university of groningen

Identification of Prioritized SNPs Involved in Human Reproductive Behavior

Genetic Epidemiology and Bioinformatics

Life Science & Technology, Biomedical Sciences

Bachelor Research Project

Author: Daniils Zilkins (S3757722)

Contributor: Andrei Guriencu (S3959872)

Coordinator: Ilja Nolte

Supervisor: Ahmad Vaez

May/July, 2022

Abstract

Background

In previously conducted studies, 371 specific germline substitution mutations, also known as SNPs, that are linked to human reproductive behavior have been identified through the GWAS analysis. Applied GWAS method was based on the linkage disequilibrium (LD), which is the association network that arises between DNA variants within the human genome.

Research Aim & Question

Aim of this study is to identify functional SNPs through the post-GWAS analysis, and create a list of prioritized genes underlying the GWAS signals. This list of genes could be the foundation for future laboratory studies and that may result in a therapeutic applications and implementation of knowledge on the most relevant SNPs with a possible pathogenic effect on human reproductive behavior. The following research question is set to be answered – What are the most significant genes that are involved in human reproductive behavior, particularly in the context of AFB and AFS phenotypic traits?

Methods

Post-GWAS analysis consisted of two phases. The goal of the first In-Silico Sequencing Phase was to identify all of the SNPs linked to the 371 gSNPs. Aim of the second eQTL phase was to analyze whether the associated missense SNPs, identified in the previous phase, impact expression levels in adjacent genes.

Results

Two lists, each contains the five most significant SNPs and their correspondent gene associations were created. In-Silico phase list is composed of five missense gSNPs with the highest CADD score and LD value closest to 1. The eQTL analysis phase list consists of five most significant associations of previously identified missense SNPs with the cis-eQTLs and associated genes, with the highest Z-score.

Conclusion

The study was concluded with the two lists, each consisting of five prioritized genes, that though further bibliography analysis, indeed were found to be potential triggers of various disorders. Four of five genes from the first phase were linked to reproductive behavioral traits, as well as to different pathologies, of which some potentially affecting reproduction. Although, further research is needed in order to confirm and test whether there is a certain connection between the increased disease risk and single nucleotide polymorphisms within the GWAS-associated gene, which are involved in human reproductive behavior.

Introduction

The method of finding associations between genetic variants across the genome and expressed phenotypes in a studied population is known as GWAS analysis. Its core objective is to gain a better understanding of the disease biology so that more effective prevention and treatment strategies could be designed. The GWAS method is based on the linkage disequilibrium (LD), an association network that occurs between DNA variants in human genome, which arises as a result of prior evolutionary changes, such as population size constraints, mutations, recombination rate, as well as natural selection. (Visscher P. M. et al., 2017*)* Systemic biases produced by marginal sources of error can magnify the number of false-positive and false-negative associations, therefore precise genotyping is crucial for feasibility through any large-scale GWAS study. (Anderson C. A. et al., 2017)

Since the objective of GWAS is the study of genetic variation and their associations with complex traits and diseases, it worth mentioning the most prevalent genetic variation occurring within the human genome, known as single nucleotide polymorphisms. SNPs are germline substitution mutations of a single nucleotide at a particular location within the genome. SNP array-based GWAS studies as research method have limitations due to required high significance threshold as a result of multiple testing correction, although these could easily be compensated by increased sampling size. (Kim S. et al., 2007)

In previously conducted studies, GWAS techniques have been used to identify specific SNPs, which are associated with the human reproductive behavior. 371 SNPs have been found to have an associated with the two particular reproductive behavioral parameters – the age at first sexual intercourse or AFS, and the age of the first child birth or AFB. One of the most important is the one associated with the elevated levels of C-Reactive Protein, a protein involved in variety of diseases in humans. For example, the CRP related genes with their linked genetic pathways and heritability may be associated to the specific genetic variants, relevant for reproductive human behaviors. (Vaez A. et al., 2015)

When one amino acid is replaced by another, a missense SNP, which is the type of nonsynonymous mutation (nsSNPs), occurs leading to a mutant protein with structural and functional impairments. This faulty protein could later cause onset of an illness. Finding SNPs that are pathogenic or associated to a specific phenotype's expression in individuals is one of the most challenging tasks for the research. (Dakal T. C. et al., 2017) Pleiotropy arises if a certain SNP linked to changes in the expression of the target genes. The nonsynonymous missense SNPs commonly associated with not only with the pleiotropic effect on gene expression, but also as likely deleterious, thereby harmful. These SNPs are often analyzed through the expression quantitative trait loci (eQTL). (Gratten J. et al., 2016)

Many human gene expression levels are influenced by common DNA variants. The previously mentioned eQTLs are the genetic variants which govern that genetic regulation (see Figure 1). The subsequent difference in gene expression amongst individuals has been established as a determinant for phenotypic variation and vulnerability towards the onset of the complex genetic disorders. (Majewski J. et al., 2011) There are two mechanisms of action for DNA genetic variants on a given gene within the human genome. Cis- acting DNA variants modulate the gene expression levels at adjacent gene sites, while the trans-acting variants modulate the gene expression levels at the distal gene sites. (Cheung et al., 2009) (Xu Z. et al., 2017) Figure 1 shows the visual representation of the variant activity modes. Both the cis- and trans- acting variants are may have a profound genomic effect, however I am considering only significant cis-acting eQTLs since they interact in a more direct manner.

Since the link between these 371 SNPs and their biological-medical associations are still not well-established, the main goal of this study is to identify novel genes associated with the human reproductive behavior through in-silico sequencing and eQTL post-GWAS analyses. (Vaez A. et al., 2015)

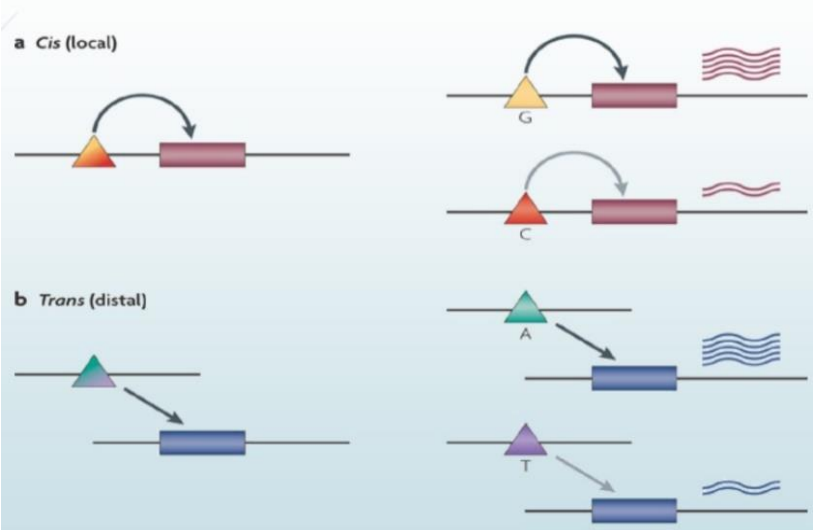


Figure 1:" Cis- and Trans- acting DNA variants". Changes in cis- and trans-acting DNA have different effect on gene expression levels. Polymorphic variants of regulators that operate in cis (a) or trans (b) (also known as local and distal regulators, respectively) to the target genes cause changes in target gene expression. Cis-acting variants are located near by the target genes, whilst trans-acting variants are found much farther away, sometimes on another chromosome.

The following research question is set to be answered – What are the most significant genes that are involved in human reproductive behavior, particularly in the context of AFB and AFS phenotypic traits?

To find and fully establish the link between the pathology and heritability of gene involved in a human reproductive behavior, one approach is to identify what are the functional SNPs (missense, pleiotropic and eQTLs) among the original 371 GWAS SNPs (gSNPs) as well as SNPs in linkage disequilibrium (LD) with these gSNPs. Identification of functional SNPs will allow us to create a list of prioritized genes underlying the GWAS signals. This list of genes could be the foundation for future laboratory studies and that may result in a therapeutic applications and implementation of knowledge on the most relevant SNPs with a possible pathogenic effect on human reproductive behavior.

Methods

For this project, the open-source software R was used for the post GWAS analysis of the 371 gSNPs.

Open-source software is software that has been published freely or is publicly shared and is accessible to everyone (Corbly J. E. et al., 2014). Due to these advancements in the software industry, open-source software provides a unique outlook to big data analysis for the studies conducted by academics from public institutions to scientific research with R programming (Thorbergsson H. et al., 2007).

R is a data processing, analysis, and visualization programming language (Krotov V. et al., 2017). The most significant aspects of R and R-Studio, which is a graphical user interface developed explicitly for R using the webserver or an app (Rstudio.com, 2011), are similar to those of other computation languages. R studio allows statistical assessment and in-/out-put instructions. The R environment includes features such as graphical utilities for data retrieval, manipulation, and storage, among many others. This computer software is especially well-suited for data analysis (Krotov V. et al., 2017) since it assists in handling a large set of instructions simultaneously, stores all data and progress, and allows analysts to effortlessly fix minor errors (Zumel N. et al., 2016).

To analyze the 371 SNPs, R-studio required an additional package known as SNPannotator used only in the first phase of the experiment.

The SNPannotator package can be found in the CRAN repository, which is a list of installable packages in the R-studio environment. An advantage of using R-studio is that it provides access to all transcripts, with no functional or practical bias. Therefore, genes could be analyzed in a broader context and compared to components of a homologous gene ensemble or genes involved in similar physiological pathways (Aubourg & Rouzé, 2001).

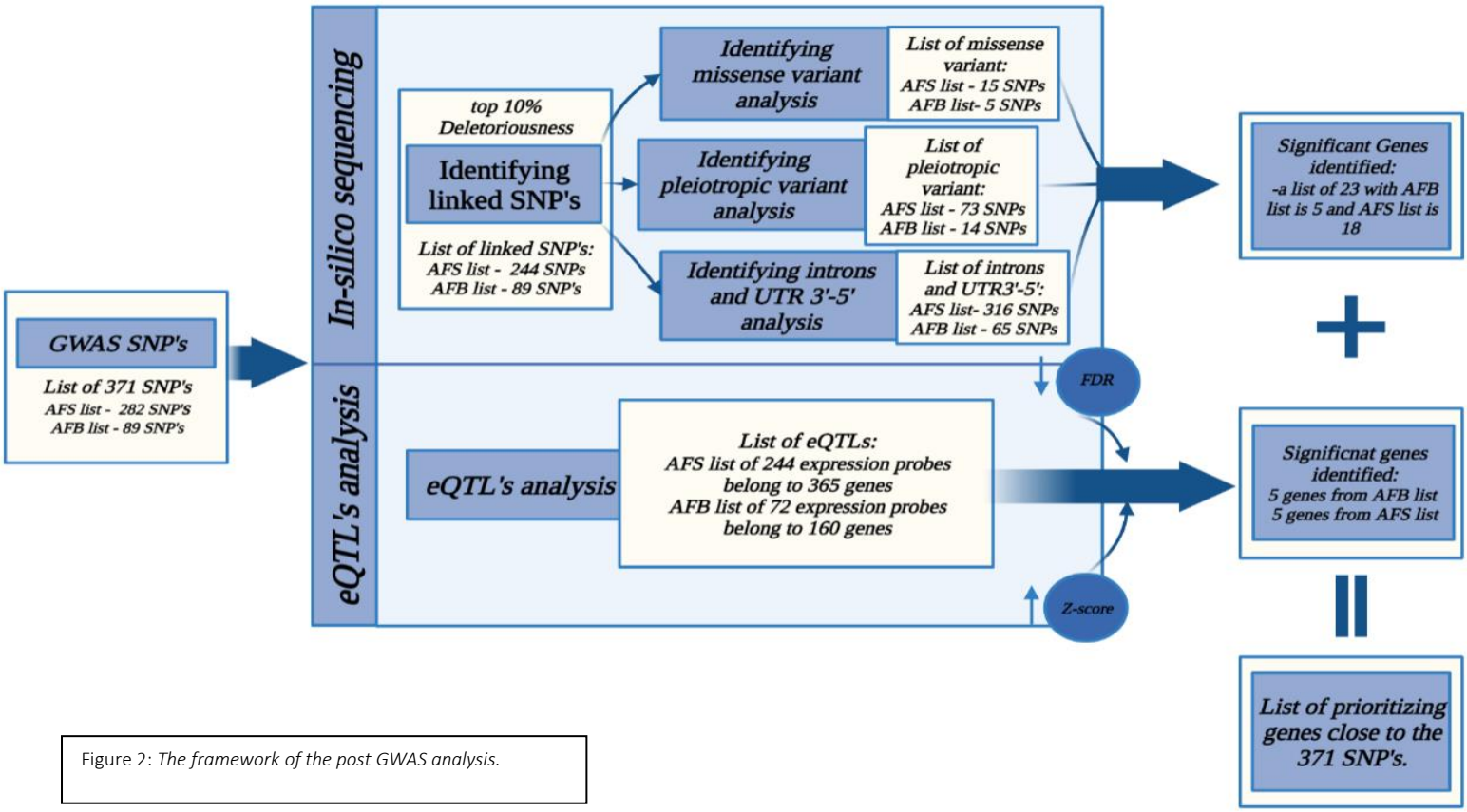


Figure 2: *The framework of the post GWAS analysis.*

This package requires data from the Ensembl server, which is a website that provides genomic information, containing genetic variants, multi-species alignments, orthologous and paralogous gene descriptions, and substantial polymorphism and regulatory data (Flicek et al., 2009). This server is essential to this study, as it provides useful insights into variation between the sample and the reference genome, facilitating the discovery of conserved regions (Crosswell & Thornton, 2014).

From the Ensemble server, two datasets were used to process of 371 SNPs which are:

- Gene_Names_Ensembl_104_GRCh8
- homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20210107

The first Gene Names dataset has a table of 57623 items and 6 columns related to SNP rs numbers, chromosomes, start-end locations, gene type, and gene name. Aside from the gene names of rs SNPs, the second dataset homo sapiens comprises of a table with 600908 elements and 5 columns, which are similar to the first dataset.

The study took a bioinformatics-based strategy, with two separate phases, each with several steps, as mentioned subsequently ***(Figure 2)***.

Phase 1: In-silico sequencing analysis

Identifying Linked Variants

The experiment began with a list of 371 SNPs acquired from a previous study (with 89 from AFB and 282 from AFS). The goal of this phase was to identify all of the SNPs linked to the 371 GWAS SNPs (gSNP's), which are SNP's found in the GWAS Catalog. The software SNPannotator uses rs numbers of the SNP's input to analyze the data, the list utilized in this study was filtered for SNPs lacking an rs number and the final number of AFB associated gSNPs were 72. The analyses also made use of data from the 1000 Genomes Project Full Phase 1 November 2010 release (using alignments from August 2010), which included European ancestry. After that, the r2 between each gSNP and all neighboring SNPs was determined as a marker of linkage disequilibrium (LD). LD is a single gene non-random alleles association at different loci. In the following stage of the study, only SNPs with high (r2>0.80) LD with the matching gSNP were utilized ***(Figure. 2)***.

Identifying Linked Nonsynonymous SNPs

All SNPs in LD with any of the gSNPs were evaluated using the SNPannotator package and then sorted sequentially BASED ON. More essential arguments include LDlist, which is a variable that, if set to TRUE, finds and adds variants with high LD to the output, and caad, which is an expression that, if set to TRUE, adds CADD scores to variant details. The Combined Annotation-Dependent Depletion (CADD) is a common measurement of variant deleteriousness that can efficiently highlight responsible mutations in genetic analyses, particularly those that are high drivers of severe Mendelian syndromes (Rentzsch et al., 2018).

The second stage was to generate three separate tables based on the type of variations (missense, introns and phenotypes) using the information from the previously created table. The remaining types were left out of the analysis.

The final step was to filter the tables using the CADD scores for deleteriousness, which is a measure of the negative effect of SNPs, to the top 10% of deleteriousness. This final step was taken to improve the analysis of the most relevant SNPs.

In Silico Pleiotropy Analysis

To broaden our understanding of the potential functions of the SNPs discovered in the preceding steps, the study filters the results identified in the previously generated table for age at first birth (AFB) and age of first intercourse (AFS). These parameters (AFS, AFB) were taken into account due to their significance in human reproductive behaviour. The final step in the research was to search on the University of California Santa Cruz (UCSC) Genome Project website for genes that are close to the SNPs reported in the three tables found in the Data Supplement ***(Tables 1,2,3)***. The genes discovered will be useful in establishing patterns between SNPs and determining the activities of genes relevant to human reproductive behaviour.

Phase 2: eQTLs analysis

Expression-Quantitative-Trait-Locus, also known as eQTL analysis is carried out in order to see whether the associated SNPs, previously identified by the post GWAS analysis.

There are two mechanisms of action for DNA genetic variants on a given gene. *Cis*- acting DNA variants modulate the gene expression levels at adjacent gene sites, while the *Trans*- acting variants modulate the gene expression levels at the distal gene. (Cheung et al., 2009). The research is going to focus on the expression levels in an adjacent gene, hence in a cis- manner. The list of 10,507,664 significant cis-eQTLs was downloaded from the eQTLGen portal, which contains data on 31,684 individuals' levels of gene expression in blood.

The cis-eQTL list was compared with the AFS and AFB lists of 371 SNPs to determine which gSNPs influencing

expression of nearby genes. Then, the top 5 linked SNPs of the highest significance were chosen. On the basis of having the lowest False Discovery Rate, FDR, and the highest absolute Z-score, the top 5 SNPs were chosen. Z- score is determined by subtraction of the total average gene frequency out of each raw expression of the gene. Then, this result is divided by the standard deviation (SD) from all counts obtained throughout all samples. (E. Khabirova, 2017)

To sum up, these top five SNPs are the most likely to alter expression of nearby genes.

Results

Phase 1: Post-GWAS In-Silico Sequencing Analysis

The GWAS analysis on human reproductive behavior originally identified 282 SNPs related to the age of first sexual intercourse (AFS) and 89 SNPs related to the age of first child birth (AFB) (Mills M et al. 2021). Since there are two characteristics, the original dataset was split into two sets and this part of the project is focused on the SNPs associated with AFS.

Composed table of linked SNPs (SNPAnnotator dataset) with dataset of 282 gSNPs, originally consisted of 19,471 outputs, with the linkage disequilibrium (LD) threshold value of r^2 =0.5. To narrow down the large dataset, the LD threshold was changed to r^2 = 0.8. Changing the r^2 decreased the sample data to 7,525 outputs, with already computed CADD scores.

Data was further processed via applying the deleteriousness level of below 10%, which filtered out 244 AFS associated probes. The 244 AFS gSNP dataset was further sorted out based on the types of SNP, with only missense and intronic being considered.

First, two tables, one with 15 missense and the other one with 316 intronic and 5'-3'-UTR SNPs were composed. Secondly, a table listing all of the pleiotropic associations of linked gSNPs from the 244 dataset was made, which contained 73 phenotypic associations. The pleiotropic table included some overlap with both the missense and intronic tables. Proceeding that, the SNPs rs-id numbers of linked gSNPs were used to find the associated genes and their names on the UCSC genome Browser. Finally, the five gSNPs with the highest CADD score and LD value closest to 1 were chosen to compose a top 5 missense gSNPs table, represented by the figure 3.

Phase 2: Post-GWAS eQTL Analysis

For the eQTL analysis, the large population dataset from the eQTLGen portal was downloaded and compared with the list of 244 linked gSNPs.

Figure 3: "Diagram of top 5 gSNPs from In-Silico Sequencing", the highest CADD score, also shows for each variant allele. Outer circle shows chromosome locations highlighted with different colors. Then there are gene names, then rs-ids of SNPs in LD with gSNPs within the blue circle, CADD score and respective alleles are on peach colored circle. In the central yellow cercle three are phenotypic associations of these five prioritized genes.

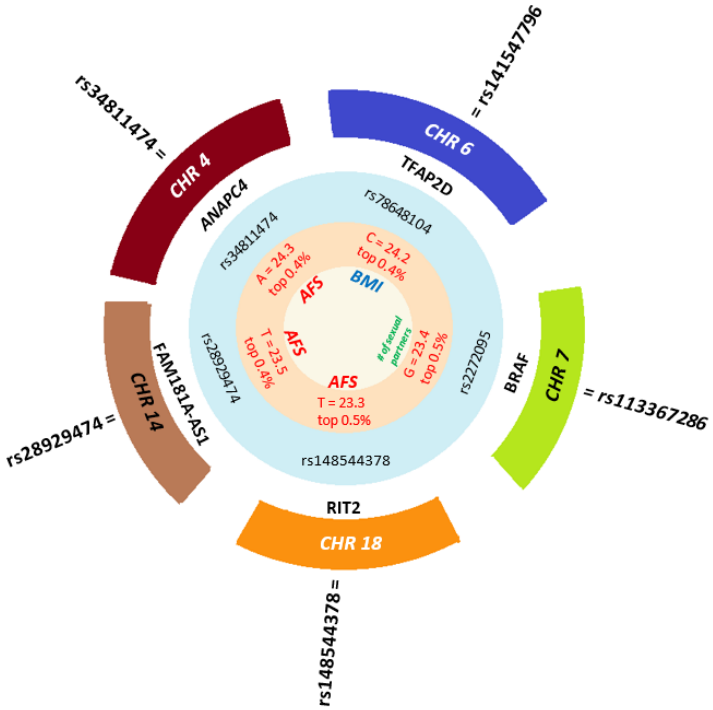
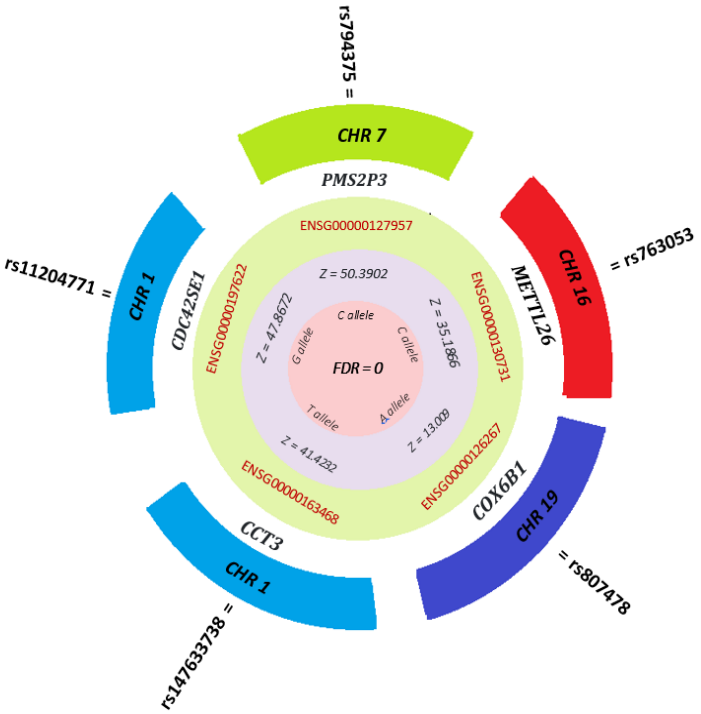


Figure 4:" Diagram of top 5 most significant gSNPs for eQTL". Outer circle shows chromosome locations highlighted with different colors. Then there are gene names, then gene-ids within the green circle, Z-scores are on purple circle. In the central pink cercle three is alleles for different variants and the FDR value is zero for all five gene associations.



The auxiliary code in R studio (see Appendix) produced a list of 365 gene association based on the 244 gSNP probes. Then, the missense, pleiotropic and intronic gSNPs tables were made, similar to those in Phase 1.

The top 5 eQTLs of the missense table were chosen based on the highest Z-score and with a zero or close to zero score of False Discovery Rate (FDR). The table, shown by the figure 4 was composed.

Discussion

In the present study, the post-GWAS analysis of 282 gSNPs, that have been found to be associated with the AFS reproductive behavior traits was conducted. In the first phase of the study, in-silico sequencing found 244 association between gSNPs and SNPs in linkage disequilibrium (LD) with them and then evaluating the top 10% deleterious SNPs. This dataset was used to create missense and intronic SNP lists, however, only the missense table was considered for selection of the top five significant gene associations, since the missense SNPs are known to possess detrimental deleterious character. In the second phase, the eQTL analysis, which was aimed at finding associations between 371 linked-gSNPs and the most significant cis-eQTLs. Resulting list of 365 eQTLs was processed based on the highest z-score and zero FDR. Moreover, similar to phase one, the list of five most significant gSNPs was evaluated further.

In the following paragraphs each of five gSNP-gene associations from the first phase of the study will be discussed in terms of gene function, potential deleterious effect as well as its phenotypic associations.

Phase 1: In Silico Sequencing Analysis of most significant missense gSNPs

Reproductive Behavior – Gene Associations

As a result of In-Silico Sequencing Analysis, all five identified gSNPs have a CADD score of above twenty, implying that all of these polymorphisms are predicted to have the most deleterious impact on the gene expression of their target genes. Moreover, each of the five gSNPs has been linked to one or more phenotypic associations. Four gSNPs have been linked to reproductive behavior characteristics such as the main subject of this study – Age of First Sexual Intercourse (AFS) and the Number of Sexual Partners (NSP). One of the gSNPs has been linked to changes in BMI and other related health risks. Four reproductive behavior-associated gSNPs were linked to ANAPC4, BRAF, FAM181A-AS1, and RIT2 genes, with each of them having several various phenotypic associations. Nevertheless, the BRAF gene has only been linked to the NSP.

Based on the bibliography assessment, autism spectrum disorders (ASD), schizophrenia, and PD, at least to some degree, could all be leading to a lack of individual reproductive success. A large-scale study performed on schizophrenia suffering patients showed that these patients are roughly twice as prone to partake in significantly riskier and most frequently unprotected sexual intercourse. The study also depicted those schizophrenic individuals have a 60 % higher likelihood of contracting STDs. (Cournos F. et al., 2013)

During the following year, the first quantitative analysis of reproductive stoppage in households affected by ASD had has conducted, in order to identify the potential hereditary risk factors. Since the release of the research results, there has been substantial progress in genetic screening of ASD risk-posing genetic variants. When assessing birth rates for ASD-affected families were contrasted with the unaffected ASD group. It shows that the number of births in ASD-affected families is about 0.668 times smaller than that of the control group. (Hoffmann T. J. et al., 2014) Considering the presented arguments, there is a foundation aiding the conclusion, that neurological disorders can affect the success of human reproductive behavior. For instance, individual reproductive success is reduced due to abnormal or hostile behavior, which was caused by an individual unstable cognitive state.

Bibliography-Aided Results Analysis

Upon further assessment of the gSNPs rs-ids and their associated genes throughout the literature and via the GeneCards database, the presence of at least three of the five gSNPs were confirmed as possible risk factors for variety of disorders.

TFAP2D

One of the gSNPs tied to the TFAP2D gene, with an rs-id of rs141547796, has been linked to being potentially a risk factor for developing abnormalities in expression various phenotypic traits such as body mass index (BMI), systolic blood pressure, and even cognitive function, as well as educational achievement and an increased chance of developing ADHD (attention deficit & hyperactivity disorder). Since there was no precise match in bibliography for the rs141547796 gSNP linked to a TFAP2D gene, it was examined separately in light of CADD score-derived phenotypic associations from its post-GWAS in-silico sequencing analysis phase of the study.

One of the key phenotypic associations of this particular gSNP is its impact on the expression of the gene, responsible for the determination of BMI. The research conducted in 2013 by J.N. Painter et al. found that missense SNPs in the TFAP2D gene are linked to an onset of metabolic syndrome and an overall elevated BMI, which is consistent with my In-silico sequencing analysis results, as well as the CADD-score prediction of gSNP-gene phenotypic associations. Given that a

high BMI is often tied to an increased risk of endometrial cancer, a form of uterus cancer, the gSNP (rs141547796) within the TFAP2D gene may well be associated with this particular type of cancer. Thus, further research into this gSNP is required. (Painter J.N. et al., 2013) Both link to the development of female reproductive system conditions, and the association with the increased risk of oncology development related to mutations within the TFAP2D gene, as well as a high CADD score of 24. This indicates high deleterious effect of rs141547796 gSNP, thus, it may be concluded that this association fits well within the context of this study. As a suggestion for future research, this particular missense SNP within the TFAP2D gene could be explored as a potential genomic biomarker for the detection and early-stage treatment of female endometrial cancer.

RIT2

Protein product RIT2 is a member of the Ras protein superfamily of small GTPases, which is involved in a number of key cellular functions such as survival and cell differentiation. (Wennerberg K. et al., 2005) RIT2 gene has recently been found as a novel Parkinson's disease (PD) associated gene, as well as a candidate gene for other neurological and developmental disorders such as schizophrenia and autism. (Daneshmandpour Y. et al., 2018) Previously conducted research

Furthermore, merely this year, a new study on axial impairment following deep brain stimulation in Parkinson's disease identified a RIT2 gene SNP variant with the rs-id of rs148544378, with a CADD score of 31, which illustrates the highly deleterious character of this particular variant. (Visanji N. P. et al., 2022) This study supports the results of my post-GWAS analysis of linked-gSNPs, since the RIT2 gene gSNP variant has an identical rs-id. The phenotypic analysis identified this gSNP variant to be associated with AFS and NSP reproductive behavior traits. Associations for this gSNP include quite an unorthodox phenotypic characteristic measurement "Leisure sedentary behavior: television watching". The link between this variant being a potential risk factor for the onset of PD and simultaneously being involved in human reproductive behavior makes it an important matter of focus from the perspective of future clinical research.

ANAPC4

Another, yet even more significant variant gSNP with an rs-id rs34811474, was found to be associated with the anaphase-promoting complex subunit 4. ANAPC4 is an E3 ubiquitin ligase that regulates mitosis and mediates the G1 phase of the cell cycle as part of the anaphase-promoting complex/cyclosome (APC/C). Previous research on the ANAPC4 gene found that its SNP variants may be clinically implicated as biomarkers of the early-stage development of oral squamous cell carcinoma. (Diniz M.G. et al., 2015) In 2019, a study aimed at the identification of novel therapeutic targets for osteoarthritis through the GWAS analysis using the UK Biobank archive, with over 95% posterior probability, identified three causal nonsynonymous missense SNP variants. One of these three missense SNPs, with an rs-id rs34811474, was linked to an ANAPC4 gene. (Tachmazidou I. et al., 2019) Analysis of the phenotypic associations identified gSNP (rs34811474) variant that possess a highly pleiotropic effect, since it is associated not only with AFS along with BMI, but additionally, it has been linked to several cognitive performance traits, as well as a variety of chronic conditions. Furthermore, this gSNP variant has the highest CADD score of 24.3, pointing at the significance of its highly deleterious character. A summarized amount of data strongly suggests that gSNP (rs34811474) variant of the ANAPC4 gene has to be considered to be of prime importance from the perspective of future research, as well as potential clinical implementation, especially considering this variant's potential in the diagnostics of oral squamous cell carcinoma.

BRAF

B-Raf proto-oncogene, serine/threonine kinase is a member of the RAF family of protein kinases, which are key players in the MAPK signaling pathway that governs cellular proliferation and differentiation. (Ciampi R. et al., 2005) From 50% to 70% of all human cancer melanomas are caused by the gain-of-function mutation within the BRAF oncogene region. (Garnett M. J. et al., 2004). Post-GWAS In-Silico Sequencing only states a single phenotypic association found for this missense gSNP. Nevertheless, single missense phenotypic association, being a number of sexual partners (NSP), which is an important reproductive behavior determinant. Additionally, there have been records of a variety of disorders linked to deleterious SNP within this oncogene, such as cardiofaciocutaneous, Leopard, and Noonan syndromes, as well as posing as onset risk factors of both colorectal and lung cancers. BRAF gene has a CADD score of C = 23, which signifies the very deleterious character of this gSNP, the rs-id is rs113367286, within the BRAF proto-oncogene region.

FAM181A-AS1

FAM181A Antisense RNA-1, abbreviated as FAM181A-AS1, is the RNA gene, part of the Long non-coding RNA class. There are a few types of cancer associated with missense SNPs within the FAM181A-AS1 gene. Since the missense SNP disrupts normal FAM181A-AS1 gene function, it then loses control over

the cell cycle. This frequently promotes the growth and division of thyroid cancer cells. During the de novo thyroid cancer pathway research, the gene mutation could also be used to a benefit, since the missense SNP variant within a gene could serve as a diagnostic marker. (Tian J. et al., 2020)

The phenotypic analysis states that this particular gene possesses a very strong pleiotropic character since there are over 35 various phenotypic associations. Among these associations, are the AFS, as well as determinants of various sex hormones like testosterone, or a variety of physical trait associations, such as BMI/Free-Fat-Mass and blood pressure, or the important association with determinants of breast size. Obviously, there are quite a few associations with a health risk, especially for the lifestyle and addictions association like alcohol consumption and smoking. Lastly, it is worth pointing out that there are a few associations, that were not mentioned here, and some of these associations are various blood nutrient content and internal organs health, with three separate liver-disruptions associations.

GeneCards database identified a single association of the FAM181A-AS1 missense SNP variant, being the increased glioma susceptibility. The majority of potentially harmful FAM181A-AS1 variant is stored within the testis. Although, the testis is directly linked to human reproductive behavior, possibly even playing a role in reproduction success and long-term survival. (Dessen P., 2014)
According to Liu Z. Q. et al. research done merely last year, the missense SNP within the FAM181A-AS1 gene will most likely lead to loss of control over the expression, thereby it was proposed that downregulated FAM181A Antisense RNA-1 influences the development of BRCA mutations in elderly people. With age, BRCA loses its heterozygous character and becomes as well susceptible to transformation into breast cancer in the elderly. (Liu Z. Q. et al., 2021)

Phase 2: eQTL Analysis Interpretation

In the following section, I will be discussing the outcomes of the eQTL analysis, specifically on three out of the five most significant gene associations since these have the highest Z-score. Notably, the 3 most significant genes, CDC42SE1, PMS2P3, and COX6B1, were all linked to an increased risk of cancer onset.

CDC42 is a GTPase protein encoded by the CDC42SE1 gene. This protein's main function is to increase cell adhesion and spreading, which in turn promotes ECM and cell actin cytoskeleton remodeling. (Price L. S. et al., 1998) The CDC42SE1 gene promotes carcinogenesis when it is downregulated, making it an effective diagnostic biomarker for skin cancer treatment. (Kalailingam P. et al., 2019)

The PMS2P3 gene is a product of the PMS2 gene, which has been attributed to the development of breast cancer and is tied to a poor prognosis. (Wang X. et al., 2022) Out of the five most significant linkages, this gene had the highest Z-score of 50.4, indicating that the rs794375 gSNP was the most important for this gene and could be considered for the future clinical-based research.

The human respiratory chain is comprised with several subunits, features cytochrome c oxidase, encoded by the COX6B1 gene. It was found that a mutation in the Cox6B1 gene can cause COX deficiency and a reduction in the protein's overall regulatory function. As a consequence, the COX6B1 gene has been associated with the increased risk of colorectal cancer. (Lascorz J. et al., 2012)

Conclusion

Since the main aim of this study was to find a link between pathology and the heritability of genes involved in human reproductive behavior, it was possible to conclude that the majority of the genes were indeed found to be triggers of various disorders. Although, further research is needed in order to confirm and test whether there is a certain connection between the increased disease risk and single nucleotide polymorphisms within the GWAS-associated gene, which are involved in human reproductive behavior.

Critic of the Study & Future Outlook

Gene enrichment analysis is be done, in order to compose a complete picture of the genome-wide interactions of the key identified genes. However, this most important part of the study, which would also be the third and concluding part of this research was not performed, due to time limitations. Therefore, the future outlook of this research would be a completion of the gene enrichment analysis phase and identification of the most significant interactions and pathways, of genes involved with human reproductive behavior, specifically in the context of the AFS and AFB phenotypical traits.

References

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics, 101(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Anderson, C. A. et al. Data quality control in genetic case–control association studies. Nat. Protoc. 5, 1564–1573 (2010).

Xu, Z., Wu, C., Wei, P., & Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. Genetics, 207(3), 893-902.

Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends in Genetics, 27(2), 72-79.

Vaez, A., Jansen, R., Prins, B. P., Hottenga, J. J., De Geus, E. J., Boomsma, D. I., Penninx, B. W., Nolte, I. M., Snieder, H., & Alizadeh, B. Z. (2015). In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons. Circulation: Cardiovascular Genetics, 8(3), 487–497. https://doi.org/10.1161/circgenetics.114.000714

Kim, S., & Misra, A. (2007). SNP Genotyping: Technologies and Biomedical Applications. Annual Review of Biomedical Engineering, 9(1), 289–320. https://doi.org/10.1146/annurev.bioeng.9.060906.152037

Dakal, T. C., Kala, D., Dhiman, G., Yadav, V., Krokhotin, A., & Dokholyan, N. V. (2017). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in IL8 gene. Scientific reports, 7(1), 6525. https://doi.org/10.1038/s41598-017-06575-4

Gratten, J., & Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome medicine, 8(1), 78. https://doi.org/10.1186/s13073-016-0332-x

Krotov, Vlad. (2017). A Quick Introduction to R and RStudio. 10.13140/RG.2.2.10401.92009. https://www.researchgate.net/publication/321111041_A_Quick_Introduction_to_R_and_RStudio\

Rstudio.com. (2011). *RStudio, new open-source IDE for R*. Available at: <https://www.rstudio.com/blog/rstudio-new-open-source-ide-for-r/>

Efdal Kaya, Muge Agca, Fatih Adiguzel & Mehmet Cetin (2019) Spatial data analysis with R programming for environment, Human and Ecological Risk Assessment: An International Journal, 25:6, 1521-1530, DOI: 10.1080/10807039.2018.1470896

Thorbergsson, Helgi & Björgvinsson, Tryggvi & Valfells, Ársæll. (2007). Economic benefits of free and open source software in electronic governance. 183-186. 10.1145/1328057.1328095.

Corbly, James Edward (25 September 2014). "The Free Software Alternative: Freeware, Open Source Software, and Libraries". *Information Technology and Libraries*. **33** (3): 65. doi:10.6017/ital.v33i3.5105. ISSN 2163-5226\

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., … Searle, S. M. (2010). Ensembl's 10th year. *Nucleic acids research*, *38*(Database issue), D557–D562. https://doi.org/10.1093/nar/gkp972

Nina Zumel, (March 2016), Preparing data for analysis using R, Win-Vector LLC-Microsoft, 2, https://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-Whitepaper-Data-Prep-Using-R.pdf#:~:text=One%20of%20the%20advantages%20of%20data%20analysis%20in,run%20summary%28%29%20on%20our%20data%20and%20get%20this%3A1

Aubourg, S., & Rouzé, P. (2001). Genome annotation. *Plant Physiology and Biochemistry*, *39*(3-4), 181–193. https://doi.org/10.1016/s0981-9428(01)01242-6

Ani, A., Kamali, Z., & Vaez, A. (2022, April 27). *SNPannotator: Investigating the Functional Characteristics of Selected SNPs and Their Vicinity Genomic Region*. R-Packages. https://cloud.r-project.org/web/packages/SNPannotator/index.html

L.C. Crosswell, J.M. Thornton, 6.12 - EBI and ELIXIR, Anders Brahme, Comprehensive Biomedical Physics, Elsevier, 2014, Pages 175-190,ISBN 9780444536334, https://doi.org/10.1016/B978-0-444-53632-7.01123-0. (http ://www.sciencedirect.com/science/article/pii/B9780444536327011230)

Cheung, V. G., & Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature reviews. Genetics*, *10*(9), 595–604. https://doi.org/10.1038/nrg2630

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, [online] 47(D1), pp.D886–D894. doi:10.1093/nar/gky1016.

E. Khabirova. (2017). Review of RNA-seq normalization methods. Software Solutions for Life Science Data Management and Integration. https://genestack.com/news/blog/review-of-rna-seq-normalisation-methods/#:%7E:text=Z%2Dscore%3A%20calculated%20by%20subtracting,measured%20counts%20across%20all%20samples.

Diniz, M.G., de Fatima Correia Silva, J., de Souza, F.T.A. *et al.* Association between cell cycle gene transcription and tumor size in oral squamous cell carcinoma. *Tumor Biol.* 36, 9717–9722 (2015). https://doi.org/10.1007/s13277-015-3735-1

Painter, J. N., Macgregor, S., Tomlinson, I., Nyholt, D. R., Zondervan, K., Thompson, D., Dunning, A., Easton, D., Montgomery, G. W., & Spurdle, A. B. (2013). Abstract 3182: A GWAS-based cross-disease approach suggests genes predisposing to risk of endometriosis and endometrial cancer. *Molecular and Cellular Biology*. https://doi.org/10.1158/1538-7445.am2013-3182

Ciampi, R., & Nikiforov, Y. E. (2005). Alterations of the <I>BRAF</I> Gene in Thyroid Tumors. *Endocrine Pathology*, *16*(3), 163–172. https://doi.org/10.1385/ep:16:3:163

Daneshmandpour, Y., Darvish, H. & Emamalizadeh, B. *RIT2*: responsible and susceptible gene for neurological and psychiatric disorders. *Mol Genet Genomics* **293,** 785–792 (2018). https://doi.org/10.1007/s00438-018-1451-4

Tian, J., & Luo, B. (2022). Identification of Three Prognosis-Related Differentially Expressed lncRNAs Driven by Copy Number Variation in Thyroid Cancer. *Journal of Immunology Research*, *2022*, 1–18. https://doi.org/10.1155/2022/9203796

Diniz, M.G., de Fatima Correia Silva, J., de Souza, F.T.A. *et al.* Association between cell cycle gene transcription and tumor size in oral squamous cell carcinoma. *Tumor Biol.* **36,** 9717–9722 (2015). https://doi.org/10.1007/s13277-015-3735-1

Warren, H., Evangelou, E., Cabrera, C. *et al.* Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet* **49,** 403–415 (2017). https://doi.org/10.1038/ng.3768

Tachmazidou, I., Hatzikotoulas, K., Southam, L. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat Genet* **51,** 230–236 (2019). https://doi.org/10.1038/s41588-018-0327-1

Visanji, N. P., Ghani, M., Yu, E., Kakhki, E. G., Sato, C., Moreno, D., Naranian, T., Poon, Y. Y., Abdollahi, M., Naghibzadeh, M., Rajalingam, R., Lozano, A. M., Kalia, S. K., Hodaie, M., Cohn, M., Statucka, M., Boutet, A., Elias, G. J., Germann, J., Fasano, A. (2022). Axial Impairment Following Deep Brain Stimulation in Parkinson's Disease: A Surgicogenomic Approach. *Journal of Parkinson's Disease*, *12*(1), 117–128. https://doi.org/10.3233/jpd-212730

Wennerberg, K., Rossman, K. L., & Der, C. J. (2005). The Ras superfamily at a glance. *Journal of Cell Science*, *118*(5), 843–846. https://doi.org/10.1242/jcs.01660

Liu, Z. Q., Zhang, G. T., Jiang, L., Li, C. Q., Chen, Q. T., & Luo, D. Q. (2021). Construction and Comparison of ceRNA Regulatory Network for Different Age Female Breast Cancer. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.603544

Price, L. S., Leng, J., Schwartz, M. A., & Bokoch, G. M. (1998). Activation of Rac and Cdc42 by Integrins Mediates Cell Spreading. *Molecular Biology of the Cell*, *9*(7), 1863–1871. https://doi.org/10.1091/mbc.9.7.1863

Kalailingam, P., Tan, H. B., Pan, J. Y., Tan, S. H., & Thanabalu, T. (2019). Overexpression of CDC42SE1 in A431 Cells Reduced Cell Proliferation by Inhibiting the Akt Pathway. *Cells*, *8*(2), 117. https://doi.org/10.3390/cells8020117

Wang, X., Chen, H., Kapoor, P. M., Su, Y. R., Bolla, M. K., Dennis, J., Dunning, A. M., Lush, M., Wang, Q., Michailidou, K., Pharoah, P. D., Hopper, J. L., Southey, M. C., Koutros, S., Freeman, L. E. B., Stone, J., Rennert, G., Shibli, R., Murphy, R. A., . . . Lindström, S. (2022). A Genome-Wide Gene-Based Gene–Environment Interaction Study of Breast Cancer in More than 90,000 Women. *Cancer Research Communications*, *2*(4), 211–219. https://doi.org/10.1158/2767-9764.crc-21-0119

Massa, V., Fernandez-Vizarra, E., Alshahwan, S., Bakhsh, E., Goffrini, P., Ferrero, I., Mereghetti, P., D'Adamo, P., Gasparini, P., & Zeviani, M. (2008). Severe infantile encephalomyopathy caused by a mutation in COX6B1, a nucleus-encoded subunit of cytochrome c oxidase. *American journal of human genetics*, *82*(6), 1281–1289. https://doi.org/10.1016/j.ajhg.2008.05.002

Cournos F., Coomaraswamy S., Guido J.R., Meyer-Bahlburg H.F.; Sexual activity and risk of HIV infection among patients with schizophrenia. Am J. Psychiatry. 2013. https://doi.org/10.1176/ajp.151.2.228

Hoffmann, T. J., Windham, G. C., Anderson, M., Croen, L. A., Grether, J. K., & Risch, N. (2014). Evidence of Reproductive Stoppage in Families with Autism Spectrum Disorder. *JAMA Psychiatry*, *71*(8), 943. https://doi.org/10.1001/jamapsychiatry.2014.420

Bronner, G., Hassin-Baer, S., & Gurevich, T. (2017). Sexual Preoccupation Behavior in Parkinson's Disease. *Journal of Parkinson's Disease*, *7*(1), 175–182. https://doi.org/10.3233/jpd-160926

Dhomen, N., Reis-Filho, J. S., Da Rocha Dias, S., Hayward, R., Savage, K., Delmas, V., Larue, L., Pritchard, C., & Marais, R. (2009). Oncogenic Braf Induces Melanocyte Senescence and Melanoma in Mice. *Cancer Cell*, *15*(4), 294–303. https://doi.org/10.1016/j.ccr.2009.02.022

Garnett, M. J., & Marais, R. (2004). Guilty as charged: B-RAF is a human oncogene. *Cancer cell*, *6*(4), 313-319.

Liu, F., Ma, D., Chen, W., Chen, X., Qian, Y., Zhao, Y., Hu, T., Yin, R., Zhu, Y., Zhang, Y., Zhang, Y., & Zhao, W. (2019). Gold Nanoparticles Suppressed Proliferation, Migration, and Invasion in Papillary Thyroid Carcinoma Cells via Downregulation of CCT3. *Journal of Nanomaterials*, *2019*, 1–12. https://doi.org/10.1155/2019/1687340

Lascorz, J., Bevier, M., Schönfels, W. V., Kalthoff, H., Aselmann, H., Beckmann, J., Egberts, J., Buch, S., Becker, T., Schreiber, S., Hampe, J., Hemminki, K., Försti, A., & Schafmayer, C. (2012). Polymorphisms in the mitochondrial oxidative phosphorylation chain genes as prognostic markers for colorectal cancer. *BMC Medical Genetics*, *13*(1). https://doi.org/10.1186/1471-2350-13-31

Dessen P., *FAM181A-AS1 (FAM181A antisense RNA,* Atlas Genet Cytogenet Oncol Haematol. 2014-11-01. Online version: http://atlasgeneticsoncology.org/gene/63085/

GeneCards the human gene database. https: www.genecards.org
Safran M, Rosen N, Twik M, BarShir R, Iny Stein T, Dahary D, Fishilevich S, and Lancet D. *The GeneCards Suite* Chapter, *Practical Guide to Life Science Databases* (2022) pp 27-56 [PDF]'

Supplements

Scripts

Scripts used for Phase 1 post-GWAS In-Silico sequencing analysis:

```
setwd("C:/Users/Daniil/Desktop/B.R.P.35.#2 ")

install.packages("SNPannotator")

# load the library
library(SNPannotator)

# select server for GRCh38 or GRCh37
# server <- "https://grch37.rest.ensembl.org" ### GRCh37
server <- "https://rest.ensembl.org" ### GRCh38

# select database population for LD calculation
# db <- "1000GENOMES:phase_3:ALL" ### all samples in 1000G study
db <- "1000GENOMES:phase_3:EUR" ### European super population in 1000G study

# create a vector from variant rs numberss
rslist=c('rs1962545','rs803679','rs7533341','rs1392816','rs111991969','rs7525548','rs77214504','rs140681455','rs141655075','rs1156981','rs10922907','rs1931262','rs1146566'
,'rs2274568','rs79764489','rs11204771','rs113142203','rs147633738','rs11240331','rs10157166','rs6586405','rs1320330','rs2091377','rs2014149','rs138850767','rs12463727'
,'rs4952343','rs35508442','rs985919','rs1516172','rs9789483','rs6719762','rs359243','rs70959844','rs62180269','rs11688027','rs1368546','rs145846361','rs10496949','rs1226414'
,'rs12692596','rs11678980','rs10165889','rs13387970','rs13009323','rs7575189','rs56306056','rs147725178','rs76253389','rs76536952','rs875097','rs6748341','rs9809849'
,'rs2084572','rs6550942','rs114456303','rs2278480','rs67723420','rs562462868','rs2188151','rs2612029','rs186723454','rs6764919','rs6445264','rs7618715','rs4334682'
,'rs2317603','rs12714592','rs112523595','rs6797231','rs12714702','rs369789482','rs57945129','rs705240','rs56392241','rs752703112','rs767462343','rs34811474','rs702'
,'rs72631060','rs7671317','rs993700','rs10516875','rs11729080','rs809955','rs435538','rs12517438','rs8185308','rs7381195','rs71301503','rs12653396','rs34073570','rs2406374'
,'rs35080996','rs561029885','rs2910032','rs1298310','rs4868800','rs11955430','rs245753','rs12203592','rs767943','rs7746553','rs141547796','rs222440','rs370705844','rs1925686'
,'rs2397678','rs72990858','rs28858382','rs13216871','rs12204714','rs369498508','rs7785195','rs56066200','rs198310','rs35851551','rs10233473','rs12701263','rs794375'
,'rs540866996','rs13307225','rs2694934','rs7783012','rs11767283','rs11772444','rs6966898','rs113367286','rs1991651','rs7008955','rs7828172','rs2923407','rs7824756'
,'rs1585634','rs10104523','rs11382985','rs372420182','rs72674824','rs10955084','rs9643087','rs13280592','rs8180995','rs7033296','rs4961705','rs12554512','rs10746578'
,'rs10992812','rs777352006','rs9886840','rs11255903','rs11458658','rs2093623','rs2650705','rs61856978','rs3896224','rs10886022','rs7897631','rs7079070','rs1866710'
,'rs10835387','rs34804222','rs780814398','rs199507956','rs4439537','rs590414','rs111709606','rs34880764','rs11428242','rs10743299','rs10770452','rs35179240','rs202149107'
,'rs285582','rs59957074','rs7955865','rs7972441','rs2279574','rs752331861','rs9554165','rs7987501','rs61960829','rs9538248','rs341521','rs2174752','rs10646652','rs12894029'
,'rs139881447','rs34902169','rs12147463','rs3007104','rs12878359','rs551086366','rs10134692','rs28929474','rs76715069','rs783544','rs4702','rs1014403','rs763053','rs9923553'
,'rs2870488','rs4985127','rs7188873','rs12448731','rs76513770','rs12446652','rs200005647','rs11866420','rs410520','rs3853548','rs28406364','rs6504551','rs7503604','rs4800204'
,'rs2849767','rs56393977','rs148544378','rs34155040','rs9964201','rs776081653','rs7236339','rs10853981','rs4804512','rs807478','rs117831144','rs2145108','rs6058613'
,'rs35852264','rs1609598','rs4809346','rs375909440','rs7473421','rs62599791','rs146852038','rs6637831','rs961522','rs7608187','rs34481141','rs7024334','rs11038866'
,'rs76702070','rs590648','rs11392435','rs1435757')

# run the pipeline
# the result will be returned as a data frame and also saved as an excel filers2174752
''
# fetch information for the rslist, add cadd score and regulatory type
output <- annotate(rslist,server,db, 'sampleOutput.xlsx',
                   LDlist = FALSE,
                   cadd = TRUE,
                   geneNames.file = 'Gene_Names_Ensembl_104_GRCh38 (1).rds',
                   regulatoryType.file = 'homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20210107.rds'
)
```

Script for the Phase 2 eQTL analysis:

```
setwd("C:/Users/Daniil/Desktop/B.R.P.35.#2 ")

install.packages("SNPannotator")

# load the library
library(SNPannotator)

# select database population for LD calculation
# db <- "1000GENOMES:phase_3:ALL" ### all samples in 1000G study
db <- "1000GENOMES:phase_3:EUR" ### European super population in 1000G study

# create a vector from variant rs numberss
rslist=c('rs1962545','rs803679','rs7533341','rs1392816','rs111991969','rs7525548','rs77214504','rs140681455','rs141655075','rs1156981','rs10922907','rs1931262','rs1146566'
,'rs2274568','rs79764489','rs11204771','rs113142203','rs147633738','rs11240331','rs10157166','rs6586405','rs1320330','rs2091377','rs2014149','rs138850767','rs12463727'
,'rs4952343','rs35508442','rs985919','rs1516172','rs9789483','rs6719762','rs359243','rs70959844','rs62180269','rs11688027','rs1368546','rs145846361','rs10496949','rs1226414'
,'rs12692596','rs11678980','rs10165889','rs13387970','rs13009323','rs7575189','rs56306056','rs147725178','rs76253389','rs76536952','rs875097','rs6748341','rs9809849'
,'rs2084572','rs6550942','rs114456303','rs2278480','rs67723420','rs562462868','rs2188151','rs2612029','rs186723454','rs6764919','rs6445264','rs7618715','rs4334682','rs2317603'
,'rs12714592','rs112523595','rs6797231','rs12714702','rs369789482','rs57945129','rs705240','rs56392241','rs752703112','rs767462343','rs34811474','rs702','rs72631060'
,'rs7671317','rs993700','rs10516875','rs11729080','rs809955','rs435538','rs12517438','rs8185308','rs7381195','rs71301503','rs12653396','rs34073570','rs2406374','rs35080996'
,'rs561029885','rs2910032','rs1298310','rs4868800','rs11955430','rs245753','rs12203592','rs767943','rs7746553','rs141547796','rs222440','rs370705844','rs1925686','rs2397678'
,'rs72990858','rs28858382','rs13216871','rs12204714','rs369498508','rs7785195','rs56066200','rs198310','rs35851551','rs10233473','rs12701263','rs794375','rs540866996'
,'rs13307225','rs2694934','rs7783012','rs11767283','rs11772444','rs6966898','rs113367286','rs1991651','rs7008955','rs7828172','rs2923407','rs7824756','rs1585634','rs10104523'
,'rs11382985','rs372420182','rs72674824','rs10955084','rs9643087','rs13280592','rs8180995','rs7033296','rs4961705','rs12554512','rs10746578','rs10992812','rs777352006'
,'rs9886840','rs11255903','rs11458658','rs2093623','rs2650705','rs61856978','rs3896224','rs10886022','rs7897631','rs7079070','rs1866710','rs10835387','rs34804222','rs780814398'
,'rs199507956','rs4439537','rs590414','rs111709606','rs34880764','rs11428242','rs10743299','rs10770452','rs35179240','rs202149107','rs285582','rs59957074','rs7955865'
,'rs7972441','rs2279574','rs752331861','rs9554165','rs7987501','rs61960829','rs9538248','rs341521','rs2174752','rs10646652','rs12894029','rs139881447','rs34902169','rs12147463'
,'rs3007104','rs12878359','rs551086366','rs10134692','rs28929474','rs76715069','rs783544','rs4702','rs1014403','rs763053','rs9923553','rs2870488','rs4985127','rs7188873'
,'rs12448731','rs76513770','rs12446652','rs200005647','rs11866420','rs410520','rs3853548','rs28406364','rs6504551','rs7503604','rs4800204','rs2849767','rs56393977'
,'rs148544378','rs34155040','rs9964201','rs776081653','rs7236339','rs10853981','rs4804512','rs807478','rs117831144','rs2145108','rs6058613','rs35852264','rs1609598','rs4809346'
,'rs375909440','rs7473421','rs62599791','rs146852038','rs6637831','rs961522','rs7608187','rs34481141','rs7024334','rs11038866','rs76702070','rs590648','rs11392435','rs1435757')

for (snp in rslist) {
  first <- snp==rslist[1]
  write.table(cis.eQTLs.txt[cis.eQTLs.txt$SNP == snp,], "eQTLS.txt",      col.names=first, row.names=F, quote=F, sep="\t", append=!first)
}
```

Tables

Supplementary table 1: Missense gSNPs for In-Silico Phase 1. *Following table consists of list of rs-ids of GWAS SNPs, as well as rs-ids of the SNPs in a LD with these gSNPs, chromosomes and positions within these chromosomes, CADD score, showing also an allele prediction, deleteriousness level and associations with linked SNP. All of these SNPs are missense ones. rs-ids and candidate gene names are highlighted in yellow. Blue color indicates chosen top five missense SNPs with the highest CADD score. Deleteriousness up to top 10%.*

gSNP	Linked_SNP	chr	Pos_37	LD	gene	Associations	cadd	Deleteriousness
rs11204771	rs4971007	1	151135661	0,8168	LYSMD1		G = 20.1	top 1.0%
rs147633738	rs11548200	1	156320865	0,980451	LYSMD1	Household income, red cell distribution width, Walking pace	C = 22.5	top 0.6%
rs875097	rs55760516	2	219489386	0,907658	PLCD4		G = 21.5	top 0.7%
rs12714702	rs9813894	3	88139270	0,949671	CGGBP1/ZNF654		A = 11.97	top 6.4%
rs12714702	rs7653652	3	88140191	0,949671	CGGBP1/ZNF654		C = 17.41	top 1.8%
rs34811474	rs34811474	4	25407216	1	ANAPC4	AFB, AFS, BMI, Body size at age 10, Cognitive aspects of educational attainment, Cognitive performance, Heel bone mineral density, Height, Highest math class taken, Intelligence, Leisure sedentary behavior television watching, Lung function, Male-pattern baldness, Menarche age at onset, Multisite chronic pain, Osteoarthritis, predicted visceral adipose tissue, Snoring, Urate levels, Verbal-numerical reasoning, White blood cell count	A = 24.3	top 0.4%
rs141547796	rs78648104	6	50715296	0,886266	TFAP2D	Attention deficit hyperactivity disorder, BMI, Cognitive performance, Cystatin C levels, educational attainment, Systolic blood pressure	C = 24.2	top 0.4%
rs794375	rs6947307	7	75494199	0,968015	RHBDD2		T = 19.7	top 1.1%
rs113367286	rs2272095	7	140459051	0,896095	BRAF	NSP	G = 23.4	top 0.5%

rs2279574	rs2279574	12	89351700	1	DUSP6	AFS, Cognitive performance Highest math class taken, Hypogonadotropic hypogonadism 19 with or without anosmia, Self-reported math ability	A = 22.9	top 0.5%
rs28929474	rs28929474	14	94378610	1	FAM181A-AS1	AFS, Alanine aminotransferase levels, Alanine transaminase levels, Alanine transaminase levels in high alcohol intake, Alcohol consumption (drinks per week), Alcohol consumption drinks per week, ALPHA-1-ANTITRYPSIN DEFICIENCY, Antineutrophil cytoplasmic antibody-associated vasculitis, Appendicular lean mass, Aspartate aminotransferase levels, Aspartate aminotransferase to alanine aminotransferase ratio, Aspartate transaminase levels in high alcohol intake, Bioavailable testosterone levels, Bitter alcoholic beverage consumption, Blood protein levels, Brain morphology MOS Test, Breast size, C-reactive protein levels, Calcium levels, Chronic obstructive pulmonary disease, Cirrhosis (alcohol related), Cirrhosis multi-trait analysis, Cystatin C levels, Direct bilirubin levels, Fat-free mass, FRAXE, Gallstone disease, Gamma glutamyl transferase levels, Glucagon levels in response to oral glucose tolerance test (fasting), Heel bone mineral density, Height, Hip circumference adjusted for BMI, Inborn genetic diseases, Insulin-like growth factor 1 levels, Liver enzyme levels (alanine transaminase), Liver enzyme levels (alkaline phosphatase), Liver enzyme levels (gamma-glutamyl transferase), Low density lipoprotein cholesterol levels, Metabolite levels (small molecules and protein measures), Osteoprotegerin levels, PI Z, PI Z(AUGSBURG), PI Z(TUN), Post bronchodilator FEV1, Post bronchodilator FEV1/FVC ratio, Post bronchodilator FEV1/FVC ratio in smoking, Post bronchodilator percent predicted FEV1 in smoking, Problematic alcohol use MTAG, Serum albumin level, Serum alkaline phosphatase levels, Serum alpha-fetoprotein levels, Serum total protein level, Sex hormone-binding globulin levels, Sex hormone-binding globulin	T = 23.5	top 0.4%

						levels adjusted for BMI, Systolic blood pressure, Testosterone levels, TNF-related apoptosis-inducing ligand levels, Total testosterone levels, Urea levels		
rs763053	rs1139897	16	670986	0,802361	RAB40C	Age of smoking initiation (MTAG), Smoking initiation (ever regular vs never regular), Smoking initiation (ever regular vs never regular) (MTAG)	A = 16.32	top 2.3%
rs148544378	rs148544378	18	42743602	1	RIT2	AFS, Leisure sedentary behavior television watching, NSP	T = 23.3	top 0.5%
rs807478	rs231591	19	35733804	0,873839	KMT2B		G = 11.41	top 7.2%
rs11038866	rs1317826	11	46366318	0,949319	DGKZ		G = 11.06	top 7.8%

Supplementary table 2: Phenotypes table of gSNPs for In-Silico Phase 1. *Following table consists of list of rs-ids of GWAS SNPs, as well as rs-ids of the SNPs in a LD with these gSNPs, chromosomes and positions within these chromosomes, CADD score, showing also an allele prediction, deleteriousness level and associations with linked SNP, and most importantly the phenotypic associations. All of these SNPs are of different type, as listed in a table. rs-ids and candidate gene names are highlighted in yellow. Deleteriousness up to top 10%.*

gSNP	Linked_SNP	Chr	Pos_37	LD	Gene	Type	Associations	cadd	Deleteriousness
rs7525548	rs12041912	1	74538026	0,995998	LRRIQ3	intron_variant	Body Mass Index	A = 10.16	top 9.6%
rs7525548	rs3895907	1	74540343	0,988036	LRRIQ3	intron_variant	Body Mass Index	G = 12.09	top 6.2%
rs7525548	rs1514175	1	74525960	0,892618	LRRIQ3	intron_variant	Body Mass Index, Smoking initiation, Smoking initiation (ever regular vs never regular)	G = 18.15	top 1.5%
rs7525548	rs3845345	1	74536983	0,861009	LRRIQ3	intron_variant	Body Mass Index	T = 12.33	top 5.8%
rs7525548	rs6703637	1	74532111	0,845463	LRRIQ3	intron_variant	Body Mass Index	A = 12.06	top 6.2%
rs140681455	rs140681455	1	77979081	1	AK5	5_prime_UTR_variant	Age at first sexual intercourse, Body shape index, Hip circumference adjusted for BMI, Leisure sedentary behavior computer use, Leisure sedentary behavior television watching, mean corpuscular volume, mean spheric corpuscular volume, Waist-hip index, Waist-to-hip ratio adjusted for BMI	GACCGG = 20.6	top 0.9%
rs141655075	rs141655075	1	87328913	1	SELENOF	5_prime_UTR_variant	Age at first sexual intercourse	T = 19.63	top 1.1%
rs10922907	rs12042107	1	90730619	0,996032	_	intergenic_variant	Educational attainment, Number of sexual partners, Smoking status (ever vs never smokers)	C = 22.2	top 0.6%
rs147633738	rs11548200	1	156320865	0,980451	CCT3	missense_variant	Household income MTAG, Red cell distribution width, Walking pace	C = 22.5	top 0.6%
rs6586405	rs1329125	1	234605134	0,99095	TARBP1	3_prime_UTR_variant	Educational attainment (MTAG), Educational attainment (years of education)	T = 16.17	top 2.4%
rs1320330	rs11127491	2	646145	0,93788	_	intergenic_variant	Body Mass Index	C = 12.34	top 5.8%
rs12463727	rs1631026	2	26730982	0,972412	OTOF	3_prime_UTR_variant	Adult body size	T = 17.65	top 1.7%
rs12463727	rs1731259	2	26730714	0,972412	OTOF	3_prime_UTR_variant	Body Mass Index	G = 10.72	top 8.5%
rs35508442	rs12998046	2	44653786	0,942716	CAMKMT	intron_variant	Chronotype	A = 10.81	top 8.3%
rs62180269	rs62180269	2	63093389	1	EHBP1	intergenic_variant	Age at first sexual intercourse	C = 14.18	top 3.8%
rs10496949	rs2381473	2	143394233	0,99189	_	intron_variant	Trauma exposure	A = 13.83	top 4.1%
rs11678980	rs11678980	2	161244750	1	RBMS1	non_coding_transcript_exon_variant	Age at first sexual intercourse, Cognitive aspects of educational attainment, Cognitive performance, Cognitive performance (MTAG), Cognitive traits MTAG, Educational attainment (MTAG), Educational	A = 12.69	top 5.4%

							attainment (years of education), Highest math class taken, Highest math class taken (MTAG), Self-reported math ability, Self-reported math ability (MTAG), Smoking initiation (ever regular vs never regular), Smoking initiation (ever regular vs never regular) (MTAG), Verbal-numerical reasoning		
rs875097	rs1050816	2	219493476	0,890128	PLCD4	3_prime_UTR_variant	Estimated glomerular filtration rate	T = 15.1	top 3.1%
rs114456303	rs114456303	3	24631667	1	THRB-AS1	intron_variant	Age at first sexual intercourse	A = 13.71	top 4.3%
rs2278480	rs2278480	3	25594252	1	RARB	intron_variant	Age at first sexual intercourse	C = 14.2	top 3.8%
rs186723454	rs186723454	3	54122571	1	_	5_prime_UTR_variant	Age at first sexual intercourse	G = 18.1	top 1.5%
rs6445264	rs6445264	3	62368750	1	_	intron_variant	Age at first sexual intercourse, Body Mass Index	A = 16.48	top 2.2%
rs112523595	rs34495106	3	85601986	0,995741	CADM2	intron_variant	Smoking status, Smoking status (ever vs never smokers)	G = 19.62	top 1.1%
rs57945129	rs62264764	3	117920728	1	_	intron_variant	Smoking initiation (ever regular vs never regular)	A = 14.02	top 4.0%
rs56392241	rs3851353	3	132170357	0,894299	DNAJC13	intron_variant	Core binding factor acute myeloid leukemia	G = 15.15	top 3.1%
rs34811474	rs34811474	4	25407216	1	ANAPC4	missense_variant	Adult body size, Age at first sexual intercourse, Body Mass Index, Body size at age 10, Cognitive aspects of educational attainment, Cognitive performance, Cognitive performance (MTAG), Educational attainment (MTAG), Educational attainment (years of education), General cognitive ability, Heel bone mineral density, Height, Highest math class taken (MTAG), Intelligence, Intelligence (MTAG), Leisure sedentary behavior television watching, Lung function (FVC), Male-pattern baldness, Menarche age at onset, Multisite chronic pain, Osteoarthritis, Predicted visceral adipose tissue, Self-reported math ability (MTAG), Snoring, Urate levels, Verbal-numerical reasoning, Waist circumference adjusted for body mass index, White blood cell count	A = 24.3	top 0.4%
rs11729080	rs72678864	4	111500989	0,984627	_	intergenic_variant	Age of smoking initiation (MTAG), Educational attainment (MTAG), Lifetime smoking index, Smoking cessation (MTAG), Smoking initiation (ever regular vs never regular) (MTAG), Smoking status, Smoking status (ever vs never smokers)	A = 20.3	top 0.9%
rs12653396	rs12653396	5	88551455	1	MEF2C-AS1	intron_variant	Age at first sexual intercourse, Attention deficit hyperactivity disorder or caudate nucleus volume (pleiotropy), Body Mass Index, Educational attainment (MTAG), Educational attainment (years of education), Highest math class taken (MTAG), Noncognitive aspects of educational attainment	A = 19.09	top 1.2%
rs11955430	rs11955430	5	167993291	1	PANK3	intron_variant	Age at first sexual intercourse	G = 12.73	top 5.3%

rs245753	rs7730898	5	171032671	0,802441	_	intron_variant	Body Mass Index, HDL cholesterol	A = 16.63	top 2.2%
rs12203592	rs12203592	6	396321	1	IRF4	intron_variant	Age at first sexual intercourse, Aging traits health span, parental lifespan or longevity multivariate analysis, Balding type 1, Basal cell carcinoma, Black vs. blond hair color, Black vs. red hair color, cutaneous squamous cell carcinoma, Eye color, Eye color (brightness), Eye color traits, Facial pigmentation, Feeling nervous, Freckling, Hair color, Hair greying, Hair morphology traits, Keratinocyte cancer MTAG, Low tan response, Lymphocyte counts, Male puberty timing age at voice breaking MTAG, Male puberty timing early vs. average onset facial hair, Male puberty timing late vs. average onset facial hair, Male-pattern baldness, Monobrow, Neuroblastoma, Nevus count, Non-melanoma skin cancer, Progressive supranuclear palsy, Rosacea symptom severity, Skin aging (microtopography measurement), Skin color saturation, Skin pigmentation, Skin pigmentation traits, Skin, hair and eye pigmentation (multivariate analysis), Skin/hair/eye pigmentation, variation in, 8, Smoking cessation, Smoking cessation (MTAG), Squamous cell carcinoma, Sunburns, Tanning, Vitiligo, White blood cell count, Youthful appearance self-reported	T = 14.22	top 3.8%
rs767943	rs767943	6	23446463	1	_	intron_variant	Cognitive performance (MTAG), Educational attainment (MTAG), Educational attainment (years of education), Highest math class taken (MTAG)	A = 17.68	top 1.7%
rs767943	rs2022330	6	23446327	0,800427	_	intron_variant	Educational attainment (years of education)	G = 10.47	top 9.0%
rs141547796	rs78648104	6	50715296	0,886266	TFAP2D	missense_variant	Attention deficit hyperactivity disorder MTAG, Body Mass Index, Cognitive performance (MTAG), Cystatin C levels, educational attainment (MTAG), Educational attainment (years of education), Systolic blood pressure	C = 24.2	top 0.4%
rs1925686	rs2031522	6	87111783	0,991515	_	intergenic_variant	Atrial fibrillation	G = 12.6	top 5.5%
rs72990858	rs72990858	6	104699909	1	_	intergenic_variant	Age at first sexual intercourse, educational attainment (MTAG), Smoking initiation (ever regular vs never regular) (MTAG), Sporadic neuroblastoma	A = 21.9	top 0.6%
rs12204714	rs12204714	6	151914204	1	CCDC170	intron_variant	Age at first sexual intercourse	T = 17.63	top 1.7%
rs12204714	rs6557171	6	151913458	0,889834	CCDC170	intron_variant	Educational attainment (years of education)	C = 12.45	top 5.7%
rs12701263	rs1045530	7	32868523	0,880174	AVL9/DPY19L1P2	3_prime_UTR_variant	Mean corpuscular volume	G = 14.78	top 3.3%
rs7783012	rs71149745	7	114416441	0,909981	_	intron_variant	Age at first birth	AATTTCATAATTTCAT = 16.66	top 2.2%
rs6966898	rs4732129	7	135540408	0,813515	_	intergenic_variant	Cognitive ability, years of educational attainment or schizophrenia pleiotropy	C = 13.78	top 4.2%

rs113367286	rs2272095	7	140459051	0,896095	BRAF	missense_variant	Number of sexual partners	G = 23.4	top 0.5%
rs7828172	rs4739558	8	38479746	1	_	intergenic_variant	Adult body size, Age of smoking initiation (MTAG), Body size at age 10, Smoking initiation (ever regular vs never regular) (MTAG)	G = 14.41	top 3.6%
rs7828172	rs36061954	8	38472132	0,978997	_	regulatory_region_variant	Body Mass Index, Triglyceride levels, Waist-hip ratio	T = 18.83	top 1.3%
rs7828172	rs3213849	8	38468528	0,879855	_	5_prime_UTR_variant	Craniosynostosis syndrome, HYPOGONADOTROPIC HYPOGONADISM 2 WITH OR WITHOUT ANOSMIA, OSTEOGLOPHONIC DYSPLASIA, TRIGONOCEPHALY 1, Waist-hip ratio	A = 16.36	top 2.3%
rs7828172	rs62505473	8	38466247	0,852217	_	intron_variant	Electrocardiogram morphology amplitude at temporal datapoints	G = 12.24	top 6.0%
rs11382985	rs6471482	8	86667075	0,983457	_	stop_gained	Achromatopsia 3, ClinVar: phenotype not specified, Heart failure, STARGARDT DISEASE 1	C = 18.82	top 1.3%
rs72674824	rs957448	8	94529074	0,809655	LINC00535	synonymous_variant	No syndromic cleft lip with cleft palate	G = 11.65	top 6.8%
rs12554512	rs7029201	9	23358083	0,979807	_	intron_variant	Educational attainment (years of education)	A = 13.4	top 4.6%
rs9886840	rs10818604	9	121846330	0,971053	_	intron_variant	Educational attainment (years of education)	A = 11.16	top 7.7%
rs3896224	rs3896224	10	104708095	1	CCNM2	intron_variant	Age at first sexual intercourse, Childhood maltreatment, Cognitive performance (MTAG), Intelligence, Lifetime smoking index, Smoking status, Trauma exposure	G = 18.61	top 1.4%
rs1866710	rs2099744	11	12844599	0,985212	TEAD1	intron_variant	Cognitive performance (MTAG), General cognitive ability	A = 14.42	top 3.6%
rs59957074	rs4759073	12	54259474	0,995815	_	5_prime_UTR_variant	Adult body size	A = 18.19	top 1.5%
rs7955865	rs772921	12	56009793	0,912063	_	intron_variant	Heel bone mineral density, Smoking status, Smoking status (ever vs never smokers)	T = 15.14	top 3.1%
rs2279574	rs2279574	12	89351700	1	_	missense_variant	Age at first sexual intercourse, Cognitive performance (MTAG), Highest math class taken, Highest math class taken (MTAG), Hypogonadotropic hypogonadism 19 with or without anosmia, Self-reported math ability (MTAG)	A = 22.9	top 0.5%
rs7987501	rs2165985	13	53402639	0,920861	_	intron_variant	Body Height	C = 15.22	top 3.0%
rs10646652	rs10612751	13	106993685	0,925872	_	intergenic_variant	Lobe attachment (rater-scored or self-reported)	TG = 12.18	top 6.1%
rs28929474	rs28929474	14	94378610	1	FAM181A-AS1	missense_variant	Age at first sexual intercourse, Alanine aminotransferase levels, Alanine transaminase levels, Alanine transaminase levels in high alcohol intake, Alcohol consumption (drinks per week) (MTAG), Alcohol consumption drinks per week, ALPHA-1-ANTITRYPSIN DEFICIENCY, Antineutrophil cytoplasmic antibody-associated vasculitis, Appendicular lean mass, Aspartate aminotransferase levels, Aspartate aminotransferase to alanine aminotransferase ratio, Aspartate transaminase levels in high alcohol intake, Bioavailable testosterone levels, Bitter alcoholic beverage consumption, Blood	T = 23.5	top 0.4%

							protein levels, Brain morphology MOSTest, Breast size, C-reactive protein levels, Calcium levels, Chronic obstructive pulmonary disease, Cirrhosis (alcohol related), Cirrhosis multi-trait analysis, ClinVar: phenotype not specified, Cystatin C levels, Direct bilirubin levels, Fat-free mass, FRAXE, Gallstone disease, Gamma glutamyl transferase levels, Glucagon levels in response to oral glucose tolerance test (fasting), Heel bone mineral density, Height, Hip circumference adjusted for BMI, Inborn genetic diseases, Insulin-like growth factor 1 levels, Liver enzyme levels (alanine transaminase), Liver enzyme levels (alkaline phosphatase), Liver enzyme levels (gamma-glutamyl transferase), Low density lipoprotein cholesterol levels, Metabolite levels (small molecules and protein measures), Osteoprotegerin levels, PI Z, PI Z(AUGSBURG), PI Z(TUN), Post bronchodilator FEV1, Post bronchodilator FEV1/FVC ratio, Post bronchodilator FEV1/FVC ratio in smoking, Post bronchodilator percent predicted FEV1 in smoking, Problematic alcohol use MTAG, Serum albumin level, Serum alkaline phosphatase levels, Serum alpha-fetoprotein levels, Serum total protein level, Sex hormone-binding globulin levels, Sex hormone-binding globulin levels adjusted for BMI, Systolic blood pressure, Testosterone levels, TNF-related apoptosis-inducing ligand levels, Total testosterone levels, Urea levels		
rs783544	rs783544	15	82571543	1	SAXO2	5_prime_UTR_variant	Age at first sexual intercourse	C = 16.52	top 2.2%
rs4702	rs4702	15	90883330	1	_	3_prime_UTR_variant	Age at first sexual intercourse, anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome pleiotropy, Autism spectrum disorder or schizophrenia, Bipolar disorder MTAG, Childhood maltreatment, Cognitive ability, years of educational attainment or schizophrenia pleiotropy, Feeling hurt, General risk tolerance (MTAG), INSOMNIA, Neuropsychiatric disorders, Number of sexual partners, Schizophrenia, Schizophrenia MTAG	A = 12.84	top 5.2%

rs763053	rs1139897	16	670986	0,802361	RAB40C	missense_variant	Age of smoking initiation (MTAG), Smoking initiation (ever regular vs never regular), Smoking initiation (ever regular vs never regular) (MTAG)	A = 16.32	top 2.3%
rs9923553	rs7189389	16	5758519	0,863752	_	intron_variant	Smoking initiation (ever regular vs never regular) (MTAG)	G = 16.7	top 2.1%
rs76513770	rs76513770	16	72471635	1	AC004158.3	intron_variant	Age at first sexual intercourse, Body Mass Index, Risk-taking tendency (4-domain principal component model)	C = 17.96	top 1.6%
rs28406364	rs16948048	17	49363104	0,928621	UTP18	intron_variant	Atopic asthma, Atopic dermatitis, Brain morphology MOSTest, Cancer, Coronary Artery Disease, Diastolic blood pressure, Height, Systolic blood pressure	G = 10.66	top 8.6%
rs28406364	rs112502960	17	49361940	0,908263	UTP18	5_prime_UTR_variant	Asthma, Asthma (moderate or severe), Cortical surface area MOSTest	A = 14.7	top 3.4%
rs28406364	rs9889262	17	49320708	0,904211	MBTD1	intron_variant	Allergic disease (asthma, hay fever or eczema), Eosinophil counts, Eosinophil percentage of granulocytes, Respiratory diseases	A = 16.09	top 2.5%
rs28406364	rs35587648	17	49340816	0,899785	UTP18	intron_variant	Height, Thyroid stimulating hormone levels	A = 11.47	top 7.1%
rs4800204	rs4800204	18	25067306	1	RP11-739N10.1	intron_variant	Age at first sexual intercourse	T = 15.85	top 2.6%
rs4800204	rs8089996	18	25068541	0,983761	RP11-739N10.1	intron_variant	Educational attainment (years of education)	A = 17.9	top 1.6%
rs148544378	rs148544378	18	42743602	1	_	missense_variant	Age at first sexual intercourse, Leisure sedentary behaviour television watching, Number of sexual partners	T = 23.3	top 0.5%
rs146852038	rs146852038	X	129984833	1	ENOX2	intron_variant	Age at first sexual intercourse	A = 18.66	top 1.4%
rs11038866	rs1317826	11	46366318	0,949319	DGKZ	missense_variant	ClinVar: phenotype not specified	G = 11.06	top 7.8%
rs11392435	rs1163627	13	111573354	0,858012	_	regulatory_region_variant	Waist-hip ratio	A = 13.78	top 4.2%

Supplementary table 3: In-Silico Phase Intronic gSNP list. *Following table consists of list of rs-ids of GWAS SNPs, as well as rs-ids of the SNPs in a LD with these gSNPs, chromosomes and positions within these chromosomes, CADD score, showing also an allele prediction, deleteriousness level and associations with linked SNP. All of these SNPs are all intronic. rs-ids with no found candidate gene names are highlighted in black in the "Gene" column. Deleteriousness up to top 10%.*

gSNP	Linked_SNP	ch	Pos_37	LD	Gene	Type	cadd	Deleteriousness
rs112523595	rs12638798	3	85502670	0,90848	CADM2	intron_variant	C = 21	top 0.8%
rs1991651	rs11250078	8	108090071	0,88665	PINX1	intron_variant	A = 21	top 0.8%
rs57945129	rs62264768	3	1,18E+08	0,944684	RP11-384F7.2	intron_variant	C = 20.3	top 0.9%
rs809955	rs769657	4	1,4E+08	0,855786	MAML3	intron_variant	G = 20.3	top 0.9%
rs35508442	rs4952715	2	44579481	0,85086 2	[redacted]	intron_variant	T = 19.92	top 1.0%
rs35508442	rs934777	2	44618119	0,837765	[redacted]	intron_variant	C = 19.83	top 1.0%
rs7024334	rs12555670	9	1,06E+08	0,820462	[redacted]	intron_variant	C = 20.2	top 1.0%
rs6764919	rs13097786	3	60916643	0,906631	[redacted]	intron_variant	C = 19.46	top 1.1%
rs112523595	rs34495106	3	85601986	0,995741	[redacted]	intron_variant	G = 19.62	top 1.1%
rs112523595	rs1463205	3	85547282	0,987238	[redacted]	intron_variant	A = 19.68	top 1.1%

rs112523595	rs35827242	3	85433511	0,946053	[redacted]	intron_variant	A = 19.72	top 1.1%
rs10134692	rs996661	14	93471650	0,979838		intron_variant	G = 19.6	top 1.1%
rs9923553	rs57105172	16	5740082	0,864938		intron_variant	T = 19.67	top 1.1%
rs2145108	rs7272651	20	30295985	0,927033		intron_variant	T = 19.62	top 1.1%
rs112523595	rs2029130	3	85594897	0,991478		intron_variant	A = 19.22	top 1.2%
rs112523595	rs17456820	3	85533789	0,987238		intron_variant	A = 19.16	top 1.2%
rs112523595	rs6549039	3	85556853	0,987238		intron_variant	G = 19.28	top 1.2%
rs12653396	rs12653396	5	88551455	1		intron_variant	A = 19.09	top 1.2%
rs245753	rs10475963	5	1,71E+08	0,990846		intron_variant	A = 19.22	top 1.2%
rs372420182	rs4734227	8	92406024	1		intron_variant	C = 19.31	top 1.2%
rs139881447	rs76548359	14	30273670	1		intron_variant	A = 19.39	top 1.2%
rs1435757	rs8027136	15	47547588	0,886316		intron_variant	T = 19.32	top 1.2%
rs112523595	rs4502590	3	85547329	0,987238		intron_variant	T = 18.96	top 1.3%
rs245753	rs2121124	5	1,71E+08	1		intron_variant	T = 18.87	top 1.3%
rs245753	rs1366206	5	1,71E+08	1		intron_variant	A = 18.89	top 1.3%
rs12894029	rs8008023	14	27139911	0,990277		intron_variant	G = 18.94	top 1.3%
rs3896224	rs3896224	10	1,05E+08	1		intron_variant	G = 18.61	top 1.4%
rs146852038	rs146852038	X	1,3E+08	1	BCORL1	intron_variant	A = 18.66	top 1.4%
rs7525548	rs1514175	1	74525960	0,892618	[redacted]	intron_variant	G = 18.15	top 1.5%
rs35508442	rs4953110	2	44579477	0,850862	CAMKMT	intron_variant	A = 18.16	top 1.5%
rs2084572	rs6442680	3	17377041	0,952463	TBC1D5	intron_variant	C = 18.32	top 1.5%
rs705240	rs697377	3	1,19E+08	0,960875	ENSG00000243276	intron_variant	A = 18.12	top 1.5%
rs34804222	rs11037653	11	43803816	0,979244	HSD17B12	intron_variant	A = 18.33	top 1.5%
rs76513770	rs76513770	16	72471635	1	AC004158.3/LINC01572	intron_variant	C = 17.96	top 1.6%
rs4800204	rs8089996	18	25068541	0,983761	ZNF521	intron_variant	A = 17.9	top 1.6%
rs7533341	rs852764	1	57819741	0,945125	DAB1	intron_variant	C = 17.59	top 1.7%
rs6719762	rs2419405	2	59924047	0,918164	RP11-444A22.1	intron_variant	C = 17.65	top 1.7%
rs2084572	rs9813532	3	17424212	0,86669	TBC1D5	intron_variant	A = 17.82	top 1.7%
rs112523595	rs2875908	3	85548543	0,987238	CADM2	intron_variant	C = 17.63	top 1.7%
rs767943	rs767943	6	23446463	1	[redacted]	intron_variant	A = 17.68	top 1.7%

rs12204714	rs12204714	6	1,52E+08	1	ESR1	intron_variant	T = 17.63	top 1.7%
rs34155040	rs10502880	18	47244375	0,951931	SKOR2	intron_variant	C = 17.81	top 1.7%
rs70959844	rs6747099	2	60550363	0,802294	BCL11A	intron_variant	C = 17.24	top 1.9%
rs2084572	rs2733502	3	17232930	0,98396	TBC1D5	intron_variant	T = 17.16	top 1.9%
rs112523595	rs17459563	3	85560215	0,987238	CADM2	intron_variant	A = 17.27	top 1.9%
rs112523595	rs76034006	3	85535085	0,987238	CADM2	intron_variant	AA = 17.17	top 1.9%
rs1435757	rs11633288	15	47584513	0,98813	SEMA6D	intron_variant	T = 17.23	top 1.9%
rs2274568	rs11102050	1	1,1E+08	0,883389	[redacted]	intron_variant	C = 17.04	top 2.0%
rs11240331	rs4950976	1	2,05E+08	0,935468	NFASC	intron_variant	G = 17.03	top 2.0%
rs112523595	rs3086190	3	85548550	0,983021	CADM2	intron_variant	AATAATAATAA = 16.99	top 2.0%
rs2910032	rs1438946	5	1,53E+08	0,821722	LINC01470	intron_variant	C = 16.93	top 2.0%
rs2084572	rs13318609	3	17353828	0,991977	TBC1D5	intron_variant	C = 16.72	top 2.1%
rs112523595	rs62250715	3	85466846	0,987238	CADM2	intron_variant	A = 16.69	top 2.1%
rs112523595	rs12492753	3	85552290	0,90446	CADM2	intron_variant	A = 16.84	top 2.1%
rs2279574	rs10506971	12	89364160	0,991977	[redacted]	intron_variant	A = 16.85	top 2.1%
rs9923553	rs7189389	16	5758519	0,863752	RBFOX1	intron_variant	G = 16.7	top 2.1%
rs35508442	rs4952716	2	44579592	0,850862	CAMKMT	intron_variant	C = 16.55	top 2.2%
rs6445264	rs6445264	3	62368750	1	PTPRG-AS1	intron_variant	A = 16.48	top 2.2%
rs245753	rs13180996	5	1,71E+08	0,977252	RANBP17	intron_variant	G = 16.5	top 2.2%
rs245753	rs7730898	5	1,71E+08	0,802441	RANBP17	intron_variant	A = 16.63	top 2.2%
rs7783012	rs71149745	7	1,14E+08	0,909981	FOXP2/AC073626.2	intron_variant	AATTTCATAATTTCAT = 16.66	top 2.2%
rs7783012	rs7785701	7	1,14E+08	0,909981	FOXP2	intron_variant	G = 16.5	top 2.2%
rs112523595	rs68001049	3	85591268	0,991478	CADM2	intron_variant	G = 16.32	top 2.3%
rs112523595	rs34467301	3	85538920	0,987238	CADM2	intron_variant	C = 16.37	top 2.3%
rs112523595	rs2033526	3	85447608	0,946053	CADM2	intron_variant	C = 16.32	top 2.3%
rs7671317	rs7655188	4	62106917	0,98651	ADGRL3	intron_variant	C = 16.3	top 2.3%
rs11688027	rs78876578	2	77732532	0,88534	LRRTM4	intron_variant	C = 16.11	top 2.4%
rs112523595	rs17515586	3	85520022	0,987238	CADM2	intron_variant	G = 16.28	top 2.4%
rs2406374	rs34644687	5	1,08E+08	1	FBXL17	intron_variant	A = 16.24	top 2.4%
rs875097	rs2385539	2	2,2E+08	0,80049	[redacted]	intron_variant	C = 16.06	top 2.5%
rs2084572	rs2060628	3	17233681	0,979988	TBC1D5	intron_variant	C = 16.09	top 2.5%

rs809955	rs769671	4	1,4E+08	0,860522	NOCT	intron_variant	T = 15.99	top 2.5%
rs28406364	rs9889262	17	49320708	0,904211	MBTD1	intron_variant	A = 16.09	top 2.5%
rs2145108	rs4097052	20	30306086	0,927033	BCL2L1	intron_variant	C = 15.99	top 2.5%
rs4800204	rs4800204	18	25067306	1	RP11-739N10.1	intron_variant	T = 15.85	top 2.6%
rs35508442	rs4953107	2	44569974	0,818008	PREPL	intron_variant	G = 15.75	top 2.7%
rs6764919	rs6804218	3	60899004	0,985756	FHIT	intron_variant	C = 15.69	top 2.7%
rs112523595	rs17023019	3	85565122	0,90848	CADM2	intron_variant	G = 15.65	top 2.7%
rs7608187	rs6728741	2	50348161	0,879499	NRXN1	intron_variant	C = 15.73	top 2.7%
rs35508442	rs11451478	2	44579232	0,820573	PREPL	intron_variant	TTTTTTTTTT = 15.54	top 2.8%
rs705240	rs705225	3	1,19E+08	0,934221	IGSF11	intron_variant	C = 15.52	top 2.8%
rs222440	rs2744452	6	53087523	0,893466	[redacted]	intron_variant	G = 15.59	top 2.8%
rs7608187	rs7574552	2	50387685	0,98790 4	NRXN1	intron_variant	C = 15.51	top 2.8%
rs112523595	rs35789162	3	85538913	0,987238	CADM2	intron_variant	AAAA = 15.35	top 2.9%
rs57945129	rs62264780	3	1,18E+08	0,938128	[redacted]	intron_variant	C = 15.32	top 2.9%
rs61856978	rs11597197	10	96153399	0,981423	[redacted]	intron_variant	G = 15.44	top 2.9%
rs34804222	rs6485465	11	43804045	0,983373	HSD17B12	intron_variant	A = 15.33	top 2.9%
rs113142203	rs11581644	1	1,54E+08	0,812	DENND4B	intron_variant	G = 15.26	top 3.0%
rs13009323	rs28780764	2	1,71E+08	0,939985	UBR3	intron_variant	C = 15.23	top 3.0%
rs112523595	rs77852438	3	85560262	0,987238	CADM2	intron_variant	T = 15.17	top 3.0%
rs809955	rs769675	4	1,4E+08	0,855786	NOCT	intron_variant	C = 15.22	top 3.0%
rs7783012	rs12533005	7	1,14E+08	0,909981	[redacted]	intron_variant	C = 15.16	top 3.0%
rs7987501	rs2165985	13	53402639	0,920861	[redacted]	intron_variant	C = 15.22	top 3.0%
rs7024334	rs13295012	9	1,06E+08	0,820462	[redacted]	intron_variant	A = 15.17	top 3.0%
rs112523595	rs62793760	3	85591419	0,912878	CADM2	intron_variant	ACACA = 15.05	top 3.1%
rs56392241	rs3851353	3	1,32E+08	0,894299	DNAJC13	intron_variant	G = 15.15	top 3.1%
rs7671317	rs7655208	4	62106934	0,98651	ADGRL3	intron_variant	C = 15.15	top 3.1%
rs34804222	rs1518816	11	43803596	0,983373	HSD17B12	intron_variant	C = 15.07	top 3.1%
rs7955865	rs772921	12	56009793	0,912063	[redacted]	intron_variant	T = 15.14	top 3.1%
rs76715069	rs8019512	14	98109162	0,933159	LINC02291	intron_variant	A = 15.12	top 3.1%
rs62180269	rs17432775	2	63239328	0,893382	EHBP1	intron_variant	T = 15	top 3.2%

rs1226414	rs12997268	2	1,56E+08	0,896144		intron_variant	G = 14.9	top 3.2%
rs112523595	rs72615727	3	85560248	0,983021	CADM2	intron_variant	C = 14.94	top 3.2%
rs57945129	rs6778926	3	1,18E+08	0,81597		intron_variant	G = 14.83	top 3.3%
rs245753	rs10076357	5	1,71E+08	0,995432	FGF18	intron_variant	C = 14.85	top 3.3%
rs2612029	rs7373253	3	53755567	0,8899	CACNA1D	intron_variant	C = 14.64	top 3.4%
rs245753	rs3849709	5	1,71E+08	0,995424		intron_variant	T = 14.71	top 3.4%
rs35508442	rs12712920	2	44654081	0,950713	CAMKMT	intron_variant	G = 14.56	top 3.5%
rs809955	rs797090	4	1,4E+08	0,986807	NOCT	intron_variant	G = 14.4	top 3.6%
rs1866710	rs2099744	11	12844599	0,985212	TEAD1	intron_variant	A = 14.42	top 3.6%
rs35508442	rs5830797	2	44612193	0,83345	CAMKMT	intron_variant	TT = 14.3	top 3.7%
rs1925686	rs6938885	6	87297461	0,991498		intron_variant	G = 14.32	top 3.7%
rs7987501	rs4883678	13	53368497	0,874184		intron_variant	A = 14.37	top 3.7%
rs12446652	rs8192516	16	75566587	0,980451	RP11-77K12.7/CHST5	intron_variant	C = 14.3	top 3.7%
rs35508442	rs13400118	2	44644735	0,828904	CAMKMT	intron_variant	C = 14.2	top 3.8%
rs13009323	rs28892917	2	1,71E+08	0,939985	UBR3	intron_variant	C = 14.23	top 3.8%
rs2278480	rs2278480	3	25594252	1	RARB	intron_variant	C = 14.2	top 3.8%
rs112523595	rs724304	3	85558989	0,987238	CADM2	intron_variant	C = 14.23	top 3.8%
rs12203592	rs12203592	6	396321	1	IRF4	intron_variant	T = 14.22	top 3.8%
rs1925686	rs9444491	6	87292667	0,995751		intron_variant	G = 14.22	top 3.8%
rs10955084	rs2319923	8	96759480	0,818647	C8orf37-AS1	intron_variant	G = 14.2	top 3.8%
rs9886840	rs10760192	9	1,22E+08	0,979488		intron_variant	T = 14.18	top 3.8%
rs9923553	rs12051184	16	5741333	0,869843		intron_variant	G = 14.16	top 3.8%
rs794375	rs236607	7	75441843	0,913249	CCL24	intron_variant	C = 14.05	top 3.9%
rs7987501	rs1017539	13	53467531	0,879497		intron_variant	A = 14.12	top 3.9%
rs7024334	rs16925382	9	1,06E+08	0,820462		intron_variant	G = 14.08	top 3.9%
rs112523595	rs62250750	3	85501819	0,987238	CADM2	intron_variant	C = 13.94	top 4.0%
rs57945129	rs62264764	3	1,18E+08	1		intron_variant	A = 14.02	top 4.0%
rs7824756	rs4873442	8	50267317	0,862984		intron_variant	A = 14.02	top 4.0%
rs61856978	rs10882743	10	96201992	0,849417	TBC1D12	intron_variant	A = 13.98	top 4.0%
rs10496949	rs2381473	2	1,43E+08	0,99189		intron_variant	A = 13.83	top 4.1%
rs2084572	rs1449881	3	17245325	0,988002	TBC1D5	intron_variant	G = 13.86	top 4.1%

rs112523595	rs7609594	3	85433445	0,946053	CADM2	intron_variant	A = 13.91	top 4.1%
rs56392241	rs1010899	3	1,32E+08	0,894299	DNAJC13	intron_variant	A = 13.91	top 4.1%
rs7987501	rs2197304	13	53373330	0,866385		intron_variant	A = 13.82	top 4.1%
rs7533341	rs852786	1	57835819	0,945125	DAB1	intron_variant	C = 13.76	top 4.2%
rs11772444	rs147977828	7	1,34E+08	0,948595	EXOC4	intron_variant	TTAGTTT = 13.73	top 4.2%
rs34155040	rs17785603	18	47243163	0,951931		intron_variant	A = 13.75	top 4.2%
rs114456303	rs114456303	3	24631667	1		intron_variant	A = 13.71	top 4.3%
rs702	rs138129526	4	28577330	0,835074		intron_variant	AAAAA = 13.7	top 4.3%
rs702	rs2458628	4	28577398	0,835074		intron_variant	T = 13.67	top 4.3%
rs1435757	rs5812402	15	47566914	0,882501	SEMA6D	intron_variant	- = 13.67	top 4.3%
rs2084572	rs1867772	3	17349847	0,911401	TBC1D5	intron_variant	A = 13.54	top 4.4%
rs2084572	rs2596649	3	17265903	0,908298	TBC1D5	intron_variant	A = 13.52	top 4.4%
rs9886840	rs7046409	9	1,22E+08	0,840472		intron_variant	T = 13.61	top 4.4%
rs222440	rs2744451	6	53087519	0,868374		intron_variant	T = 13.43	top 4.5%
rs34804222	rs10838157	11	43724014	0,933896	HSD17B12	intron_variant	G = 13.44	top 4.5%
rs7987501	rs9536408	13	53390211	0,885732		intron_variant	T = 13.46	top 4.5%
rs13009323	rs1362487	2	1,71E+08	0,820177	UBR3	intron_variant	C = 13.41	top 4.6%
rs12554512	rs7029201	9	23358083	0,979807		intron_variant	A = 13.4	top 4.6%
rs113142203	rs12043350	1	1,54E+08	0,816429	GATAD2B	intron_variant	T = 13.25	top 4.7%
rs375909440	rs9984518	21	39211240	0,915085	KCNJ6	intron_variant	C = 13.29	top 4.7%
rs35508442	rs734016	2	44641997	0,814326	CAMKMT	intron_variant	C = 13.15	top 4.8%
rs12714702	rs4374552	3	88135733	0,949671	CGGBP1	intron_variant	A = 13.17	top 4.8%
rs245753	rs10041523	5	1,71E+08	1		intron_variant	C = 13.2	top 4.8%
rs7785195	rs12700808	7	3344512	0,867313	SDK1	intron_variant	G = 13.21	top 4.8%
rs1925686	rs9450630	6	87189442	0,991498		intron_variant	A = 13.14	top 4.9%
rs10134692	rs1910517	14	93394857	0,910667	CHGA	intron_variant	G = 13.14	top 4.9%
rs35508442	rs4953111	2	44580267	0,832535	PREPL	intron_variant	A = 13	top 5.0%
rs590414	rs539238	11	1,06E+08	0,825703	KBTBD3	intron_variant	G = 12.89	top 5.1%
rs34481141	rs13396624	2	1,85E+08	0,993059		intron_variant	G = 12.92	top 5.1%
rs35508442	rs62132285	2	44593102	0,836178	CAMKMT	intron_variant	G = 12.81	top 5.2%

rs1516172	rs1878135	2	51650587	0,826504		intron_variant	G = 12.86	top 5.2%
rs2084572	rs5846958	3	17350731	0,991977	TBC1D5	intron_variant	G = 12.88	top 5.2%
rs2084572	rs2733509	3	17222863	0,988002	TBC1D5	intron_variant	G = 12.86	top 5.2%
rs2406374	rs12656108	5	1,08E+08	0,855807	FBXL17	intron_variant	G = 12.8	top 5.2%
rs6504551	rs12947658	17	67906890	0,816001		intron_variant	G = 12.87	top 5.2%
rs2084572	rs2470577	3	17271819	1	TBC1D5	intron_variant	G = 12.79	top 5.3%
rs11955430	rs11955430	5	1,68E+08	1	PANK3	intron_variant	G = 12.73	top 5.3%
rs875097	rs2010528	2	2,19E+08	0,812511	PLCD4	intron_variant	G = 12.67	top 5.4%
rs11955430	rs1345735	5	1,68E+08	0,873867	PANK3	intron_variant	C = 12.69	top 5.4%
rs7783012	rs10249234	7	1,14E+08	1		intron_variant	A = 12.65	top 5.4%
rs72674824	rs12678305	8	94508911	0,813737	LINC00535	intron_variant	C = 12.71	top 5.4%
rs1866710	rs11022519	11	12866689	0,97544	TEAD1	intron_variant	C = 12.68	top 5.4%
rs113142203	rs11590099	1	1,54E+08	0,816429	GATAD2B	intron_variant	T = 12.54	top 5.6%
rs245753	rs7735245	5	1,71E+08	0,990846		intron_variant	C = 12.53	top 5.6%
rs12204714	rs6557171	6	1,52E+08	0,889834	CCDC170	intron_variant	C = 12.45	top 5.7%
rs1435757	rs28505872	15	47551862	0,886425	SEMA6D	intron_variant	C = 12.42	top 5.7%
rs7525548	rs3845345	1	74536983	0,861009	LRRIQ3	intron_variant	T = 12.33	top 5.8%
rs1226414	rs1226422	2	1,56E+08	0,900229		intron_variant	G = 12.31	top 5.9%
rs34155040	rs4534948	18	47267048	0,915459		intron_variant	A = 12.32	top 5.9%
rs2278480	rs59134881	3	25661909	0,842012	TOP2B	intron_variant	C = 12.21	top 6.0%
rs7828172	rs62505473	8	38466247	0,852217		intron_variant	G = 12.24	top 6.0%
rs34481141	rs13422256	2	1,85E+08	1		intron_variant	T = 12.24	top 6.0%
rs7533341	rs852759	1	57818564	0,949276	DAB1	intron_variant	G = 12.14	top 6.1%
rs2084572	rs9824952	3	17406945	0,86669	TBC1D5	intron_variant	A = 12.13	top 6.1%
rs222440	rs222449	6	53051264	0,845947		intron_variant	T = 12.15	top 6.1%
rs7525548	rs3895907	1	74540343	0,988036	LRRIQ3	intron_variant	G = 12.09	top 6.2%
rs7525548	rs6703637	1	74532111	0,845463	LRRIQ3	intron_variant	A = 12.06	top 6.2%
rs245753	rs4868049	5	1,71E+08	0,995424		intron_variant	C = 12.08	top 6.2%
rs1925686	rs6940325	6	87197198	0,995751		intron_variant	G = 12.06	top 6.2%
rs12714702	rs6551273	3	88127039	0,985535	CGGBP1	intron_variant	T = 11.93	top 6.4%

rs72674824	rs12679345	8	94552305	0,813737	LINC00535	intron_variant	G = 11.95	top 6.4%
rs9886840	rs7027567	9	1,22E+08	0,839671		intron_variant	C = 11.96	top 6.4%
rs34481141	rs10206254	2	1,85E+08	1		intron_variant	A = 11.94	top 6.4%
rs112523595	rs4856273	3	85549110	0,982976	CADM2	intron_variant	A = 11.9	top 6.5%
rs11382985	rs6471476	8	86664130	0,991712		intron_variant	C = 11.82	top 6.6%
rs112523595	rs11127899	3	85545412	0,987238	CADM2	intron_variant	G = 11.77	top 6.7%
rs12714702	rs7650707	3	88109416	0,949671	CGGBP1	intron_variant	C = 11.72	top 6.7%
rs1925686	rs1188817	6	87346001	0,833454		intron_variant	G = 11.73	top 6.7%
rs10134692	rs8003519	14	93360326	0,89216		intron_variant	G = 11.75	top 6.7%
rs7533341	rs706409	1	57816494	0,949276	DAB1	intron_variant	G = 11.68	top 6.8%
rs35508442	rs17032420	2	44598574	0,835416	CAMKMT	intron_variant	A = 11.61	top 6.9%
rs186723454	rs73083946	3	54130128	0,951907		intron_variant	T = 11.63	top 6.9%
rs7671317	rs13151071	4	62115046	0,916148	ADGRL3	intron_variant	G = 11.57	top 7.0%
rs7671317	rs10003184	4	62086125	0,886535	ADGRL3	intron_variant	A = 11.54	top 7.0%
rs245753	rs10042357	5	1,71E+08	0,995424		intron_variant	A = 11.54	top 7.0%
rs245753	rs4362957	5	1,71E+08	0,995424		intron_variant	G = 11.48	top 7.1%
rs10134692	rs1503958	14	93392264	0,910667	CHGA	intron_variant	T = 11.47	top 7.1%
rs28406364	rs35587648	17	49340816	0,899785	UTP18	intron_variant	A = 11.47	top 7.1%
rs3007104	rs2933223	14	46889271	0,943024	LINC00871	intron_variant	C = 11.44	top 7.2%
rs375909440	rs9979936	21	39280463	1	KCNJ6	intron_variant	G = 11.42	top 7.2%
rs6764919	rs7356063	3	60944685	0,82632	FHIT	intron_variant	A = 11.39	top 7.3%
rs245753	rs10475962	5	1,71E+08	0,995432	FGF1	intron_variant	G = 11.37	top 7.3%
rs11038866	rs3802890	11	46512996	0,935713	AMBRA1	intron_variant	G = 11.31	top 7.4%
rs6748341	rs60714794	2	2,25E+08	0,951713		intron_variant	AAAAAAAAAA = 11.23	top 7.5%
rs11428242	rs61909696	11	1,34E+08	0,813377	GLB1L2	intron_variant	C = 11.23	top 7.5%
rs113142203	rs71697078	1	1,54E+08	0,816429	DENND4B	intron_variant	AATTAATTA = 11.18	top 7.6%
rs35508442	rs10204480	2	44637302	0,828106	CAMKMT	intron_variant	G = 11.21	top 7.6%
rs13009323	rs7594247	2	1,71E+08	0,939985	UBR3	intron_variant	T = 11.2	top 7.6%
rs2084572	rs2348005	3	17254264	0,904105	TBC1D5	intron_variant	C = 11.22	top 7.6%
rs112523595	rs1865252	3	85439590	0,946053	CADM2	intron_variant	G = 11.22	top 7.6%

rs9923553	rs13335882	16	5775964	0,985247		intron_variant	T = 11.17	top 7.6%
rs809955	rs769673	4	1,4E+08	0,860522	NOCT	intron_variant	A = 11.14	top 7.7%
rs9886840	rs10818604	9	1,22E+08	0,971053		intron_variant	A = 11.16	top 7.7%
rs9964201	rs12957045	18	53080790	0,844416	TCF4	intron_variant	G = 11.13	top 7.7%
rs35508442	rs79542623	2	44594567	0,831757	CAMKMT	intron_variant	C = 11.09	top 7.8%
rs1226414	rs13002285	2	1,56E+08	0,888022		intron_variant	T = 11.06	top 7.8%
rs2084572	rs13093375	3	17394106	0,948378	TBC1D5	intron_variant	G = 11.05	top 7.9%
rs67723420	rs13097782	3	35698083	0,922311	ARPP21	intron_variant	T = 11.05	top 7.9%
rs11688027	rs375312965	2	77795127	0,987299		intron_variant	C = 10.9	top 8.1%
rs2084572	rs1597393	3	17218265	0,984039	TBC1D5	intron_variant	C = 10.91	top 8.1%
rs7671317	rs10013024	4	62137318	0,894051	ADGRL3	intron_variant	A = 10.89	top 8.1%
rs34804222	rs10838173	11	43799169	0,983373	HSD17B12	intron_variant	G = 10.89	top 8.1%
rs7987501	rs1342669	13	53363570	0,878011		intron_variant	C = 10.86	top 8.2%
rs76702070	rs77664243	11	85581160	0,975731	CCDC83	intron_variant	T = 10.86	top 8.2%
rs7533341	rs1323828	1	57780268	0,911314	DAB1	intron_variant	A = 10.82	top 8.3%
rs35508442	rs12998046	2	44653786	0,942716	CAMKMT	intron_variant	A = 10.81	top 8.3%
rs705240	rs798581	3	1,19E+08	0,896873	IGSF11	intron_variant	C = 10.83	top 8.3%
rs7824756	rs4873443	8	50267341	0,85348		intron_variant	T = 10.8	top 8.3%
rs10955084	rs4735438	8	96817689	0,964735	C8orf37-AS1	intron_variant	T = 10.82	top 8.3%
rs11955430	rs12522181	5	1,68E+08	0,950634	PANK3	intron_variant	G = 10.78	top 8.4%
rs245753	rs2161216	5	1,71E+08	0,972639		intron_variant	A = 10.74	top 8.4%
rs12554512	rs4977836	9	23355666	0,979807		intron_variant	A = 10.74	top 8.4%
rs34155040	rs57971954	18	47245998	0,951931		intron_variant	GGAG = 10.78	top 8.4%
rs35508442	rs6726493	2	44568225	0,821131	PREPL	intron_variant	C = 10.72	top 8.5%
rs6764919	rs6763967	3	60898944	0,980938	FHIT	intron_variant	A = 10.73	top 8.5%
rs112523595	rs59417256	3	85571331	0,987238	CADM2	intron_variant	G = 10.71	top 8.5%
rs57945129	rs1456196	3	1,18E+08	0,938128		intron_variant	A = 10.69	top 8.5%
rs56392241	rs1847832	3	1,32E+08	0,894299	DNAJC13	intron_variant	C = 10.73	top 8.5%
rs7024334	rs12685887	9	1,06E+08	0,80259		intron_variant	G = 10.71	top 8.5%
rs1226414	rs2695440	2	1,56E+08	0,896144		intron_variant	G = 10.64	top 8.6%

rs2084572	rs2596673	3	17229331	0,98396	TBC1D5	intron_variant	G = 10.64	top 8.6%
rs28406364	rs16948048	17	49363104	0,928621	UTP18	intron_variant	G = 10.66	top 8.6%
rs1516172	rs1606972	2	51647649	0,93895		intron_variant	T = 10.6	top 8.7%
rs2084572	rs283911	3	17290212	0,991977	TBC1D5	intron_variant	T = 10.59	top 8.7%
rs341521	rs341530	13	59805313	0,927724		intron_variant	T = 10.55	top 8.8%
rs112523595	rs59073108	3	85560139	0,98723	CADM2	intron_variant	G = 10.5	top 8.9%
rs12714702	rs959048	3	88062157	0,949671		intron_variant	C = 10.5	top 8.9%
rs7608187	rs11125301	2	50371712	0,98790	NRXN1	intron_variant	T = 10.53	top 8.9%
rs767943	rs2022330	6	23446327	0,800427		intron_variant	G = 10.47	top 9.0%
rs7785195	rs4416733	7	3332329	0,86274		intron_variant	G = 10.47	top 9.0%
rs7824756	rs34205061	8	50261562	0,931179		intron_variant	A = 10.44	top 9.0%
rs2084572	rs201959300	3	17344471	0,911401	TBC1D5	intron_variant	C = 10.42	top 9.1%
rs12714702	rs9847019	3	88078842	0,949671		intron_variant	G = 10.41	top 9.1%
rs4868800	rs883322	5	1,68E+08	0,826554	TENM2	intron_variant	T = 10.41	top 9.1%
rs12714702	rs4402954	3	88097433	0,942549		intron_variant	A = 10.36	top 9.2%
rs112523595	rs74384786	3	85535537	0,98723	CADM2	intron_variant	TTTTT = 10.27	top 9.4%
rs2084572	rs283915	3	17288276	0,98794	TBC1D5	intron_variant	C = 10.23	top 9.5%
rs7987501	rs7336148	13	53452920	0,924937		intron_variant	A = 10.21	top 9.5%
rs7525548	rs12041912	1	74538026	0,995998	LRRIQ3	intron_variant	A = 10.16	top 9.6%
rs11772444	rs2542264	7	1,34E+08	0,803605		intron_variant	G = 10.16	top 9.6%
rs12554512	rs7467480	9	23354942	0,967899		intron_variant	A = 10.19	top 9.6%
rs10992812	rs3750354	9	93653371	0,916718	SYK	intron_variant	T = 10.18	top 9.6%
rs10134692	rs57184074	14	93478812	0,826182	ITPK1	intron_variant	T = 10.18	top 9.6%
rs7608187	rs1402128	2	50389875	0,83856	NRXN1	intron_variant	C = 10.17	top 9.6%
rs112523595	rs62252513	3	85538387	0,98723	CADM2	intron_variant	T = 10.12	top 9.7%
rs138850767	rs67446571	2	22331776	0,942025		intron_variant	G = 10.07	top 9.8%
rs76715069	rs79504488	14	98120119	0,933159	LINC02291	intron_variant	C = 10.07	top 9.8%
rs2084572	rs2733500	3	17234657	0,98396	TBC1D5	intron_variant	A = 10.05	top 9.9%
rs6764919	rs9826649	3	60899621	0,976244	FHIT	intron_variant	C = 10.04	top 9.9%
rs112523595	rs62250500	3	85578541	0,991478	CADM2	intron_variant	T = 10.05	top 9.9%

Supplementary table 4: eQTL Phase list of top 5 most significant gSNPs.

Values for this table were chosen based on the highest Z-score. The False Discovery Rate (FDR) is zero in all five cases.

gSNP	Alleles	Gene	Z-score
rs11204771	G	CDC42SE1	47.8672
rs794375	C	PMS2P3	50.3902
rs763053	C	METTL26	35.1866
rs807478	A	COX6B1	41.4232
rs147633738	T	CCT3	13.009

Supplementary table 5: In-Silico Phase list of top 5 most significant gSNPs.

Table shows rs-ids, candidate gene names, values were chosen based on the highest CADD score, shown for each variant allele.

gSNP	Linked_SNP	Chr	Pos_37	LD	gene	cadd	Deleteriousness
rs34811474	rs34811474	4	25407216	1	ANAPC4	A = 24.3	top 0.4%
rs141547796	rs78648104	6	50715296	0,886266	TFAP2D	C = 24.2	top 0.4%
rs113367286	rs2272095	7	140459051	0,896095	BRAF	G = 23.4	top 0.5%
rs28929474	rs28929474	14	94378610	1	FAM181A-AS1	T = 23.5	top 0.4%
rs148544378	rs148544378	18	42743602	1	RIT2	T = 23.3	top 0.5%

Supplementary table 6: eQTL Phase table of most significant missense gSNPs from eQTL analysis.

Table shows rs-ids of missense variants, their corresponding allele, predicted candidate gene name and the Z-score. The yellow color represents the failure in fining and establishing association between eQTLs and gSNP variants.

gSNP	Alleles	Gene	Z-score
rs11204771	G	CDC42SE1	47.8672
rs147633738	T	CCT3	13.009
rs875097			
rs12714702	A	C3orf38	-14.209
rs12714702	A	C3orf38	-14.209
rs34811474	A	ANAPC4	-35.412
rs141547796			
rs794375	C	PMS2P3	50.3902
rs113367286	T	SLC37A3	-7.3043
rs2279574	C	RP11-981P6.1	7.9902
rs28929474	T	SERPINA1	10.0133
rs763053	C	C16orf13	35.1866
rs148544378			
rs807478	A	COX6B1	41.4232
rs11038866	G	MADD	9.3395