные исследования.

Introduction

One of the main tasks of human genetics is to clarify the mechanisms by which genome variations lead to phenotypic differences, primarily to individual differences in susceptibility to diseases and in sensitivity to drugs. Single-nucleotide polymorphisms (SNPs) are the most common type of genome variation (Chanock, 2001). Currently, due to the development of next-generation sequencing (NGS) technologies, more than 950 million SNPs of the human genome are registered in database dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi) (Sherry et al., 2001); furthermore, rare SNPs with a frequency (prevalence in the population) of less than 1 % constitute more than 90 % of the total number of SNPs. It seems unlikely that the vast majority of identified variations can be important for the phenotype, but some of them certainly form the genetic basis of phenotypic traits, including predisposition to various diseases. In recent decades, major efforts have been applied to the search for such SNPs.

The most popular approach (which started back in the 1980s) to the identification of trait-related SNPs has been the determination of associations with diseases for SNPs found in candidate genes (Lander, Schork, 1994; Ring, Kroetz, 2002), and in this context, only SNPs affecting the protein-coding part of a gene have been investigated (Cooper, 1998). Somewhat later, there has been some interest in elucidating the functionality of variants located in noncoding regions of genes, i. e., in determining an effect of such variants on a certain molecular phenotype. In particular, it has been shown that such SNPs affect transcription factor-binding sites (TFBSs), thereby leading to changes in the expression of the respective genes (Ludlow et al., 1996; Piedrafita et al., 1996; Knight et al., 1999; Vasiliev et al., 1999). Nonetheless, these few functional studies have remained almost invisible against the backdrop of a huge wave of research aimed at identifying associations.

The productivity of detection of disease-associated SNPs increased dramatically with the advent of the genome-wide association study (GWAS) technology in the mid-2000s (Fig. 1, *a*), which is based on an unbiased – not based on any ideas about the formation of a trait – principle of a genome-wide search for SNPs associated with the trait (Visscher et al., 2012; Tam et al., 2019). To date, more than 72 thousand associations of genetic variants with traits have been found by GWASs (GWAS Catalog, https://www.ebi.ac.uk/gwas/), thus allowing to find many new genes and systems of genes associated with predisposition to various diseases (Buniello et al., 2019; Tam et al., 2019; Claussnitzer et al., 2020).

Nevertheless, GWAS technology does not provide any information about the functionality of the detected variants, thereby making it very difficult to elucidate the molecular mechanisms underlying the development of a pathology and hence to develop effective methods for its treatment and prevention. Additionally, based on results of GWASs, it is almost impossible to distinguish a truly disease-associated variant from the many marker variants that are detected due to linkage disequilibrium (Lappalainen, 2015; Tam et al., 2019; Zhao et al., 2020). It is also known that most SNPs identified by GWASs are located in the noncoding part of the genome and, as a rule, in its regulatory regions (e. g., promoters and enhancers) (Hindorff et al., 2009; Maurano et al., 2012; Bryzgalov et al., 2013; Farh et al., 2015); these data imply an influence of such SNPs on the binding of transcription factors (TFs) and on gene expression. Thus, a need for research on functional interpretation of data from GWASs – both at

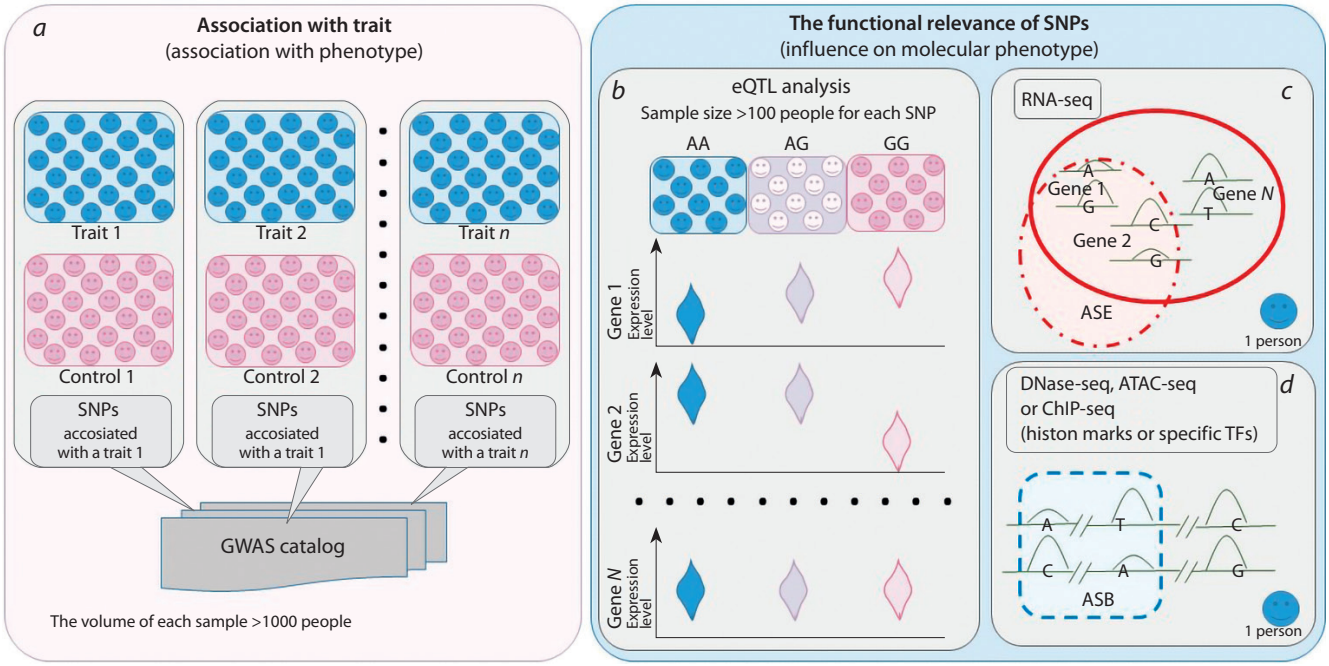

Fig. 1. Genome-wide approaches to the analysis of SNPs.

a, The principle of GWASs; *b*, the scheme of eQTL analysis for each SNP. A search for allele-specific events: *c*, expression (allele-specific expression, ASE) and *d*, binding (allele-specific binding, ASB).

the level of individual potentially regulatory SNPs (rSNPs) and at the level of all such variants collectively – has become obvious. In addition, because the GWAS technology greatly underestimates the actual number of associations owing to the required strict thresholds in statistical processing (Tam et al., 2019), investigators have recognized the need to develop GWAS-unrelated large-scale function-based approaches to the search for rSNPs (Westra, Franke, 2014; Maurano et al., 2015; Cavalli et al., 2016b; Korbolina et al., 2018).

Of the genome-wide function-based approaches, the first in timing of emergence (almost simultaneous with the start of the application of the GWAS technology) is analysis of expression quantitative trait loci (eQTLs) (see Fig. 1, *b*). This analysis, by means of transcriptome data (earlier, microarray data have been employed, then data from high-throughput RNA sequencing [RNA-seq] started to be used), determines for each SNP a difference in the level of expression of individual genes among homozygotes and heterozygotes for different alleles of this SNP (Westra, Franke, 2014; GTEx Consortium, 2020). Somewhat later, techniques were devised for finding allele-specific events both in data from RNA-seq (allele-specific expression [ASE] events) (see Fig. 1, *c*) (Castel et al., 2020; Fan et al., 2020) and in data from ChIP-seq, DNase-seq, and ATAC-seq (allele-specific binding [ASB] events) (see Fig. 1, *d*) (Maurano et al., 2015; Cavalli et al., 2016a, b; Xu et al., 2020; Korbolina et al., 2021). Massive function-based approaches also include massively parallel reporter assay (MPRA), SNP-seq, and SNP-SELEX (Zhang et al., 2018; Lu et al., 2021; Yan et al., 2021).

At the beginning of this review, using the latest studies as examples, we examine rSNPs that are (i) associated with the development of pathologies according to results of GWASs and are (ii) characterized in detail in terms of their influence on an interaction with a TF and on the expression of nearby or distant genes. Next, the review describes the application of functional genomics methods to interpretation of data from GWASs and to the search for new regulatory variants without GWASs. In the course of the presentation, strengths and weaknesses of these approaches are shown, as is the importance of comprehensive use of GWASs and functional genomics methods to reveal the role of SNPs in molecular aberrations underlying the development of a pathology.

Functional interpretation of data from GWASs at the level of individual rSNPs

A wide range of experimental methods are utilized to functionally study individual possible rSNPs. At initial stages, classical methods are usually used: analysis of DNA-probe retardation in a gel by nuclear extract proteins (electrophoretic mobility shift assay, EMSA) and a reporter assay (analysis of reporter gene expression under the control of allelic variants of the SNP region), which allow a researcher to detect an influence of an SNP on the binding of some TF or on reporter gene expression, respectively (Antontseva et al., 2015; Fang et al., 2017). To identify the TFs the binding sites of which are affected by nucleotide substitution, scientists perform EMSA using appropriate antibodies or purified TFs (Piedrafita et al., 1996; Knight et al., 1999; Vasiliev et al., 1999; Jiang et al., 2020), chromatin immunoprecipitation with detection of allele asymmetry in a PCR product (ChIP-AS-qPCR) (Gao et al., 2018; Choi et al., 2020; Thynn et al., 2020; Protze et al., 2022), and mass-spectrometric analysis of proteins isolated from complexes with oligonucleotides containing minor alleles (Fang et al., 2017; Liu D. et al., 2018; Choi et al., 2020).

An impact of a nucleotide substitution in the identified TFBSs on their potential target genes is confirmed by experiments with downregulation or artificial upregulation of genes of the corresponding TFs or by studying the effect of point mutations introduced into the TFBS using CRISPR/Cas9 technology (Prestel et al., 2019; Gutierrez-Arcelus et al., 2020; Pan et al., 2020; Thynn et al., 2020; Wang Y. et al., 2020; Wang X. et al., 2021). Lately, an increasingly popular approach in this field of research on SNPs has also been the determination of allelic imbalance of the expression for SNPs located in transcribed regions of genes. Nonetheless, it is worth noting that such an SNP can be either an rSNP proper (Syddall et al., 2013; Klein et al., 2019) or a marker SNP in a linkage group with an rSNP located in a nontranscribed region (Fang et al., 2017; Li X.-X. et al., 2019; Peng et al., 2020).

Various combinations of these techniques are employed in modern research, as illustrated by the examples below.

rs36115365

A textbook example of a well-studied rSNP is rs36115365 (G/C), which is situated at a locus associated with various types of cancer according to data from GWASs (chr5p15.33: region 2). At this locus, a correlation analysis has revealed nine SNPs ($r^2 > 0.60$, 1000G EUR population) associated with pancreatic cancer, testicular germ cell tumor, lung cancer, and melanoma. To screen all nine SNPs for regulatory activity, EMSA and reporter assay were performed by means of eight human cell lines (Fang et al., 2017). The use of several cell lines is a common practice in such studies (Bryzgalov et al., 2013; Boldes et al., 2020) and is aimed at the highest possible coverage of events of interaction of TFs with their binding sites; the reason is substantial differences in the sets of TFs expressed in different cell types (Tobias et al., 2021). The screening analyses identified only SNP rs36115365 as potentially regulatory, and its C allele showed both much better binding to a certain protein in the EMSA and greater activation of reporter gene expression as compared to the G allele (Fang et al., 2017).

rs36115365 is located ~18 kilobase pairs (kbp) upstream of the start of the *TERT* gene (encoding reverse transcriptase of the telomerase complex) and approximately 5 kbp downstream of the end of the *CLPTM1L* gene. According to the results of ChIP-seq (chromatin immunoprecipitation followed by sequencing of DNA from the precipitates) from project ENCODE (Moore et al., 2020), this SNP's location overlaps with many TFBSs, and this region is enriched with active chromatin histone marks, which is typical for enhancer regions. Inactivation of this region by small-interfering-RNA-mediated transcriptional silencing (Malecová, Morris, 2010) results in downregulation of only the *TERT* gene. To identify the TF the binding site of which changes as a consequence of the substitution of G with C, the binding of proteins from nuclear extracts to oligonucleotides containing minor alleles was implemented, followed by mass-spectrometric analysis of the bound proteins. After an analysis of the obtained peptides, four TF candidates that prefer the C allele were proposed: ZNF148, VEZF1/ZNF161, ZNF281, and ZNF740. In EMSA involving specific antibodies to these TFs, only ZNF148 was confirmed, which was subsequently verified by means of the purified ZNF148 protein in an experiment. A small-interfering-RNA-mediated knockdown of ZNF148 gave a definitive answer because it reduced *TERT* expression and telomerase activity and shortened telomere length. Thus, the C allele corrects the binding site of ZNF148, enhances the expression of *TERT* and, as a consequence, increases the risk of carcinogenesis (Fang et al., 2017).

rs174575

According to GWASs, rs174575 (C/G) correlates with an elevated risk of colorectal cancer (Tian et al., 2020). This SNP is located in the first intron of the delta-6-desaturase gene (*FADS2*) at a distance of +41.5 kbp from its transcription start site and at a distance of –178.8 kbp from the transcription start site of the gene of long noncoding RNA AP002754.2. The rs174575 region in chromatin is enriched with histone modifications characteristic of active regulatory regions (H3K4me1, H3K4me3, or H3K27ac), and judging by DNase-seq data (identification of sites of hypersensitivity to DNase I) and findings of ATAC-seq (assay for transposase-accessible chromatin), corresponds to open chromatin (ENCODE). The hypothesis of a regulatory role of rs174575 is supported by eQTL analysis data, which show an association of the G allele with overexpression of *FADS2* and AP002754.2.

Computer analysis of DNA motifs using Web services Cistrome (Zheng et al., 2019) and JASPAR (Fornes et al., 2020) has revealed that in the case of rs174575, the replacement of G with C damages the binding site of TF E2F1, and the data of ChIP-seq from ENCODE, obtained on colorectal cancer LoVo cells, indicate that E2F1 is mapped to the location of this SNP. A cross-competitive EMSA has confirmed better binding of a certain nuclear extract protein to the G allele, and that this protein is E2F1 has been demonstrated by ChIP-qPCR with appropriate antibodies. For instance, in cell lines with different genotypes – HCT116 (CG), SNU-C1 (CG), and HT115 (CC) – a stronger binding of E2F1 is observed in cells carrying the G allele. A study involving reporter constructs has also confirmed a higher enhancer activity of a DNA fragment containing the G allele (Tian et al., 2020).

Direct contact between the region containing rs174575 and promoters of *FADS2* and AP002754.2 has been detected by the chromosome conformation capture (3C) method, and it turned out that the interaction was much more pronounced in cell lines carrying the G allele. Further experiments indicated that overexpression of AP002754.2 sharply raises the level of *FADS2* expression in HCT116 and LoVo cells, and a knockdown of AP002754.2 by microRNA causes a decrease in the expression of this gene, indicating a stimulatory role of AP002754.2 in the regulation of the *FADS2* gene. On the other hand, overexpression of *FADS2* or AP002754.2 significantly increases the proliferation rate of HCT116 and LoVo cells, whereas a knockdown of *FADS2* or AP002754.2 significantly reduces it. It has also been shown that overexpression of *FADS2* and AP002754.2 accelerates tumor growth in *in vivo* experiments in mice. It is known that the product of the *FADS2* gene is a key enzyme in the biosynthesis of polyunsaturated fatty acids, including arachidonic acid, which in turn is a precursor of prostaglandin E2 (PGE2), which promotes tumor growth and metastasis. Thus, rs174575 acts as an allele-specific enhancer that stimulates the transcription of *FADS2* and AP002754.2, leading to an elevated risk of colorectal cancer in the case of the G allele (Tian et al., 2020).

rs4903064

According to GWASs, rs4903064 (T/C) is associated with renal cell carcinoma, the most common type of kidney cancer (Scelo et al., 2017). rs4903064 is located in the third intron of the *DPF3* gene, which encodes a protein of the BAF subfamily of the SWI/SNF chromatin-remodeling complex. The predominance of the C allele of rs4903064 in tumor tissue samples from patients with clear-cell renal carcinoma has been demonstrated, in contrast to both normal kidney tissues and tumor tissue from individuals with papillary renal cell carcinoma and individuals with chromophobe renal cell carcinoma. In this context, a measurement of the ratio of alleles in *DPF3* pre-mRNA in patients with clear-cell renal carcinoma heterozygous for rs4903064 has confirmed a skew toward the C allele (Protze et al., 2022).

According to ATAC-seq data obtained on primary renal cancer cells, rs4903064 is located in an open chromatin region, in a putative enhancer. Given that the substitution of T with C (rs4903064) creates a potential binding site for TF HIF, a reporter analysis has been performed on HeLa and MCF-7 cell lines, showing that an increase in reporter gene expression takes place only in the case of the C allele and only when cells are treated with a specific stabilizer of HIF: dimethyloxalylglycine. Knockouts of various isoforms of HIF have revealed that the increase in reporter activity upon stimulation with dimethyloxalylglycine depends on HIF-1α. By means of ChIP-qPCR in primary renal tubular cells with different genotypes of rs4903064 (TT, CT, and CC), enhanced binding of HIF-1α and HIF-1β to risk allele C has been confirmed. To elucidate the role of overexpression of *DPF3* in the development of clear-cell renal carcinoma, a knockout of *DPF3* has been performed in cells of proximal renal tubules using CRISPR/Cas9 technology. It was demonstrated in that report that cells with defective expression of *DPF3* grow more slowly than control clones of the corresponding cells, suggesting that increased expression of *DPF3* in proximal tubule cells stimulates proliferation (Protze et al., 2022).

rs17114036

rs17114036 (T/C) is associated with coronary heart disease and ischemic stroke according to GWASs (Dichgans et al., 2014). This SNP is located in the 5th intron of the *PLPP3* gene, encoding phospholipid phosphatase 3, which inhibits inflammation of endothelium and contributes to the integrity of its monolayer (Panchatcharam et al., 2014; Wu et al., 2015). Experiments with ATAC-seq and ChIP-seq (H3K27ac and H3K4me2) performed on human aortic endothelial cells (HAECs) have identified the region containing rs17114036 as a potential enhancer. The enhancer activity of this region was confirmed by reporter analysis, whereas protective allele C (as compared to the T allele) significantly increased the activity of luciferase upon transfection of vector constructs into HAECs. A deletion of a 66-bp region containing rs17114036 by means of CRISPR/Cas9 significantly reduced the expression of *PLPP3* as compared to the unedited genome and enhanced the permeability of the monolayer of edited HAECs. As a result of modeling hemodynamic processes, an increase in the activity of the studied enhancer (containing the C allele) in HAECs was shown with an 18-hour "atheroprotective" flow as compared to an "atherogenic" flow, while no effect of the T allele was found. It turned out that the substitution of the T nucleotide with C creates a binding site (CACC) for the KLF2 protein, as confirmed by ChIP-AS-qPCR analysis in HAECs heterozygous for rs17114036. Cotransfection experiments with a plasmid causing overexpression of KLF2 also showed higher luciferase activity in the case of the tested enhancer carrying the C allele (Krause et al., 2018).

rs4407214

rs4407214 (T/G) is associated with estrogen receptor-negative breast cancer. In the HMEC cell line, an analysis of ChIP-seq data from the ENCODE project regarding locations of marks of active chromatin (H3K4m1, H3K4m2, H3K4m3, H3K9ac, H3K27ac, H3K36m3, H3K79m2, H4K20m1, EZH2, and H2AZ) has helped to find in the rs4407214 region a regulatory locus in intron 1 of the *WDR43* gene (a protein-coding gene associated with rRNA processing and ribosomal biogenesis). By EMSA, the researchers showed ASB of nuclear proteins from MCF10A and CAL-51 cells to DNA probes mimicking the region of this SNP's location in the genome. Reporter gene expression under the control of the identified regulatory region on the same cell lines was also found to be allele-dependent (Couch et al., 2016; Fachal et al., 2020).

Bioinformatic analysis was then performed using the JASPAR database and available ChIP-seq data for this region from the ENCODE project to identify the TFs the binding sites of which are altered by the G-to-T substitution (rs4407214). As a result, USF1 was identified as such a TF. A competitive EMSA with nuclear proteins isolated from CAL-51 and MCF10A cells confirmed the ASB of USF1 in the case of the G allele. It was also demonstrated that CRISPR/Cas9-mediated removal of the presumed regulatory region containing rs4407214 results in underexpression of *PLB1* (phospholipase B1), which is located at a distance of ~400 kbp from *WDR43*.

Mass interpretation of GWAS data by functional genomics methods

Because results of GWASs tend to associate a trait (disease) with multiple loci (Goldstein, 2009; Boyle et al., 2017), most of which in turn contain many often coinherited SNPs (Tak, Farnham, 2015; Schaid et al., 2018), it is very difficult to choose among them the potential rSNPs involved in a disease's pathogenesis. To date, several effective experimental solutions to this problem have been developed. First of all, these are scaled up versions of the approaches utilized to investigate individual rSNPs: a reporter assay (MPRA) and methods for studying protein–nucleic acid interactions (Reel-seq, SNP-seq, and SNP-SELEX) (Zhao et al., 2020; Lu et al., 2021; Yan et al., 2021).

In particular, MPRA has been successfully used to find the rSNPs that play a key role in the genetic predisposition to lupus erythematosus (Lu et al., 2021). According to GWAS findings, 3,073 SNPs in 91 loci are associated with this disease. Those researchers constructed a barcoded library containing 12,396 170-bp oligonucleotides containing in the middle all known variants of these 3,073 SNPs; these oligos were inserted upstream of a minimal promoter placed before the *eGFP* gene. An influence of a minor allele on enhancer activity of inserts for 51 SNPs from 27 loci was demonstrated in the GM12878

cell line. The small number of identified rSNPs can most likely be explained by transfection of only one cell line; if several cell lines had been tested, the number of rSNPs would have been much larger due to expansion of the set of TFs involved in the study. Similarly, in MPRA, 30 out of 832 SNPs (associated with melanoma risk according to GWASs) showed a significant difference in the impact of minor alleles on reporter gene expression in the UACC903 melanoma cell line (Choi et al., 2020). For one of these 30 SNPs (rs398206), which is located in intron 1 of the *MX2* gene, the difference was the largest. It turned out that risk allele A of rs398206 significantly enhances the binding of TF YY1 *in vitro* (in EMSA) and *in vivo* (according to ChIP-AS-qPCR), thereby upregulating *MX2* and thus contributing to the initiation of melanoma (Choi et al., 2020). Other examples can be found in refs (Ulirsch et al., 2016; Liu S. et al., 2017; Kalita et al., 2018; Klein et al., 2019).

By methodically similar Reel-seq and SNP-seq, which are based on a comparative analysis of the binding of a TF to oligonucleotides containing minor alleles, 521 potential rSNPs have been selected out of 4,316 SNPs correlating with breast cancer according to GWASs (Zhao et al., 2020) as well as 403 possible rSNPs out of 903 SNPs associated with prostate cancer (Zhang et al., 2018). The largest study involving the approach from this research group was published in 2021 (Yan et al., 2021). The approach was named SNP-SELEX. To implement it, those authors used a library of 383,544 40-bp oligonucleotides containing in the middle all possible alleles of 95,886 SNPs. SNPs were chosen based on either their association with type 2 diabetes mellitus in GWASs or localization within a 500-kbp window containing a variant associated with this pathology. By means of 270 recombinant TFs, those authors conducted a multiplex analysis of their binding to the oligonucleotides and identified 11,079 SNPs having an appreciable allele effect on binding to at least one TF.

For mass interpretation of the results of GWASs, functional genomics data from available databases are also widely used, such as data on genome-wide profiles (ChIP-seq) of TFs' binding (Li S. et al., 2020) and of histone modifications (Jones et al., 2020), on open-chromatin genome-wide profiles (ATAC-seq) (Corces et al., 2020), on three-dimensional chromatin contacts (Corces et al., 2020), and on locations of enhancer and superenhancer regions (Gong et al., 2018; Sun W. et al., 2018; Nasser et al., 2021) as well as data from eQTL analysis (Gamazon et al., 2018; Zheng et al., 2019; Barbeira et al., 2021).

For example, 8,005 SNPs – either directly associated by GWASs with major depressive disorder or present in the same linkage group – have been mapped to ChIP-seq peaks obtained from a brain tissue or cells of neuronal origin by means of antibodies to 34 various TFs (Li S. et al., 2020). After that, a search was performed for binding sites of the corresponding TFs in the regions of the SNPs using a database containing 7,699 position weight matrices (Whitington et al., 2016), and it was revealed that 34 SNPs disrupt the binding sites of 15 TFs. A reporter assay confirmed the effect of an allele on gene expression for 29 SNPs. One of them, rs3101339, proved to be located at a potential binding site of TF REST in the promoter region of the *NEGR1* gene, whereas the substitution of A with C considerably damaged the structure of this site, as evidenced by a decrease in reporter gene expression under the control of an insert carrying the C allele. The influence of rs3101339 on *NEGR1* gene expression *in vivo* was confirmed by elimination of an appropriate DNA fragment via CRISPR-Cas9 genomic editing. Since the product of *NEGR1* plays an important part in the maintenance of required density of dendritic spines, its underexpression when A is replaced by C may substantially contribute to the development of a depressive state (Li S. et al., 2020).

Data on genome-wide profiles of active-chromatin histone marks are also turning out to be very informative for the functional interpretation of GWAS results. For example, to analyze many SNPs correlating with epithelial ovarian cancer, they have been mapped in the region of ChIP-seq peaks for H3K27Ac; these peaks were obtained by the researchers in a study on 26 tissue samples of this type of cancer (Jones et al., 2020). Then, using the motifbreakR tool (Coetzee et al., 2015), among the mapped SNPs, 469 SNPs were selected in which the nucleotide substitution substantially altered a binding site of some TF. The most frequent was the change in the sequence of the binding site of TF REST, for which there are data on its tumor-suppressive and oncogenic functions (Jones et al., 2020). Besides, the use of ChIP-seq datasets on various histone modifications from relevant databases in combination with data and tools from a database of potential regulatory variants (rVarBase) (Guo et al., 2016) has made it possible to detect in superenhancers 286 and 366 possible rSNPs associated with type 2 diabetes mellitus (Sun W. et al., 2018) and coronary heart disease (Gong et al., 2018), respectively, that alter the predicted TFBSs.

GWAS-unrelated function-based approaches to identifying potential rSNPs

Modern massive function-based approaches to the identification of potential rSNPs are mainly based on the registration of an effect of a nucleotide substitution on some molecular phenotype. This may be (i) determination (in transcriptomes) of a difference in the expression level of individual genes among homozygotes and heterozygotes for different alleles of each SNP (eQTL analysis), (ii) identification of SNPs showing asymmetry of enrichment within transcriptome data (RNA-seq: ASE events) or within epigenomic data (DNase-seq, ChIP-seq, and ATAC-seq: ASB events), or (iii) determination of an influence of an allele on reporter gene expression by MPRA.

eQTL analysis

The term "eQTL" either means that there is a correlation between a variant (eVariant) and the expression level of a certain gene(s) (eGene[s]) (GTEx Consortium, 2017, 2020) or refers directly to an SNP the alleles of which show such a correlation; the term is used much more frequently in the latter sense (Fairfax et al., 2014; Fan et al., 2020; Jiang et al., 2020; Werling et al., 2020). Transcriptomic data obtained using either microarrays (Fairfax et al., 2014; Westra, Franke, 2014) or RNA-seq (GTEx Consortium, 2020) are suitable for finding eQTLs. These data are quite sufficient for the detection of eQTLs located in transcribed regions (Göring et al., 2007), whereas the identification of their entire set also requires genome sequencing data (GTEx Consortium, 2020;

Werling et al., 2020). In contrast to GWASs, which require biological samples from many thousands of individuals (Tam et al., 2019), several hundred participants are sufficient for eQTL analysis (Westra et al., 2013; Fairfax et al., 2014; GTEx Consortium, 2020). Nonetheless, just as in GWASs, in eQTL analyses, the problem of distinguishing an SNP that is indeed relevant to the formation of a trait – among marker variants detected through linkage disequilibrium – is still relevant (Zou et al., 2019; Umans et al., 2021).

The largest-scale project on obtaining transcriptomic data and identifying eQTLs is international consortium GTEx, within which RNA-seq data on 15,201 postmortem samples of 49 tissues collected from 838 donors have been collected, allowing to identify 4,278,636 eQTLs associated with changes in the expression of 18,262 and 5,006 genes encoding proteins and long intergenic noncoding RNAs, respectively (GTEx Consortium, 2020).

There are other datasets of eQTLs, including those obtained not on postmortem but on biopsy materials (Fairfax et al., 2014; Stolze et al., 2020). For example, in a study by Stolze et al., in an analysis of transcriptomes (RNA-seq) of the aortic endothelium of 157 donors, the investigators identified thousands of eQTLs not registered in the GTEx Consortium data (Stolze et al., 2020). Fairfax et al. have employed $CD14^+$ monocytes (derived from healthy individuals) treated *in vitro* with either interferon gamma (for 24 h, 367 individuals) or bacterial cell wall lipopolysaccharide (LPS) (2 and 24 h, 261 and 322 individuals, respectively). $CD14^+$ monocytes from 414 people served as controls there. Through microarray RNA profiling and genotyping, 609,704 SNPs (minor allele frequency > 0.04) were examined and 21,516 eQTLs were detected, 24.6 % of which manifested themselves in control cells, 21.6 % of which manifested themselves after 2 h of treatment with LPS, and 25.4 and 28.3 % after 24 h treatment with LPS and IFN-γ, respectively. The results of this work point to an important role of genomic variants in the nature of transcriptomic response to a drug (Fairfax et al., 2014). In conclusion, it should be noted that any dataset of transcriptomic data obtained from hundreds or more individuals can be used to search for eQTLs, as, for example, has been done by us (Korbolina et al., 2021) with the help of data from RNA-seq analysis of postmortem brain samples from 96 individuals (Ramaker et al., 2017).

At present, quantitative expression trait loci analysis is mainly utilized to identify groups of genes participating in trait formation (Hormozdiari et al., 2016; Morrow et al., 2018; Gamazon et al., 2019; Ratnapriya et al., 2019; Jaffe et al., 2020). Additionally, its results are often used to prioritize GWAS-identified SNPs for their subsequent rigorous experimental investigation. An example is rs13239597, which is located in the *TNPO3* gene promoter and associated with lupus erythematosus and multiple sclerosis according to GWASs. eQTL analysis of transcriptomes of lymphoblastoid cell lines derived from 373 individuals has not revealed any impact of rs13239597 on *TNPO3* expression but uncovered a significant association of the A allele of this SNP with overexpression of the *IRF5* gene [which is located 118 kbp away (Thynn et al., 2020)], in agreement with findings of the GTEx Consortium (GTEx Consortium, 2017). Next, an analysis of available Hi-C data showed that *IRF5* is one of 12 genes that are in direct contact with the rs13239597 region. A computational analysis of motifs that potentially change affinity for a TF as a result of a nucleotide substitution (Coetzee et al., 2015) has revealed four such TFs: EVI1, ERF, GATA1, and TAL1. By ChIP-AS-qPCR, it has been demonstrated that EVI1 binds much better to the rs13239597 region in the case of the A allele as compared to the C allele (Thynn et al., 2020). Similar examples can also be found in refs (Roca-Ayats et al., 2019; Jiang et al., 2020; Tian et al., 2020).

A large-scale search for ASE and ASB events

The development of next-generation-sequencing-based methods of transcriptomic analysis (RNA-seq) and epigenomic analysis (ChIP-seq, DNase-seq, and ATAC-seq) has opened up a unique opportunity for quantifying a difference in the enrichment of two alleles (an allele imbalance) of each heterozygous polymorphic site of a diploid organism within the respective dataset (Maurano et al., 2015; Cavalli et al., 2016a, b; Castel et al., 2020; Fan et al., 2020; Xu et al., 2020; Korbolina et al., 2021). An important feature of these approaches to the identification of potential rSNPs – in contrast to eQTL analysis (and even more so in contrast to GWASs investigating SNPs of many individuals in various genomic contexts and living conditions) – is that allele-asymmetric events are recorded for each individual and the backgrounds are identical. This arrangement enables researchers to obtain reliable data when studying a very small number of individuals, down to one (Harvey et al., 2015). An increase in sample size is required only for involving in the analysis a larger number of SNPs that are in a heterozygous state. For instance, calculations show that data from 20 individuals theoretically allow to determine ASE or ASB events for 65–70 % of SNPs that have a population frequency of ≥5 % (Cavalli et al., 2016a).

For this reason, possibilities of pharmacogenetic and pharmacogenomic projects become much more abundant. The most striking example of such a project is a simultaneous analysis of allele-specific effects of 50 substances (steroid and peptide hormones, nutrients, commonly used drugs, and a number of environmental pollutants) in primary cultures of five cell types (LCLs, PBMCs, HUVECs, SMCs, and melanocytes), each of which is represented by cell samples from three individuals (Moyerbrailean et al., 2016). An analysis of the resultant transcriptomic data helped to identify more than 300 SNPs, an imbalance in the enrichment of the alleles of which within transcriptomes emerged or increased significantly in response to treatment with one or another drug. Via the same approach, inducer (LPS of the bacterial cell wall)-dependent ASE events have been identified in 19 immune response genes by an analysis of transcriptomes of blood mononuclear cells from eight individuals (Edsgärd et al., 2016) as well as 561 ASE events responsive to treatment of $CD4^+$ T cells (from 24 genotyped individuals) with immobilized anti-CD3/CD28 antibodies (Gutierrez-Arcelus et al., 2020). These results open up a new – unrelated to any *a priori* hypothesis – way to elucidate mechanisms of individual sensitivity to drugs.

The largest dataset of ASE events at present, containing 431 million such events, is derived from data on RNA-seq and whole-genome sequencing from the GTEx Consortium (GTEx Consortium, 2020) and is published in a paper by Castel et

al. (Castel et al., 2020). This dataset can serve as a source for mass discovery of rSNPs, for example, in a comparison with data from GWASs or from eQTL analysis. It is worth noting that in the absence of whole-genome sequencing data, relevant information can be acquired by more sophisticated methods of bioinformatic search for allele-specific events directly in RNA-seq data (Harvey et al., 2015; Moyerbrailean et al., 2016; Fan et al., 2020; Korbolina et al., 2021).

ChIP-seq experiments based on antibodies to various TFs make it possible to directly register events of allele-asymmetric interaction of these proteins with their binding sites in the case of a heterozygous state of SNPs at these sites. In a pioneering work in the laboratory of Claus Wadelius, they analyzed all the then-available data from the ENCODE project on binding profiles of TFs in cell lines GM12878 (B cells), H1-hESC, K562, and SK-N-SH, thereby revealing 9,962 SNPs featuring an allelic imbalance in the binding of a TF (ASB) (Cavalli et al., 2016b). By the same approach, 3,713 SNPs have been found showing an allelic imbalance in the binding of TFs in HepG2 and HeLa-S3 cells; testing 39 of them in a luciferase reporter system has confirmed the effect of an allele on reporter gene expression for 27 SNPs (Cavalli et al., 2016a). A detailed analysis of one of them, rs953413, indicates that the A allele disrupts the binding site of TF FOXA, resulting in reduced binding of not only this TF but also of TF HNF4α cooperatively interacting with it, thereby ultimately leading to underexpression of *ELOVL2* and possibly serving as a factor in the pathogenesis of non-alcoholic fatty liver disease (Pan et al., 2020).

Because allele-asymmetric alterations in profiles of histone modifications and of open chromatin can reflect SNP-induced changes in the binding of TFs (Kar et al., 2014; Hatayama, Aruga, 2018; Huang et al., 2018; Yi et al., 2020), these data are also widely utilized to find ASB events. For instance, Maurano et al. (Maurano et al., 2015) have examined 493 open-chromatin profiles (DNase-seq) obtained in various cell lines, where 64,599 SNPs with ASB have been found. Their bioinformatic analysis using position weight matrices for 2,203 motifs of TFBSs from different sources indicates that most of the identified SNPs can affect the binding of TFs and hence the accessibility of the respective DNA regions to DNase I (Maurano et al., 2015). A newer technique for detecting open chromatin, ATAC-seq, is based on the ability of a hyperactive mutant of Tn5 transposase to detect open DNA regions in chromatin (Marinov, Shipony, 2021) and is also used to search for ASB events. In particular, it has been employed to identify 53 rSNPs in breast cancer MCF-7 cells and 125 rSNPs in a line of human mesenchymal stem cells (MSCs); in total, 30 % of the rSNPs found in MCF-7 cells and 43 % of those found in MSCs have been identified as eQTLs in GTEx data, indicating their influence on gene expression (Xu et al., 2020). Examples of use of data from ChIP-seq (involving antibodies to histone marks) for registering ASB events can be found in refs (Sun J. et al., 2016; D'Oliveira Albanus et al., 2021; Li M. et al., 2021).

The combination of searches for ASE events and ASB events can be considered the most productive approach to identifying rSNPs. For example, in our work (Korbolina et al., 2018), at first, ASB events were identified in ChIP-seq data from the ENCODE project for histone modifications H3K27ac, H3K4me1, H3K4me2, H3K4me3, and H3K27me3 as well as for 456 TFs and their associated proteins in human cell lines K562, MCF-7, and HCT-116. Then, by means of RNA-seq data obtained from the same cell lines, SNPs (rSNPs) that are associated with changes in gene expression levels were identified. According to GWASs, out of 1,633 rSNPs found in this way, 27 have shown associations with cancers (Korbolina et al., 2018), and 14, with cognitive disorders (Bryzgalov et al., 2018). Another 30 rSNPs have been implicated in colorectal cancer with the help of data from the International Cancer Genome Consortium (ICGC) (Seshagiri et al., 2012). Genotyping of patients with colorectal cancer and healthy individuals for six of these SNPs has revealed an association of rs590352, rs4796672, and rs2072580 with this disease (Leberfarb et al., 2020). A correlation with breast cancer has been found for rs2072580 (Degtyareva et al., 2020). Later, the same approach has allowed to identify 14,266 rSNPs during the processing of data obtained in a study by Reyes-Palomares et al. (Reyes-Palomares et al., 2020) on H3K4me3 histone marker profiling (ChIP-seq) and RNA-seq data on pulmonary-artery epithelial samples from 19 individuals (Korbolina et al., 2021).

MPRA

Research into the effect of polymorphic-site alleles on reporter gene expression via simultaneous transfection of hundreds and thousands of barcoded plasmid constructs into eukaryotic cells with subsequent transcriptome sequencing is also an informative approach to finding rSNPs (Vockley et al., 2015; Tewhey et al., 2016; Movva et al., 2019). The largest-scale study based on this approach has been conducted in the laboratory of Bas van Steensel (van Arensbergen et al., 2019). Using a promoterless plasmid and fragmented genomes (fragment length 150–500 bp) of four individuals belonging to different ethnic groups, two barcoded libraries were constructed for each individual, where inserts were expected to play the role of a promoter. At the same time, on the basis of data on transcription initiation in enhancer regions (Natoli, Andrau, 2012; van Arensbergen et al., 2017), those authors expected to detect not only promoters but also enhancers. The use of DNA from humans of genetically distant ethnic groups allowed those investigators to hope for an analysis of the largest possible number of polymorphic sites that are homozygous for different alleles in at least two of those people.

After transfection of K562 and HepG2 cells with the resulting libraries, 19 and 14 thousand potential rSNPs, respectively, were found, most of which did not overlap, once again indicating tissue specificity of the supragenomic (protein) regulatory machine. The identified SNPs showed significant enrichment within regulatory regions of the genome. In this case, the enrichment (approximately 15-fold) was three times higher in promoter regions than in enhancer regions (approximately 5-fold); this outcome is obviously due to the design of the reporter constructs. For several rSNPs, by mass-spectrometric analysis of proteins interacting with oligonucleotides containing minor alleles, those authors were able to identify TFs the binding sites of which are altered by a nucleotide substitution. In particular, the A allele of rs623853 was found to disrupt the binding of TFs of the ELF family, whereas the C allele of rs554591 weakens the binding of ZNF787 while enhancing the binding of KLF and SP (van Arensbergen et al., 2019).

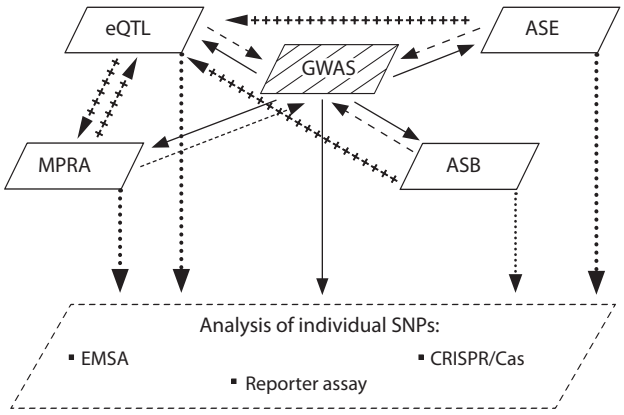

Fig. 2. The scheme of an integrative approach to the search for SNPs functionally important for the development of a trait (pathology) on the basis of two opposite principles: from an association with a trait to its function (solid arrows) and vice versa, from a function to a trait (dashed arrows). Cross-containing arrows show ways to increase the evidence base for functional significance of sets of SNPs, and dotted arrows represent ways to study individual SNPs in detail.

Conclusion

Programs – of coordinated switching on, switching off, and changes of the expression of various genes – that underlie (1) ontogenesis events, (2) the existence of many types of differentiated cells, and (3) the abilities of cells to respond to various factors of the external and internal environment are implemented by the regulatory part of the genome of multicellular organisms. The information encoded in the regulatory regions is converted into a desired pattern of gene expression primarily via the binding of TFs to specific sequences in the regulatory regions (e. g., promoters, enhancers, and silencers) (Lan et al., 2012; Merkulova et al., 2013; Dubois-Chevalier et al., 2018; Chen, Pugh, 2021; Tobias et al., 2021). According to present-day concepts, the SNPs located in regulatory regions of genes, affecting binding sites of TFs, and changing the level of gene expression play a central part in the variation of phenotypic traits, including predisposition/resistance to multifactorial diseases (Maurano et al., 2015; Deplancke et al., 2016; Carrasco Pro et al., 2020). In this regard, there is a strong interest both in functional interpretation of SNPs having an association with various diseases (primarily according to GWAS data) and in the development of massive function-based approaches to the discovery of rSNPs. Interpretation of data from GWASs is carried out either at the level of individual SNPs or for all SNPs collectively by a variety of functional genomics techniques (Fig. 2). Meanwhile, the same methods of functional genomics are used for personal searches for rSNPs, but at the same time, it is necessary to solve the inverse problem: determining a relation between the found rSNPs and a trait (disease). The most popular solution to this problem is to compare the obtained data with available information from GWASs (see Fig. 2). On the other hand, in this way, usually only 1.5–3.0 % of found rSNPs are implicated in various traits (Cavalli et al., 2016b, 2019; Korbolina et al., 2021). In this regard, it seems very promising to take advantage of results of eQTL analysis, enabling an investigator to determine an influence of many specific rSNPs on the expression of fairly large groups of genes, the subsequent analysis of which by modern functional annotation tools (Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and others) allows to get an idea about a possible affected trait (Korbolina et al., 2021).

To sum up, there are two approaches – based on opposite principles – to finding SNPs that are important for the development of a trait (pathology): on the one hand, starting from data on an association of an SNP with some trait, and on the other hand, starting from determination of allele-specific changes at the molecular level (in the transcriptome or regulome). It can be concluded that comprehensive use of the two approaches appreciably enriches our knowledge about the participation of genetic determinants in molecular mechanisms of trait formation, including predisposition to multifactorial diseases.

References

Antontseva E.V., Matveeva M.Y., Bondar N.P., Kashina E.V., Leberfarb E.Y., Bryzgalov L.O., Gervas P.A., Ponomareva A.A., Cherdyntseva N.V., Orlov Y.L., Merkulova T.I. Regulatory single nucleotide polymorphisms at the beginning of intron 2 of the human *KRAS* gene. *J. Biosci.* 2015;40(5):873-883. DOI 10.1007/s12038-015-9567-8.

Barbeira A.N., Bonazzola R., Gamazon E.R., Liang Y., Park Y., Kim-Hellmuth S., Wang G., Jiang Z., Zhou D., Hormozdiari F., Liu B., Rao A., Hamel A.R., Pividori M.D., Aguet F., Bastarache L., Jordan D.M., Verbanck M., Do R., Stephens M., Ardlie K., McCarthy M., Montgomery S.B., Segrè A.V., Brown C.D., Lappalainen T., Wen X., Im H.K. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* 2021;22(1):49. DOI 10.1186/s13059-020-02252-4.

Boldes T., Merenbakh-Lamin K., Journo S., Shachar E., Lipson D., Yeheskel A., Pasmanik-Chor M., Rubinek T., Wolf I. R269C variant of ESR1: high prevalence and differential function in a subset of pancreatic cancers. *BMC Cancer*. 2020;20(1):531. DOI 10.1186/s12885-020-07005-x.

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

Bryzgalov L.O., Antontseva E.V., Matveeva M.Y., Shilov A.G., Kashina E.V., Mordvinov V.A., Merkulova T.I. Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. *PLoS One*. 2013;8(10):e78833. DOI 10.1371/journal.pone.0078833.

Bryzgalov L.O., Korbolina E.E., Brusentsov I.I., Leberfarb E.Y., Bondar N.P., Merkulova T.I. Novel functional variants at the GWAS-implicated loci might confer risk to major depressive disorder, bipolar affective disorder and schizophrenia. *BMC Neurosci.* 2018;19(S1):22. DOI 10.1186/s12868-018-0414-3.

Buniello A., MacArthur J.A.L., Cerezo M., Harris L.W., Hayhurst J., Malangone C., McMahon A., Morales J., Mountjoy E., Sollis E., Suveges D., Vrousgou O., Whetzel P.L., Amode R., Guillen J.A., Riat H.S., Trevanion S.J., Hall P., Junkins H., Flicek P., Burdett T., Hindorff L.A., Cunningham F., Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019;47(D1):D1005-D1012. DOI 10.1093/nar/gky1120.

Carrasco Pro S., Bulekova K., Gregor B., Labadorf A., Fuxman Bass J.I. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. *Sci. Rep.* 2020;10(1):17632. DOI 10.1038/s41598-020-74793-4.

Castel S.E., Aguet F., Mohammadi P., Aguet F., Anand S., Ardlie K.G., Gabriel S., Getz G.A., Graubert A., Hadley K., … Moore H.M., Nierras C.R., Rao A.K., Vaught J.B., Volpi S., Ardlie K.G., Lappalainen T. A vast resource of allelic expression data spanning human tissues. *Genome Biol.* 2020;21(1):234. DOI 10.1186/s13059-020-02122-z.

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.

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.

Chanock S. Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis. Markers*. 2001;17(2): 89-98. DOI 10.1155/2001/858760.

Chen H., Pugh B.F. What do transcription factors interact with? *J. Mol. Biol.* 2021;433(14):166883. DOI 10.1016/j.jmb.2021.166883.

Choi J., Zhang T., Vu A., Ablain J., Makowski M.M., Colli L.M., Xu M., Hennessey R.C., Yin J., Rothschild H., Gräwe C., Kovacs M.A., Funderburk K.M., Brossard M., Taylor J., Pasaniuc B., Chari R., Chanock S.J., Hoggart C.J., Demenais F., Barrett J.H., Law M.H., Iles M.M., Yu K., Vermeulen M., Zon L.I., Brown K.M. Massively parallel reporter assays of melanoma risk variants identify *MX2* as a gene promoting melanoma. *Nat. Commun.* 2020;11(1):2718. DOI 10.1038/s41467-020-16590-1.

Claussnitzer M., Cho J.H., Collins R., Cox N.J., Dermitzakis E.T., Hurles M.E., Kathiresan S., Kenny E.E., Lindgren C.M., MacArthur D.G., North K.N., Plon S.E., Rehm H.L., Risch N., Rotimi C.N., Shendure J., Soranzo N., McCarthy M.I. A brief history of human disease genetics. *Nature*. 2020;577(7789):179-189. DOI 10.1038/s41586-019-1879-7.

Coetzee S.G., Coetzee G.A., Hazelett D.J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*. 2015;31(23):3847-3849. DOI 10.1093/bioinformatics/btv470.

Cooper D. The human gene mutation database. *Nucleic Acids Res.* 1998;26(1):285-287. DOI 10.1093/nar/26.1.285.

Corces M.R., Shcherbina A., Kundu S., Gloudemans M.J., Frésard L., Granja J.M., Louie B.H., Eulalio T., Shams S., Bagdatli S.T., Mumbach M.R., Liu B., Montine K.S., Greenleaf W.J., Kundaje A., Montgomery S.B., Chang H.Y., Montine T.J. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* 2020;52(11): 1158-1168. DOI 10.1038/s41588-020-00721-x.

Couch F.J., Kuchenbaecker K.B., Michailidou K., Mendoza-Fandino G.A., Nord S., Lilyquist J., Olswold C., Hallberg E., Agata S., Ahsan H., … Slager S., Chenevix-Trench G., Pharoah P.D.P., Monteiro A.A.N., García-Closas M., Easton D.F., Antoniou A.C. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nat. Commun.* 2016;7(1):11375. DOI 10.1038/ncomms11375.

Degtyareva A.O., Leberfarb E.Y., Efimova E.G., Brusentsov I.I., Usova A.V., Lushnikova E.L., Merkulova T.I. *rs2072580T>A* polymorphism in the overlapping promoter regions of the *SART3* and *ISCU* genes associated with the risk of breast cancer. *Bull. Exp. Biol. Med.* 2020;169(1):81-84. DOI 10.1007/s10517-020-04829-2.

Deplancke B., Alpern D., Gardeux V. The genetics of transcription factor DNA binding variation. *Cell*. 2016;166(3):538-554. DOI 10.1016/j.cell.2016.07.012.

Dichgans M., Malik R., König I.R., Rosand J., Clarke R., Gretarsdottir S., Thorleifsson G., Mitchell B.D., Assimes T.L., Levi C., … Willenborg C., Laaksonen R., Voight B.F., Stewart A.F.R., Rader D.J., Hall A.S., Kooner J.S. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*. 2014;45(1):24-36. DOI 10.1161/STROKEAHA.113.002707.

D'Oliveira Albanus R., Kyono Y., Hensley J., Varshney A., Orchard P., Kitzman J.O., Parker S.C.J. Chromatin information content landscapes inform transcription factor and DNA interactions. *Nat. Commun.* 2021;12(1):1307. DOI 10.1038/s41467-021-21534-4.

Dubois-Chevalier J., Mazrooei P., Lupien M., Staels B., Lefebvre P., Eeckhoute J. Organizing combinatorial transcription factor recruitment at *cis*-regulatory modules. *Transcription*. 2018;9(4):233-239. DOI 10.1080/21541264.2017.1394424.

Edsgärd D., Iglesias M.J., Reilly S.-J., Hamsten A., Tornvall P., Odeberg J., Emanuelsson O. GeneiASE: detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci. Rep.* 2016;6(1):21134. DOI 10.1038/srep21134.

Fachal L., Aschard H., Beesley J., Barnes D.R., Allen J., Kar S., Pooley K.A., Dennis J., Michailidou K., Turman C., … Edwards S.L., Antoniou A.C., Chenevix-Trench G., Simard J., Easton D.F., Kraft P., Dunning A.M. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* 2020; 52(1):56-73. DOI 10.1038/s41588-019-0537-1.

Fairfax B.P., Humburg P., Makino S., Naranbhai V., Wong D., Lau E., Jostins L., Plant K., Andrews R., McGee C., Knight J.C. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science.* 2014;343(6175):1246949. DOI 10.1126/science.1246949.

Fan J., Hu J., Xue C., Zhang H., Susztak K., Reilly M.P., Xiao R., Li M. ASEP: gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.* 2020; 16(5):e1008786. DOI 10.1371/journal.pgen.1008786.

Fang J., Jia J., Makowski M., Xu M., Wang Z., Zhang T., Hoskins J.W., Choi J., Han Y., Zhang M., … Amos C.I., Iles M.M., Nathanson K.L., Landi M.T., Vermeulen M., Brown K.M., Amundadottir L.T. Functional characterization of a multi-cancer risk locus on chr5p15.33 reveals regulation of *TERT* by ZNF148. *Nat. Commun.* 2017;8(1):15034. DOI 10.1038/ncomms15034.

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.

Fornes O., Castro-Mondragon J.A., Khan A., van der Lee R., Zhang X., Richmond P.A., Modi B.P., Correard S., Gheorghe M., Baranašić D., Santana-Garcia W., Tan G., Chèneby J., Ballester B., Parcy F., Sandelin A., Lenhard B., Wasserman W.W., Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2020;48(D1):D87-D92. DOI 10.1093/nar/gkz1001.

Gamazon E.R., Segrè A.V., van de Bunt M., Wen X., Xi H.S., Hormozdiari F., Ongen H., Konkashbaev A., Derks E.M., Aguet F., Quan J., Nicolae D.L., Eskin E., Kellis M., Getz G., McCarthy M.I., Dermitzakis E.T., Cox N.J., Ardlie K.G. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* 2018;50(7):956-967. DOI 10.1038/s41588-018-0154-4.

Gamazon E.R., Zwinderman A.H., Cox N.J., Denys D., Derks E.M. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nat. Genet.* 2019;51(6):933-940. DOI 10.1038/s41588-019-0409-8.

Gao P., Xia J.-H., Sipeky C., Dong X.-M., Zhang Q., Yang Y., Zhang P., Cruz S.P., Zhang K., Zhu J., Lee H.-M., Suleman S., Giannareas N., Liu S., Tammela T.L.J., Auvinen A., Wang X., Huang Q., Wang Liguo, Manninen A., Vaarala M.H., Wang Liang, Schleutker J., Wei G.-H. Biology and clinical implications of the 19q13

aggressive prostate cancer susceptibility locus. *Cell*. 2018;174(3): 576-589.e18. DOI 10.1016/j.cell.2018.06.003.

Goldstein D.B. Common genetic variation and human traits. *N. Engl. J. Med.* 2009;360(17):1696-1698. DOI 10.1056/NEJMp0806284.

Gong J., Qiu C., Huang D., Zhang Y., Yu S., Zeng C. Integrative functional analysis of super enhancer SNPs for coronary artery disease. *J. Hum. Genet.* 2018;63(5):627-638. DOI 10.1038/s10038-018-0422-2.

Göring H.H.H., Curran J.E., Johnson M.P., Dyer T.D., Charlesworth J., Cole S.A., Jowett J.B.M., Abraham L.J., Rainwater D.L., Comuzzie A.G., Mahaney M.C., Almasy L., MacCluer J.W., Kissebah A.H., Collier G.R., Moses E.K., Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 2007;39(10):1208-1216. DOI 10.1038/ng2119.

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213. DOI 10.1038/nature 24277.

GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318-1330. DOI 10.1126/science.aaz1776.

Guo L., Du Y., Qu S., Wang J. rVarBase: an updated database for regulatory features of human variants. *Nucleic Acids Res.* 2016;44(D1): D888-D893. DOI 10.1093/nar/gkv1107.

Gutierrez-Arcelus M., Baglaenko Y., Arora J., Hannes S., Luo Y., Amariuta T., Teslovich N., Rao D.A., Ermann J., Jonsson A.H., Navarrete C., Rich S.S., Taylor K.D., Rotter J.I., Gregersen P.K., Esko T., Brenner M.B., Raychaudhuri S. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.* 2020;52(3):247-253. DOI 10.1038/s41588-020-0579-4.

Harvey C.T., Moyerbrailean G.A., Davis G.O., Wen X., Luca F., Pique-Regi R. QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics*. 2015;31(8):1235-1242. DOI 10.1093/bioinformatics/btu802.

Hatayama M., Aruga J. Role of Zic family proteins in transcriptional regulation and chromatin remodeling. In: Aruga J. (Ed.) Zic Family. Advances in Experimental Medicine and Biology. Vol. 1046. Singapore: Springer, 2018;353-380. DOI 10.1007/978-981-10-7311-3_18.

Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA.* 2009;106(23):9362-9367. DOI 10.1073/pnas.0903103106.

Hormozdiari F., van de Bunt M., Segrè A.V., Li X., Joo J.W.J., Bilow M., Sul J.H., Sankararaman S., Pasaniuc B., Eskin E. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 2016;99(6):1245-1260. DOI 10.1016/j.ajhg.2016.10.003.

Huang Q., Ma C., Chen L., Luo D., Chen R., Liang F. Mechanistic insights into the interaction between transcription factors and epigenetic modifications and the contribution to the development of obesity. *Front. Endocrinol.* (*Lausanne*). 2018;9:370. DOI 10.3389/fendo.2018.00370.

Jaffe A.E., Hoeppner D.J., Saito T., Blanpain L., Ukaigwe J., Burke E.E., Collado-Torres L., Tao R., Tajinda K., Maynard K.R., Tran M.N., Martinowich K., Deep-Soboslay A., Shin J.H., Kleinman J.E., Weinberger D.R., Matsumoto M., Hyde T.M. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. *Nat. Neurosci.* 2020;23(4):510-519. DOI 10.1038/s41593-020-0604-z.

Jiang Z., Huang Y., Zhang P., Han C., Lu Y., Mo Z., Zhang Z., Li X., Zhao S., Cai F., Huang L., Chen C., Shi Z., Zhang Y., Ling F. Characterization of a pathogenic variant in GBA for Parkinson's disease with mild cognitive impairment patients. *Mol. Brain*. 2020;13(1): 102. DOI 10.1186/s13041-020-00637-x.

Jones M.R., Peng P.-C., Coetzee S.G., Tyrer J., Reyes A.L.P., Corona R.I., Davis B., Chen S., Dezem F., Seo J.-H., Kar S., Dareng E., Berman B.P., Freedman M.L., Plummer J.T., Lawrenson K., Pharoah P., Hazelett D.J., Gayther S.A. Ovarian cancer risk variants are enriched in histotype-specific enhancers and disrupt transcription factor binding sites. *Am. J. Hum. Genet.* 2020;107(4):622-635. DOI 10.1016/j.ajhg.2020.08.021.

Kalita C.A., Brown C.D., Freiman A., Isherwood J., Wen X., Pique-Regi R., Luca F. High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. *Genome Res.* 2018;28(11):1701-1708. DOI 10.1101/gr.237354.118.

Kar S., Parbin S., Deb M., Shilpi A., Sengupta D., Rath S.K., Rakshit M., Patra A., Patra S.K. Epigenetic choreography of stem cells: the DNA demethylation episode of development. *Cell. Mol. Life Sci.* 2014;71(6):1017-1032. DOI 10.1007/s00018-013-1482-2.

Klein J.C., Keith A., Rice S.J., Shepherd C., Agarwal V., Loughlin J., Shendure J. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* 2019;10(1): 2434. DOI 10.1038/s41467-019-10439-y.

Knight J.C., Udalova I., Hill A.V.S., Greenwood B.M., Peshu N., Marsh K., Kwiatkowski D. A polymorphism that affects OCT-1 binding to the *TNF* promoter region is associated with severe malaria. *Nat. Genet.* 1999;22(2):145-150. DOI 10.1038/9649.

Korbolina E.E., Brusentsov I.I., Bryzgalov L.O., Leberfarb E.Y., Degtyareva A.O., Merkulova T.I. Novel approach to functional SNPs discovery from genome-wide data reveals promising variants for colon cancer risk. *Hum. Mutat.* 2018;39(6):851-859. DOI 10.1002/humu.23425.

Korbolina E.E., Bryzgalov L.O., Ustrokhanova D.Z., Postovalov S.N., Poverin D.V., Damarov I.S., Merkulova T.I. A panel of rSNPs demonstrating allelic asymmetry in both ChIP-seq and RNA-seq data and the search for their phenotypic outcomes through analysis of DEGs. *Int. J. Mol. Sci.* 2021;22(14):7240. DOI 10.3390/ijms22147240.

Krause M.D., Huang R.-T., Wu D., Shentu T.-P., Harrison D.L., Whalen M.B., Stolze L.K., Di Rienzo A., Moskowitz I.P., Civelek M., Romanoski C.E., Fang Y. Genetic variant at coronary artery disease and ischemic stroke locus 1p32.2 regulates endothelial responses to hemodynamics. *Proc. Natl. Acad. Sci. USA.* 2018;115(48):E11349-E11358. DOI 10.1073/pnas.1810568115.

Lan X., Farnham P.J., Jin V.X. Uncovering transcription factor modules using one- and three-dimensional analyses. *J. Biol. Chem.* 2012; 287(37):30914-30921. DOI 10.1074/jbc.R111.309229.

Lander E.S., Schork N.J. Genetic dissection of complex traits. *Science.* 1994;265(5181):2037-2048. DOI 10.1126/science.8091226.

Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* 2015;25(10): 1427-1431. DOI 10.1101/gr.190983.115.

Leberfarb E.Y., Degtyareva A.O., Brusentsov I.I., Maximov V.N., Voevoda M.I., Autenshlus A.I., Morozov D.V., Sokolov A.V., Merkulova T.I. Potential regulatory SNPs in the *ATXN7L3B* and *KRT15* genes are associated with gender-specific colorectal cancer risk. *Per. Med.* 2020;17(1):43-54. DOI 10.2217/pme-2019-0059.

Li M., Huang H., Li L., He C., Zhu L., Guo H., Wang L., Liu Jiafeng, Wu S., Liu Jingxin, Xu T., Mao Z., Cao N., Zhang K., Lan F., Ding J., Yuan J., Liu Y., Ouyang H. Core transcription regulatory circuitry orchestrates corneal epithelial homeostasis. *Nat. Commun.* 2021;12(1):420. DOI 10.1038/s41467-020-20713-z.

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.

Li X.-X., Peng T., Gao J., Feng J.-G., Wu D.-D., Yang T., Zhong L., Fu W.-P., Sun C. Allele-specific expression identified rs2509956 as a novel long-distance *cis*-regulatory SNP for *SCGB1A1*, an important gene for multiple pulmonary diseases. *Am. J. Physiol. Lung. Cell. Mol. Physiol.* 2019;317(4):L456-L463. DOI 10.1152/ajplung.00275.2018.

Liu D., Qin S., Ray B., Kalari K.R., Wang L., Weinshilboum R.M. Single nucleotide polymorphisms (SNPs) distant from xenobiotic response elements can modulate Aryl hydrocarbon receptor function: SNP-dependent CYP1A1 induction. *Drug Metab. Dispos.* 2018;46(9):1372-1381. DOI 10.1124/dmd.118.082164.

Liu S., Liu Y., Zhang Q., Wu J., Liang J., Yu S., Wei G.-H., White K.P., Wang X. Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* 2017;18(1):194. DOI 10.1186/s13059-017-1322-z.

Lu X., Chen X., Forney C., Donmez O., Miller D., Parameswaran S., Hong T., Huang Y., Pujato M., Cazares T., Miraldi E.R., Ray J.P., de Boer C.G., Harley J.B., Weirauch M.T., Kottyan L.C. Global discovery of lupus genetic risk variant allelic enhancer activity. *Nat. Commun.* 2021;12(1):1611. DOI 10.1038/s41467-021-21854-5.

Ludlow L.B., Schick B.P., Budarf M.L., Driscoll D.A., Zackai E.H., Cohen A., Konkle B.A. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting in the Bernard–Soulier syndrome. *J. Biol. Chem.* 1996;271(36):22076-22080. DOI 10.1074/jbc.271.36.22076.

Malecová B., Morris K.V. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. *Curr. Opin. Mol. Ther.* 2010;12(2):214-222.

Marinov G.K., Shipony Z. Interrogating the accessible chromatin landscape of eukaryote genomes using ATAC-seq. In: Shomron N. (Ed.) Deep Sequencing Data Analysis. Methods in Molecular Biology. Vol. 2243. New York: Humana, 2021;183-226. DOI 10.1007/978-1-0716-1103-6_10.

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.

Maurano M.T., Haugen E., Sandstrom R., Vierstra J., Shafer A., Kaul R., Stamatoyannopoulos J.A. Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*. *Nat. Genet.* 2015;47(12):1393-1401. DOI 10.1038/ng.3432.

Merkulova T.I., Ananko E.A., Ignatieva E.V., Kolchanov N.A. Transcription regulatory codes of eukaryotic genomes. *Russ. J. Genet.* 2013;49(1):29-45. DOI 10.1134/S1022795413010079.

Moore J.E., Purcaro M.J., Pratt H.E., Epstein C.B., Shoresh N., Adrian J., Kawli T., Davis C.A., Dobin A., Kaul R., … Snyder M.P., Bernstein B.E., Wold B., Hardison R.C., Gingeras T.R., Stamatoyannopoulos J.A., Weng Z. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*. 2020;583(7818):699-710. DOI 10.1038/s41586-020-2493-4.

Morrow J.D., Cho M.H., Platig J., Zhou X., DeMeo D.L., Qiu W., Celli B., Marchetti N., Criner G.J., Bueno R., Washko G.R., Glass K., Quackenbush J., Silverman E.K., Hersh C.P. Ensemble genomic analysis in human lung tissue identifies novel genes for chronic obstructive pulmonary disease. *Hum. Genomics*. 2018;12(1):1. DOI 10.1186/s40246-018-0132-z.

Movva R., Greenside P., Marinov G.K., Nair S., Shrikumar A., Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*. 2019;14(6):e0218073. DOI 10.1371/journal.pone.0218073.

Moyerbrailean G.A., Richards A.L., Kurtz D., Kalita C.A., Davis G.O., Harvey C.T., Alazizi A., Watza D., Sorokin Y., Hauff N., Zhou X., Wen X., Pique-Regi R., Luca F. High-throughput allele-specific expression across 250 environmental conditions. *Genome Res.* 2016;26(12):1627-1638. DOI 10.1101/gr.209759.116.

Nasser J., Bergman D.T., Fulco C.P., Guckelberger P., Doughty B.R., Patwardhan T.A., Jones T.R., Nguyen T.H., Ulirsch J.C., Lekschas F., Mualim K., Natri H.M., Weeks E.M., Munson G., Kane M., Kang H.Y., Cui A., Ray J.P., Eisenhaure T.M., Collins R.L., Dey K., Pfister H., Price A.L., Epstein C.B., Kundaje A., Xavier R.J., Daly M.J., Huang H., Finucane H.K., Hacohen N., Lander E.S., Engreitz J.M. Genome-wide enhancer maps link risk variants to disease genes. *Nature*. 2021;593(7858):238-243. DOI 10.1038/s41586-021-03446-x.

Natoli G., Andrau J.-C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* 2012;46(1):1-19. DOI 10.1146/annurev-genet-110711-155459.

Pan G., Cavalli M., Carlsson B., Skrtic S., Kumar C., Wadelius C. rs953413 regulates polyunsaturated fatty acid metabolism by modulating *ELOVL2* expression. *iScience*. 2020;23(2):100808. DOI 10.1016/j.isci.2019.100808.

Panchatcharam M., Salous A.K., Brandon J., Miriyala S., Wheeler J., Patil P., Sunkara M., Morris A.J., Escalante-Alcalde D., Smyth S.S. Mice with targeted inactivation of *Ppap2b* in endothelial and hematopoietic cells display enhanced vascular inflammation and permeability. *Arterioscler. Thromb. Vasc. Biol.* 2014;34(4):837-845. DOI 10.1161/ATVBAHA.113.302335.

Peng T., Zhong L., Gao J., Wan Z., Fu W.-P., Sun C. Identification of rs11615992 as a novel regulatory SNP for human *P2RX7* by allele-specific expression. *Mol. Genet. Genomics*. 2020;295(1):23-30. DOI 10.1007/s00438-019-01598-0.

Piedrafita F.J., Molander R.B., Vansant G., Orlova E.A., Pfahl M., Reynolds W.F. An Alu element in the myeloperoxidase promoter contains a composite SP1-thyroid hormone-retinoic acid response element. *J. Biol. Chem.* 1996;271(24):14412-14420. DOI 10.1074/jbc.271.24.14412.

Prestel M., Prell-Schicker C., Webb T., Malik R., Lindner B., Ziesch N., Rex-Haffner M., Röh S., Viturawong T., Lehm M., Mokry M., den Ruijter H., Haitjema S., Asare Y., Söllner F., Najafabadi M.G., Aherrahrou R., Civelek M., Samani N.J., Mann M., Haffner C., Dichgans M. The atherosclerosis risk variant rs2107595 mediates allele-specific transcriptional regulation of *HDAC9* via E2F3 and Rb1. *Stroke*. 2019;50(10):2651-2660. DOI 10.1161/STROKEAHA.119.026112.

Protze J., Naas S., Krüger R., Stöhr C., Kraus A., Grampp S., Wiesener M., Schiffer M., Hartmann A., Wullich B., Schödel J. The renal cancer risk allele at 14q24.2 activates a novel hypoxia-inducible transcription factor-binding enhancer of DPF3 expression. *J. Biol. Chem.* 2022;298(3):101699. DOI 10.1016/j.jbc.2022.101699.

Ramaker R.C., Bowling K.M., Lasseigne B.N., Hagenauer M.H., Hardigan A.A., Davis N.S., Gertz J., Cartagena P.M., Walsh D.M., Vawter M.P., Jones E.G., Schatzberg A.F., Barchas J.D., Watson S.J., Bunney B.G., Akil H., Bunney W.E., Li J.Z., Cooper S.J., Myers R.M. Post-mortem molecular profiling of three psychiatric disorders. *Genome Med.* 2017;9(1):72. DOI 10.1186/s13073-017-0458-5.

Ratnapriya R., Sosina O.A., Starostik M.R., Kwicklis M., Kapphahn R.J., Fritsche L.G., Walton A., Arvanitis M., Gieser L., Pietraszkiewicz A., Montezuma S.R., Chew E.Y., Battle A., Abecasis G.R., Ferrington D.A., Chatterjee N., Swaroop A. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* 2019;51(4):606-610. DOI 10.1038/s41588-019-0351-9.

Reyes-Palomares A., Gu M., Grubert F., Berest I., Sa S., Kasowski M., Arnold C., Shuai M., Srivas R., Miao S., Li D., Snyder M.P., Rabinovitch M., Zaugg J.B. Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. *Nat. Commun.* 2020;11(1):1673. DOI 10.1038/s41467-020-15463-x.

Ring H.Z., Kroetz D.L. Candidate gene approach for pharmacogenetic studies. *Pharmacogenomics*. 2002;3(1):47-56. DOI 10.1517/14622416.3.1.47.

Roca-Ayats N., Martínez-Gil N., Cozar M., Gerousi M., Garcia-Giralt N., Ovejero D., Mellibovsky L., Nogués X., Díez-Pérez A., Grinberg D., Balcells S. Functional characterization of the *C7ORF76* genomic region, a prominent GWAS signal for osteoporosis in 7q21.3. *Bone*. 2019;123:39-47. DOI 10.1016/j.bone.2019.03.014.

Scelo G., Purdue M.P., Brown K.M., Johansson M., Wang Z., Eckel-Passow J.E., Ye Y., Hofmann J.N., Choi J., Foll M., … Deleuze J.-F., McKay J.D., Parker A.S., Wu X., Houlston R.S., Brennan P., Chanock S.J. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* 2017;8(1):15724. DOI 10.1038/ncomms15724.

Schaid D.J., Chen W., Larson N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 2018;19(8):491-504. DOI 10.1038/s41576-018-0016-z.

Seshagiri S., Stawiski E.W., Durinck S., Modrusan Z., Storm E.E., Conboy C.B., Chaudhuri S., Guan Y., Janakiraman V., Jaiswal B.S., Guillory J., Ha C., Dijkgraaf G.J.P., Stinson J., Gnad F., Huntley M.A., Degenhardt J.D., Haverty P.M., Bourgon R., Wang W., Koeppen H., Gentleman R., Starr T.K., Zhang Z., Largaespada D.A., Wu T.D., de Sauvage F.J. Recurrent R-spondin fusions in colon cancer. *Nature*. 2012;488(7413):660-664. DOI 10.1038/nature11282.

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311. DOI 10.1093/nar/29.1.308.

Stolze L.K., Conklin A.C., Whalen M.B., López Rodríguez M., Õunap K., Selvarajan I., Toropainen A., Örd T., Li J., Eshghi A., Solomon A.E., Fang Y., Kaikkonen M.U., Romanoski C.E. Systems genetics in human endothelial cells identifies non-coding variants modifying enhancers, expression, and complex disease traits. *Am. J. Hum. Genet.* 2020;106(6):748-763. DOI 10.1016/j.ajhg.2020.04.008.

Sun J., Zhao Y., McGreal R., Cohen-Tayar Y., Rockowitz S., Wilczek C., Ashery-Padan R., Shechter D., Zheng D., Cvekl A. Pax6 associates with H3K4-specific histone methyltransferases Mll1, Mll2, and Set1a and regulates H3K4 methylation at promoters and enhancers. *Epigenetics Chromatin*. 2016;9(1):37. DOI 10.1186/s13072-016-0087-z.

Sun W., Yao S., Tang J., Liu S., Chen J., Deng D., Zeng C. Integrative analysis of super enhancer SNPs for type 2 diabetes. *PLoS One*. 2018;13(1):e0192105. DOI 10.1371/journal.pone.0192105.

Syddall C.M., Reynard L.N., Young D.A., Loughlin J. The identification of *trans*-acting factors that regulate the expression of *GDF5* via the osteoarthritis susceptibility SNP rs143383. *PLoS Genet.* 2013;9(6):e1003557. DOI 10.1371/journal.pgen.1003557.

Tak Y.G., Farnham P.J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin*. 2015;8:57. DOI 10.1186/s13072-015-0050-4.

Tam V., Patel N., Turcotte M., Bossé Y., Paré G., Meyre D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 2019;20(8):467-484. DOI 10.1038/s41576-019-0127-1.

Tewhey R., Kotliar D., Park D.S., Liu B., Winnicki S., Reilly S.K., Andersen K.G., Mikkelsen T.S., Lander E.S., Schaffner S.F., Sabeti P.C. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*. 2016;165(6):1519-1529. DOI 10.1016/j.cell.2016.04.027.

Thynn H.N., Chen X.-F., Hu W.-X., Duan Y.-Y., Zhu D.-L., Chen H., Wang N.-N., Chen H.-H., Rong Y., Lu B.-J., Yang M., Jiang F., Dong S.-S., Guo Y., Yang T.-L. An allele-specific functional SNP associated with two systemic autoimmune diseases modulates *IRF5* expression by long-range chromatin loop formation. *J. Invest. Dermatol.* 2020;140(2):348-360.e11. DOI 10.1016/j.jid.2019.06.147.

Tian J., Lou J., Cai Y., Rao M., Lu Z., Zhu Y., Zou D., Peng X., Wang H., Zhang M., Niu S., Li Y., Zhong R., Chang J., Miao X. Risk SNP-mediated enhancer–promoter interaction drives colorectal cancer through both *FADS2* and *AP002754.2*. *Cancer Res.* 2020;80(9):1804-1818. DOI 10.1158/0008-5472.CAN-19-2389.

Tobias I.C., Abatti L.E., Moorthy S.D., Mullany S., Taylor T., Khader N., Filice M.A., Mitchell J.A. Transcriptional enhancers: from prediction to functional assessment on a genome-wide scale. *Genome*. 2021;64(4):426-448. DOI 10.1139/gen-2020-0104.

Ulirsch J.C., Nandakumar S.K., Wang L., Giani F.C., Zhang X., Rogov P., Melnikov A., McDonel P., Do R., Mikkelsen T.S., Sankaran V.G. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*. 2016;165(6):1530-1545. DOI 10.1016/j.cell.2016.04.048.

Umans B.D., Battle A., Gilad Y. Where are the disease-associated eQTLs? *Trends Genet.* 2021;37(2):109-124. DOI 10.1016/j.tig.2020.08.009.

van Arensbergen J., FitzPatrick V.D., de Haas M., Pagie L., Sluimer J., Bussemaker H.J., van Steensel B. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* 2017;35(2):145-153. DOI 10.1038/nbt.3754.

van Arensbergen J., Pagie L., FitzPatrick V.D., de Haas M., Baltissen M.P., Comoglio F., van der Weide R.H., Teunissen H., Võsa U., Franke L., de Wit E., Vermeulen M., Bussemaker H.J., van Steensel B. High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* 2019;51(7):1160-1169. DOI 10.1038/s41588-019-0455-2.

Vasiliev G.V., Merkulov V.M., Kobzev V.F., Merkulova T.I., Ponomarenko M.P., Kolchanov N.A. Point mutations within 663–666 bp of intron 6 of the human *TDO2* gene, associated with a number of psychiatric disorders, damage the YY-1 transcription factor binding site. *FEBS Lett.* 1999;462(1-2):85-88. DOI 10.1016/S0014-5793(99)01513-6.

Visscher P.M., Brown M.A., McCarthy M.I., Yang J. Five years of GWAS discovery. *Am. J. Hum. Genet.* 2012;90(1):7-24. DOI 10.1016/j.ajhg.2011.11.029.

Vockley C.M., Guo C., Majoros W.H., Nodzenski M., Scholtens D.M., Hayes M.G., Lowe W.L., Reddy T.E. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* 2015;25(8):1206-1214. DOI 10.1101/gr.190090.115.

Wang X., Hayes J.E., Xu X., Gao X., Mehta D., Lilja H.G., Klein R.J. Validation of prostate cancer risk variants rs10993994 and rs7098889 by CRISPR/Cas9 mediated genome editing. *Gene*. 2021;768:145265. DOI 10.1016/j.gene.2020.145265.

Wang Y., Ma R., Liu B., Kong J., Lin H., Yu X., Wang R., Li L., Gao M., Zhou B., Mohan M., Yu H., Hou Z., Shen H., Qian B. SNP rs17079281 decreases lung cancer risk through creating an YY1-binding site to suppress DCBLD1 expression. *Oncogene*. 2020;39(20):4092-4102. DOI 10.1038/s41388-020-1278-4.

Werling D.M., Pochareddy S., Choi J., An J.-Y., Sheppard B., Peng M., Li Z., Dastmalchi C., Santpere G., Sousa A.M.M., Tebbenkamp A.T.N., Kaur N., Gulden F.O., Breen M.S., Liang L., Gilson M.C., Zhao X., Dong S., Klei L., Cicek A.E., Buxbaum J.D., Adle-Biassette H., Thomas J.-L., Aldinger K.A., O'Day D.R., Glass I.A., Zaitlen N.A., Talkowski M.E., Roeder K., State M.W., Devlin B., Sanders S.J., Sestan N. Whole-genome and RNA sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell Rep.* 2020;31(1):107489. DOI 10.1016/j.celrep.2020.03.053.

Westra H.-J., Franke L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta.* 2014;1842(10):1896-1902. DOI 10.1016/j.bbadis.2014.04.024.

Westra H.-J., Peters M.J., Esko T., Yaghootkar H., Schurmann C., Kettunen J., Christiansen M.W., Fairfax B.P., Schramm K., Powell J.E., … Psaty B.M., Ripatti S., Teumer A., Frayling T.M., Metspalu A., van Meurs J.B.J., Franke L. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 2013;45(10):1238-1243. DOI 10.1038/ng.2756.

Whitington T., Gao P., Song W., Ross-Adams H., Lamb A.D., Yang Y., Svezia I., Klevebring D., Mills I.G., Karlsson R., Halim S., Dun-

ning M.J., Egevad L., Warren A.Y., Neal D.E., Grönberg H., Lindberg J., Wei G.-H., Wiklund F. Gene regulatory mechanisms underpinning prostate cancer susceptibility. *Nat. Genet.* 2016;48(4): 387-397. DOI 10.1038/ng.3523.

Wu C., Huang R.-T., Kuo C.-H., Kumar S., Kim C.W., Lin Y.-C., Chen Y.-J., Birukova A., Birukov K.G., Dulin N.O., Civelek M., Lusis A.J., Loyer X., Tedgui A., Dai G., Jo H., Fang Y. Mechanosensitive PPAP2B regulates endothelial responses to atherorelevant hemodynamic forces. *Circ. Res.* 2015;117(4):e41-e53. DOI 10.1161/CIRCRESAHA.117.306457.

Xu S., Feng W., Lu Z., Yu C.Y., Shao W., Nakshatri H., Reiter J.L., Gao H., Chu X., Wang Y., Liu Y. regSNPs-ASB: a computational framework for identifying allele-specific transcription factor binding from ATAC-seq data. *Front. Bioeng. Biotechnol.* 2020;8:886. DOI 10.3389/fbioe.2020.00886.

Yan J., Qiu Y., Ribeiro dos Santos A.M., Yin Y., Li Y.E., Vinckier N., Nariai N., Benaglio P., Raman A., Li X., Fan S., Chiou J., Chen F., Frazer K.A., Gaulton K.J., Sander M., Taipale J., Ren B. Systematic analysis of binding of transcription factors to noncoding variants. *Nature*. 2021;591(7848):147-151. DOI 10.1038/s41586-021-03211-0.

Yi M., Tan Y., Wang L., Cai J., Li Xiaoling, Zeng Z., Xiong W., Li G., Li Xiayu, Tan P., Xiang B. TP63 links chromatin remodeling and enhancer reprogramming to epidermal differentiation and squamous cell carcinoma development. *Cell. Mol. Life Sci.* 2020;77(21):4325-4346. DOI 10.1007/s00018-020-03539-2.

Zhang P., Xia J.-H., Zhu J., Gao P., Tian Y.-J., Du M., Guo Y.-C., Suleman S., Zhang Q., Kohli M., Tillmans L.S., Thibodeau S.N., French A.J., Cerhan J.R., Wang L.-D., Wei G.-H., Wang L. High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.* 2018;9(1):2022. DOI 10.1038/s41467-018-04451-x.

Zhao Y., Wu D., Jiang D., Zhang X., Wu T., Cui J., Qian M., Zhao J., Oesterreich S., Sun W., Finkel T., Li G. A sequential methodology for the rapid identification and characterization of breast cancer-associated functional SNPs. *Nat. Commun.* 2020;11(1):3340. DOI 10.1038/s41467-020-17159-8.

Zheng R., Wan C., Mei S., Qin Q., Wu Q., Sun H., Chen C.-H., Brown M., Zhang X., Meyer C.A., Liu X.S. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 2019;47(D1):D729-D735. DOI 10.1093/nar/gky1094.

Zou J., Hormozdiari F., Jew B., Castel S.E., Lappalainen T., Ernst J., Sul J.H., Eskin E. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLoS Genet.* 2019;15(12):e1008481. DOI 10.1371/journal.pgen.1008481.

ORCID ID
E.V. Antontseva orcid.org/0000-0002-4214-7153
A.O. Degtyareva orcid.org/0000-0001-8586-2256
E.E. Korbolina orcid.org/0000-0002-7460-5892
I.S. Damarov orcid.org/0000-0002-3883-3054
T.I. Merkulova orcid.org/0000-0002-2707-0127

Acknowledgements. This work was supported by the Russian Science Foundation, grant No. 23-15-00113.
Conflict of interest. The authors declare no conflict of interest.
Received January 17, 2023. Revised March 24, 2023. Accepted March 30, 2023.