Whole Exome Sequencing Cases: Association Testing with External Controls

Audrey E. Hendricks on behalf of the UK10K Statistics Group

Introduction

- Sequencing only cases enables researchers to sequence a larger number of cases
- Focusing resources on sequencing cases can be especially valuable when the cases are unique and rare
- However, finding a suitable control set and developing an appropriate QC plan for case-control analysis is necessary when sequencing only cases

Objectives

- Compare use of various sample sets as controls, including those sequenced at a much different depth, for whole exome sequenced (WES) cases
- Explore various aspects of data QC and preparation when using external controls

Methods

Samples

- All samples are part of the UK10K project (http://www.uk10k.org/)
- Severe Childhood Onset Obesity Project (SCOOP)
  - BMI Standard Deviation Score > 3 and obesity onset < 10 years
- Neuro Aberdeen Schizophrenia (NEURO)
  - Schizophrenia samples gathered from Aberdeen Scotland
- Cohort (COHORT)
  - 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	NEURO	COHORT
N	667	347	2432
Sequencing Type	WES	WES	WGS
Mean Depth	~60x	~60x	~6x

Sequencing & Informatics

- WES and Whole Genome Sequencing (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 4060 UK10K exomes together and across 2432 UK10K genomes together
- Variants were filtered using GATK VQSR
- For WES
  - Variants were called within baits ± 100bp
  - Sites for individuals were set to missing when genotype quality < 20
  - Imputation performed within each sample set using Beagle v3.3
- All initial analyses on PASSed variants on chr. 20
- For WGS, genotype likelihoods were set with Beagle and imputed with IMPUTE2

Association Analysis

- Single marker association case-control analysis using dosage genotypes was run using SNPTEST v2.4.0
- Within sample case-control analysis
  - Ran to find filters that retain the appropriate null distribution of test statistics
  - SCOOP: 334 cases vs 333 controls
  - NEURO: 174 cases 173 controls
- Between sample case-control analysis
  - Ran to determine if filters found through within sample analysis retained the appropriate null distribution of test statistics between samples as well
  - WES vs WES: 667 SCOOP cases vs 347 NEURO controls
  - WES vs WGS
    - 667 SCOOP cases vs 2432 COHORT controls
    - 347 NEURO cases vs 2432 COHORT controls

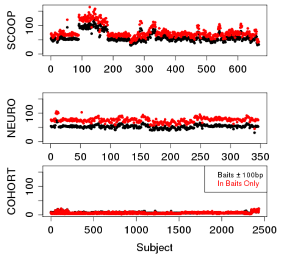

Figure 1. Mean Sample Depth.

Table 2. Variant Filters

Filter	Levels
Minor Allele Frequency (MAF)	0.01, 0.05, 0.1
Genotype Call Rate (GCR)	0.5, 0.8, 0.9, 0.95, 0.99
Imputation r^2 (r^2)	0.8, 0.9, 0.95
SNPTEST Info Score (Info)	0.9
Baits	Baits ± 100bp, Baits Only

Results

- n: number of markers retained for analysis
- Lambda: ratio of median expected χ2 statistic to observed χ2 statistic

Within SCOOP Analysis (Figure 2)

- Moderate to high inflation seen at median and in tail of test statistic distribution
- Both median and tail inflation controlled by variant filters (i.e. MAF, imputation quality r^2, etc.) within genotype data (Fig. 2A) and within imputed data (Fig. 2B)
- Stricter MAF filter needed for genotype analysis (MAF > 5%) compared to imputation analysis (MAF > 1%)
- GCR filter in genotype data is necessary but not sufficient to control for inflation
- Imputation quality r^2 filter is sufficient to control for inflation
- When using variant filters (specifically GCR or r^2), filtering to the variants only called within bait regions is not necessary

Within NEURO Analysis (Figure 3)

- Less inflation seen in NEURO analyses
- Inflation removed just by MAF filter

Between Sample Analysis (Figure 4)

- Extreme inflation seen at median and in tail
- Inflation in tail removed by using MAF, r^2, and an additional Info filter
- High inflation around the median remains although filtering to only the bait regions lowers the inflation slightly

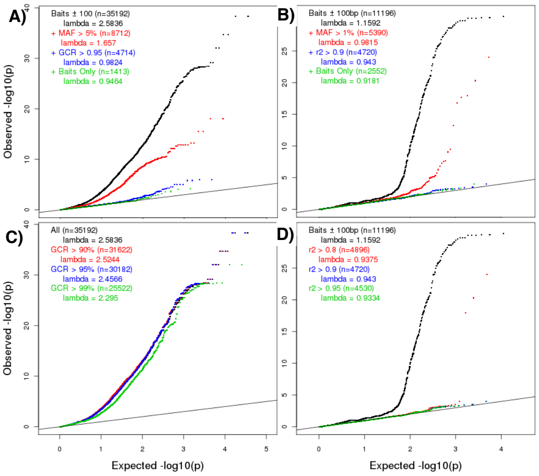

Figure 2. SCOOP WES. Within SCOOP sample analysis. A) Filtering within genotype data; B) Filtering within imputed data; C) Genotype Call Rate (GCR) Filtering within genotype data; D) Imputation r^2 filtering within imputed data.

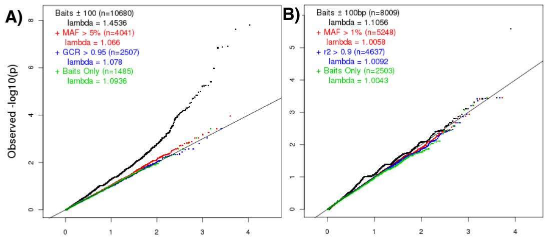

Figure 3. NEURO WES. Within NEURO sample analysis. First 174 NEURO exomes chosen as cases vs. second 173 as controls. A) Filtering within genotype data; B) Filtering within imputed data.

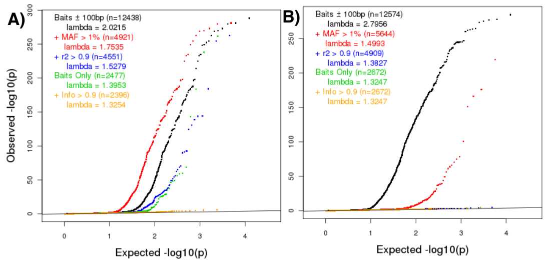

Figure 4. Between Sample Analysis. A) 667 SCOOP cases vs 2432 COHORT controls; B) 667 SCOOP cases vs 347 NEURO controls.

Discussion

Differences in WES Depth

- Initial inflation much more severe for SCOOP exomes compared to NEURO exomes
- Difference in mean sequencing depth within SCOOP exomes may be related to severe inflation in both the median and tail of the test statistic distribution
- Inflation appeared to be removed by using variant level filters such as MAF, GCR, or imputation r^2

Baits ± 100bp vs Baits only

- Using variant filters such as GCR and imputation r^2 appeared to remove inflation seen when including variants called outside of the bait regions
- GCR and imputation r^2 likely remove variants that would only be called in samples sequenced at a higher depth
- 2-3 times as many variants are retained when including filtered variants called within 100bp of the baits

WES vs WGS Control Sets

- Extreme inflation in the tail is removed by variant filters
- Adjusting for possible population stratification or other subject level filters may help to alleviate the large inflation that remains at the median after variant filters

Future Work

- Apply strict individual level filters for between sample analysis & include covariates to adjust for population stratification

Conclusions

- Adequate variant filters correct for large inflation at the tail due to sequencing differences between cases and controls
- Variant filters alone do not adjust for inflation at the median and more research is needed to address this residual inflation
- MAF filters are necessary for single marker tests but exclude rare variants that high depth WES studies are designed to detect; thus additional research is needed for using external controls with methods that aggregate rare variants

Acknowledgements

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

For Further Information

Please contact Audrey Hendricks ah16@sanger.ac.uk