Supplementary Information

Melanocortin-4 receptor pathway dysfunction in obesity: Patient stratification aimed at MC4R agonist treatment

Kristin L. Ayers[1,2], Benjamin S. Glicksberg[1], Alastair S. Garfield[3], Simonne Longerich[4], Joseph A. White[4], Pengwei Yang[4], Lei Du[4], Thomas W. Chittenden[4], Jeffery R. Gulcher[4], Sophie Roy[3], Fred Fiedorek[3], Keith Gottesdiener[3] Sarah Cohen[5], Kari E. North[6], Eric E. Schadt[1,2], Shuyu D. Li[1,2*], Rong Chen[1,2*], Lex H.T. Van der Ploeg[3*].

[1]Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[2]Sema4, a Mount Sinai venture, Stamford, CT 06902, USA

[3]Rhythm Pharmaceuticals, Boston, MA 02116, USA.

[4]WuXiNextCode, Cambridge, MA 02142, USA

[5]EpidStat Institute, Ann Arbor MI 48105, USA

[6]University of North Carolina, Chapel Hill, NC 27599, USA

[*]Corresponding authors: Dan Li <shuyu.li@sema4genomics.com>; Rong Chen <rong.chen@sema4genomics.com>; Lex Van der Ploeg <lvanderploeg@rhythmtx.com>

Supplementary Information

Supplementary Introduction

Obesity and MC4R pathway LoF variants

Obesity, characterized by an excessive accumulation of body fat has lacked effective pharmacotherapies, due principally to poor efficacy and/or corollary side-effects (1-3). Obesity significantly increases the risk of type 2 diabetes, cardiovascular disease, hypertension, and certain cancers, and is associated with elevated morbidity and mortality (4). As such, obesity has become an increasingly significant public health concern. Obesity prevalence varies among ethnicities with significantly higher rates in non-Hispanic blacks and Hispanics than non-Hispanic whites or non-Hispanic Asians (5). The hypothalamic MC4R-pathway plays a critical role in controlling food intake through brain-periphery axes, with the hypothalamus receiving, and responding to, signals from peripheral tissues including adipose tissue (e.g. leptin), pancreas (e.g. insulin) and gastrointestinal tract (e.g. ghrelin) (6,7). Genetic studies of several monogenic forms of early-onset obesity have revealed the significance of loss of function (LoF) mutations within the MC4R pathway, including leptin (*LEP),* the leptin receptor (*LEPR),* pro-opiomelanocortin (*POMC)*, prohormone convertase 1 (*PCSK1)* and the melanocortin 4 receptor (*MC4R)* gene (8-11). The main goal of this research is to thoroughly investigate several genes responsible for monogenic forms of extreme obesity that are upstream of MC4R, where individuals suffering from deficiency in these gene may be responsive to treatment with an MC4R agonist. We begin by selecting 3 candidate genes: *LEPR*, *POMC* and *PCSK1* genes, and categorize variants in these genes by likely functional impact and population frequency. We evaluate whether these variants are likely to impact phenotype by (1) doing extensive research into previous studies, (2) using tools to predict whether a variant is likely to affect protein functions, and (3) looking at the association of a variant with the phenotype. Once we collect this information, we attempt to estimate the likely number of individuals with monogenic forms of obesity for these 3 genes as a baseline count for existing individuals that can potentially benefit from treatment from an MC4R agonist. We also search a biobank for potential

candidates for treatment in a clinical trial to test the effect of setmelanotide on individuals harboring specific variants. This also gives us a sense of the yield we might expect to find by using existing biobanks to find individuals that can benefit from treatment. We then investigate other more common variants which may affect phenotype as these individuals may still benefit from treatment with an MC4R agonist. The long term goal is to identify additional genes and variants where individuals suffering from morbid obesity are likely to greatly benefit from treatment with an MC4R agonist where other treatments such as bariatric surgery, attempts at lifestyle changes in diet or exercise, or other obesity drugs have failed. The hope is that one day personalized medicine will allow these individuals to be identified by their genetics before they undergo suboptimal treatments that may place unnecessary harm on or cause discouragement in the patient.

To accomplish these goals, we first compiled and curated a comprehensive list of LoF variants in the *LEPR*, *POMC* and *PCSK1* genes, including variants reported in the literature, obesity pathogenic variants in human mutation databases, and computationally identified LoF variants based on predicted functional impact. Leveraging this list against the Genome Aggregation (gnomAD) database we estimate the prevalence of homozygous and compound heterozygous carriers, across the United States (USA). To further expand understanding of the genotype-phenotype correlation we next interrogated several large genotyped and sequence datasets with BMI measurements to assess the association of these LoFs variants (and other variants in these genes) with BMI and obesity, and carried out an allele burden test to evaluate if the cumulative number of affected alleles in these three genes is associated with a risk for increased BMI. Based on these findings we then queried the Mount Sinai (MtSH) Bio*Me* Biobank (12) for novel, informative MC4R-pathway variant carriers aimed at initiating phase-2 proof of concept clinical studies with setmelanotide.

Supplementary Materials and Methods

Genetic datasets

Below is a summary of the datasets used in the analyses.

Supplementary Table 4. The analyzed datasets.

Name (Abbreviation)	Number of Individuals	Type of Data	Individual Level data	Used in BMI analysis	Ethnicity
gnomAD	~120K (WES) ~15.5K (WGS)	WGS/WES	no	no	Mixed
1000 genomes	2500	WGS	yes	no	Mixed
UK10K Twins (UK10K)	1600	WGS	yes	yes	British Females
UK Biobank (UKBB)	150K	Genotype array + imputation	yes	yes	Only British analyzed
MtSH (BB)	11K	Genotype Array + Imputation	yes	yes	Mixed
(BM)	5K	Targeted Sequencing	yes	yes	Mixed

gnomAD

The publicly available gnomAD database (http://gnomad.broadinstitute.org/about) entails a compilation of datasets, from a variety of large-scale sequencing projects. The database contains information about variant frequencies in various race/ethnicity groups.

1000 genomes

The 1000 genome project (13,14), has individual whole genome sequencing data for 2504 individuals from 26 populations. In this analysis, we used the 1000 genomes to find allele frequencies in various ethnic populations.

MtSH Biobank

The Mount Sinai Hospital (MtSH) Biobank currently has collected DNA samples from more than 30,000 enrolled participants. Genotyping (BB) data has been generated for more than 14,000 patient samples using Illumina OmniExpressExome-8 v1.1 BeadChip that covers approximately one million genetic markers. Available clinical information for all the participants such as disease diagnosis, laboratory test results and medication history were obtained from Mount Sinai electronic medical record (EMR) databases. Individuals with discordant sex, call rates below 98%, or out-lying heterozygosity were removed. SNPs with call rates below 95% or with deviation from Hardy-Weinberg equilibrium (HWE) with p-value<5e-5 were also excluded. The genotype dataset was prephased using SHAPEITv2 (15) and imputation was performed with IMPUTE2 (16) using the 1000 genomes phase III integrated variant set as the haplotype reference panel. Imputed data and phenotype information was available for 11,091 individuals. Related individuals were also removed by randomly choosing one individual from approximate first or second-degree relatives (PI_HAT>0.25 in PLINK --genome analysis). BMI was recorded as the maximum reading for each individual. Unrealistic readings were excluded. Individuals with large shifts in BMI were examined and likely incorrect readings or elevated readings due to pregnancy were excluded. Principle components (PCs) we computed to estimate genetic ethnicity and clustering. We were left with 10,338 individuals with genotype, PC, and phenotype information for association analysis. Imputed variants were represented as estimated alternate allele count determined by the imputed genotype posterior probabilities. The Mount Sinai sequencing data includes (BM) 5543 individuals with sequencing data from a gene panel that included the *POMC*, *PCSK1*, and *LEPR* genes. As this is not WGS data, we are missing the information for the up/downstream variants. PCs were computed using a set of ancestry informative markers included in the gene panel. Children under the age of 18 and those missing reported ancestry were removed from any association analysis. For both datasets, ethnicity was estimated using reported race, reported ethnicity, country of birth, and genetic clustering. Individuals that did not have ancestry information or did not cluster with their reported ancestry in PC analysis were also removed from any ancestry informed analysis. There were enough individuals and information to do analysis for 6 ethnic groups: 1) Caucasians

of European descent (EA), 2) Caucasians of Jewish descent (EA_AJ), 2) Hispanic from Central or South American not identifying as white (HA_LAT), 3) Hispanic from the Dominican Republic (HA_DOM), 4) Hispanic from Puerto Rico (HA_PUR), 5) Black or African American from the USA (AA), and 6) Black or African American not from the USA with high African ancestry (AFR). Each of these groups has very distinct genetics, diets, and rates of obesity confounding analyses, which is why we chose to analyze them separately. For example, for those identifying as Black or African American, individuals born in the USA have morbid obesity rates around 14%, while those not claiming to be born in the USA have morbid obesity rates in the 6-9% range. Among Hispanic Latinos, individuals from Puerto Rico have a morbid obesity rate close to 10%, while those from the Dominican Republic have rates closer to 4%.

The UK Biobank

The UK Biobank (https://www.ukbiobank.ac.uk/about-biobank-uk/) is a national and international database, aimed at improving the prevention, diagnosis and treatment of life-threatening illnesses. The UK Biobank recruited 500,000 people (between 40-69 years of age in 2006-2010) from across the U.K. The first release of the UK BioBank consists of around 150K individuals with extensive phenotyping. The largest ethnic group consists of around 120K white British individuals with BMI information with genotyping information including selected coding variants. The UK Biobank was genotyped on two very similar arrays, the Biobank (BB) array and the BiLEVE (BL) array. PCs and genetic outliers were pre-computed (excludes those with a genotype missing rate >0.05 and heterozygosity >0.196) and an unrelated set of individuals was selected for analysis. Those without PC information or with missing genetic ethnicity were not analyzed. Many variants that were not genotyped have been imputed using a large reference panel of sequencing data. In this analysis, only the imputed variants that have high information scores (>0.7) are taken into account due to the inherent inaccuracy in using such analysis for rare variants. A high score implies they were reasonably well imputed.

The UK10K

Other data sets examined include the UK10K sequencing control data sets, including the TWINS 1692 female individuals and ALSPAC children. We did not analyze the ALSPAC data since the obesity rate was low and it is more difficult to examine BMI in children. Relateds (PI_HAT>0.2 in PLINK --genome analysis) and heterozygosity outliers were excluded from analyses.

Selection of variants

We first compiled and curated a comprehensive list of LoF variants in the *LEPR*, *POMC* and *PCSK1* genes, including variants reported in the literature, obesity pathogenic variants in human mutation databases such HGMD (17) and ClinVar (downloaded March 2017) (18) and computationally identified LoF variants based on the predicted functional impact. Group 1 variants consist of manually evaluated published variants based on the LoF criteria described in the main text. Group 1 also includes additional variants observed in the above data sets that are predicted to be nonsense, frameshift or splice site mutations. Group 2 variants consist of additional likely impactful missense variants seen in the above data sets, as determined by the criteria described in the main text. Only the UniProt canonical transcript was used for each gene to determine variant mutation type.

DeepCODE Methods for classifying Group 2 variants

We used a novel deep artificial neural network to predict functional relevance of missense variants in *LEPR*, *POMC* and *PCSK1*. This deepCODE deep learning algorithm was developed using high-confidence pathogenic variants curated by Clinvar (18), and another independent set of variants from the Exome Sequencing Project, ESP6500 (https://esp.gs.washington.edu/c) predicted to be likely benign based on their high allele frequency. Classification performance of the DeepCODE algorithm was near perfect (Supplementary Figure 1. AUC = 0.9933).

Fisher's Exact tests were then used to assess whether highly relevant (≥0.9) DeepCODE scores are more likely to occur in functional domains of these proteins. Domain annotations were obtained from NCBI for each of the three proteins, limited to domains with official designations (e.g. PFAM identifier) and strong literature evidence (see Supplementary Table 6). All observed rare variants with a maximum allele frequency of less than or equal to 1% (1089 variants collected from allele frequency databases including 1000 Genomes, ESP/EVS, Genomes of the Netherlands, DeCODE Iceland, gnomAD, and Kyoto Japanese) were mapped to protein positions. Variants with high (≥0.9) DeepCODE scores are significantly enriched in functional domains (p = 1.3e-03, 6.8e-07 and 1.9e-06 for *LEPR*, *PCSK1* and *POMC*, respectively, when all functional domains are pooled for each protein), indicating that variants with a predicted functional impact are more likely to occur in functionally annotated protein domains. We also assessed whether variants with high DeepCODE scores were enriched or depleted in each of 20 individual domains across the three genes; there were five individual domains with a significant association between the score and domain location. With the exception of the signal peptide domain of PCSK1, all domains are enriched with DeepCODE high-scoring variants (Supplementary Table 5).

Supplementary Table 5. Enrichment of variants with high DeepCODE scores in protein domains.

Protein	Domain	Protein region (amino acids)	p-value	BH adj. p-value
LEPR	Leptin Receptor/Ig-like C2-type	333-420	0.000652	0.000652
PCSK1	Peptidase S8/Subtilase family	158-432	1.30 e-14	6.5 e-14
PCSK1	Signal peptide	1-27	6.85 e-06	1.71 e-05*
POMC	Corticotrophin, ACTH, Melanotropin	136-176	0.00104	0.00519
POMC	Melanotropin gamma	77-87	0.01381	0.03451

*Statistically significant depletion of high-scoring variants in the signal peptide region as compared with the rest of the protein

DeepCODE variant-scoring model development

Two classification models were built for predicting the pathogenicity of human missense single-nucleotide variants (SNVs) across the genome (Yang et al., 2017, Manuscript in Preparation). Prediction scores from the deep artificial neural network and the Least Absolute Shrinkage and Selection Operator (LASSO) models are designated DeepCODE and lassoCODE, respectively. Here we only use the proposed DeepCODE model.

A deep neural network, "DeepCODE", was trained as described below to predict functional relevance of human missense single-nucleotide variants (SNVs). The algorithm was built using a non-linear deep neural network of 310 features derived from 59 of the 115 annotation columns from a published annotation resource, the Combined Annotation Dependent Depletion data set (CADD: http://cadd.gs.washington.edu/home; 19). Data sources for CADD (version 1.3) include ENSEMBL (v.75), variant-effect predictor (VEP, v.76), regulatory data from Encode, and missense prediction scores from Polyphen and SIFT. CADD C-scores for functional prediction were not used for training the DeepCODE DANN model. The model was trained with non-synonymous missense variants derived from the intersection of two data sources: (1) whole genome variants obtained from CADD, and (2) exonic coordinate regions for hg19 obtained from the UCSC genome browser. This classification scheme was trained and tested with a total of 2100 missense variants: 1050 missense variants from ClinVar (annotated by multiple labs as pathogenic), and 1050 common missense variants with allelic frequencies of 5 to 10%, randomly selected from the Exome Sequencing Project, ESP6500 (https://esp.gs.washington.edu/). The Clinvar "pathogenic" missense variants submitted by multiple labs served as "true values" for functional missense variants in the DeepCODE model. Similarly, the 1050 ESP6500 variants served as "true values" for neutral missense variants. For model training purposes, 80% of the 2100 total variants were used. The model was tested by predicting functional relevance for the remaining 20% of the total 2100 variants. The

DeepCODE model was evaluated with ROC curves and AUC metrics; the model had AUCs greater than 0.99 for both training and testing sets (Supplementary Figure 1).

The non-linear DeepCODE model was trained with a deep neural network in a CUDA-enabled GPU computing platform. The "lasagna" and "nolearn" python modules were used to construct the deep learning model with the "Theano" compiler. The neural network was initialized with an input layer, three hidden layers using the Rectify non-linear activation function for artificial neurons:

$$\varphi(x) = max(0, x)$$

and an output layer using the Softmax activation function:

$$\varphi(\mathbf{x})_j = \frac{e^{\mathbf{x}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}_k}}$$

where *K* is the total number of neurons in the layer. Stochastic Gradient Descent (SGD) was performed for parameter updates with Nesterov momentum (22) under the categorical cross-entropy loss function:

$$L_i = -\sum_j t_{i,j} \log\left(p_{i,j}\right)$$

where *t* is the target giving the correct class index per data point and *p* is the softmax output of the neural network with class probabilities. A dropout technique was applied to prevent neural networks from overfitting, as previously described (23). Model parameters such as the update learning rate, number of units, dropout rate and max epoch number were optimized by cross-validated grid-search over the parameter grid.

Variants in *LEPR*, *POMC* and *PCSK1*

Variants in *LEPR*, *POMC* and *PCSK1* from population allele frequency databases (including 1000 Genomes, ESP/EVS, Genomes of the Netherlands, DeCODE Iceland, gnomAD, and Kyoto Japanese) harmonized and maintained in GORdb (48) were obtained via Sequence Miner, a JAVA-based interface (WuXi NextCODE). These were cross-referenced to DeepCODE scores for all the missense variants. Variants were also mapped to protein domains. Domain information for the three proteins was obtained from GenPept records hosted at the NCBI (https://www.ncbi.nlm.nih.gov/protein/; LEPR - NP_002294.2; POMC – NP_000930.1; PCSK1 – NP_002294.2); overlapping domains were merged into one segment. Protein domain/segment boundaries were mapped to genomic locations using Alamut Visual (v 2.7.2) (Supplementary Table 6).

Supplementary Table 6. Protein Functional Domains.

Chrom	Genomic Start Pos	Genomic End Pos	Gene	Domain	Domain start (aa position)	Domain end (aa position)
chr1	66031249	66036176	LEPR	signal peptide	1	21
chr1	66062136	66064481	LEPR	FN3 (Fibronectin type 3 domain)	237	330
chr1	66067077	66067338	LEPR	"Lep_receptor_Ig"/"Ig-like C2-type domain; pfam06328	333	420
chr1	66067639	66070767	LEPR	Leptin-binding	467	484
chr1	66074441	66075756	LEPR	FN3 (Fibronectin type 3 domain)	537	627
chr1	66083655	66085649	LEPR	Fibronectin type III domain; pfam00041	741	812
chr1	66087062	66087128	LEPR	transmembrane region	840	862
chr1	66088602	66088626	LEPR	Box 1 motif	871	879
chr1	66101877	66101892	LEPR	Required for JAK2 activation	893	898
chr1	66101892	66101916	LEPR	Required for STAT3 phosphorylation	898	906
chr2	25383955	25384219	POMC	Lipotropin beta, melanocyte SF, ACTH, beta-endorphin, Op_neuropeptide, Met-enkephalin	179	267
chr2	25384228	25384348	POMC	Corticotropin, ACTH, melanotropin	136	176
chr2	25384495	25384525	POMC	Melanotropin gamma	77	87
chr2	25384546	25387560	POMC	NPP; Pro-opiomelanocortin, N-terminal region; pfam0838	28	70

chr2	25387566	25387641	POMC	Signal peptide	1	26
chr5	95728716	95728824	PCSK1	Proho_convert; Prohormone convertase enzyme; pfam12177	715	751
chr5	95730681	95734661	PCSK1	P_proprotein; Proprotein convertase P-domain; pfam01483	504	591
chr5	95735793	95759088	PCSK1	Peptidase_S8; Subtilase family; pfam00082	158	432
chr5	95761592	95768647	PCSK1	S8_pro-domain; Peptidase S8 pro-domain; pfam16470	34	110
chr5	95768668	95768746	PCSK1	Signal peptide	1	27

Prevalence Estimation

Leveraging the list of Group 1 and 2 variants against the Genome Aggregation (gnomAD) database, we estimate the prevalence of homozygous and compound heterozygous carriers, and across the United States. 40% of the first release of the gnomAD data of 60K individuals, known as the Exome Aggregation Consortium (ExAc) data are derived from US population and have similar estimated ancestry composition to US Census results (Supplementary Figure 3A). Similarly, for the full gnomAD database we have shown excellent concordance with the US Census results (Supplementary Figure 3B). Note however, while race/ethnic distribution in the gnomAD database appears representative of the US census, heterogeneity within race/ethnic groups has not been studied and may skew results obtained for a race/ethnic representation. We assume no negative selection of LoF variants even though there is evidence that loss of function in these genes may be associated with early lethality. In addition, new mutations or recurrent mutations will continue to arise in each generation which we do not account for. We assume random mating between obese and non-obese populations and between different races/ethnicities; however, non-random mating in these populations may increase the LoF frequencies within the obese population or in some races/ethnicities. Furthermore, because most of the variants are rare with MAF < 0.1%, we assume the mutations originally occurred on separate haplotype backgrounds. As the variants are in close proximity and it is rare for them to occur in the same individual, we also assume that recombination has not occurred between them since the original mutation event. The estimated number of homozygotes for a particular

variant is simply N*p^2, where N is the population size and p is the frequency of the allele of interest for the variant. The estimated number of compound heterozygotes for 2 variants is 2*N*p_1*p_2, where p_1 is the frequency of the allele of interest for the first variant, p_2 for the second variant, and the multiple of 2 from the fact that the first variant can be from either the mother or father and vice versa. The confidence intervals for each individual genotype can be approximated using formula for variance in Chakraborty et al (24). We roughly approximated the total variance as the sum of the variances.

We also computed the frequency and prevalence for non-random mating within each ancestry group assuming mating is restricted to individuals within the same race/ethnicity and use these frequencies to predict prevalence for non-random mating (NRM) across the whole population given the relative proportions of race/ethnic group. For the non-random mating within ethnicity calculations in Figure 3, we assumed the following proportion of individual ancestry from the gnomAD in the US (roughly estimated from a 2010-2014 censuses): African 12%, Ashkenazi Jewish 1.4%, East Asian 4%, Finnish 0.2%, European (not Finnish) 60%, Latino 16%, South Asian 3%, and Other/Unknown 3.4%. We report the prevalence for the different race/ethnicity groups as the prevalence per 100K individuals as some of these groups only comprise a small proportion of the population and would be not be visible in a bar plot if absolute numbers were used. This allows one to see which populations are enriched for carriers.

Association in Known, Novel Variants, and Genes

To further expand understanding of the genotype-phenotype correlation, we interrogated several large genotyped and sequence datasets with BMI measurements to assess the association of the Group 1 and 2 variants (along with other variants in these genes) with BMI and obesity, and carried out an allele burden test to evaluate if the cumulative number of affected alleles in these three genes is associated with a risk for increased BMI. Association of

variants with BMI was done by regressing BMI on age, age squared, gender, and 6 genetic principle components which help control for BMI and variant frequency differences in populations with substructures (for example, there may be differences between BMI and variant frequencies for those from the South versus those from the Northern UK) which can lead to false positive or false negative associations. When the two UK Biobank arrays were analyzed together, we also included a covariate for the array (Biobank array versus BiLEVE array which may have subtle differences in genotyping error or missing rates). For the UK Biobank, we computed obesity odds ratios by dichotomizing cases and controls to morbidly obese cases with BMI>40 and normal controls with BMI<25. Logistic regression was used to analyze the common variants while firth regression was used for rare variants. BMI is influenced by many environment and genetic variants. Due to confounding (reported race/ethnicity associated with trait, genetic burden associated with genetic race/ethnicity, and genetic PCs associated with trait and reported race/ethnicity, trait associated with sex differently in different races/ethnicities, etc.), we analyzed each reported race/ethnicity individually and combined results. We did not log transform BMI to retain the original scale. (The ENGAGE study performed in inverse normal transformation of BMI, Supplementary Table 11). The meta-analyses and forest plots were done in the R package metaphor using a random-effects weighted model (Wolfgang Viechtbauer, 2010 36:3. Jstatsoft: https://www.jstatsoft.org/article/view/v036i03). The forest plots represent the estimated effect sizes and confidence intervals for each data set and subgroup. Each dataset is assigned a weight based on the inverse of the variance in the estimate and thus larger or more homogenous data sets tend to have larger weights. The overall effect size and p-value are a weighted average of these studies. A test of heterogeneity is also performed to determine if the different studies have significantly different estimated effects (the I^2 statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance). Beta reflects the BMI shift for the carriers of the variant set. For imputed variants (genotypes for unsequenced variants estimated from a reference set of individuals), only variants with an IMPUTE information score >0.7 were analyzed. For testing for a heterozygote carrier, an individual was required to have an allele dosage ≥0.9. For tests for carriers of two or variants or

homozygotes, and individual needed an allele dosage sum ≥1.8. An individual was only considered a non-carrier (WT) when the sum of their allele dosages over all variants tested was <0.5. For all rare variants, a missing variant was considered as homozygote for the major allele for any burden test. For tests involving *PCSK1* N221D or S690T, any missing genotypes were imputed.

For the association analyses, we only analyzed the 120K British individuals from the UK Biobank, as this was the largest ethnic population. This analysis has several limitations: First working in a selected ethnicity can obscure BMI association signals due to the presence of interacting alleles, prevalent in this population. Secondly, due to population specific prevalence distributions, many alleles will not be represented. In line with these considerations, we noted that most of the Group 1 and Group 2 variants were too rare to test for association with BMI even for the heterozygotes.

To determine if the common variants reported in literature were true associations and/or the possible causal locus, when available, we also looked at the reports of association of the known variants in two European genome-wide meta-analyses consortia: GIANT (25) and ENGAGE (26) (http://diagram-consortium.org/2015_ENGAGE_1KG/) and in the UK Biobank cohort (27). The GIANT data set consists of around 124K genotyped individuals imputed with the HapMap data, while the ENGAGE data set is comprised of around 87K genotyped individuals imputed on the 1000g phase I.

We then hypothesized, that an aggregation of known rare (or both rare and common) variants in these genes within an individual would result an elevated BMI. Previous studies have looked at the effect of the overall burden of BMI associated variants common in obesity related traits (28) and in the GIANT study with the identified risk alleles (25). A combined analysis of common SNPs along with copy number variation was also found to be significantly associated with BMI (29). In addition to published variants, we determine which group of variants best correlate with BMI and obesity from available datasets and identify carriers of these variants as potential patients for a clinical trial.

LoF Carriers in Mount Sinai BioMe Biobank

Based on these findings we then queried the Mount Sinai (MtSH) Bio*Me* Biobank (12) which has a high percentage of obese (39% with BMI>30) and morbidly obese (8% with BMI>40) patients, for informative MC4R-pathway variant carriers with the future goal to initiate a phase-2 proof of concept clinical studies with setmelanotide in selected MC4R-pathway deficient patients. As LoF/LoF are very rare, we are also particularly interested in those carrying LoF/PLoF variants, especially those that are homozygous or compound heterozygous within a gene, or those that are what we will call composite, i.e. compound heterozygotes across genes. However, it may be that some heterozygotes or individuals carrying common variants will have increased response to treatment as well.

Supplementary Results

For *POMC*, *PCSK1*, and *LEPR* we identified 30, 31, and 22 potential variants in the literature, respectively (Supplementary Tables 1A, 2A and 3A). For references where only the amino acid change position was reported, we searched the available cohorts for each possible change (Supplementary Tables 1-3). We were unable to identify a handful of the variants from the literature with the given notation. In total, we found 587 known and predicted LoFs, including 135, 231, and 221 in *POMC*, *PCSK1* and *LEPR* respectively (Supplementary Table 1-3, Table 1). These include 166 Group 1 variants and 421 Group 2 variants. There were 104 Group 1 and 392 Group 2 variants in the gnomAD data set. There were 43 Group 1 and 67 Group 2 variants in the datasets analyzed for BMI. 121 of the Group 1 variants were observed in any data set analyzed. (These counts do not include T640A).

Prevalence Estimation

The cumulative allele frequencies for the Group 1 and Group 2 variants are presented in Supplementary Figure 4.

Impact of Ethnicity

The mating preference considerations do not greatly impact the overall U.S. prevalence estimates (maximum absolute change roughly 25%), and the true prevalence and mating preferences likely lie between the random mating and non-random mating values. Due to several variants with modest prevalence in only one or several ethnicities, the prevalence of homozygous and compound heterozygous carriers is predicted to differ significantly among ethnicities and may help identify target populations.

For example, while *POMC* has high frequency of LoF carriers and predicted homozygous deficient populations (Supplementary Figure 4A) of European, African, and Ashkenazi Jewish ancestry, the number of carriers in East and South Asians approaches zero. By contrast, the South Asian population has the highest prevalence of homozygous and compound heterozygous carriers of LoFs in *PCSK1* (Supplementary Figure 4B). Some of these differences can be attributed to a few specific variants. For example, in POMC, the Ashkenazi population has higher frequencies of both R236G (Group 1, MAF=0.8%) and H143Y (Group 2, MAF=0.3%) than many of the other populations. The Group 1 R80Q variant in *PCSK1* has a frequency of 2.8% in East Asians, but is rare in other populations with MAF < 0.1%.

Rare Variant Analysis

We use the 120K British individuals from the UK Biobank to analyze the impact of variants within these 3 genes on BMI. However, because the UK Biobank is not nucleotide sequencing data but genotyping data (direct identification of predefined single nucleotide polymorphisms), it will be missing many important rare variants and indels (MAF <0.1%). The sheer size of the data set allows us to test for evidence of associations to BMI for variants in the 0.1-0.5% MAF range where there are several hundred heterozygous individuals. For each analysis, we report (1) the average increase in BMI (effect size or β) and (2) the odds ratio (OR) for BMI<25 vs BMI>40 for the heterozygous (or homozygous) individuals versus those who are wildtype (WT), i.e. do not have the variant. Of the Group 1 and 2 variants, none stand out as statistically

significant at the α=0.05 level, except the well-studied *PCSK1* N221D variant (Supplementary Figure 5B, 6A, 6B), though some variants are correlated with an increase in BMI (a summary of the effects sizes and ORs for the Group 1 and 2, and other rare variants, see Supplementary Figure 5). Analysis of additional rare missense variants found a potentially novel LoF variant in *PCSK1*, T640A (MAF 0.3%) significantly associated with BMI (Supplementary Figures 5B and 6C). Since the effect sizes for N221D and the newly identified T640A variant were consistent across other populations as well and the overall effect was strongly statistically significant (Supplementary Figures 5B and 6), these variants are included in relevant MC4R-pathway allele burden tests.

PCSK1 T640A

There was only one homozygote, a non-British individual with a BMI just over 40. The effect of this variant does not seem to be caused by nearby significant genotyped SNPs and based on the observed genotype frequencies, the variant does not appear to be on the N221D or S690T/Q665E haplotype. The frequency of *PCSK1* T640A variant is 0.3% in the UK Biobank British individuals but higher in Puerto Ricans where the MAF is closer to 1-2% in the MtSH Biobank, and is also higher for some of the South American 1000 genomes populations

PCSK1 R80Q

The variant that is driving the high prevalence in *PSCK1* in East Asians (EAS) discussed above, R80Q (30), (Group 1 variant) has frequency 2.8% in EAS in gnomAD (mostly seen in Chinese and Vietnamese individuals in the 1000 genomes (see Materials and Methods)) and is rare in other populations (<0.1% in the gnomAD data). A functional study saw a 30-40% reduction of enzyme activity due to this variant (30). In the UK Biobank Chinese population, an association test supports an increase in BMI, with an average BMI increase (effect size) of 2.91 with p-value 0.015. Of course, this is a small sample size and this abundant variant in this ethnic group is based on imputed data with information score just below our cutoff, but it is reassuring that this variant might impact the BMI phenotype.

POMC β-MSH, including *R236G*

Some POMC β-MSH alleles have been correlated with obesity in prior studies, including 2 studies in human and a study in obese Labrador dogs (31-33). We identified several β-MSH variants, including R236G (a weak partial MC4R agonist), Y221C (imputed) and E206X, but in our analysis, could not confirm an association with obesity in the UK Biobank (Supplementary Figure 5A). If these variants are indeed LoF, with a causative role in severe obesity, our lack of detection may be due to limited sample size and a less obese UK Biobank population (Supplementary Figure 12). None of the 7 identified R236G homozygotes have BMI> 40. However, these individuals are not available for resequencing to confirm these results. Considering the biochemical evidence supporting the reduced functionality of the R236G variant (34) we are currently testing whether R236G allele carriers respond uniquely to setmelanotide, in order to determine their functional relevance in the MC4R-pathway. In addition, based on evaluation of biobanks with > 1400 severely obese and hyperphagic children (data not shown), we have identified several severely obese R236G homozygous deficient patients. Their analysis is ongoing. Several rare POMC alleles in heterozygous form (E105X and E206X) were correlated with increased BMI (Supplementary Figure 5A). Overall the effect size of heterozygous POMC deficiency for the other alleles identified, could not be associated with obesity in this cohort.

In the UK Biobank data, the R236G frequency differs regionally within the UK, as does BMI. The allele frequency reaches as high as 0.82% in the northwest of England and as low as 0.43% in the southwest. However, the variant is still not associated with increased BMI when we analyze these regions separately. R236G appears to be on the same haplotype as rs934778 in the UK Biobank, a common allele very significantly associated with BMI (Supplementary Figure 7A).

Compound Heterozygotes

We also analyzed the Group 1 and 2 potential compound heterozygotes for each of the three genes. We evaluated effect size and odds ratio (OR) in UK Biobank of individual either homozygous or compound heterozygotes for Group 1 plus Group 2 variants (Supplementary Figure 7A and 7B). The effect size β and OR was most significant in the *PCSK1* N221D/T640A compound heterozygote, with an estimated OR of 7.6 (SE=0.59) and p-value of 0.003. Including

T640A in the US, *PCSK1* prevalence estimate adds over 33,000 individuals, mainly in the form of N221D/T640A compound heterozygotes. In *PCSK1*, the genotype counts for the combination of the two variants in the entire UK Biobank for N221D/T640A, Y181H/M125I, M125I/N221D, T175M/N221D, E250X/N221D and S307L/N221D all provide support that these variants are on different haplotypes and thus true compound heterozygotes (i.e. the compound heterozygote frequency is consistent with the estimate assuming the variants are on different haplotypes). However, we cannot rule out that a rare recombination event may have taken place in an individual. Additionally, T640A/S690T carriers also seem to be true compound heterozygotes. The *POMC* Group 1 and 2 compound heterozygotes were too rare to draw any conclusions, and we did not find any *LEPR* compound heterozygotes (though there were few genotyped LoF variants in LEPR). In *POMC*, all R236G homozygotes are also rs934778 homozygotes implying that these variants are on the same haplotype in these individuals. Thus, it is likely that most of the individual carrying 1 copy of each variant are not compound heterozygotes and the expected and observed counts of the genotype combinations support R236G occurring on the rs934778 haplotype the majority of the time.

For common variants, we recognize that without trios, we cannot determine at an individual level if a common/rare variant combination is compound heterozygous. However, when we analyze population data under the assumption that the two variants lie on different haplotypes, the population counts for carriers of both variants either agree or disagree with the population expected counts. When these counts agree, it is most likely that all or nearly all individuals are true compound heterozygotes. For example, in the case of PCSK1 T640A (a rare variant) and S690T (a common variant), given the allele frequencies of these 2 variants, the expected number of individuals that would carry both these 2 variants would be 199 in UK Biobank if they were not on the same haplotype. If they were on the same haplotype, we would expect this number to be closer to 717, the total number of T640A carriers. We observed 189 individuals carry both T640A and S690T, strong suggesting they are very likely compound heterozygotes. On the contrary, as described above that in POMC, all R236G (a rare variant) homozygotes are also rs934778 (a common variant) homozygotes implying that they are on the

same haplotype. We rationalize that most of the individual carrying 1 copy of R236G and 1 copy of rs934778 are unlikely compound heterozygotes, and the expected and observed counts of the genotype combinations support R236G occurring on the rs934778 haplotype the majority of the time.

Burden Test

For most of the analyses, the results are dominated by the UK Biobank results due to the large sample size (assigned high weights in the meta-analyses in Figure 4 and Supplementary Figure 6) and small data sets are included mainly for the purpose of visualizing the effect size trend across the datasets and populations. Below is a table of the results, including for *PCSK1* compound heterozygotes (Forest plot not shown).

Supplementary Table 8. LoF allele burden test.

Single variant (S) or two or more variants in 1 or more genes (M)	**Analysis - high level description**	**β (p-value)**	**Comparator**

S	PCSK1 N221D Het	0.18 (0.010)	WT
S/M	> 1 allele Group 1 or 2 alleles	0.35 (0.012)	WT
M	*PCSK1* N221D Het, plus one Group 1 or 2 allele vs N221D Het. (Excluding N221D Hom contribution)	0.57 (0.058)	N221D
M	>2 Group 1 plus Group 2 alleles in different genes (composites)	0.52 (0.098)	WT
M	>2 Group 1 or 2 alleles	2.79 (0.016)	WT
M	*PCSK1* compound heterozygotes	1.75 (0.005)	WT
S	N221D homozygous	3.22(0.047)	WT

*Note that these estimates have large confidence intervals.

Additional Composite analysis in Common Variants

To provide additionally evidence that variants in different genes can contribute to disease burden more than a single heterozygous variant, we looked at the composite effect in two common variants (MAF ~30%) where we have enough composites to have statistical power to detect composite effects. Using the two most statistically significant BMI associated variants in the UK Biobank individuals genotyped on the Biobank array (UKBB_BB): *PCSK1* rs6235 (S690T) and *POMC* rs934778, we compare the effects of: 1) heterozygous for one variant; 2) heterozygous for both variants; 3) homozygous for 1 variant: 4) homozygous for one variant and heterozygous for the other; and 5) homozygous for both versus WT for both. In this example, the effect for genotypes in # 2 (heterozygous for both variants) seem to be intermediate that of #1 and #3, and has a larger effect than either single variant effect. In the UK Biobank the case of homozygosity for both variants at both genes versus wildtype at both genes, shows a significant effect size of 0.85 and this result is corroborated for those genotyped on the UK BiLEVE (UKBB_BL) array (Supplementary Table 9, Supplementary Figure 8A).

Supplementary Table 9. Composite analysis of PCSK1 rs6235 and POMC rs934778.

Genotype Combo	UK Biobank Array			UK BiLEVE Array		
	Effect Size β	P-value	Bin Size	Effect Size β	P-value	Bin Size
1	0.17	6.97e-5	32716	0.08	0.192	10110
2	0.25	1.44e-6	13494	0.15	0.056	17055
3	0.32	1.20e-6	6785	0.29	0.002	7071
4	0.41	7.90e-09	5639	0.24	0.024	2916
5	0.85	2.11e-5	566	0.64	0.024	321

To more broadly evaluate MC4R-pathway allele burden in the UK Biobank, we looked at the effect of carrying 2 or more Group 1 or Group 2 variants versus WT by BMI category. Please note this analysis includes homozygotes, compound heterozygotes and composites among any of the genes being analyzed. In Figure 5 and Supplementary Figure 9 when the Group 1 plus Group 2 alleles, (including the Group 1 alleles, *PCSK1* N221D and T640A), are included in the burden test, the carriers of 2 or more alleles had significantly higher numbers of morbid obesity patients (BMI>40) when compared to the low BMI (BMI<20) population (p=0.0008, Figure 5; specifically: 8/2680 vs. 26/2410, or about 3 times higher than expected). This effect appeared to be larger in males than in females (Supplementary Figure 9A and 9B).

Burden of Rare and Common Variants

Our data on common variants in the MC4R-pathway indicate that it will be worthwhile evaluating to what extent these common alleles can contribute to severe obesity when in combination with Group 1 or 2 alleles and thus predict sensitivity to MC4R-pathway supplementation therapy, to mitigate severe obesity and its comorbidities. While these common alleles are predicted to predispose to higher BMI it is clear, that taken in isolation these highly prevalent alleles cannot significantly contribute to the predisposition to severe obesity. We, therefore, aim to analyze setmelanotide in patients with a high total allele burden

in severe obesity, starting with the alleles that are proven to contribute to severe obesity (presence of Group 1 and/or Group 2 alleles) and evaluate these when superimposed on more common alleles. For example, there is a severely obese individual in the UK Biobank with *POMC* E105X who is also homozygous for rs934778 (Supplementary Figure 7A). Here we show how a combination of rare and common variants could be used to compute a genetic risk score for an individual and identify individuals with high risk of obesity due to their genetic risk in these rare and more common variants of these MC4R pathway genes.

One way of choosing patients would be to find those with the highest rare and common variant burden or risk over the 3 genes. As an example, for the UK BioBank individuals, we construct a weighted risk score, known as polygenic risk score (PRS) based on the estimated effect size β, estimated from the Group 1 and 2 variants and for several common variants from independent association signals. In the UK Biobank (for the individuals genotyped on the UK Biobank chip), we estimate the effect size β for each of the variants as a weight (we required at least 5 individuals to carry the variant on the UK Biobank chip in order to have a more informative weight). This is a weighted risk score constructed from a linear model. It gives a higher weight to variants that have larger impact, and negative weight to protective alleles. Then, for each individual i, we simply sum their genotype $G_{i,j} = \{0,1,2\}$ at each locus j times the weight of that genotype over all L variants in the model. The PRS for an individual is:

$$PRS_i = \sum_{i=1}^{L} \beta_j G_{i,j}$$

This will give us a genetic risk score for the estimated increase in BMI for an individual due to their genotypes. Those with a higher score are expected to have a higher BMI based on their genetics. We then binned individuals into quantitative risk groups based on their score and computed the effect sizes by risk group (Risk Bin 3 being the highest risk). We applied the estimates from the Biobank chip to the individuals on the UK BiLEVE chip and to the UK Twins. In Supplementary Table 10 below, we demonstrate that the correlation between the risk bins and BMI constructed from one British data set holds in the two other British datasets. However, we do not expect this to hold with other races/ethnicities due to differences in LD patterns

between the associated common variant and whatever the true causal variant(s) may be. The idea is that those individuals with the highest risk are most likely to have variants in these 3 genes that affect their BMI. These types of scores could be used to help prioritize the patients selected for treatment.

Supplementary Table 10. Polygenic risk scores associated with BMI.

Risk Bin	UK Biobank Chip			UK BiLEVE Chip			UK Twins		
	Effect Size β	P-value	Bin Size	Effect Size β	P-vale	Bin Size	Effect Size β	P-value	Bin Size
0	-	-	12499	-	-	6543	-	-	237
1	0.18	1.98e-4	44285	0.13	0.068	22999	0.37	0.31	879
2	0.38	1.85e-13	21373	0.21	0.006	11342	0.65	0.11	438
3	0.97	5.17-08	732	0.78	0.007	318	1.56	0.28	13

*Note: Bin 0 is the low risk baseline bin. Effect sizes are estimated against these individuals.

Based on these analyses we conclude that in the MC4R-pahtway increased allele burden due to the accumulation of diverse loss of function alleles, can significantly predispose to a higher BMI. This in turn, forecasts that patients with a higher MC4R-pathway allele burden may show increased sensitivity to pharmacotherapy aimed at restoring MC4R agonist tone. As more studies are done, weights could instead be estimated based on patients' response to treatment.

Extensive Review of Common variants

Many of the common variants in these three genes have been heavily studied in published genetic association studies for BMI and other anthropometric traits. We evaluated associations and hence benchmarked our populations under analysis against published literature by reviewing the summary statistics from the ENGAGE (European Network for Genetic and Genomic Epidemiology) (26) and GIANT (Genetic Investigation of Anthropometric Traits) (25) meta-analyses for the common variants with MAF > 5% for variants within any of the three genes. These studies include data from many cohorts, aggregated to make up datasets with 87K and 124K individuals, respectively. For these 2 studies, we did not see strong evidence for

association of any of these known common variants with BMI (Supplementary Table 9). Additionally, we analyzed common variation in the British individuals from the UK BioBank.

In association studies, the most associated variant at a locus is not necessarily within the gene nor does it carry a known functional mechanism explaining the association. Generally, it is assumed that these variants are in linkage disequilibrium (LD) with functional variants (rare variants are not generally genotyped in these studies) or that there are haplotypic effects. We suspect this may be true for many of the reported common SNPs for obesity in the literature. Manhattan plots for the UK Biobank for all 3 genes are shown if Supplementary Figure 10 (including 5kB up/downstream).

POMC gene-wide BMI association

The *POMC* locus reached genome-wide significance (p-value<$5x10^{-8}$) for an impact on BMI in both the ENGAGE and GIANT studies and near genome-wide significance in our analysis. The lead SNP near the *POMC* locus in the ENGAGE meta-analysis was an intergenic SNP rs6749422. In our analysis of the UK BioBank, the intronic variants rs934778, rs1009388, and rs6713532 are significantly associated with BMI with p-values < 10^{-6} (Supplementary Figure 10, 11A). rs1009388 is imputed and is in high linkage disequilibrium (LD) with the genotyped rs934778 (r^2=0.8 in Europeans). rs934778 has a frequency around 30% in Europeans, but closer to 5% in Africans. The rs934778 alternate allele appears to be on the rs6713532 reference allele background and vice versa (r^2=0.16 in Europeans) and has a frequency near 20% in Europeans but closer to 50% in other populations. rs6713532 has previously been reported to be associated with waist-hip ratio, visceral fat and abdominal fat in Dutch males (35). rs934778 and rs6713532 are modestly associated with BMI in the GIANT and ENGAGE datasets as well, with p-values 0.004 and 0.008 in GIANT and 0.016 and 0.017 in ENGAGE respectively. The 3' UTR rs1042571 alternate allele (MAF~3-30%, 20% in Europeans) often occurs with the rs934778 alternate allele (r^2=0.51 in Europeans), though rs934778 is more common. *POMC* rs1042571 was modestly associated with BMI in the UK BioBank (p-value near 0.01), and has been associated with obese (BMI ≥ 30 kg/m^2) versus non-obese individuals (BMI < 30 kg/m^2) in a

small population of North Indians (36) and in a small population of European Americans (37). Currently, it is difficult to determine which of these common variants if any, are driving a putative association signal at this locus, but analysis in additional populations and ethnicities can aid in fine mapping of this region.

PCSK1- gene wide BMI association

The *PCSK1* locus was not reported as genome-wide significant in the ENGAGE or GIANT dataset. However, consistent with our ethnicity analysis, the *PCSK1* locus did reach genome-wide significance in a meta-analysis of nearly 28K East Asians (38). There is a set of variants in *PCSK1* that showed some evidence of association with BMI in the GIANT dataset, though these variants did not reach the genome-wide significance level. The 3 variants in *PCSK1*: rs6234 (Q665E), rs6235 (S690T), and rs6232 (N221D discussed above) have been heavily studied for implication in obesity and adiposity, and show association in Caucasians but mixed results in other ethnicities (39-43). Functional analysis showed the rs6234/ rs6235 pair elicits a non-significant 6% decrease in enzymatic activity of PCSK1, while rs6232 resulted in a 10% reduction in activity versus wild type (44). rs6234 and rs6235 are the more common variants (frequency around 0.23-0.29) and tend to be in high LD with each other. rs6232 is rarer (MAF ~0.01-0.05) and is also in LD with the other pair, especially in individuals of European descent. The largest study on these variants consists of more than 300K individuals from more than 30 cohorts of diverse ethnicity and provides convincing evidence that these variants have a small but real effect on BMI and obesity (45). Thus, it may be worth considering the effects of an MC4R agonist treatment for carriers of these variants in a severely obese population. While these variants might not have large effects in the overall population, they may be having a large impact on some obese individuals with a predisposing genetic background. We also performed a standard association test (under an additive model or genotype model) on the common known variants with MAF>0.05 in the UK Biobank (Supplementary Figures 11B). Most of the previously reported variants showed little evidence of association with BMI, but rs6232/4/5 are significantly associated with BMI and obesity in the UK Biobank as well as rs271919, rs271939, rs3811942, rs3811951, rs155981, rs155982 and the rare variant rs139453594. rs6232 is

associated with BMI in most of the datasets examined (Supplementary Figures 5B and 7). In European populations, rs6234 and rs6235 are in nearly perfect LD with each other and are in high LD with the intronic variant rs3811951 (r^2=0.84). In the UK Biobank, the effect of rs3811951 is completely mitigated in non- rs6234/5 carriers. rs3811951 also was modestly significant in the GIANT dataset as well, likely due to LD. rs6232 is usually on the same haplotype as rs6234/5. We repeated the rs6234/rs6235 association removing all rs6232 carriers and found the effect size and association still held, implying these 2 signals carry independent effects, as previously seen. rs6232 has a prevalence around 0.05 in European populations, but closer to 0.01 in African populations. rs6232 also tends to occur on a rs271919/rs271939 GC background, and often occurs with the rs155982 G allele in the UK Biobank British individuals. The intronic variant rs271919 and downstream variant rs271939 are in nearly perfect LD with each other and in LD with rs3811942 (r^2=0.65). However, in the UK Biobank, the effects of rs155982 with either rs6232 or rs6234/5 seem to be independent or additive signals. The rare variant rs155981 (minor allele A, MAF=0.01) tends to occur with the rs155982 minor allele G, but rs155982 is likely responsible for any signal from rs155981 in the UK Biobank (these variants are also modestly associated with BMI in the UK Twins). The effect of rs271919 disappears when we exclude carriers of rs6232/4/5, rs155982, and rs139453594 and is likely driven by the LD patterns with these variants.

LEPR- gene wide BMI association

The *LEPR* locus was not reported as significant in the ENGAGE or GIANT study. For the ENGAGE, GIANT, and UK Biobank, few, if any, variants within these genes achieved a p-value below 0.01 (Supplementary Figure 11C). It may be that there are no common variants with even small effects on BMI in this gene.

Results from additional studies of rs6232/rs6234/rs6235

Kilpeläinen et al., examined rs6232 (N221D) and rs6235 (S690T) SNPs in 20,249 individuals of European descent from Norfolk, UK (40). They found neither of the SNPs was significantly associated with obesity, BMI or waist circumference under the additive genetic model (P > 0.05)

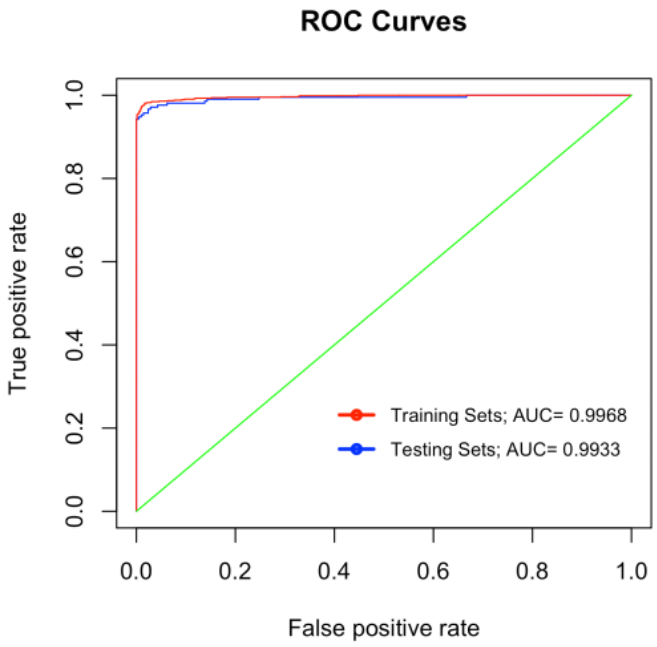
but observed an interaction between rs6232 and age on the level of BMI (P = 0.010) and risk of obesity (P = 0.020). The rs6232 SNP was associated with BMI (P = 0.021) and obesity (P = 0.022) in the younger individuals [less than median age (59 years)], but not among the older age group (P = 0.81 and P = 0.68 for BMI and obesity, respectively).

Choquet et al., using subjects from CARDIA (8,359 subjects), found that in European-American subjects, only rs6232 (not rs6235) was associated with BMI (P = 0.006) and obesity (P = 0.018) but also increased the obesity incidence during the 20 years of follow-up (HR = 1.53 [1.07-2.19], P = 0.019) (39). They also found that alternatively, in African-American subjects from CARDIA, rs6235 (but not rs6232) was associated with BMI (P = 0.028) and obesity (P = 0.018). However, due to the low frequency of this variant in Africans, it would be unlikely to find association in a study of that magnitude. In a Mexican -Mestizo population of 2382 individuals, Villalobos-Comparán et al found rs6232 was significantly associated with childhood obesity and adult class III obesity (OR = 3.01, 95%CI 1.64–5.53; $P = 4\times10^{-4}$ in the combined analysis) (41). In contrast, rs6235 showed no significant association with obesity or with glucose homeostasis parameters in any group.

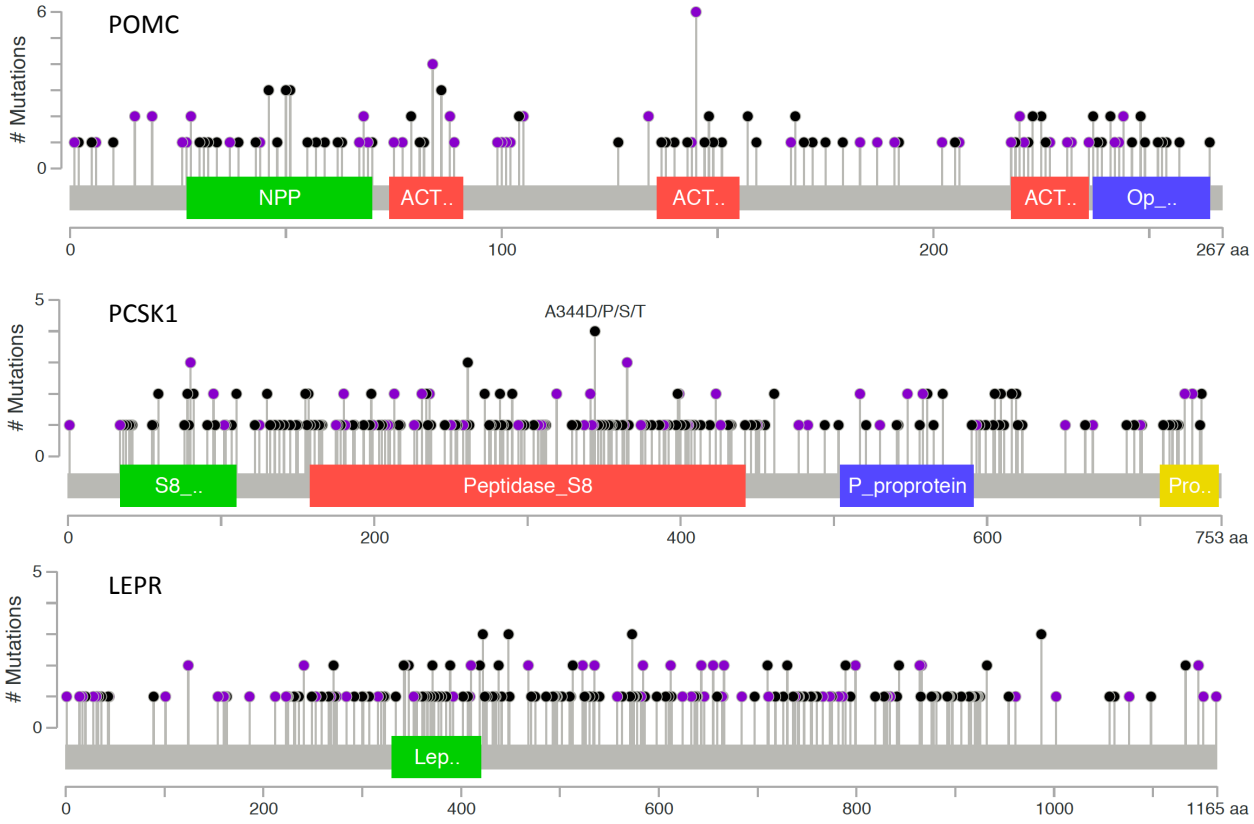
It could be that rs6234/rs6235 are sometimes detected since they are in high LD with rs6232 but are more common (though in the UK Biobank, the effect of rs6234/rs6235 remains after removal of carriers of rs6232). It may also be that the rs6232/rs6234/rs6235 has a larger effect that any one variant alone, but this is harder to untangle due to the high LD. The LD may be lower in non-European ethnicities and we could potential use these groups to determine interactions. rs6234/rs6235 are also in high LD with rs17085675 and rs2882298 (44).

Supplementary Figures

Supplementary Figure 1. Receiver Operator Characteristic (ROC) Curve. The ROC curve indicates near-perfect classification of functional relevance for human missense variants. This model was used to score variants within LEPR, POMC and PCSK1. Note no significant statistical shrinkage (overfitting) observed between training and testing of the deepCODE model.

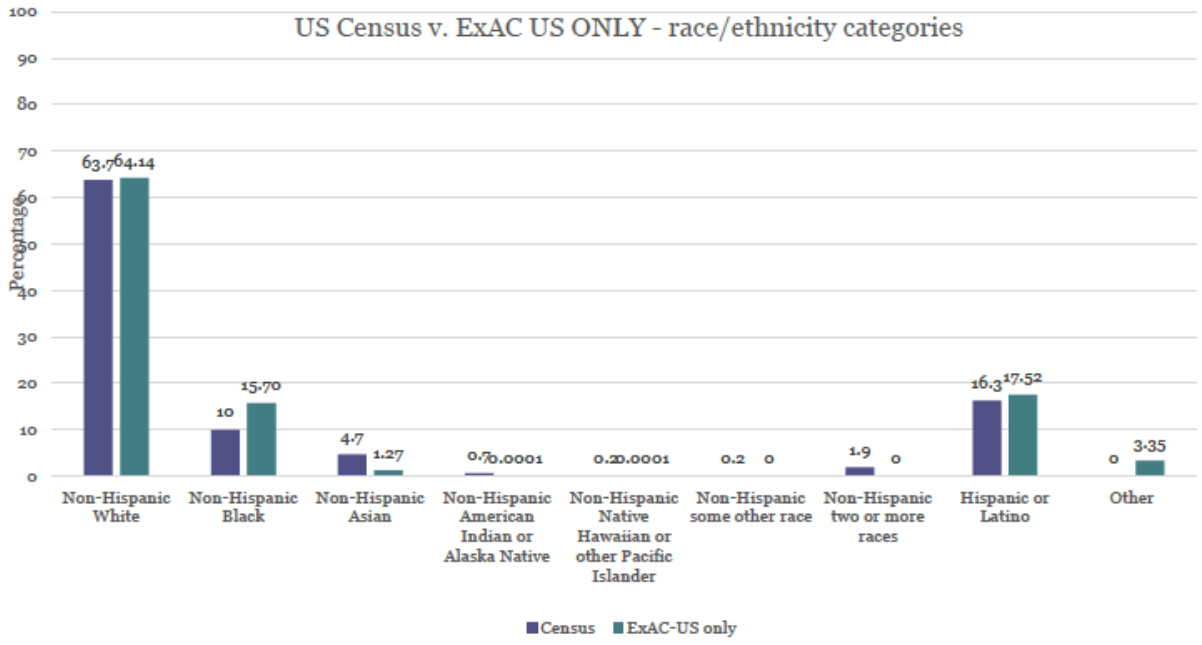


Supplementary Figure 2. Distribution of Group 1 (purple) and Group 2 (black) variants in POMC, PCSK1 and LEPR coding regions. The lollipop plots were generated using cBioportal's MutationMapper tool (http://www.cbioportal.org/mutation_mapper.jsp). Protein domains shown in the plots are PFAM domains based on UniProt annotation (www.uniprot.org/).

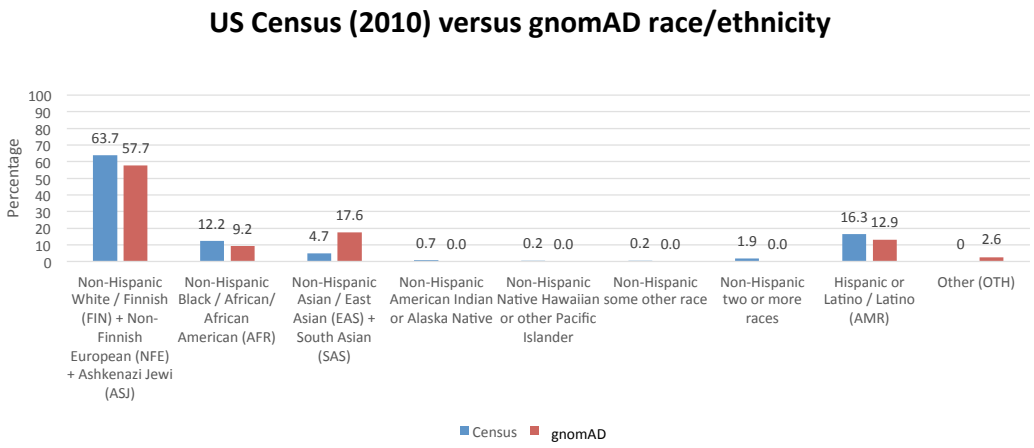


Supplementary Figure 3. Ethnicities in US Census vs. the ExAc cohort.

A



B

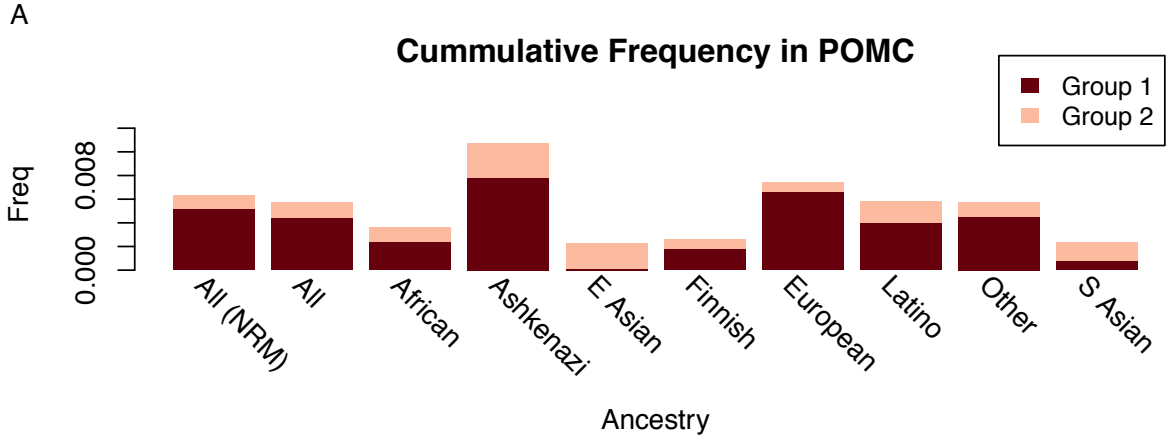


Note: Entire ExAC included (N=60,706) within gnomAD)N=141,352.

.

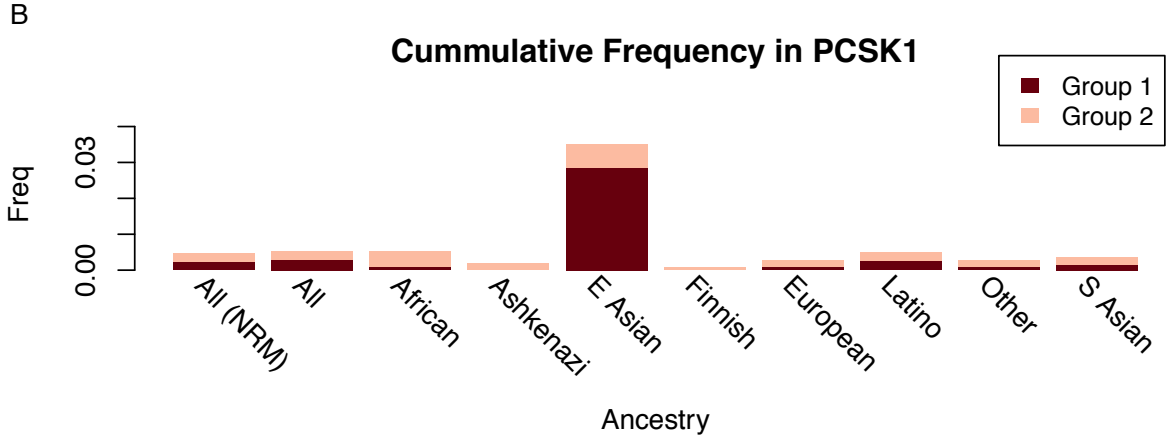
Supplementary Figure 4. Cumulative allele frequencies of Group 1 and Group 2 variants in *POMC* (A), *PCSK1* (B), and *LEPR* (C) in various ethnicities along with the list of Group 1 variants. Variants listed red font occur more than once in the gnomAD data. The 'All' group is based on the overall gnomAD allele frequencies, while the 'All (NRM)' group is based on non-random mating within ethnicities given the estimated percentage of individuals within each ethnic group in the US. Group 1 and Group 2 variants are listed below each plot. Those in red occur in more than one individual in gnomAD.

A



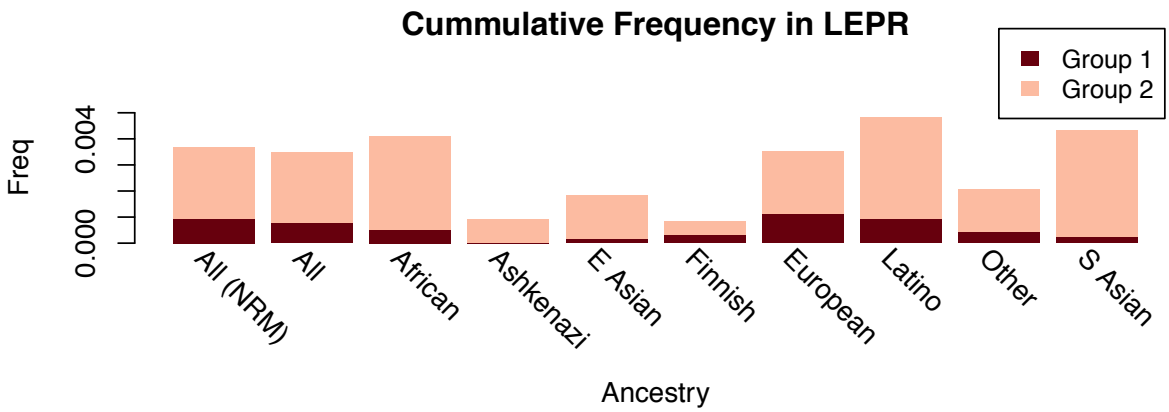
Group 1	Group 1 cont	Group 2 cont	Group 2 cont	Group 2 cont
Glu244ArgfsTer71	Ala15GlyfsTer105	Gly219Asp	Trp84Gly	Glu30Asp
Ser243ProfsTer9	Ala15Gly	Leu205Gln	Trp84Arg	Cys28Tyr
Arg236Gly	Cys6SerfsTer65	Asp192Tyr	Phe82Val	Gln19Arg
Pro231Leu	c.−11C>A	Glu179Gln	His81Arg	Gly10Val
Tyr221Cys		Glu175Gln	Met79Ile	Cys5Tyr
Tyr221LeufsTer94	**Group 2**	Phe172Leu	Met79Ile	Pro2Ser
Glu218Ter	Lys264Asn	Ser168Leu	Leu70Gln	
Glu206Ter	Phe254Ile	Ser168Pro	Gln68Lys	
Asp192ThrfsTer50	Leu253Gln	Val159Ala	Gly63Glu	
Glu187Ter	Thr252Met	Gly151Ser	Pro62Leu	
Glu167Ter	Pro249Ser	Pro149Thr	Pro59Leu	
Arg145Leu	Thr248Met	Lys148Asn	Glu57Lys	
Arg145His	Thr248Ala	Lys148Glu	Ser55Leu	
Arg145Cys	Ser246Arg	Gly147Asp	Lys51Asn	
Phe144Leu	Glu244Asp	Arg145Gly	Lys51Thr	
Gln102ArgfsTer56	Met241Thr	His143Tyr	Cys50Ser	
Gly101LysfsTer58	Met241Lys	Ser140Phe	Cys50Tyr	
Gly99AlafsTer59	Gly239Ser	Ser138Phe	Cys50Arg	
Arg89GlnfsTer57	Gly238Ser	Arg137Pro	Arg48Pro	
Gly88AlafsTer97	Tyr237Cys	Glu134Lys	Cys46Tyr	
Arg86Ter	Tyr237Asn	Gly127Cys	Cys46Arg	
Trp84Ter	Phe226Leu	Glu105Gln	Cys46Ser	
Trp84Ter	His225Gln	Arg104His	Leu43Met	
132+2T>C	His225Pro	Arg104Cys	Thr39Met	
Leu37Phe	Met223Thr	Gly88Arg	Cys34Arg	
Cys28Phe	Met223Val	Arg86Gln	Ser32Arg	
Gln19Ter	Arg222Trp	Arg86Gly	Ser31Cys	

B



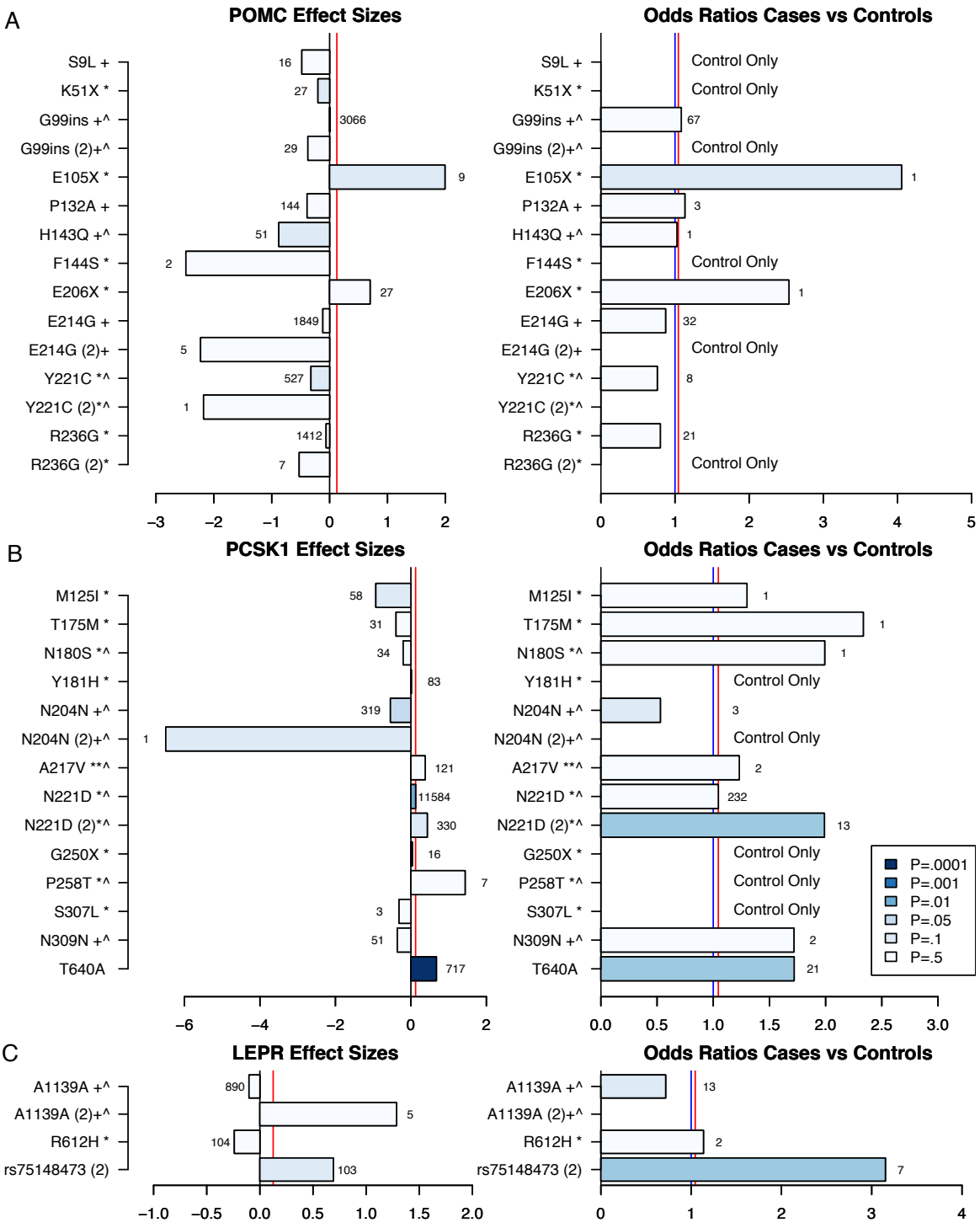
Group 1	Group 1 cont	Group 2 cont	Group 2 cont	Group 2 cont	Group 2 cont	Group 2 cont	Group 2 cont
Tyr729LeufsTer2	Tyr181His	Asp616His	Ala446Val	Ser348Phe	Val261Gly	Thr164Ile	Ser79Cys
Phe700SerfsTer8	Asn180Ser	Asp616Asn	Gly442Arg	Lys346Asn	Val261Ala	Val161Ala	Arg78Ser
Arg669Ter	Thr175Met	Asn611Ser	Leu433Trp	Ala344Asp	Val261Met	Gly158Ala	Pro76Ser
Thr558Ala	Met125Ile	Ser609Phe	Gly432Ser	Ala344Pro	His260Arg	Thr157Met	Gly59Asp
Phe548Ile	Gln102Ter	Ser609Tyr	Ala431Glu	Ala344Thr	Ser253Thr	Thr157Ser	Gly59Ser
1588+2T>C	285+1G>A	Thr608Met	Pro419Leu	Ala344Ser	Gly237Ser	Ile156Thr	Asp56His
Arg517Ter	Arg80Gln	Arg605Pro	Trp413Arg	Pro341Leu	Val235Ala	Gly155Asp	Tyr55Cys
Trp426Ter	Arg80Ter	Arg605His	His409Tyr	Ser332Gly	Lys234Ile	Gly155Ser	Gly42Asp
Asn423Lys	Glu34Ter	Pro604Arg	Met407Ile	Ile329Val	Lys234Thr	Pro150Thr	Gly41Arg
Arg405Ter		Glu599Gly	Thr403Asn	Cys319Tyr	Tyr231Cys	Ile149Thr	Pro40His
Trp404Ter	**Group 2**	Ser595Cys	Pro400Ser	Arg312His	Gly228Val	Asp145Asn	Glu38Lys
1197−1G>A	Arg740Gln	Ile590Thr	Ala398Glu	Gly311Arg	Ala217Val	Leu141Gln	Ala36Thr
1196+2T>A	Arg740Trp	Ile571Asn	Ala398Thr	Gly310Arg	Ile216Thr	Thr138Met	
Ala389LeufsTer45	Asp739Asn	Ile571Phe	Glu397Asp	Val304Ile	Ala213Val	Thr135Ile	
Cys374Ter	Tyr734His	Glu561Gly	Phe392Ser	Gly298Ala	Thr210Ser	Leu132Phe	
1095+1G>C	Tyr729Asp	Thr558Lys	Gly390Ser	Arg296Ile	Glu205Lys	Trp130Leu	
Gln363Ter	Val725Asp	Val556Ile	Pro386Leu	Tyr290Cys	Thr203Ala	Trp130Ser	
Tyr343Ter	Tyr721His	Asp542Gly	Thr381Ile	Tyr290His	Pro198Leu	Arg110His	
Asp320Ter	Ser719Arg	Arg541Trp	Thr377Met	Ala287Thr	Pro198Ser	Arg110Cys	
Ser307Leu	Asp715Tyr	Leu521Phe	Thr375Lys	Ala284Pro	Asp193Gly	Arg107Lys	
882+2T>C	Tyr701Cys	Arg517Gln	Thr366Met	Arg282Gln	Tyr187His	Tyr103Asn	
Pro258Thr	Pro696Arg	Ser503Phe	Asp362Asn	Arg282Trp	Ser186Asn	Val96Ala	
709+2T>C	Ser664Phe	Cys494Gly	Gly358Arg	Thr276Ile	Asn180Lys	Arg95His	
Gly236Ter	Met623Ile	Val461Leu	Tyr355His	Lys275Gln	Tyr178Cys	Ser91Pro	
Gly226Arg	Val620Met	Val461Met	Ser354Cys	Asp272Val	Asp176Tyr	Ser82Thr	
Gly209Arg	Gly619Trp	Asp451Asn	Thr353Ile	Asp272Gly	Leu166Val	Ser82Cys	
Arg199Ter	Gly619Arg	Leu449Val	Thr350Ile	Asp262Asn	Val165Ile	Arg80Leu	

C

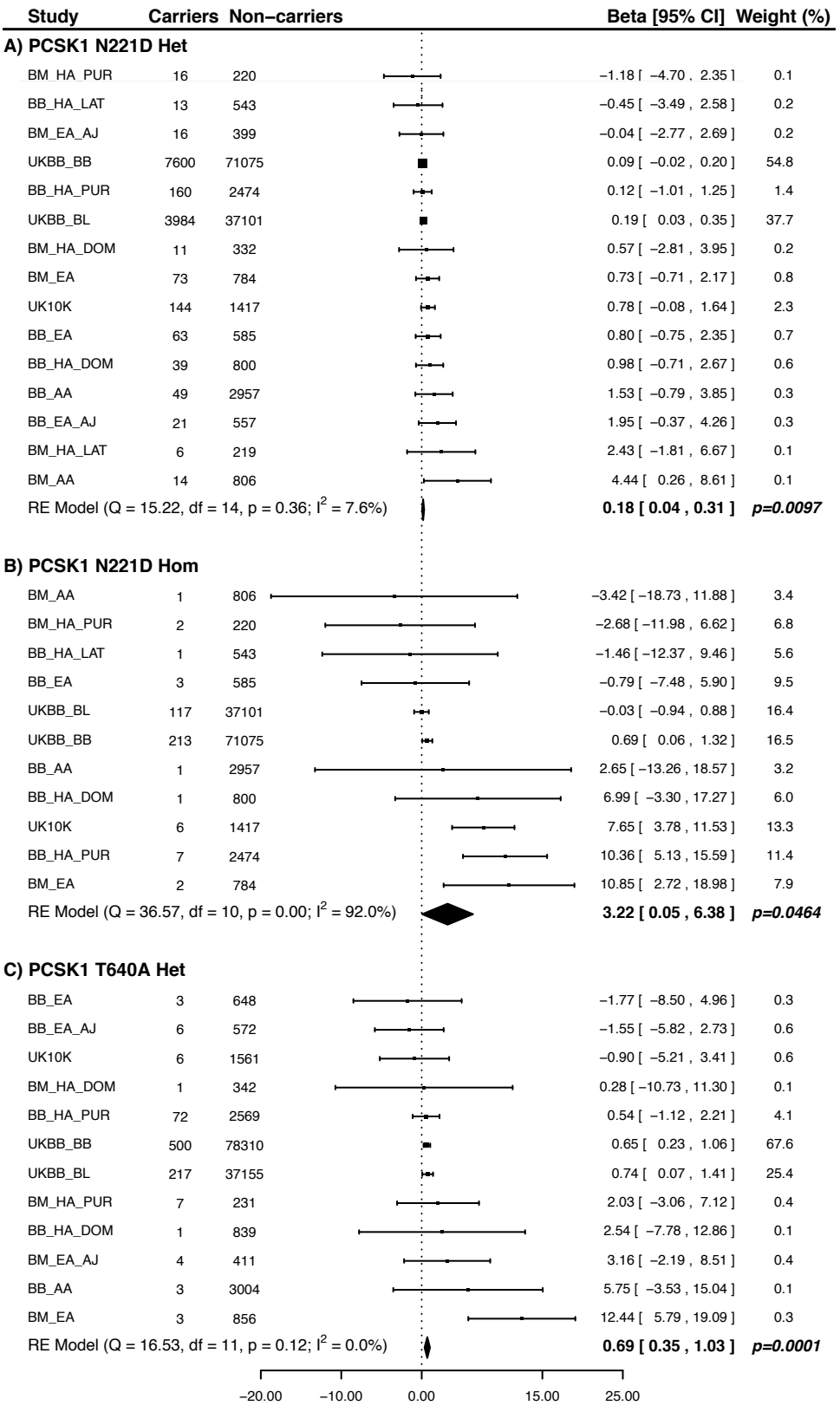


Group 1	Group 1 cont	Group 2 cont	Group 2 cont	Group 2 cont	Group 2 cont	Group 2 cont
Met1?	2396–1G>T	Lys272Ile	Arg402Gln	Met511Val	Asn697Ser	Cys881Phe
Thr29TyrfsTer6	2592_2597+3delCCAA...	Ser292Tyr	Tyr407Cys	Ile513Val	His710Arg	Pro892Thr
Tyr46Ter	2597+1G>T	Pro300His	His419Arg	Asp523Val	His710Leu	Thr894Met
Leu101TyrfsTer15	Val962Ter	Gly301Glu	His419Leu	Pro525Leu	Asn718Ser	His897Arg
Tyr155IlefsTer13	Gln1002Ter	Gln307His	Tyr422Cys	Pro526Ala	Asn726Ser	Val906Gly
Cys212Ter	Tyr1078IlefsTer2	Trp319Cys	Tyr422Phe	Leu530Pro	Thr730Ala	Glu914Gln
Gln223Ter	Thr1147AsnfsTer8	Trp322Cys	Glu424Gln	Val534Glu	Thr730Asn	Pro915Leu
Leu241Ter	Gln1151ArgfsTer13	Ser323Asn	Tyr426Asn	Pro540Thr	Ser736Arg	Glu916Gln
Gln268Ter	Thr1164ValfsTer15	Ile334Thr	Ile432Val	Phe563Val	Val741Met	Glu920Gly
850–1G>A		Thr342Ser	Ile434Met	Gln571His	Gln742Leu	Asp921Asn
Pro316Thr	**Group 2**	Thr342Arg	Ser435Pro	Arg573Cys	Tyr747Asp	Ile922Thr
Phe394SerfsTer3	Phe33Leu	Ser343Asn	Thr438Pro	Arg573His	Val754Met	Ser923Gly
Tyr411LeufsTer4	Ser36Tyr	Asn347His	Thr438Asn	Arg573Leu	Ile755Thr	Asp925His
c.1403+1_1403+2dupGT	Thr43Ala	Asn347Ser	Tyr441Cys	Tyr574Cys	Leu760Arg	Asp932Asn
1752+1G>A	Cys89Phe	Tyr354Cys	Arg448Thr	Gly575Val	Lys775Glu	Asp932His
Arg612His	Leu161Val	Lys355Glu	Arg448Ile	Gly578Val	Ser789Tyr	Cys954Phe
Glu644LeufsTer6	Tyr163Asn	Val361Phe	Trp449Leu	Tyr586Cys	Ser789Phe	Ala987Thr
Trp646Ter	Ser224Leu	Ser363Pro	Arg468Lys	Leu598Arg	Lys794Thr	Ala987Asp
Glu657GlyfsTer15	Pro225Ser	Ile366Phe	Leu471Arg	Tyr607Cys	Asp799Glu	Ala987Val
Trp664Arg	Pro231Ser	Trp369Ser	Asp475Gly	Arg612Cys	Val819Gly	Gly1056Arg
1996–10_1998delTTT...	Ile232Arg	Asn371Lys	Lys486Arg	Cys613Tyr	Phe828Tyr	Ser1098Leu
His684Pro	Lys236Glu	Ala373Thr	Asp493Val	Ile636Lys	Gly841Arg	Ser1133Tyr
2213–1G>T	Leu241Ser	Pro377Ser	Gly494Asp	Val638Phe	Tyr843Ser	Ser1133Phe
Lys766Ter	Asp249Gly	Gln380Arg	Tyr496Asn	Pro639Ser	Tyr843Cys	Gln1146Leu
Glu773Ter	Ser259Ile	Ser385Arg	Glu497Ala	Pro643Ser	Pro876Leu	
Ile783SerfsTer37	Pro266Ser	Ser389Gly	Pro502Ala	Asn659His	Asn877Asp	
Leu786Pro	Val271Met	Ser389Asn	Thr510Ile	Lys665Thr	Pro878Ser	

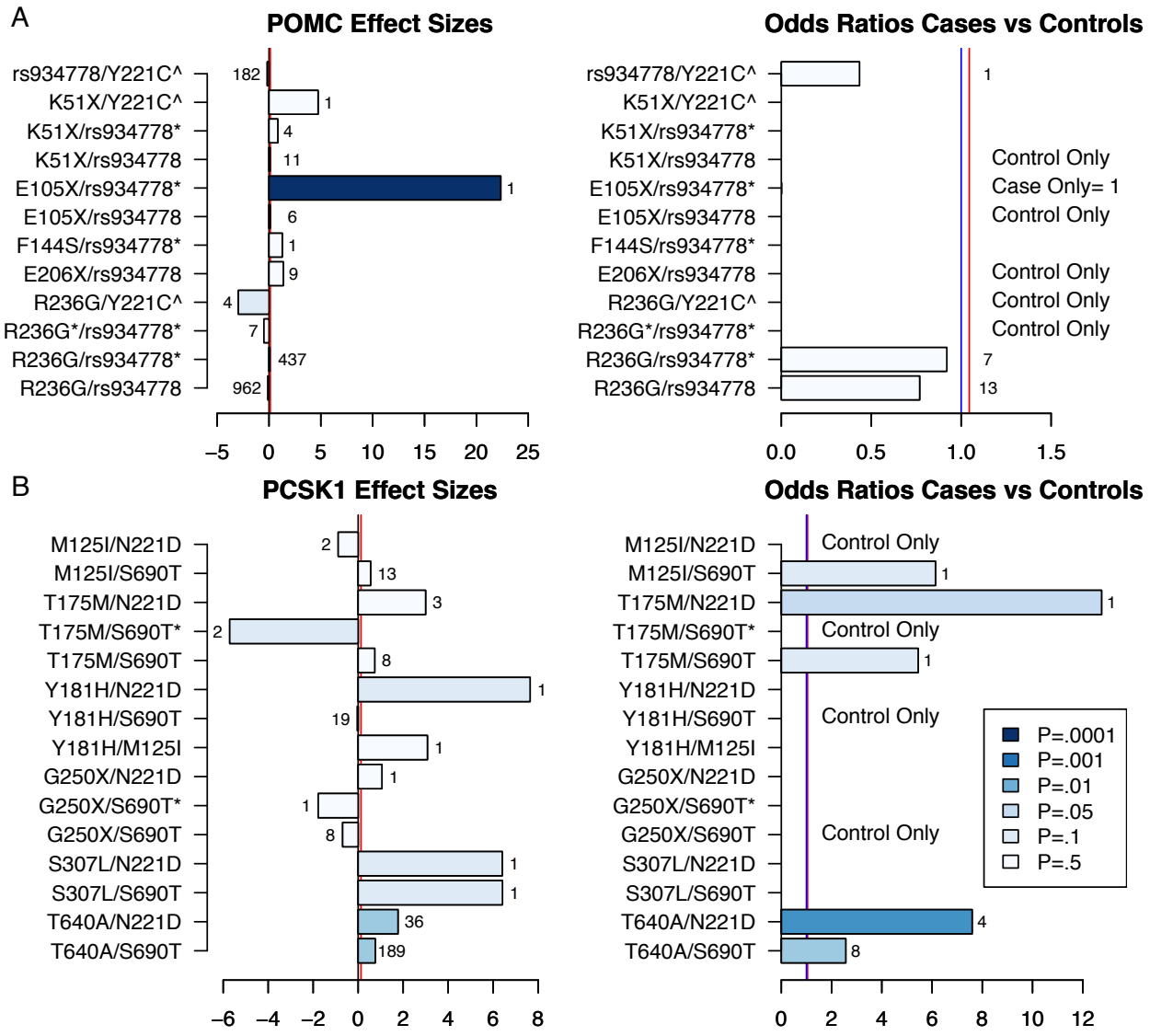
Supplementary Figure 5. Estimated BMI effect sizes (left) and case/control odds ratios (right) for homozygous (denoted by (2)) or heterozygous carriers of individual rare Group 1, Group 2, literature reviewed variants, or other significant variants in *POMC* (A) and *PCSK1* (B) and *LEPR* (C) vs. controls in UK Biobank. For effect size estimates, the number of carriers is denoted next to the bar. For odds ratio estimates, cases are defined as BMI>40, controls as BMI<25, and the number of cases is denoted next to the bar (a notation is made when carriers are either only cases or controls). Coding variants are referred to by their amino acid change while non-coding variants are reported with dbSNP ids. +Denotes a variant studied in literature not in Group1/2. *Denotes a Group 1 variant. **Denotes a Group 2 variant. ^Denotes an imputed variant. Any other additional associations with p-values<0.01 are also reported. The red line represents the *PCSK1* N221D heterozygote effect size or odds ratios. The blue line represents an odds ratio of 1. The darker the blue bar, the more statistically significant the association.


A
POMC Effect Sizes
Odds Ratios Cases vs Controls
S9L +
K51X *
G99ins +^
G99ins (2)+^
E105X *
P132A +
H143Q +^
F144S *
E206X *
E214G +
E214G (2)+
Y221C *^
Y221C (2)*^
R236G *
R236G (2)*
16
27
3066
29
9
144
51
2
27
1849
5
527
1
1412
7
Control Only
Control Only
67
Control Only
1
3
1
Control Only
1
32
Control Only
8
21
Control Only
−3
−2
−1
0
1
2
0
1
2
3
4
5
B
PCSK1 Effect Sizes
Odds Ratios Cases vs Controls
M125I *
T175M *
N180S *^
Y181H *
N204N +^
N204N (2)+^
A217V **^
N221D *^
N221D (2)*^
G250X *
P258T *^
S307L *
N309N +^
T640A
58
31
34
83
319
1
121
11584
330
16
7
3
51
717
1
1
1
Control Only
3
Control Only
2
232
13
Control Only
Control Only
Control Only
2
21
P=.0001
P=.001
P=.01
P=.05
P=.1
P=.5
−6
−4
−2
0
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
C
LEPR Effect Sizes
Odds Ratios Cases vs Controls
A1139A +^
A1139A (2)+^
R612H *
rs75148473 (2)
890
5
104
103
13
2
7
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4

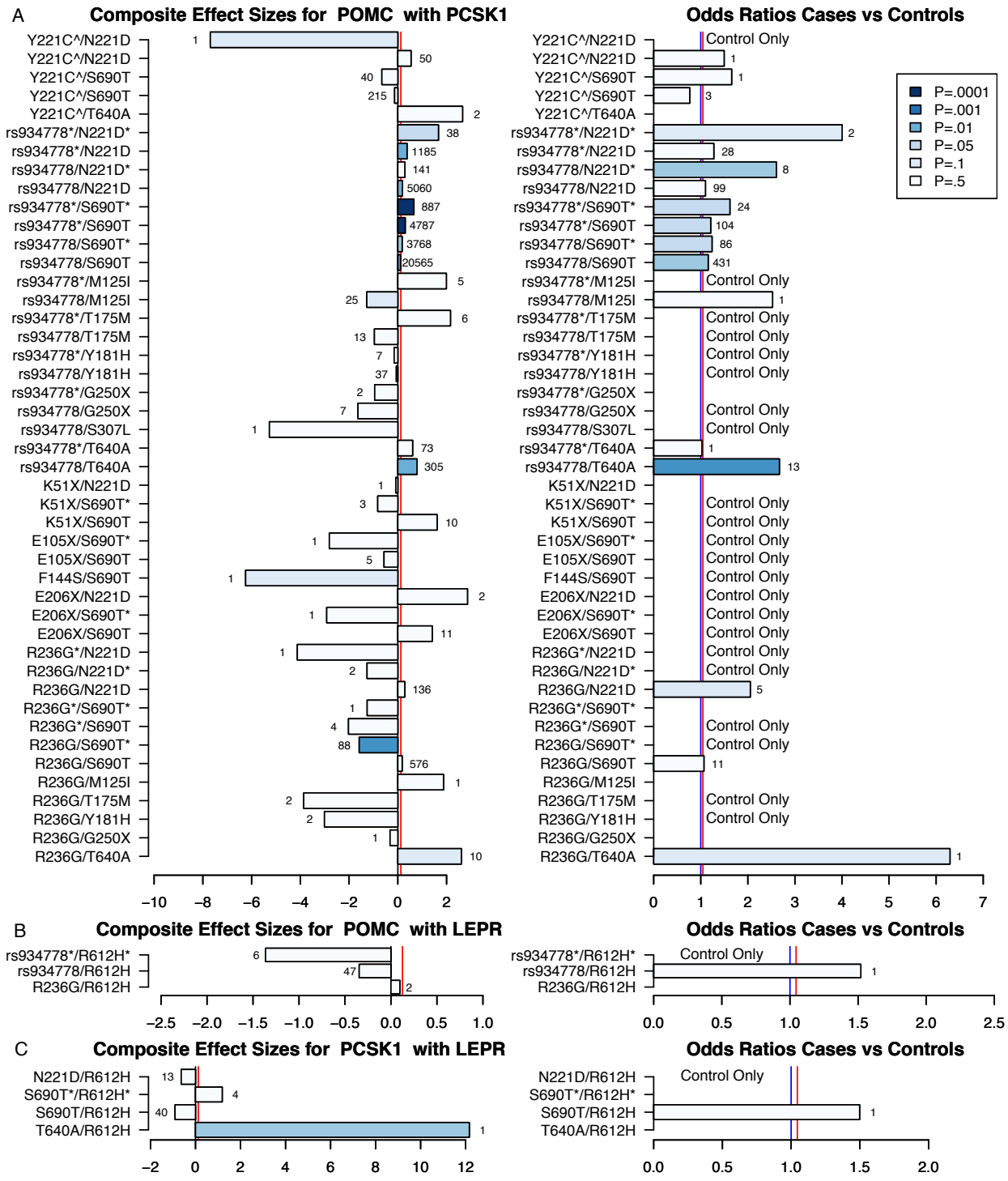
Supplementary Figure 6. Forest plot of effect sizes of *PCSK1* N221D (A) heterozygotes and (B) homozygotes, and (C) *PCSK1* T640A carriers versus non-carriers across cohorts. The UK Biobank data is split by genotyping array: BL (BiLEVE array) and BB (Biobank Axiom array). For the MHS data, both the sequences (BM) and genotype (BB) data set were split by reported race and ethnicity: 1) White Caucasian was divided into EA-European American or EA_AJ- Jewish; 2) Black or African was divided into AA-African American or AFR-African; and 3) HA-Hispanic/Latino was divided in PUR-Puerto Rican, LAT-Central and South American, and DOM- Dominican Republic. Ethnic outliers were removed using PCs (see Supplementary Materials and Methods).


Study
Carriers
Non−carriers
Beta [95% CI]
Weight (%)
A) PCSK1 N221D Het
BM_HA_PUR 16 220 −1.18 [−4.70 , 2.35] 0.1
BB_HA_LAT 13 543 −0.45 [−3.49 , 2.58] 0.2
BM_EA_AJ 16 399 −0.04 [−2.77 , 2.69] 0.2
UKBB_BB 7600 71075 0.09 [−0.02 , 0.20] 54.8
BB_HA_PUR 160 2474 0.12 [−1.01 , 1.25] 1.4
UKBB_BL 3984 37101 0.19 [0.03 , 0.35] 37.7
BM_HA_DOM 11 332 0.57 [−2.81 , 3.95] 0.2
BM_EA 73 784 0.73 [−0.71 , 2.17] 0.8
UK10K 144 1417 0.78 [−0.08 , 1.64] 2.3
BB_EA 63 585 0.80 [−0.75 , 2.35] 0.7
BB_HA_DOM 39 800 0.98 [−0.71 , 2.67] 0.6
BB_AA 49 2957 1.53 [−0.79 , 3.85] 0.3
BB_EA_AJ 21 557 1.95 [−0.37 , 4.26] 0.3
BM_HA_LAT 6 219 2.43 [−1.81 , 6.67] 0.1
BM_AA 14 806 4.44 [0.26 , 8.61] 0.1
RE Model (Q = 15.22, df = 14, p = 0.36; I² = 7.6%) 0.18 [0.04 , 0.31] p=0.0097
B) PCSK1 N221D Hom
BM_AA 1 806 −3.42 [−18.73 , 11.88] 3.4
BM_HA_PUR 2 220 −2.68 [−11.98 , 6.62] 6.8
BB_HA_LAT 1 543 −1.46 [−12.37 , 9.46] 5.6
BB_EA 3 585 −0.79 [−7.48 , 5.90] 9.5
UKBB_BL 117 37101 −0.03 [−0.94 , 0.88] 16.4
UKBB_BB 213 71075 0.69 [0.06 , 1.32] 16.5
BB_AA 1 2957 2.65 [−13.26 , 18.57] 3.2
BB_HA_DOM 1 800 6.99 [−3.30 , 17.27] 6.0
UK10K 6 1417 7.65 [3.78 , 11.53] 13.3
BB_HA_PUR 7 2474 10.36 [5.13 , 15.59] 11.4
BM_EA 2 784 10.85 [2.72 , 18.98] 7.9
RE Model (Q = 36.57, df = 10, p = 0.00; I² = 92.0%) 3.22 [0.05 , 6.38] p=0.0464
C) PCSK1 T640A Het
BB_EA 3 648 −1.77 [−8.50 , 4.96] 0.3
BB_EA_AJ 6 572 −1.55 [−5.82 , 2.73] 0.6
UK10K 6 1561 −0.90 [−5.21 , 3.41] 0.6
BM_HA_DOM 1 342 0.28 [−10.73 , 11.30] 0.1
BB_HA_PUR 72 2569 0.54 [−1.12 , 2.21] 4.1
UKBB_BB 500 78310 0.65 [0.23 , 1.06] 67.6
UKBB_BL 217 37155 0.74 [0.07 , 1.41] 25.4
BM_HA_PUR 7 231 2.03 [−3.06 , 7.12] 0.4
BB_HA_DOM 1 839 2.54 [−7.78 , 12.86] 0.1
BM_EA_AJ 4 411 3.16 [−2.19 , 8.51] 0.4
BB_AA 3 3004 5.75 [−3.53 , 15.04] 0.1
BM_EA 3 856 12.44 [5.79 , 19.09] 0.3
RE Model (Q = 16.53, df = 11, p = 0.12; I² = 0.0%) 0.69 [0.35 , 1.03] p=0.0001
−20.00
−10.00
0.00
15.00
25.00

Supplementary Figure 7. Estimated effect sizes (left) and odds ratios (right) of potential compound heterozygotes for carriers of individual rare Group 1 plus Group 2 variants and the 2 BMI associated variants POMC rs934778 and *PCSK1* S690T in *POMC* (A) and *PCSK1* (B) vs. controls in UK Biobank. For effect size estimates, the number of carriers is denoted next to the bar. For odds ratio estimates, cases are defined as BMI>40, controls as BMI<25, and the number of cases is denoted next to the bar (a notation is made when carriers are either only cases or controls). Coding variants are referred to by their amino acid change while non-coding variants are reported with dbSNP ids. *Denotes a homozygous copy for that variant. ^Denotes an imputed variant. Homozygote carriers of a variant are denoted with (2). The blue line represents an odds ratio of 1. The darker the blue bar, the more statistically significant the association. Note: for combinations of rare alleles with rs934778 or S690T, some of the individuals may be compound heterozygotes while for others the variants will lie on the same haplotype. Combinations of N221/S690T were excluded, as they are known to be on the same haplotype in the British population.

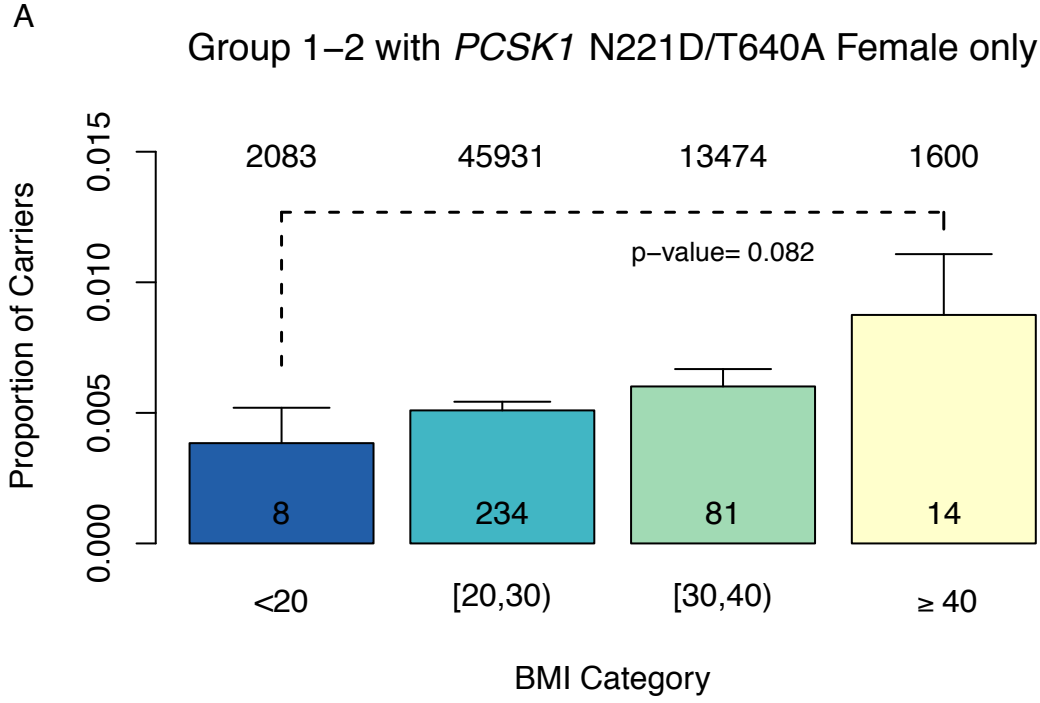

A
POMC Effect Sizes
rs934778/Y221C^
K51X/Y221C^
K51X/rs934778*
K51X/rs934778
E105X/rs934778*
E105X/rs934778
F144S/rs934778*
E206X/rs934778
R236G/Y221C^
R236G*/rs934778*
R236G/rs934778*
R236G/rs934778
182
1
4
11
1
6
1
9
4
7
437
962
−5 0 5 10 15 20 25
Odds Ratios Cases vs Controls
1
Control Only
Case Only= 1
Control Only
Control Only
Control Only
Control Only
7
13
0.0 0.5 1.0 1.5
B
PCSK1 Effect Sizes
M125I/N221D
M125I/S690T
T175M/N221D
T175M/S690T*
T175M/S690T
Y181H/N221D
Y181H/S690T
Y181H/M125I
G250X/N221D
G250X/S690T*
G250X/S690T
S307L/N221D
S307L/S690T
T640A/N221D
T640A/S690T
2
13
3
2
8
1
19
1
1
1
8
1
1
36
189
−6 −4 −2 0 2 4 6 8
Odds Ratios Cases vs Controls
Control Only
Control Only
Control Only
Control Only
1
1
1
4
8
P=.0001
P=.001
P=.01
P=.05
P=.1
P=.5
0 2 4 6 8 10 12

Supplementary Figure 8. Estimated effect sizes (left) and odds ratios (right) of composites for carriers of individual rare Group 1 plus Group 2 variants and the 2 BMI associated variants POMC rs934778 and PCSK1 S690T in *POMC/PCSK1* (A), *POMC*/*LEPR* (B), and *PCSK1/LEPR* (C) vs. controls in UK Biobank. For effect size estimates, the number of carriers is denoted next to the bar. For odds ratio estimates, cases are defined as BMI>40, controls as BMI<25, and the number of cases is denoted next to the bar (a notation is made when carriers are either only cases or controls). Coding variants are referred to by their amino acid change while non-coding variants are reported with dbSNP ids. *Denotes a homozygous copy for that variant. ^Denotes an imputed variant. The variant listed first in the pair belongs to the gene listed first in the title. The blue line represents an odds ratio of 1. The darker the blue bar, the more statistically significant the association.

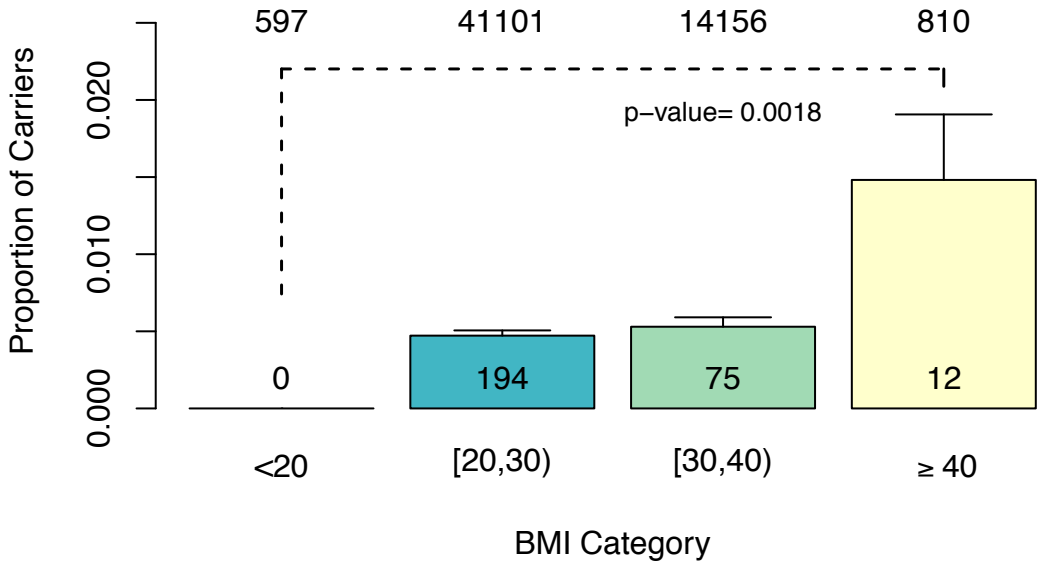

A
Composite Effect Sizes for POMC with PCSK1
Odds Ratios Cases vs Controls
P=.0001
P=.001
P=.01
P=.05
P=.1
P=.5
B
Composite Effect Sizes for POMC with LEPR
Odds Ratios Cases vs Controls
C
Composite Effect Sizes for PCSK1 with LEPR
Odds Ratios Cases vs Controls

Supplementary Figure 9. In the UK Biobank, the proportion of individuals carrying two or more Group 1 and 2 variants (plus *PCSK1* N221D and T640A) (bottom) across these 3 genes increases as BMI increases. (A) Females only. (B) Males only.

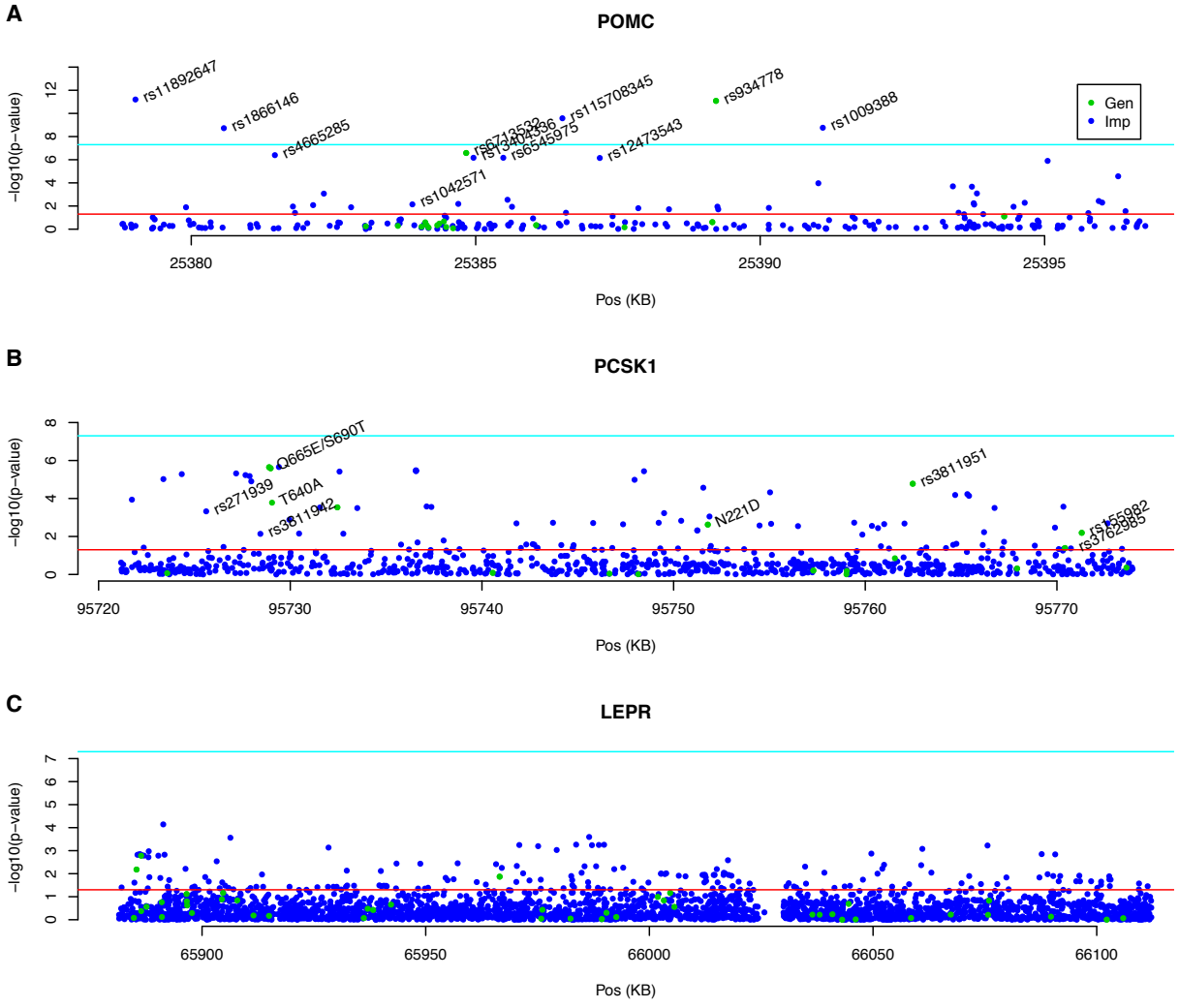
A


Group 1–2 with *PCSK1* N221D/T640A Female only
Proportion of Carriers
0.000
0.005
0.010
0.015
2083
45931
13474
1600
p–value= 0.082
8
234
81
14
<20
[20,30)
[30,40)
≥ 40
BMI Category

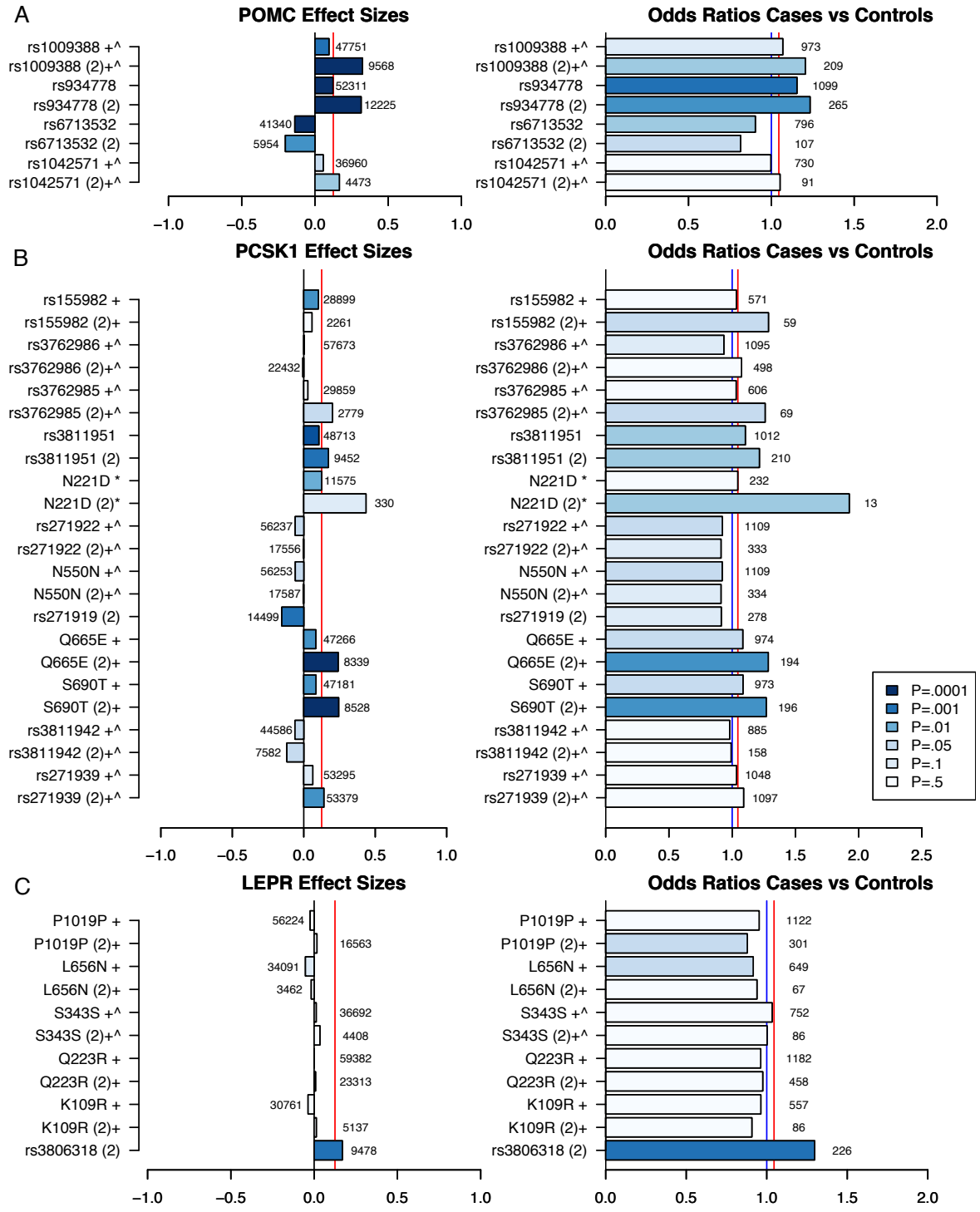
B


Group 1–2 with *PCSK1* N221D/T640A Male only
Proportion of Carriers
0.000
0.010
0.020
597
41101
14156
810
p–value= 0.0018
0
194
75
12
<20
[20,30)
[30,40)
≥ 40
BMI Category

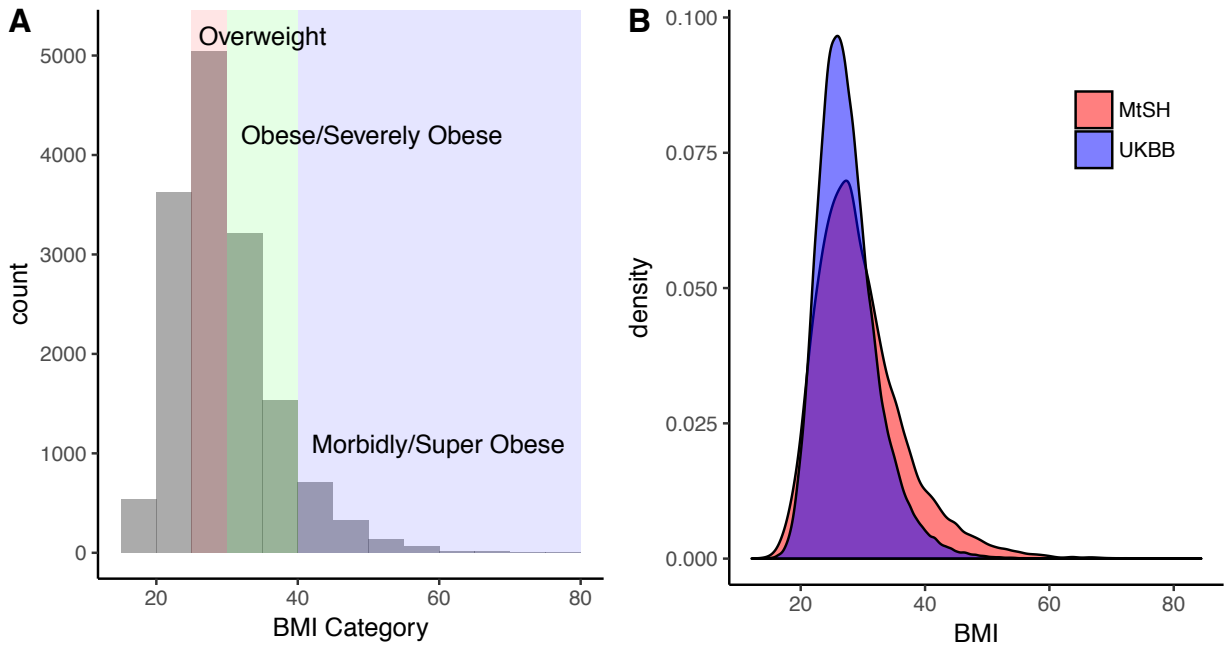
Supplementary Figure 10. Manhattan plot (p-values versus position) for the UK Biobank of variant in *POMC* (A), *PCSK*1 (B), and *LEPR* (C). The red line denotes a p-value of 0.05. Significant variants are labeled with either the amino acid change or dbSNP id. Genotyped (Gen) variants are represented by green dots, while imputed variants (Imp) are represented by blue dots. The y axis is the –log10 p-values while the x axis corresponds to base pair position in Kb.



Supplementary Figure 11. Estimated BMI effect sizes (left) and case/control odds ratios (right) for homozygous (denoted by (2)) or heterozygous carriers of individual common literature variants or other significant variants in *POMC* (A) and *PCSK1* (B) and *LEPR* (C) vs. controls in UK Biobank. For effect size estimates, the number of carriers is denoted next to the bar. For odds ratio estimates, cases are defined as BMI>40, controls as BMI<25, and the number of cases is denoted next to the bar. Coding variants are referred to by their amino acid change while non-coding variants are reported with dbSNP ids. +Denotes a variant studied in literature not in Group1/2. ^Denotes an imputed variant. * Denotes a group 1 variant. Tests on the homozygous genotype are denote with (2). Any other additional associations with p-values<0.01 are also reported. The red line represents the *PCSK1* N221D heterozygote effect size or odds ratios. The blue line represents an odds ratio of 1. The darker the blue bar, the more statistically significant the association.


A
POMC Effect Sizes
rs1009388 +^
rs1009388 (2)+^
rs934778
rs934778 (2)
rs6713532
rs6713532 (2)
rs1042571 +^
rs1042571 (2)+^
47751
9568
52311
12225
41340
5954
36960
4473
−1.0
−0.5
0.0
0.5
1.0
Odds Ratios Cases vs Controls
973
209
1099
265
796
107
730
91
0.0
0.5
1.0
1.5
2.0
B
PCSK1 Effect Sizes
rs155982 +
rs155982 (2)+
rs3762986 +^
rs3762986 (2)+^
rs3762985 +^
rs3762985 (2)+^
rs3811951
rs3811951 (2)
N221D *
N221D (2)*
rs271922 +^
rs271922 (2)+^
N550N +^
N550N (2)+^
rs271919 (2)
Q665E +
Q665E (2)+
S690T +
S690T (2)+
rs3811942 +^
rs3811942 (2)+^
rs271939 +^
rs271939 (2)+^
28899
2261
57673
22432
29859
2779
48713
9452
11575
330
56237
17556
56253
17587
14499
47266
8339
47181
8528
44586
7582
53295
53379
−1.0
−0.5
0.0
0.5
1.0
Odds Ratios Cases vs Controls
571
59
1095
498
606
69
1012
210
232
13
1109
333
1109
334
278
974
194
973
196
885
158
1048
1097
0.0
0.5
1.0
1.5
2.0
2.5
P=.0001
P=.001
P=.01
P=.05
P=.1
P=.5
C
LEPR Effect Sizes
P1019P +
P1019P (2)+
L656N +
L656N (2)+
S343S +^
S343S (2)+^
Q223R +
Q223R (2)+
K109R +
K109R (2)+
rs3806318 (2)
56224
16563
34091
3462
36692
4408
59382
23313
30761
5137
9478
−1.0
−0.5
0.0
0.5
1.0
Odds Ratios Cases vs Controls
1122
301
649
67
752
86
1182
458
557
86
226
0.0
0.5
1.0
1.5
2.0

Supplementary Figure 12. (A) Distribution of BMI in Mount Sinai Biobank. (B) Distribution of BMI in Mount Sinai Biobank (MtSH) cohort versus the UK Biobank (UKBB) cohort.



Supplementary Table (in Microsoft excel) Description

Supplementary Tables 1-3. Potential LoF variants in POMC (Supplementary Table 1), PCSK1 (Supplementary Table 2), and LEPR (Supplementary Table 3) genes. All tables include the dbSNP id, the predicted effect, the amino acid change, the hg19 left normalized position, the reference allele, the alternate allele, the gnomAD allele frequency, the predicted Wuxi deepCODE score, the Mutation assessor (46) predicted effect, and the SIFT (47) predicted effects. Tables 1A, 2A, and 3A are the results of literature curation and include additional information about the functional evidence, hgmd and clinvar notation, and references. Tables 1B, 2B, and 3B list all Group 1 and Group2 variants.

Supplementary Table 7. The maximum number of sequenced individuals for any variant in gnomAD for each ethnicity.

Supplementary Table 11. Association results from different studies for common variants studied in the literature along with allele frequencies amongst different cohorts.

Supplementary References

1. Heymsfield SB, Wadden TA. Mechanisms, Pathophysiology, and Management of Obesity. The New England journal of medicine 2017; 376:254-266

2. Jackson VM, Breen DM, Fortin JP, Liou A, Kuzmiski JB, Loomis AK, Rives ML, Shah B, Carpino PA. Latest approaches for the treatment of obesity. Expert opinion on drug discovery 2015; 10:825-839

3. Jackson VM, Price DA, Carpino PA. Investigational drugs in Phase II clinical trials for the treatment of obesity: implications for future development of novel therapies. Expert opinion on investigational drugs 2014; 23:1055-1066

4. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. Jama 2013; 309:71-82

5. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. Jama 2014; 311:806-814

6. Muller TD, Nogueiras R, Andermann ML, Andrews ZB, Anker SD, Argente J, Batterham RL, Benoit SC, Bowers CY, Broglio F, Casanueva FF, D'Alessio D, Depoortere I, Geliebter A, Ghigo E, Cole PA, Cowley M, Cummings DE, Dagher A, Diano S, Dickson SL, Dieguez C, Granata R, Grill HJ, Grove K, Habegger KM, Heppner K, Heiman ML, Holsen L, Holst B, Inui A, Jansson JO, Kirchner H, Korbonits M, Laferrere B, LeRoux CW, Lopez M, Morin S, Nakazato M, Nass R, Perez-Tilve D, Pfluger PT, Schwartz TW, Seeley RJ, Sleeman M, Sun Y, Sussel L, Tong J, Thorner MO, van der Lely AJ, van der Ploeg LH, Zigman JM, Kojima M, Kangawa K, Smith RG, Horvath T, Tschop MH. Ghrelin. Mol Metab 2015; 4:437-460

7. Cone RD. Anatomy and regulation of the central melanocortin system. Nature neuroscience 2005; 8:571-578

8. Farooqi IS, O'Rahilly S. Mutations in ligands and receptors of the leptin-melanocortin pathway that lead to obesity. Nature clinical practice Endocrinology & metabolism 2008; 4:569-577

9. Walley AJ, Asher JE, Froguel P. The genetic contribution to non-syndromic human obesity. Nature reviews Genetics 2009; 10:431-442

10. Yeo GS, Heisler LK. Unraveling the brain regulation of appetite: lessons from genetics. Nature neuroscience 2012; 15:1343-1349

11. Collet T-H, Dubern B, Mokrosinski J, Connors H, Keogh JM, Mendes de Oliveira E, Henning E, Poitou-Bernert C, Oppert J-M, Tounian P, Marchelli F, Alili R, Le Beyec J, Pépin D, Lacorte J-M, Gottesdiener A, Bounds R, Sharma S, Folster C, Henderson B, O'Rahilly S, Stoner E, Gottesdiener K, Panaro BL, Cone RD, Clément K, Farooqi IS, Van der Ploeg LHT. Evaluation of a melanocortin-4 receptor (MC4R) agonist (Setmelanotide) in MC4R deficiency. Molecular Metabolism 2017;

12. Streicher SA, Sanderson SC, Jabs EW, Diefenbach M, Smirnoff M, Peter I, Horowitz CR, Brenner B, Richardson LD. Reasons for participating and genetic information needs among racially and ethnically diverse biobank participants: a focus group study. Journal of community genetics 2011; 2:153-163

13. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491:56-65

14. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stutz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G,

Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature 2015; 526:75-81

15. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods 2012; 9:179-181

16. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda, Md) 2011; 1:457-470

17. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics 2014; 133:1-9

18. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research 2014; 42:D980-985

19. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 2014; 46:310-315

20. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2015 update. Nucleic acids research 2015; 43:D670-681

21. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science (New York, NY) 2012; 337:64-69

22. Sutskever I, Martens J, Dahl G, Hinton G. 2013 On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28; 2013; Atlanta, GA, USA.

23. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15:1929-1958

24. Chakraborty R, Srinivasan MR, Daiger SP. Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. American journal of human genetics 1993; 52:60-70

25. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segre AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpelainen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li S, Moffatt MF, Myers RH, Narisu N, Perry JR, Peters MJ, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM, White CC, Wiklund F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi J, Thiering E, Alavere H, Alibrandi MT, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AI, Boes T,

Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA, Busonero F, Campbell H, Cappuccio FP, Cavalcanti-Proenca C, Chen YD, Chen CM, Chines PS, Clarke R, Coin L, Connell J, Day IN, den Heijer M, Duan J, Ebrahim S, Elliott P, Elosua R, Eiriksdottir G, Erdos MR, Eriksson JG, Facheris MF, Felix SB, Fischer-Posovszky P, Folsom AR, Friedrich N, Freimer NB, Fu M, Gaget S, Gejman PV, Geus EJ, Gieger C, Gjesing AP, Goel A, Goyette P, Grallert H, Grassler J, Greenawalt DM, Groves CJ, Gudnason V, Guiducci C, Hartikainen AL, Hassanali N, Hall AS, Havulinna AS, Hayward C, Heath AC, Hengstenberg C, Hicks AA, Hinney A, Hofman A, Homuth G, Hui J, Igl W, Iribarren C, Isomaa B, Jacobs KB, Jarick I, Jewell E, John U, Jorgensen T, Jousilahti P, Jula A, Kaakinen M, Kajantie E, Kaplan LM, Kathiresan S, Kettunen J, Kinnunen L, Knowles JW, Kolcic I, Konig IR, Koskinen S, Kovacs P, Kuusisto J, Kraft P, Kvaloy K, Laitinen J, Lantieri O, Lanzani C, Launer LJ, Lecoeur C, Lehtimaki T, Lettre G, Liu J, Lokki ML, Lorentzon M, Luben RN, Ludwig B, Manunta P, Marek D, Marre M, Martin NG, McArdle WL, McCarthy A, McKnight B, Meitinger T, Melander O, Meyre D, Midthjell K, Montgomery GW, Morken MA, Morris AP, Mulic R, Ngwa JS, Nelis M, Neville MJ, Nyholt DR, O'Donnell CJ, O'Rahilly S, Ong KK, Oostra B, Pare G, Parker AN, Perola M, Pichler I, Pietilainen KH, Platou CG, Polasek O, Pouta A, Rafelt S, Raitakari O, Rayner NW, Ridderstrale M, Rief W, Ruokonen A, Robertson NR, Rzehak P, Salomaa V, Sanders AR, Sandhu MS, Sanna S, Saramies J, Savolainen MJ, Scherag S, Schipf S, Schreiber S, Schunkert H, Silander K, Sinisalo J, Siscovick DS, Smit JH, Soranzo N, Sovio U, Stephens J, Surakka I, Swift AJ, Tammesoo ML, Tardif JC, Teder-Laving M, Teslovich TM, Thompson JR, Thomson B, Tonjes A, Tuomi T, van Meurs JB, van Ommen GJ, Vatin V, Viikari J, Visvikis-Siest S, Vitart V, Vogel CI, Voight BF, Waite LL, Wallaschofski H, Walters GB, Widen E, Wiegand S, Wild SH, Willemsen G, Witte DR, Witteman JC, Xu J, Zhang Q, Zgaga L, Ziegler A, Zitting P, Beilby JP, Farooqi IS, Hebebrand J, Huikuri HV, James AL, Kahonen M, Levinson DF, Macciardi F, Nieminen MS, Ohlsson C, Palmer LJ, Ridker PM, Stumvoll M, Beckmann JS, Boeing H, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Collins FS, Cupples LA, Smith GD, Erdmann J, Froguel P, Gronberg H, Gyllensten U, Hall P, Hansen T, Harris TB, Hattersley AT, Hayes RB, Heinrich J, Hu FB, Hveem K, Illig T, Jarvelin MR, Kaprio J, Karpe F, Khaw KT, Kiemeney LA, Krude H, Laakso M, Lawlor DA, Metspalu A, Munroe PB, Ouwehand WH, Pedersen O, Penninx BW, Peters A, Pramstaller PP, Quertermous T, Reinehr T, Rissanen A, Rudan I, Samani NJ, Schwarz PE, Shuldiner AR, Spector TD, Tuomilehto J, Uda M, Uitterlinden A, Valle TT, Wabitsch M, Waeber G, Wareham NJ, Watkins H, Wilson JF, Wright AF, Zillikens MC, Chatterjee N, McCarroll SA, Purcell S, Schadt EE, Visscher PM, Assimes TL, Borecki IB, Deloukas P, Fox CS, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, Mohlke KL, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, van Duijn CM, Wichmann HE, Frayling TM, Thorsteinsdottir U, Abecasis GR, Barroso I, Boehnke M, Stefansson K, North KE, McCarthy MI, Hirschhorn JN, Ingelsson E, Loos RJ. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics 2010; 42:937-948

26. Horikoshi M, Mgi R, van de Bunt M, Surakka I, Sarin AP, Mahajan A, Marullo L, Thorleifsson G, Hgg S, Hottenga JJ, Ladenvall C, Ried JS, Winkler TW, Willems SM, Pervjakova N, Esko T, Beekman M, Nelson CP, Willenborg C, Wiltshire S, Ferreira T, Fernandez J, Gaulton KJ, Steinthorsdottir V, Hamsten A, Magnusson PK, Willemsen G, Milaneschi Y, Robertson NR, Groves CJ, Bennett AJ, Lehtimki T, Viikari JS, Rung J, Lyssenko V, Perola M, Heid IM, Herder C, Grallert H, Muller-Nurasyid M, Roden M, Hypponen E, Isaacs A, van Leeuwen EM, Karssen LC, Mihailov E, Houwing-Duistermaat JJ, de Craen AJ, Deelen J, Havulinna AS, Blades M, Hengstenberg C, Erdmann J, Schunkert H, Kaprio J, Tobin MD, Samani NJ, Lind L, Salomaa V, Lindgren CM, Slagboom PE, Metspalu A, van Duijn CM, Eriksson JG, Peters A, Gieger C, Jula A, Groop L, Raitakari OT, Power C, Penninx BW, de Geus E, Smit JH, Boomsma DI, Pedersen NL, Ingelsson E, Thorsteinsdottir U, Stefansson K, Ripatti S, Prokopenko I, McCarthy MI, Morris AP. Discovery and

Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. PLoS genetics 2015; 11:e1005230

27. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 2015; 12:e1001779

28. Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G, Barroso I, Nordstrom P, Franks PW. Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden. Human molecular genetics 2009; 18:1489-1496

29. Peterson RE, Maes HH, Lin P, Kramer JR, Hesselbrock VM, Bauer LO, Nurnberger JI, Jr., Edenberg HJ, Dick DM, Webb BT. On the association of common and rare genetic variation influencing body mass index: a combined SNP and CNV analysis. BMC genomics 2014; 15:368

30. Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, Nelson SF, Martin MG, Lindberg I. Functional consequences of a novel variant of PCSK1. PloS one 2013; 8:e55065

31. Biebermann H, Castaneda TR, van Landeghem F, von Deimling A, Escher F, Brabant G, Hebebrand J, Hinney A, Tschop MH, Gruters A, Krude H. A role for beta-melanocyte-stimulating hormone in human body-weight regulation. Cell metabolism 2006; 3:141-146

32. Lee YS, Challis BG, Thompson DA, Yeo GS, Keogh JM, Madonna ME, Wraight V, Sims M, Vatin V, Meyre D, Shield J, Burren C, Ibrahim Z, Cheetham T, Swift P, Blackwood A, Hung CC, Wareham NJ, Froguel P, Millhauser GL, O'Rahilly S, Farooqi IS. A POMC variant implicates beta-melanocyte-stimulating hormone in the control of human energy balance. Cell metabolism 2006; 3:135-140

33. Raffan E, Dennis RJ, O'Donovan CJ, Becker JM, Scott RA, Smith SP, Withers DJ, Wood CJ, Conci E, Clements DN, Summers KM, German AJ, Mellersh CS, Arendt ML, Iyemere VP, Withers E, Soder J, Wernersson S, Andersson G, Lindblad-Toh K, Yeo GS, O'Rahilly S. A Deletion in the Canine POMC Gene Is Associated with Weight and Appetite in Obesity-Prone Labrador Retriever Dogs. Cell metabolism 2016; 23:893-900

34. Challis BG, Pritchard LE, Creemers JW, Delplanque J, Keogh JM, Luan J, Wareham NJ, Yeo GS, Bhattacharyya S, Froguel P, White A, Farooqi IS, O'Rahilly S. A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. Human molecular genetics 2002; 11:1997-2004

35. Ternouth A, Brandys MK, van der Schouw YT, Hendriks J, Jansson JO, Collier D, Adan RA. Association study of POMC variants with body composition measures and nutrient choice. European journal of pharmacology 2011; 660:220-225

36. Srivastava A, Mittal B, Prakash J, Srivastava P, Srivastava N. Analysis of MC4R rs17782313, POMC rs1042571, APOE-Hha1 and AGRP rs3412352 genetic variants with susceptibility to obesity risk in North Indians. Annals of human biology 2016; 43:285-288

37. Wang F, Gelernter J, Kranzler HR, Zhang H. Identification of POMC exonic variants associated with substance dependence and body mass index. PloS one 2012; 7:e45300

38. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L, Chen CH, Delahanty RJ, Okada Y, Tabara Y, Gu D, Zhu D, Haiman CA, Mo Z, Gao YT, Saw SM, Go MJ, Takeuchi F, Chang LC, Kokubo Y, Liang J, Hao M, Le Marchand L, Zhang Y, Hu Y, Wong TY, Long J, Han BG, Kubo M, Yamamoto K, Su MH, Miki T, Henderson BE, Song H, Tan A, He J, Ng DP, Cai Q, Tsunoda T, Tsai FJ, Iwai N, Chen GK, Shi J, Xu J, Sim X, Xiang YB, Maeda S, Ong RT, Li C, Nakamura Y, Aung T, Kamatani N, Liu JJ, Lu W, Yokota M, Seielstad M, Fann CS, Wu JY, Lee JY, Hu FB, Tanaka T, Tai ES, Shu XO. Meta-analysis identifies common variants associated with body mass index in east Asians. Nature genetics 2012; 44:307-311

39. Choquet H, Kasberger J, Hamidovic A, Jorgenson E. Contribution of common PCSK1 genetic variants to obesity in 8,359 subjects from multi-ethnic American population. PloS one 2013; 8:e57857

40. Kilpelainen TO, Bingham SA, Khaw KT, Wareham NJ, Loos RJ. Association of variants in the PCSK1 gene with obesity in the EPIC-Norfolk study. Human molecular genetics 2009; 18:3496-3501

41. Villalobos-Comparan M, Villamil-Ramirez H, Villarreal-Molina T, Larrieta-Carrasco E, Leon-Mimila P, Romero-Hidalgo S, Jacobo-Albavera L, Liceaga-Fuentes AE, Campos-Perez FJ, Lopez-Contreras BE, Tusie-Luna T, Del Rio-Navarro BE, Aguilar-Salinas CA, Canizales-Quinteros S. PCSK1 rs6232 is associated with childhood and adult class III obesity in the Mexican population. PloS one 2012; 7:e39037

42. Loffler D, Behrendt S, Creemers JW, Klammt J, Aust G, Stanik J, Kiess W, Kovacs P, Korner A. Functional and clinical relevance of novel and known PCSK1 variants for childhood obesity and glucose metabolism. Mol Metab 2017; 6:295-305

43. Stijnen P, Tuand K, Varga TV, Franks PW, Aertgeerts B, Creemers JW. The association of common variants in PCSK1 with obesity: a HuGE review and meta-analysis. American journal of epidemiology 2014; 180:1051-1065

44. Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, Durand E, Guerardel A, Boutin P, Jouret B, Heude B, Balkau B, Tichet J, Marre M, Potoczna N, Horber F, Le Stunff C, Czernichow S, Sandbaek A, Lauritzen T, Borch-Johnsen K, Andersen G, Kiess W, Korner A, Kovacs P, Jacobson P, Carlsson LM, Walley AJ, Jorgensen T, Hansen T, Pedersen O, Meyre D, Froguel P. Common nonsynonymous variants in PCSK1 confer risk of obesity. Nature genetics 2008; 40:943-945

45. Nead KT, Li A, Wehner MR, Neupane B, Gustafsson S, Butterworth A, Engert JC, Davis AD, Hegele RA, Miller R, den Hoed M, Khaw KT, Kilpelainen TO, Wareham N, Edwards TL, Hallmans G, Varga TV, Kardia SL, Smith JA, Zhao W, Faul JD, Weir D, Mi J, Xi B, Quinteros SC, Cooper C, Sayer AA, Jameson K, Grontved A, Fornage M, Sidney S, Hanis CL, Highland HM, Haring HU, Heni M, Lasky-Su J, Weiss ST, Gerhard GS, Still C, Melka MM, Pausova Z, Paus T, Grant SF, Hakonarson H, Price RA, Wang K, Scherag A, Hebebrand J, Hinney A, Franks PW, Frayling TM, McCarthy MI, Hirschhorn JN, Loos RJ, Ingelsson E, Gerstein HC, Yusuf S, Beyene J, Anand SS, Meyre D. Contribution of common non-synonymous variants in PCSK1 to body mass index variation and risk of obesity: a systematic review and meta-analysis with evidence from up to 331 175 individuals. Human molecular genetics 2015; 24:3582-3594

46. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research 2011; 39:e118

47. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research 2003; 31:3812-3814

48. Guðbjartsson H, Georgsson GF, SGuðjónsson SA, Valdimarsson RT, Sigurðsson JH, Stefánsson SK, Másson G, Magnússon G, Pálmason V, Stefánsson K. *Bioinformatics*, Volume 32, Issue 20, 15 October 2016, Pages 3081–3088)

49. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hedge M, Lyon E, Spector E, Voelkerding K and Rehm H. Standards and guidelines for the interpretation of sequence variants: a joint concensus recommendation of the American College of Medical Genetics and Genomics and the Association of Molecular Pathology. Genetics in Medicine, 2015; 1-20