Quality control is essential to ensuring reproducibility in genotype and non-genetic data

Molly Hall, PhD
Assistant Professor
Veterinary and Biomedical Sciences
June 18, 2018
Genetic Association

Cases (n=1,000) (express the trait) vs. Controls (n=1,000) (do not express the trait)

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>62%</td>
<td>38%</td>
</tr>
<tr>
<td>Controls</td>
<td>49%</td>
<td>51%</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 34.2, \text{ p-value} = 4.9 \times 10^{-9} \]

http://www.discoveryandinnovation.com/
GWAS of Body Mass Index (BMI)

Biomedical Informatics

Predict Disease
Genome → Predict Disease
Why do we perform quality control (QC) methods?

• We want to make sure signals we find in GWAS (or other analyses) are not due to:
  • Poor quality samples (sample call rate)
  • Poor quality markers (marker call rate)
  • Relatedness in your sample set (Identity by decent (IBD))
  • Differences due to population structure (principal component analysis (PCA))

• We want to reduce the number of tests we need to adjust for:
  • Minor allele frequency (MAF)
  • Linkage disequilibrium (LD)
PLINK Files

• *.fam files contain information about samples, one sample per line
• *.bim files contain information about markers, one marker per line
• *.bed files contain binary genotype information. You should not be viewing this file directly.

• Here is an example command to read these files in PLINK: plink --bfile myfile
Input Genotype Data

Samples passing sex check

Discard samples with problems from sex check
QC:

Pre-QC:

Samples: 3,896  
Markers: 561,490

SEX CHECK:

plink --bfile T2D --check-sex

awk '{if ($5=="PROBLEM")print}' plink.sexcheck

16230834 111584@1018348317 2 0  PROBLEM 0.4688
16228083 119785@1018342676 2 0  PROBLEM 0.2364
16222319 137237@1018343183 2 0  PROBLEM 0.7259
16231930 108172@1018342658 2 0  PROBLEM 0.4507
16228204 119481@1018298703 2 1  PROBLEM 1
16233113 104569@1018301292 1 0  PROBLEM 0.4823
16214881 159853@1018299011 1 2  PROBLEM 0.1067

Dropped 2 samples: plink --bfile T2D --remove drop_sex_check --make-bed
MARKER:

plink --bfile T2D_sex-check --geno 0.01 --make-bed --out T2D_sex-check_genotypefiltered

Dropped 8416 markers.

Remaining:
Markers: 553,074
Samples: 3,894

SAMPLE:

plink --bfile T2D_sex-check_genotypefiltered --mind 0.01 --make-bed --out T2D_sex-genotypefiltered

Dropped 7 samples.

Remaining:
Markers: 553,074
Samples: 3,887
MAF:

plink --bfile T2D_sex-check_geno99_mind99 --maf 0.05 --make-bed --out T2

52,401 markers dropped.

Remaining:
Markers: 500,673
Samples: 3,887
<table>
<thead>
<tr>
<th>Z0</th>
<th>Z1</th>
<th>Z2</th>
<th>Kinship</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>MZ twin or duplicate</td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.50</td>
<td>Parent-offspring</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.50</td>
<td>Full siblings</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>0.0</td>
<td>0.25</td>
<td>Half siblings</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
<td>0.0</td>
<td>0.125</td>
<td>Cousins</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>Unrelated</td>
</tr>
</tbody>
</table>
runQC -p 4 -f T2D_sex-check_geno99_mind99_maf05 -l 0.3 -C -P8

runQC -p 5 -f T2D_sex-check_geno99_mind99_maf05_prune03 -C -P8
Dropped 513 samples.
Input Genotype Data

Samples passing sex check
- Marker Call Rate
- Sample Call Rate
- Minor Allele Frequency

Initial Quality Control

Discard samples with problems from sex check

Discard markers and samples not passing initial quality control

IBD Estimate (on approx. 100,000 markers)
- Remove samples with cryptic relatedness
- Re-do Initial Quality Control on unrelated samples
Any time samples are dropped - REPEAT
Genome → Predict Disease
Future Directions

Genome

Exposome

Predict Disease
Issues to consider for exposure (and phenotype) data

- Outliers
- Survey data (bias, missingness, incorrect)
- Faulty lab measurement
- Small sample size
- Desperate data types (continuous, categorical, binary) combined
- Skew
- Can’t look by hand at big data!
CLARITE

Start by Choosing a Data File

Output Folder: GUI_Output

Descriptive  Quality Control  Association

General
- Get Unique Values
- Sample Size

Categorical/Binary
- Frequency Table
- Chi-squared Test
- Bar Plot

Continuous
- Correlations
- Outliers
- Histogram
- Box Plot
- QQPlot

Live Logs

LIVE LOG REPORT - PROGRAM OPENED: 03:02PM on June 13, 2018
LOGS INPUT VALUES, STDOUT, STDERR, R SCRIPT & FUNC CALLS (GUI PROGRAM)
Click to take a virtual tour of NHANES
Example of QC Protocol (NHANES HDL-C)

• Drop any sample that’s missing a covariate value or phenotype
• Split by variable type:
  • Split into 4 tables by variable type: binary, continuous (min values = 15), categorical (3-6), and ambiguous (6-15)
  • By hand, determine ambiguous variable type and merge into appropriate file
  • Drop any variables that are indeterminant according to the NHANES data dictionary
• Sample Size Filter:
  • Drop variables < 200 samples (“Min # Samples”) and < 200 samples in a category (“Min Category Size”)  
• Remove any variable with > 90% of the samples with a 0 value.
• Log(x+1) transformation all exposures and phenotypes
Environment-Wide Association Studies (EWAS)

- Test a variety of environmental variables in a high-throughput manner for association with phenotype(s)

- Analogous to GWAS method of testing loci across the genome

- Enable agnostic exposure assessment

Diet
Pollution
Alcohol
Smoking
Exercise
Pesticides
Residence
UV

Type 2 diabetes

EWAS Discovery and Replication

• Following QC...
• Linear regression (HDL-C)
• Covariates: Sex, Age, BMI, SES, Race, Series

• Repeat QC and EWAS in Replication dataset (2 later surveys in NHANES)
MANHATTAN PLOT OF REPLICATING EWAS RESULTS TO BE ADDED
Summary

• The first step in data reproducibility is ensuring high quality data.

• To do this, rigorous and well-documented (so it can be reproduced!) QC is essential.

• Standardized QC protocols and tools are well-established and utilized in genomics.

• Few standardized protocols and tools are established for environment and phenotype data but are needed.
Key QC Papers:

• Stephen Turners GWAS QC paper:
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066182/

• Other useful papers:
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025522
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061487/
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592376/
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896766/

• PLINK paper:
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/
Tools:

• PLINK: Most commonly used in Hall Lab for genotype quality control (http://pngu.mgh.harvard.edu/~purcell/plink/)

• PLATO: Most commonly used in Hall Lab for complex association studies (http://ritchielab.psu.edu/software/plato-download)

• Eigensoft/smartpca: For principal component analysis (PCA) (http://www.hsph.harvard.edu/alkes-price/software/)

Acknowledgements:

Molly Hall, PhD
Anastasia Lucas *
Nicole Palmiero
Bryan Almonte
Tuyen Pham