Published in final edited form as:
Nat Genet. 2013 May ; 45(5): 513–517. doi:10.1038/ng.2607.

Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity

Eleanor Wheeler[1], **Ni Huang**[1,*], **Elena Bochukova**[2,*], **Julia M. Keogh**[2], **Sarah Lindsay**[1], **Sumedha Garg**[2], **Elana Henning**[2], **Hannah Blackburn**[1], **Ruth J.F. Loos**[3], **Nick J Wareham**[3], **Stephen O'Rahilly**[2], **Matthew E. Hurles**[1], **Inês Barroso**[1,2], and **I. Sadaf Farooqi**[2]

[1]Wellcome Trust Sanger Institute, Cambridge, UK

[2]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK

[3]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK

Common and rare variants associated with body mass index (BMI) and obesity account for <5% of the variance in BMI. We performed SNP and CNV association analyses in 1,509 children with obesity at the extreme tail (>3 standard deviations) of the BMI distribution and 5380 controls. Evaluation of 29 SNPs ($p<10^{-5}$) in an additional 971 severely obese children and 1990 controls identified four new loci associated with severe obesity (*LEPR*, *PRKCH*, *PACS1*, *RMST*). A previously reported 43kb deletion at the *NEGR1* locus was significantly associated with severe obesity ($p=6.6\times10^{-7}$). However, this signal was entirely driven by a flanking 8kb deletion; absence of this deletion increased risk for obesity ($p=6.1\times10^{-11}$). We found a significant burden of rare, single CNVs in severely obese cases ($p<0.0001$). Integrative gene network pathway analysis of rare deletions indicated enrichment of genes affecting GPCRs involved in the neuronal regulation of energy homeostasis.

Within a population that shares the same environment, a significant proportion of the variance in body mass index (BMI) is genetically determined[1]. Meta-analyses of genome-wide association studies (GWAS) for obesity-related traits (including BMI, waist-hip ratio and obesity) have led to the discovery of at least 52 loci to date[2]. However, the fraction of BMI variance that can be attributed to common SNPs is at the most 2%[2]. Candidate gene studies in severe obesity have led to the identification of several rare, highly penetrant mutations involving the hypothalamic leptin-melanocortin pathway, with variants in the melanocortin 4 receptor (*MC4R*) being the most common, with a population prevalence of 1/1000[1]. Although epistatic and gene-environment interactions may contribute to the unexplained heritability of obesity, it seems likely that a significant fraction is due to missing loci or established loci that have not yet been fully characterised[3].

Correspondence should be addressed to: I. Sadaf Farooqi (isf20@cam.ac.uk) and Inês Barroso (ib1@sanger.ac.uk)..
[*]These authors contributed equally.

AUTHOR CONTRIBUTIONS
I.S.F., I.B., M.E.H., E.W., N.H. and E.G.B. designed the study, analysed the data and wrote the paper. All the authors contributed to data collection and/or processing, as well as the current version of the paper. The authors have no competing financial interests.

As studying individuals at the extremes of a quantitative phenotype distribution has established utility[4] we performed a genome-wide association study (GWAS) in 1509 UK Caucasian children with severe early onset obesity (SCOOP-UK) and 5380 WTCCC2 UK controls[5] to identify novel associations of common and low-frequency SNPs and CNVs with obesity (**Methods**). Analysis of ~2 Million directly genotyped or imputed SNPs, under a log-additive model, revealed 29 loci with evidence of association with obesity ($p<1\times10^{-5}$) (**Supplementary Table 1**). These 29 SNPs were taken forward for validation in additional 971 severely obese children and 1990 controls (**Methods**). Joint analysis of discovery and validation stages, including up to 2480 cases and 7370 controls identified nine genome-wide significant signals in eight loci ($p<5\times10^{-8}$; **Table 1**; **Figure 1**). Three mapped in or near *PRKCH*, *PACS1* and *RMST* and have not been previously associated with obesity **(Figure 1 and Figure 2)**. These genes are ubiquitously expressed (**Supplementary Figure 1**) and to date their role in energy homeostasis is unknown. A significant finding was that of a novel association with an intermediate frequency allele (rs11208659, 6% MAF in CEU HapMap) in the gene encoding the leptin receptor (*LEPR)*. Homozygous loss of function *LEPR* variants can lead to monogenic obesity[1]. To test whether this signal was driven by rare causal mutations, we sequenced all cases carrying at least one predisposing (C) allele on whom DNA was available (286/295); none had homozygous *LEPR* mutations and the three rare non-synonymous changes identified are unlikely to be sufficient to drive the association signal (data not shown), arguing against the *LEPR* signal being driven by causal mutations and suggests that a more common *LEPR* allele can influence predisposition to severe obesity. An analysis collapsing directly genotyped low-frequency SNPs (MAF <5%) across the discovery sample set did not find a statistically significant accumulation of rare SNPs in any loci apart from *LEPR*. However, the rare variant association signal at the LEPR locus was driven by the lead SNP (rs11208659); adjusting for this SNP abolished the association signal (data not shown). We also found that the risk allele (C) at the *LEPR* locus is associated with lower LEPR expression in monocytes[6] (p = 0.0321, **Supplementary Figure 2**). Given the established importance of LEPR-mediated signalling in energy homeostasis, this might suggest that the association with obesity may be mediated by a reduction in levels of LEPR expression.

The remaining four obesity associated loci (in or near, *FTO*, *MC4R*, *TMEM18* and *NEGR1*) have been reported previously[2]. As our cases lie at the very extreme end of the BMI distribution, there are very few other cohorts of sufficient size and comparable severity of obesity for replication. Consistent with this, on comparing our genome-wide significant results to other GWAS for childhood obesity[7,8] and for varying degrees of obesity in adults and children[9-16] **(Supplementary Table 2),** we found there was relatively little overlap, other than for *FTO*, *TMEM18 and MC4R*. To explore the differences between our results and those from large-scale meta-analysis of BMI in population-based cohorts, we compared our results with those obtained from the GIANT consortium (Supplementary Tables 3 and 4, Supplementary Figures 3 and 4). For several GIANT loci we did not obtain genome-wide significant support for association in our cohort with severe early onset obesity although 14 loci had directionally consistent associations at nominal significance levels (p<0.05), more than expected by chance $p_{\text{binomial}}= 6.10\times10^{-5}$ (**Supplementary Table 4**, **Supplementary Figure 4.** Some of these differences are accounted for by differences in power as our sample size is much smaller than that of GIANT; power calculations showed that we had < 80% power to detect signals other than *FTO*, *TMEM18* and *MC4R* at genome-wide levels of significance (**Methods**). This observation may also reflect differences in susceptibility to early versus adult onset obesity or in alleles associated with BMI in the obese vs severely obese range **(Supplementary Table 2).** Indeed the data suggest that while there is significant overlap between the loci influencing BMI and common obesity and those influencing severe obesity, this overlap is incomplete and the relative contribution of each locus to common and severe forms of obesity also differs. For example, while *FTO* is the locus with the largest reported effect size in several population studies, our results show that there are other loci with comparable or greater effect sizes on severe early onset obesity (**Table 1**). Also, while *NEGR1* (tagging the 40kb deletion) and *SH2B1*, both have similar effect sizes in GIANT, the effect of *SH2B1* is much smaller in SCOOP (**Supplementary Table 4, Supplementary Figure 4**). Construction of a risk score with the SNPs from GIANT[17] available in our data also suggests that the severely obese cases seen in SCOOP do not have an increased burden of established BMI loci (mean risk score = 27.2), suggesting they may be due to different risk alleles. Similarly, data from GIANT do not support the association of the new SCOOP loci with BMI in population-derived cohorts (**Supplementary Table 3, Supplementary Figure 3**). However, in the recent GIANT discovery sample (N=123,865), only 167 individuals would have a comparable BMI to SCOOP cases, so the lack of replication does not preclude an effect of these variants on severe obesity.

We analysed the SNP data to examine whether there was an excess in cases of homozygosity by descent (HBD), either genome-wide, or in specific genomic locations (**Methods**). We did not observe any significant genome-wide homozygosity burden in cases (**Supplementary Figure 5**). When the distribution of regions of HBD across individual genes was considered, there were a number of overlapping regions of HBD in cases, across genes in regions of chromosomes 2, 8, 10, with nominal p<0.01, although these did not survive correction for multiple testing (**Supplementary Figure 5**). These findings suggest that individually, and cumulatively, homozygously-acting genes are unlikely to play a major role in the etiology of severe obesity in our outbred UK cohort.

We assessed the role of common and rare CNVs in severe early onset obesity. Our strongest association signal ($p = 6.1\times10^{-11}$), also supported by the SNP data, was of a protective ~8kb deletion upstream of *NEGR1*, with a weaker positive association signal (p=6.6×10^{-7}) from a larger non-overlapping ~43kb deletion (72,541-72,584 kb) flanking the ~8kb deletion (**Figure 3**). These deletions are known to segregate on distinct haplotypes, and as the three alleles (43kb deletion, 8kb deletion and undeleted) are mutually exclusive, we tested whether the two deletion alleles were independently associated with severe obesity. The frequency of the undeleted allele is approximately the same in cases and controls and is not associated with the phenotype (OR (95% CI) =1 (0.90-1.1); p = 0.93, two-sided Fisher's exact test). When conditioned on the smaller deletion allele, the association of the larger deletion allele was completely abolished (OR (95% CI) = (0.97-1.22); p=0.16). However, when conditioned on the larger allele, the association of the smaller deletion remained significant (OR (95% CI) =0.70 (0.60-0.82); p=6.93×10^{-6}), suggesting that the association signal at the *NEGR1* locus in our cohort is largely driven by the protective effect of the ~8kb deletion allele. In keeping with these findings, we found that the SNP with the strongest association signal for obesity at this locus (rs1993709) tags the 8Kb deletion (**Table 1**) and conditional analyses performed at the SNP level are consistent with these findings (data not shown), highlighting the strength of the combined SNP and CNV analysis.

Although the 8kb deletion does not disrupt the coding sequence of any gene, it encompasses a single conserved transcription factor binding site for NKX6.1[18] (**Figure 4a and 4b**), which is known to be involved in neuronal development in the mid and hindbrain. We used Electrophoretic Mobility Shift Assay (EMSA) to confirm binding of NKX6.1 to its predicted binding site within the 8kb deletion (**Figure 4c**). NKX6.1 is a potent transcriptional repressor, so loss of binding would be predicted to lead to increased gene expression. In existing eQTL data, the small deletion appears to be associated with increased expression of *NEGR1* in lymphoblastoid cell lines (p=0.0427, **Supplementary Figure 6**). Further studies will be needed to establish whether *NEGR1* is transcriptionally regulated by NKX6.1 in relevant human tissues. As NEGR1 is involved in neurite outgrowth[19], a process known to be involved in human obesity[1], these observations could form the basis for a molecular explanation for this association, which needs to be formally tested.

We previously demonstrated a significant burden of large rare deletions in patients with severe-early onset obesity and developmental delay[20]. Here we used a permutation-based method to assess the statistical significance of genome-wide CNV burden **(Methods)**. A significant enrichment in cases was observed for all CNVs >100kb in size and < 1% in frequency (**Table 2; Supplementary Table 5**). The most significant enrichment was observed in duplications in the range of 100-200kb and deletions >100kb in size. The enrichment of singleton CNVs is stronger than that of rare recurrent CNVs (1.8-fold increase, p=0.0001). We focused on the large number of obesity-specific deletions that were not detected in controls (765).-Whilst CNVs may contribute to disease by affecting regulatory regions of the genome, given the large number of rare CNVs identified in this study, we focused on CNVs that disrupt the exon sequence of 490 genes (in 336 patients) where we can more directly infer pathogenicity due to dosage sensitivity. To examine whether these 490 genes converged on specific biological pathways, we conducted pathway enrichment analysis using Ingenuity Pathways Analysis (IPA); genes involved in nervous system development and/or function were significantly overrepresented (p=9.83×10 $^{-5}$) (**Supplementary Table 6**). The most significant enrichment involved a network of 13 genes encoding G-protein coupled receptors (GPCRs) (**Supplementary Figure 7**); these CNVs were validated by an independent method, MLPA (data not shown). Several of the genes disrupted by these CNVs play an important role in energy homeostasis in animals, the mu-opioid receptor 1 (OPRM1), the follicle-stimulating hormone receptor (FSHR) and the oxytocin receptor (OXTR). In addition, genetic variation in the dopamine receptor D2 (DRD2) has been associated with eating behaviour and brain activation responses to images of food in humans[21]. We independently analysed the same dataset using another method which corrects for CNV size[22] and replicated the observation that rare CNVs deleting genes encoding GPCRs were significantly associated with severe obesity (**Supplementary Table 7**).

In conclusion, by combining SNP and CNV analysis and focussing on severe obesity, we have identified four novel obesity susceptibility loci, including an intermediate frequency variant in *LEPR*. These findings add to the body of evidence suggesting that both common and rare variants around specific genes/loci (*LEPR, POMC, MC4R, BDNF, SH2B1*) are involved in the pathogenesis of obesity. We show that there is an incomplete overlap between loci influencing risk of severe obesity and those that influence more common obesity, as detected by studies in population-based cohorts, and that the relative contribution of each locus to severe versus common obesity also differs. Furthermore, we provide evidence that severe obesity without developmental delay is associated with a significantly increased burden of rare, typically singleton CNVs, in parallel with findings in intellectual disability. Using pathway analysis, we found that rare CNVs that delete genes involved in the neuronal regulation of energy homeostasis contribute to severe obesity; looking for rare coding variants in these genes may be fruitful. As we observed a significant enrichment for CNVs that deleted GPCRs, which are key targets for drug development, these findings may have potential therapeutic implications.

URLs

Wellcome Trust Case-Control Consortium , http://www.wtccc.org.uk/; Wellcome Trust Case Control Consortium 2, http://www.wtccc.org.uk/ccc2/; Stata 11, http://www.stata.com/; Quanto v 1.2.3, http://hydra.usc.edu/gxe/; CCRaVAT, Case-Control Rare Variant Analysis Tool, http://www.sanger.ac.uk/resources/software/rarevariant/; Database for Annotation, Visualization and Integrated Discovery (DAVID), http://david.abcc.ncifcrf.gov/; PANTHER classification system, http://www.pantherdb.org/; Ingenuity Pathway Analysis (IPA, version 9.0, Ingenuity Systems), http://www.ingenuity.com/.

Online Methods

Cohort information

The Severe Childhood Onset Obesity Project (SCOOP)—The SCOOP cohort is a UK Caucasian (self-reported) subset of the Genetics of Obesity Study (GOOS), which includes individuals with severe, early-onset obesity (BMI standard deviation score (SDS) > 3; onset of obesity before the age of 10 years.

WTCCC2 UK controls—Publicly available Affymetrix Human SNP Array 6.0 was available for 6,000 individuals (3,000 from the 1958 British Birth Cohort and 3,000 from the UK Blood Service Collection) recruited for the Wellcome Trust Case Control Consortium 2 (see URLs).

EPIC Norfolk controls—Population-based controls from the EPIC-Norfolk (European Prospective Investigation into Cancer and Nutrition - Norfolk) study[24] were genotyped in the follow-up stage. Samples, not previously genotyped as part of the EPIC-Obesity Study[25] were selected to have BMI greater than or equal to 20 and less than or equal to 25.

SNP Analysis

Discovery genotyping and quality control

Genome-wide genotyping on 1386 patients from the SCOOP cohort was carried out using the Affymetrix Human SNP Array 6.0 chip at the Wellcome Trust Sanger Institute (WTSI). An additional 334 UK Caucasian GOOS samples described elsewhere[20] previously genotyped using the Affymetrix Human SNP Array 6.0 chip at Aros, Inc. were re-called at the WTSI; both sets of cases were used in the discovery stage.

For the common SNP association analysis the following criteria were applied for exclusion of SNPs prior to imputation: (i) minor allele frequency (MAF)<0.01, (ii) Hardy-Weinberg equilibrium (HWE)$p<10^{-4}$, (iii) call-rate<0.95 if MAF>0.05, call-rate<0.97 if 0.02<=MAF<=0.05 and call-rate<0.99 if 0.01<=MAF<0.02. Samples were excluded prior to imputation based on the following criteria: (i) call-rate<0.95, (ii) heterozygosity outside the population-specific bounds ((iii) relatedness – 500 SNPs with 0.45 ≤minor allele frequency<0.5 and call-rate>0.95 and 1000000 bp apart were used to estimate the concordance between pairs of samples. If the genotype concordance for two individuals is >0.7 and <0.97 (related) or ≥0.97 (duplicate) then the one with the lowest call rate was excluded. (iv) gender checks: gender was inferred from the intensity of the A-allele probes on the X chromosome normalised against the autosomal intensities. Individuals for which gender could not be inferred were excluded. (v) intensity failure: individuals were excluded if the mean of their A and B allele intensities from 10000 SNPs on chromosome 22 was outlying when compared to the sample at large. (vi) identity: prior to being genotyped, ~30 SNPs were typed using Sequenom® at the WTSI. If the concordance between the genome-wide & Sequenom® genotypes was <90% then the individual is excluded on the basis of unknown identity. (vii) ethnic outliers: cases with a distance to CEU > 30 based on a principal components analysis were excluded.

Imputation

After sample and SNP quality control, data was available for 1509 patients (333 with developmental delay) and 5380 controls from the 1958BC and UKBS collections used in WTCCC2 and up to 862,722 autosomal SNPs. These were separately imputed to the HapMap Phase 2 reference panel (release 22) in IMPUTE v1.0.0[26]. After imputation, SNPs were excluded based on the same MAF, HWE & call-rate thresholds used prior to imputation.

Principal components analysis

The genotypes of 10k SNPs at least 20kb apart along the genome were taken to calculate the pairwise distance matrix, which was used as the input for classical multidimensional scaling. Individuals were projected using the first two dimensions that correspond to the largest eigenvalues, and the ‘genetic distance’ to Europeans was calculated as the distance in the projected space between individual and the center (median) of the CEU cluster. An empirical distance threshold was adopted, above which individuals were regarded as ethnic outliers (**Supplementary Figure 8**).

Follow-up genotyping

In the validation stage, 980 SCOOP cases and 2,000 EPIC-Norfolk controls were genotyped at the WTSIby Sequenom® or Taqman® (Applied Biosystems). Sequenom genotyping was performed using the iPLEX™ Gold Assay (Sequenom® Inc.) according to manufacturer's instructions. Variants that failed or could not be designed on the Sequenom® platform were genotyped individually as TaqMan® assays (Applied Biosystems, California USA) on 8ng DNA per sample following the standard protocol. Replication samples were excluded on the basis of (i) call-rate<80% or mismatch between inferred and supplied gender. SNPs all had call-rate>90%, Hardy-Weinberg equilibrium $P > 10^{-4}$ and MAF ≥0.01 in cases and controls apart from rs17025867 and rs7255638 indicated in **Supplementary Table 1**.

Statistical analysis

Case control analyses under a log-additive model were performed using SNPTEST v1.1.5 adjusting for gender. A list of independent loci was generated using the clumping procedure implemented in PLINK, whereby loci were considered to be independent if the pair-wise linkage disequilibrium (LD, r^2) was less than 0.1 and if they were at least 500kb from the index SNP. Manual inspection of cluster plots was carried out to eliminate spurious associations due to poor genotyping prior to selecting variants for follow-up genotyping. Replication analysis was performed in Stata 11 (see URLs). Power calculations were performed using Quanto v 1.2.3 (see URLs) using the effect size estimate and control effect allele frequency from the validation stage.

Rare variant analysis

A rare variant analysis method implemented in CCRaVAT (Case-Control Rare Variant Analysis Tool, see URLs) was used to test for association with rare variants (MAF<0.05 in cases), collapsing within genes. Only directly genotyped SNPs for the discovery sample set were used, removing those that failed quality control. Cluster plots for SNPs contributing to any potential signals were manually re-called and checked.

Analysis of HBD burden

In order to remove non-random missingness (SNP failures correlating with individual datasets and experimental error rather than random failures) a chi-squared test was carried out on missingness per SNP between datasets. The SNPs with the top 5% of the resultant *p*-values were removed. In addition, in order to remove possible inflation of HBD from samples with non-European ancestry, only samples that clustered tightly with other European samples in the principle components analysis were included (**Supplementary Figure 8**). Consequently, 1472 case and 5380 controls were used in the Beagle analysis. Beagle was run using the default settings, and memory requirements dictated that the control cohorts were run as separate datasets. HBD burden was obtained by calculating the proportion of individuals that had one or more regions of HBD above a number of different size thresholds in each dataset. *P*-values for these thresholds were obtained by carrying out a Fishers Exact Test.

CNV analysis

Case and control CNVs were called by plate using Birdsuite-1.53 with CEU parameters. Autosomal calls made by Birdseye and Canary with LOD>=10, number of probes>=5, size>=1kb, probe density>=1 per 10kb were kept. Normalized probe intensities produced by apt-probe set-summarize were converted into $\log_2$ratios using the plate median as the reference. For each sample, the median-of-absolute-deviation (MAD) and the sum of autocorrelation (SAC) between pairs of probe 1-5 probes apart were calculated from $\log_2$ratios, which reflect the level of noise and long range waviness in the data, respectively. Samples yielding excessive CNV calls were filtered out by fitting the number of kept CNV calls per sample as a linear function of the sample's MAD in samples with a SAC in the lower 90% and removing outliers of the fitted linear model. Adjacent CNVs in the same sample were merged if (i) they were called with the same genotype, (ii) the distance between them was covered by <100 probes with a density > 0.2 probe per kb, (iii) the ratio of the number of probes between them to the number of probes of the merged call<0.1 and (iv) the difference in average $\log_2$ratio <0.15. Affy6 SNPs of all cases, controls, the HapMap1 populations (CEU, CHB+JPT, YRI) and an Indian population in HapMap3 (GIH) were called using Birdseed.

Common CNV analysis

Case and control CNVs were pooled and clustered into CNV events (CNVEs) as previously described [20]. 587 CNVEs with a carrier frequency >1% were defined as common CNVEs. Each CNVE was analyzed using the R package CNV tools which implements a likelihood ratio test that models the distribution of per sample CNV measurements as a Gaussian mixture and compares the goodness of fit with or without association to affected status. Test p values could be obtained for 481 of the 587 common CNVEs, whereas the rest failed to run through CNV tools due to poor genotype clustering. For further analysis of NEGR1 deletions, association tests were performed using Fisher's exact test based on discrete genotype calls, as the distribution of CNV measures of both cases and controls were in distinct and non-overlapping copy number clusters.

Rare CNV burden

Conditional permutation test[20] was performed to test the hypothesis that cases exhibit a greater burden of rare (carrier frequency<1%) and large (>100kb) CNVs relative to controls. Statistical significance was established by 10,000 rounds of permutation.

Biological Pathway Enrichment Analysis

Using the Database for Annotation, Visualization and Integrated Discovery (DAVID, see URLs), gene lists were mapped onto curated biological processes gene categories using Gene Ontology (GO, The Gene Ontology Consortium). Gene lists were also mapped to pathways and biological processes using the PANTHER classification system (see URLs) and Ingenuity Pathway Analysis (IPA, version 9.0, Ingenuity Systems, see URLs). The most significantly enriched categories of molecules and pathways (smallest p-values, p<0.05) were identified using these tools.

Electrophoretic Mobility Shift Assay (EMSA)

Perfect match (wild type, WT) or mutated (MUT) oligonucleotides to the predicted NKX6.1 binding site within the 8-kb deletion were designed (**Supplementary Table 8**) and used in an EMSA assay. Non-radioactive chemiluminescent detection was employed, and specifically the Light Shift Chemiluminescent EMSA Kit (Fisher-Thermo Scientific, UK). Briefly, labelled or unlabeled oligonucleotides were incubated with cell lysates of HEK293 cells expressing the human NKX6.1 protein (Novus Biologicals, EU), in the presence or absence of anα-NKX6.1 antibody (Abcam, UK). The DNA-protein complexes were run on 4% DNA retardation polyacrylamide gels (Invitrogen, UK), transferred to a nylon membrane using iBlot DNA transfer stacks and iBlot transfer system (Invitrogen, UK), and the resulting chemiluminesce detected by exposure to an X-ray film (Kodak, UK).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are indebted to the patients and their families for their participation and to the physicians involved in the Genetics of Obesity Study (GOOS). This work was supported by the Wellcome Trust (085475; WT098051), Medical Research Council and NIHR Cambridge Biomedical Research Centre. This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available from the Wellcome Trust Case-Control Consortium website (see URLs). The authors would like to thank Sarah Edkins, Emma Gray, Douglas Simpkin, Sarah Hunt, Sue Bumpstead and Serge Dronov of the WTSI Sample Logistics, Genotyping and Variation Informatics Facilities; Felicity Payne & Emma Gayton for Taqman genotyping and Eva Gonçalves Serra for imputation of the *GPR120* region.

References

1. Farooqi IS, O'Rahilly S. Genetic factors in human obesity. Obes Rev. 2007; 8(Suppl 1):37–40. [PubMed: 17316299]
2. Loos RJ. Genetic determinants of common obesity and their value in prediction. Best Pract Res Clin Endocrinol Metab. 2012; 26:211–26. [PubMed: 22498250]
3. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012; 109:1193–8. [PubMed: 22223662]
4. Cohen JC, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–72. [PubMed: 15297675]
5. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–78. [PubMed: 17554300]
6. Stranger BE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012; 8:e1002639. [PubMed: 22532805]
7. Bradfield JP, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet. 2012; 44:526–31. [PubMed: 22484627]
8. Dina C, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet. 2007; 39:724–6. [PubMed: 17496892]
9. Frayling TM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007; 316:889–94. [PubMed: 17434869]
10. Scherag A, et al. Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. PLoS Genet. 2010; 6:e1000916. [PubMed: 20421936]

11. Loos RJ, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet. 2008; 40:768–75. [PubMed: 18454148]
12. Jiao H, et al. Genome wide association study identifies KCNMA1 contributing to human obesity. BMC Med Genomics. 2011; 4:51. [PubMed: 21708048]
13. Benzinou M, et al. Common nonsynonymous variants in PCSK1 confer risk of obesity. Nat Genet. 2008; 40:943–5. [PubMed: 18604207]
14. Meyre D, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet. 2009; 41:157–9. [PubMed: 19151714]
15. Ichimura A, et al. Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. Nature. 2012; 483:350–4. [PubMed: 22343897]
16. Paternoster L, et al. Genome-wide population-based association study of extremely overweight young adults--the GOYA study. PLoS One. 2011; 6:e24303. [PubMed: 21935397]
17. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–48. [PubMed: 20935630]
18. Hafler BP, Choi MY, Shivdasani RA, Rowitch DH. Expression and function of Nkx6.3 in vertebrate hindbrain. Brain Res. 2008; 1222:42–50. [PubMed: 18586225]
19. Schafer M, Brauer AU, Savaskan NE, Rathjen FG, Brummendorf T. Neurotractin/kilon promotes neurite outgrowth and is expressed on reactive astrocytes after entorhinal cortex lesion. Mol Cell Neurosci. 2005; 29:580–90. [PubMed: 15946856]
20. Bochukova EG, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature. 463:666–70. [PubMed: 19966786]
21. Stice E, Spoor S, Bohon C, Small DM. Relation between obesity and blunted striatal response to food is moderated by TaqIA A1 allele. Science. 2008; 322:449–52. [PubMed: 18927395]
22. Raychaudhuri S, et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet. 2010; 6
23. Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–7. [PubMed: 20634204]
24. Day N, et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. Br J Cancer. 1999; 80(Suppl 1):95–103. [PubMed: 10466767]
25. Loos RJ, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet. 2008; 40:768–75. [PubMed: 18454148]
26. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39:906–13. [PubMed: 17572673]

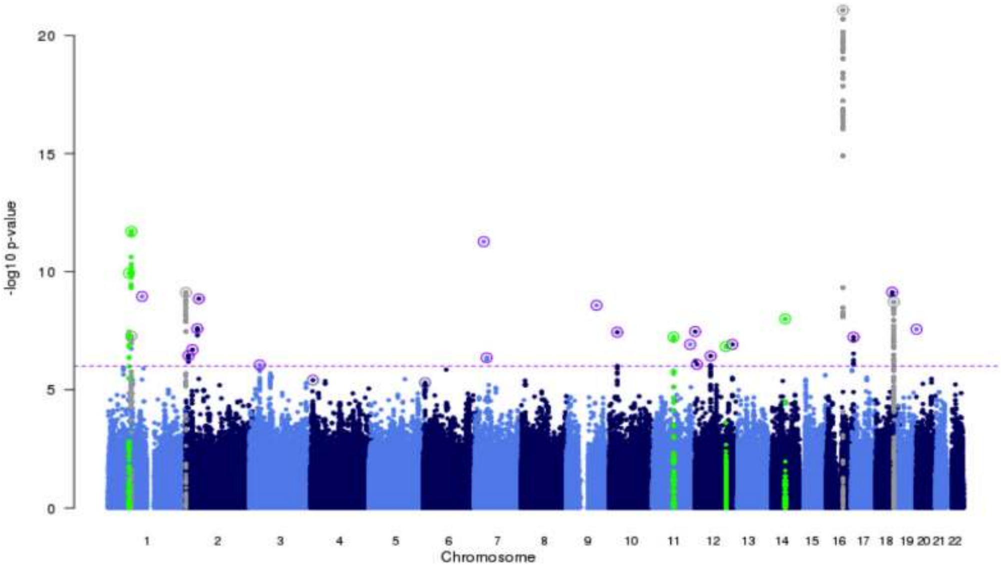

Figure 1.

Genome wide SNP association study in severe early onset obesity. Manhattan plot showing the significance ($-\log_{10}(p)$) of all SNPs in the discovery analysis. SNPs taken for replication (circled), established loci (dark grey) and novel loci (green) are shown. Horizontal dashed line indicates p=1×10^{-6}.

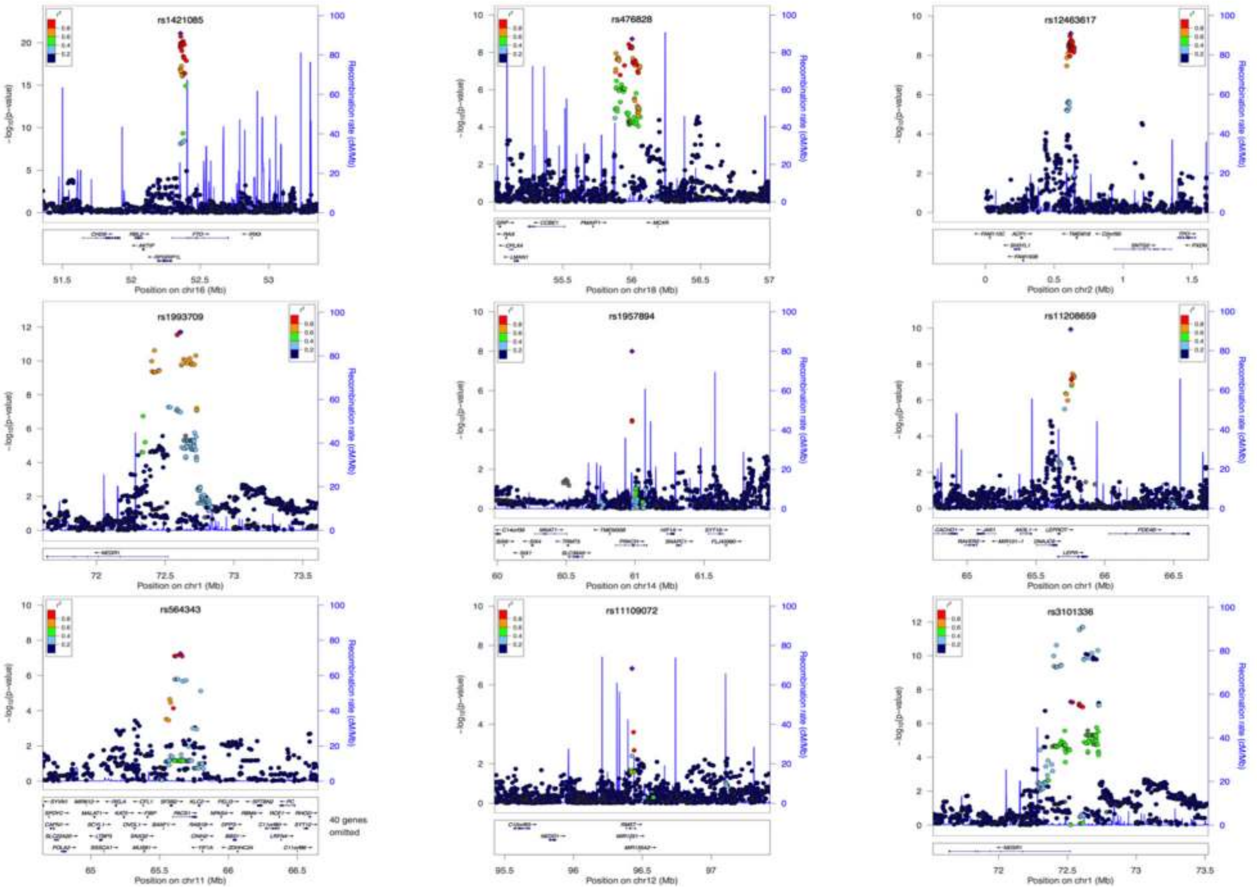

Figure 2.

Regional association plots of the novel loci generated using LocusZoom[23].

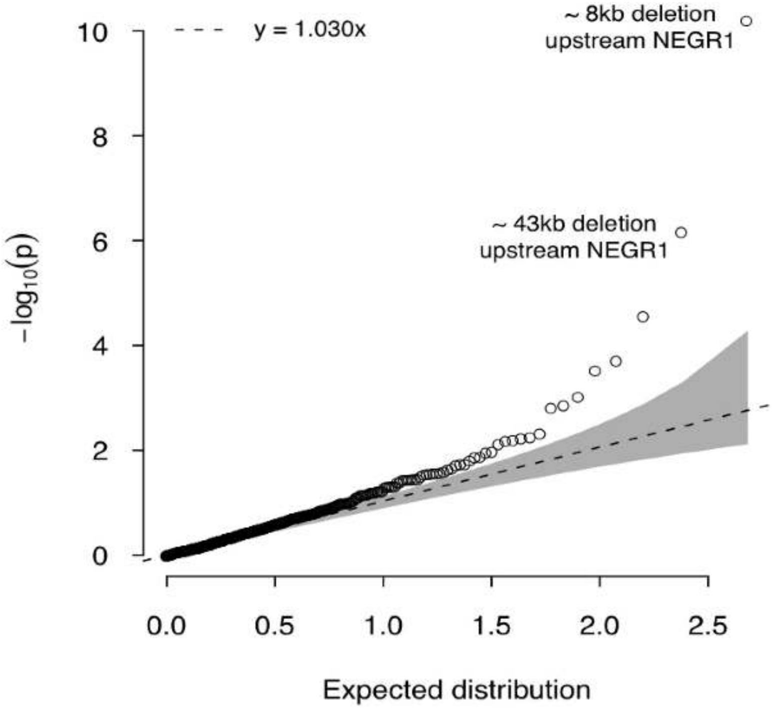

Figure 3.

A quantile-quantile plot of $-\log_{10}(p)$ of 481 common CNVs. Concentration band represents 95% confidence intervals. The data generally conform to the $-\log_{10}$ transformed uniform distribution expected under the null hypothesis of no association, with the exception of strong associations with the two deletions upstream of *NEGR1*.

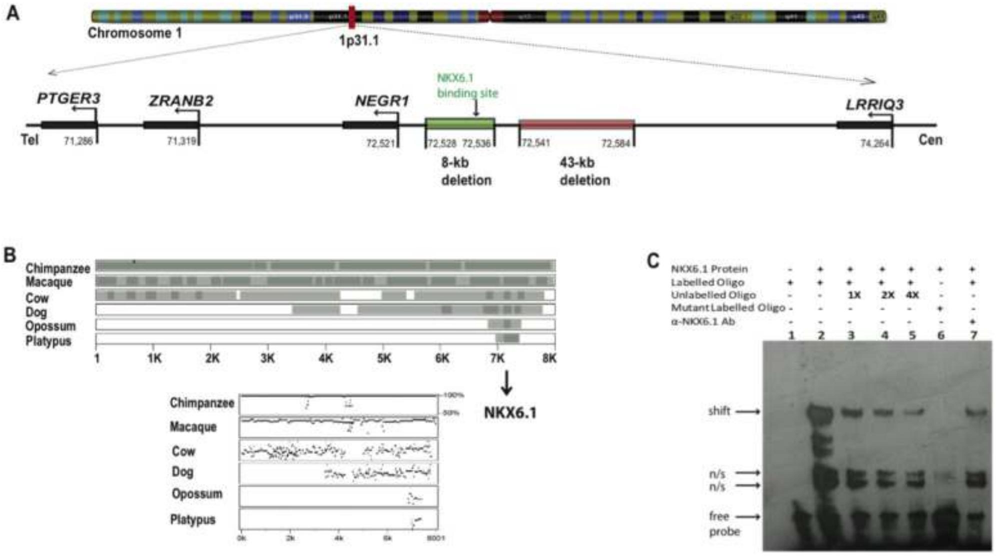

Figure 4.

NEGR1 locus on chromosome 1p31.1**. a.** Schematic of the locus with the positions of the two deletions (8-kb and 43-kb) and neighbouring genes. **b.** Sequence conservation of the small 8-kb deletion sequence. Nucleotide Percent Identity Plot (PIP) for the region is shown. Orthologous sequences from multiple species were aligned to the human sequence using MultiPIP maker, and sequence identity >50% (light grey) and >75% (dark grey) is highlighted, with the details shown below. The predicted NKX6.1 binding site is located in the most conserved region (arrow). **c.** Transcription factor NKX6.1 binding was confirmed using EMSA assay, electromobility shift (arrow) was specific to the sequence contained in the 8-kb deletion. n/s (non-specific band).

Table 1

SNPs associated with severe, early-onset obesity at genome-wide significance levels

							Discovery 1509 cases, 5380 controls				**Follow-up 971 cases, 1990 controls**				**Combined 2480 cases, 7370 controls**	
SNP	**Nearest gene**	**Novel**	**Chr.**	**Position (bp)[a]**	**Effect allele**	**Other allele**	**OR (95% CI)**	**P-value**	**EAF cases**	**EAF controls**	**OR (95% CI)**	**P-value**	**EAF cases**	**EAF controls**	**OR (95% CI)**	**P-value**
rs1421085	FTO		16	52358455	C	T	1.49 (1.37 - 1.61)	8.38×10^{-22}	0.5	0.41	1.36 (1.21 - 1.51)	4.66×10^{-08}	0.48	0.4	1.44 (1.35 - 1.54)	$\mathbf{3.07\times10^{-28}}$
rs476828[b]	MC4R		18	56003567	C	T	1.33 (1.21 - 1.46)	1.91×10^{-09}	0.29	0.24	1.32 (1.16 - 1.50)	1.64×10^{-05}	0.27	0.22	1.33 (1.23 - 1.43)	$\mathbf{9.38\times10^{-14}}$
rs12463617	TMEM18		2	619244	C	A	1.43 (1.27 - 1.60)	7.70×10^{-10}	0.89	0.85	1.4 (1.19 - 1.65)	4.80×10^{-05}	0.87	0.83	1.42 (1.29 - 1.56)	$\mathbf{1.70\times10^{-13}}$
rs1993709	NEGR1 *(8kb deletion)*	1	1	72611117	G	A	1.46 (1.32 - 1.63)	1.98×10^{-12}	0.87	0.81	1.22 (1.05 - 1.41)	0.011	0.85	0.82	1.38 (1.26 - 1.50)	$\mathbf{5.09\times10^{-13}}$
rs1957894	PRKCH	1	14	60977864	T	G	1.64 (1.39 - 1.95)	1.01×10^{-08}	0.08	0.06	1.34 (1.11 - 1.61)	0.002	0.1	0.08	1.5 (1.32 - 1.70)	$\mathbf{3.35\times10^{-10}}$
rs11208659	LEPR	1	1	65751868	C	T	1.63 (1.40 - 1.88)	1.16×10^{-10}	0.11	0.08	1.15 (0.96 - 1.38)	0.118	0.1	0.09	1.42 (1.27 - 1.59)	$\mathbf{1.84\times10^{-09}}$
rs564343	PACS1	1	11	65651742	A	G	1.25 (1.16 - 1.36)	5.81×10^{-08}	0.47	0.41	1.17 (1.05 - 1.30)	0.006	0.44	0.4	1.22 (1.15 - 1.31)	$\mathbf{2.03\times10^{-09}}$
rs11109072	RMST	1	12	96425401	A	C	1.79 (1.44 - 2.22)	1.48×10^{-07}	0.05	0.03	1.5 (1.13 - 1.99)	0.004	0.04	0.03	1.67 (1.41 - 1.99)	$\mathbf{4.21\times10^{-09}}$
rs3101336	NEGRI *(43kb deletion)*		1	72523773	C	T	1.26 (1.16 - 1.37)	5.28×10^{-08}	0.66	0.6	1.12 (1.00 - 1.25)	0.047	0.64	0.61	1.21 (1.13 - 1.29)	$\mathbf{2.21\times10^{-08}}$

Genome-wide loci for association with severe, early-onset obesity aligned to the risk-increasing (effect) allele. SNPs rs12463617, rs1993709, rs11208659, rs11109072 and rs3101336 were imputed. OR = Odds ratio; 95% CI = 95% confidence interval for the odds ratio; EAF = effect allele frequency.

[a] Positions according to Build 36.

[b] rs2168711 used as a proxy in the follow-up stage.

Table 2

Global CNV burden analysis of rare CNVs >100kb.

Type	Frequency	Case rate	Control rate	Case/control ratio	P_{MAD}	P_{NCPS}
Losses and gains	All < 1%	1.7439	1.5271	1.1419	$<1\times10^{-4}$	$<1\times10^{-4}$
	Single occurrence	0.4002	0.2849	1.4048	$<1\times10^{-4}$	$<1\times10^{-4}$
	Recurrent < 0.1%	0.4627	0.3968	1.1659	1.7×10^{-3}	2×10^{-4}
Losses	All < 1%	0.5972	0.5438	1.0983	1.23×10^{-2}	3.5×10^{-3}
	Single occurrence	0.1259	0.1001	1.2574	7.2×10^{-3}	7.0×10^{-3}
	Recurrent < 0.1%	0.1710	0.1415	1.2085	9.4×10^{-3}	4.3×10^{-3}
Gains	All < 1%	1.1467	0.9834	1.1661	$<1\times10^{-4}$	$<1\times10^{-4}$
	Single occurrence	0.2743	0.1848	1.4846	$<1\times10^{-4}$	$<1\times10^{-4}$
	Recurrent < 0.1%	0.2917	0.2553	1.1423	3.17×10^{-2}	8.0×10^{-3}

P_{MAD}, P_{NCPS}: P values derived from permutation conditioned on MAD of samples log2 ratio or number of calls per sample