www.nature.com/ijo

International Journal of Obesity

ARTICLE

Genetics and Epigenetics

Weighted burden analysis in 200,000 exome-sequenced subjects characterises rare variant effects on BMI

David Curtis [1,2]✉

INTRODUCTION: A number of genes have been identified in which rare variants can cause obesity. Here we analyse a sample of exome sequenced subjects from UK Biobank using BMI as a phenotype with the aims of identifying genes in which rare, functional variants influence BMI and characterising the effects of different categories of variant.
METHODS: There were 199,807 exome sequenced subjects for whom BMI was recorded. Weighted burden analysis of rare, functional variants was carried out, incorporating population principal components and sex as covariates. For selected genes, additional analyses were carried out to clarify the contribution of different categories of variant. Statistical significance was summarised as the signed log 10 of the p value (SLP), given a positive sign if the weighted burden score was positively correlated with BMI.
RESULTS: Two genes were exome-wide significant, *MC4R* (SLP = 15.79) and *PCSK1* (SLP = 6.61). In *MC4R*, disruptive variants were associated with an increase in BMI of 2.72 units and probably damaging nonsynonymous variants with an increase of 2.02 units. In *PCSK1*, disruptive variants were associated with a BMI increase of 2.29 and protein-altering variants with an increase of 0.34. Results for other genes were not formally significant after correction for multiple testing, although *SIRT1*, *ZBED6* and *NPC2* were noted to be of potential interest.
CONCLUSION: Because the UK Biobank consists of a self-selected sample of relatively healthy volunteers, the effect sizes noted may be underestimates. The results demonstrate the effects of very rare variants on BMI and suggest that other genes and variants will be definitively implicated when the sequence data for additional subjects becomes available.

International Journal of Obesity (2022) 46:782–792; https://doi.org/10.1038/s41366-021-01053-4

INTRODUCTION

Genome wide association studies (GWAS) detect large numbers of common variants showing statistically significant association with obesity although it can be difficult to interpret the biological processes underlying these signals [1]. In addition, a small number of genes have been identified in which very rare variants can have a major effect on body mass index (BMI) and their contribution and mechanisms have recently been reviewed [2]. In some of these, such as *LEP*, *LEPR*, *PCSK1* and *SIM1*, recessively acting variants cause deficiency of the gene product and this can result in obesity. In others, including *POMC* and *MC4R*, heterozygous variants have been reported to be causative. Dominantly and recessively acting *MC4R* variants together constitute the commonest causes of inherited early-onset obesity, with a prevalence of 0.5–0.6%. It is also recognised that other nonsynonymous variants in *MC4R* can be associated with lower BMI and can be protective against obesity [3, 4].

As sequence data becomes available for larger numbers of subjects it is possible to explore the contribution of rare genetic variants to traits in the general population and we recently reported results obtained from analysing the association between rare variants and BMI in 50,000 exome-sequenced UK Biobank subjects [5]. Although no gene was exome wide significant, the analysis did highlight some which were potentially of interest, including *LYPLAL1* and *NSDHL*. Since then, additional data has been released meaning that exome sequence data is now available for 200,000 of the 500,000 UK Biobank subjects to approved researchers [6]. Analyses of this larger dataset shows that it is better powered to detect rare variant effects and such analyses were successful in implicating, at exome-wide significance, genes previously recognised as risk factors for both hyperlipidaemia and type 2 diabetes [7, 8]. Here, we apply the same approach as previously, using BMI as the phenotype in the enlarged sample.

Early access to exome sequence data from the remaining UK Biobank subjects was granted to Regeneron Pharmaceuticals Inc. and their collaborators and a study using data from 429,000 UK Biobank subjects of European origin along with 217,000 from other samples has recently been published [9]. This study of over 640,000 exomes used BMI as a phenotype and performed burden analyses of rare variants to implicate 16 genes at exome-wide significance: *UHMK1*, *GPR75*, *ROBO1*, *KIAA1109*, *PCSK1*, *GPR151*, *SPARC*, *UBR2*, *CALCR*, *PDE3B*, *ANO4*, *KIAA0586*, *MC4R*, *DPP9*, *ANKRD27* and *GIPR*. The approach used in the present study differs in a number of ways. The 640 K exome study excluded UK

[1] UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT, UK. [2] Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK. ✉email: d.curtis@ucl.ac.uk

Received: 20 January 2021 Revised: 29 November 2021 Accepted: 13 December 2021
Published online: 24 January 2022

Biobank subjects of non-European ancestry, whereas we have previously shown that the methods used here are robust against population stratification, allowing the inclusion of subjects of all ancestries without inflation of the test statistic [5]. The 640 K exome study applied a simple burden analysis whereby counts of different categories of variant within each gene, such as predicted loss of function and missense predicted to be deleterious, were totalled together to test for association with BMI. Seven different variant selection models were used, requiring an additional correction for multiple testing. By contrast, we apply a weighted burden analysis which incorporates all variants in a single analysis but with higher weights assigned to those expected to have a larger effect. For example, predicted loss of function variants are given higher weights than missense variants predicted to be deleterious, which in turn have a higher weight than other missense variants. This removes the requirement to correct for multiple testing and, more importantly, is expected to yield higher power. From a statistical point of view, this is because likelihood ratio tests have higher power when the model for the alternative hypothesis more closely resembles the real situation. Thus, if predicted loss of function variants do in fact have a larger effect on the phenotype than missense variants, then one expects to gain power by weighting them differently. We have previously shown that predicted loss of function variants in *LDLR* are associated with a very high odds ratio for developing hyperlipidaemia whereas variants annotated as deleterious have a much more modest, though still statistically significant, effect [7]. In addition, we assign higher weights to variants which are rarer, under the assumption that variants which are extremely rare may have larger effect sizes. Work on real world data confirms that weighting on variant annotation and allele frequency in this way does indeed yield increased power [10].

The aims of this study were to identify genes in which rare sequence variants had an effect on BMI and to characterise the effect sizes of different categories of variant. Attention was focused on rare variants because the effects of common variants would have been well established from previous GWASs.

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 [6]. 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 [11–13]. 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 [14, 15]. The phenotype was obtained from data field 21001-0.0, which records BMI at first assessment. Although BMI is an imperfect indicator of adiposity it has been widely used in similar investigations and also has the advantage that this data field is missing in very few subjects.

Using the same approach as described previously, the SCOREASSOC program was used to carry out a weighted burden analysis to test whether, in each gene, the weighted burden of sequence variants which were rarer and/or predicted to have more severe functional effects correlated with BMI [5]. Attention was restricted to rare variants with minor allele frequency (MAF) < = 0.01. 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 [5]. 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 [16, 17]. The weights allocated are to some extent arbitrary but took account of the analysis of the effects of different categories of variant in *LDLR* on hyperlipidaemia risk [7]. A systematic exploration of different weighting schemes has shown than no one scheme is optimal in all circumstances but that the one used here has reasonable performance [10]. Variants predicted to cause complete loss of function (LOF) of the gene were assigned a weight of 100. Nonsynonymous variants were assigned a weight of 5 but if PolyPhen annotated them as possibly or probably damaging then 5 or 10 was added to this and if SIFT annotated them as deleterious then 20 was added. 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

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 [11–13].

VEP/SIFT/Polyphen annotation	Weight	Category
intergenic_variant	0	Unused
feature_truncation	0	Intronic, etc.
regulatory_region_variant	0	Intronic, etc.
feature_elongation	0	Intronic, etc.
regulatory_region_amplification	1	Intronic, etc.
regulatory_region_ablation	1	Intronic, etc.
TF_binding_site_variant	1	Intronic, etc.
TFBS_amplification	1	Intronic, etc.
TFBS_ablation	1	Intronic, etc.
downstream_gene_variant	0	Intronic, etc.
upstream_gene_variant	0	Intronic, etc.
non_coding_transcript_variant	0	Intronic, etc.
NMD_transcript_variant	0	Intronic, etc.
intron_variant	0	Intronic, etc.
non_coding_transcript_exon_variant	0	Intronic, etc.
3_prime_UTR_variant	1	3 prime UTR
5_prime_UTR_variant	1	5 prime UTR
mature_miRNA_variant	5	Unused
coding_sequence_variant	0	Unused
synonymous_variant	0	Synonymous
stop_retained_variant	5	Unused
incomplete_terminal_codon_variant	5	Unused
splice_region_variant	1	Splice region
protein_altering_variant	5	Protein altering
missense_variant	5	Protein altering
inframe_deletion	10	InDel, etc
inframe_insertion	10	InDel, etc
transcript_amplification	10	InDel, etc
start_lost	10	Unused
stop_lost	10	Unused
frameshift_variant	100	Disruptive
stop_gained	100	Disruptive
splice_donor_variant	100	Splice site variant
splice_acceptor_variant	100	Splice site variant
transcript_ablation	100	Disruptive
SIFT deleterious	20	Deleterious
PolyPhen possibly damaging	5	Possibly damaging
PolyPhen probably damaging	10	Probably damaging

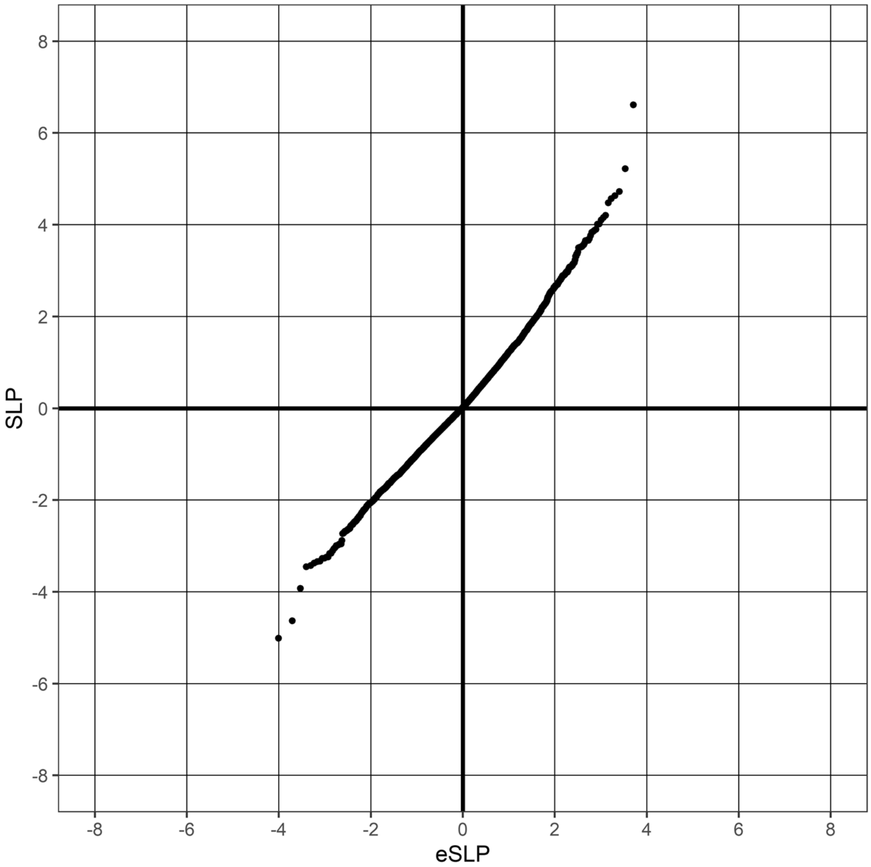

Fig. 1 QQ plot of gene-level SLPs testing association with BMI. QQ plot of SLPs obtained for weighted burden analysis of association with BMI showing observed against expected SLP for each gene, omitting results for *MC4R*, which has SLP = 15.79.

together to provide an overall weight for each variant. Variants were excluded if there were >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 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, multiple linear regression analysis was carried out including the first 20 population principal components and sex as covariates and a likelihood ratio test was performed comparing the likelihoods of the models with and without the gene-wise burden score. 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 BMI. 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 BMI.

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 [18]. 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 BMI.

For selected genes, additional analyses were carried out to clarify the contribution of different categories of variant. As described previously, multiple linear regression analyses were performed on the counts of the separate categories of variant as listed in Table 1, again including principal components and sex as covariates, to estimate the effect size for each category [7]. The mean effect on BMI for each category was estimated along with the standard error and the Wald statistic was used to obtain a *p* value. The associated p value was converted to an SLP, again with the sign being positive if the mean count was positively correlated with BMI. In these analyses, stop variants and frameshift variants were considered jointly as "disruptive variants" and splice site variants were considered separately, although all three types of variant might generally be expected to have a similar LOF effect.

Data manipulation and statistical analyses were performed using GENEVARASSOC, SCOREASSOC and R [19]. Code availability: Software and scripts used to carry out the analyses are available at https://github.com/davenomiddlenamecurtis.

RESULTS

There were 199,807 exome sequenced subjects for whom BMI was recorded. There were 110,092 male subjects with mean age 56.3 (SD = 8.0) and mean BMI 27.0 (SD = 5.1). There were 89,715 female subjects with mean age 56.7 (SD = 8.2) and mean BMI 27.8 (SD = 4.2). There were 20,384 genes for which there were qualifying variants, meaning that the critical threshold for the absolute value of the SLP to declare a result as formally statistically significant is -log10(0.05/20384) = 5.61. This threshold was met by two genes, *MC4R* (SLP = 15.79) and *PCSK1* (SLP = 6.61). The quantile-quantile (QQ) plot for the SLPs obtained for all genes except *MCR4* is shown in Fig. 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 1.23 with intercept at −0.0005 and the gradient for negative SLPs is 1.03 with intercept at 0.02, indicating only moderate inflation of the test statistic for those genes showing a positive correlation.

Table 2. Results from regression analysis showing the effects on BMI of different categories of variant within the two exome-wide significant genes, *MC4R* and *PCSK1*.

Category	Number of different variants	Total number of variants	Average variant load per subject	BMI mean in carriers	BMI SD in carriers	SLP	Effect on mean BMI (95% CI)
(A) Results for *MC4R*.							
Intronic, etc	0	0					
5 prime UTR	25	286	0.001431	27.82	5.08	1.02	0.47 (−0.09–1.03)
Synonymous	60	883	0.004419	28.14	5.19	−0.66	−0.19 (−0.50–0.12)
Splice region	0	0					
3 prime UTR	7	49	0.000245	27.57	5.26	0.12	0.20 (−1.15–1.55)
Protein altering	140	1355	0.006782	28.24	5.42	0.03	0.02 (−0.34–0.37)
InDel, etc	1	4	0.000020	28.76	8.08	0.33	1.70 (−3.03–6.42)
Disruptive	19	80	0.000400	30.16	4.93	6.55	2.72 (1.66–3.79)
Splice site variant	0	0					
Deleterious	70	452	0.002262	28.70	5.88	−0.54	−0.50 (−1.44–0.44)
Possibly damaging	23	201	0.001006	28.42	5.51	1.68	0.92 (0.13–1.72)
Probably damaging	55	425	0.002127	28.98	5.86	4.29	2.02 (1.02–3.02)
Subjects with no variant		197329	0.987598	27.37	4.75		
(B) Results for *PCSK1*.							
Intronic, etc	592	20821	0.104206	27.67	4.95	0.86	0.04 (−0.01–0.08)
5 prime UTR	21	3081	0.015420	27.74	5.14	0.05	0.01 (−0.16–0.18)
Synonymous	136	3616	0.018097	28.22	5.17	0.08	0.02 (−0.15–0.18)
Splice region	35	164	0.000821	27.87	5.21	0.54	0.39 (−0.35–1.12)
3 prime UTR	27	79	0.000395	27.46	4.62	−0.09	−0.12 (−1.19–0.94)
Protein altering	292	2970	0.014864	27.73	4.96	2.74	0.34 (0.12–0.56)
InDel, etc	4	14	0.000070	26.93	6.16	−0.14	−0.45 (−2.97–2.08)
Disruptive	22	51	0.000255	29.66	6.29	3.28	2.29 (0.97–3.62)
Splice site variant	3	8	0.000040	29.59	8.40	0.64	2.01 (−1.33–5.35)
Deleterious	153	990	0.004955	27.87	5.06	0.56	0.34 (−0.28–0.95)
Possibly damaging	53	451	0.002257	27.44	4.54	−0.41	−0.27 (−0.90–0.36)
Probably damaging	102	602	0.003013	27.85	5.36	−0.21	−0.17 (−0.87–0.52)
Subjects with no variant		176391	0.882807	27.34	4.73		

For each category of variant, the table shows the number of different variants of that category (at different locations) and the total number of times a variant of that category occurred. Also shown is the mean and SD of the BMI for all subjects carrying at least one variant of that category. The SLP is the signed log10 *p* value from the regression analysis and the estimated effect for each category is the fitted mean change in BMI after incorporating principal components and sex as covariates.

For the two exome-wide significant genes, *MC4R* (SLP = 15.79) and *PCSK1* (SLP = 6.61), logistic regression analysis of different categories of variants was carried out to elucidate their relative contributions. The results are shown in Table 2, which shows differences between the genes relating to the implicated pattern of variants. In *MC4R*, disruptive variants (stop and frameshift) are associated with a highly significant (SLP = 6.55) increase in BMI by 2.72 units, equivalent to about 8 kg for somebody of average height, and carriers have an average BMI of 30.16. These variants occur a total of 80 times at 19 separate positions. There are no splice site variants. In addition, nonsynonymous variants annotated by PolyPhen as probably damaging are also significantly (SLP = 4.29) associated with an average increase in BMI of 2.02 units. These occur in total 425 times at 55 positions. By contrast, other variants, including those annotated as deleterious by SIFT, are not associated with BMI changes. The estimated effect of the probably damaging variants represents an average across all the variants in this category and of course it is possible that some have major effects whereas other do not. However inspection of the detailed results showed that all of these variants were very rare (MAF < 0.001) and so it was not possible to reliably assess the effect of any individual variant. In *PCSK1*, disruptive variants are also significantly (SLP = 3.28) associated with an increase in BMI of 2.29 units and carriers have a mean BMI of 29.66. The estimated effect of splice site variants, which are also predicted to cause LOF, is similar, an increase of 2.01 units, but they only occur 8 times and this effect is not statistically significant. In contrast with *MC4R*, there is no suggestion that variants in *PCSK1* annotated as probably damaging have any effect on BMI. However the much larger general category of protein-altering variants is associated with a modest (0.34 units) but statistically significant (SLP = 2.74) increase in BMI. In total these occur 2,970 times, meaning that there is an average burden per subject of 0.015.

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

Gene symbol	SLP	Gene name
MC4R	15.79	Melanocortin 4 Receptor
PCSK1	6.61	Proprotein Convertase Subtilisin/ Kexin Type 1
PTOV1	5.22	PTOV1 Extended AT-Hook Containing Adaptor Protein
GALNT14	4.72	Polypeptide N-Acetylgalactosaminyltransferase 14
LOC112268007	4.63	GRM3 Antisense RNA 1
RNF187	4.57	Ring Finger Protein 187
LOC102724050	4.48	Uncharacterised LOC102724050
DYNC1H1	4.21	Dynein Cytoplasmic 1 Heavy Chain 1
SMARCE1	4.15	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily E, Member 1
SMPD1	4.10	Sphingomyelin Phosphodiesterase 1
SATL1	4.02	Spermidine/Spermine N1-Acetyl Transferase Like 1
GALNT9	4.01	Polypeptide N-Acetylgalactosaminyltransferase 9
ZDHHC17	3.90	Zinc Finger DHHC-Type Palmitoyltransferase 17
LOC101927911	3.88	Uncharacterised LOC101927911
CFP	3.85	Complement Factor Properdin
TRIP12	3.83	Thyroid Hormone Receptor Interactor 12
FOXK2	3.76	Forkhead Box K2
SHROOM2	3.70	Shroom Family Member 2
ADNP	3.66	Activity Dependent Neuroprotector Homeobox
HSFX1	3.66	Heat Shock Transcription Factor Family, X-Linked 1
CTAGE1	3.66	Cutaneous T Cell Lymphoma-Associated Antigen 1
NUDT16L1	3.65	Nudix Hydrolase 16 Like 1
PRR36	3.59	Proline Rich 36
BCLAF3	3.56	BCLAF1 And THRAP3 Family Member 3
DPP8	3.55	Dipeptidyl Peptidase 8
SRPK2	3.52	SRSF Protein Kinase 2
ZC3H8	3.52	Zinc Finger CCCH-Type Containing 8
ACSL3	3.51	Acyl-CoA Synthetase Long Chain Family Member 3
FAM19A1	3.51	TAFA Chemokine Like Family Member 1
OCRL	3.50	OCRL Inositol Polyphosphate-5-Phosphatase
FLJ44635	3.50	TPT1-Like Protein
OS9	3.41	OS9 Endoplasmic Reticulum Lectin
UBR3	3.37	Ubiquitin Protein Ligase E3 Component N-Recognin 3
CPA5	3.36	Carboxypeptidase A5
OR6C3	3.31	Olfactory Receptor Family 6 Subfamily C Member 3
PTPRG	3.31	Protein Tyrosine Phosphatase Receptor Type G
H2AFZ	3.23	H2A.Z Variant Histone 1
AMOT	3.18	Angiomotin
SIRT1	3.16	Sirtuin 1
CRYBG3	3.15	Crystallin Beta-Gamma Domain Containing 3
RNASE7	3.14	Ribonuclease A Family Member 7
ATP12A	3.11	ATPase H+/K+ Transporting Non-Gastric Alpha2 Subunit
SLC17A9	3.11	Solute Carrier Family 17 Member 9
CITED2	3.11	Cbp/P300 Interacting Transactivator With Glu/Asp Rich Carboxy-Terminal Domain 2
NMI	3.08	N-Myc And STAT Interactor
CACNA1I	3.08	Calcium Voltage-Gated Channel Subunit Alpha1 I
TGIF2LX	3.08	TGFB Induced Factor Homeobox 2 Like X-Linked
BMP10	3.08	Bone Morphogenetic Protein 10
FOXD4L1	3.07	Forkhead Box D4 Like 1
UBE4B	3.05	biquitination Factor E4B
SCN8A	3.04	Sodium Voltage-Gated Channel Alpha Subunit 8
AEBP1	−3.04	AE Binding Protein 1
ANTXRL	−3.06	ANTXR Like
ATP8B2	−3.10	ATPase Phospholipid Transporting 8B2
ITLN2	−3.16	Intelectin 2
POPDC3	−3.17	Popeye Domain Containing 3
MIR6881	−3.24	MicroRNA 6881
CFAP97D1	−3.24	CFAP97 Domain Containing 1
USP4	−3.27	Ubiquitin Specific Peptidase 4
CLUH	−3.27	Clustered Mitochondria Homolog
HS6ST3	−3.33	Heparan Sulfate 6-O-Sulfotransferase 3
ZBED6	−3.34	Zinc Finger BED-Type Containing 6
FAM171B	−3.37	Family With Sequence Similarity 171 Member B
PKP4	−3.43	Plakophilin 4
GIT2	−3.45	GIT ArfGAP 2
NOP14	−3.93	NOP14 Nucleolar Protein
DEFB4B	−4.63	Defensin Beta 4B
BAIAP3	−5.01	BAI1 Associated Protein 3

One would expect that by chance 20 genes would produce SLPs with absolute value greater than 3, equivalent to $p < 0.001$, whereas in fact there are 68, suggesting that some might have an effect on BMI while failing to reach exome-wide significance after correction for multiple testing. These genes are listed in Table 3 and the SLPs for all genes are listed in Supplementary Table S1. Variant category analyses were carried out for those which seemed biologically plausible as well as for genes previously reported to be causative of obesity as listed in the introduction. These analyses yielded some findings of possible interest, discussed as follows.

It is perhaps striking that two similar genes, *GALNT14* (SLP = 4.72) and *GALNT9* (SLP = 4.01), fall within the top 13 genes. These enzymes catalyze the transfer of N-acetyl-D-galactosamine (GalNAc) to the hydroxyl groups on serines and threonines in target peptides. The *GALNT9* intronic SNP rs11247009-A has been reported to be associated with BMI ($p = 6 \times 10^{-9}$) [20]. A study of broiler chickens claimed that in unpublished data one of the six most highly significant variants in a genome-wide study of abdominal fat was in *GALNT9* and reported that *GALNT9* expression in liver differed between lean and fat lines [21]. However, overall there seems to be little prior evidence to implicate these genes as affecting BMI and they have mostly been studied in the context of cancer progression, although there is also a report of a homozygous frameshift variant of *GALNT14* being found in a patient with nonsyndromic keratoconus. The results of variant-wise analysis of these two genes are shown in Table 4A, B. This shows that *GALNT14* there are 302 disruptive variants associated with a significant (SLP = 2.89) increase in BMI of 0.88 units, while in *GALNT9* there are 12 splice site variants associated with an increase in BMI of 3.97 units (SLP = 2.44) and 9 indels associated with an increase in BMI of 4.84 units (SLP = 2.65). 36 disruptive variants in *GALNT9* are also associated with an increase in BMI of 1.14 units but this is not statistically significant (SLP = 0.86).

The results for *SIRT1* (SLP = 3.16) are potentially of interest because SIRT1 and other sirtuins have effects similar to calorie restriction and reduced expression of *SIRT1* and *SIRT2* promotes adipogenesis and accumulation of visceral fat [22, 23]. From these findings one might well predict that genetic variants damaging *SIRT1* might lead to increased BMI. The results from variant-wise analysis are shown in Table 4C, which shows only weakly significant effects from disruptive (SLP = 1.34) and possibly damaging (SLP = 1.61) variants.

ZBED6 (SLP = −3.33) codes for a transcriptional inhibitor of *IGF2* which has a major impact on muscle development in placental mammals and CRISPR/Cas9 disruption of its binding site is being used commercially to produce strains of pigs which are leaner and have enhanced muscle development [24, 25]. The results for variant-wise analysis are shown in Table 4D, showing that disruptive variants are associated with a reduction in BMI of 1.59 units (SLP = −2.48) and deleterious nonsynonymous variants with a reduction of 0.37 units (SLP = −1.49).

The gene with the most negative SLP, *BAIAP3* (SLP = −5.01), may have some role in insulin secretion but does not in general seem to be an obvious candidate to have effects on BMI [26]. Splice site variants are associated with a reduction in BMI of 1.41 units (SLP = −3.47).

It is well established that variants in *LEP* (SLP = 0.61) and *LEPR* (SLP = 0.13) can cause obesity but the gene-based analyses produced no evidence to implicate them. The results of variant-wise analyses are shown in Table 5A, B. It can be seen that disruptive and splice site variants in *LEP* do indeed have substantially higher BMIs but because there are only 6 of them this does not produce a statistically significant effect, at least if one corrects for the numbers of categories tested. There is no suggestion that any other type of variant has an effect. By contrast, in *LEPR* there are a total of 88 disruptive and splice site variants but their effect on mean BMI is negligible, as is also the case for other types of variant.

A common nonsynonymous variant *BDNF*, rs6265, causes a Val66Met substitution which was originally reported to be associated with anorexia nervosa and minimum BMI in anorexia nervosa patients and whose effect on BMI was subsequently confirmed in large GWAS samples [27, 28]. This variant shows highly significant association in the current sample (SLP = −21.86). The number of subjects with Val/Val, Val/Met and Met/Met genotypes is 132,003, 60,639 and 7165 with uncorrected mean BMIs of 27.47, 27.22 and 26.96. The per-allele effect size on BMI as estimated from multiple linear regression analysis including principal components and sex as covariates is −0.19 (−0.23 to −0.15). However the gene-wise weighted burden analysis of *BDNF* using rare variants produced no evidence for association (SLP = 0.41) and variant-wise analyses likewise failed to show any effect from any category of rare variant. The mean effect size for protein-altering variants was 0.23 but there were only 1910 of these in total and the result does not approach statistical significance.

Of the remaining genes implicated by the analysis of the 640 K exome study, some produced some evidence for association which did not survive correction for multiple testing consisting of *UHMK1* (SLP = −1.53), *GPR75* (SLP = −2.98), *ROBO1* (SLP = 2.21), *KIAA1109* (SLP = 1.84), *UBR2* (SLP = 2.69), *PDE3B* (SLP = 2.06), *ANO4* (SLP = 1.50), *DPP9* (SLP = −2.09) and *GIPR* (SLP = −2.63). However other genes showed no overall evidence for association, consisting of *GPR151* (SLP = −0.80), *SPARC* (SLP = 0.14), *CALCR* (SLP = 0.41), *KIAA0586* (SLP = −0.84) and *ANKRD27* (SLP = −0.37). Detailed variant category analyses for these genes are presented in Supplementary Table S2. For some genes, it was possible to identify particular variant categories which appeared to be associated with BMJ. These consisted of disruptive variants in *GPR75* (SLP = −4.87), *ROBO1* (SLP = 2.91), *KIAA1109* (SLP = 3.20), *GPR151* (SLP = −2.17) and *ANO4* (SLP = 1.31) whereas the broad category of protein-altering variants produced the strongest signal in two other genes, *SPARC* (SLP = 4.93) and *GIPR* (SLP = −4.29). For other genes, no category of variant was associated.

Other genes previously implicated in obesity which likewise failed to show evidence of association in either gene-wise analyses or variant category analyses include *SIM1* (SLP = 0.89), *NTRK2* (SLP = 0.88), *KSR2* (SLP = 0.17), *CPE* (SLP = −0.35), *SH2B1* (SLP = 0.78), *TUB* (SLP = −0.08) and *FTO* (SLP = 1.02). Variant category analyses for all genes of interest are presented in Supplementary Table S3.

In order to see if any additional genes were highlighted by analysing gene sets, gene set analysis was performed as described above after first removing all genes with absolute SLP value greater than 3. In order to correct for the observed inflation of the positive SLPs, the absolute value of each SLP was divided by an average inflation factor of 1.13 before being utilised to contribute to the set-wise chi-squared statistic. Following this adjustment, no gene set produced a result significant after correction for multiple testing. The highest MLP was 2.45, achieved by the set Specific Transcriptional Repressor Activity. Out of 1454 sets, the fifth highest ranked was Regulation Of Lipid Metabolic Process (MLP = 1.93). This contains 12 genes including *NPC2* (SLP = 2.80), which is involved in cholesterol transport and recessively acting variants in *NPC2* are a cause of Niemann-Pick C disease in which lipid accumulation causes neurodegeneration [29]. NPC2 presents cholesterol to NPC1 and rare LOF variants in *NPC1* are known to cause obesity although *NPC1* does not demonstrate association with BMI in the current sample (SLP = 0.15) [30]. In a GWAS of obesity in F2 pigs a variant within *NPC2*, rs81396056, produced the most highly significant result ($p = 10^{-16}$) [31]. The results of variant category analysis of *NPC2* are shown in Table 4E and it can be seen that there is significant (SLP = 3.10) association of 3119 splice site variants, occurring at 3 different positions, with an average increase in BMI of 0.28. Disruptive variants are also associated with higher BMI but there are only 111 of them and this result is not statistically significant. Results for all gene sets are presented in Supplementary Table S4.

DISCUSSION

These analyses help to elucidate the impact of rare genetic variants on a complex phenotype such as BMI and also illustrate some of the challenges of dealing with exome sequence data. The gene-wise weighted burden analyses successfully identify two genes already known to impact BMI, *MC4R* and *PCSK1*, but fail to detect effects of other known obesity genes. In due course sequence data will become available for all 500,000 UK Biobank participants and it is reasonable to expect that this larger dataset will produce additional results. For example, the subjects with LOF variants in *LEP* do have notably higher BMIs but there are so few of them that they do not produce a statistically significant result in this sample. Obviously, the power to detect association depends both on the effect size and the frequency of variants, and power will improve with increased sample size. To take another example

Table 4. Results from variant category regression analyses for other genes of possible interest.

Category	Number of different variants	Total number of variants	Average variant load per subject	BMI mean in carriers	BMI SD in carriers	SLP	Effect on mean BMI (95% CI)
(A) Results for *GALNT14*.							
Intronic, etc	877	20167	0.100932	27.73	4.98	−0.02	−0.00 (−0.07–0.06)
5 prime UTR	38	697	0.003488	28.16	5.27	0.05	0.02 (−0.34–0.38)
Synonymous	121	5558	0.027817	27.49	4.79	−0.66	−0.08 (−0.22–0.05)
Splice region	46	1204	0.006026	28.94	5.02	−0.33	−0.10 (−0.38–0.18)
3 prime UTR	22	243	0.001216	27.22	4.71	−0.25	−0.17 (−0.78–0.43)
Protein altering	299	9851	0.049303	27.47	4.80	0.03	0.00 (−0.10–0.11)
InDel, etc	5	6	0.000030	24.72	3.11	−0.81	−2.74 (−6.59–1.12)
Disruptive	30	302	0.001511	28.24	5.35	2.89	0.88 (0.33–1.42)
Splice site variant	13	50	0.000250	27.76	4.82	0.22	0.35 (−0.99–1.69)
Deleterious	176	2302	0.011521	27.72	4.91	1.00	0.32 (−0.07–0.71)
Possibly damaging	46	695	0.003478	27.93	4.84	0.26	0.15 (−0.36–0.66)
Probably damaging	140	1335	0.006681	27.58	5.04	−0.18	−0.09 (−0.52–0.34)
Subjects with no variant		171281	0.857232	27.34	4.73		
(B) Results for *GALNT9*.							
Intronic, etc	613	25032	0.125281	27.69	4.99	−1.08	−0.05 (−0.10–0.01)
5 prime UTR	34	132	0.000661	28.21	4.57	−0.24	−0.22 (−1.01–0.57)
Synonymous	183	10134	0.050719	27.90	5.20	0.69	0.06 (−0.04–0.17)
Splice region	44	231	0.001156	27.45	4.87	0.24	0.17 (−0.45–0.79)
3 prime UTR	228	11273	0.056419	27.82	5.00	−0.12	−0.01 (−0.10–0.08)
Protein altering	362	2952	0.014774	27.66	4.95	0.43	0.13 (−0.16–0.41)
InDel, etc	5	9	0.000045	32.21	10.83	2.65	4.84 (1.68–8.01)
Disruptive	19	36	0.000180	28.17	7.12	0.86	1.14 (−0.40–2.68)
Splice site variant	7	12	0.000060	31.20	12.28	2.44	3.97 (1.24–6.70)
Deleterious	207	1626	0.008138	27.63	4.99	−0.08	−0.04 (−0.45–0.37)
Possibly damaging	83	951	0.004760	27.68	4.79	0.46	0.21 (−0.24–0.66)
Probably damaging	109	736	0.003684	27.69	5.05	0.50	0.25 (−0.25–0.75)
Subjects with no variant		172467	0.863168	27.34	4.73		
(C) Results for *SIRT1*.							
Intronic, etc	590	12438	0.062250	27.35	4.81	0.10	0.01 (−0.08–0.10)
5 prime UTR	43	777	0.003889	27.39	4.75	−1.16	-0.31 (−0.65–0.03)
Synonymous	168	3064	0.015335	27.64	4.98	0.62	0.10 (−0.07–0.27)
Splice region	29	54	0.000270	28.04	4.19	0.44	0.56 (−0.68–1.80)
3 prime UTR	21	376	0.001882	28.69	4.97	−0.04	−0.03 (−0.53–0.47)
Protein altering	320	8020	0.040139	27.48	4.79	0.13	0.03 (−0.14–0.19)
InDel, etc	19	432	0.002162	27.81	5.32	0.85	0.33 (−0.12–0.79)
Disruptive	16	26	0.000130	29.44	6.52	1.34	1.79 (0.00–3.57)
Splice site variant	2	3	0.000015	31.03	6.93	0.84	3.99 (−1.47–9.44)
Deleterious	130	4552	0.022782	27.43	4.76	0.11	0.03 (−0.19–0.25)
Possibly damaging	33	98	0.000490	28.54	5.35	1.61	1.08 (0.12–2.04)
Probably damaging	57	257	0.001286	27.62	4.76	0.33	0.22 (−0.38–0.83)
		177451	0.888112	27.37	4.75		

Table 4. continued

Category	Number of different variants	Total number of variants	Average variant load per subject	BMI mean in carriers	BMI SD in carriers	SLP	Effect on mean BMI (95% CI)
Subjects with no variant							
(D) Results for *ZBED6*.							
Intronic, etc	0	0					
5 prime UTR	0	0					
Synonymous	145	1316	0.006586	28.12	5.36	0.39	0.11 (−0.15–0.37)
Splice region	0	0					
3 prime UTR	19	104	0.000521	28.26	4.34	1.24	0.88 (−0.05–1.81)
Protein altering	322	4785	0.023948	27.44	4.80	−0.07	−0.02 (−0.18–0.15)
InDel, etc	11	102	0.000510	27.13	4.54	−0.17	−0.20 (−1.13–0.74)
Disruptive	40	74	0.000370	25.72	3.48	−2.48	−1.59 (−2.68 –0.51)
Splice site variant	0	0					
Deleterious	121	914	0.004574	27.36	4.92	−1.49	−0.37 (−0.72–0.02)
Possibly damaging	69	410	0.002052	27.68	4.57	0.25	0.14 (−0.35–0.63)
Probably damaging	99	451	0.002257	27.44	4.61	−0.01	−0.01 (−0.48–0.47)
Subjects with no variant		193605	0.968960	27.37	4.75		
(E) Results for *NPC2*.							
Intronic, etc	118	2755	0.013788	27.50	4.81	0.41	0.08 (−0.10–0.26)
5 prime UTR	41	361	0.001807	27.75	4.57	−1.54	−0.55 (−1.05–0.05)
Synonymous	35	299	0.001496	26.68	4.53	−1.51	−0.59 (−1.14–0.04)
Splice region	11	907	0.004539	27.79	4.98	−0.10	−0.04 (−0.36–0.28)
3 prime UTR	85	1461	0.007312	27.91	4.89	0.46	0.12 (−0.13–0.37)
Protein altering	72	1573	0.007873	27.69	4.96	−0.43	−0.18 (−0.58–0.22)
InDel, etc	1	1	0.000005	23.91		−0.28	
Disruptive	10	111	0.000556	28.94	5.74	1.27	0.87 (−0.03–1.76)
Splice site variant	3	3119	0.015610	27.63	4.95	3.10	0.28 (0.11–0.45)
Deleterious	41	742	0.003714	28.45	5.10	0.96	0.40 (−0.10–0.90)
Possibly damaging	12	525	0.002628	26.97	4.56	−0.68	−0.33 (−0.86–0.19)
Probably damaging	19	68	0.000340	27.51	4.59	−0.08	−0.13 (−1.33–1.07)
Subjects with no variant		189490	0.948365	27.36	4.75		
(F) Results for *BAIAP3*.							
Intronic, etc	1486	30454	0.152417	27.47	4.82	−1.44	−0.05 (−0.10–0.00)
5 prime UTR	29	272	0.001361	27.39	4.29	0.35	0.22 (−0.35–0.79)
Synonymous	390	4376	0.021901	27.70	4.96	0.09	0.02 (−0.12–0.16)
Splice region	144	1224	0.006126	27.75	5.09	0.18	0.06 (−0.21–0.33)
3 prime UTR	79	3417	0.017102	27.72	5.09	0.20	0.04 (−0.12–0.20)
Protein altering	700	14175	0.070943	27.37	4.77	−0.90	−0.10 (−0.23–0.03)
InDel, etc	6	26	0.000130	26.15	4.84	−0.87	−1.39 (−3.24–0.46)
Disruptive	59	293	0.001466	27.08	4.69	−0.61	−0.32 (−0.87–0.23)
Splice site variant	25	145	0.000726	26.02	3.99	−3.47	−1.41 (−2.19–0.62)
Deleterious	355	7254	0.036305	27.32	4.79	−0.01	−0.00 (−0.17–0.16)
Possibly damaging	139	1475	0.007382	27.32	4.60	−0.19	−0.06 (−0.32–0.20)
Probably damaging	178	3773	0.018883	27.12	4.56	−0.96	−0.15 (−0.33–0.04)
Subjects with no variant		157851	0.790017	27.37	4.75		

The tables show the numbers of variant of each category, their total numbers and the mean and SD of BMI observed in variant carriers along with the SLP and estimated effect size.

Table 5. Results from variant category regression analyses for *LEP* and *LEPR*.

Category	Number of different variants	Total number of variants	Average variant load per subject	BMI mean in carriers	BMI SD in carriers	SLP	Effect on mean BMI (95% CI)
(A) Results for *LEP*.							
Intronic, etc.	58	7313	0.036600	27.61	5.00	−0.16	−0.02 (−0.13–0.09)
5 prime UTR	4	13	0.000065	29.44	5.61	1.00	2.16 (−0.46–4.78)
Synonymous	45	1153	0.005771	27.77	5.13	1.71	0.32 (0.05–0.60)
Splice region	2	4	0.000020	30.91	6.67	0.83	3.41 (−1.31–8.14)
3 prime UTR	10	2486	0.012442	27.63	4.85	−0.20	−0.05 (−0.24–0.14)
Protein altering	50	1235	0.006181	28.62	5.34	−0.21	−0.07 (−0.37–0.22)
InDel, etc	1	1	0.000005	27.70		−0.01	−0.14 (−9.59–9.30)
Disruptive	4	5	0.000025	32.05	3.90	1.64	4.80 (0.58–9.03)
Splice site variant	1	1	0.000005	33.46		0.68	5.91 (−3.54–15.35)
Deleterious	18	59	0.000295	27.10	4.00	−0.48	−0.96 (−2.95–1.03)
Possibly damaging	8	91	0.000455	27.30	4.91	0.01	0.01 (−1.06–1.09)
Probably damaging	15	49	0.000245	27.27	4.05	0.42	0.95 (−1.21–3.11)
Subjects with no variant		188116	0.941489	27.36	4.74		
(B) Results for *LEPR*.							
Intronic, etc	2481	51218	0.256337	27.43	4.80	−0.65	−0.02 (−0.06–0.02)
5 prime UTR	48	3200	0.016015	27.25	4.73	−0.01	−0.00 (−0.17–0.17)
Synonymous	145	4298	0.021511	27.67	4.96	−0.13	−0.02 (−0.18–0.13)
Splice region	43	147	0.000736	27.11	4.57	−0.22	−0.20 (−0.98–0.58)
3 prime UTR	19	117	0.000586	28.26	5.55	1.24	0.83 (−0.05–1.70)
Protein altering	396	3862	0.019329	27.64	4.82	−0.26	−0.07 (−0.32–0.17)
InDel, etc	3	4	0.000020	27.31	4.59	0.02	0.14 (−4.58–4.86)
Disruptive	25	53	0.000265	27.49	5.13	0.06	0.11 (−1.19–1.40)
Splice site variant	8	35	0.000175	26.47	5.69	−0.47	−0.77 (−2.37–0.83)
Deleterious	159	1950	0.009759	27.76	4.96	0.11	0.05 (−0.30–0.40)
Possibly damaging	77	990	0.004955	27.55	4.77	0.57	0.21 (−0.17–0.60)
Probably damaging	83	1189	0.005951	27.96	4.99	0.31	0.14 (−0.27–0.55)
Subjects with no variant		153180	0.766640	27.36	4.74		

The tables show the numbers of variant of each category, their total numbers and the mean and SD of BMI observed in variant carriers along with the SLP and estimated effect size.

of this issue, although the results for the *BDNF* Val66Met variant are highly statistically significant, other protein altering variants in *BDNF* are associated with a larger average effect size but do not produce a statistically significant result because they are cumulatively so much rarer.

The results provide some indication about the quantitative effects of sequence variants but we should first note that the UK Biobank is not completely representative. It consists of volunteer participants who are on average older and healthier than the population as a whole. One implication of this is that subjects with more severe phenotypes will be less likely to be included and an overall effect of this will be to underestimate the effect size of rare variants which can cause morbidity and premature mortality. For example, we can observe that LOF variants in *MC4R* and *PCSK1* are associated with an average increase of 2 or more BMI units but that this estimate may well represent a floor for the real effect size, and indeed much larger effects have been reported in a birth cohort characterised at age 18 [32].

The public health impact of genetic variants depends on their effect and on how many people carry them. For those categories of variant which are rare, the proportion of people carrying such a variant will be approximated by the average variant load because few people will have more than one variant. Thus, we may say that 0.04% of this sample has a LOF variant in *MC4R* associated with an increase of 2.7 in BMI while 0.2% have a variant annotated as probably damaging by PolyPhen associated with an average BMI increase of 2.0. Likewise, <0.03% of the sample has a LOF variant in *PCSK1* which tends to increase BMI by 2.3 units whereas 1.5% carry a protein altering variant associated with an average BMI increase of 0.3.

The analyses fail to conclusively implicate novel genes as influencing BMI. The three which are arguably biologically the most plausible are *SIRT1*, *ZBED6* and *NPC2* but it must be acknowledged that the statistical evidence supporting their involvement is fairly weak. Conversely, there are other genes with higher statistical significance but whose function, as far as it is known, does not immediately suggest that they would have a prominent role in influencing BMI. It is clear that additional data will be needed to arrive at definitive solutions, whether it be from the remaining UK Biobank subjects or from alternative sources.

The results from these analyses would seem to point to very rare variants in a fairly small number of genes as having detectable effects on BMI but there are some caveats which are worth stating. First, the approach used makes the assumption that when variants are considered jointly then they will tend to have the same direction of effect on the phenotype. This seems a reasonable assumption for LOF variants, expected to reduce the functioning of a gene, but the method would fail if some non-synonymous variants reduced function but were balanced out by others which produced gain of function. While we may expect that on average a non-synonymous change, especially one annotated as damaging or deleterious, will be more likely to impair than improve function it is important to acknowledge that if there is a good deal of heterogeneity of effect then genes and classes of variant will fail to achieve statistical significance. Thus, these results should not be taken to exclude the possibility that there may be very large numbers of individually rare variants in many genes which might cumulatively make a substantial contribution to the overall variance of BMI in the population.

Another point to make is that association studies such as this, especially those based on population samples, are not expected to necessarily identify genes which affect BMI but rather genes in which naturally occurring variation affects BMI. For example, there are large variations in the frequency with which LOF variants are observed in different genes, reflecting partly the size of the gene but also selection pressures. Only 6 subjects have LOF variants in *LEP* compared to thousands in *NPC2* and so *LEP* does not produce a detectable signal. However it may well be that by recognising *LEP* as potentially having a major and direct impact on BMI, functional studies will yield useful understanding of the underlying physiology. It should be noted that the selection pressures reducing variation in a particular gene may relate to the phenotype under consideration, here BMI, but may also involve other biological processes impacting on fitness.

To conclude, the study of very large, exome-sequenced samples such as the UK Biobank can afford us further insights into the relationship between genetic variation and a quantitative, health-related phenotype such as BMI.

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. 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.

REFERENCES

1. Müller MJ, Geisler C, Blundell J, Dulloo A, Schutz Y, Krawczak M, et al. The case of GWAS of obesity: does body weight control play by the rules? [Internet]. Vol. 42, Int J Obes. Nature Publishing Group; 2018 [cited 2021 Jan 13]. p. 1395–405. Available from: https://doi.org/10.1038/s41366-018-0081-6
2. Thaker VV. Genetic and epigenetic causes of obesity. Adolesc Med State Art Rev [Internet]. 2017 [cited 2021 Jan 13];28:379–405. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30416642
3. Stutzmann F, Vatin V, Cauchi S, Morandi A, Jouret B, Landt O, et al. Non-synonymous polymorphisms in melanocortin-4 receptor protect against obesity: the two facets of a Janus obesity gene. Hum Mol Genet [Internet]. 2007 Aug 1 [cited 2021 Jan 13];16:1837–44. Available from: https://pubmed.ncbi.nlm.nih.gov/17519222/
4. Brouwers B, Oliveira EM de, Marti-Solano M, Monteiro FBF, Laurin S-A, Keogh JM, et al. Human MC4R variants affect endocytosis, trafficking and dimerization revealing multiple cellular mechanisms involved in weight regulation. Cell Rep [Internet]. 2021 Mar 23 [cited 2021 Sep 24];34. Available from: http://www.cell.com/article/S2211124721001765/fulltext
5. Curtis D. Multiple linear regression allows weighted burden analysis of rare coding variants in an ethnically heterogeneous population. Hum Hered [Internet]. 2020 Jan 7 [cited 2021 Jan 8];1–10. Available from: https://www.karger.com/Article/FullText/512576
6. Szustakowski JD, Balasubramanian S, Sasson A, Khalid S, Bronson PG, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. medRxiv [Internet]. 2020 Jan 1;2020.11.02.20222232. Available from: http://medrxiv.org/content/early/2020/11/04/2020.11.02.20222232.abstract
7. Curtis D. Analysis of 200,000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. J Med Genet [Internet]. 2021 Apr 28 [cited 2021 Apr 30];jmedgenet-2021-107752. Available from: https://doi.org/10.1136/jmedgenet-2021-107752
8. 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 [Internet]. 2021 [cited 2021 Sep 24]; Available from: https://pubmed.ncbi.nlm.nih.gov/34216101/
9. Akbari P, Gilani A, Sosina O, Kosmicki JA, Khrimian L, Fang YY, et al. Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. Science (80−) [Internet]. 2021 [cited 2021 Sep 24];373. Available from: https://doi.org/10.1126/science.abf8683
10. Curtis D. Exploration of weighting schemes based on allele frequency and annotation for weighted burden association analysis of complex phenotypes. Gene [Internet]. 2021 Oct [cited 2021 Nov 29];809:146039. Available from: https://pubmed.ncbi.nlm.nih.gov/34688815/
11. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS.Thormann A,et al. The ensembl variant effect predictor. Genome Biol [Internet]. 2016 Jun 6 [cited 2017 May 9] 17:122. Available from: https://doi.org/10.1186/s13059-016-0974-4.
12. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet [Internet]. 2013[cited 2017 May 17];7 Unit7.20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23315928
13. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc [Internet]. 2009 Jun 25 [cited 2017 May 17];4:1073–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19561590
14. 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 [Internet]. 2015 Dec 25 [cited 2017 Sep 19];4:7. Available from: https://doi.org/10.1186/s13742-015-0047-8
15. 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 [Internet]. 2016 Mar 3 [cited 2020 Dec 14];98:456–72. Available from: https://pubmed.ncbi.nlm.nih.gov/26924531/
16. 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.
17. 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 [Internet]. 2016;26:223–7. http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00041444-900000000-99634
18. 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 USA [Internet]. 2005;102:15545–50. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16199517
19. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria.: R Foundation for Statistical Computing; 2014. Available from: http://www.r-project.org
20. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. PLoS Genet [Internet]. 2015;11:1–42. https://pubmed.ncbi.nlm.nih.gov/26426971/
21. Jin P, Wu X, Xu S, Zhang H, Li Y, Cao Z, et al. Differential expression of six genes and correlation with fatness traits in a unique broiler population. Saudi J Biol Sci [Internet]. 2017 May 1 [cited 2021 Jan 18];24:945–9. Available from: https://doi.org/10.1016/j.sjbs.2015.04.014
22. Chang HC, Guarente L. SIRT1 and other sirtuins in metabolism. Vol. 25, Trends in Endocrinology and Metabolism. Elsevier Current Trends; 2014. p. 138–45.
23. Perrini S, Porro S, Nigro Pasquale, Cignarelli A, Caccioppoli C, Valentina, et al. Reduced SIRT1 and SIRT2 expression promotes adipogenesis of human visceral adipose stem cells and associates with accumulation of visceral fat in human obesity. Int J Obes [Internet]. 2020;44:307–19. Available from: https://doi.org/10.1038/s41366-019-0436-7

24. Liu X, Liu H, Wang M, Li R, Zeng J, Mo D, et al. Disruption of the ZBED6 binding site in intron 3 of IGF2 by CRISPR/Cas9 leads to enhanced muscle development in Liang Guang Small Spotted pigs. Transgenic Res [Internet]. 2019;28:141–50. https://pubmed.ncbi.nlm.nih.gov/30488155/
25. Younis S, Schönke M, Massart J, Hjortebjerg R, Sundström E, Gustafson U, et al. The ZBED6-IGF2 axis has a major effect on growth of skeletal muscle and internal organs in placental mammals. Proc Natl Acad Sci USA [Internet]. 2018 Feb 27 [cited 2021 Jan 18];115:E2048–57. Available from: https://pubmed.ncbi.nlm.nih.gov/29440408/
26. Zhang X, Jiang S, Mitok KA, Li L, Attie AD, Martin TFJ. BAI AP3, a C2 domain-containing Munc 13 protein, controls the fate of dense-core vesicles in neuroendocrine cells. J Cell Biol [Internet]. 2017 [cited 2021 Jan 18];216:2151–66. Available from: https://pubmed.ncbi.nlm.nih.gov/28626000/
27. Ribasés M, Gratacòs M, Armengol L, De Cid R, Badía A, Jiménez L, et al. Met66 in the brain-derived neurotrophic factor (BDNF) precursor is associated with anorexia nervosa restrictive type. Mol Psychiatry [Internet]. 2003 Jul 30 [cited 2021 Jan 19];8:745–51. Available from: www.nature.com/mp
28. Pulit SL, Stoneman C, Morris AP, Wood AR, Glastonbury CA, Tyrrell J, et al. Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet [Internet]. 2019 Jan 1 [cited 2021 Jan 19];28:166–74. Available from: https://pubmed.ncbi.nlm.nih.gov/30239722/
29. Xu Y, Zhang Q, Tan L, Xie X, Zhao Y. The characteristics and biological significance of NPC2: Mutation and disease. Vol. 782, Mutation Research - Reviews in Mutation Research. Elsevier B.V.; 2019. p. 108284.
30. Liu R, Zou Y, Hong J, Cao M, Cui B, Zhang H, et al. Rare loss-of-function variants in npc1 predispose to human obesity. Diabetes [Internet]. 2017;66:935–47. https://pubmed.ncbi.nlm.nih.gov/28130309/
31. Kogelman LJA, Pant SD, Fredholm M, Kadarmideen HN. Systems genetics of obesity in an F2 pig model by genome-wide association, genetic network, and pathway analyses. Front Genet [Internet]. 2014 [cited 2021 Jan 19];5(Jul). Available from: https://pubmed.ncbi.nlm.nih.gov/25071839/
32. Wade KH, Lam BYH, Melvin A, Pan W, Corbin LJ, Hughes DA, et al. Loss-of-function mutations in the melanocortin 4 receptor in a UK birth cohort. Nat Med 2021 276 [Internet]. 2021 May 27 [cited 2021 Sep 24];27:1088–96. Available from: https://www.nature.com/articles/s41591-021-01349-y

ACKNOWLEDGEMENTS

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

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41366-021-01053-4.

Correspondence and requests for materials should be addressed to David Curtis.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.