Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Allele Frequency Analysis Suggests Potentially Protective Effect in the Lithuanian Population

Gabriele Zukauskaite (gabriele.zukauskaite@mf.vu.lt)
Vilnius University: Vilniaus Universitetas https://orcid.org/0000-0003-1726-5825
Ingrida Domarkiene
Vilnius University: Vilniaus Universitetas
Tautvydas Rancelis
Vilnius University: Vilniaus Universitetas
Ingrida Kavaliauskiene
Vilnius University: Vilniaus Universitetas
Karolis Baronas
Vilnius University: Vilniaus Universitetas
Vaidutis Kucinskas
Vilnius University: Vilniaus Universitetas
Laima Ambrozaityte
Vilnius University: Vilniaus Universitetas

Research article

Keywords: protective alleles, effect variants, genotyping, allele frequency analysis, complex diseases

Posted Date: November 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-100922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

Background. In the scientific literature, a wide range of effect variants that protect against complex disease phenotypes has been identified. Analysis of these variants and overall genetic structure of isolated or, in our case, small populations is important in association analysis. When analysing admixture populations during GWAS, one could expect inaccuracies, which could be eliminated by choosing distinct populations as one of the interests of study. Population genetic structure determines similarities and differences between individuals or different groups of individuals and the factors that may lead to those differences.

Results. In our study, we identified six missense effect variants in the Lithuanian population having frequencies that were significantly different compared to other European populations. Three of these effect variants may potentially protect against type 2 diabetes and coronary heart disease.

Conclusions. Even though high rates of these diseases in the Lithuanian population and other populations indicates the presence of environmental factors and the lack of knowledge about the interactions between regulatory regions and other effect variants. Identification of these effect variants is important not only to provide a better understanding of the microevolutionary processes and etiopathogenetic mechanisms, but also to develop disease prevention programs and novel, personalised therapies using genome editing or other genetic tools.

Background

Specific genomic loci and variants associated with survival vary between populations due to microevolutionary processes. In a changing environment, variants that once were protective may become deleterious, and therefore microevolutionary processes are ongoing and lead to transformations in the genetic architecture of a population that is adapting [1]. Natural selection mediates adaptation process, which can have various consequences, such as the prevalence of complex diseases leading to high mortality: hypertension [2], coagulation changes [3] and hyperlipidemia [4]. Such research findings have implications for population-specific (geographically and ethnically) diagnosis worldwide [5], which is important for developing new prevention and treatment strategies in an era of personalised medicine.

To understand the mechanisms of complex diseases and traits, it is important to answer the question of natural selection and adaptation through the genomic variation fluctuation process in the population during periods of time. Some genomic variants cannot simply be categorized as ‘risk’ or ‘protective’ because of the conflicting interpretation of their effect. Thus, we refer to these variants as ‘effect variants’. Effect variants are usually rare but those that provide selective advantage tend to become common in the population. That is why our analysis includes rare or previously rare effect variants. In addition, most of these variants are likely to be common in biologically redundant genes, thereby escaping the effects of purifying selection, preserving these variants at high frequencies in different populations [6]. For example, if a person has an effect variant that protects against obesity, it is possible that this person will be less likely obese and more likely to pass this variation to his offspring due to positive selection. Based on this logic, complex disease rates in the population would drop in the future. However, complex disease rates are steady and one of the reasons is the exploding growth of the human population, which results in an accumulation of extremely rare variants [7]. GWASs under-represent low-frequency ($0.5\% \leq MAF < 5\%$) and rare ($MAF < 0.5\%$) variants that could underlie much of the unexplained heritability of many complex traits [8]. In addition, minor alleles are more likely to be characterised as risk alleles in the published GWASs on complex diseases because minor alleles are more easily detected as risk alleles in GWASs [9].

Frequency was not the only criterion for selecting effect variants. Mostly non-synonymous single nucleotide effect variants were chosen for the study in order to analyse ones that affect the structure of the protein and may have a function-altering effect.

Many effect variants protect against disease by disrupting protein function, typically via loss-of-function or gene knockout effects, and have an impact on clinically relevant phenotypic effects. In this case, most of the functionally relevant loss-of-function variants should be removed through purifying selection. One of the examples is effect variants in *IFIH1* and *IL23R* genes, which are thought to protect against immune-mediated disease through impairment of the host-pathogen response. These variants are likely to have been subjected to negative selection pressures in the past, which would account for their current scarcity in the population [10]. Recent studies have however shown that synonymous mutations can influence the amount of protein that is produced; so-called optimal codons are faster for cells to process and lead to increased protein production [11]. This reveals that synonymous mutations most likely play an underappreciated role in human variation. That is why we also included some of the synonymous effect variants as well.

The identification of effect variants and a better understanding of interactions between them could provide the possibility to characterise candidate genomic regions and specify their functional significance across different populations [12–14]. The aim of this study was to identify and analyse single nucleotide effect variants (risk and protective) in the Lithuanian population.

Results

After using the compiled catalogue of effect variants (144 variants in total) as a reference, the sequencing and genotyping data of our sample group were filtered. Filtered variants were tested for Hardy-Weinberg equilibrium and 70 genome variants passed (39 variants from genotyping and sequencing data; 7 variants from genotyping data alone; 24 variants from sequencing data alone). Sample sizes used for calculations of allele frequencies in the Lithuanian and various European populations are presented in Table 1.

Table 1
Sample sizes used for calculations of allele frequencies in the studied Lithuanian and European populations.

Variant ID	Gene	Related condition	LTU	CEU	FIN	GBR	IBS	TSI	EUR
rs1801282	PPARG	T2D	168	99	99	92	107	107	504
rs13266634	SLC30A8	T2D	464						
rs11556924	ZC3HC1	CHD	465						
rs2274223	PLCE1	Esophageal cancer	463						
rs7498665	SH2B1	Obesity	98						
rs698	ADH1C	Alcohol dependence	98						

CEU – Utah Residents (CEPH) with Northern and Western European Ancestry; CHD – coronary heart disease, FIN – Finnish in Finland; GBR – British in England and Scotland; LTU – Lithuanian population; TSI – Tuscans in Italy, T2D – type 2 diabetes.

Frequencies of six missense variants were significantly different between the study group and other European populations (Table 2).

Table 2
Distribution of variant genotypes in studied the Lithuanian population.

Variant ID	Gene	Change (from reference to effect allele)*	Homozygous alternative genotype count	Heterozygous genotype count	Homozygous reference genotype count	MAF in LTU	MAF in EUR	χ2	p
rs1801282	PPARG	NM_001354668.2:c.34C > G	4	46	118	0.161	0.120	3.632	0.05
rs13266634	SLC30A8	NM_001172815.2:c.826C > T	50	195	219	0.318	0.283	4.839	0.03 (FIN)
								4.787	0.03 (GBR)
								4.387	0.04 (CEU)
rs11556924	ZC3HC1	NM_001282190.1:c.1025G > A	77	225	163	0.408	0.377	4.146	0.04 (GBR)
								16.642	< 0.01 (FIN)
rs2274223	PLCE1	NM_001165979.2:c.4856A > G	65	223	175	0.381	0.338	3.919	0.04
rs7498665	SH2B1	NM_001145812.1:c.1450A > G	19	36	43	0.378	0.329	7.015	0.03
rs698	ADH1C	NM_000669.5:c.1048A > G	25	48	25	0.5	0.405	5.878	0.05

* In our case, all effect alleles were minor alleles as well. CEU – Utah Residents (CEPH) with Northern and Western European Ancestry; EUR – European population; FIN – Finnish in Finland population; GBR – British in England and Scotland population; LTU – Lithuanian population; MAF – minor allele frequency.

According to the scientific literature, these variants may have protection against alcohol dependence (*ADH1C*, rs698, p = 0.05), type 2 diabetes (*PPARG*, rs1801282, p = 0.05; *SLC30A8*, rs13266634, p = 0.03 (FIN), p = 0.03 (GBR), p = 0.04 (CEU)), coronary heart disease (*ZC3HC1*, rs11556924, p = 0.04 (GBR), p < 0.01 (FIN)), obesity (*SH2B1*, rs7498665, p = 0.03), and oesophageal cancer (*PLCE1*, rs2274223, p = 0.04). Distribution of variant genotypes in the studied Lithuanian population is presented in Table 2.

Filtered effect variants were compared with primate species to ascertain which allele is derived and which is ancestral in order to avoid the erroneous assumption in some cases that the rare allele is the derived allele for common variants. The analysis showed that several of our catalogue-selected effect variants (in *PLCE1*, *ADH1C*, and *SH2B1* genes) in humans are in fact ancestral.

Discussion

According to free access *in silico* analysis tools, five of the six effect variants for which frequencies in the Lithuanian population differed significantly from European populations are considered benign (regarding *Varsome* or *UniProt*) or a risk factor (*Ensembl*). All these five variants (*PPARG*: rs1801282, *SLC30A8*: rs13266634, *ZC3HC1*: rs11556924, *PLCE1*: rs2274223, *SH2B1*: rs7498665) were selected from the scientific articles for our catalogue of effect variants. In these articles, these variants were identified as candidate protective genome variants after GWAS data was filtered for nonsynonymous SNPs to increase the likelihood of them being functional and after bioinformatic analyses were performed to detect evidence of positive natural selection for the effect variant and to estimate the probability of the mutation being damaging. In addition, a variant was considered protective when it was more frequent in controls than cases [5].

Variant rs698 in the *ADH1C* gene is known as protective (according to the *Ensembl*, *ClinVar* and *OMIM*) and has an impact on ethanol metabolism. Even though databases define this variant as protective, various studies suggest that this variant is associated with slower ethanol metabolism, which could lead to a longer period of consuming alcohol and the consumption of greater quantities. Therefore, people carrying the variant have a higher risk of heavy and excessive drinking [15, 16]. According to one study, common SNPs are responsible for as much as 30% of the variance in alcohol dependence, but few have been identified [17]. Power analyses however indicate that additional SNPs associated with alcohol dependence are likely to have small effect sizes and are more consistent with more common psychiatric disorders [18]. This shows that an understanding of the molecular mechanisms involved in excessive alcohol consumption and other complex conditions are still unresolved and that the collection of large numbers of well-characterised cases and controls is needed.

Besides function, the origin of effect alleles must also be addressed. Every disease-associated SNP consists of two alleles, of which one is considered as risk-associated and the other as disease-protective. A common practice to ascertain whether a nonsynonymous SNP is protective (i.e. the respective derived allele is protective) is to deduce which allele is derived and which is ancestral, since a minor allele does not necessarily equal the derived (mutant) one. The origin of the allele could be determined by using genomic alignments with primate species. The effect variants we analysed did not have a very low (< 1%) minor allele frequency, and we cannot assume that the rare allele is the derived allele. However, if a derived allele provides a protective function and gives an individual a selective advantage, one might expect positive selection to sweep it to become the most common allele in the population [5]. This may be the reason why the effect variants we analysed have allele frequencies greater than 1%. Moreover, this could be the reason why databases and SNP analysis tools call these variants polymorphisms. Comparison with primate species showed that variants analysed in *PLCE1*, *ADH1C*, *SH2B1* are indeed ancestral. The protective nature of genomic variants can be considered when the allele is derived, which is why we did not interpret these variants as protective. Despite contradicting data, significant variants may have some effects on the aetiopathogenesis of particular complex diseases.

In our study, effect variants may have an impact on protection against type 2 diabetes (variants in *PPARG* and *SLC30A8* genes) and coronary heart disease (*ZC3HC* variant). It is important to keep in mind the effects of environmental factors. According to data from The Lithuanian Department of Statistics (Statistics Lithuania) [19], the highest number of deaths (55.4%) in 2018 was caused by diseases of the cardiovascular system. In 2014, 4.4% of the population had type 2 diabetes and 7.5% had coronary heart disease. Even though our population have effect variants that may protect against these diseases, lifestyle and other environmental factors may influence the frequency of morbidity. Also, many studies concentrate on effect variants of coding genomic parts, but interactions between coding and non-coding variants are as important but are not examined enough. Although these effect variants may reduce the risk of disease (or maintain health), there are additional genetic mechanisms that control this process. Not only are the effects of single genomic variants important, but their interactions and the interactions between regulatory regions are also consequential [9].

Butler et al. [5] estimated an integrated haplotype score for the effect variants that we have analysed in the *PPARG*, *SLC30A8*, and *ZC3HC1* genes that showed that these variants may have undergone recent positive selection [20]. This shows that a derived allele is beneficial for an individual's fitness and may be protective. However, the functional impact of these variants has to be confirmed and additional analysis is needed. According to Plenge et al., most alleles associated with complex diseases (approximately 85%) fall outside the protein-coding sequence, and thus each disease-associated allele should be evaluated to see whether it is in linkage disequilibrium with a variant that changes protein structure. If it is, then these findings should be fast-tracked for functional studies in human cells and animal models to assess the gain-of-function or loss-of-function. For non-coding effect variants, the effect on gene expression should be evaluated in a relevant human cell type. For example, if a risk allele is associated with higher gene expression, then pharmacological inhibition may be effective in treating the disease [21].

Conclusions

During our study, we identified three effect variants in the Lithuanian population group that may protect against type 2 diabetes and coronary heart disease. A better understanding of common variants and their effects can help build better databases, because sometimes the effect of the variant can be incorrectly described, as was previously demonstrated in this study regarding the variant in the *ADH1C* gene. Detection of these effect variants is important not only to provide a better understanding of the aetiopathogenetic mechanisms and microevolutionary processes. It could broaden the knowledge about the differences between populations and this way move towards personalised medicine as well. Knowledge

of effect variants in specific populations can be used as targets for the development of disease prevention programs and novel, personalised therapies and for the use of genome editing tools.

Methods

The aim and design of the study

The aim of this study was to identify genome effect variants in the Lithuanian population. Genotypes were identified using high-throughput genotyping and sequencing technologies. A catalogue of effect variants was compiled and used as a reference to filter effect variants characteristic to our sample group. Allele frequency and statistical analysis was performed. Comparison of filtered effect variants with primate species variants was carried out to ascertain which allele is derived and which is ancestral.

Participants and samples

The study group included 475 unrelated, self-reported healthy subjects (239 women and 236 men) of Lithuanian descent.

DNA was extracted from peripheral blood leukocytes using the phenol-chlorophorm-isoamyl alcohol method according to a laboratory-approved methodology or using an automated *TECAN Freedom EVO®200* system (manufactured by Tecan Group Ltd., Switzerland) using *Promega* beads assay according to the manufacturer's user guidelines. Concentration and purity of the DNA were determined with a *NanoDrop®* spectrophotometer.

Catalogue of effect variants

A catalogue of 144 effect variants from the *ClinVar* [22] and *OMIM* [23] databases as well as scientific publications was compiled. The criteria for including a variant from the databases were 1) clinical significance review status (protective or uncertain) and 2) count of submissions (more than 1). The criteria for including a variant from scientific publications were 1) influence of the variant on gene function (i.e., variant was expected to alter gene function; mostly loss-of-function) and 2) the frequency of the variant (i.e., rare or previously rare alleles, which increased in frequency possibly because of positive effects on the phenotype). This catalogue was used as a reference to filter out the effect variants in the sample group from the Lithuanian population. The catalogue of effect variants is presented in Additional file 1.

Genotyping data and statistical analysis

The genotypes were extracted from different types of data: exome sequencing and genome-wide genotyping arrays. Whole exome sequencing was performed using 5500 series *SOLiD™* systems protocol guides for 98 subjects of Lithuanian descent. High-throughput genotyping (*Illumina HiScanSQ System*) was performed using *Illumina Infinium® HD* and *HTS* assay protocol guides (beadchip arrays *Illumina 770 HumanOmniExpress-12 v1.0, v.1.1.* and *Infinium OmniExpress-24v1-2)* for 475 individuals of Lithuanian descent.

Quality control of exome sequencing data was performed using *LifeScope™ Genomic Analysis Software v2.5*. Sequence coverage value of more than 10-fold was considered acceptable (mean quality score of the reference allele: 28 (± 2.3), mean quality score of the new allele: 28.4 (± 1.8)) [24].

Primary genotyping results were examined and prepared for further analysis by using *GenomeStudio v2011.1* software (*Illumina Inc*.). DNA sample quality parameters were used [25]: 465 samples out of 475 had a call rate > 97, p10GC > 0.7, and therefore 10 DNA samples were excluded. To perform SNP quality evaluation, call frequency, *GenTrain* and *ClusterSep* as quality control parameters were used [26]. Call frequency range is from 0.13 to 1 (10262 SNPs were eliminated), *GenTrain* range is from 0.35 to 0.98. SNPs with *ClusterSep* scores less than 0.27 score were eliminated. Subsequent data analysis (Hardy-Weinberg equilibrium), SNP filtering, and SNP frequency calculations were performed using the *PLINK v1.9* software [27].

Allele frequencies of effect variants were calculated and compared to the general European population and to distinct European populations (Utah Residents with Northern and Western European Ancestry; Finnish in Finland; British in England and Scotland; Iberian populations in Spain; and Tuscans in Italy) based on 1000 Genomes project data, [28] accessible in the *NCBI dbSNP* database [29]. Statistical analysis (χ^2 or Fisher's exact test [when the sample size was ≤ 5], $\alpha = 0.05$) was performed using *Rstudio* v3.5.2. software [30].

To define the possible impact on the genome, effect variants were analysed using *in silico* tools and databases: *Varsome* [31], *Uniprot* [32], *Ensembl* [33], *ClinVar*, and *OMIM*.

The *Ensembl* database was used to compare filtered effect variants with primate species variants (*Gorilla gorilla*, *Pongo abelii*, *Theropithecus gelada*, and *Chlorocebus sabaeus*) to ascertain which allele is derived and which is ancestral.

Abbreviations

CEU – Utah Residents (CEPH) with Northern and Western European Ancestry

CHD – coronary heart disease

EUR – European population

FIN – Finnish in Finland population

GBR – British in England and Scotland population

GWAS – genome-wide association study

LTU – Lithuanian population

MAF – minor allele frequency

SNP – single nucleotide polymorphism

T2D – type 2 diabetes

TSI – Tuscans in Italy

Declarations

Ethics approval and consent to participate

This research is a part of the LITGEN (VP1-3.1-ŠMM-07-K-01-013) and the ADAPT (S-MIP-20-35) projects, which were approved by the Vilnius Regional Research Ethics Committee (approval No. 158200–05–329–79 and 2019/4□1119□612, respectively). Written informed consent was received from all study participants.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

This research is a part of the LITGEN (funded by the European Social Fund under the Global Grant measure, agreement No VP1-3.1-ŠMM-07-K-01-013) and the ADAPT (has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-20-35) projects. Financial means to allow the authors to carry out the study (sample collection, sequencing and high-throughput genotyping) was provided by LITGEN project. Funding for the analyses of data and writing the manuscript was provided by ADAPT project. The funding bodies played no role in the design of the study.

Authors' contributions

ID, IK and LA collected the samples and performed high-throughput genotyping and sequencing. Data quality control check was performed by ID and IK. TR performed sequencing data annotation. GZ and KB compiled effect variant catalogue. GZ analysed and interpreted sequencing and genotyping data using effect variant catalogue. VK designed the project, LA and ID contributed to the design of the research. GZ, ID and LA were the major contributors in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Merilä J, Sheldon BC, Kruuk LEB. Explaining stasis: microevolutionary studies in natural populations. Genetica. 2001;112(1):199–222.
2. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 2008;4(2):e32.
3. Dahlbäck B. Advances in understanding pathogenic mechanisms of thrombophilic disorders. Blood. 2008;112(1):19–27.
4. Stengård JH, Zerba KE, Pekkanen J, Ehnholm C, Nissinen A, Sing CF. Apolipoprotein E polymorphism predicts death from coronary heart disease in a longitudinal study of elderly Finnish men. Circulation. 1995;91(2):265–9.
5. Butler JM, Hall N, Narendran N, Yang YC, Paraoan L. Identification of candidate protective variants for common diseases and evaluation of their protective potential. BMC Genomics. 2017;18(1):575.
6. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8.
7. Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. Population genetics of rare variants and complex diseases. Human Hered. 2012;74(3–4):118–28.
8. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics. 2014;95(1):5–23.
9. Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, et al. Are minor alleles more likely to be risk alleles? BMC Med Genom. 2018;11(1):1–11.
10. Harper AR, Nayee S, Topol EJ. Protective alleles and modifier variants in human health and disease. Nat Rev Genet. 2015;16(12):689–701.
11. Dhindsa RS, Copeland BR, Mustoe AM, Goldstein DB. Natural selection shapes codon usage in the human genome. The American Journal of Human Genetics. 2020;107(1):83–95.
12. Slim L, de Foucauld H, Chatelain C, Azencott CA. A systematic analysis of gene-gene interaction in multiple sclerosis. *bioRxiv*. 2020.
13. Chattopadhyay A, Lu TP. Gene-gene interaction: the curse of dimensionality. *Annals of Translational Medicine*. 2019; 7(24).
14. Li Y, Cho H, Wang F, Canela-Xandri O, Luo C, et al. Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. Journal of the American Heart Association. 2020;9(7):e014146.
15. Tolstrup JS, Nordestgaard BG, Rasmussen S, Tybjaerg-Hansen A, Grønbæk M. Alcoholism and alcohol drinking habits predicted from alcohol dehydrogenase genes. Pharmacogenomics J. 2008;8(3):220–7.
16. Edenberg HJ. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. Alcohol Res Health. 2007;30(1):5.
17. Palmer RH, McGeary JE, Heath AC, Keller MC, Brick LA, Knopik VS. Shared additive genetic influences on DSM-IV criteria for alcohol dependence in subjects of European ancestry. Addiction. 2015;110:1922–31.
18. Walters RK, Polimanti R, Johnson EC, McClintick JN, Adams MJ, Adkins AE, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat Neurosci. 2018;21:1656–69.
19. The Lithuanian Department of Statistics (Statistics Lithuania). Available from: https://www.stat.gov.lt/en/ [Accessed 28th July 2020].
20. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):e72.
21. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nature reviews Drug discovery. 2013;12(8):581–94.
22. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4. PubMed PMID: 29165669.
23. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), [2020-08-12]. Available from: https://omim.org/.
24. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. PLoS Genet. 2013;9(9):e1003815.
25. Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, et al. "Illumina human exome genotyping array clustering quality control" Nature protocols. 2014;9(11):2643.
26. Illumina. Infinium genotyping data analysis–a guide for analyzing infinium genotyping data using the genomestudio genotyping module. 2010. Available from: https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
28. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
29. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. Available from: http://www.ncbi.nlm.nih.gov/SNP/.

30. Team RC. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013. http://www.R-project.org.
31. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Aguilera MA, et al. VarSome: The Human Genomic Variant Search Engine. Oxford Bioinformatics. 2019;35(11):1978.
32. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic acids research. 2019;47(D1):D506–15.
33. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, et al. Ensembl 2020. Nucleic acids research. 2020;48(D1):D682–8.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1catalogueofeffectvariants.xlsx