Exploration of weighting schemes based on allele frequency and annotation for weighted burden association analysis of complex phenotypes

David Curtis

UCL Genetics Institute, University College London, London UK

Centre for Psychiatry, Queen Mary University of London, London UK

Correspondence:

David Curtis

d.curtis@ucl.ac.uk

00 44 7973 906 143

UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT.

ORCID 0000-0002-4089-9183

Abstract

Weighted burden analysis can incorporate variants with different frequencies and annotations into a combined test for association between a gene and a phenotype. However there has not been a systematic exploration of which weighting schemes provide maximum power to detect association. Here we assess different weighting schemes using a number of genes for which exome-wide evidence of association with common phenotypes was obtained in 200,000 exome-sequenced UK Biobank participants. We find that there are marked differences in optimal weighting schemes between genes, both with respect to allele frequency and to annotation, implying that there is no "one-size-fits-all" scheme which is generally optimal. It seems helpful to weight rare variants more highly than common ones, to give loss of function variants higher weights than protein-altering variants and to assign higher weights to protein-altering variants predicted to have more severe effects. However with the data currently available it does not seem possible to make more specific recommendations. This research has been conducted using the UK Biobank Resource.

Keywords

Exome; rare variant; association; loss of function; nonsynonymous.

Introduction

Weighted burden analysis can be used in large samples to detect gene-level association between rare DNA variants and a phenotype (Curtis, 2016). It is applied to variants which are too rare to individually generate a statistically significant effect and hence need to be analysed jointly. It is expected that many variants, such as synonymous and intronic variants, will often have little or no effect and so broadly speaking it makes sense to either restrict attention to a particular category of variant, such as loss of function (LOF) variants, or to include other types of variant but provide higher weights to variants thought more likely *a priori* to exert an effect. We have applied weighting

schemes which incorporate both annotation and allele frequency to test for association with a variety of different phenotypes (Curtis, 2020; Curtis et al., 2019, 2018). However an obvious problem is how to choose a weighting scheme which will have maximal power to detect association, this being the scheme which most accurately mirrors the different biological effects of different types of variant. The availability of large exome-sequenced samples which have been characterised for multiple phenotypes allows the empirical investigation of the contribution which different categories of variant make to the observed gene-phenotype associations and hence could be used to inform choice of weighting scheme (Szustakowski et al., 2020).

It is worth addressing some of the issues involved in greater depth. Firstly, it can be stated explicitly that weighted burden analysis models the situation where all variants have the same direction of effect, which broadly speaking is to impair the function of the gene. LOF variants (comprising stop-gained, frame shift and essential splice site variants) produce haploinsufficiency and may be expected to result in less protein product while other rare variants having an effect are expected to be more likely to yield a product which in some way or another works less effectively. Of course, this does not reflect the real situation since some nonsynonymous variants may in fact produce a gain of function but if variants are too rare to be considered individually then it is necessary to group them together and assume a shared effect. It is then reasonable to assume that a variant occurring at random within a gene is more likely to impair its function than to enhance it. The method also assumes that all variants within a category share a similar magnitude of effect. Again this is clearly not correct for the category of nonsynonymous variants, where effects may vary dramatically between variants, and even for LOF variants there may be important differences, for example depending on which transcripts are impacted. The fact that it is assumed that the variant burden acts to reduce gene functioning does not imply that a direction of effect on the phenotype is also assumed. The method is agnostic as to whether impaired function of a gene increases or decreases a quantitative phenotype or risk of disease.

The approach that we use incorporates both a weight related to the predicted functional impact of the variant and a weight related to its rarity. In the original conception, the approach was proposed as a method to analyse common and rare variants jointly, taking as its example phenotype Crohn's disease, in which both common and rare variants had been shown to contribute to risk (Curtis, 2012). A previous study had proposed weighting variants according to their frequency in controls but had a number of disadvantages, including that the weights depended on sample size and that the null hypothesis distribution was unknown, necessitating permutation testing (Madsen and Browning, 2009). To address these issues, a parabolic weighting function was implemented, with variants having minor allele frequency (MAF) = 0.5 given a weight of 1 while very rare variants were assigned a higher weight, typically 10. The weight assigned for functional annotation was then multiplied by the weight assigned for MAF to produce a combined weight for each variant. Subsequently, as it became clear that common variants generally do not have large effects on phenotype, attention was restricted to variants with MAF <= 0.01 and the weighting scheme was modified to give a weight of 1 to variants with MAF = 0.01, again increasing to 10 for very rare variants. However until now there has been little empirical exploration of the optimal values to use to weight by allele frequency.

Broadly speaking, it seems reasonable to begin from a position that most LOF variants will have a similar effect whereas most intronic and intergenic variants will have little effect but it is clear that the effects of different nonsynonymous variants can vary greatly. As a consequence, considerable effort has been devoted to developing methods which seek to predict the likely impact of nonsynonymous variants. These methods can utilise a variety of different approaches. Some may use information about the nature of the amino acid change within its local context. Others may take account of conservation of the DNA sequence and/or amino acid sequence between species. Since the different methods produce different results there are also different approaches to developing

predictions based on combining their outputs. The database dbNSFP contains annotations for all potential nonsynonymous and splice site single nucleotide variants using 37 different prediction algorithms (Liu et al., 2020). In previous analyses we have incorporated predictions from both SIFT and PolyPhen to contribute to weighting schemes (Adzhubei et al., 2013; Kumar et al., 2009).

Naturally, attempts have been made to compare the performance of these methods on real world data (Hassan et al., 2019; Liu et al., 2020). However one can argue that such comparisons are in some ways limited. They have tended to rely on the ability of a method to classify a known pathogenic variant as being pathogenic but the results obtained will depend on which known pathogenic variants are tested and these tend to be variants identified as causing severe disorders. Typically the identification of a pathogenic variant in a clinical setting will utilise at least some of the criteria which are implemented in the prediction tools, meaning that "known" pathogenic variants entered into databases such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) will tend to have characteristics similar to those used by prediction tools even if steps are taken to avoid explicit circularity. In clinical situations a substantial number of cases fail to yield a genetic diagnosis and this may be because the causative variant matches poorly to the usual criteria for identifying pathogenicity. Such a variant will not then become identified as a "known" pathogenic variant. Thus, to some extent we may be in the position that we are unaware of the failure of predictive tools to identify pathogenic variants because variants failing to match the expect profile are absent from testing and training datasets. A somewhat separate consideration is that there may be differences between the kinds of variant which cause a severe genetic condition in an individual, which the prediction tools are designed to detect, and the kinds of variant which have a more moderate impact on a common phenotype and which might be helpful to incorporate in weighted burden analyses. This issue has been addressed explicitly regarding the application of PolyPhen-2, where it is recommended that the version trained on HumVar be used to assist the diagnosis of Mendelian disorders while the version trained on HumDiv should be used to evaluate rare alleles for complex genotypes (Adzhubei et al., 2010).

The availability of large, exome-sequenced samples means that it has become possible to empirically investigate the distribution of variants within a gene known to affect a phenotype and to see which types of variant are most strongly associated with the phenotype. We have identified a number of genes demonstrating statistically significant association with common phenotypes within the UK Biobank (Szustakowski et al., 2020). Here, we study how changing the weightings assigned on the basis for allele frequency and variant annotation impacts the results of weighted burden analysis.

Methods

As previously described, the UK Biobank dataset was downloaded along with the variant call files for 200,632 subjects who had undergone exome-sequencing and genotyping by the UK Biobank Exome Sequencing Consortium using the GRCh38 assembly with coverage 20X at 95.6% of sites on average (Szustakowski et al., 2020). UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001). To obtain population principal components reflecting ancestry, version 2.0 of *plink* (https://www.cog-genomics.org/plink/2.0/) was run with the options *--maf 0.1 --pca 20 approx* (Chang et al., 2015; Galinsky et al., 2016).

Weighted burden analyses were carried out using the SCOREASSOC and GENEVARASSOC programs (Curtis, 2016). Attention was restricted to rare variants with minor allele frequency (MAF) <= 0.01 in both cases and controls. As previously described, variants were weighted by overall MAF so that

variants with MAF=0.01 were given a weight of 1 while very rare variants with MAF close to zero were given a weight equal to a weighting factor, WF, with a parabolic function used to assign weights with intermediate MAFs (Curtis, 2020). Additionally each variant has a weight assigned according to annotation and the overall weight for each variant consists of the frequency weight multiplied by the annotation weight. For each subject and each gene, the weights for the variants carried by the subject are summed to provide an overall weighted burden score. Regression modelling is done to calculate the likelihood for the phenotype data given covariates consisting of sex and the first 20 principal components and then the likelihood is recalculated for the model additionally incorporating the weighted burden score. Twice the natural log of the ratio of these likelihoods is a likelihood ratio statistic taken to be distributed as a chi-squared statistic with 1 degree of freedom. The evidence for association is summarised as the signed log p value (SLP) taken as the log base 10 of the p value and given a positive sign if there is a positive correlation between the weighted burden score and the phenotype. For the current analysis, the programs were modified to allow the specification of multiple annotation weights to be considered simultaneously, as detailed below.

Variant annotation was performed in two stages. First, a primary categorisation was made using Variant Effect Predictor (VEP), which uses information based on the reference sequence and coordinates of known transcripts to report findings such as whether variants occur within exons, if so whether they change amino acid sequence, etc (McLaren et al., 2016). For purposes of the present analyses, variants predicted to have a similar kind of effect were grouped together so that, for example, stop gained, frameshift and essential splice site variants were all treated as LOF. The full list of annotations as reported by VEP and the category they were assigned to is shown in Table 1, along with the weights which were used for the original weighted burden analyses, which had been arbitrarily assigned based on expectations of the likely biological importance of each annotation. Each of the annotation categories was used to generate a separate burden score, so that for example the burden score relating to the category LOF for a subject would consist of the number of LOF variants carried by that subject.

For the nonsynonymous and splice site variants listed in dbNSFP v4, secondary annotation scores were obtained consisting of the rank scores for the different prediction and conservation methods (Liu et al., 2020). Thus, a subject's secondary score for the SIFT annotation would consist of the sum of all the SIFT rank scores of the variants carried by that subject. For ease of processing, special characters in dbNSFP annotations were replaced, for example GERP++ was changed to GERPPP. A total of 43 such scores were used, as presented below.

The genes selected for this study consisted of those which had previously produced exome-wide significant results in weighted burden analyses using a number of common phenotypes and for which there was additional, independent evidence to support their involvement. These genes and phenotypes are listed in Table 2. The BMI phenotype was calculated directly from the fields for height and weight whereas the case definitions for hyperlipidaemia, hypertension and type 2 diabetes were derived from a mixture of self-report, recorded diagnoses and medication reports (Curtis, 2021a, 2021b, 2021c, 2020). In the case of *PCSK9* and *ANGPTL3* the original SLPs obtained were negative, indicating that variants impairing the function of these genes were protective and were associated with lower risk of developing hyperlipidaemia. For the purpose of the current study, in order to make it easier to interpret the results for these genes alongside the others, the phenotype of interest for these two genes is taken to be "not having hyperlipidaemia", meaning that all associated variants will tend to generate positive SLPs. Each of the genes selected contains hundreds of variants across different categories, providing empirical data to assess the ability of different variant annotations to inform tests for association.

In order to assess the impact of varying the allele frequency weighting factor, WF, a range of values were generated by having log10(WF) range from 0 to 2, generating values for WF from 1 to 100, and these were then used to derive a frequency weight for each variant. For each value of WF, these frequency weights were summed for each variant category to produce a category-specific burden score for each subject. The association of each of these scores with the phenotype was tested to produce an SLP based on the Wald statistic for each category and these SLPs were compared across the different values for WF.

The results from these analyses suggested that there was some benefit in weighting variants according to MAF and that a reasonable weighting scheme was obtained with log10(WF)=1.4, WF=25.12. Using this value, the SLPs produced by the different variant categories for each gene were tabulated and compared.

Similar analyses were performed for secondary annotations obtained from dbNSFP, except that for these analyses the weighted burden score produced by the ProteinAltering category was included as an additional covariate. This is because the overall burden for each dbNSFP annotation to some extent reflects the number of ProteinAltering variants each subject carries. The SLPs produced for each annotation were tabulated and compared. The results of these analyses were used to select a small group of different annotations which each produced a signal close to maximal in at least one of the genes and including an annotation based on conservation rather than predicted effect on the product. These annotations were then entered jointly into a regression models and the strength of evidence obtained for association was compared to that obtained for the analyses using the annotations individually.

Finally, the variant categories and secondary annotations were included jointly in a multivariate analysis along with the principal components and sex as covariates and for each gene a likelihood ratio test was performed to compare the likelihoods of the models with and without these predictor variables. The SLPs produced from these tests were compared with those previously reported in the original analyses.

Results

The only variant categories to generate SLPs of a large magnitude were LOF and ProteinAltering and the SLPs generated using different values of log10(WF) are shown in Table 3. The results for all categories are presented in Supplementary Table S1. What can be seen from Table 3 is that varying the value of WF has a negligible effect on the results obtained for the LOF variants but that large differences can be seen with some genes for the ProteinAltering variants, in particular for *MC4R*, *PCSK1* and *LDLR*. The results for these genes are plotted in Figure 1. This shows that the SLP for *MC4R* is strongly negative for small values of WF but increases steeply at first and then more gradually until it reaches a value of 5.92 at WF=100. For *PCSK1* the SLP is 6.52 at WF=1, rises to a maximum of 9.87 at WF=10 and then gradually falls back to 6.00 at WF=100. For *LDLR* the SLP rapidly rises from 21.03 at WF=1 to 30.43 at WF=4 and then stays almost unchanged. These results clearly demonstrate that if the goal is to choose a value for WF which will maximise the SLP then there is no single value which is optimal for every gene and phenotype combination. There does seem to be an advantage in setting WF to be greater than 1, that is to accord greater weight to rarer variants, and the default value of WF=10 performs reasonably well over these few examples. However *MC4R* does produce higher SLPs with somewhat higher values of WF and so

log10(WF)=1.4, WF=25.12, was chosen as a reasonable compromise and was used in the subsequent analyses to explore the contributions of different variant annotations.

Table 4 shows the SLPs produced by each variant category for each gene using WF=25.12. The same information is displayed graphically by the heatmap shown in Figure 2, in which the sizes of the dots for each gene are proportional to the SLP for each variant category relative to the maximum SLP obtained by any category for that gene. This clearly shows the way that different categories of variant make different contributions to the evidence for association for each gene. Thus, for *HNF4A* and to a lesser extent *PCSK1* and *LDLR* there is a high relative contribution from ProteinAltering rather than LOF variants but this situation is reversed for *GIGFY1* and to a lesser extent for *PCSK9*, *FES* and *GCK*. For *ANGPTL3* there is an additional contribution from SpliceRegion variants but this is not the case for the other genes. Detailed inspection of the results for *ANGPTL3* revealed that the contribution for the SpliceRegion category was driven by a single variant, 1:62598067T>C (rs372257803) which had frequency 0.0011 in controls and 0.00064 in cases. Again, these results suggest that there is no single weighting scheme which would be optimal for every gene considered individually.

Figure 3 illustrates the pattern of correlations between the different annotations obtained from dbNSFP for the variants used in this study. It can be seen that some annotations are strongly positively correlated with each other while others are not and that there are even pairs of annotations which are negatively correlated. This is reflected in the fact that different variant annotations make different contributions to evidence for association, as presented in Table 5 and Figure 4. Table 5 presents the SLP produced by each annotation individually while Figure 4 presents the same information graphically, again with the dot size indicating the SLP relative to the maximum SLP obtained by any annotation for the gene in question. It can be seen that both the absolute and relative magnitudes of the SLPs obtained vary between genes. For *PCSK1*, *ANGPLT3* and *GIGFY1* no annotation produces an SLP greater than 3 so the relative magnitudes may not be very meaningful but for the other genes there are clear differences in terms of which annotations produce the strongest evidence for association. For example, if we compare the results produced by MutationTaster and MutationAssessor we can see that the *PCSK9*, *DNMT3A* and *GCK* produce relatively large SLPs with MutationTaster and much smaller ones with MutationAssessor but for *MC4R*, *LDLR* and *HNF4A* the situation is reversed (Reva et al., 2011; Schwarz et al., 2014). No annotation produces consistently high SLPs across all genes.

Based on the information in Table 5 and Figure 4, a small group of annotations was selected to be incorporated jointly into multivariate analyses. These were the annotations producing the maximum SLP for each of the genes apart from *PCSK1*, *ANGPLT3* and *GIGFY1* (which were minimally informative), and consisted of VEST4, MutationTaster, Polyphen2_HVAR, MutationAssessor and PROVEAN (Adzhubei et al., 2013; Carter et al., 2013; Choi et al., 2012; Reva et al., 2011; Schwarz et al., 2014). In order to incorporate an annotation based on conservation, phastCons30way_mammalian was also selected since it produced moderately positive SLPs for a number of genes (Siepel et al., 2005). The scores for these annotations were incorporated along with the scores for the categories SpliceRegion, ProteinAltering and LOF in a multivariate analysis. The scores for these categories and annotations were treated as predictor variables and included along with sex and principal components as covariates. The Wald statistic was used to produce an SLP for each of these predictor variables and in order to assess the overall evidence for association a likelihood ratio test was performed to compare the likelihoods of the models which did and did not include the predictor variables. Since a total of 8 predictor variables were used, twice the difference in log likelihoods was taken to be a chi-squared statistic with 8 degrees of freedom. The SLPs for

individual predictors and for the overall analysis are shown in Table 6. Once more, it can be seen that the different annotation scores make different contributions to the evidence for association for different genes. For each gene, the multivariate model achieved a higher log likelihood maximised over the different predictor variables than the log likelihood achieved by maximising over a single weighted burden score. However, the fact that the multivariate model had more degrees of freedom meant that the likelihood ratio test using multiple predictors produced higher SLPs for some genes and lower SLPs for others. The overall results for the combined multivariate analysis can be compared to those previously reported for the original weighted burden tests which were shown in Table 2. It can be seen that for *PCSK1*, *LDLR* and *PCSK9* the evidence for association is substantially stronger using the multivariate analysis including the selected categories and annotations whereas for *DNMT3A*, *FES* and *HNF4A* the original weighted burden analysis using a single combined score produced substantially higher SLPs.

Discussion

If one knew in advance the effect size of every variant then a model which pre-specified these relative effects would have maximal power to detect association. However these simple exploratory analyses are sufficient to demonstrate that there is no single frequency-based weighting factor, selection of variant categories or group of secondary annotations which will consistently maximise the power to detect association for every gene. Armed with this knowledge, a number of approaches are available to the investigator. One is to choose a single weighting scheme which is hoped to be "good enough". Another is to perform repeated analyses with different choices of frequency weighting factors, annotation weights and variant selections and then to apply some kind of correction for multiple testing, for example by combining p values to yield a test statistic with a known null hypothesis distribution (Liu et al., 2019). A third is to perform likelihood maximisation over a multivariate model which includes a number of different annotations in order to produce a test statistic with a larger number of degrees of freedom. All of these approaches are sub-optimal in one way or another and there is not currently sufficient information available to make any firm recommendation. As further knowledge accrues as to which types of variant affect susceptibility to which types of phenotype in which types of gene then it is possible that the situation may change in the future. At present all we might say that it seems sensible to include both ProteinAltering and LOF variants, that the latter should be given more weight and that some kind of secondary annotations should be used to distinguish which ProteinAltering variants are more or less likely to be relevant. Also, some kind of weighting for MAF should be included so that rare variants are given higher weights than common ones.

It is worth stating a number of points explicitly. This study aims to investigate which types of variant most strongly support association in the context of attempting to identify genes affecting susceptibility to complex, common phenotypes. This differs from the problem of attempting to decide whether a given variant is "pathogenic" and might be responsible for a phenotype in an individual patient for whom one is attempting to make a genetic diagnosis. The models tested are simple linear combinations of rank scores and of course it is possible that more sophisticated models would have improved performance in particular situations. However we believe that the results obtained suggest that a complex model which performed well in one situation would likely perform less well in another. Given the heterogeneity of results we obtain across just a small number of genes it seems doubtful that applying more sophisticated approaches would produce a worthwhile benefit. Another point to emphasise is that the task addressed here is simply to establish that association exists between a phenotype and overall impairment of function of a particular gene. Once such an association is established then follow-up analyses might explore in more detail the

contribution of different variants and in this context the analysis of different annotations might well yield useful additional insights.

Conclusion

Given all the above considerations, we make no firm recommendations as to what weighting scheme should be used for weighted burden analysis. Hopefully the results presented here will be useful for investigators in order to inform a thoughtful approach to devising the analytic schemes which they feel are most appropriate for the situations they face.

Competing interests

The author declares he has no competing interests.

Acknowledgments

This research has been conducted using the UK Biobank Resource. The author wishes to acknowledge the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. This work was carried out in part using resources provided by BBSRC equipment grant BB/R01356X/1. The author wishes to thank the participants who volunteered for the UK Biobank project.

Ethics statement

UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001).

Author contributions

DC carried out the analyses and prepared the manuscript.

References

Adzhubei, I., Jordan, D.M., Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. 7 Unit7.20.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7, 248–249.

Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., Karchin, R. (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P. (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 7.

Curtis, D. (2012) A rapid method for combined analysis of common and rare variants at the level of a region, gene, or pathway. Adv Appl Bioinform Chem 5, 1–9.

Curtis, D. (2016) Pathway analysis of whole exome sequence data provides further support for the

involvement of histone modification in the aetiology of schizophrenia. Psychiatr. Genet. 26, 223–7.

Curtis, D. (2020) Multiple Linear Regression Allows Weighted Burden Analysis of Rare Coding Variants in an Ethnically Heterogeneous Population. Hum. Hered. 1–10.

Curtis, D. (2021a) Analysis of 200 000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. J. Med. Genet. jmedgenet-2021-107752.

Curtis, D. (2021b) Analysis of 200,000 Exome-Sequenced UK Biobank Subjects Implicates Genes Involved in Increased and Decreased Risk of Hypertension. Pulse 9, 17–29.

Curtis, D. (2021c) Weighted burden analysis in 200,000 exome-sequenced subjects characterises rare variant effects on risk of type 2 diabetes. medRxiv 2021.01.08.21249453.

Curtis, D., Bakaya, K., Sharma, L., Bandyopadhay, S. (2019) Weighted burden analysis of exome-sequenced late onset Alzheimer's cases and controls provides further evidence for involvement of PSEN1 and demonstrates protective role for variants in tyrosine phosphatase genes. Ann Hum Genet 84, 291–302.

Curtis, D., Coelewij, L., Liu, S.-H., Humphrey, J., Mott, R. (2018) Weighted Burden Analysis of Exome-Sequenced Case-Control Sample Implicates Synaptic Genes in Schizophrenia Aetiology. Behav. Genet. 43, 198–208.

Galinsky, K.J., Bhatia, G., Loh, P.R., Georgiev, S., Mukherjee, S., Patterson, N.J., Price, A.L. (2016) Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. 98, 456–472.

Hassan, M.S., Shaalan, A.A., Dessouky, M.I., Abdelnaiem, A.E., ElHefnawi, M. (2019) A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene.

Kumar, P., Henikoff, S., Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081.

Liu, X., Li, C., Mou, C., Dong, Y., Tu, Y. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 12.

Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., Lin, X. (2019) ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am. J. Hum. Genet. 104, 410.

Madsen, B.E., Browning, S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5, e1000384.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F. (2016) The Ensembl Variant Effect Predictor. Genome Biol. 17, 122.

Reva, B., Antipin, Y., Sander, C. (2011) Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 39.

Schwarz, J.M., Cooper, D.N., Schuelke, M., Seelow, D. (2014) Mutationtaster2: Mutation prediction for the deep-sequencing age. Nat. Methods.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.D.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Szustakowski, J.D., Balasubramanian, S., Sasson, A., Khalid, S., Bronson, P.G., Kvikstad, E., Wong, E., Liu, D., Davis, J.W., Haefliger, C., Loomis, A.K., Mikkilineni, R., Noh, H.J., Wadhawan, S., Bai, X., Hawes, A., Krasheninina, O., Ulloa, R., Lopez, A., Smith, E.N., Waring, J., Whelan, C.D., Tsai, E.A., Overton, J., Salerno, W., Jacob, H., Szalma, S., Runz, H., Hinkle, G., Nioi, P., Petrovski, S., Miller, M.R., Baras, A., Mitnaul, L., Reid, J.G. (2020) Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. medRxiv 2020.11.02.20222232.

Table 1. Table showing annotations produced by VEP, the weights assigned to them for weighted burden analysis and the categories they were assigned to for the current analyses. Annotations marked as unused were not applied to any of the variants in the genes studied.

VEP annotation	Weight	Category
intergenic_variant	0	Unused
feature_truncation	0	IntronicEtc
regulatory_region_variant	0	IntronicEtc
feature_elongation	0	IntronicEtc
regulatory_region_amplification	1	IntronicEtc
regulatory_region_ablation	1	IntronicEtc
TF_binding_site_variant	1	IntronicEtc
TFBS_amplification	1	IntronicEtc
TFBS_ablation	1	IntronicEtc
downstream_gene_variant	0	IntronicEtc
upstream_gene_variant	0	IntronicEtc
non_coding_transcript_variant	0	IntronicEtc
NMD_transcript_variant	0	IntronicEtc
intron_variant	0	IntronicEtc
non_coding_transcript_exon_variant	0	IntronicEtc
3_prime_UTR_variant	1	ThreePrime
5_prime_UTR_variant	1	FivePrime
mature_miRNA_variant	5	Unused
coding_sequence_variant	0	Unused
synonymous_variant	0	Synonymous
stop_retained_variant	5	Synonymous
incomplete_terminal_codon_variant	5	Unused
splice_region_variant	1	SpliceRegion
protein_altering_variant	5	ProteinAltering
missense_variant	5	ProteinAltering
inframe_deletion	10	InDelEtc
inframe_insertion	10	InDelEtc
transcript_amplification	10	InDelEtc
start_lost	10	ProteinAltering
stop_lost	10	ProteinAltering
frameshift_variant	100	LOF
stop_gained	100	LOF
splice_donor_variant	100	LOF
splice_acceptor_variant	100	LOF
transcript_ablation	100	LOF

Table 2. List of genes used for these analyses along with the SLP obtained in the original analysis with the corresponding phenotype. Variants which impaired functioning of *PCSK9* and *ANGPTL3* were found to be protective against hyperlipidaemia so for convenience the phenotype of interest is stated to be "Not hyperlipidaemia".

Phenotype	**Gene symbol**	**Gene name**	**SLP**
BMI	*MC4R*	Melanocortin 4 Receptor	15.79
BMI	*PCSK1*	Proprotein Convertase Subtilisin/Kexin Type 1	6.61
Hyperlipidaemia	*LDLR*	Low Density Lipoprotein Receptor	50.08
Not hyperlipidaemia	*PCSK9*	Proprotein Convertase Subtilisin/Kexin Type 9	10.42
Not hyperlipidaemia	*ANGPTL3*	Angiopoietin Like 3	5.67
Hypertension	*DNMT3A*	DNA Methyltransferase 3 Alpha	8.21
Hypertension	*FES*	FES Proto-Oncogene, Tyrosine Kinase	6.10
Type 2 diabetes	*GCK*	Glucokinase	22.25
Type 2 diabetes	*HNF4A*	Hepatocyte Nuclear Factor 4 Alpha	6.82
Type 2 diabetes	*GIGYF1*	GRB10 Interacting GYF Protein 1	6.22

Table 3. SLPs for weighted burden tests for the studied genes generated individually by variant categories ProteinAltering and LOF using different values for WF, the weighting function based on MAF. Values of log10(WF) range from 0 (WF=1) to 2 (WF=100). Sex and principal components were included as covariates.

Table 3A. SLPs for LOF variants.

log10(WF)	*MC4R*	*PCSK1*	*LDLR*	*PCSK9*	*ANGPTL3*	*DNMT3A*	*FES*	*GCK*	*HNF4A*	*GIGYF1*
0.0	6.60	3.63	12.79	7.52	3.50	5.38	3.26	13.61	-0.06	11.16
0.2	6.61	3.63	12.74	7.50	3.49	5.38	3.26	13.61	-0.06	11.16
0.4	6.61	3.62	12.71	7.49	3.49	5.38	3.26	13.61	-0.06	11.16
0.6	6.61	3.62	12.69	7.48	3.48	5.38	3.26	13.61	-0.06	11.16
0.8	6.62	3.61	12.69	7.48	3.47	5.38	3.26	13.61	-0.06	11.16
1.0	6.62	3.61	12.68	7.48	3.47	5.38	3.26	13.61	-0.06	11.16
1.2	6.62	3.60	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16
1.4	6.62	3.60	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16
1.6	6.62	3.59	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16
1.8	6.62	3.60	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16
2.0	6.62	3.60	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16

Table 3B. SLPs for ProteinAltering variants.

log10(WF)	*MC4R*	*PCSK1*	*LDLR*	*PCSK9*	*ANGPTL3*	*DNMT3A*	*FES*	*GCK*	*HNF4A*	*GIGYF1*
0.0	-11.32	6.52	21.03	4.24	1.99	2.32	-0.01	3.99	2.74	0.98
0.2	-7.78	6.90	26.21	4.13	2.06	2.58	0.23	4.11	2.80	0.92
0.4	-3.78	7.46	29.23	4.03	2.03	2.75	0.50	4.18	2.83	0.87
0.6	-0.89	8.24	30.43	3.97	1.93	2.87	0.72	4.23	2.85	0.82
0.8	0.26	9.17	30.73	3.92	1.83	2.95	0.89	4.26	2.86	0.79
1.0	1.56	9.87	30.69	3.89	1.74	2.99	1.01	4.28	2.87	0.77
1.2	3.00	9.81	30.57	3.87	1.68	3.02	1.09	4.29	2.88	0.76
1.4	4.17	8.91	30.45	3.86	1.64	3.04	1.14	4.30	2.88	0.75
1.6	5.00	7.73	30.37	3.85	1.62	3.05	1.17	4.31	2.88	0.74
1.8	5.56	6.72	30.30	3.84	1.60	3.06	1.19	4.31	2.88	0.74
2.0	5.92	6.00	30.26	3.84	1.59	3.07	1.21	4.31	2.88	0.74

Table 4 SLPs produced individually by each variant category for each gene using WF=25.12, including sex and principal components as covariates.

Variant category	*MC4R*	*PCSK1*	*LDLR*	*PCSK9*	*ANGPTL3*	*DNMT3A*	*FES*	*GCK*	*HNF4A*	*GIGYF1*
FivePrime	1.00	-0.11	0.78	-0.18	-0.01	-0.22	0.00	0.66	-0.30	0.02
InDelEtc	0.32	-0.15	0.12	-0.35	0.02	0.59	0.37	1.63	-0.06	0.24
IntronicEtc	0.00	1.37	-0.40	-0.33	-0.31	-0.75	-0.13	0.29	0.05	-0.13
LOF	6.62	3.60	12.68	7.47	3.47	5.38	3.26	13.61	-0.06	11.16
ProteinAltering	4.17	8.91	30.45	3.86	1.64	3.04	1.14	4.30	2.88	0.75
SpliceRegion	0.00	0.53	0.27	-0.20	3.02	0.50	-0.32	0.10	-0.94	0.09
Synonymous	-0.59	0.76	-1.13	-0.99	0.18	0.40	-0.37	-0.17	-0.53	-0.07
ThreePrime	0.12	-0.08	0.00	0.43	0.10	-0.97	-0.18	-0.25	-0.02	-0.41

Table 5 SLPs produced individually by each secondary annotation from dbNSFP for each gene using WF=25.12, including sex and principal components as covariates.

Annotation	*MC4R*	*PCSK1*	*LDLR*	*PCSK9*	*ANGPTL3*	*DNMT3A*	*FES*	*GCK*	*HNF4A*	*GIGYF1*
Polyphen2_HDIV_rankscore	9.56	-0.73	17.12	1.95	0.39	1.36	4.32	5.12	3.40	0.11
SIFT_converted_rankscore	6.21	-0.14	10.15	5.08	0.83	0.42	2.41	5.36	3.97	0.05
SIFT4G_converted_rankscore	6.81	0.10	14.40	2.04	0.27	2.44	1.61	5.77	3.66	0.10
Polyphen2_HVAR_rankscore	8.64	-0.37	19.14	2.39	0.50	0.98	4.41	5.39	3.27	0.12
LRT_converted_rankscore	4.49	-0.07	14.65	3.95	-0.10	4.46	4.29	8.44	2.10	0.70
MutationTaster_converted_rankscore	4.77	0.38	9.38	9.57	0.30	6.93	1.90	13.90	1.90	0.56
MutationAssessor_rankscore	7.95	-0.45	18.40	1.68	1.61	1.61	1.32	4.35	3.03	0.00
FATHMM_converted_rankscore	0.32	2.42	0.08	2.30	0.13	0.40	-1.01	-0.58	-0.20	0.42
PROVEAN_converted_rankscore	5.92	0.28	10.89	4.45	1.08	0.66	1.67	2.85	4.49	-0.11
VEST4_rankscore	12.48	-0.06	15.27	2.91	0.44	4.58	3.16	12.51	3.34	0.04
MetaSVM_rankscore	7.04	-0.14	10.81	2.39	0.86	0.48	1.43	-0.62	0.51	0.15
MetaLR_rankscore	8.44	-0.25	7.14	1.80	0.71	0.63	1.23	1.01	0.29	0.31
M_CAP_rankscore	11.50	0.02	1.15	3.03	0.44	2.50	0.76	2.12	2.26	0.15
REVEL_rankscore	11.28	0.13	11.09	3.02	0.77	1.69	1.29	1.98	1.40	0.14
MutPred_rankscore	2.73	2.08	6.29	0.33	0.25	1.22	1.81	1.20	1.76	-0.03
MVP_rankscore	0.72	-2.63	-2.45	0.35	-0.01	-0.64	-1.32	-0.11	-0.52	0.12
MPC_rankscore	4.20	-1.80	1.26	1.17	0.05	-0.76	-0.44	1.59	0.21	0.13
PrimateAI_rankscore	0.16	-2.70	2.71	1.90	-0.49	0.09	-0.34	2.24	1.06	0.03
DEOGEN2_rankscore	-0.21	-0.88	-1.20	0.81	0.12	-0.29	-1.26	-0.09	-0.24	0.05
BayesDel_addAF_rankscore	7.04	-0.29	2.38	2.43	0.38	3.74	1.30	9.18	1.38	2.55
BayesDel_noAF_rankscore	3.34	-0.61	-0.06	3.41	0.41	2.18	0.23	4.89	-0.03	1.76
ClinPred_rankscore	4.96	-2.10	9.19	0.57	0.41	0.94	0.52	2.71	2.24	0.04
LIST_S2_rankscore	0.83	-2.81	0.30	1.77	0.11	-0.03	-0.40	1.49	0.02	0.23
CADD_raw_rankscore	2.33	-0.78	7.63	5.30	0.54	2.83	1.19	11.21	0.48	2.44
CADD_raw_rankscore_hg19	1.26	-0.68	8.04	7.66	0.59	2.62	1.54	13.51	0.59	2.32
DANN_rankscore	0.91	-1.78	6.90	5.92	0.94	3.08	1.91	5.35	0.92	0.55
fathmm_MKL_coding_rankscore	3.16	-0.07	11.35	3.37	0.23	4.24	1.50	5.87	0.91	0.98
fathmm_XF_coding_rankscore	0.70	0.16	2.71	0.55	0.25	1.95	1.86	1.45	0.49	0.17
Eigen_raw_coding_rankscore	4.98	0.28	9.19	5.18	0.71	5.69	2.24	5.52	1.52	1.21
Eigen_PC_raw_coding_rankscore	2.59	0.28	6.78	4.99	0.52	4.20	2.05	7.48	1.71	1.20
GenoCanyon_rankscore	0.75	0.72	2.73	2.46	0.46	3.07	-0.21	5.14	-0.06	1.30
integrated_fitCons_rankscore	5.63	1.01	2.30	0.29	-0.19	1.19	-0.34	3.52	2.35	0.17
GM12878_fitCons_rankscore	1.23	0.63	0.27	-0.25	0.01	0.48	0.20	0.01	-0.07	0.48
H1_hESC_fitCons_rankscore	-0.40	0.50	-0.06	0.10	-0.03	0.99	-0.09	-0.15	-0.21	0.39
HUVEC_fitCons_rankscore	0.53	0.48	0.08	0.12	-0.23	0.29	-0.04	2.36	0.36	0.42
GERPPP_RS_rankscore	-0.18	0.12	2.54	-0.06	0.97	0.41	0.46	0.42	-0.03	0.26
phyloP100way_vertebrate_rankscore	2.13	1.05	2.12	-0.13	0.18	1.00	0.85	1.63	1.05	-0.28
phyloP30way_mammalian_rankscore	-0.16	0.74	2.63	-0.19	-0.21	-0.16	1.64	0.96	1.21	-0.15
phyloP17way_primate_rankscore	1.34	1.34	0.14	1.26	0.90	0.29	1.62	6.12	0.76	-0.21
phastCons100way_vertebrate_rankscore	0.51	-0.04	5.18	2.67	0.00	1.55	1.11	0.25	0.12	0.75

phastCons30way_mammalian_rankscore	2.41	1.11	0.61	2.97	0.56	1.23	1.88	2.55	2.47	-0.39
phastCons17way_primate_rankscore	0.18	1.05	2.62	0.84	0.81	0.12	0.59	2.16	0.94	0.40
SiPhy_29way_logOdds_rankscore	0.38	0.46	9.83	1.75	1.47	0.46	0.89	2.28	0.19	2.32

Table 6 Results of multivariate analysis including selected categories and annotations with sex and principal components as covariates. The SLP obtained from the Wald statistics for each predictor variable is shown followed by the SLP for the likelihood ratio test incorporating all variables. Also shown for comparison are the SLPs obtained from the original weighted burden tests, as first presented in Table 2.

Category / Annotation	*MC4R*	*PCSK1*	*LDLR*	*PCSK9*	*ANGPTL3*	*DNMT3A*	*FES*	*GCK*	*HNF4A*	*GIGYF1*
SpliceRegion	NA	0.53	0.26	-0.12	3.03	0.47	-0.33	0.10	-0.43	0.09
ProteinAltering	-1.58	1.48	-0.70	-0.54	0.07	-0.74	-1.05	-1.47	-0.18	0.76
LOF	7.08	2.75	13.15	3.85	3.56	0.52	3.01	4.54	-0.06	10.55
MutationTaster_converted_rankscore	-0.40	-0.08	-0.40	1.45	-0.36	1.05	-0.29	1.94	-0.47	-0.47
Polyphen2_HVAR_rankscore	1.80	-0.74	3.71	-0.88	-0.25	0.56	2.79	0.75	0.44	0.58
MutationAssessor_rankscore	1.59	-1.36	3.20	-0.24	1.16	0.06	-0.21	1.79	-0.05	NA
PROVEAN_converted_rankscore	-0.77	2.44	-0.11	1.76	-0.06	-0.08	0.00	-2.31	1.52	-0.10
phastCons30way_mammalian_rankscore	0.79	1.21	-0.51	0.62	0.58	0.08	0.76	0.25	1.17	-0.60
SLP for likelihood ratio test	14.73	10.26	65.73	13.22	5.40	5.90	4.67	20.73	4.16	5.80
SLP previously obtained from standard weighted burden analysis	15.79	6.61	50.08	10.42	5.67	8.21	6.1	22.25	6.82	6.22

Figure 1 SLPs for *LDLR*, *MC4R* and *PCSK1* generated by the ProteinAltering variants using different values for WF, the weighting function based on MAF.

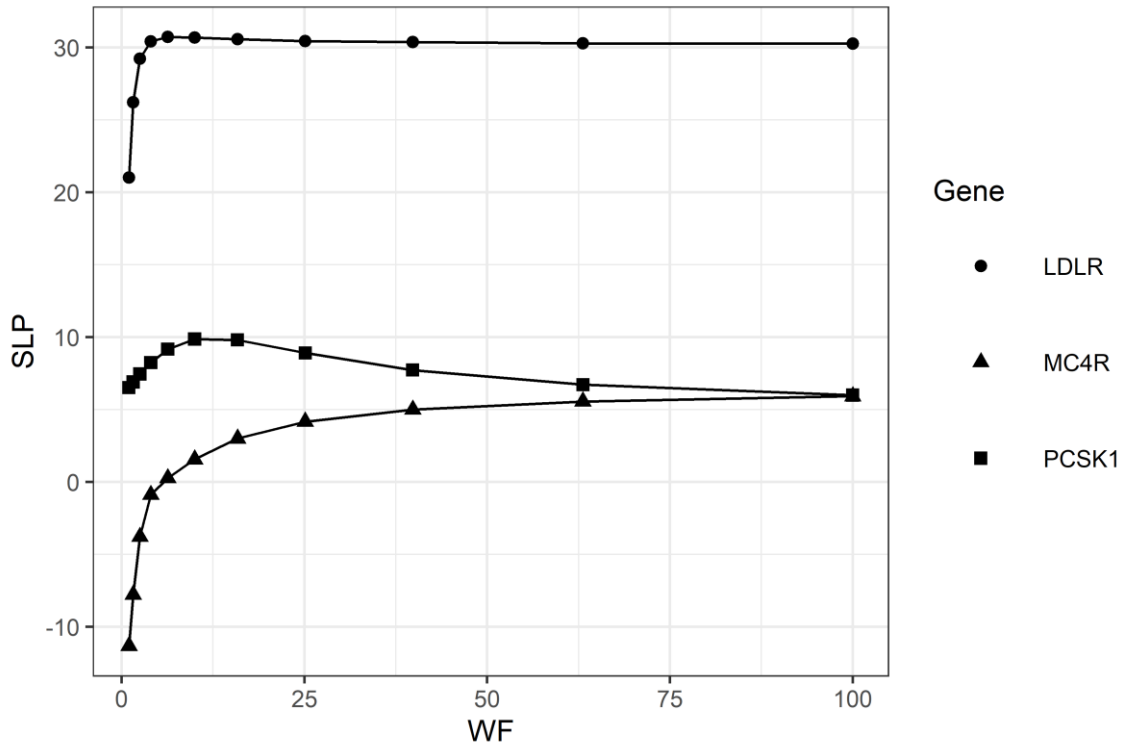

Figure 2 Heatmap of SLPs produced by each variant category for each gene. The size of the dots for each gene are proportional to the SLP for each variant category relative to the maximum SLP obtained for that gene. White circles indicate negative SLPs.

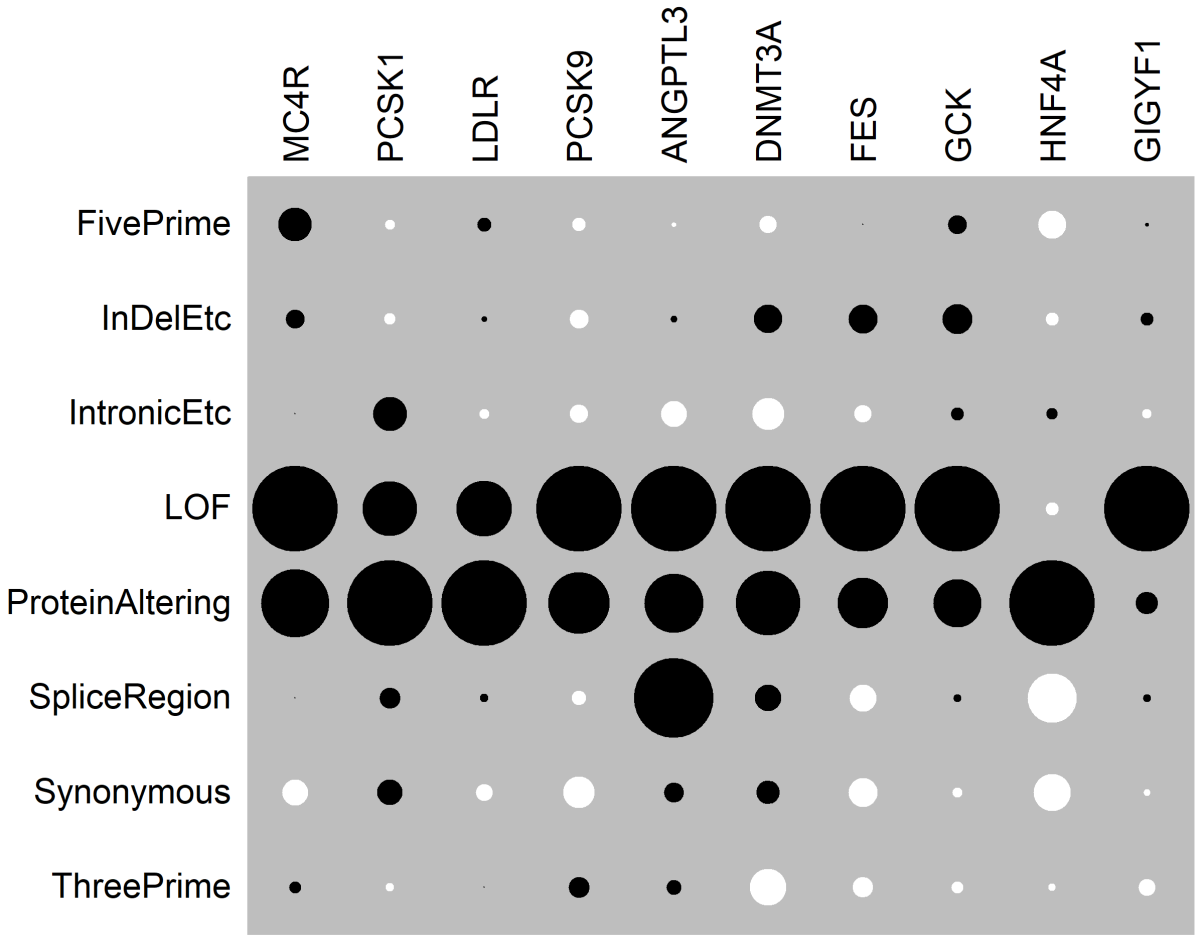

Figure 3 Plot of pairwise correlations between dbNSFP annotations for the variants used in this study. Black circles indicate positive correlations and white circles negative correlations.

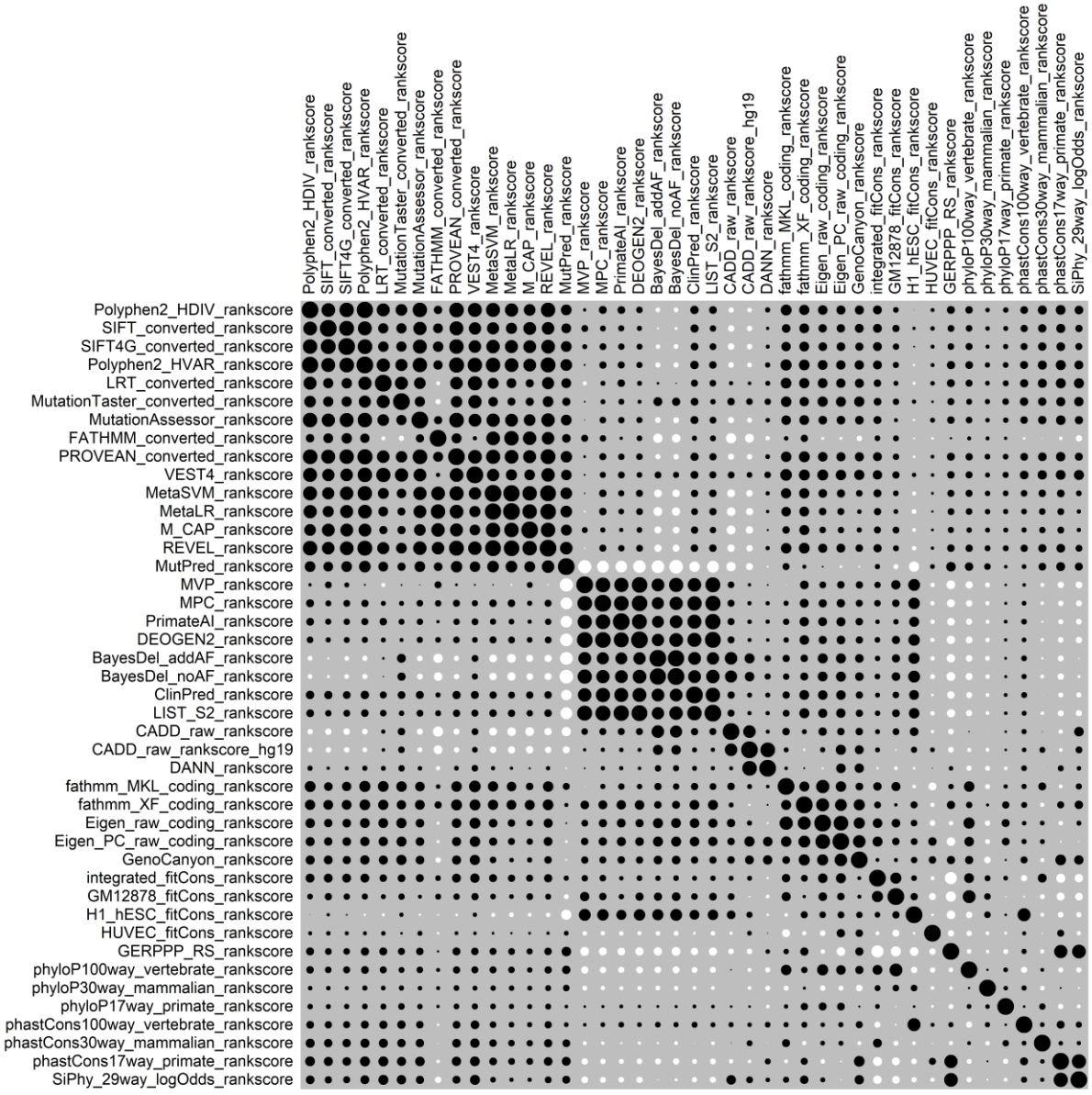

Figure 4 Heatmap of SLPs produced by each dbNSFP annotation for each gene. The size of the dots for each gene are proportional to the SLP for each annotation relative to the maximum SLP produced by any annotation for that gene. White circles indicate negative SLPs.