Multiple linear regression allows weighted burden analysis of rare coding variants in an ethnically heterogeneous population

David Curtis

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

Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ.

Short title: **Weighted burden analysis in an ethnically heterogeneous population**

Correspondence:

David Curtis

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

d.curtis@ucl.ac.uk

00 44 7973 906 143

Abstract

Weighted burden analysis has been used in exome-sequenced case-control studies to identify genes in which there is an excess of rare and/or functional variants associated with phenotype. Implementation in a ridge regression framework allows simultaneous analysis of all variants along with relevant covariates such as population principal components. In order to apply the approach to a quantitative phenotype, a weighted burden score is derived for each subject and included in a linear regression analysis. The weighting scheme is adjusted in order to apply differential weights to rare and very rare variants and a score is derived based on both the frequency and predicted effect of each variant. When applied to an ethnically heterogeneous dataset consisting of 49,790 exome-sequenced UK Biobank subjects and using BMI as the phenotype the method produces a very inflated test statistic. However this is almost completely corrected by including 20 population principal components as covariates. When this is done the top 30 genes include a few which are quite plausibly associated with the phenotype, including *LYPLAL1* and *NSDHL*. This approach offers a way to carry out gene-based analyses of rare variants identified by exome sequencing in heterogeneous datasets without requiring that data from ethnic minority subjects be discarded. This research has been conducted using the UK Biobank Resource.

Keywords

Exome; BMI; ethnicity; *LYPLAL1*; *NSDHL*.

Introduction

We have previously developed a method of weighted burden analysis which allows all variants within a gene to be included in a case-control analysis to test whether there is on average an excess of highly weighted variants among cases. In the original conception, implemented in the SCOREASSOC program, both common and rare variants were included but a parabolic function based on minor allele frequency (MAF)was applied such that rare variants would be assigned a higher weight than common ones [1]. This approach was subsequently extended in a number of ways. Functional weights were assigned based on the predicted effect of each variant on the function of the gene, so that for example variants predicted to produce a truncated protein product would be weighted more highly than synonymous variants, and an overall weight for each variant was derived as the product of the frequency and functional weights [2]. Unlike other approaches applied to exome sequence data, this means that all variants can be included in a single combined analysis without having to dichotomise according to allele frequency or predicted impact [3,4]. Sets of genes within a metabolic pathway could be jointly analysed by testing whether the overall burden of rare, functional variants varied between cases and controls across the set of genes rather than within an individual gene [2]. The comparison of variant burden was then implemented in a ridge logistic regression framework and this allowed the inclusion of covariates such as population principal components as well as additional risk factors such as pathogenic copy number variants and polygenic risk scores [5]. These approaches were applied to large samples of exome sequenced cases and controls and implicated genes affecting functioning of the glutamatergic NMDA receptor in schizophrenia and genes coding for tyrosine phosphatases in late onset Alzheimer’s disease [6,7].

Additional exome sequenced datasets are becoming available and some of these, such as the UK Biobank, have been phenotyped for a number of quantitative traits [8]. It is plausible that variants disrupting the functioning of particular genes might be associated with changes in the mean value of some of these traits and an obvious approach would be to carry out linear regression rather than logistic regression in order to test for this. However there are some important considerations which need to be addressed.

In contrast to a targeted case-control study, biobank samples may not be ethnically well-matched and this is expected to impact testing for an excess of rare, functional variants in exome-sequenced sample. Even in studies which seek to match cases and controls there may be residual stratifications which affect the results and this occurred in the Swedish schizophrenia study [9]. Here, a higher proportion of cases than controls had a substantial Finnish component to ancestry and at the same time there was a lower frequency of rare, damaging variants in those with Finnish ancestry generally, across both cases and controls. In fact, there was an excess of rare damaging variants among the schizophrenia cases but this only became apparent when ancestry was included as a covariate or when the analysis was restricted to subjects without Finnish ancestry [6,9]. This provides a practical example of how unrecognised population stratification may distort the results of rare variant burden analyses. Of particular relevance to biobanks is that there is an expectation that such stratification will produce artefactual results even if there is in fact no difference in the general distribution of variant allele frequencies between subjects with different ancestries. If the bulk of subjects share similar ancestry but there is a minority cohort with different ancestry then we can expect by chance that some variants will have different frequencies between the main and minority cohorts. However, a variant which is relatively common in the minority cohort will have a lower frequency in the sample overall. This results in the expectation that there will appear to be an excess of apparently rarer variants in minority cohorts. If the phenotype in question has a different mean value in the minority cohort then there will be an artefactual association between the phenotype and rare variants, whether rarity is defined using a threshold or is used in a weighting scheme.

One approach to dealing with heterogenous ancestry within biobank samples is to restrict attention only to those of a particular ethnicity. The UK Biobank sample contains 503,317 subjects of whom 94.6% are of white ethnicity, somewhat higher than for the population as a whole [10]. Regrettably, it seems to have become standard practice to simply ignore the non-white subjects. To give two recent examples, a genome-wide meta-analysis of problematic alcohol use only considered 435,563 European-ancestry individuals and a genome-wide association study of susceptibility to keratitis only considered 337,199 subjects of European ancestry [11,12]. Likewise, a recent analysis of the 49,960 exome-sequenced UK Biobank subjects only used information from 45,596 European subjects [3]. This latter study applied six different methods of burden analysis and although it incorporated population principal components as covariates it still restricted itself to ancestry-matched case-controls. Restricting attention to subjects with white European ancestry within UK Biobank discards information from thousands of citizens who have volunteered personal information, donated biological samples and undergone uncomfortable investigations with the aim of contributing to knowledge about disease. We believe that simply ignoring data from ethnic minority subjects is ethically indefensible [13].

A second issue to be addressed in weighted burden testing is the nature of the weighting which is to be applied according to MAF. In its original conception, the method was intended to incorporate variants of all allele frequencies and an example application used Crohn's diseases as the phenotype, since susceptibility is influenced by both common and rare variants [1]. However over the course of countless genome wide association studies it has become apparent that it is in fact very unusual for common variants to exert substantial effects on risk and when dealing with next generation sequence data it does not make sense to include common variants since they will essentially produce noise which may swamp any real signal. However with the weighting scheme originally proposed, described by a parabola with value of 1 at MAF = 0.5 increasing to 10 at MAF ~= 0, the allocated weight is almost the same for a variant with MAF = 0.01 as it is for ultra-rare variants. Since selection pressures mean that very rare variants can have larger effect sizes than less rare ones, it would be desirable to have a revised weighting scheme which would distinguish rare from ultra-rare variants [14].

A third issue to address when dealing with quantitative traits is how information from different genes within a set should be combined. The original assumptions were that rare variants with a functional impact were more likely to impair normal function of a gene than enhance it and hence that such variants were more likely to be associated with a disease phenotype across different genes. This would mean that variant scores could simply be added across genes within a set [2]. With a quantitative trait we may still expect that rare, functionally impactful variants may impair gene functioning but we need to recognise that impairing the function of some genes may tend to reduce the value of a trait whereas impairing the function of other genes in the same pathway may increase the value. Since the effects may be in opposite directions, we will not wish to simply add up variant scores across related genes unless we can be very confident of the direction of the effect we are expecting. Since we are intending to use this approach for gene discovery we will not generally be in a position to predict the likely direction of effect and so we require a method which is agnostic regarding this.

Here we develop the weighted burden approach previously applied to case-control data in three different ways. We produce a modified scheme for weighting by allele frequency to give additional weight to extremely rare variants. We implement linear regression rather than logistic regression and incorporate population principal components as covariates. We combine evidence across different genes within a set using p values rather than variant scores. We apply the revised method to analyse a quantitative trait, BMI, in an ethnically heterogeneous dataset, the UK Biobank.

Methods

Variant weight according to allele frequency

The original formula we proposed for assigning a weight according to allele frequency was

$$W_i=(4f-4)q_i^2-(4f-4)q_i+f$$

where q_i is the allele frequency of the *i*th variant and f is a weighting factor equal to or greater than 1. The formula describes a parabola with a minimum value of 1 at q=0.5 and rising to f at q=0 and q=1. In typical applications a value of 10 would be chosen for f.

In order to provide a scheme which more strongly distinguishes rare from very rare variants, we will choose q' to mean the frequency of the rarer allele and we will set a threshold of t for the MAF. We will either discard variants with $q'>t$ or else we will assign them a weight equal to 1. For those with $q'<=t$ we will assign a weight as:

$$W_i=(4f-4)(0.5*q'_i/t)^2-(4f-4)(0.5*q'_i/t)+f$$

This simply produces weights which increase parabolically from a value of 1 at q=t or q=1-t towards a value of f at q=0 and q=1.

Implementation of linear regression

As described previously, testing for an effect of gene variants on a case-control phenotype involved constructing a weighted burden score for each subject and then carrying out a likelihood ratio test using ridge regression comparing the likelihoods for models which did and did not include the burden scores [5]. Here, we simply calculate the likelihoods using linear regression rather than logistic regression. Each variant is assigned a weight based on its predicted likely effect on gene function and then this functional weight is multiplied by the weight based on the variant frequency as described above. For each subject a gene-wise weighted burden score is derived as the sum of the variant-wise weights, each multiplied by the number of alleles of the variant which the given subject possesses. If a subject is not genotyped for a variant then they are assigned the subject-wise average score for that variant. A number of covariates can be included for each subject, typically consisting of population principal components though if desired factors such as age, polygenic risk score and presence or absence of other known risk factors can also be used. The score and covariates are entered into a standard linear regression model with quantitative phenotype as the outcome variable and after variable normalisation the likelihood of the model is maximised using the L-BFGS quasi-newton method, implemented using the dlib library (King, 2009). The contribution of different variables to risk is assessed using standard likelihood ratio tests by comparing twice the difference in maximised log likelihoods between models with and without the variables of interest. This likelihood ratio statistic is then taken as a chi-squared statistic with degrees of freedom equal to the difference between models in number of variables fitted. The coefficients for each variable can be varied to maximise the likelihood or can be fixed, for example if the effect size of a particular risk factor is known from epidemiological studies. This approach was implemented in a modified version of SCOREASSOC, which outputs the coefficients for the fitted models along with their estimated standard errors and the results of the likelihood ratio test. When association with the gene-wise weighted burden score alone is tested, i.e. when the two models differ only in whether or not the score is included, then the statistical significance is summarised as a signed log p value (SLP) which is the log base 10 of the p value given a positive sign if the score correlates positively with the quantitative phenotype and negative if it correlates negatively. For other analyses the minus log base 10 of the p value (MLP) is output. The support program for SCOREASSOC, called GENEVARASSOC, was also modified to deal with quantitative phenotypes and the linear regression tests.

Gene set analysis

In case-control analyses the assumption was made that rare, functional variants tended to impair gene function and that impaired gene function increased disease risk, meaning that weighted burden scores could simply be added across genes. For a quantitative trait, impaired function of a gene within a metabolic pathway might either increase or decrease the value of the trait and since genes might have opposite effects it is not appropriate to simply sum the scores of genes within a set. Instead, Fisher's method for combining p values can be applied. This assumes that, under the null hypothesis that no genes within a set influence the value of the trait, the sum of the natural logs of their *p* values multiplied by -2 will follow a chi-squared distribution with degrees of freedom equal to the twice number of genes. This can conveniently be tested by summing the absolute values of the SLPs and multiplying by -2ln(10) to use as the chi-squared statistic. The statistical significance of the test that one or more genes in the set influence phenotype can then be expressed as minus log base 10 of the p value (MLP) of the chi-squared test.

Practical application to an example dataset

The UK Biobank dataset was downloaded along with the variant call files for 49,953 subjects who had undergone exome-sequencing and genotyped using the GRCh38 assembly with coverage 20X at 94.6% of sites on average [8]. UK Biobank had obtained ethics approval from the Research Ethics Committee (REC; approval number: 11/NW/0382) and written informed consent from all participants. Analysis of the data was approved by the UCL Research Ethics Committee (approval number 11527/001). All variants were annotated using VEP, PolyPhen and SIFT [15–17]. To obtain population principal components reflecting ancestry, version 1.90beta of *plink* (https://www.cog-genomics.org/plink2) was run on these variants with the options *--maf 0.1 --pca header tabs --make-rel* [18–20].

The example quantitative phenotype chosen was BMI, which is known to be moderately heritable and which was available for 49,790 subjects. In order to better understand the structure of the data a number of other variables were studied including self-declared ethnicity, birth coordinates within the UK, year of birth and Townsend index of deprivation. The relationships between these variables and the principal components were investigated using multivariate analyses implemented in R and visualised using *ggplot2* and *ggpubr* (https://cran.r-project.org/web/packages/ggpubr/index.html) [21,22].

Analysis was restricted to variants have MAF<=0.01 and weighting based on frequency was applied as described above with a weighting factor, *f*, of 10 and a threshold, *t*, of 0.01. Previous association studies have demonstrated that there are no common variants with a substantial effect on BMI. Additionally, the weighted burden approach explicitly assumes that variants disrupting the gene have a similar direction of effect whereas the minor allele of a common variant might be associated with either increased or reduced gene function and hence including common variants could add unwanted noise to the analysis. Variants were also weighted according to their functional annotation using the default weights provided with the GENEVARASSOC program, which was used to generate input files for weighted burden analysis by SCOREASSOC. This weighting scheme is intended to allocate a higher weight to variants which are expected to have a major effect on gene function than those which are expected to have little effect. For example, a weight of 5 was assigned for a synonymous variant, 10 for a non-synonymous variant and 20 for a stop gained variant or frameshift variant, which would be expected to completely prevent one copy of the gene from being expressed. Additionally, 10 was added to the weight of a variant if the PolyPhen annotation was “possibly” or “probably” damaging and also if the SIFT annotation was “deleterious”, meaning that a non-synonymous variant annotated as both damaging and deleterious would be assigned a weight of 30. The full set of weights is shown in Supplementary Table 1, copied from the previous reports which used this method [6,7]. Variants were excluded if there were more than 10% of genotypes missing in the controls or if the heterozygote count was smaller than both homozygote counts. Each variant was then assigned an overall rate consisting of the product of the frequency weight and the functional weight. For each subject a gene-wise weighted burden score was derived as the sum of the variant-wise weights, each multiplied by the number of alleles of the variant which the given subject possessed.

Gene-wise weighted burden tests for association with BMI were carried out for every autosomal and X chromosome gene listed in the RefSeq GRCh38 release. For each gene, three analyses were carried out. In the first, weighted burden score was used to predict BMI in a simple linear regression model and in the second analysis the first 20 principal components were additionally included as covariates, with likelihood ratio tests used to assess statistical significance as described above. To assess the effects of introducing the new scheme for weighting by allele frequency, a third analysis was carried out also including the principal components but using the original scheme for weighting by frequency, meaning that the frequency weights for all variants would be very similar since all had MAF<=0.01. The third analysis was performed simply for the purposes of making this comparison and the definitive results are intended to be those obtained from using the new weighting scheme, as well as incorporating the principal components.

Gene set analyses were carried out using the 1454 "all GO gene sets, gene symbols" pathways as listed in the file *c5.all.v5.0.symbols.gmt* downloaded from the Molecular Signatures Database at http://www.broadinstitute.org/gsea/msigdb/collections.jsp [23]. For each set of genes the SLPs from the analyses incorporating principal components and new weighting scheme were combined using Fisher’s method as described above, yielding an MLP as a test of association of the set with the BMI.

Results

Some principal components were associated with self-declared ethnicity, with the first principal component distinguishing African from European ancestry, the second Asian from European and the fourth South Asian from East Asian (Supplementary Figures 1 and 2). Principal components also varied with place of birth, year of birth and Townsend deprivation index (Supplementary Figures 3 to 6). Likewise, BMI also varied with self-declared ethnicity as shown in Figure 1 and Table 1, as well as with demographic variables (Supplementary Figure 7). Finally, BMI varied with the principal components (Supplementary Figure 8). Thus BMI is a phenotype which can vary between subjects with different ancestries and could be liable to produce artefactual results according to mechanisms such as those outlined above.

There were 21,644 genes for which there were qualifying variants and the QQ plots for the SLPs are displayed in Figure 2. It can be seen that when principal components are not included there is a very marked inflation of the positive SLPs and that this is almost entirely corrected when the principal components are used as covariates. With the principal components included in the analysis, if the highest and lowest 100 SLPs are excluded (since they might capture a true biological effect) then the negative SLPs have an intercept of -0.022 and a gradient of 0.996 while the positive SLPs have an intercept of -0.006 and a gradient of 1.08, indicating only a fairly modest amount of inflation of the positive SLPs. If the original scheme for weighting by allele frequency was used instead then the results obtained were overall fairly similar, as seen by comparing Figure 2B and 2C. The correlation between the two sets of SLPs was 0.91 and there were only 182 genes for which the difference between the two SLPs was greater than 1. The largest change was for *TXNDC16*, which produced an SLP of 3.89 with the new weighting scheme compared to only 0.51 with the original one, whereas the second largest change was for *ABCA13*, which had an SLP of 1.43 with the new scheme and 3.96 with the original one.

Applying a Bonferroni correction for the number of genes tested would mean that the absolute value of the SLP would need to exceed log10(21,644*20)=5.6 or, allowing for the inflation referred to above, log10(21,644*20)*1.08=6.08, in order to be regarded as statistically significant. This was only achieved by one gene, *CCDC140*, which codes for a small effector of CDC42 and does not seem a particularly plausible candidate to have a direct biological influence on BMI. If the test were well behaved then under the null hypothesis one would expect 11 genes to have SLP of 3 or more and 11 to have SLP of -3 or less. The observed numbers are 22 and 7, possibly reflecting the modest inflation of the positive SLPs. Genes with absolute value of SLP of 3 or more, equivalent to p<=0.001, are listed in Table 2. Although these do not meet formal standards of statistical significance after correction for multiple testing, there are a few for which disruption of functioning might plausibly affect BMI.

The result for *ACAD11* may be of some interest because its product is involved in mitochondrial beta-oxidation of lipids and energy production. Additionally, two common variants in *ACAD10*, which has a similar function, are associated with obesity and type 2 diabetes in Pima Indians and *Acad10* knockout mice have excess weight gain which increases with age [24,25]. However in the present study the SLP for *ACAD10* itself is only -0.51.

There is good evidence from previous GWASs that common variants in or near *LYPLAL1* are associated with a variety of metabolic traits including central obesity, fatty liver and waist-to-hip ratio, although ablation of *Lyplal1* in mice does not lead to any significant abnormality in phenotype or metabolic physiology [26]. *LYPLAL1* is also one of the top 10 genes implicated in insulin resistance from two GWASs [27,28]. Knock out of *LYPLAL1* has recently been shown to cause reduced insulin-induced AKT2 phosphorylation and glucose uptake in human adipocytes [29].

The product of *BRSK2*, SAD-A kinase, is involved in the regulation of pancreatic islet β-cell size and mass and insulin secretion in response to glucose levels [30]. However there does not seem to be any other evidence that it might have an influence on BMI.

NT5C1A codes for a 5′-nucleotidase which catalyses the hydrolysis of AMP and silencing it in mouse tibialis anterior muscle results in reduced protein content and increased glucose uptake [31].

The product of *NSDHL* is involved in cholesterol biosynthesis and different allelic variants in it are known to cause the X-linked disorders CK syndrome, characterised by intellectual disability and an asthenic build, and CHILD syndrome, characterised by hemidysplasia, erythroderma and limb defects, which is typically lethal in males [32,33]. A non-synonymous polymorphism in *Nsdhl* has been reported to be associated with reduced HDL cholesterol levels in mice [34].

Among other functions, *RAB35* may regulate the insulin-stimulated translocation of glucose transporter SLC2A4/GLUT4 in adipocytes and SNPs near *RAB35* are associated with backfat thickness at 100 kg in pigs [35,36].

The results from the gene set analyses, are summarised in Figure 3A. After the top 20 scoring sets are removed the MLPs have a gradient of 1.26 with an intercept at 0.02, indicating quite marked inflation. It seemed possible that this might represent the cumulative effect of the modest inflation noted in the individual SLPs and so the procedure of combining them according to Fisher's method was repeated after first dividing them by the inflation factor of 1.08. This produced the results displayed in Figure 3B. Here the gradient is 0.70 with an intercept at -0.03, indicating some deflation of the statistic. Applying a Bonferroni correction for the number of sets tested would mean that the value of the MLP would need to exceed log10(1,454*20)=4.5 and this was achieved by only one set, PHOSPHOINOSITIDE_BINDING with MLP=4.71. This set includes 19 genes and the result is driven by the fact that it contains both *PIGK* (SLP=4.81) and *RAB35* (SLP=-4.92). Only two other genes in the set are individually significant at p<0.05, *CYTH3* (SLP=-1.55) and *ZFYVE16* (SLP=-1.38). Although all these genes fall into the category of phosphoinositide binding they do not seem to be involved in any shared metabolic processes and, apart from *RAB35,* their functions are not such that one would expect them to have a direct effect on BMI. There were 8 gene sets with MLP>2 (equivalent to p<0.01) and these are listed in Table 3. None look particularly plausible as candidates to be involved in directly influencing BMI.

Discussion

The results demonstrate that weighted burden tests are very sensitive to artefactual false positive results arising from population stratification even when, as here, only a small proportion of subjects report a minority ethnicity. Perhaps surprisingly, this problem is almost completely resolved by inclusion of population principle components as covariates. In fact, population stratification is intrinsically less challenging for weighted burden approaches than for variant-wise association analyses because artefacts arise from differences in the distribution of allele frequencies rather than from the differences in frequencies of individual alleles. Hence, inclusion of the principal components as covariates seems to be quite effective. By contrast, principal components are well-recognised not to fully capture population structure in the context of variant-wise tests of association and polygenic risk scores [37]. It may be worth noting that in the current dataset some categories of self-reported ethnicity included very small numbers, meaning that it would not be possible to use these categories as covariates. Rather, the principal components seem to adequately capture ethnicity and other sources of stratification.

In terms of the findings from the example analysis, only one gene (just) reaches conventional criteria for genome-wide significance and from what is known of its function it does not seem likely that this represents a real biological effect. On the other hand, there are a few highly ranked, though not formally statistically significant, genes where it is quite plausible that rare, functional variants might be exerting an influence on BMI. Perhaps the two most notable examples would be *LYPLAL1*, for which it is well established that common variants are associated with obesity and related traits, and *NSDHL*, for which it is known that some rare allelic variants can produce a syndrome with asthenia as part of the phenotype. These findings could be explored in other datasets and it is reasonable to expect that as sequence data becomes available for additional UK Biobank subjects it will be possible to distinguish true positive results. The analysis of gene sets has on this occasion failed to yield any further insights or even any suggestive findings. It is possible that such analyses might be more successful when applied to other phenotypes or with gene classifications which better captured biological function.

The weighting schemes for variant annotation and for allele frequency are chosen *a priori* according to some reasonable working assumptions. However, given that we do not know which genes and variants are actually influencing the phenotype in this sample, it is impossible to say how appropriate are these schemes and one might hope that they could be improved upon as more data becomes available. We note that the introduction of the new weighting scheme based on allele frequency has only minor effects on the results obtained for most genes but it does at least allow researchers to accord more weight to extremely rare variants if they feel that this is a model which they wish to implement.

To conclude, the approach presented here seems to show promise as being statistically fairly well-behaved and, importantly, as allowing heterogeneous datasets to be analysed without having to discard data from ethnic minority subjects.

Software availability

SCOREASSOC and GENEVARASSOC along with related scripts and documentation can be downloaded from https://github.com/davenomiddlenamecurtis?tab=repositories.

Data availability

The raw data is available on application to UK Biobank. The summary gene-wise results for all analyses are included in the supplementary file *UKBB.BMI.allSLPs.20200611.xlsx*. Detailed results with variant counts cannot be made available because they might be used for subject identification.

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.

Statement of ethics

UK Biobank had obtained ethics approval from the Research Ethics Committee (REC; approval number: 11/NW/0382) and written informed consent from all participants. Analysis of the data was approved by the UCL Research Ethics Committee (approval number 11527/001).

Conflict of Interest Statement

The author declares he has no conflict of interest.

Funding Sources

This work did not receive any external funding but was carried out in part using resources provided by BBSRC equipment grant BB/R01356X/1.

Author Contributions

DC conceived and executed the project and wrote the manuscript.

References

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

[2] Curtis D. Pathway analysis of whole exome sequence data provides further support for the involvement of histone modification in the aetiology of schizophrenia. Psychiatr Genet 2016;26:223–7. https://doi.org/10.1097/YPG.0000000000000132.

[3] Zhao Z, Bi W, Zhou W, VandeHaar P, Fritsche LG, Lee S. UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. Am J Hum Genet 2020;106:3–12. https://doi.org/10.1016/j.ajhg.2019.11.012.

[4] Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun 2020;11:1–10. https://doi.org/10.1038/s41467-020-14288-y.

[5] Curtis D. A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. Eur J Hum Genet 2019;27:114–24. https://doi.org/10.1038/s41431-018-0272-6.

[6] Curtis D, Coelewij L, Liu S-H, Humphrey J, Mott R. Weighted Burden Analysis of Exome-Sequenced Case-Control Sample Implicates Synaptic Genes in Schizophrenia Aetiology. Behav Genet 2018;43:198–208. https://doi.org/10.1007/s10519-018-9893-3.

[7] Curtis D, Bakaya K, Sharma L, Bandyopadhay S. 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 2019;84:291–302. https://doi.org/10.1101/596007.

[8] Hout CV Van, Tachmazidou I, Backman JD, Hoffman JX, Ye B, Pandey AK, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. BioRxiv 2019:572347. https://doi.org/10.1101/572347.

[9] Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci 2016;19:1433–41. https://doi.org/10.1038/nn.4402.

[10] Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol 2017;186:1026–34. https://doi.org/10.1093/aje/kwx246.

[11] Zhou H, Sealock JM, Sanchez-Roige S, Clarke T-K, Levey DF, Cheng Z, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. Nat Neurosci 2020. https://doi.org/10.1038/s41593-020-

0643-5.

[12] Xu Y, Yang X-L, Yang X-L, Ren Y-R, Zhuang X-Y, Zhang L, et al. Functional Annotations of Single-Nucleotide Polymorphism (SNP)-Based and Gene-Based Genome-Wide Association Studies Show Genes Affecting Keratitis Susceptibility. Med Sci Monit 2020;26. https://doi.org/10.12659/msm.922710.

[13] Curtis D, Balloux F. Editorial: Topical ethical issues in the publication of human genetics research. Ann Hum Genet 2020. https://doi.org/10.1111/ahg.12382.

[14] Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat Genet 2017;49:568–78. https://doi.org/10.1038/ng.3809.

[15] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4.

[16] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;7 Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76.

[17] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81. https://doi.org/10.1038/nprot.2009.86.

[18] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. https://doi.org/10.1086/519795.

[19] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8.

[20] Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–52. https://doi.org/10.1038/nature08185.

[21] R Core Team. R: A language and environment for statistical computing. Vienna, Austria., Austria.: R Foundation for Statistical Computing; 2014.

[22] Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.

[23] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.

[24] Bian L, Hanson RL, Muller YL, Ma L, Kobes S, Knowler WC, et al. Variants in ACAD10 are associated with type 2 diabetes, insulin resistance and lipid oxidation in Pima Indians. Diabetologia 2010;53:1349–53. https://doi.org/10.1007/s00125-010-1695-y.

[25] Bloom K, Mohsen AW, Karunanidhi A, El Demellawy D, Reyes-Múgica M, Wang Y, et al. Investigating the link of ACAD10 deficiency to type 2 diabetes mellitus. J Inherit Metab Dis 2018;41:49–57. https://doi.org/10.1007/s10545-017-0013-y.

[26] Watson RA, Gates AS, Wynn EH, Calvert FE, Girousse A, Lelliott CJ, et al. Lyplal1 is dispensable for normal fat deposition in mice. DMM Dis Model Mech 2017;10:1481–8. https://doi.org/10.1242/dmm.031864.

[27] Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 2012;44:991–1005. https://doi.org/10.1038/ng.2385.

[28] Lotta LA, Gulati P, Day FR, Payne F, Ongen H, van de Bunt M, et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. Nat Genet 2017;49:17–26. https://doi.org/10.1038/ng.3714.

[29] Chen Z, Yu H, Shi X, Warren CR, Lotta LA, Friesen M, et al. Functional Screening of Candidate Causal Genes for Insulin Resistance in Human Preadipocytes and Adipocytes. Circ Res 2020;126:330–46. https://doi.org/10.1161/CIRCRESAHA.119.315246.

[30] Nie J, Han X, Shi Y. SAD-A and AMPK kinases: The “yin and yang” regulators of mTORC1 signaling in pancreatic β cells. Cell Cycle 2013;12:3366–9. https://doi.org/10.4161/cc.26496.

[31] Kulkarni SS, Karlsson HKR, Szekeres F, Chibalin A V., Krook A, Zierath JR. Suppression of 5’-nucleotidase enzymes promotes AMP-activated protein kinase (AMPK) phosphorylation and metabolism in human and mouse skeletal muscle. J Biol Chem 2011;286:34567–74. https://doi.org/10.1074/jbc.M111.268292.

[32] McLarren KW, Severson TM, Du Souich C, Stockton DW, Kratz LE, Cunningham D, et al. Hypomorphic temperature-sensitive alleles of NSDHL cause CK syndrome. Am J Hum Genet 2010;87:905–14. https://doi.org/10.1016/j.ajhg.2010.11.004.

[33] Ramphul K, Kota V, Mejias SG. Child Syndrome 2019.

[34] Bautz DJ, Broman KW, Threadgill DW. Identification of a novel polymorphism in X-linked sterol-4-alpha-carboxylate 3-dehydrogenase (Nsdhl) associated with reduced high-density lipoprotein cholesterol levels in i/LnJ mice. G3 Genes, Genomes, Genet 2013;3:1819–25. https://doi.org/10.1534/g3.113.007567.

[35] Chen D, Wu P, Yang Q, Wang K, Zhou J, Yang X, et al. Genome-wide association study for backfat thickness at 100 kg and loin muscle thickness in domestic pigs based on genotyping by sequencing. Physiol Genomics 2019;51:261–6. https://doi.org/10.1152/physiolgenomics.00008.2019.

[36] Davey JR, Humphrey SJ, Junutula JR, Mishra AK, Lambright DG, James DE, et al. TBC1D13 is a RAB35 Specific GAP that Plays an Important Role in GLUT4 Trafficking in Adipocytes. Traffic 2012;13:1429–41. https://doi.org/10.1111/j.1600-0854.2012.01397.x.

[37] Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? Hum Genet 2020;139:23–41. https://doi.org/10.1007/s00439-019-02014-8.

Figure 1

Plot of mean and SE(mean) of BMI against self-declared ethnicity in exome-sequenced UK Biobank subjects.

Figure 2

QQ plots of SLPs obtained for weighted burden analysis of 21,644 genes for association with BMI. 2A shows the results for regression of the weighted burden score against BMI alone and 2B shows the results when the population principal components are included as covariates. 2C shows the results with principal components included but using the old scheme for weighting variants by frequency, which does not strongly distinguish rare from very rare variants.

Figure 3 QQ plots of MLPs for 1,454 gene set analyses performed by combining SLPs obtained from the gene-wise analyses incorporating population principal components. 3A shows the results obtained for the uncorrected analyses and 3B shows the results obtained after first dividing the SLPs by an inflation factor of 1.08.