Whole exome sequencing case-control using 1,000 severe obesity cases identifies putative new loci and replicates previously established loci

Audrey E. Hendricks on behalf of the UK10K Obesity Group

Introduction

- The genetic architecture of severe obesity in childhood is different to common obesity and BMI including some overlapping variants as well as unique variants (Wheeler et al., Nature Genetics 2013.)
- Here, case-control analysis using 737 high depth (~60x) whole exome sequenced (WES) severe obesity cases and 3,621 low depth (~6x) whole genome sequenced (WGS) controls identifies putative new and previously established locus
- Further, an uninflated set of test statistics as shown in QQplots and detection of an established loci support use of WGS controls for WES cases in single-variant case-control analysis

Methods

Samples

- All samples are part of the UK10K project (http://www.uk10k.org/)
- Cases
  - Severe Childhood Onset Obesity Project (SCOOP)
  - BMI Standard Deviation Score > 3 and obesity onset < 10 years
- Controls
  - UK10K cohort sample
  - Avon Longitudinal Study of Parents and Children (ALSPAC)
  - TwinsUK study from the Department of Twin Research and Genetic Epidemiology (DTR) at King's College London

Table 1. Samples

Sample	SCOOP	COHORT
N (after QC)	737	3621
Sequencing Type	WES	WGS
Mean Depth	~60x	~6x

Sequencing & Informatics

- WES and WGS were sequenced using a paired end HiSeq platform (Illumina)
- WES target enrichment using Agilent Technologies; Human All Exon 50 Mb array
- Realigned around known indels and recalibrated base quality scores
- For variant stability and accuracy, variants were called using SAMtools across 4723 UK10K exomes together and across 3781 UK10K genomes together
- For WES, variants were called within baits ± 100bp

Quality Control & Association Analysis

Variant QC
- Filtered variants using GATK VQSR with a 99.84% and 99.85% truth sensitivity tranche for WES and WGS respectively
- For WES, sites for individuals were set to missing when genotype quality < 20

↓

Sample QC
- Removed contaminated samples (freemix fraction > 0.03 & Fraction Extreme Alternate Sites > 0.0027) (Jun et al., AJHG 2012)
- Removed samples with low concordance with GWAS data
- Removed samples with non-European ancestry as identified using principal components projected from 1000 Genomes data
- Removed related samples (i.e. samples with cryptic estimated relatedness > 0.125)

↓

Genotype Refinement
- Genotype refinement (GR) performed within SCOOP and Cohort samples separately using using Beagle v3.3.2 and Beagle v4 respectively

↓

Single Variant Association Analysis
- Single marker association case-control analysis using dosage genotypes run using SNPTEST v2.4.0

↓

Remove regions less accessible to sequencing (1000 Genomes)
- Removed regions identified by 1000 Genomes as less accessible to sequencing using short reads (Strict Mask = P)
  - Depth of coverage is too low or too high
  - Lots of reads with zero mapping quality
  - Average mapping quality is low

↓

Filter by Genotype Refinement Quality
- Retained variants above various GR quality thresholds (i.e. 0.8, 0.9, 0.95, 0.99)

↓

Filter by Minor Allele Frequency
- Retained variants above various minor allele frequency (MAF) thresholds (i.e. none, 0.1%, 1%, 5%)

Results

- Strict mask = P: removed regions less accessible to sequencing
- Lambda: ratio of median expected $\chi 2$ statistic to observed $\chi 2$ statistic

QQplots (Figure 2)

- Severe inflation given no post association variant filtering
- Inflation in the tail and at the median is mitigated by filtering out regions less accessible to sequencing as identified by the 1000 Genomes project, GR quality, and MAF
- Given GR quality > 0.9, a low MAF threshold (MAF ≥ 0.001) may be able to be applied
- Given MAF ≥ 0.001, a more lenient GR quality threshold may be able to be applied (GR quality > 0.8)
- After strict mask = P, GR and MAF filtering, limiting to baits instead of baits ± 100bp does not appear to have a dramatic affect on the QQplot and removes nearly 50% of the variants

Manhattan Plots (Figure 3)

- *ADCY3*, an established obesity gene (Speliotes et al., Nature Genetics 2010; Wen et al., Nature Genetics 2012), is detected with a moderate p-value (p = 3.6 E-06)
- All SNVs with p-values < 1E-08 have either MAF < 0.005 or GR quality below 0.9 – genotype validation and further replication is needed
- SNPs and proxies (r2 > 0.8) from other previously established obesity and BMI loci (e.g. *FTO*, *MC4R*) were not captured by exome sequencing

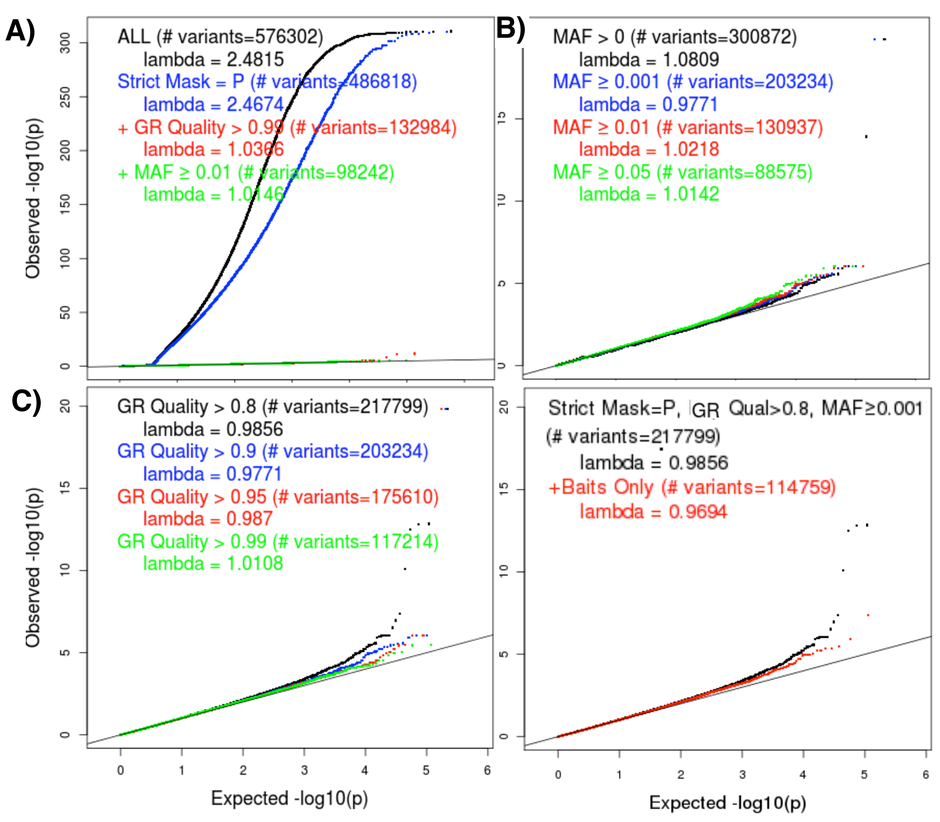

Figure 2. QQplots. Quantile-Quantile plots for various filtering thresholds and metrics. A) Filtering from all variants using several filtering metrics ; B) Filtering over various MAF thresholds at GR quality > 0.9 & strict mask = P; C) Filtering over various GR quality thresholds at MAF ≥ 0.001 and strict mask = P; D) Including a baits only filter (red) vs all called variants at baits ± 100bp (black) after filtering for GR quality > 0.8, MAF ≥ 0.001, and strict mask = P.

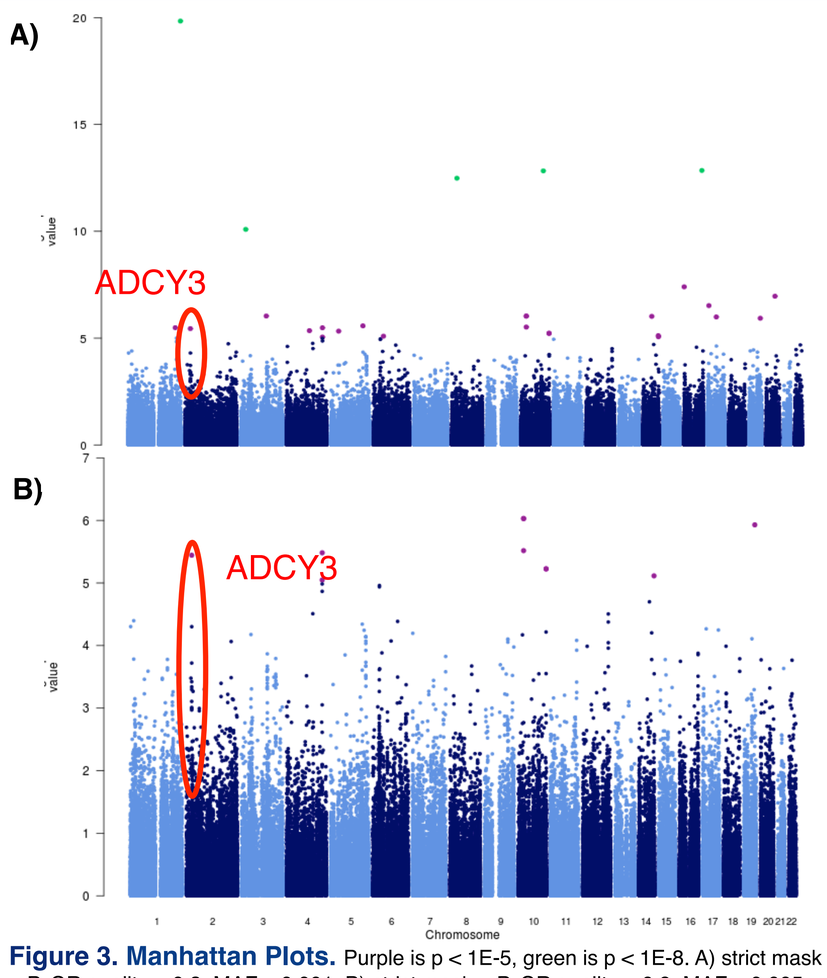

Figure 3. Manhattan Plots. Purple is p < 1E-5, green is p < 1E-8. A) strict mask = P, GR quality > 0.8, MAF ≥ 0.001; B) strict mask = P, GR quality > 0.9, MAF ≥ 0.005.

Discussion

Using WGS Controls for WES Cases

- After GR and filtering by GR quality and MAF, the distribution of inflated test statistics appears to be mitigated
- GR quality and MAF filters do not appear to have to be overly stringent
- Suggests that low depth WGS samples can be used as controls for WES cases in single variant tests after GR, filtering out regions less accessible to short read sequencing, and moderate filtering by GR quality and MAF

Association to Severe Childhood Obesity

- *ADCY3*, the only previously established obesity gene from GWAS with SNPs or proxies (r2 > 0.8) within the exome sequencing set, was detected with a moderate p-value
- Variants with p-values < 1 E-08 may represent novel obesity loci, but genotypes need to be validated and then the loci need to be replicated with independent samples

Next Steps

- Genotype SNVs with low p-values to validate that sequenced and imputed genotypes are accurate
- Replicate SNVs signals and peaks using an independent set of clinically extreme obese cases and controls
- Given off-target reads of 0.5x on average, impute exome set genome-wide using sequencing panels as reference set (e.g. 1000Genomes and UK10K cohorts) to capture obesity signals not directly sequenced through exome sequencing

Conclusions

- Genotype refinement and adequate variant filters appear to correct for large inflation at the median and tail due to sequencing differences between cases and controls
- Case-control analysis using exome sequenced cases can capture known obesity loci within the sequenced region
- More work (validation and replication) is needed to confirm the top variants as novel obesity variants/loci

Acknowledgements

This study is part of the UK10K project (http://www.uk10k.org/). Special thanks to the UK10K production team, the UK10K obesity group, and the UK10K postdocs.

For Further Information

Please contact Audrey Hendricks ah16@sanger.ac.uk