Human Heredity

Research Article

Hum Hered 2024;89:1–7
DOI: 10.1159/000537771

Received: September 13, 2023
Accepted: February 7, 2024
Published online: February 10, 2024

Investigation of Recessive Effects of Coding Variants on Common Clinical Phenotypes in Exome-Sequenced UK Biobank Participants

David Curtis

UCL Genetics Institute, University College London, London, UK

Keywords
Exome · Recessive effects · Compound heterozygote

Abstract
Introduction: Previous studies have demonstrated effects of rare coding variants on common, clinically relevant phenotypes although the additive burden of these variants makes only a small contribution to overall trait variance. Although recessive effects of individual homozygous variants have been studied, little work has been done to elucidate the impact of rare coding variants occurring together as compound heterozygotes. ***Methods:*** In this study, attempts were made to identify pairs of variants likely to be occurring as compound heterozygotes using 200,000 exome-sequenced subjects from the UK Biobank. Pairs of variants, which were seen together in the same subject more often than would be expected by chance, were excluded as it was assumed that these might be present in the same haplotype. Attention was restricted to variants with minor allele frequency ≤0.05 and to those predicted to alter amino acid sequence or prevent normal gene expression. For each gene, compound heterozygotes were assigned scores based on the rarity and predicted functional consequences of the constituent variants and the scores were used in a logistic regression analysis to test for association with hypertension, hyperlipidaemia, and type 2 diabetes. ***Results:*** No statistically significant associations were observed and the results conformed to the distribution, which would be expected under the null hypothesis. The average number of apparently compound heterozygous subjects for each gene was only 282.2. ***Conclusion:*** It seems difficult to detect an effect of compound heterozygotes on the risk of these phenotypes. Even if recessive effects from compound heterozygotes do occur, they would only affect a small number of people and overall would not make a substantial contribution to phenotypic variance. This research has been conducted using the UK Biobank Resource.

Introduction

With the availability of results from large-scale exome-sequencing projects, it has become possible to gain insights into the contribution of rare coding variants to human phenotypes [1, 2]. By treating this contribution simply as an additive burden of variants, it explains on average only 1.3% of the variance of 22 common traits and diseases [2]. If we consider the situation in more detail, we can note that there are some gene-phenotype pairs where protein-truncating variants (PTVs, consisting of stop gained, frameshift, and essential splice site variants) are very rare but have large effect sizes while some categories of nonsynonymous variant are less rare and have more modest effects on risk [3–6]. To take one concrete example, in 200,000 exome-sequenced UK Biobank subjects, fewer than 0.1% of subjects with hyperlipidaemia were observed to carry a PTV in *LDLR* and

karger@karger.com
www.karger.com/hhe

Correspondence to:
David Curtis, d.curtis@ucl.ac.uk

these variants were estimated to confer increased risk with OR >20 whereas over 1% carried a nonsynonymous variant annotated as deleterious by SIFT and for these variants the OR was estimated to be 1.7 [4]. If variants within a particular category, such as nonsynonymous, have differential effects on risk, then classing them all together and characterising their average effect size will tend to underestimate proportion of the phenotypic variance, which they explain. Given that in silico annotation methods cannot consistently predict variant effects, it is not possible to accurately characterise the effect sizes of the individual variants within a category [7]. This means that the actual contribution to trait liability made by rare coding variants might be considerably larger than the estimate of 1.3%, which is obtained if they are all lumped together and considered to have an additive burden.

A complementary scenario whereby coding variants might influence phenotypes, which would not be captured by simply assessing rare variant burden, would be through variants, which acted recessively. If a class of variant in a gene had a negligible effect when only present in one copy of the gene but a substantial effect if in both copies, then there would only be a weak association between the phenotype and the variant burden. Such variants might even be not especially rare. If we consider the example of cystic fibrosis, until recently a lethal condition, we can note that the carrier rate in Europeans of around 1 in 30 implies a cumulative risk allele frequency of greater than 0.01. It seems plausible that variants having recessive effects producing less severe phenotypes might cumulatively be even more common. We have argued elsewhere that intuitively one might expect that unrecognised recessive effects could be important risk factors for disease [8]. When considering rare variant effects in isolation, PTVs are generally expected to have large effect sizes and in general this is supported empirically, although there are also examples of particular non-synonymous variants having large effects. However, although this is clearly an over-simplification, the expectation of the effect of a PTV in biological terms is that it produces haploinsufficiency of the gene – that one copy does not function but the other remains intact and that the overall product yield might be halved. By contrast, it does not seem hard to imagine that non-synonymous variants affecting both copies of the gene could have an effect as large as, or even greater than, a single PTV affecting only one copy.

In analyses of individual variants in exome-sequenced participants in UK Biobank, it was possible to detect significant recessively acting variant-trait associations for 1,088 binary traits and 10,770 quantitative traits, of which 21% and 12%, respectively, were not detectable using a dominant model [1]. That is, for these variants there was significant association of the trait with the homozygous genotype. However, one might expect that in an outbred population the frequency of homozygotes for a single variant would be low relative to the number of compound heterozygotes formed by combinations of two variants present in *trans* and indeed it has been shown that compound heterozygotes can be important causes of disease [9]. However, detecting compound heterozygotes is by no means straightforward. In the same study as the one showing associations with single variant homozygotes, a gene-wise recessive model was defined as one in which two qualifying variants were observed in the same gene in the same subject, without any indication as to whether they occurred in *cis* or *trans* [1]. No significant gene-wise recessive associations were reported. Although it may seem superficially reasonable to assume that when two rare variants are observed in the same subject, then they are likely to be on different chromosomes; in fact, this cannot be relied upon. Earlier attempts to detect recessive effects on schizophrenia risk foundered when it became apparent that rare variants occurred together in the same haplotype surprisingly frequently and that, without information about phase, simple approaches to treating pairs of rare variants observed in the same subject as compound heterozygotes were unlikely to be successful [8].

The analyses reported here set out to investigate whether recessive effects on common, clinically relevant phenotypes could be detected by observing increased risk in subjects who are apparently compound heterozygotes for qualifying variants. The hope was that there might be a relatively large number of such subjects and that compound heterozygotes of coding variants might make a substantial contribution to risk of developing common phenotypes.

One approach to determining whether variants within a gene occur together in *cis* or *trans* is to first carry out phasing to estimate the underlying haplotypes forming the observed genotypes. Phasing algorithms can be computationally intensive and earlier software, developed primarily for phasing common SNPs, proved inaccurate for phasing rare variants within a gene although newer methods offer improved performance [8, 10]. However, for the present purpose of identifying probably compound heterozygotes a simpler procedure may be adequate if we restrict attention to variants with relatively low MAF and if we assume that, because variants are within the same gene, it will be rare for recombination to occur between them. In this situation, denoting a pair of variants as Aa and Bb, if within the population there are no haplotypes bearing both minor alleles a and b then we will expect to sometimes observe

joint genotypes AABb and AaBB but only rarely the compound heterozygote AaBb. On the other hand, if the b allele was formed on a chromosome bearing the a allele then we will rarely see the b allele in the absence of the a allele and we expect to see joint genotypes AaBB and AaBb, or for the reverse scenario we expect so see AABb and AaBb. To operationalise this, we can say that if there is a haplotype within the population bearing both the a and b allele then we will expect to see the AaBb joint genotype much more frequently than would be predicted if the a and be allele were segregating independently. We can also note that if any individual is homozygous for one minor allele and heterozygous for the other then both alleles must be present together in one haplotype. This simple approach can be used to identify pairs of alleles, which sometimes occur together in *cis*, and such pairs can be discounted when attempting to identify subjects who are likely compound heterozygotes.

This process was used to identify participants in the UK Biobank who were probable compound heterozygotes for coding variants within a gene and then scores were assigned to these compound heterozygotes based on the constituent variants. Weighted burden analysis was applied to these scores to assess whether compound heterozygote status for a gene was associated with risk for a number of clinically relevant phenotypes.

Methods

Exome variants were identified and annotated as previously described [11]. 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 [12]. 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 written 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). All variants were annotated using the standard software packages VEP, PolyPhen, and SIFT [13–15]. 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* [16, 17].

As described previously, the SCOREASSOC program was used to assign a score to each variant in each gene, such that variants, which were rarer and/or predicted to have more severe functional effects, were given higher scores [11]. Attention was restricted to variants with minor allele frequency (MAF) ≤0.05. A rarity score for each variant was assigned as a parabolic function of MAF, such that variants with MAF approaching 0 have a score of 10 while variants with MAF of 0.05 have a score of 1. Variants were also scored according to their functional annotation using the GENEVARASSOC program, which was used to generate input files for the analysis by SCOREASSOC [18, 19]. Attention was restricted to protein-altering variants, likely to either alter the amino acid sequence of a protein or prevent normal gene expression. Variants predicted to cause complete loss of function were given a score of 100 while nonsynonymous, in-frame indels and start or stop loss variants were given a score of 10. For nonsynonymous variants, 20 was added to the score if they were annotated as deleterious by SIFT, 5 if they were annotated as possibly damaging by PolyPhen, and 10 if they annotated as probably damaging by PolyPhen. Other variants were ignored. The overall score for each variant consisted of the product of the rarity score and the functional score, meaning that the score could range from 1,000 for a very rare loss of function variant down to 10 for a common nonsynonymous variant with no additional annotations.

Variants were excluded if there were more than 10% of genotypes missing or if the heterozygote count was smaller than both homozygote counts. If a subject was not genotyped for a variant, then they were assumed to be homozygous for the reference allele. For variants on the X chromosome, only female subjects were considered.

The method used aimed to focus only on compound heterozygotes and to ignore homozygotes. There were a number of reasons for this. One is that a homozygous call might be a reflection of increased autozygosity for a subject across their genome rather than any specific effect of the variant or gene under consideration. Another is that a homozygous call might be more likely to be the result of a genotyping error than a heterozygote call. A third is the expectation that, in an outbred population, if a number of different variants could in combination have a recessive effect then one might expect that the number of homozygotes would be low relative to the number of compound heterozygotes.

For each gene in turn, the next step was to identify subjects who carried two different valid variants in whom the variants seemed likely to occur as a compound heterozygote, i.e., were not both present on the same chromosome. In order to do this, it was first necessary to exclude pairs of variants, which were seen to occur together more often than would be expected by chance. Since enumerating counts for all pairs is $O(n^2)$, the list of potentially valid variants was first restricted to a maximum of 200 variants for each gene. If there were initially more than 200 potentially valid variants, then the aim would be to restrict attention to those which would be most likely to provide information and these would be variants which occurred more frequently and those which had higher scores. In order to achieve this, variants were ranked in order of "importance," which was taken to be for each variant the product of its score with its frequency, and then any variants with a rank lower than 200 were discarded.

Out of all possible pairs of valid variants, a pair was considered to be invalid if any subject was observed who carried one variant and was homozygous for the other since this would indicate that both variants at least sometimes occurred in the same haplotype. Additionally, a pair was considered invalid if both were observed together in the same subject in at least half as many subjects as carried one member of the pair, again because this would suggest that they sometimes occurred in the same haplotype.

Once a list of valid pairs of variants was obtained, any subject carrying both variants of a valid pair, presumed to be a compound heterozygote, was allocated a score consisting of the sum of the scores of each variant. If a subject carried more than one valid pair, then the highest scoring pair was used. Thus, each subject was assigned a score indicating whether they appeared to carry different

variants in both copies of the gene in question and if so reflecting the rarity and predicted functional impacts of the two variants.

This procedure was coded and implemented in a modified form of the SCOREASSOC program [11]. Three different common and clinically relevant phenotypes were used, consisting of hypertension, hyperlipidaemia, and type 2 diabetes. Using the same processes as previously described, for each phenotype cases were defined using a combination of recorded diagnosis and prescribed medication, with the other subjects used as controls [3–5].

For each gene and each phenotype, multiple logistic regression analysis was carried out with caseness as the dependent variable and including the first 20 population principal components and sex as covariates. A likelihood ratio test was performed comparing the likelihoods of the models, which did and did not additionally include the compound heterozygote score as a predictor of caseness. Under the null hypothesis, the distribution of the natural log of this likelihood ratio should follow a χ^2 distribution with 1 degree of freedom. For convenience, the statistical significance is expressed as a signed log p value (SLP), which is the log base 10 of the p value given a positive sign if the score is positively correlated with caseness. This means strongly positive or negative values for the SLP indicate results, which are statistically significant, while the sign indicates whether impaired functioning of the gene is positively or negatively associated with caseness.

Although the primary analyses utilised the compound heterozygote score as described above as the predictive variable, two subsidiary analyses were also performed. For the first of these, the rank of the score across subjects was used instead of the absolute value and for the second of these a simple indicator variable for having a non-zero score or not (being a compound heterozygote or not) was used. Although all genes were analysed, the primary analyses were restricted to genes in which at least 20 subjects were classified as compound heterozygotes.

Data manipulation and statistical analyses were performed using GENEVARASSOC, SCOREASSOC, and R [20]. Software and scripts used to carry out the analyses are available at https://github.com/davenomiddlenamecurtis.

Results

200,627 subjects were analysed, of whom 66,123 were classified as cases with hypertension, 44,054 cases with hyperlipidaemia, and 13,938 as cases with type 2 diabetes. 19,763 genes were analysed, of which 15,665 had at least one subject qualifying as a compound heterozygote. Of these, the mean number of variants per gene was 296.4 (SD = 303.3) with a mean total minor allele count for these variants of 12,401.2 (SD = 20,829.7). The mean number of subjects qualifying as compound heterozygotes per gene was 282.2 (SD = 1,351.5) and 9,677 genes had 20 subjects or more qualifying as compound heterozygotes. These were the genes, which were used for the primary analyses. Given that three phenotypes were analysed, the critical threshold for the absolute value of the SLP to declare a result as formally statistically significant is –log10(0.05/[3*9,677]) = 5.76.

Logistic regression analysis for the three phenotypes using the compound heterozygote scores was carried out as described above for each gene with 20 or more compound heterozygotes. The gene-wise results for each phenotype conformed closely to the distribution under the null hypothesis. This is illustrated by the QQ plots of the SLPs, which are displayed in Figure 1. None of the results are statistically significant after correction for multiple testing and only 3 results are significant at $p < 10^{-4}$, these being *XRCC4* (SLP = 4.95) and *DGCR6L* (SLP = 4.66) for type 2 diabetes and *CASP4* (SLP = –4.04) for hyperlipidaemia. The detailed results for these genes are shown in Table 1. None of these genes seem to be obvious candidates to affect the associated trait from a biological point of view. Examination of all the results obtained using rank scores or simply dichotomised as being a compound heterozygote or not showed that these results also complied closely with the null hypothesis expectations. Likewise, the analysis of genes with fewer than 20 compound heterozygotes did not produce any results with higher levels of significance. The full results for all genes using the raw scores, the rank scores, or dichotomised status as compound heterozygote are provided in online supplementary Table 1 (for all online suppl. material, see https://doi.org/10.1159/000537771).

Discussion

Overall, the results seem entirely negative, with no suggestion that the method detects genes in which recessive effects of coding variants have important effects on risk of the phenotypes studied. These results stand in marked contrast to those obtained from weighted burden analysis assuming additive effects of variants impacting single copies of each gene, with exome-wide significant results being obtained for all three of these phenotypes using the same dataset [3–5]. Of course, it remains plausible that recessively acting variants in compound heterozygotes do sometimes have effects on risk of these phenotypes and the approach used, applied to unphased data, is not expected to infallibly detect every compound heterozygote. Likewise, methods used to characterise the likely functional effect of variants are not completely accurate [7]. Thus, it is possible that slightly different results would have been obtained if different scoring methods and/or predictors of impact of nonsynonymous variants had been used. However, the fact that neither the rank scores nor dichotomised carrier status produced

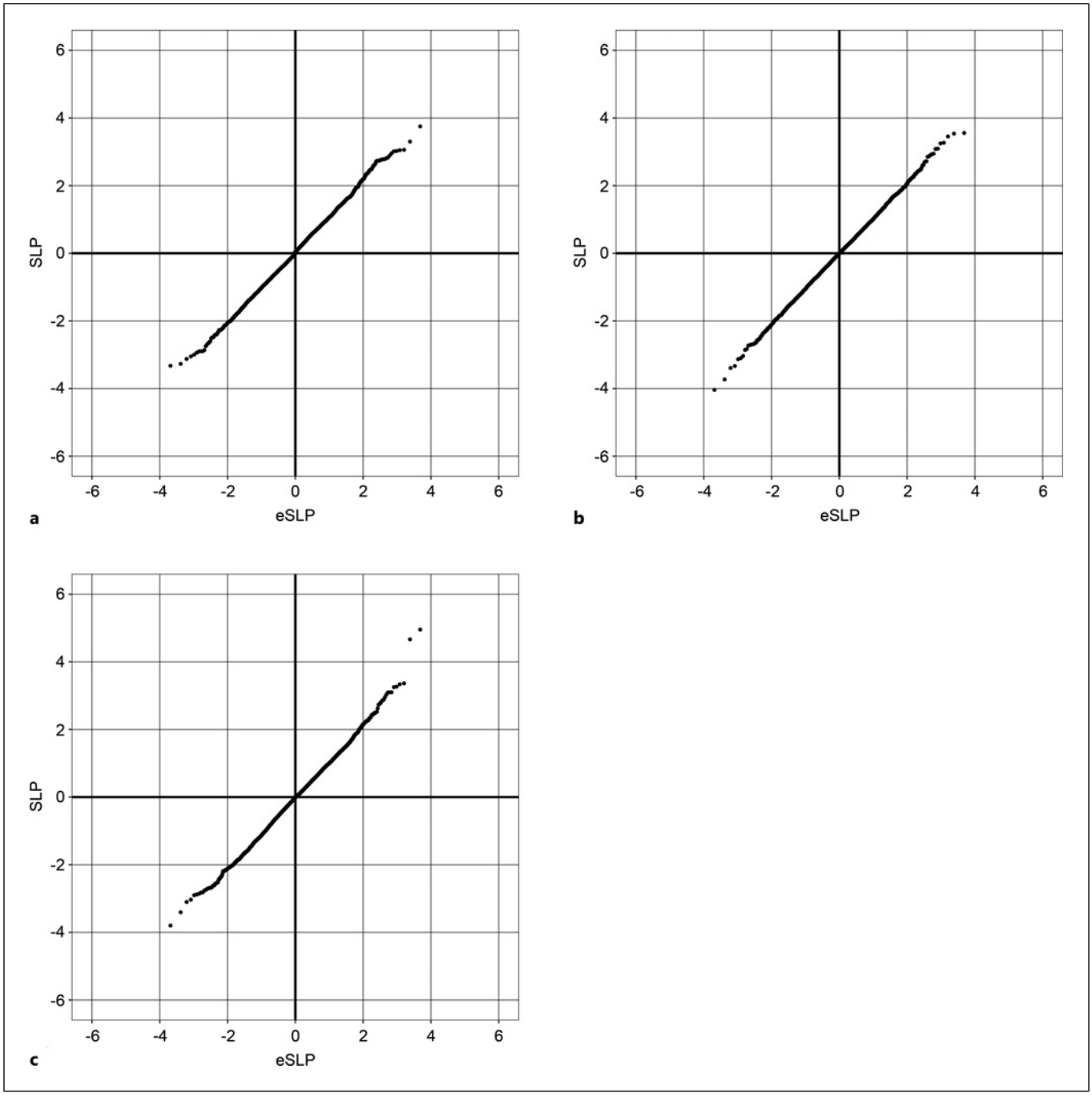

Fig. 1. QQ plots obtained from logistic regression analysis of compound heterozygote score showing observed against expected SLP for each gene. **a** Results for hypertension. **b** Results for hyperlipidaemia. **c** Results for type 2 diabetes.

any signal would seem to suggest that it is unlikely that different scoring methods would have a very major effect. Overall, it seems clear that recessively acting coding variants do not represent "low hanging fruit" in terms of the potential for the discovery of novel associations.

Although the sample size might seem large and for each gene the minor allele counts of protein-altering

Table 1. Detailed results for three genes significant at $p < 10^{-4}$, showing SLPs obtained using score or dichotomised compound heterozygote status along with odds ratio associated with compound heterozygote status

Gene symbol	Gene name	Phenotype	Number of compound heterozygotes	SLP for score	SLP for compound heterozygote status	OR (95% CI) for compound heterozygote status
XRCC4	X-ray repair cross-complementing 4	Type 2 diabetes	275	4.95	2.80	1.9 (1.3–2.7)
DGCR6L	DiGeorge syndrome critical region gene 6 like	Type 2 diabetes	21	4.66	3.51	7.1 (2.8–18.3)
CASP4	Caspase 4	Hyperlipidaemia	116	−4.04	−2.79	0.44 (0.24–0.78)

variants with MAF ≤0.05 tends to run into thousands, in fact the number of subjects apparently carrying two such variants on different copies of the same gene averages only a few hundred. This is broadly in line with the expectation given the overall frequency of qualifying variants. With an average about 12,000 qualifying variants in a gene observed in about 200,000 participants, we can say that the probability for one copy of the gene to carry a variant is about 0.03 and for both copies to carry a variant, assuming independence, the probability would be 0.009. This means one would expect only about one in a thousand subjects to have variants in both copies of the gene. And many of these variants, especially those seen most commonly, might have little or no effect on gene function. This expected rarity of compound heterozygotes has two implications. One is that compound heterozygotes may have effects on risk but that the sample is underpowered to detect them. The other is that even if such effects do occur then, for common phenotypes, they could only be relevant to a small fraction of cases. These results suggest that recessively acting effects of coding variants do not make a substantial contribution overall to the variance of the liability to these phenotypes.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 51119. The author wishes to thank the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. The author wishes to thank the participants who volunteered for the UK Biobank project. This work uses data provided by patients and collected by NHS England as part of their care and support. This research also used data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grants MC_PC_20029 and MC_PC_20058).

Statement of Ethics

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 written 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).

Conflict of Interest Statement

The author declares he has no conflict of interest.

Funding Sources

No external funding was received for this study.

Author Contributions

D.C. conceived the study, carried out the analyses, and wrote up the manuscript.

Data Availability Statement

The raw data are available on application to UK Biobank. Requests for access can be made at this site: https://ams.ukbiobank.ac.uk/ams/. Detailed results with variant counts cannot be made available because they might be used for subject identification. Scripts and relevant derived variables will be deposited in UK Biobank. Software and scripts used to carry out the analyses are available at https://github.com/davenomiddlenamecurtis. Further enquiries can be directed to the corresponding author.

References

1 Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021;597:527–32.
2 Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, et al. Polygenic architecture of rare coding variation across 394,783 exomes. Nature. 2023;614(7948):492–99.
3 Curtis D. Analysis of 200,000 exome-sequenced UK Biobank subjects implicates genes involved in increased and decreased risk of hypertension. Pulse. 2021;9(1–2):17–29.
4 Curtis D. Analysis of 200 000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. J Med Genet. 2021;59(6):597–604.
5 Curtis D. Analysis of rare coding variants in 200,000 exome-sequenced subjects reveals novel genetic risk factors for type 2 diabetes. Diabetes Metab Res Rev. 2022;38(1):e3482.
6 Curtis D. Analysis of rare variants in 470,000 exome-sequenced UK Biobank participants implicates novel genes affecting risk of hypertension. Pulse. 2023;11(1):9–16.
7 Curtis D. Exploration of weighting schemes based on allele frequency and annotation for weighted burden association analysis of complex phenotypes. Gene. 2022;809:146039.
8 Curtis D Investigation of recessive effects in schizophrenia using next-generation exome sequence data. Ann Hum Genet. 2015;79(5):313–19.
9 Miller DB, Piccolo SR. A survey of compound heterozygous variants in pediatric cancers and structural birth defects. Front Genet. 2021;12:640242.
10 Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat Genet. 2023;55(7):1243–9.
11 Curtis D. Multiple linear regression allows weighted burden analysis of rare coding variants in an ethnically heterogeneous population. Hum Hered. 2020;85(1):1–10.
12 Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet. 2021;53(7):942–8.
13 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
14 Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20.
15 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(7):1073–81.
16 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(1):7.
17 Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, et al. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am J Hum Genet. 2016;98(3):456–72.
18 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.
19 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(5):223–7.
20 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.r-project.org.