GWAS QC - theory and steps

H3ABioNet Data Management Workshop
Shaun Aron
Presentation credit: Dr. Ananyo Choudhury
June 2014
Quality control for GWAS studies

• Genotyping experiments often generate a mixed bag of results

• Errors might originate at many different steps
  – Sample selection related issues
  – Sample handling related issues
  – Genotyping chip related issues
  – Batch effect related issues

• Steps
  – QC by SNP
  – QC by sample
German MI family study  Affymetrix 500K Array Set
SNPs on chips: 493,840

SNPs passing QC: 270,701

Roadmap

Discordant sex information
High Missingness
Excess or deficiency of heterozygosity
Duplicate or related
Divergent ancestry

Low minor allele frequency
Missingness
Differential missingness
Hardy-Weinberg outliers
Software

• Programs required for QC
  – PLINK (Purcell, 2007)
  – Scripts for processing results files
  – R (Statistical Software) for plotting results

• Programs for population structure analysis
  – SmartPCA, PLINK
  – Admixture
Sample QC - identifying and removing individuals with:

- Discordant sex information
- High Missingness and outlying heterozygosity rate
- Duplicate and related
- Divergent ancestry
Why quality control individuals?

- Sample handling related issues:
  - Poor DNA quality/concentration
  - Contamination
  - Error in labeling/plating
- Sample selection related issues:
  - Cryptic relatedness
  - Population structure
- Measures to remove individuals not genotyped properly
Sample QC steps

Discordant sex information

High Missingness and outlying heterozygosity rate

Duplicate and related

Divergent ancestry
Using genotypic data to estimate sample sex

- Males have a single X chromosome and therefore can be estimated to be homozygous for all the X chromosome SNPs (other than those in the pseudo autosomal region (PAR)).
- Therefore, X chromosome homozygosity estimate for males (XHE) is 1

- Plink assigns sex based on XHE estimate (F or inbreeding coefficient):
  - Male (1): XHE > 0.80
  - Female (2): XHE < 0.20
  - No sex (0): 0.20 < XHE < 0.80

- Comparisons of predicted and observed sex can be used to identify miscoded sex or sample mix-ups, etc.

- Samples with discordant sex information are removed
Identify individuals with discordant sex information

```
plink --bfile example --check-sex --out sexstat --noweb
```

Creates a file named `sexstat.sexcheck`

<table>
<thead>
<tr>
<th>FID</th>
<th>IID</th>
<th>PEDSEX</th>
<th>SNPSEX</th>
<th>STATUS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>P554</td>
<td>P554</td>
<td>2</td>
<td>2</td>
<td>OK</td>
<td>-0.02654</td>
</tr>
<tr>
<td>P555</td>
<td>P555</td>
<td>1</td>
<td>0</td>
<td>PROBLEM</td>
<td>0.5685</td>
</tr>
<tr>
<td>P557</td>
<td>P557</td>
<td>2</td>
<td>2</td>
<td>OK</td>
<td>0.1264</td>
</tr>
<tr>
<td>P558</td>
<td>P558</td>
<td>2</td>
<td>2</td>
<td>OK</td>
<td>-0.0007684</td>
</tr>
</tbody>
</table>

Select individuals with Status=“PROBLEM” in the file `sexstat.sexcheck`
Try to identify the problem. If the problem cannot be resolved write the IDs of the individuals with discordant sex information to a file “fail_sex_check-example.txt”

```
grep "PROBLEM" sexstat.sexcheck > fail_sex_check-example.txt
```
Sample QC steps

High Missingness and outlying heterozygosity rate

Duplicate and related

Divergent ancestry
**Genotyping call rate**
- Per sample (individual) rate
- Number of non-missing genotypes divided by the total number of genotyped markers.
- Low genotyping call rate indicate problem with sample DNA like low concentration.
- Thresholds used generally vary between 3% and 7%

**Heterozygosity Rate**
- Per sample (individual) rate
- Number of (total non-missing genotypes(N) – homozygous(0)) genotypes divided by total non-missing genotypes(N)
- Excess heterozygosity - Possible sample contamination
- Less than expected heterozygosity- Possibly inbreeding
- Threshold for inclusion is generally Mean ± 3 std.dev. over all samples

Genotyping call rate and heterozygosity rate are generally plotted together. Cutoffs are selected so as to identify outlier individuals based on both the statistics.
Identification of individuals with elevated missing data rates

```
plink --bfile example --missing --out example_miss
```

Before frequency and genotyping pruning, there are 98604 SNPs
646 founders and 0 non-founders found
34704 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ example_miss.hh ]
3452 SNPs with no founder genotypes observed
Warning, MAF set to 0 for these SNPs (see --nonfounders)
Writing list of these SNPs to [ example_miss.nof ]
Writing individual missingness information to [ example_miss.imiss ]
Writing locus missingness information to [ example_miss.lmiss ]

<table>
<thead>
<tr>
<th>FID</th>
<th>IID</th>
<th>MISS_PHENO</th>
<th>N_MISS</th>
<th>N_GENO</th>
<th>F_MISS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P554</td>
<td>P554</td>
<td>N</td>
<td>4096</td>
<td>97722</td>
<td>0.04191</td>
</tr>
<tr>
<td>P557</td>
<td>P557</td>
<td>N</td>
<td>4011</td>
<td>97722</td>
<td>0.04105</td>
</tr>
<tr>
<td>P558</td>
<td>P558</td>
<td>N</td>
<td>4327</td>
<td>97722</td>
<td>0.04428</td>
</tr>
<tr>
<td>P562</td>
<td>P562</td>
<td>N</td>
<td>4099</td>
<td>97722</td>
<td>0.04195</td>
</tr>
</tbody>
</table>

```
HR                      SNP   N_MISS | N_GENO | F_MISS
1                       vh_1_1108138   9   646  0.01393
1                       vh_1_1110294   4   646  0.006192
1                       rs7515488    1   646  0.001548
1                       rs6603785    9   646  0.01393
```
Identification of individuals with extremely high or low heterozygosity rate

plink --bfile example --het --out example_het

Before frequency and genotyping pruning, there are 98604 SNPs
646 founders and 0 non-founders found

Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 98604 SNPs
646 founders and 0 non-founders found
34704 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ example_het.hh ]
3452 SNPs with no founder genotypes observed
Warning, MAF set to 0 for these SNPs (see --nonfounders)
Writing individual heterozygosity information to [ example_het.het ]

<table>
<thead>
<tr>
<th>FID</th>
<th>IID</th>
<th>O(HOM)</th>
<th>E(HOM)</th>
<th>N(NM)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>P554</td>
<td>P554</td>
<td>67663</td>
<td>6.725e+04</td>
<td>86305</td>
<td>0.02173</td>
</tr>
<tr>
<td>P557</td>
<td>P557</td>
<td>66873</td>
<td>6.731e+04</td>
<td>86388</td>
<td>-0.02301</td>
</tr>
<tr>
<td>P558</td>
<td>P558</td>
<td>67155</td>
<td>6.707e+04</td>
<td>86091</td>
<td>0.004538</td>
</tr>
<tr>
<td>P562</td>
<td>P562</td>
<td>68367</td>
<td>6.724e+04</td>
<td>86306</td>
<td>0.05891</td>
</tr>
</tbody>
</table>
Based on the plot we need to decide reasonable thresholds at which to exclude individuals based on elevated missing or extreme heterozygosity.

We decided to exclude all individuals with a genotype failure rate $\geq 0.06$ and/or heterozygosity rate $\pm 3$ standard deviations from the mean.
Sample QC steps

Duplicate and related

Divergent ancestry
Identify related and duplicate individuals

- A basic assumption of standard population-based case-control association studies is that all the samples are **unrelated** (i.e. the maximum relatedness between any pair of individuals is less than a second degree relative)

- Presence of duplicate and related individuals in the dataset may **introduce bias** and cause **genotypes in families** to be over-represented.

- To identify duplicate and related individuals, a metric (identity by state, IBS) is calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosomes)
- The IBS method works best when only **independent SNPs** are included in the analysis.
- Independent SNP set for IBS calculation is generally prepared by **removing regions of extended LD** and **pruning the remaining regions** so that no pair of SNPs within a given window (say, 50kb) is correlated.
- Following the calculation of IBS between all pairs of individuals, duplicates are denoted as those with an IBS of 1.
- Related individuals will share more alleles IBS than expected by chance, with the **degree of additional sharing proportional to the degree of relatedness**.

**Identity by State (IBS)**

$$IBS = \frac{IBS2 + 0.5 \times IBS1}{N \text{ SNP pairs}}$$

- $IBS2 = \text{number of loci in which the two individuals have two alleles in common}$
- $IBS1 = \text{number of loci in which the two individuals have one allele in common}$
- $N \text{ SNP pairs} = \text{number of common, nonmissing SNPs}$
• The degree of **recent shared ancestry for a pair of individuals** (identity by descent, IBD) can be estimated using genome-wide IBS data using Plink. (IBD shown as pi_hat in plink)

• The expectation is that:
  • IBD = 1 for duplicates or monozygotic twins
  • IBD = 0.5 for first-degree relatives,
  • IBD = 0.25 for second-degree relatives
  • IBD = 0.125 for third-degree relatives

• Genotyping error, LD and population structure cause variation around these theoretical values and it is typical to remove one individual from each pair with an IBD > 0.1875 (halfway between the expected IBD for third- and second-degree relatives).

• For same reasons an IBD > 0.98 identifies duplicates.
Identification of duplicated or related individuals

PREPROCESSING

plink --bfile example --exclude high-LD-regions.txt --range --indep-pairwise 50 5 0.2 --out example

CALCULATING IBD

plink --bfile example --extract example.prune.in --genome --out example

FILTERING RELATED INDIVIDUALS

perl run-IBD-QC.pl example.genome

create a file example.genome containing pairwise IBS for all pairs of individuals

As this step is highly computationally intensive, it is a good option to remove regions of high LD (pre-calculated and stored in the file, high-LD-regions.txt) before the IBS run.

creates the file example.prune.in, containing the list of SNPs to be kept in the analysis.

identify all pairs of individuals with an IBD > 0.185. Looks at the individual call rates stored in example_miss.imiss and output the ids of the individual with the lowest call-rate to 'fail_IBD_example.txt' for subsequent removal.
Population structure

- Population substructure or stratification occurs when samples have different genetic ancestries.
- Can lead to spurious associations due to differences in ancestry rather than true associations.
- Imperative to check for population structure within samples.
- Can control for structure if identified, in downstream analysis.
Approaches to identify population structure

- Methods to measure the ancestry of each sample in the data

- Structure based approach:
  - Admixture, CLUMPP

- Principle component based approach
  - SmartPCA, SNPRelat, PLINK

- Comparison of cases and controls in sample

- Can also compare with other known populations
Population structure

• Outcome of approaches
  – Identify if there is population structure in the dataset – apply appropriate measures to control for this in association test/selection of association test
  – Identification of samples that are significant outliers in the dataset based on population ancestry – exclude those individuals
Population structure - PCA
Discordant sex information

High Missingness and outlying heterozygosity rate

Duplicate and related

Divergent ancestry

fail_ancestry_example.txt

fail_IBD_example.txt

fail_miss_het_example.txt

JOIN FILES
cat fail_* | sort -k1 | uniq > fail_example_inds.txt

REMOVE FROM DATA
plink --bfile example --remove fail_example_inds.txt --make-bed --out clean_inds_example --noweb

Sample QC completed
SNP QC

- Low MAF
- High Missingness
- Differential missingness
- HWE outliers
Genotype resolution is often challenging

Bad calls can lead to false associations!
What causes low quality SNP genotyping?

- Genotype clusters of many SNPs demonstrate low quality genotyping due to:
  - Low DNA concentration
  - Poor binding and competitive binding by other sequences
  - Structural and copy number variants
- Not possible to QC each SNP manually
- Measures to remove low quality SNPs are required
Low MAF

High Missingness

Differential missingness

HWE outliers

SNP/Marker QC steps
Genotype calling algorithms perform poorly for SNPs with low MAF.
Clustering depends largely on sample size for low MAF SNPs.
Power for detecting associations to SNPs with low MAF is low (unless the sample size is very large).
Low MAF SNPs are therefore excluded.
An often used exclusion threshold is MAF 1% to 2%.
Identify low minor allele frequency SNPs

GET ALLELE FREQUENCIES

plink --bfile clean-inds-example --freq --out clean inds example freq

Generates the file “clean inds example freq.frq” containing minor allele frequency of each SNP

GENERATE MAF PLOT USING R SCRIPT

maf_plot.R

CHOOSE STANDARD MAF CUTOFF (MAF>0.01) OR SELECT ONE ON THE BASIS OF THE PLOT
Low MAF

High Missingness

Differential missingness

HWE outliers

SNP/Marker QC steps
Missingness frequency

- SNPs which cannot be assigned definitively to a cluster are assigned “missing” status during genotype calling.

- **Missing frequency** (also termed 1 minus SNP call rate) is the fraction of total genotype calls for a SNP which has been assigned missing status.

- High missingness often implies that the **cluster separation** for a particular SNP has been poor and the SNP needs to be removed.

- A missingness cutoff of 1%-5% is generally used.
Identify SNPS with high missingness

**GET ALLELE FREQUENCIES**

plink --bfile clean_ind_example --missing --out clean_ind_example_missing --noweb

Generates the file “clean_ind_example_missing.miss” containing missingness value for each SNP

**GENERATE PLOT USING R SCRIPT**

snpmiss_plot.R

CHOOSE THE STANDARD MISSINGNESS (F_MISS) CUTOFF >0.05

OR

CHOOSE CUTOFF ON THE BASIS OF THE PLOT
SNP/Marker QC steps

Low MAF

High Missingness

Differential missingness

HWE outliers
Missing frequency is also assessed separately in cases and in controls because differential missingness is a common source of false positive associations.

SNPs showing highly differential missingness (P<0.00001) are excluded.
Identify SNPs with high differential missingness in case and controls

**GET ALLELE FREQUENCIES**

```bash
plink --bfile clean_ind_example --test-missing --out clean_ind_example_test_missing --noweb
```

Generates the file “example_test_missing.missing” containing differential missingness statistics for each SNP

**GENERATE PLOT USING R SCRIPT**

diffmiss_plot.R

**CHOOSE STANDARD DIFFERENTIAL MISSINGNESS P-VALUE CUTOFF (0.00001)**

**OR CHOOSE ON THE BASIS OF THE PLOT**

To identify SNPs showing differential missingness P-value greater than cutoff:

```bash
perl run_difmiss.pl clean-Inds_example
```

Creates the file “fail_difmiss_example.txt”
SNP/Marker QC steps

High Missingness

Low MAF

Differential missingness

HWE outliers
Hardy Weinberg Equilibrium

Assumptions

– Diploid organisms
– Infinite population size
– Non-overlapping generations
– Random mating
– No selection, mutation or migration

Testing for HWE

• Calculate the allele frequency \( (p) \)
  • Using observed genotype counts
• Calculate the expected genotype counts
  • Using the allele frequency \( (p) \)
• Compare the observed to the expected counts
  • \( \chi^2 \) test
HWE Example

Step I

- Observed Genotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>GG</th>
<th>Gg</th>
<th>gg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

1. Calculate the allele frequency \( (p) \):

\[
p = \frac{2(12) + 2}{2(22)} = 0.59
\]
Step II

Observed Genotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>GG</th>
<th>Gg</th>
<th>gg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

2. Calculate the expected genotype counts:

\[
E(GG) = np^2 = 22(0.59^2) = 7.66
\]

\[
E(Gg) = n2pq = 22(0.59)(1 - 0.59) = 10.64
\]

\[
E(gg) = nq^2 = 22((1 - 0.59)^2 = 3.68
\]
Step III

<table>
<thead>
<tr>
<th>Genotype</th>
<th>GG</th>
<th>Gg</th>
<th>gg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

• Observed Genotypes

3. Compare the observed and expected counts:

\[
\chi^2 = \frac{(12 - 7.66)^2}{7.66} + \frac{(2 - 10.64)^2}{10.64} + \frac{(8 - 3.69)^2}{3.69} = 14.50
\]

REJECT THE NULL!
Reasons for HW Deviations

- **Genotyping Error**
- **Subdivided Population**
  - Excess homozygotes = “Wahlund Effect”
- Excess homozygotes= “Allele dropout in old samples”
- Any violations of the HW assumptions

- SNPs are excluded if substantially more or fewer samples heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency)
- Threshold for significance $10^{-3}$ to $10^{-6}$

GENEVA alcohol-dependence project: Quality control report
Identify SNPS which show extreme HWE deviations

**GET ALLELE FREQUENCIES**

```
plink --bfile clean_ind_example --hardy --out clean_ind_example_hwe --noweb
```

Generates the file "clean_ind_example_hwe.hwe" containing Hardy Weinberg statistics for each SNP separately in cases, controls and all samples.

**SELECT UNAFFECTED**

```
head -1 clean_ind_example_hwe.hwe > example_clean_ind_example_hweu.hwe | grep "UNAFF"
clean_ind_example_hwe.hwe >> example_clean_ind_example_hweu.hwe
```

**GENERATE PLOT USING R SCRIPT**

```
hwe_plot.R
```

(based only on controls)

**CHOOSE THE STANDARD HWE P-VALUE CUTOFF (0.00001) OR SELECT ONE ON THE BASIS OF THE PLOT**
SNP/Marker QC final

Differential missingness

HWE outliers

High Missingness

Low MAF

plink --bfile clean-inds-example
--maf 0.01
--geno 0.05
--exclude fail_diffmiss_example.txt
--hwe 0.00001
--make-bed --out clean-example

plink --no-web --bfile clean-example --chr X --make-bed --out xsnps
plink --no-web --bfile clean-example --exclude x_snps --make-bed --out qced_example

QCed data ready for assoc !!

in most cases you would need additionally to remove the X and Y chromosomes
thank you!