GWAS Exercises 5 - Quality Control for Genomic Data

Peter Castaldi - adapted from Merry-Lynn McDonald

January 31, 2013

1 Elements of Quality Control - QC

Quality control, or QC, is a critical element in doing GWAS. Since hundreds of thousands of genotypes are generated, even a small percentage of genotyping error can, if unidentified, lead to spurious GWAS results. QC can be considered to have two aspects - QC related to genotyping chips (i.e. issues related to making genotype calls from intensity measurements) and downstream QC issues. In this exercise, we focus exclusively on downstream QC approaches, i.e. data cleaning procedures that can be applied once you already have genotype calls.

Downstream QC covers two major areas of quality - subject-based quality measures and variant-based quality measures. The specific QC measures for the two domains are:

Subject-Based Measures

- Subject-Specific Missingness Rate - proportion of missing genotypes per subject
- Gender - check that self-reported gender matches genotyped gender
- Relatedness - undisclosed familial relationships, duplicate enrollment
- Replicate Discordance - agreement with independent genotyping
- Population Outliers - subjects with significantly different genetic background from rest of study sample

Variant-Based Measures

- Variant-Specific Missingness Rate - proportion of failed assays for a variant
- Minor Allele Frequency - very low frequency alleles are more likley to represent genotyping error and can give spurious association results

- Hardy-Weinberg equilibrium
- Mendelian Errors - in family data evidence of non-Mendelian transmission
- Replicate Discordance

2 Implementing QC

We're going to review how to actually do QC in PLINK. For this exercise we are going to use genomewide data from the HapMap CEU population.

2.1 Check Gender Assignment

The first step is to check to make sure that the gender of a subject matches with their number of X chromosomes.

```
plink --file /cluster/tufts/cbicourse/GAS/DATA/hapmap3_pop/hapmap3_r2_b36_fwd.CEU.qc.poly
  --check-sex --out sexcheck --noweb
```

Look at the file by typing:

less sexcheck.sexcheck

Are there any subjects where the X-chromsome data disagrees with the reported gender?

Problematic subjects can be removed from the dataset. There is a PLINK command for this, but we need to make a text file of the FID and IIDs of the individuals to remove first. Open another window to run R.

```
module add R/2.15.0

bsub -Ip -q int_public6 R
```

```
> temp1 <- "1349"
> temp2 <- "NA10854"
> exc <- data.frame(FID = temp1, IID = temp2)
> write.table(exc, file = "exclude.txt", row.names = F, col.names = T,
+     quote = F)
```

In the PLINK window, now remove this subject from the data using the '–remove' flag, and write a compressed BED file to your directory. As you can

see, working with genome-wide data can be a bit tedious because it takes a while for each step to run. For convenience, let's also use the '–chr' flag to reduce the genotype data to just one chromosome.

```
plink --file /cluster/tufts/cbicourse/GAS/DATA/hapmap3_pop/hapmap3_r2_b36_fwd.CEU.qc.poly
--remove exclude.txt --chr 17 --make-bed --out CEU_chr17
```

```
Check the log file to confirm that one person was removed.
```

2.2 Remove Subjects with More than 10 Percent Missing genotypes

If the quality of the DNA sample from an individual is poor, there will be a higher rate of uncalled genotypes. This is a marker for poor DNA quality overall. To screen out these subjects use the '–mind' command. 'Mind' sets the maximum rate of per-individual missingness.

```
plink --bfile CEU_chr17 --mind 0.1 --make-bed --out CEU_chr17_clean
```

Question 1: Look at the plink log. Is there anyone with more than 10 percent missing genotype data?

2.3 Mendel Errors

If you have families in your data, you can use the familial relationships to do some detailed quality control with the '–mendel' flag. We're not going to go into detail on this, but the command to run is:

```
plink --bfile CEU_chr17_clean --mendel
```

The output files are 'plink.mendel', 'plink.lmendel' and 'plink.imendel'. They have the plink prefix because we did not use the '–out' flag. The PLINK documentation has a nice discussion of what is included in these files under the 'Summary Stats' section.

Most GWAS studies of common diseases are not family-based, so we are going to remove the related individuals from this dataset using the '–founder' flag.

```
plink --bfile CEU_chr17_clean --filter-founders --make-bed --out CEU_chr17_clean
```

Question 2: How do you think PLINK defines a founder (i.e. in relation to variables in the first six columns of a pedfile)?

We now have 112 unrelated individuals in the data set.

2.4 Removing SNPs with high rate of missing genotype calls

We already used the –mind flag to remove samples with a high rate of poor genotypes. We now are going to use the –geno flag to remove SNPs that have a high genotyping failure rate. This can be due to poor primer design and non-specific DNA binding to a particular SNP probe.

```
plink --bfile CEU_chr17_clean --geno 0.05 --make-bed --out CEU_chr17_clean
```

Question 3: How many SNPs have a missing genotype rate greater than 5 percent?

2.5 Removing SNPs out of Hardy-Weinberg equilibrium

Population genetic theory suggests that under 'normal' conditions, there is a predictable relationship between allele frequencies and genotype frequencies. In cases where the genotype distribution is different from what one would expect based on the allele frequencies, one potential explanation for this is genotyping error. Natural selection is another explanation. For this reason, we typically check for deviation from Hardy-Weinberg equilibrium in the controls for a case-control study. For a quantitative trait, PLINK just uses everyone. The following command generates p-values for deviation from HWE for each SNP. Low p-values indicate that a SNP is out of HWE.

```
plink --bfile CEU_chr17_clean --hardy --out CEU_chr17_hardy
```

When you run this command, each SNP has three rows of results corresponding to the HWE test for SNPs in all subjects, cases (AFF) and controls (UNAFF). You can sort and parse the results in the R window. We will read in the results, look at the structure of the hwe object, check how many rows are in the data frame, then filter out just the HWE results which are calculated using the entire cohort. We will also order the results by p-value and look at the top five results.

```
> hwe <- read.table("CEU_chr17_hardy.hwe", header = T, stringsAsFactors = F)
> str(hwe)
```

```
'data.frame':        109791 obs. of  9 variables:
 $ CHR   : int  17 17 17 17 17 17 17 17 17 17 ...
 $ SNP   : chr  "rs8069278" "rs8069278" "rs8069278" "rs6565733" ...
 $ TEST  : chr  "ALL" "AFF" "UNAFF" "ALL" ...
 $ A1    : chr  "G" "G" "G" "G" ...
 $ A2    : chr  "A" "A" "A" "A" ...
 $ GENO  : chr  "1/38/73" "0/0/0" "0/0/0" "1/29/82" ...
 $ O.HET.: num  0.339 NaN NaN 0.259 NaN ...
 $ E.HET.: num  0.293 NaN NaN 0.238 NaN ...
 $ P     : num  0.19 NA NA 0.69 NA ...
```

```
> nrow(hwe)

[1] 109791

> hweall <- hwe[which(hwe$TEST == "ALL"), ]
> nrow(hweall)

[1] 36597

> hweall <- hweall[order(hweall$P), ]
> hweall[1:5, ]
```

```
      CHR        SNP TEST A1 A2    GENO  O.HET.  E.HET.         P
709    17   rs379248  ALL  C  T 0/65/47 0.58040 0.41190 1.722e-06
34519  17  rs6505076  ALL  G  T 4/3/102 0.02752 0.09583 8.168e-06
56734  17  rs2959971  ALL  G  A 0/59/53 0.52680 0.38800 4.101e-05
69394  17 rs17817901  ALL  G  A 0/59/53 0.52680 0.38800 4.101e-05
57652  17   rs155733  ALL  T  C 0/58/53 0.52250 0.38600 4.350e-05
```

Question 4: How many SNPs have a HWE p-value of 10-5 or less?

You can also use the –hwe command to automatically remove SNPs above a certain HWE p-value threshold.

```
plink --bfile CEU_chr17_clean --hwe 1E-4 --out CEU_chr17_clean
```

2.6 Setting a minimum minor allele frequency

Genetic associations with SNPs with a low minor allele frequency can give spurious results. Any ideas why? It's common practice to remove SNPs with very low minor allele frequency prior to analysis. This is achieved with with '–maf' flag.

```
plink --bfile CEU_chr17_clean --maf 0.05 --out CEU_chr17_clean_maf
```

You can see that in the HapMap data, there are quite a lot of low MAF SNPs.