This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: http://orca.cf.ac.uk/102999/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:


10.1038/s41598-017-03054-8 file

Publishers page: http://dx.doi.org/10.1038/s41598-017-03054-8 <http://dx.doi.org/10.1038/s41598-017-03054-8>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Rare Variant Analysis of Human and Rodent Obesity Genes in Individuals with Severe Childhood Obesity

Audrey E. Hendricks1,2, Elena G. Bochukova3,4, Gaëlle Marenne1, Julia M. Keogh3, Neli Atanassova3, Rebecca Bounds1, Eleanor Wheeler1, Vanisha Mistry1, Elana Henning1, Understanding Society Scientific Group1, Antje Körner5,6, Dawn Muddymann1, Shane McCarthy1, Anke Hinney1, Johannes Hebebrand7, Robert A. Scott4, Claudia Langenberg8, Nick J. Wareham9, Praveen Surendran10, Joanna M. Howson9, Adam S. Butterworth9,10, John Danesh4,9,10, EPIC-CVD Consortium11, Børge G Nordestgaard11,12, Sune F Nielsen11,12, Shoia Afzal11,12, Sofia Papadia3, Sofie Ashford3, Sumedha Garg3, Glenn L. Millhauser13, Rafael I. Palomino13, Alexandra Kwasnielwska3, Ioanna Tachmazidou1, Stephen O’Rahilly3, Eleftheria Zeggini3, UK10K Consortium1, Inês Barroso2,3 & I. Sadaf Farooqi3

Obesity is a genetically heterogeneous disorder. Using targeted and whole-exome sequencing, we studied 32 human and 87 rodent obesity genes in 2,548 severely obese children and 1,117 controls. We identified 52 variants contributing to obesity in 2% of cases including multiple novel variants in GNAS, which were sometimes found with accelerated growth rather than short stature as described previously. Nominally significant associations were found for rare functional variants in BBS1, BBS9, GNAS, MKKS, CLOCK and ANGPTL6. The p.S284X variant in ANGPTL6 drives the association signal (rs201622589, MAF=0.1%, odds ratio = 10.13, p-value = 0.042) and results in complete loss of secretion in cells. Further analysis including additional case-control studies and population controls (N = 260-642) did not support association of this variant with obesity (odds ratio = 2.34, p-value = 2.59 x 10^-3), highlighting the challenges of testing rare variant associations and the need for very large sample sizes. Further validation in cohorts with severe obesity and engineering the variants in model organisms will be needed to explore whether human variants in ANGPTL6 and other genes that lead to obesity when deleted in mice, do contribute to obesity. Such studies may yield druggable targets for weight loss therapies.

Studies focused on severe early onset obesity alone, or obesity with developmental delay and/or dysmorphic features have identified a number of genes harbouring highly penetrant causal mutations1-3. The further characterisation of rare, highly penetrant variants identified in such individuals can provide insights into the cellular and

1Wellcome Trust Sanger Institute, Cambridge, UK. 2Department of Mathematical and Statistical Sciences, University of Colorado-Denver, Denver, CO, 80204, USA. 3University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. 4The Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. Center for Pediatric Research, University Children’s Hospital Leipzig, Leipzig, Germany. 5IFB Adiposity Diseases Medical Faculty, University of Leipzig, Leipzig, Germany. 6Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, University Hospital Essen and University of Duisburg-Essen, Essen, Germany. 7MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK. 8Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. 9The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge, UK. 10Department of Clinical Biochemistry and The Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark. 11Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 12Department of Chemistry & Biochemistry, University of California Santa Cruz, Santa Cruz, CA, 95064, USA. Audrey E. Hendricks and Elena G. Bochukova contributed equally to this work. A comprehensive list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to I.B. (email: ib1@sanger.ac.uk) or I.S.F. (email: isf20@cam.ac.uk)
physiological mechanisms involved in energy homeostasis and human obesity, and can identify and/or validate targets for therapeutic intervention. The aim of our study was to examine the prevalence of rare/novel variants in human and mouse obesity genes using high-throughput next-generation sequencing in a large cohort of individuals with severe early onset obesity. Here we describe results from an analysis of 119 candidate genes sequenced in 2,548 individuals with severe, early-onset obesity from the Severe Childhood Onset Obesity Project (SCOOP) (UK individuals of European ancestry recruited to the Genetics of Obesity Study, GOOS; BMI standard deviation score (SDS) > 3; onset of obesity before the age of 10 years; Methods) from the UK10K project. Our analysis includes 737 SCOOP individuals with whole-exome sequence, and 1,811 additional SCOOP individuals, not consented for whole-exome analysis, in whom we performed targeted sequencing. As this work was performed as part of a consortium (UK10K project), this data was compared to 1,117 individuals with other disorders (e.g. neurodevelopmental and rare disease) in whom exome sequencing and analysis were performed using the same methods. For the purposes of this analysis, these individuals are designated as “controls” (Methods), although we recognise that there are limitations with this design.

Study Design
SCOOP individuals likely to have congenital leptin deficiency, a treatable cause of severe obesity, were excluded by measurement of serum leptin, and individuals with mutations in the melanocortin 4 receptor gene (MC4R) (the most common genetic form of penetrant obesity) were excluded by prior Sanger sequencing. We focused on six tiers of genes (Methods and Supplementary Table 1): (1) genes known to harbour variants causing human obesity alone (Obesity Alone; n = 6); (2) genes in which known variants cause human obesity combined with developmental delay and/or dysmorphology (Obesity and Delay; n = 26); (3) genes in which loss-of-function leads to obesity in mice (LoF Mice; n = 51); (4) genes in which gain-of-function leads to obesity in mice (GoF Mice; n = 5); (5) genes encoding anorectic molecules and their receptors (Anorectic Molecules; n = 7), and (6) genes in which loss-of-function is associated with other metabolic phenotypes in mice (Complex Metabolic Effects, n = 24). Our aim was to include genes where there was sufficient evidence to indicate disorders with Mendelian inheritance (obesity syndromes) or where complete deletion or overexpression (as opposed to conditional knockouts) causes an obesity phenotype in mice (Methods) (i.e. genes in which inherited loss/gain of function variants might exist). To identify variants more likely to be causally linked to obesity, we focused on rare (MAF < 1%) and novel (not seen in the data we used for filtering) variants predicted to be functional (i.e. nonsense variants, missense amino acid substitutions, alterations of conserved splice sites or small insertions/deletions (indels) that introduced a frameshift) (Methods). We filtered our data against approximately 8,000 publicly available sequenced samples, as well as 2,097 exomes (sequenced in parallel on the same platform as our samples) and 3,781 whole-genome sequenced samples also from the UK10K project (Methods).

Human Obesity Syndrome Genes.
Firstly, we sought to identify rare and novel functional variants in the 32 genes known to cause human obesity, with or without, additional developmental delay and/or dysmorphology features (Obesity Alone, or Obesity and Delay) (Methods). We identified 11 rare potentially functional variants in ClinVar with pathogenic/likely pathogenic status and 321 novel functional variants, which we confirmed by Sanger sequencing (Fig. 1, Methods). Based on inheritance patterns (where available) and the functional properties of variants that have previously been characterised, variants in these genes may contribute to obesity, sometimes in a non-fully penetrant manner, in 52 (2%) individuals (Supplementary Figure 1 & Supplementary Table 2; Fig. 2, Methods). As variants in MC4R account for approximately 5% of severe obesity in this cohort, these findings indicate that > 90% of patients within this cohort do not have their phenotype explained by variants in known human obesity genes. Further analysis exploring the whole-exome in an agnostic manner, with appropriately matched non-disease controls not available in the UK10K project, will be an aim of future investigations.

Notably, we identified thirteen clinically associated GNAS variants in fourteen SCOOP individuals (12 of these confirmed on Sanger sequencing; 0.5%) (Supplementary Table 3). GNAS is an imprinted gene in which heterozygous loss-of-function variants are associated with obesity, short stature and skeletal abnormalities, and, when maternally inherited, hormone resistance syndromes. As GNAS sequencing has traditionally only been performed in individuals with classical clinical features, our findings suggest that the true prevalence in childhood obesity may be underappreciated. Three of the variants found in our study have been described previously in patients with classical features (p.Y163X; p.R258W, and p.R265H). We identified a novel nonsense variant (p.Y169X) predicted to remove the entire Ras-like GTPase domain and several missense variants predicted to affect downstream signalling by affecting the interaction with G-protein coupled receptors (GPCRs), G-protein β- and γ-subunits, or downstream adenylyl cyclase when mapped onto the protein structure of GNAS (Fig. 3). We confirmed maternal transmission of variants in three out of the five families where parental samples were available for genotyping. Although four GNAS variant carriers exhibited endocrinopathies and nine had developmental delay as anticipated, unexpectedly four individuals had accelerated linear growth in childhood (height SDS > 2) rather than short stature (defined as height SDS < 2.0) (Supplementary Table 3). Further molecular and physiological studies will be needed to investigate potential genotype-phenotype correlations. As studies in rodents have shown that Gnas is imprinted in the paraventricular nucleus of the hypothalamus, the location of the majority of neurons expressing the G-protein coupled receptor MC4R, it is plausible to hypothesize that some GNAS variants may contribute to obesity and accelerated linear growth by reducing melanocortin signalling.

Next we formally tested for enrichment of rare, or novel, functional variants in case-control analysis (Methods). Although no gene attained experiment-wide threshold of p-value < 2 × 10−4 (Methods and Supplementary Tables 4–9), four genes (BBS1, BBS9, GNAS, and MKKS) known to cause obesity and developmental delay and/or dysmorphology (Obesity and Delay) had nominally significant burden test p-values (p-value < 0.05) and a higher burden of variants in cases than controls (Table 1).
Mouse Obesity Genes. We found a higher burden of rare variants in cases than controls for two genes known to cause obesity when disrupted in mice - *ANGPTL6* and *CLOCK* (Table 1). Testing the functional consequences of *CLOCK* variants in cells is not straightforward as the molecular mechanisms that lead to obesity are not fully understood. As such, we focused on verifying the accuracy of the computationally predicted function of all twelve rare *ANGPTL6* variants (Methods, Fig. 4, Supplementary Table 10). All variants found in cases (but not the one variant found exclusively in controls, V143L) were predicted to affect the fibrinogen-like domain of the protein (Fig. 4a). *ANGPTL6* is predominantly expressed in the liver but is also expressed in white adipose tissue and kidney. To evaluate the effect of these variants, HEK293 cells were transiently transfected with constructs encoding wild type and variant forms of *ANGPTL6*. Whereas wild-type *ANGPTL6* was readily detected in the medium, several mutants reduced, and the S286X mutant abolished, secretion of *ANGPTL6* (Fig. 4b). Incorporation of this functional information did not qualitatively change the results of the *ANGPTL6* gene-based test (Supplementary Table 10). However, limiting the analysis to the single variant that completely abolished protein secretion (p.S286X, rs201622589) led to an adjusted OR = 10.13 and p-value = 0.028 (Fisher’s p-value = 0.041) suggesting that the association signal at this gene is primarily driven by this variant (Supplementary Table 10). To increase power, and given that there were no additional cases of European descent from our obesity cohort, we analysed existing exome-chip data from an additional 253,587 unrelated European, non-Finnish population controls (Methods, Supplementary Table 11). In an updated analysis with the 2,548 cases and 253,587 controls, we observed an odds ratio (OR) = 2.90 and a Fisher’s p-value = 0.0022, which did not reach experiment or exome-wide significance. Further analysis of 1,436 non-overlapping obesity cases and 1,954 non-overlapping controls from two studies did not lend additional support, and when combined with the original data, yielded an overall OR = 2.34 and chi-squared p-value = 0.0060 (total 3,984 cases vs 256,658 controls, Supplementary Table 11). The results are similar when limiting to non-obese controls (Supplementary Table 12). These findings highlight the difficulty in studying very rare variants in complex diseases, and potentially the effect of winner’s curse, reinforcing the need for extremely large sample sizes. Of interest, mice with targeted deletion of *Angptl6* that survive to birth (20%) develop marked obesity, have increased food intake, reduced energy expenditure, exhibit lipid accumulation in liver and muscle, and develop insulin resistance. Conversely, mice with targeted overexpression of *Angptl6* are lean, insulin sensitive and are protected from diet-induced obesity. As *ANGPTL6* is a liver-derived circulating peptide, and thus could potentially be manipulated for therapeutic purposes, further genetic studies in larger cohorts and experimental studies in mice and humans are necessary to explore its role in obesity and potential utility as an anti-obesity drug target.

Power and Gene Set Analysis. Assuming an experiment-wide significance level of $2 \times 10^{-4}$ and that 30% of rare (MAF $< 1\%$) variants within a given gene are causal we have very limited power ($\sim 20\%$) to detect an association to a gene region using our sample of 2,548 cases and 1,117 controls (Supplementary Figure 2, Online Methods, Supplementary Note 4). As there is evidence that looking across group of genes with stronger priors may increase power, we tested for association of rare or novel functional variants within each of the six
candidate gene sets. We further filtered variants with respect to their likelihood of being deleterious and performed six nested tests for each gene set (Methods). Although none of the analyses of tiered gene sets reached experiment-wide significance (Supplementary Table 13, Fig. 5), we see an increase in OR as we focus on rarer variants with more evidence of predicted deleterious effect within the Obesity Alone gene set. This suggests that focusing on extremely rare variants with strong evidence of being deleterious may be a good strategy for diseases with a complex genetic architecture, which likely includes rare, highly penetrant alleles. Our power analyses also suggests that restricting the MAF threshold to be near the very rare allele frequency of that expected for the causal variants, rather than a more lenient MAF threshold of 1%, increases power (Supplementary Figure 3). Finally, there are nominally significant signals (p-value < 0.05) within the LoF Mice candidate gene set for the novel, functional variant tests suggesting that within this rather large gene set (n = 51) there may be additional genes associated with human early onset obesity (Fig. 5).

Discussion

In summary, analysis of sequence data in 119 genes across 2,548 severely obese children and 1,117 other disease controls from the UK10K project identified variants in known obesity genes, contributing to obesity in 2% of children within this cohort, in whom mutations in MC4R and leptin had been excluded. It also highlighted that GNAS mutations may be more prevalent in human obesity than previously thought, and that the spectrum of phenotypic consequences may be broader and more heterogeneous than previously described (6). Additional detailed genotype-phenotype studies will be required to further elucidate the molecular underpinnings of the physiological consequences of these variants.

Taking the information generated in this study together with the published literature, we suggest that molecular genetic investigations, including testing known obesity genes, or agnostic exome-wide or genome-wide approaches (as these become more established at the point of care), should become part of the assessment for a child presenting with severe obesity (BMI SDS > 3) in the absence of other syndromic features. Genetic studies can allow the identification of congenital leptin deficiency which is entirely treatable with recombinant leptin therapy18. Also, early reports suggest that patients with mutations that disrupt POMC signalling may be effectively
treated with Setmelanotide (a MC4R agonist)\textsuperscript{19}. Finally, our data suggest that focusing on very rare, deleterious variants may increase power to find genetic association. Further detailed investigation of genes that lead to obesity when deleted in mice may identify new pathophysiological mechanisms involved in human obesity that can be targeted for drug discovery.

**Methods**

**Sample Sets.** All studies were approved by the Cambridge Local Research Ethics Committee and all participants and their parents gave written informed consent. All methods were performed in accordance with the relevant laboratory/clinical guidelines and regulations.

The SCOOP cohort\textsuperscript{3} was studied as part of the UK10K consortium\textsuperscript{4}. Data was compared to subsets within the neurodevelopmental and rare disease groups that were consented for use as controls. Details and further

---

**Figure 3.** Structural model of variants identified in GNAS. (A) Structure of the active state ternary protein complex of G-protein coupled receptor (GPCR) beta 2-adrenergic receptor (ADRB2), and the nucleotide-free Gs heterotrimer, composed of Gs-\(\alpha\)(GNAS), Gs-\(\beta\), and Gs-\(\gamma\) subunits (based on Rasmussen et al.\textsuperscript{20}, pdb file: 3sn6). The two major domains of GNAS are noted, \(\alpha\)-helical domain and Ras-like GTPase domain. The principal interactions between GNAS and GPCRs involve the amino- and carboxy-terminal \(\alpha\)-helices. The Ras-like GTPase domain contains most of the catalytic residues necessary for GTP hydrolysis, as well as the G\(\beta\)\(\gamma\) and effector binding regions (switch regions I, II and III), which change confirmation upon binding to GTP or GDP). (B) Structure of the complex of Gs-alpha with the catalytic domains of mammalian adenylyl cyclase (based on Tesmer et al.\textsuperscript{21}, pdb file: 1azs). Variant residues (purple), and the different components of the complex are highlighted.
info about the UK10K project can be found at [http://www.uk10k.org/](http://www.uk10k.org/) and in the UK10K consortium paper from 2015.4

We gathered the genotypes for the variant S284X (rs201622589) in 202,981 unrelated population controls with European non-Finnish ancestry from existing cohorts (UKHLS22, Fenland, EPIC Norfolk23, CCHS24–26, CGPS24–26, CHD524–26, EPIC-CVD37, UK Biobank28 and ExAC29) and in two additional studies of obese children and healthy or thin controls30,31 including one from the Leipzig Childhood Obesity Cohort32. (Supplementary Note 1)

### Sequencing, Variant Calling, and Quality Control

Targeted Sequencing (TS) and Whole Exome Sequencing (WES) was performed as described elsewhere (refs 33 and 4 respectively). All UK10K WES samples from all arms of the exome study available at the time of the variant calling and SCOOP TS samples (N = 5,233 and 2,819 respectively) were called together on the non-redundant targets +/− 100 bp using multi-sample calling. Using SAMtools34, a BCF file was created and the site genotype likelihoods were calculated. Variants (SNPs and Indels) were called using BCFtools. Indels were left-aligned using ‘vcf norm’ from the htslib package. Variants were filtered for sequencing and genotype quality at the site and genotype level using vcf-annotate35. Additional details in Supplementary Note 2.

### Variant Annotation

All variant annotation was applied using the GRCh37 human reference. Variants were annotated with rsIDs from dbSNP 137, and allele frequencies from the final 1000Genomes Phase 1 integrated (v3) callset36, the NHLBI Exome Sequencing Project (ESP) v237, and the UK10K WGS sample set4. The Ensembl Variant Effect Predictor (http://www.ensembl.org/info/docs/variation/vep/index.html)38 v.2.8 with Ensembl 66 was used to add variant consequence annotations including the predicted deleteriousness of each missense variant as predicted by SIFT39,40, PolyPhen41,42, and Condel43. For the 32 known human obesity genes, only the variant consequences on the most clinically relevant transcript were considered for further analysis. For the other 87 candidate obesity genes, the most severe consequence on any transcript was retained for each variant (Supplementary Table 1). Using Sequence Ontology terms44,45, variant consequences were defined as functional (i.e. essential splice site, stop gained, stop lost, complex indel, frameshift coding, non synonymous coding, within mature miRNA, partial codon).

To identify rare variants, we used the thirteen UK10K WES sample sets that were not obese sample sets and were not included as controls in this study (Supplementary Note 1). These sample sets were processed in parallel with our cases and controls and thus provide a similar sequencing coverage and depth across the called regions. We removed variants that had a MAF > 1% across all or a MAF > 10% in any of the thirteen sample sets. We then removed variants with a MAF > 1% in any of seven additional sample sets: 1. UK10K WGS sample set4, 2–3. European and African American NHLBI ESP v2 sample sets37, and 4–7, the four continent sample sets (AFR, AMR, ASN, EUR) from 1000 Genomes Phase 1 integrated v3 call set38. Novel variants were identified as sites not seen in any of the internal or external datasets used for MAF annotation, i.e. novel at the time of this study.

### Sample quality control

Samples were identified as contaminated using a combination of two methods: VerifyBamID v1.0.046 and “fraction skewed hets”47. For the TS samples, the FREEMIX value was estimated using 11,250 high quality, autosomal, biallelic SNPs with an alternate AF ≥ 0.01, and a call rate ≥ 0.5 in both the TS sample set and the 1000Genomes Project Phase 1 v3.36. Of the 2,819 TS SCOOP samples, the 784 WES SCOOP samples, and the 1427 WES samples used as controls 37, one, and nine were excluded due to contamination

<table>
<thead>
<tr>
<th>Gene</th>
<th>Tier</th>
<th># Variants</th>
<th>Rare alleles in SCOOP TS (N = 1,811)</th>
<th>Rare alleles in SCOOP WES (N = 737)</th>
<th>Rare alleles in Control WES (N = 1,117)</th>
<th>OR</th>
<th>Adjusted OR</th>
<th>p-value SKATO</th>
<th>p-value BURDEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare Functional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBS1</td>
<td>Obesity and Delay</td>
<td>18</td>
<td>36</td>
<td>18</td>
<td>10</td>
<td>2.368</td>
<td>0.022</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>BBS9</td>
<td>Obesity and Delay</td>
<td>20</td>
<td>24</td>
<td>5</td>
<td>5</td>
<td>2.543</td>
<td>0.081</td>
<td>0.045</td>
<td></td>
</tr>
<tr>
<td>GNAS</td>
<td>Obesity and Delay</td>
<td>14</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>Inf</td>
<td>14.505</td>
<td>0.015</td>
<td>0.008</td>
</tr>
<tr>
<td>MKKS</td>
<td>Obesity and Delay</td>
<td>19</td>
<td>90</td>
<td>25</td>
<td>28</td>
<td>1.801</td>
<td>0.006</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>ANGPTL6</td>
<td>Mouse LOF</td>
<td>12</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>3.073</td>
<td>0.043</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Novel Functional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNAS</td>
<td>Obesity and Delay</td>
<td>13</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>Inf</td>
<td>13.629</td>
<td>0.019</td>
<td>0.010</td>
</tr>
<tr>
<td>CLOCK</td>
<td>Mouse LOF</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>Inf</td>
<td>9.208</td>
<td>0.066</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 1. Summary of case-control results. Summary of case-control results in genes with nominally significant (p-value < 0.05, bold) burden of Rare (top) or Novel (bottom) functional variants in SCOOP cases (N = 1,811 with targeted sequence and N = 737 with whole-exome sequence), compared to 1,117 controls with whole-exome sequence data from the UK10K project. The number of variants (# variants) per gene are shown, as well as the number of alleles in cases and controls, odds ratios (OR) and p-values from SKAT-O (p-value SKATO) and from burden tests (p-value BURDEN). For variants not detected in controls an adjusted odds ratio (Adjusted OR) was calculated by adding 0.5 to the number of alleles in each cell of the two-by-two table.
respectively. Sixteen non-contaminated TS samples were excluded based on a mean sample read-depth 3 SD below the average for all samples (i.e. mean sample read-depth < 12.09). Genotype concordance for 436 WES SCOOP samples and 1,035 TS SCOOP samples for which we had both sequence and GWAS data was also calculated. Four WES samples and six TS samples with a concordance rate below 90% were identified and excluded. A set of highly polymorphic markers (MAF > 0.3) was genotyped and compared to the sequencing calls from each sample. Non-concordant WES samples were removed prior to variant calling and are not included in the original sequencing numbers. Eighty-one TS samples with low concordance were removed after variant calling. Three TS samples were removed due to having an extremely high genotype missing rate of > 50%.

To identify non-European samples, we calculated principal components (PCs) from the 1000Genomes Phase I integrated call set using either EIGENSTRAT v4.2 or LASER 2.0 for the WES and TS samples respectively (Supplementary Note 3). Of the 2,676 TS samples, 837 were classified as non-European due to genetic ancestry and three were excluded due to a reported non-European ancestry resulting in 1,836 good quality TS samples of European ancestry. Of the 779 WES SCOOP samples, 37 samples were classified as non-European due to genetic ancestry resulting in 742 good quality WES SCOOP samples of European ancestry. Of the 1,418 control samples, 146 were identified as being non-European leaving 1,272 high quality WES controls.

Known and cryptic relatedness was identified by estimating pairwise identity by descent using PLINK v1.07. Genetic relationships within the WES sample set is described in detail elsewhere. We removed individuals sequentially by: (1) largest number of relationships, (2) diseased controls, non-diseased controls, cases, (3) lower mean depth of sequencing. Five SCOOP cases and 155 controls were removed resulting in 737 SCOOP WES cases and 1,117 WES controls all of high quality, unrelated, and European ancestry. A similar process was used to identify genetic relationships within the TS sample set and between the TS and WES sample sets (Supplementary Note 3). This resulted in 25 additional exclusions (18 TS cases related to WES cases and seven related within the TS sample set) for a total of 1,811 high quality, unrelated, TS samples of European ancestry.
Candidate Gene Sets. We performed database searches using keywords ‘obesity’, ‘growth’, ‘size’, ‘adipose tissue’, (details below) and manually curated the results to arrive at a set of six tiers of candidate gene with definite or likely links to obesity. The Online Mendelian Inheritance of Man (OMIM database, http://www.omim.org/) accessed in February, 2013 was used to identify 32 genes (Supplementary Table 1) which directly lead to human obesity. These were further categorised into two groups based on consistent clinical features: genes characterised by “Obesity alone” (Obesity Alone; n = 6), and “Obesity and developmental delay and/or dysmorphism” (Obesity and Delay; n = 26). Any putative list of candidate genes has limitations and other genes that contribute to energy homeostasis/obesity in animal models deserve interrogation in the future.

Genes that cause obesity when disrupted in rodents were identified using the Mouse Genome Informatics Database (MGI, http://www.informatics.jax.org/) and the Rat Genome Database (RGD, http://rgd.mcw.edu) both accessed in February 2013, as well as published research (www.ncbi.nlm.gov/Pubmed). Eighty-seven genes were associated with a growth/size phenotype when perturbed in rodent animal models. We evaluated the strength of evidence for a role of each gene in obesity phenotype and divided them into four further categories, specifically: Loss of function associated with obesity in mice (LoF Mice; n = 51); gain of function associated with obesity in mice (GoF Mice, n = 5); anorectic peptides and receptors (Anorectic Molecules, n = 7); loss of function associated with other metabolic phenotypes in mice (Complex Metabolic Effects, n = 24) (Supplementary Table 1).

Sequence validation of rare functional variants. Novel, functional variants in all 119 candidate genes were taken forward for validation by conventional Sanger sequencing (90% validation rate). Briefly, customized PCR primers were designed +/− 250 bp surrounding the variant; and sequencing was performed using BigDye Terminator v3 kit (Applied BioSystems) and analysed by capillary electrophoresis on an ABI3730 DNA Analyzer platform (Applied Biosystems), according to the manufacturer’s instructions. Familial segregation analysis of variants was performed where family samples were available.

Identification of clinically-associated variants in known human obesity genes. The strategy for identification of clinically-associated variants in known human obesity genes is outlined in Fig. 1. Briefly, novel and rare variants in the known human obesity genes were filtered using the ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/). We retained novel variants and those with a ClinVar status of Pathogenic/Likely pathogenic (N = 332).

Structural Analysis of GNAS Variants. For structural analysis of the genetic variants and prediction of their impact on protein function, we modelled their location in relation to previously solved protein crystal structures of GNAS in complex with either G-protein coupled receptor (ADRB2) or downstream effector (adenylyl cyclase). Specifically, the first model represented in Fig. 3a is based on crystal structure of the active ternary protein complex of GPCR beta-2-adrenergic receptor (ADRB2), and the nucleotide-free Gs heterotrimer, composed of Gs-α(GNAS), Gs-β, and Gs-γ subunits (ref. 20, PDB number: 3SN6, www.rcsb.org). The second model, represented in Fig. 3b, is based on complex of GNAS with the catalytic domains of mammalian adenylyl cyclase (ref. 21, PDB...

**Association Analyses.** We implemented the optimal Sequence Kernel Association Test SKAT-O\(^1\)\(^,\)\(^5\)\(^,\)\(^3\) using the SKAT R package v1.2.24\(^*\) using options method = “optimal.adj” for SKAT-O and rcorr = 1 for burden. For each gene and candidate gene set, we performed two primary tests using the burden test: 1. rare functional (Rare Func) or (2) novel functional (Novel Func). We performed secondary tests restricting either to variants that were predicted to be deleterious by at least one of the three algorithms (i.e. SIFT, PolyPhen, and Condel) (1 + del) or to variants that were predicted to be deleterious by all three algorithms (all del). For SIFT and Condel, variants were classified as deleterious if they were labelled as “deleterious”. For PolyPhen, variants were classified as deleterious if they were labelled as “probably damaging” or “damaging”. Our conservative Bonferroni adjusted significance level of the primary analyses for 119 genes and 6 gene sets was 0.05/(2\(^119\)\(\times\)6) = 2.0E-4. Once considering our additional ten secondary tests, our Bonferroni adjusted significance level was 0.05/(12\(\times\)25) = 3.3E-5. We repeated all primary and secondary tests using SKAT-O (Table 1; Supplementary Table 4–9, 13).

**Functional studies of variants in ANGPTL6.** To predict the impact of variants on the ANGPTL6 protein, we modelled their location using the crystal structure of a related protein fibrinogen in a complex with a peptide Gly-His-Pro-amide (described in Yang et al.\(^8\), PDB number: ILWU, www.rcsb.org). The structural representations using ribbon-depicted models were generated using the Open-Source PyMOL Molecular Graphics System, Version 1.7.x Schrödinger, LLC (http://pymol.org). N-terminal triple Flag tag was added to human ANGPTL6 cDNA cloned into a PEZ-M14 mammalian expression vector (Capital Bioscience). Mutations were introduced into this construct using QuikChange (Agilent technologies), and confirmed by Sanger sequencing. ANGPTL6 protein expression was studied by transient transfection in HEK293 cells. Media was collected 48 h – post transfection, the cell medium was centrifuged for 5 min (5,000 g at 4 °C) and the supernatants were collected. Cells were harvested prior to centrifugation for 15 min (15,000 g at 4 °C). Aliquots from the medium and cells were subjected to SDS-PAGE and immunoblot analysis. Primary antibodies (monoclonal Flag M2 antibody (Sigma) and a polyclonal antibody to Calnexin (Cell signaling)) were used at 1:1000, and secondary Horseradish peroxidase–conjugated anti-mouse or anti-rabbit IgG (Dako) antibody at 1:2000 dilution. After staining with ECL West Dura Substrate kit (Thermo Scientific), visualisation was performed on Chemidoc Digital Image (Bio-Rad) (Supplementary Information). The relative protein secretion of ANGPTL6 for each variant was standardized to wild-type protein secretion. Experiments were performed in triplicate and analysed using an unpaired T-test with Welch’s correction.

**ANGPTL6 statistical follow-up analyses.** We ran gene-based analyses using SKAT-O as described above on two subsets of ANGPTL6 variants (Supplementary Table 10): (1) the one variant that resulted in a complete loss of protein secretion (S284X; rs201622589), and (2) excluding S284X. For (1), we also used a Fisher’s Exact Test. We repeated the Fisher’s Exact Test of the S284X variant including additional population controls with pre-existing genotype data and replication samples (Online Methods, Sample Sets; Supplementary Tables 11 and 12).

**Power Analysis.** We calculated the power to detect association to a gene region using the burden test with the Power_Logistic_R function and the haplotype dataset within the SKAT R-package\(^8\). We performed power calculations using 500 simulations on a random 2Kb sub region. The effect sizes of the causal variants are equal to log\(_2\)(MAF) with a maximum effect size of 1.6 (MAF = 0.0001) and all have the same direction of effect. We limited the MAF threshold for causal variants to \(\leq 0.01, 0.001, 0.0005\) and varied the percentage of causal variants in the region to be between 10–90%. (More details in Supplementary Note 4).

**References**


---

**Note:** The asterisk (*) indicates the version of the software used for the analysis. The specific version of each package is provided for reproducibility purposes.
We are indebted to the patients and their families for their participation and to the physicians involved in the Genetics of Autism Study (GOAS). This work was supported by the Wellcome Trust (ISF, IB) (098497/2/Z/12/Z, WT090851/M), Medical Research Council (ISF, SOR) (MRC_MC_UK122/5), NIHR Cambridge Biomedical Research Centre (ISF, IB, SOR), Bernard Wolfe Health Neuroscience Endowment (ISF), European Research Council (ISF) and NIH grant DK064265 (GLM), the European Community's Seventh Framework Programme (FP7/2007–2013) project Beta-JUDO n°279153 (ISF, AK). This study comprises one arm of the UK10K Consortium (WT091310). The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The survey was conducted by NatCen and the genome-wide scan data were analysed and deposited by the Wellcome Trust.
Sanger Institute. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/. AH and JH were funded by the German Ministry for Education and Research (National Genome Research Net-Plus 01GS0820), the German Research Foundation (DFG: HI865/2-1), the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n°245009 and n°262055. We thank participants and staff of the Copenhagen City Heart Study, Copenhagen Ischemic Heart Disease Study, and the Copenhagen General Population Study for their important contributions (CCHS, CGPS, CIHDS). CHD case ascertainment and validation, genotyping, and clinical chemistry assays in EPIC-CVD were supported by grants awarded to the University of Cambridge from the EU Framework Programme 7 (HEALTH-F2-2012-279233), the UK Medical Research Council (G0800270) and British Heart Foundation (SP/09/002), the European Research Council (268834), the UK National Institute for Health Research Cambridge Biomedical Research Centre, Merck and Pfizer. We thank all EPIC participants and staff for their contribution to the study, the laboratory teams at the Medical Research Council Epidemiology Unit for sample management and Cambridge Genomic Services for genotyping, Sarah Spackman for data management, and the team at the EPIC-CVD Coordinating Centre for study coordination and administration (EPIC-CVD), JD is supported by a British Heart Foundation Professorship, NIHR Senior Investigator, European Research Council Senior Investigator. Leipzig LIFE Child as part of the Leipzig Childhood Obesity cohort was funded by the European Union, by the European Regional Development Fund (ERFD) by means of the Free State of Saxony within the framework of the excellence initiative. This research has been conducted using the UK Biobank Resource.

Author Contributions

Additional Information
Supplementary information accompanies this paper at doi:10.1038/s41598-017-03054-8

Competing Interests: The authors declare that they have no competing interests.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017
Consortia
Understanding Society Scientific Group

Michaela Benzeval14, Jonathan Burton14, Nicholas Buck14, Annette Jäckle14, Meena Kumari14, Heather Laurie14, Peter Lynn14, Stephen Pudney14, Birgitta Rabe14 & Dieter Wolke15

14Institute for Social and Economic Research, University of Essex, Colchester, UK. 15University of Warwick, Warwick, UK.

UK10K Consortium


1The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HW, Cambridge, UK. 16Department of Pathology, King Abdulaziz Medical City, P.O. Box 22490, Riyadh 11426, Saudi Arabia. 17Department of Psychiatry, Trinity Centre for Health Sciences, St James Hospital, James's Street, Dublin 8, Ireland. 18Genetics and Genomic Medicine and Birth Defects Research Centre, UCL Institute of Child Health, London WC1N 1EH, UK. 19Departments of Health Sciences and Genetics, University of Leicester, Leicester LE1 7RH, UK. 20Division of Developmental Disabilities, Department of Psychiatry, Queen's University, Kingston, Ontario N6C 0A7, Canada. 21Department of Cardiovascular Medicine and Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford
Genetics Service, Kennedy-Galton Centre, Northwick Park Hospital, Watford Road, Harrow HA1 3UJ, UK. 96University College London (UCL) Genetics Institute (UGI) Gower Street, London WC1E 6BT, UK. 97Connective Tissue Disorders Service, Sheffield Diagnostic Genetics Service, Sheffield Children’s NHS Foundation Trust, Western Bank, Sheffield S10 2TH, UK. 98Centre for Genomic and Experimental Medicine, Institute of Genetics and Experimental Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. 99Molecular Genetics, Vipath at Guy’s Hospital, London SE1 9RT, UK. 100ALS PAC & School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. 101Human Genetics Department, Radboudumc and Radboud Institute for Molecular Life Sciences (RIMLS), Geert Grooteplein ZN, Nijmegen 6525 HP, The Netherlands. 102Department of Clinical Genetics, Great Ormond Street Hospital, London, WC1N 3JH, UK. 103Clinical Genetics, Guy’s & St Thomas’ NHS Foundation Trust, London SE1 9RT, UK. 104Mackenzie Building, Kirsty Semple Way, Ninewells Hospital and Medical School, Dundee DD2 4RB, UK. 105Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK. 106School of Oral and Dental Sciences, University of Bristol, Lower Maudlin Street, Bristol BS1 2LY, UK. 107School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1UQ, UK. 108National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester LE3 9QP, UK. 109Maritime Medical Genetics Service, SB50/S590 University Avenue, PO Box 9700, Halifax, Nova Scotia B3K 6R8, Canada. 110Queensland Brain Institute, University of Queensland, Brisbane, Queensland 4072, Australia. 111Princess AI Jawhar Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, P.O. Box 80200, Jeddah 21589, Saudi Arabia. 112Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau 999078, China. 113Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. 114The Centre for Translational Omics – GOSgene, UCL Institute of Child Health, London WC1N 1EH, UK. 115School of Medicine and Pharmacology, University of Western Australia, Perth, Western Australia 6009, Australia. 116Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands, Western Australia 6009, Australia.

EPIC-CVD consortium


111Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus, Denmark. 112Department of Cardiology, Aalborg University Hospital, Aalborg, Denmark. 113Diet, Genes and Environment, Danish Cancer Society Research Center, Copenhagen, Denmark. 114INSERM, Centre for Research in Epidemiology and Population Health (CESP), U1018, Nutrition, Hormones, and Women’s Health Team, Institut Gustave Roussy, Villejuif, France. 115Division of Cancer Genetic Epidemiology, German Cancer Research Centre (DKFZ), im Neuenheimer Feld 281, 69121 Heidelberg, Germany. 116Department of Epidemiology, German Institute of Human Nutrition (DIfE), PotsdamRehbrücke, Germany. 117WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece. 118Hellenic Health Foundation, Athens, Greece. 119ARC, Lyon, France. 120Molecular and Nutritional Epidemiology Unit, Centro per lo Studio e la Prevenzione Oncologica-Scientific Institute of Tuscany, Florence, Italy. 121Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. 122Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy. 123Cancer Registry and Histopathology Unit, Civic- M.P. Arezzo Hospital, ASP Ragusa, Italy. 124Human Genetics Foundation, Turin, Italy. 125Department of Medical Sciences, University of Turin, Italy. 126Centre for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. 127Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. 128Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway. 129Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway. 130Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. 131Genetic Epidemiology Group, Folkhälsan Research Center, Helsinki, Finland. 132Public Health Directorate, Asturias, Spain. 133Public Health Institute of Navarra, Pamplona, Spain. 134Red de Investigación en Servicios de Salud en Enfermedades Crónicas, Madrid, Spain. 135Epidemiology Department, Murcia Health Authority, Murcia, Spain. 136Centre de Investigación Biomédica en red en Epidemiología y Salud Pública, Barcelona, Spain. 137Public Health Division of Gipuzkoa, Instituto Bio-Donostia, Basque Government, CIBERESP, Spain. 138CRR, Lund University, 20502 Malmö, Sweden. 139Department of Public Health and Clinical Medicine, Family Medicine, Umeå University, Umeå, Sweden. 140Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. 141German Cancer Research Center (DKFZ), Division of Cancer Epidemiology, Heidelberg, Germany.