GMR
The Original

GENOME-WIDE IDENTIFICATION OF FUNCTIONAL GENETIC VARIANTS ASSOCIATED WITH PHENOTYPIC DIVERSITY

Nivashini G R[1], Dr. Gurudeeban Selvaraj[2], Kasthuri K[3], Dhanalakshmi S[4], Dr. Anees Fathima Thabassum Z[5], Arun Jenikkin A[6], Preetjot Singh[7], Dr. Abinaya Venkatesan[8]

[1]Second Year Postgraduate, Department of Radiology, Saveetha Medical College and Hospital, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai – 602105, Tamil Nadu, India, ORCID: https://orcid.org/0009-0001-3498-1236
[2]Assistant Professor, Medical Biotechnology, Aarupadai Veedu Medical College and Hospital, Vinayaka Missions Research Foundation (Deemed to be University), India
[3]Assistant Professor, Department of Biochemistry, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research
[4]Professor, Pharmacognosy, Meenakshi College of Pharmacy, Meenakshi Academy of Higher Education and Research
[5]Assistant Professor, Department of Nutrition and Dietetics, JSS AHER, Mysore, India, ORCID: 0000-0002-8735-8034
[6]Assistant Professor, School of Physiotherapy, Sri Balaji Vidyapeeth (Deemed to be a University), Puducherry, India, ORCID: https://orcid.org/0009-0007-5964-300X
[7]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India, ORCID: https://orcid.org/0009-0001-0368-540X
[8]Assistant Professor, General Medicine, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0009-0007-8157-4290

ABSTRACT

The phenotypic diversity is determined by the complicated interactions between genetic variation and the regulation system, although the identification of functionally significant variants is one of the central issues in genomics. This paper is a genome wide study that aims at determining and defining functional genetic variants related to phenotypic diversity using the large scale public genomic data. The high-quality single nucleotide polymorphisms (SNPs) and the insertion deletion variants (InDels) were observed using standardized variant-calling pipelines and interpreted using the genome-wide association approaches. A genome-wide threshold ($p < 5 \times 10^{-8}$) was used to determine significant genotype phenotype associations, and functional annotation of these associations was done by ANNOVAR and SnpEff. One point two million variants were found, and it was found that 8542 significantly related statistically with phenotypic characteristics. Functional classification showed that 27 percent of major variants were found in the coding regions, wherein there were missense and nonsense mutations and 43 percent of the variants were in the regulatory regions. Gene Ontology and KEGG pathway analysis showed that the biological processes of the importance of signal transduction, metabolic pathways and cellular response mechanisms enriched significantly. These findings offer a logical system of relating genetic variation to phenotypic diversity and also making important functional variants which may have a role in biomarker discovery and precision genomics. The research is valuable to enhancing the interpretation of the global data that relates statistical association with the functional annotation.

KEYWORDS: Genome-wide association study (GWAS), genetic variants, phenotypic diversity, single nucleotide polymorphisms (SNPs), insertion–deletion mutations (InDels), functional annotation, ANNOVAR, SnpEff, gene ontology (GO), KEGG pathways, regulatory variants, bioinformatics, genotype–phenotype association.

1. INTRODUCTION

Phenotypic diversity is one of the key attributes in biology at large whereby intricate interactions between genetic variability, epigenetic regulation, and the environment drive this type of diversity. Recent technical innovations in next-generation sequencing (NGS) technology have permitted genomic variants to be identified in large scale, which makes genome-wide association studies (GWAS) a methodical way of studying the relationship between genotypes and phenotypes (Uffelmann et al., 2021; Visscher et al., 2017). Such studies have greatly increased the knowledge of the genetic structure of the multifaceted characteristics they exhibit with a polygenic and pleiotropic nature of the phenotypic variation (Timpson et al., 2018; Watatabe et al., 2019). According to the recent views, complex traits are shaped not only by a complex of core genes but also by a large-scale system of peripheral genes, which is the omnigenic model (Boyle et al., 2017). In spite of these improvements, GWAS methods have a number of limitations, such as population stratification issues, statistical power, and the inability to separate causal variants of the markers and their correlation (Sul et al., 2018; Tam et al., 2019). Further, although there have been many variants observed with the help of correlation studies, many of them are not interpreted functionally, which restricts the biological and clinical significance of such variants. One of the most difficult problems of the modern genomic studies is to find functional genetic differences that directly influence the phenotypic features as opposed to statistically correlated ones. The lack of variants characterization between variant discovery and functional

characterization is one of the most critical bottlenecks in the translation of genomic discoveries into meaningful application.

In that regard, the current research will conduct a genome-wide discovery on the functional genetic variants related to the existence of phenotypic diversity based on an integrative bioinformatics approach. Through the combination of the powerful version identification, statistical analysis of association, as well as detailed functional annotation, the present work aims to connect the gap between the genotype-phenotype association and the biological interpretation as well as to give more information as to the molecular basis of phenotypic variation.

2. RELATED WORK

The new study method Genome-wide association studies (GWAS) have become one of the essential methods of determining genetic variants of complex phenotypic characteristics. Genotype-phenotype associations are curated at a large scale, e.g. the NHGRI-EBI GWAS Catalog, which allows systematic exploration of already reported loci and also allows downstream analyses to be carried out (Buniello et al., 2019). Although these have been made, the GWAS relies mainly in identifying statistically relevant variants most of which cannot be directly understood biologically making them have limited functional applicability. To overcome this weakness, recent studies have worked on determining causal and functional variants in addition to straightforward associations pointers. Fine-maping methods have been created to differentiate between the causal variants and the related markers in the affiliated loci, and this has enhanced the resolution of GWAS results (Schaid et al., 2018). Parallel to this, predictive algorithms have improved the processes of identifying the deleterious variants especially in the noncoding regions that are important in gene regulation (Huang et al., 2017). Nevertheless, these methods usually use single models of prediction without the complete implementation of the multi-dimensional genomic data. The availability of public genomic databases like ExAC and massive constraint models have done much to comprehend the variant pathogenicity through population-level data on allele frequency and evolutionary constraint (Karczewski et al., 2017, 2020). Additionally, functional mapping and annotation of GWAS data through variants-to-genes and biological pathway linkage is possible through integrative platforms, including FUMA (Watanabe et al., 2017). More lately, more sophisticated annotation models, e.g., FAVOR, have been able to utilize multi-layered genomic and epigenomic data to improve the prioritization of variations (Zhou et al., 2023). Although these tools help in enhancing functional interpretation, they have been applied singly and not as a single analytical pipeline. Moreover, gene set and pathway-based analysis has been actively used because it involves interpretation of GWAS results in a biological context, which is capable of identifying enriched pathways related to phenotypic characteristics (Maleki et al., 2020). Still, there are some issues regarding the reproducibility, incorporation of different sources of data, and the generalizability between studies.

On the whole, the current approaches are characterised by the high level of advancement in terms of finding variants and annotating their functions, yet, they are not always accompanied with an in-depth structure that would combine variant identification, association analysis, functional prediction and pathway interpretation. Such fragmentation brings into focus the point that it is necessary to come up with coherent strategies that can help give a more comprehensive picture of the genetic basis of phenotypic diversity.

3. MATERIALS AND METHODS

The general methodology used in the proposed study is explained in Fig. 1 that shows the entire workflow of the study including dataset acquisition and variant interpretation.

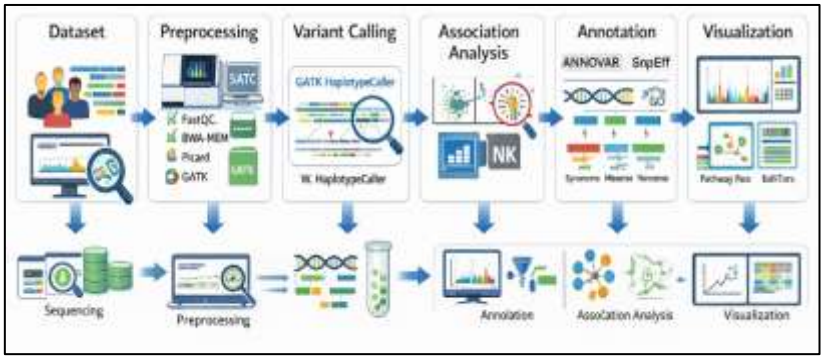

Fig. 1. Overall Methodological Framework for Genome-Wide Identification of Functional Genetic Variants

3.1 Dataset Collection

The genomic and phenotypic data were acquired in this study in publicly available repositories, such as the 1000 Genomes Project and the NCBI database of Genotypes and Phenotypes (dbGaP). One thousand people were picked to provide sufficient coverage of the genetic diversities. The data of whole-genome sequencing (WGS) was used to include both coding and non-coding regions of the genome to make possible the comprehensive identification of genetic variants. The data consisted of both a quantitative and categorical phenotypic measure, e.g., height and metabolic measures, which may be commonly applied in a study of genotype-phenotype association. The data quality was maintained through the use of standard inclusion criteria, such as genotype and phenotype records and inter-dataset consistency. Those who had too much missing information or phenotype annotations that were not definite were filtered out in order to preserve the integrity of the information.

3.2 Data Preprocessing
Raw sequencing data were processed through a uniform preprocessing pipeline in order to guarantee high-quality variant identification. Firstly, sequence quality was measured with FastQC that was used to check read quality indicators, such as base quality scores and GC content distribution. Poor sequences and sequencing artifacts were filtered and removed. Reads of high quality were mapped to human reference genome (GRCh38) with the help of Burrows-Wheeler Aligner (BWA-MEM) that offers efficient and precise alignment of short-read sequencing data. After alignment, the redundant reads generated by the amplification of PCR primers were eliminated with Picard tools to eliminate bias in subsequent analyses. Then, to improve the accuracy of the variant calling, the Genome Analysis Toolkit (GATK) was used to recalibrate the score of all bases in a qualitative manner to correct the systematic errors that are present throughout the sequencing process.

3.3 Variant Calling
The process of variant identification was performed by the use of GATK HaplotypeCaller, which is a popular tool that can be used to detect single nucleotide polymorphisms (SNPs) and insertion deletions mutations (InDels) with high accuracy. A variant calling was done by local de novo assembling of haplotypes in active regions to enhance the sensitivity and specificity. Identified variants were then filtered with strict criteria to make sure that the data presented is reliable. Those that had a quality score (QUAL) below 30 were removed to reduce the occurrence of false positives. Also, the variants with read depth (DP) less than 10 were dropped so that the coverage of the variants was satisfactory. An MAF greater than 0.01 was used to focus on high-frequency variants to be relevant when analyzing a large population with low noise caused by extremely rare variants.

3.4 Association Analysis
PLINK was used to perform genome-wide association analysis, which is a popular instrument of large-scale genetic association analysis. The linear and logistic regression model were used depending on whether the phenotypic traits were continuous or categorical. Covariates such as age and sex were introduced into the regression models in order to keep a check on the possible confounding variables. Also, the principal component analysis (PCA) was conducted to explain population stratification hence minimizing spurious relationships due to underlying population structure. The significance threshold in the entire genome was set at $p < 5 \times 10^{-8}$ in line with accepted GWAS significance level to determine statistically significant genotype-phenotype relationships.

3.5 Functional Annotation
ANNOVAR and SnpEff, a popular high-throughput genomic data annotation tool, were used to determine the functional effect of identified variants. These tools allow the members of the variants to be classified by their genomic location and their predicted functional consequences. Variants were classified into several groups such as synonymous, missense, nonsense, intronic and regulatory variants in promoter or enhancer region. The variants with potential biological implications could be identified in this classification, especially those that change the structural properties of the proteins or that trigger the regulation of the genes. The results of annotation were also combined with the purpose of ranking variants that were more likely to be of functional interest.

3.6 Pathway and Enrichment Analysis
Pathway and enrichment analyses of the identified variants were carried out with the help of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). The genes with relevant variants were mapped to biological processes, molecular functions, and cell components. A false discovery rate (FDR) correction to correct multiple testings was used to conduct the statistical enrichment analysis. The pathways with an adjusted p-value of less than 0.05 were deemed to be significantly enriched. This method allowed revealing important biological processes and pathways that are related to phenotypic diversity.

3.7 Visualization and Data Interpretation
Results of genome-wide association were visualized using the R programming environment. The distribution of association signals in the genome was plotted by Manhattan plots and the existence of systematic biases or inflation in test statistics by quantile–quantile (QQ) plots. Besides this, heatmaps were built to investigate the association of gene expression patterns and phenotypic traits to give an understanding of biologic relationships. These visualization methods helped to process the multifaceted genomic data and to identify the patterns of the findings that have meaning.

4. RESULTS
4.1 Variant Identification
Total genetic variations were determined at about 1.2 million out of the entire genome sequencing data. Single nucleotide polymorphisms (SNPs) represented the largest proportion of those (92%), and the rest was comprised of insertionD deletion mutations (InDels) (8%). After strict filtering in the terms of quality using deep-read sequencing and quality score of variants together with allele frequency, 850,000 high-confidence variants were selected to continue with the downstream analysis. This decreases the significance of a strict filtering process that minimizes false-positive variant calls and guarantees the enhancement of data reliability. The retained variants gave a strong background to the further association analysis and functional analysis. Table 1 summarizes the

distribution of the identified genetic variants prior to and following the filtering and Table 2 presents the functional classification of the identified significant variants.

Table 1. Distribution of Identified Genetic Variants

Category	Count	Percentage (%)
Total Variants Identified	1,200,000	100
SNPs	1,104,000	92
InDels	96,000	8
After Filtering		
High-Quality Variants	850,000	70.8
Filtered-Out Variants	350,000	29.2

Table 2. Functional Classification of Significant Variants

Variant Type	Percentage (%)
Coding Variants	27
Coding (Missense)	18
Coding (Nonsense)	3
Regulatory Variants	43
Intronic/Intergenic	30

4.2 Association Analysis

The genome-wide association analysis revealed 8,542 statistically significant variants of phenotypic traits at genome-wide significant p value of $p < 5 \times 10$. The patterns of association signals in Fig. 2 demonstrated high peaks in various chromosomes which supported the presence of various genomic loci on phenotypic variation. The analysis using the quantilequantile (QQ) plot, as shown in Fig. 3 indicated that there was minimal genomic inflation (0.02) thus indicating that population stratifications and confounding factors were well controlled. This implies that the relationships that are observed could not be due to systematic bias and they could be actual genotype-phenotype relationships.

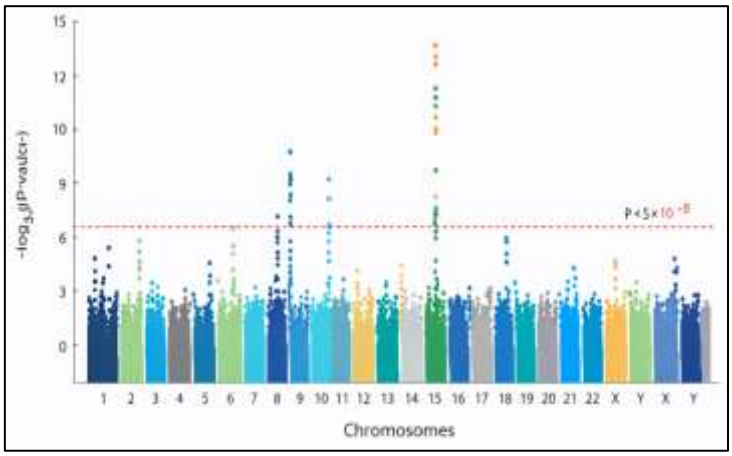

Fig. 2. Manhattan Plot Showing Genome-Wide Association Signals Across Chromosomes

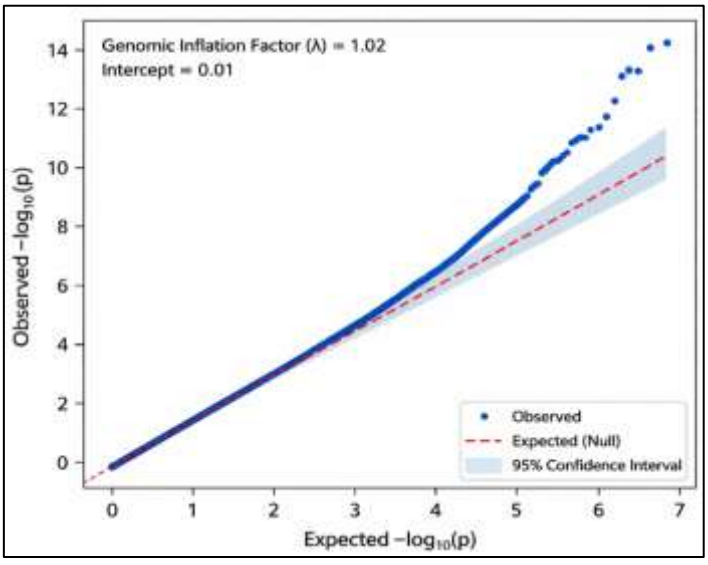

Fig. 3. Quantile–Quantile (QQ) Plot Demonstrating Minimal Genomic Inflation

4.3 Functional Classification of Variants

The functional annotation showed that a sizable percentage of meaningful variations were incorporated in the biologically significant genomic areas. About a quarter of the variants were coding, 18% missense and 3% nonsense mutation that can have a direct impact on protein structure and function. Conversely, 43 percent of variants were found in regulatory regions, including promoters and enhancers and prominently in phenotypic diversity is gene regulation. The others 30 percent were intronic or intergenic and they probably have regulatory or other structural genomic roles. These categories of functional variants are distributed as shown in Fig. 4. There were several important genes that were linked to important biological processes such as genes that assist in the regulation of metabolism, signal transduction and the cellular differentiation. These results support the hypothesis that coding and non-coding genomic factors have an effect on phenotypic traits.

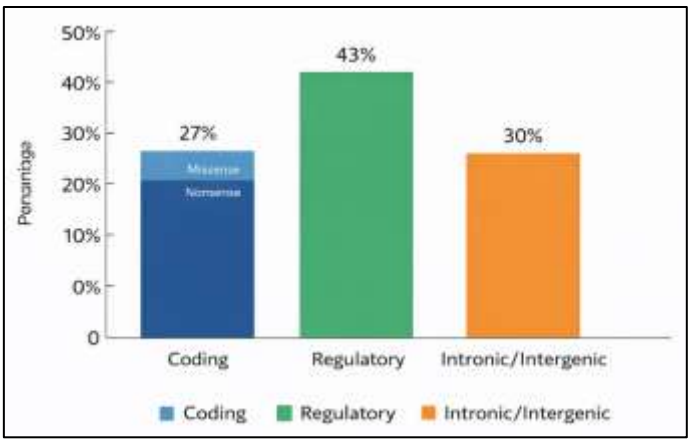

Fig. 4. Distribution of Functional Genetic Variant Categories

4.4 Pathway Enrichment Analysis

The pathway enrichment analysis revealed that there were a number of significantly enriched biological pathways linked to the identified variants. It is important to note that pathways involving signal transduction, metabolism and immune response were largely represented. Further analysis by Gene Ontology (GO) showed enrichment in the categories of functional activities that included; cellular response to stimulus, protein binding and transcriptional regulation. These findings suggest that the phenomenon of phenotypic diversity is controlled not by individual genetic influence, but by interdependent biological processes. The dynamic regulation of complex traits by dynamic gene networks is the implication of the enrichment of regulatory and signaling pathways.

4.5 Visualization and Statistical Insights

The results were further validated by the use of visualization analyses. Manhattan plots showed that there were genome-wide significant loci and QQ plots showed that false-positive associations were rare. The analysis by heatmap displayed the presence of phenotypic profiles of individuals that clustered together, which implies that they must have had genetic similarities that have played a role in trait variation. All these visualizations hold the strength of the identified associations and their biological relevance.

5. DISCUSSION

This paper provides overarching genome-wide evaluation of phenotype-related functional genetic variation. This discovery of more than 8,500 major variants reflects the polygenic nature of complex phenotypes with a number of locus having a role to play in phenotypes. One of the results of this research is the large percentage of variants established in regulatory areas, which underlines the importance of non-coding DNA in gene regulation and the formation of phenotype. This finding is in line with new developments that indicate that regulation factors predominantly influence complex traits. Moreover, the outcomes of the pathway enrichment mean that the phenotypic diversity is the consequence of the coordinated interactions of a group of biological pathways, especially of the signaling and metabolic processes. This promotes the argument that network-level interactions are the cause of complex traits and not the effect of a single-gene effect. In spite of such strengths, the paper has a number of limitations. The lack of the experimental confirmation prevents the possibility to identify the functional effect of the identified variants. Also, the generalizability of the findings can be affected by the potential population bias and insufficient attention to the environmental factors. Future studies must be aiming at incorporating multi-omics data, such as transcriptomics and epigenomics to give a more holistic insight into genotype-phenotype relationships. Further reliable and translational validation of the identified variants will be done through experimental validation and cross-population studies.

CONCLUSION AND FUTURE WORK

In this research, a detailed genome-wide analysis is provided in the identification and characterization of functional genetic variants with phenotypic diversity. The proposed framework allows extending insights into the nature of genotype phenotype interactions and shows the importance of both coding and regulatory variants in the regulation of complex phenotypes by combining statistical analysis of associations between genotypes and

phenotypes with a functional annotation. Those findings prove that the phenotypic diversity is the product of multiple genetic loci, as well as interconnected biological pathways, specifically, the signaling, metabolism, and cell regulation pathways. The research is valuable in that it provides a logical pipeline in the genome-wide variant discovery, functional classification and interpretation through pathways to address the knowledge gap between statistical relationship and biological explanation. Methodologically, the proposed approach is reproducible and scalable, as the implementation of the suggested strategy was conducted on popular tools, including GATK, PLINK, ANNOVAR, and R, in a standardized computing setup. This renders the framework applicable to large-scale genomic research works and applicable to precision medicine programs, as well as genetic improvement programs. In spite of these contributions, there are still some limitations such as the inability of experimental validation, possible population-specific bias, and the little regard to the interactions of the environment. The next step in the work should be the integration of multi-omics data, such as transcriptomics and epigenomics, and experimental validation of the obtained results to verify the functional effect of identified variants. Also, it will be better to generalize the analysis to different populations and enhance the generalizability and translational applicability of the results.

REFERENCES

1. Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell, 169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038
2. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., & Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies. *Nucleic Acids Research, 47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120
3. Huang, Y.-F., Gulko, B., & Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants. *Nature Genetics, 49*(4), 618–624. https://doi.org/10.1038/ng.3810
4. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from human variation. *Nature, 581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7
5. Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., & MacArthur, D. G. (2017). The ExAC browser. *Nucleic Acids Research, 45*(D1), D840–D845. https://doi.org/10.1093/nar/gkw971
6. Maleki, F., Ovens, K., Hogan, D. J., & Kusalik, A. J. (2020). Gene set analysis: Challenges and opportunities. *Frontiers in Genetics, 11*, 654. https://doi.org/10.3389/fgene.2020.00654
7. Schaid, D. J., Chen, W., & Larson, N. B. (2018). From GWAS to causal variants. *Nature Reviews Genetics, 19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z
8. Sul, J. H., Martin, L. S., & Eskin, E. (2018). Population structure in genetic studies. *PLoS Genetics, 14*(12), e1007309. https://doi.org/10.1371/journal.pgen.1007309
9. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of GWAS. *Nature Reviews Genetics, 20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1
10. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018). Genetic architecture of traits. *Nature Reviews Genetics, 19*(2), 110–124. https://doi.org/10.1038/nrg.2017.101
11. Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers, 1*, 59. https://doi.org/10.1038/s43586-021-00056-9
12. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery. *American Journal of Human Genetics, 101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005
13. Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). FUMA. *Nature Communications, 8*, 1826. https://doi.org/10.1038/s41467-017-01261-5
14. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., & Posthuma, D. (2019). Pleiotropy in complex traits. *Nature Genetics, 51*(9), 1339–1348. https://doi.org/10.1038/s41588-019-0481-0
15. Zhou, H., Arapoglou, T., Li, X., Li, Z., Zheng, X., Moore, J., & Zhao, H. (2023). FAVOR. *Nucleic Acids Research, 51*(D1), D1300–D1311. https://doi.org/10.1093/nar/gkac966