Analysis of 200,000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia

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

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

Correspondence:

David Curtis *d.curtis@ucl.ac.uk*

Abstract

A few genes have previously been identified in which very rare variants can have major effects on lipid levels. Weighted burden analysis of rare variants was applied to exome sequenced UK Biobank subjects with hyperlipidaemia as the phenotype, of whom 44,050 were designated cases and 156,578 controls, with the strength of association characterised by the signed log 10 p value (SLP). With principal components included as covariates there was a tendency for genes on the X chromosome to produce strongly negative SLPs, and this was found to be due to the fact that rare X chromosome variants were identified less frequently in males than females. The test performed well when both principal components and sex were included as covariates and strongly implicated *LDLR* (SLP = 50.08) and *PCSK9* (SLP = -10.42) while also highlighting other genes previously found to be associated with lipid levels. Variants classified by SIFT as deleterious have on average a two-fold effect and their cumulative frequency is such that they are present in approximately 1.5% of the population. These analyses shed further light on the way that genetic variation contributes to risk of hyperlipidaemia and in particular that there are very many protein-altering variants which have on average moderate effects and whose effects can be detected when large samples of exome-sequenced subjects are available. This research has been conducted using the UK Biobank Resource.

Keywords

Hyperlipidaemia; biobank; exome; *LDLR*; *PCSK9*; *ANGPLT3*; *ANGPLT4*; *APOC3*; *ABCG5*; *ABCD1*; *ABCA1*; *NPC1L1*.

Introduction

We recently reported the results of analysis of 50,000 exome-sequenced UK Biobank subjects aiming to identify rare variant effects in genes influencing susceptibility to hyperlipidaemia and also briefly reviewed what was known to date about the genetic contributors to this phenotype [1]. The potential advantage of studying rare variants is that they have more profound, readily interpretable impacts on biology than common variants, whose effect sizes tend to be constrained by selection pressures. Rare variants with a large dominant effect in *LDLR*, *APOB* and *PCSK9* cause 40% of cases of familial hyperlipidaemia and there are also common variants which exert small effects on hyperlipidaemia risk [2–5]. Although for most genes impaired function increases risk, the *PCSK9* variants which cause familial hyperlipidaemia produce a gain of function whereas loss of function variants cause hypobetalipoproteinemia and PCSK9 inhibitors are used as treatments to lower cholesterol levels [6].

The previous analysis of 50,000 UK Biobank identified one gene, *HUWE1*, which met criteria for statistical significance after correction for multiple testing and in which there was an excess of rare and/or damaging variants in controls, suggesting that impaired functioning of this gene was protective against hyperlipidaemia. A number of other genes which were individually significant with uncorrected $p<0.001$, were arguably of potential interest, including *LDLR*, and, in an analysis of sets of genes, the GO gene set GENERATION OF PRECURSOR METABOLITES AND ENERGY was statistically significant. The whole UK Biobank sample consists of 500,000 subjects and a new release of data means that there is now exome sequence data available for 200,000 of them [7]. We report here the results of analysis of this larger dataset, which includes the original 50,000, with the expectation that it would provide greater power to detect genes with a real biological effect. The larger sample size would also allow more sophisticated analyses which could throw light on the differential effects of different types of variant and could produce more refined estimates of the contributions to risk in the general population.

Large samples of exome sequenced subjects have only become available relatively recently and controversy remains about the optimal methods of analysis. Sequencing reveals very large numbers of genetic variants, many of which will have no biological effect and/or will be extremely rare, occurring in only a handful of subjects or just as singletons. The rarity of individual variants means that they need to be grouped together in a burden analysis and it is common practice to combine all variants which are predicted to completely disrupt the working of a gene, comprising: variants which introduce a stop codon; small insertions and deletions which are not a multiple of three bases and hence disrupt the amino acid code, termed frameshift variants; variants changing essential splice site sequences at intron-exon boundaries, disrupting normal splicing of exons. These three types of variant are predicted to all have a broadly similar effect no matter where they occur in the gene, consisting of a complete failure of the gene to produce normal product, and they may be referred to as loss of function (LOF) variants. It may then become possible to implicate a gene in the pathogenesis of phenotype by observing a general excess of LOF variants in that gene among cases relative to controls [8]. However it is certainly the case that other kinds of variant can also cause disease. A variant which changes a codon so that it codes for a different amino acid, termed a non-synonymous variant, may alter the structure or function of the protein product in a way which dramatically affects risk but alternatively a protein altering variant may have no effect at all. The impact of a non-synonymous variant will depend crucially on the nature of the amino acid change and its position in the protein and it remains a challenging task to predict the biological effect although commonly used software such as PolyPhen and SIFT attempt this [9, 10]. PolyPhen designates some variants as “possibly damaging” or “probably damaging” and SIFT designates some variants as “deleterious” but different prediction programs do not always agree with each other. Nonsynonymous variants are much more frequent than LOF variants and so it would be desirable to incorporate them into burden analyses but there is a risk that doing so may simply introduce additional noise. Identifying which specific variants are most likely to have biological effects could increase power to implicate risk genes but remains a challenging task. Even variants which do not change amino acid sequence, including synonymous and intronic variants, can through various mechanisms occasionally have effects on risk and so could potentially be included.

The approach we have taken to address these issues is to carry out weighted burden analyses, in which variants judged *a priori* to be most likely to have important effects are accorded higher weights. Since selection pressures mean that common variants are unlikely to have large effects, variants are also weighted according to rarity and the detailed scheme for doing this is described in the Methods section. However a weakness of this approach to date has been that there has been little empirical evidence to inform the exact weighting scheme which would be optimal. An advantage of the large UK Biobank dataset is that it allows some exploration of the relative average effect sizes of different categories of variant and this was carried out using multivariate analyses of

variant categories in addition to standard weighted burden analyses of genes and gene sets. These investigations were applied to the previously used hyperlipidaemia phenotype, defined as subjects with a diagnosis of hyperlipidaemia and/or taking cholesterol-lowering medication.

Methods

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 [7]. 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). All variants were annotated using the standard software packages VEP, PolyPhen and SIFT [9–11]. 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* [12, 13].

The hyperlipidaemia phenotype was determined in the same way as previously from four sources in the dataset: self-reported high cholesterol; reporting taking cholesterol lowering medication; reporting taking a named statin; having an ICD10 diagnosis for hyperlipidaemia in hospital records or as a cause of death [1]. Subjects in any of these categories were deemed to be cases with hyperlipidaemia while all other subjects were taken to be controls.

The method of analysis was the same as used previously on the smaller sample. The SCOREASSOC program was used to carry out a weighted burden analysis to test whether, in each gene, sequence variants which were rarer and/or predicted to have more severe functional effects occurred more commonly in cases than controls. 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 of 10 [14]. Variants were also weighted according to their functional annotation using the GENEVARASSOC program, which was used to generate input files for weighted burden analysis by SCOREASSOC [15, 16]. A maximum weight of 40 was assigned to variants predicted to cause complete LOF of the gene, namely stop-gained, frameshift and essential splice site variants. Other types of variant were assigned intermediate weights intended to provide an approximate measure of their likely importance, for example, a weight of 5 was assigned for a synonymous variant, 10 for a non-synonymous variant and 15 for inframe insertions and deletions. Additionally, 10 was added to the weight 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 an overall weight of 30. In order to allow exploration of the effects of different types of variant on disease risk the variants were also grouped into broader categories to be used in multivariate analyses as described below. The full set of weights and categories is displayed in Table 1.

As described previously, the weight due to MAF and the weight due to functional annotation were multiplied together to provide an overall weight for each variant. 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 in the controls. If a subject was not genotyped for a variant then they were assigned the subject-wise average score for that variant. 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. For variants on the X chromosome, hemizygous males were treated as homozygotes.

For each gene, a ridge regression analysis was carried out with lamda=1 to test whether the gene-wise variant burden score was associated with the hyperlipidaemia phenotype. To do this, SCOREASSOC first calculates the likelihood for the phenotypes as predicted by the first 20 population principal components and then calculates the likelihood using a model which additionally incorporates the gene-wise burden scores. It then carries out a likelihood ratio test assuming that twice the natural log of the likelihood ratio follows a chi-squared distribution with one degree of freedom to produce a p value. 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 is higher in cases and negative if it is higher in controls. In previous analyses it appeared that incorporating population principal components in this way satisfactorily controlled for test statistic inflation when applied to the ancestrally heterogeneous UK Biobank dataset [14]. However preliminary analyses of this new, larger dataset revealed that there was a slight tendency for more rare variants in X chromosome genes to be identified in females rather than males. Hence, sex was also included as a covariate along with the principal components and this produced a well-behaved test statistic, as detailed in the Results section.

Gene set analyses were carried out as before 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 [17]. For each set of genes, the natural logs of the gene-wise p values were summed according to Fisher's method to produce a chi-squared statistic with degrees of freedom equal to twice the number of genes in the set. The p value associated with this chi-squared statistic was expressed as a minus log10 p (MLP) as a test of association of the set with the hyperlipidaemia phenotype.

For selected genes, additional analyses were carried out to clarify the contribution of different categories of variant. To do this, each category as listed in Table 1 was assigned a weight consisting of a different power of 10 and then GENEVARASSOC and SCOREASSOC were used to obtain scores for each subject as the sum of these weights. This allowed the overall number of variants of each category possessed by a subject to be coded as a decimal number so that, for example, a score of 1000302 would indicate that the subject possessed one of one category of variant, three of another category and two of a third category. Code was written in R to read in these scores and parse them to obtain the subject-wise counts for each category of variant [18]. These were then entered into a logistic regression analysis of case-control status along with principal components and sex in order to estimate the relative contributions of different variant categories to the phenotype. The odds ratios associated with the category were estimated along with their standard errors and the Wald statistic was used to obtain a p value, except for categories in which variants occurred fewer than 50 times in which case Fisher's exact test was applied to the raw variant counts without including covariates. The associated p value was converted to an SLP, again with the sign being positive if the mean count was higher in cases than controls.

The weighted burden approach assumes that on average the effect of variants is to reduce the functioning of a gene but the variants in *PCSK9* reported to be associated with hyperlipidaemia act by causing gain of function. Therefore a list of known pathogenic and likely pathogenic variants in *PCSK9* was obtained from ClinVar (*https://www.ncbi.nlm.nih.gov/clinvar/*) in order to allow them to be analysed separately.

Results

Results of gene-wise weighted burden tests

There were 44,054 cases with a diagnosis of hyperlipidaemia and/or taking cholesterol-lowering medication and 156,578 controls. There were 22,642 genes for which there were qualifying variants and preliminary analyses showed that there was a bias towards producing strongly negative SLPs, which was confined to genes on the X chromosome. The analyses were repeated using sex as a phenotype and this confirmed that the frequency of rare, damaging variants was higher in females for genes on the X chromosome. This would occur if the genotype calling algorithm were slightly more likely to call a female as heterozygous than a male as hemizygous. Since the frequency of cases is lower in females, the overall effect is to observe an excess of rare, damaging variants in controls rather than cases for genes on the X chromosome. Therefore the analyses were repeated for hyperlipidaemia using sex as a covariate as well as the principal components. When this was done only two genes produced strongly positive or negative SLPs, *LDLR* (SLP = 50.08) and *PCSK9* (SLP = -10.42). The quantile-quantile (QQ) plot for the SLPs obtained for each of the remaining genes is shown in Figure 1. This shows that the test appears to be well-behaved and conforms fairly well with the expected distribution. Omitting the genes with the 100 highest and 100 lowest SLPs, which might be capturing a real biological effect, the gradient for positive SLPs is 1.08 with intercept at -0.019 and the gradient for negative SLPs is 1.04 with intercept at -0.013, indicating only modest inflation of the test statistic.

The role of very rare variants in both *LDLR* and *PCSK9* in the pathogenesis of familial hyperlipidaemia is already well established. However the results from the current analysis implicate a larger number of variants in these genes having a range of effects on risk of hyperlipidaemia in the population more generally. These are presented in detail below in the description of the results of the analysis of effects of different variant categories.

Given that there were 22,642 informative genes, the critical threshold for the absolute value of the SLP to declare a result as formally statistically significant is -log10(0.05/22642) = 5.66 and this was achieved by three other genes, *ANGPTL3* (SLP = -5.67), *LOC102723729* (SLP = -5.77) and *IFITM5* (SLP = -5.86). Loss of function variants in *ANGPTL3* have previously been shown to cause combined hypolipidaemia and it is the target of evinacumab, a human monoclonal antibody designed to treat hypercholesterolaemia [19, 20]. However *IFITM5* and *LOC102723729* do not seem to be biologically plausible candidates. *IFITM5* is involved in bone mineralisation and variants in it are a known cause of osteogenesis imperfecta [21]. *LOC102723729* is a poorly characterised lncRNA which may act as a tumour suppressor in non-small cell lung cancer [22]. A total of 55 genes had SLP with absolute value greater than 3 (equivalent to uncorrected $p < 0.001$), whereas one would only expect around 22642/1000 = 23 by chance so a number of these may in fact be exerting some effect on risk. These are listed in Table 2 and some appear to be of particular interest and are discussed briefly as below. The results for all genes are presented in Supplementary Table S1.

G6PC (SLP = 5.55) is of interest because mutations acting recessively cause glycogen storage disease type I (GSD1, von Gierke disease, incidence ~1/100,000), which includes hyperlipidaemia as part of the phenotype [23]. Rare homozygous variants in *ABCG5* (SLP = 3.31) can produce sitosterolaemia and are a known cause of homozygous familial hypercholesterolaemia [24]. Mutations in *ABCD1* (SLP = 3.26) cause X-linked adrenoleukodystrophy, which results in elevated levels of very long chain fatty acids in plasma and tissues [25]. Variants in *GCK* (SLP = 3.04) are known to cause maturity-onset diabetes of the young (MODY) with mild hyperglycaemia and lower triglyceride levels than other forms of type 2 diabetes [26]. For genes with negative SLPs there is a higher frequency of rare, functional variants in controls than cases. Variants in *APOC3* (SLP = -4.89) have previously shown to be protective against hyperlipidaemia risk [6]. Homozygous knockout of *PPP1R3G* (SLP = -4.25) mitigates high-fat diet induced obesity in mice [27]. The product of *NPC1L1* (SLP = -3.70) is essential for intestinal sterol absorption and is the molecular target of ezetimibe, a potent cholesterol absorption inhibitor that lowers blood cholesterol [28]. Like *ANGPTL3*, *ANGPTL4* (SLP = -3.66)

modulates the activity of lipoprotein lipase (LPL) and inactivating variants in it have previously been shown to be associated with hypolipidaemia [6].

While two ATP binding cassette transporter genes, *ABCG5* and *ABCD1*, produced SLPs above 3, a third, *ABCA1* (SLP = -2.91), was only marginally less significant. The product of *ABCA1* is responsible for transporting cholesterol out of cells and homozygous or compound heterozygous variants in it cause Tangier disease, a familial HDL deficiency syndrome, while heterozygous variants are associated with reduced HDL levels [29, 30]. It has an established role in the regulation of HDL and there are reports that common variants in it are associated with plasma lipid levels [31, 32].

Results of gene set analyses

An initial run of the gene set analyses tended to highlight sets containing hundreds of genes which included one or more of the genes with absolute SLPs over 3 as listed in Table 2 so the analyses were repeated with these genes and *ABCA1* omitted to see if any additional genes of interest could be identified. Given that 1,454 sets were tested a critical MLP to achieve to declare results significant after correction for multiple testing would be log10(1454*20) = 4.46 and this was not achieved by any set. Inspection of the results for the highest scoring sets did not reveal any additional genes which might obviously be involved in hyperlipidaemia risk. The results for all sets are provided in Supplementary Table S2.

Results of variant category analyses

For the two genes showing the most definite evidence of association, *LDLR* and *PCSK9*, a logistic regression analysis of different categories of variant was carried out to elucidate their relative contributions. The results for *LDLR* are shown in Table 3A. It can be seen that disruptive variants, comprising stop variants and frameshift variants, are significantly associated with caseness (SLP = 16.95) with a large effect on risk (OR = 40.02 (11.83 - 135.33)). There were 34 of these variants, of which only 3 were seen in controls. Essential splice site variants also exerted a large effect on risk (OR = 10.4 (1.9 - 56.7)), SLP = 5.55, with 11 out of 13 being seen in cases. Stop variants, frameshift variants and essential splice variants are expected to cause LOF but the results show that other variants which do not severely disrupt the gene but which produce changes in amino acid sequence also have moderate effects on risk. There were 6,747 nonsynonymous variants and this category was associated with OR = 1.15 (1.05 - 1.25). However of these 1,175 were annotated by SIFT as “deleterious” and this category has OR = 1.74 (1.41 - 2.14) while the risk associated with an annotation by PolyPhen of “probably damaging” was smaller, 1.30 (1.03 - 1.65), and there was no significant risk associated with an annotation of “possibly damaging”. Inframe insertion/deletion variants were observed on 10 occasions and detailed inspection of the results revealed that these consisted of deletions at 4 different positions, one of which occurred in 7 different subjects. All 10 of the subjects with one of these deletions was a case (SLP=6.58). By contrast, genetic variants which did not affect protein sequence in general did not have significant effects on risk. The exception was that the “Splice Region” category seemed to exert a protective effect, with OR = 0.86 (0.81 - 0.92), SLP = -5.21. This was driven by rs72658867, which had frequency 0.012 in controls and 0.0097 in cases and which has been previously reported to be associated with lower cholesterol and lower risk of coronary artery disease [33]. When the analysis was repeated with this variant removed, there was no general tendency for splice region variants to be associated with risk (OR = 1.13 (0.99 - 1.29), SLP = 1.20).

ClinVar lists 19 variants in *PCSK9* classified as pathogenic or likely pathogenic but none of these was present in any of the cases or controls. Table 3B shows the results of the variant category analysis

for *PCSK9*. It can be seen that these are broadly similar to those obtained for *LDLR*, albeit in the opposite direction because impaired function of *PCSK9* reduces risk of hyperlipidaemia. Disruptive, essential splice site and missense variants annotated as "deleterious" by SIFT are all significantly more common in controls and have an overall OR of around 0.5. It is interesting to note that these categories of variant occur more frequently in *PCSK9* than in *LDLR*. In *LDLR* there are only 34 disruptive variants whereas in *PCSK9* there are 291 and in *LDLR* there are only 13 essential splice site variants while in *PCSK9* there are 193.

These results allow us to gain some insight into the overall impact of variants in these genes on the risk of hyperlipidaemia in the general population. For variants in *LDLR* which are nonsynonymous and annotated as "deleterious" by SIFT, the overall estimated OR is 1.15*1.74 = 2. It should be emphasised that this estimate is for the average effect of such variants and that there is likely to be considerable variation, with some of these variants exerting marked effects on risk while others may have trivial effects or may even be protective. There are 889 of these variants and, since they are rare, few people have more than one of them so that we can say that around 850 out of the 200,000 subjects, or slightly less than 0.5%, have a deleterious variant in *LDLR* which, on average, about doubles the odds of hyperlipidaemia. This compares with the 47 LOF variants which confer high risk but which occur in only 0.02% of subjects. Deleterious variants in *PCSK9* on average have OR of about 0.5 and occur in 0.8% of subjects while LOF variants have a similar OR and occur in 0.2% of subjects. Broadly speaking, it seems that about 1.5% of people will have a rare coding variant in one of these two genes which either doubles or halves the odds of developing hyperlipidaemia.

Results for selected genes

It is relevant also to report certain genes which produced negative results. With the exception of *LDLR*, none of the genes highlighted by the previous analysis of 50,000 UK Biobank exomes showed any evidence for association in this enlarged sample once sex was included as a covariate. These genes consist of *HUWE1*, *CXorf56*, *RBP2*, *STAT5B*, *NPFFR1*, *ACOT9*, *GK*, *ADIPOQ*, *SURF1*, *ADRB3*, *GYG2*, *PHKA1* and *PHKA2* [1]. *HUWE1* and a number of others are located on the X chromosome and with hindsight it appears that they may have produced strongly negative SLPs as a consequence of the reduced frequency of variants called on the X chromosome in males, while other results may have simply been due to chance. Other genes for which notably negative results are obtained are *APOB* (SLP = 0.00) a known cause of familial hypercholesterolaemia, and *HMGCR* (SLP = -0.07), which codes for the rate-limiting enzyme in cholesterol synthesis which is the target of statins [34]. Also negative was *STAP1* (SLP = -1.02), for which there were initial claims of an association with familial hypercholesterolaemia although more recent work has thrown doubt on this [35]. These three genes were also subjected to the variant category analysis and this revealed that disruptive variants in *APOB* were more frequent in controls (OR = 0.71 (0.60 - 0.85), SLP = -3.74) but that there were much large numbers of nonsynonymous variants which overall did not show association with hyperlipidaemia, accounting for the negative result of weighted burden analysis. No other category of variant within these genes showed significant association with hyperlipidaemia after correction for the number of genes and categories tested. The full results of variant category analysis for these three genes are presented in Supplementary Table S3, along with those for all genes significant at $p < 0.001$ as listed in Table 2.

Discussion

These analyses provide a broad overview of contributions of rare coding genetic variants to the risk of hyperlipidaemia. There are a number of issues worthy of further comment.

The observation that in this dataset rare X chromosome variants are called more frequently in females than in males is important to recognise. Unless this effect is allowed for, for example by incorporating sex as a covariate, artefactual results may be produced for any phenotype whose prevalence varies with sex. With hindsight, this occurred in the earlier analysis of the 50,000 exomes and led to the identification as some genes on the X chromosome as being potentially relevant. Going forward, researchers need to be aware of this phenomenon and deal with it appropriately.

Working with biobank datasets can pose particular challenges compared to traditional case-control studies. In a case-control study one can control recruitment and assess subjects against pre-specified criteria. With the UK Biobank one has a self-selected sample of volunteers along with information about a broad range of phenotypes but some measures are only available for a subset of the sample. The phenotype studied here is intended to broadly capture clinically significant hyperlipidaemia, using as it does a combination of the diagnosis and the most commonly used treatments. However this phenotype clearly differs from what one might use in a more systematically assessed sample. No attempt was made to incorporate actual measures of blood lipids, in part because these might be distorted by treatment effects. Some subjects will have been prescribed statins purely on the basis of raised lipids found during routine clinical assessment whereas other subjects with somewhat lower levels might be receiving them because they had cardiovascular disease. Likewise, some subjects classified as controls might in fact have hyperlipidaemia which has not been diagnosed. Thus, the phenotype is understood to be a quite noisy and a distant consequence of the immediate biological effects of any functional genetic variants. Another issue is that the participants represent a relatively healthy group of subjects. People with severe hyperlipidaemia which had resulted in early death would not be included, meaning that the effect sizes observed in this sample may tend to be underestimates. One approach to following up the results reported here would be to investigate the relationships between the genes of interest and more detailed aspects of the phenotype. This might include looking for associations between particular categories of variant in each gene and aspects of the phenotype such as quantitative measures of individual components of the lipid profile or clinical outcomes such as coronary artery disease.

The current analysis highlights a number of genes for which very rare variants with large effect size have previously been shown to impact lipid levels and now demonstrates that large numbers of additional nonsynonymous variants with more moderate effect also make a broader contribution to risk in the general population, which was not previously recognised. This is most clearly the case for *LDLR* and *PCSK9* but there are a few of other genes which probably also show this effect, especially *ANGPTL3* and *ANGPTL4*. Conversely, other genes which are implicated as monogenic causes of severe familial hyperlipidaemias, such as *APOB* and *STAP1*, are not identified by this approach as making broader contributions to hyperlipidaemia although it does seem that disruptive variants in *APOB* may be associated with somewhat reduced risk. The protective effects of such variants has not previously been reported but antisense inhibition of APOB expression is used to treat familial hyperlipidaemia [36, 37]. The analyses highlight three additional genes which are already the targets of lipid-lowering therapies, *PCSK9*, *ANGPTL3* and *NPC1L1*, but completely failed to detect an effect for *HMGCR*, which encodes the target of statins. *PPP1R3G* has not previously been shown to influence hyperlipidaemia risk in humans but the findings reported here are consistent with those from animal studies, suggesting that it might also be a potential target.

The approach used is intended to detect the additive effects of variants which are individually very rare but which cumulatively have an effect on the function of a gene. Hence it is not expected to be successful if the effect of some variants impairing gene function may be counterbalanced by others which produce a gain of function. It is necessary to group variants because when a variant is only observed in a handful of subjects it is not possible to draw firm conclusions about its effect. There might be scope to gain power by devising more sophisticated approaches to variant classification, for example related to the more specific predictions about effect on the protein product, and this issue will be addressed in future work.

Estimating the effect on risk of different categories of variant within *LDLR* and *PCSK9* broadly confirms what we might have expected. LOF variants, comprising stop, frameshift and splice site variants, have large effects. Variants categorised as "deleterious" by SIFT have moderate effects on risk, whereas the categorisation as "probably damaging" by PolyPhen is associated with a somewhat smaller effect. The annotation of "possibly damaging" does not seem to have much utility in this context. The analyses show that even nonsynonymous variants in *LDLR* which do not have any of these annotations are still, on average, associated with a slightly increased risk, with OR = 1.15. Synonymous variants, intronic variant and other variants which do not cause changes in amino acid sequence do not in general seem to exert an appreciable effect on risk. These findings will be of use in constructing weighting schemes for future analyses of this nature. The variants are mostly all individually extremely rare but in total there are far more nonsynonymous variants than LOF variants. The contribution of different types of variant to risk and how best to model this will be the subject of further investigation.

The general picture which emerges is that there is a relatively small number of genes in which variants which are individually extremely rare make an appreciable contribution to the overall risk of developing hyperlipidaemia. Few variants cause LOF but those which do have a large effect, whereas far larger numbers of nonsynonymous variants tend to exert more moderate effects. Nevertheless, the cumulative frequency of these variants remains low. If we confine attention to the results about which we can feel most confident, it seems that fewer than 2% of people carry a variant which might halve or double risk. It will be possible to refine estimates such as this as more data becomes available, for example from the remaining 300,000 UK Biobank subjects for whom exome sequence data is yet to be provided. With a larger dataset it will become possible to draw more definitive conclusions about individual genes and to make more accurate estimates of effect sizes. Nevertheless, given the small number of carriers it does not seem likely that identifying rare variants with moderate effects will be clinically useful for routine risk assessment or to guide treatment.

The availability of sequence data from a large number of subjects has allowed insights into the contribution which rare coding genetic variants can make to hyperlipidaemia, an important phenotype which is also associated with a variety of socioeconomic and environmental risk factors. More detailed analyses may focus on specific genes and/or variants, may investigate signals of selection pressures and may look at interactions between different genetic and environmental variables. This may lead to better understanding of the biological processes involved and improved treatment strategies. Hyperlipidaemia provides a useful paradigm of a common complex trait and similar approaches can be applied to other phenotypes.

Conflicts of interest

The author declares he has no conflict of interest.

Data availability

The raw data is available on application to UK Biobank. Detailed results with variant counts cannot be made available because they might be used for subject identification. Relevant derived variables including principal components and variant annotations will be deposited in UK Biobank. Scripts and software used to carry out the analyses are available at https://github.com/davenomiddlenamecurtis.

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.

References

1 Curtis D. Analysis of exome-sequenced UK Biobank subjects implicates genes affecting risk of hyperlipidaemia. *Mol Genet Metab* Published Online First: 29 July 2020. doi:10.1016/j.ymgme.2020.07.009

2 Sharifi M, Futema M, Nair D, Humphries SE. Genetic Architecture of Familial Hypercholesterolaemia. Curr. Cardiol. Rep. 2017;**19**:44.

3 Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, Yang J, Visscher PM, Wray NR. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun* 2019;**10**:1–10.

4 Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, Khera A V., Zhou W, Bloom JM, Engreitz JM, Ernst J, O'Connell JR, Ruotsalainen SE, Alver M, Manichaikul A, Johnson WC, Perry JA, Poterba T, Seed C, Surakka IL, Esko T, Ripatti S, Salomaa V, Correa A, Vasan RS, Kellis M, Neale BM, Lander ES, Abecasis G, Mitchell B, Rich SS, Wilson JG, Cupples LA, Rotter JI, Willer CJ, Kathiresan S, Abe N, Albert C, Allred N (Nichole) P, Almasy L, Alonso A, Ament S, Anderson P, Anugu P, Applebaum-Bowden D, Arking D, Arnett DK, Ashley-Koch A, Aslibekyan S, Assimes T, Auer P, Avramopoulos D, Barnard J, Barnes K, Barr RG, Barron-Casella E, Beaty T, Becker D, Becker L, Beer R, Begum F, Beitelshees A, Benjamin E, Bezerra M, Bielak L, Bis J, Blackwell T, Blangero J, Boerwinkle E, Borecki I, Bowler R, Brody J, Broeckel U, Broome J, Bunting K, Burchard E, Cardwell J, Carty C, Casaburi R, Casella J, Chang C, Chasman D, Chavan S, Chen BJ, Chen WM, Chen YDI, Cho M, Choi SH, Chuang LM, Chung M, Cornell E, Crandall C, Crapo J, Curran J, Curtis J, Custer B, Damcott C, Darbar D, Das S, David S, Davis C, Daya M, de Andrade M, DeBaun M, Deka R, DeMeo D, Devine S, Do R, Duan Q, Duggirala R, Durda P, Dutcher S, Eaton C, Ekunwe L, Ellinor P, Emery L, Farber C, Farnam L, Fingerlin T, Flickinger M, Fornage M, Franceschini N, Fu M, Fullerton M, Fulton L, Gabriel S, Gan W, Gao Y, Gass M, Gelb B, Geng X (Priscilla), Germer S, Gignoux C, Gladwin M, Glahn D, Gogarten S, Gong DW, Goring H, Gu CC, Guan Y, Guo X, Haessler J, Hall M, Harris D, Hawley N, He J, Heavner B, Heckbert S, Hernandez R, Herrington D, Hersh C, Hidalgo B, Hixson J, Hokanson J, Hong E, Hoth K, Hsiung C (Agnes), Huston H, Hwu CM, Irvin MR, Jackson R, Jain D, Jaquish C, Jhun MA, Johnsen J, Johnson A, Johnston R, Jones K, Kang HM, Kaplan R, Kardia S, Kaufman L, Kelly S, Kenny E, Kessler M, Khan A, Kinney G, Konkle B, Kooperberg C, Kramer H, Krauter S, Lange C, Lange E, Lange L, Laurie C, Laurie C, LeBoff M, Lee SS, Lee WJ, LeFaive J, Levine D, Levy D, Lewis J, Li Y, Lin H, Lin KH, Liu S, Liu Y, Loos R, Lubitz S, Lunetta K, Luo J, Mahaney M, Make B, Manson JA, Margolin L, Martin L, Mathai S, Mathias R, McArdle P, McDonald ML, McFarland S, McGarvey S, Mei H, Meyers DA, Mikulla J, Min N, Minear M, Minster RL, Musani S, Mwasongwe S, Mychaleckyj JC, Nadkarni G, Naik R, Nekhai S, Nickerson D, North K, O'Connor T, Ochs-Balcom H, Pankow J, Papanicolaou G, Parker M, Parsa A, Penchev S, Peralta JM, Perez M, Peters U, Peyser P, Phillips L, Phillips S, Pollin T, Post W, Becker JP, Boorgula MP, Preuss M, Prokopenko D, Psaty B, Qasba P, Qiao D, Qin Z, Rafaels N, Raffield L, Rao DC, Rasmussen-Torvik L, Ratan A, Redline S, Reed R, Regan E, Reiner A, Rice K, Roden D, Roselli C, Ruczinski I, Russell P, Ruuska S, Ryan K, Sakornsakolpat P, Salimi S, Salzberg S, Sandow K, Sankaran V, Schmidt E, Schwander K, Schwartz D, Sciurba F, Seidman C, Sheehan V, Shetty A, Shetty A, Sheu WHH, Shoemaker MB, Silver B, Silverman E, Smith J, Smith J, Smith N, Smith T, Smoller S, Snively B, Sofer T, Sotoodehnia N, Stilp A, Streeten E, Sung YJ, Sylvia J, Szpiro A, Sztalryd C, Taliun D, Tang H, Taub M, Taylor K, Taylor S, Telen M, Thornton TA, Tinker L, Tirschwell D, Tiwari H, Tracy R, Tsai M, Vaidya D, VandeHaar P, Vrieze S, Walker T, Wallace R, Walts A, Wan E, Wang FF, Watson K, Weeks DE, Weir B, Weiss S, Weng LC, Willer C, Williams K, Williams LK, Wilson C, Wong Q, Xu H, Yanek L, Yang I, Yang R, Zaghloul N, Zhang Y, Zhao SX,

Zheng X, Zhi D, Zhou X, Zody M, Zoellner S. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun* 2018;**9**:1–12.

5 Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang HY, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, Fischer K, Fontanillas P, Fraser RM, Freitag DF, Gurdasani D, Heikkilä K, Hyppönen E, Isaacs A, Jackson AU, Johansson Å, Johnson T, Kaakinen M, Kettunen J, Kleber ME, Li X, Luan J, Lyytikäinen LP, Magnusson PKE, Mangino M, Mihailov E, Montasser ME, Müller-Nurasyid M, Nolte IM, O'Connell JR, Palmer CD, Perola M, Petersen AK, Sanna S, Saxena R, Service SK, Shah S, Shungin D, Sidore C, Song C, Strawbridge RJ, Surakka I, Tanaka T, Teslovich TM, Thorleifsson G, Van Den Herik EG, Voight BF, Volcik KA, Waite LL, Wong A, Wu Y, Zhang W, Absher D, Asiki G, Barroso I, Been LF, Bolton JL, Bonnycastle LL, Brambilla P, Burnett MS, Cesana G, Dimitriou M, Doney ASF, Döring A, Elliott P, Epstein SE, Eyjolfsson GI, Gigante B, Goodarzi MO, Grallert H, Gravito ML, Groves CJ, Hallmans G, Hartikainen AL, Hayward C, Hernandez D, Hicks AA, Holm H, Hung YJ, Illig T, Jones MR, Kaleebu P, Kastelein JJP, Khaw KT, Kim E, Klopp N, Komulainen P, Kumari M, Langenberg C, Lehtimäki T, Lin SY, Lindström J, Loos RJF, Mach F, McArdle WL, Meisinger C, Mitchell BD, Müller G, Nagaraja R, Narisu N, Nieminen TVM, Nsubuga RN, Olafsson I, Ong KK, Palotie A, Papamarkou T, Pomilla C, Pouta A, Rader DJ, Reilly MP, Ridker PM, Rivadeneira F, Rudan I, Ruokonen A, Samani N, Scharnagl H, Seeley J, Silander K, Stancáková A, Stirrups K, Swift AJ, Tiret L, Uitterlinden AG, Van Pelt LJ, Vedantam S, Wainwright N, Wijmenga C, Wild SH, Willemsen G, Wilsgaard T, Wilson JF, Young EH, Zhao JH, Adair LS, Arveiler D, Assimes TL, Bandinelli S, Bennett F, Bochud M, Boehm BO, Boomsma DI, Borecki IB, Bornstein SR, Bovet P, Burnier M, Campbell H, Chakravarti A, Chambers JC, Chen YDI, Collins FS, Cooper RS, Danesh J, Dedoussis G, De Faire U, Feranil AB, Ferrières J, Ferrucci L, Freimer NB, Gieger C, Groop LC, Gudnason V, Gyllensten U, Hamsten A, Harris TB, Hingorani A, Hirschhorn JN, Hofman A, Hovingh GK, Hsiung CA, Humphries SE, Hunt SC, Hveem K, Iribarren C, Järvelin MR, Jula A, Kähönen M, Kaprio J, Kesäniemi A, Kivimaki M, Kooner JS, Koudstaal PJ, Krauss RM, Kuh D, Kuusisto J, Kyvik KO, Laakso M, Lakka TA, Lind L, Lindgren CM, Martin NG, März W, McCarthy MI, McKenzie CA, Meneton P, Metspalu A, Moilanen L, Morris AD, Munroe PB, Njølstad I, Pedersen NL, Power C, Pramstaller PP, Price JF, Psaty BM, Quertermous T, Rauramaa R, Saleheen D, Salomaa V, Sanghera DK, Saramies J, Schwarz PEH, Sheu WHH, Shuldiner AR, Siegbahn A, Spector TD, Stefansson K, Strachan DP, Tayo BO, Tremoli E, Tuomilehto J, Uusitupa M, Van Duijn CM, Vollenweider P, Wallentin L, Wareham NJ, Whitfield JB, Wolffenbuttel BHR, Ordovas JM, Boerwinkle E, Palmer CNA, Thorsteinsdottir U, Chasman DI, Rotter JI, Franks PW, Ripatti S, Cupples LA, Sandhu MS, Rich SS, Boehnke M, Deloukas P, Kathiresan S, Mohlke KL, Ingelsson E, Abecasis GR. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–85.

6 Dron JS, Hegele RA. Genetics of Lipid and Lipoprotein Disorders and Traits. *Curr Genet Med Rep* 2016;**4**:130–41.

7 Szustakowski JD, Balasubramanian S, Sasson A, Khalid S, Bronson PG, Kvikstad E, Wong E, Liu D, Davis JW, Haefliger C, Loomis AK, Mikkilineni R, Noh HJ, Wadhawan S, Bai X, Hawes A, Krasheninina O, Ulloa R, Lopez A, Smith EN, Waring J, Whelan CD, Tsai EA, Overton J, Salerno W, Jacob H, Szalma S, Runz H, Hinkle G, Nioi P, Petrovski S, Miller MR, Baras A, Mitnaul L, Reid JG. Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. *medRxiv* 2020;:2020.11.02.20222232.

8 Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, Suvisaari J, Chheda H, Blackwood D, Breen G, Pietiläinen O, Gerety SS, Ayub M, Blyth M, Cole T, Collier D, Coomber EL, Craddock N, Daly MJ, Danesh J, DiForti M, Foster A, Freimer NB, Geschwind D, Johnstone M, Joss S, Kirov G, Körkkö J, Kuismin O, Holmans P, Hultman CM, Iyegbe C, Lönnqvist J, Männikkö M, McCarroll SA, McGuffin P, McIntosh AM, McQuillin A, Moilanen JS, Moore C, Murray RM,

Newbury-Ecob R, Ouwehand W, Paunio T, Prigmore E, Rees E, Roberts D, Sambrook J, Sklar P, Clair DS, Veijola J, Walters JTR, Williams H, Sullivan PF, Hurles ME, O'Donovan MC, Palotie A, Owen MJ, Barrett JC. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* 2016;**19**:571–7.

9 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.

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

11 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;**17**:122.

12 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.

13 Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 2016;**98**:456–72.

14 Curtis D. Multiple Linear Regression Allows Weighted Burden Analysis of Rare Coding Variants in an Ethnically Heterogeneous Population. *Hum Hered* 2021;:1–10.

15 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.

16 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.

17 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 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.

18 R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria.: : R Foundation for Statistical Computing 2014. http://www.r-project.org

19 Wang X, Musunuru K. Angiopoietin-Like 3: From Discovery to Therapeutic Gene Editing. JACC Basic to Transl. Sci. 2019;**4**:755–62.

20 Doggrell SA. Will evinacumab become the standard treatment for homozygous familial hypercholesterolemia? *Expert Opin Biol Ther* 2020;:1–4.

21 Whyte MP, Aronson J, McAlister WH, Weinstein RS, Wenkert D, Clements KL, Gottesman GS, Madson KL, Stolina M, Bijanki VN, Plotkin H, Huskey M, Duan S, Mumm S. Coalescing Expansile Skeletal Disease: Delineation Of An Extraordinary Osteopathy Involving The IFITM5 Mutation Of Osteogenesis Imperfecta Type V. *Bone* 2020;**145**:115835.

22 Yang R, Liu N, Chen L, Jiang Y, Shi Y, Mao C, Liu Y, Wang M, Lai W, Tang H, Gao M, Xiao D, Wang X, Zhou H, Tang C e., Liu W, Yu F, Cao Y, Yan Q, Liu S, Tao Y. GIAT4RA functions as a tumor suppressor in non-small cell lung cancer by counteracting Uchl3–mediated deubiquitination of LSH. *Oncogene* 2019;**38**:7133–45.

23 Kishnani PS, Austin SL, Abdenur JE, Arn P, Bali DS, Boney A, Chung WK, Dagli AI, Dale D, Koeberl D, Somers MJ, Burns Wechsler S, Weinstein DA, Wolfsdorf JI, Watson MS. Diagnosis and management of glycogen storage disease type I: A practice guideline of the American College of Medical Genetics and Genomics. *Genet Med* 2014;**16**:1–29.

24 Cuchel M, Bruckert E, Ginsberg HN, Raal FJ, Santos RD, Hegele RA, Kuivenhoven JA,

Nordestgaard BG, Descamps OS, Steinhagen-Thiessen E, Tybjærg-Hansen A, Watts GF, Averna M, Boileau C, Borén J, Catapano AL, Defesche JC, Hovingh GK, Humphries SE, Kovanen PT, Masana L, Pajukanta P, Parhofer KG, Ray KK, Stalenhoef AFH, Stroes E, Taskinen MR, Wiegman A, Wiklund O, Chapman MJ. Homozygous familial hypercholesterolaemia: New insights and guidance for clinicians to improve detection and clinical management. A position paper fromthe Consensus Panel on Familial Hypercholesterolaemia of the European Atherosclerosis Society. Eur. Heart J. 2014;**35**:2146–57.

25 Engelen M, Kemp S, De Visser M, Van Geel BM, Wanders RJA, Aubourg P, Poll-The BT. X-linked adrenoleukodystrophy (X-ALD): Clinical presentation and guidelines for diagnosis, follow-up and management. Orphanet J. Rare Dis. 2012;**7**. doi:10.1186/1750-1172-7-51

26 Ma Y, Han X, Zhou X, Li Y, Gong S, Zhang S, Cai X, Zhou L, Luo Y, Li M, Liu W, Zhang X, Ren Q, Ji L. A new clinical screening strategy and prevalence estimation for glucokinase variant-induced diabetes in an adult Chinese population. *Genet Med* 2019;**21**:939–47.

27 Zhang Y, Gu J, Wang L, Zhao Z, Pan Y, Chen Y. Ablation of PPP1R3G reduces glycogen deposition and mitigates high-fat diet induced obesity. *Mol Cell Endocrinol* 2017;**439**:133–40.

28 Betters JL, Yu L. NPC1L1 and cholesterol transport. FEBS Lett. 2010;**584**:2740–7.

29 Maranghi M, Truglio G, Gallo A, Grieco E, Verrienti A, Montali A, Gallo P, Alesini F, Arca M, Lucarelli M. A novel splicing mutation in the ABCA1 gene, causing Tangier disease and familial HDL deficiency in a large family. *Biochem Biophys Res Commun* 2019;**508**:487–93.

30 Puntoni M, Sbrana F, Bigazzi F, Sampietro T. Tangier Disease. *Am J Cardiovasc Drugs* 2012;**12**:303–11.

31 Koldamova R, Fitz NF, Lefterov I. ATP-binding cassette transporter A1: From metabolism to neurodegeneration. Neurobiol. Dis. 2014;**72**:13–21.

32 Lu Z, Luo Z, Jia A, Yu L, Muhammad I, Zeng W, Song Y. Associations of the ABCA1 gene polymorphisms with plasma lipid levels. *Medicine (Baltimore)* 2018;**97**:e13521.

33 Gretarsdottir S, Helgason H, Helgadottir A, Sigurdsson A, Thorleifsson G, Magnusdottir A, Oddsson A, Steinthorsdottir V, Rafnar T, de Graaf J, Daneshpour MS, Hedayati M, Azizi F, Grarup N, Jørgensen T, Vestergaard H, Hansen T, Eyjolfsson G, Sigurdardottir O, Olafsson I, Kiemeney LA, Pedersen O, Sulem P, Thorgeirsson G, Gudbjartsson DF, Holm H, Thorsteinsdottir U, Stefansson K. A Splice Region Variant in LDLR Lowers Non-high Density Lipoprotein Cholesterol and Protects against Coronary Artery Disease. *PLoS Genet* 2015;**11**. doi:10.1371/journal.pgen.1005379

34 LaRosa JC, He J, Vupputuri S. Effect of statins on risk of coronary disease. A meta-analysis of randomized controlled trials. J. Am. Med. Assoc. 1999;**282**:2340–6.

35 Kanuri B, Fong V, Haller A, Hui DY, Patel SB. Mice lacking global Stap1 expression do not manifest hypercholesterolemia. *BMC Med Genet* 2020;**21**. doi:10.1186/s12881-020-01176-x

36 Akdim F, Tribble DL, Flaim JD, Yu R, Su J, Geary RS, Baker BF, Fuhr R, Wedel MK, Kastelein JJP. Efficacy of apolipoprotein B synthesis inhibition in subjects with mild-to-moderate hyperlipidaemia. *Eur Heart J* 2011;**32**:2650–9.

37 Parham JS, Goldberg AC. Mipomersen and its use in familial hypercholesterolemia. *Expert Opin Pharmacother* 2019;**20**:127–31.

Table 1

The table shows the weight which was assigned to each type of variant as annotated by VEP, Polyphen and SIFT as well as the broad categories which were used for multivariate analyses of variant effects [9–11].

VEP / SIFT / Polyphen annotation	Weight	Category
intergenic_variant	1	Unused
feature_truncation	3	Intronic, etc.
regulatory_region_variant	3	Intronic, etc.
feature_elongation	3	Intronic, etc.
regulatory_region_amplification	3	Intronic, etc.
regulatory_region_ablation	3	Intronic, etc.
TF_binding_site_variant	3	Intronic, etc.
TFBS_amplification	3	Intronic, etc.
TFBS_ablation	3	Intronic, etc.
downstream_gene_variant	3	Intronic, etc.
upstream_gene_variant	3	Intronic, etc.
non_coding_transcript_variant	3	Intronic, etc.
NMD_transcript_variant	3	Intronic, etc.
intron_variant	3	Intronic, etc.
non_coding_transcript_exon_variant	3	Intronic, etc.
3_prime_UTR_variant	10	3 prime UTR
5_prime_UTR_variant	5	5 prime UTR
mature_miRNA_variant	5	Unused
coding_sequence_variant	5	Unused
synonymous_variant	5	Synonymous
stop_retained_variant	5	Unused
incomplete_terminal_codon_variant	5	Unused
splice_region_variant	5	Splice region
protein_altering_variant	10	Protein altering
missense_variant	10	Protein altering
inframe_deletion	15	InDel, etc
inframe_insertion	15	InDel, etc
transcript_amplification	15	InDel, etc
start_lost	30	Unused
stop_lost	30	Unused
frameshift_variant	40	Disruptive
stop_gained	40	Disruptive
splice_donor_variant	40	Splice site variant
splice_acceptor_variant	40	Splice site variant
transcript_ablation	20	Disruptive
SIFT deleterious	10	Deleterious
PolyPhen possibly damaging	10	Possibly damaging
PolyPhen probably damaging	10	Probably damaging

Table 2

Genes with absolute value of SLP exceeding 3 or more (equivalent to p<0.001) for test of association of weighted burden score with hyperlipidaemia.

Table 2A

Genes with SLP greater than 3.

Symbol	SLP	Name
LDLR	50.08	Low Density Lipoprotein Receptor
G6PC	5.55	Glucose-6-Phosphatase Catalytic Subunit
SULT1E1	4.63	Sulfotransferase Family 1E Member 1
LOC101928415	4.50	Uncharacterized LOC101928415
SLC35G1	4.38	Solute Carrier Family 35 Member G1
PLA2G5	4.15	Phospholipase A2 Group V
CMTM7	3.99	CKLF Like MARVEL Transmembrane Domain Containing 7
MIR6716	3.95	MicroRNA 6716
COL4A2-AS2	3.85	COL4A2 Antisense 2
DEFB131A	3.66	Defensin Beta 131A
OTULIN	3.62	OTU Deubiquitinase With Linear Linkage Specificity
FAM122C	3.51	Family With Sequence Similarity 122C
CMIP	3.48	C-Maf Inducing Protein
EIF4B	3.45	Eukaryotic Translation Initiation Factor 4B
PPP2R3B	3.41	Protein Phosphatase 2 Regulatory Subunit B''Beta
HNRNPAB	3.38	Heterogeneous Nuclear Ribonucleoprotein A/B
PREB	3.37	Prolactin Regulatory Element Binding
PEX12	3.36	Peroxisomal Biogenesis Factor 12
FAM167A	3.35	Family With Sequence Similarity 167 Member A
ABCG5	3.31	ATP Binding Cassette Subfamily G Member 5
ABCD1	3.26	ATP Binding Cassette Subfamily D Member 1
PRAF2	3.21	PRA1 Domain Family Member 2
CTHRC1	3.16	Collagen Triple Helix Repeat Containing 1
SLC25A37	3.14	Solute Carrier Family 25 Member 37
CT62	3.11	Cancer/Testis Associated 62
L1TD1	3.09	LINE1 Type Transposase Domain Containing 1
PIK3R6	3.09	Phosphoinositide-3-Kinase Regulatory Subunit 6
FOXO3B	3.08	Forkhead Box O3B
FAM47A	3.06	Family With Sequence Similarity 47 Member A
MIR6806	3.06	MicroRNA 6806
GCK	3.04	Glucokinase
MAPKAPK2	3.02	MAPK Activated Protein Kinase 2

Table 2B

Genes with SLP less than -3.

Symbol	SLP	Name
PCSK9	-10.42	Proprotein Convertase Subtilisin/Kexin Type 9
IFITM5	-5.86	Interferon Induced Transmembrane Protein 5
LOC102723729	-5.77	Uncharacterized LOC102723729
ANGPTL3	-5.67	Angiopoietin Like 3
APOC3	-4.89	Apolipoprotein C3
PPP1R3G	-4.25	Protein Phosphatase 1 Regulatory Subunit 3G
LOC107985474	-3.95	Uncharacterized LOC107985474
TBC1D8	-3.93	TBC1 Domain Family Member 8
CTXN2	-3.91	Cortexin 2
NPC1L1	-3.70	NPC1 Like Intracellular Cholesterol Transporter 1
ANGPTL4	-3.66	Angiopoietin Like 4
SNX17	-3.62	Sorting Nexin 17
LOC105377994	-3.40	Uncharacterized LOC105377994
LOC101929609	-3.30	Uncharacterized LOC101929609
SV2B	-3.29	Synaptic Vesicle Glycoprotein 2B
ITM2B	-3.23	Integral Membrane Protein 2B
UBR4	-3.23	Ubiquitin Protein Ligase E3 Component N-Recognin 4
TXNL4A	-3.22	Thioredoxin Like 4A
TTR	-3.16	Transthyretin
GFPT1	-3.10	Glutamine--Fructose-6-Phosphate Transaminase 1
APPBP2	-3.08	Amyloid Beta Precursor Protein Binding Protein 2
CRYZL1	-3.06	Crystallin Zeta Like 1
HLA-A	3.01	Major Histocompatibility Complex, Class I, A

Table 3

Results from logistic regression analysis showing the contribution different categories of variant within a gene make to risk of hyperlipidaemia. Odds ratios for each category are estimated including principal components and sex as covariates.

Table 3A

Results for *LDLR.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR (95% CI)	SLP
Intronic, etc	27207	0.173760	7854	0.178281	1.01 (0.99 - 1.04)	0.47
5 prime UTR	55	0.000351	21	0.000477	1.43 (0.85 - 2.41)	0.78
Synonymous	2582	0.016490	682	0.015481	0.92 (0.84 - 1.00)	-1.31
Splice region	6649	0.042464	1735	0.039383	0.86 (0.81 - 0.92)	-5.21
3 prime UTR	506	0.003232	139	0.003155	0.98 (0.81 - 1.19)	-0.06
Protein altering	4947	0.031594	1800	0.040859	1.15 (1.05 - 1.25)	2.87
InDel, etc	0	0.000000	10	0.000227		6.58
Disruptive	3	0.000019	31	0.000704	40.02 (11.83 - 135.33)	16.95
Splice site variant	2	0.000013	11	0.000250	23.31 (4.96 - 109.42)	5.55
Deleterious	702	0.004483	473	0.010737	1.74 (1.41 - 2.14)	6.87
Possibly damaging	400	0.002555	200	0.004540	1.20 (0.96 - 1.49)	1.02
Probably damaging	539	0.003442	350	0.007945	1.30 (1.03 - 1.65)	1.66

Table 3B

Results for *PCSK9.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR (95% CI)	SLP
Intronic, etc	5317	0.033958	1540	0.034957	1.02 (0.96 - 1.08)	0.31
5 prime UTR	269	0.001718	78	0.001771	1.03 (0.80 - 1.34)	0.10
Synonymous	7230	0.046175	2145	0.048690	1.02 (0.97 - 1.07)	0.40
Splice region	291	0.001858	90	0.002043	1.04 (0.81 - 1.33)	0.12
3 prime UTR	2418	0.015443	707	0.016048	1.02 (0.94 - 1.12)	0.24
Protein altering	3388	0.021638	831	0.018863	0.96 (0.86 - 1.06)	-0.44
InDel, etc	4	0.000026	3	0.000068	1.81 (0.39 - 8.48)	0.74
Disruptive	202	0.001290	29	0.000658	0.51 (0.34 - 0.77)	-3.05
Splice site variant	179	0.001143	14	0.000318	0.29 (0.17 - 0.51)	-5.08
Deleterious	1404	0.008967	279	0.006333	0.59 (0.45 - 0.77)	-4.00
Possibly damaging	249	0.001590	68	0.001544	1.19 (0.88 - 1.61)	-0.60
Probably damaging	1050	0.006706	227	0.005153	1.29 (0.96 - 1.74)	-1.10

Figure 1

QQ plot of SLPs obtained for weighted burden analysis of association with hyperlipidaemia showing observed against expected SLP for each gene, omitting results for *LDLR* and *PCSK9*.