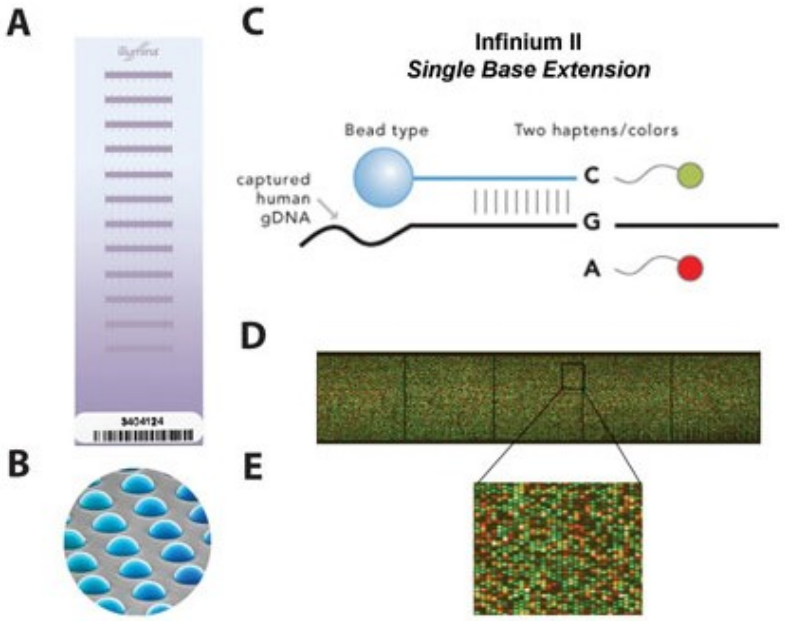
QUALITY CONTROL

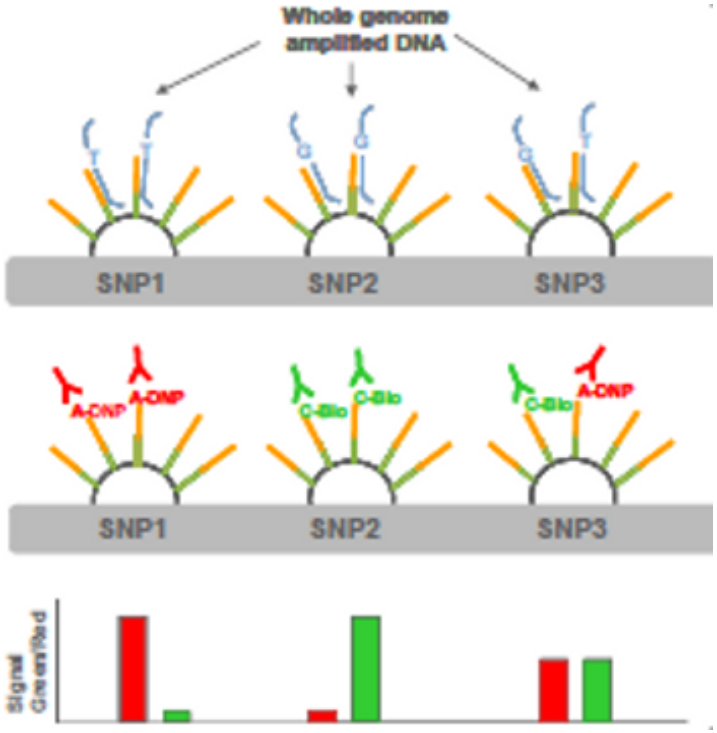
practical

Why Do Quality Control?

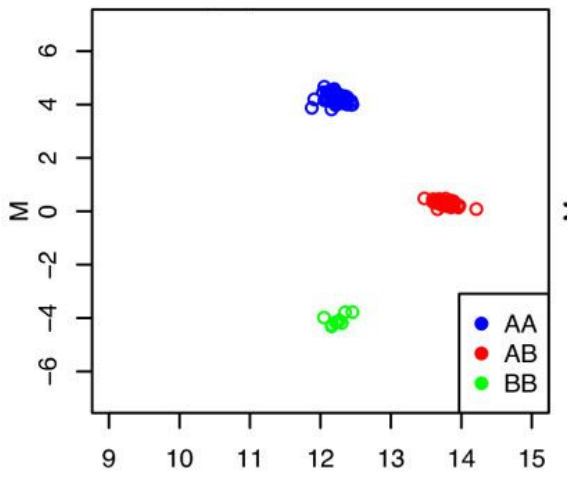
- To remove genotyping errors
  - Low quality or quantity of DNA
  - Contaminated DNA
  - Chemical or machinery failure
  - Human error
- To ensure data suitable for the analyses
  - Relatedness
- **Poor quality data ⇒ false positives / negatives**

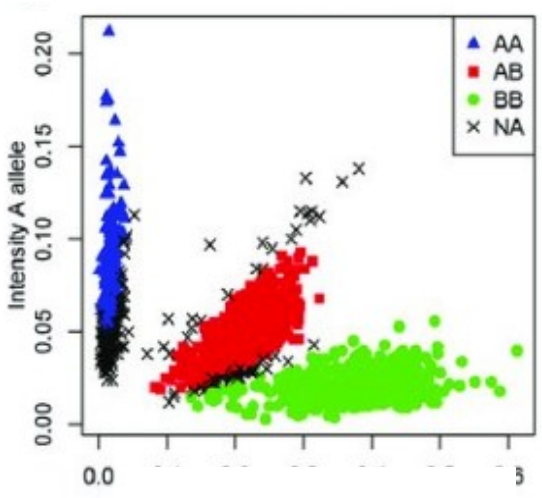
From DNA to data

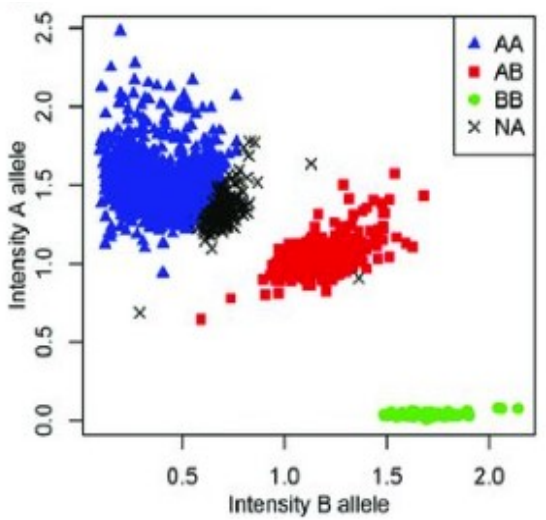


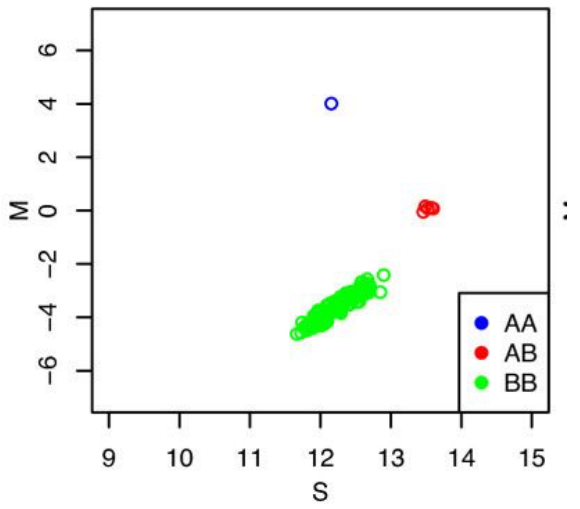


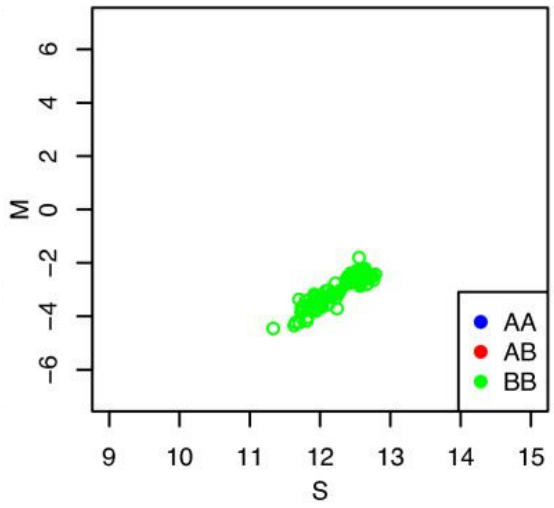
Genotyping Intensities

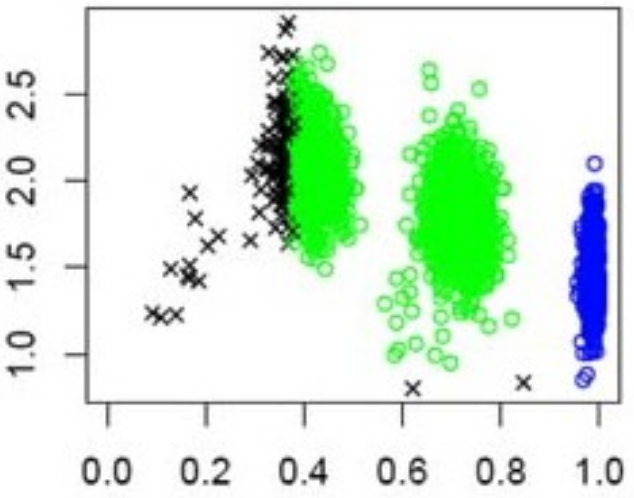












Key Steps in QC

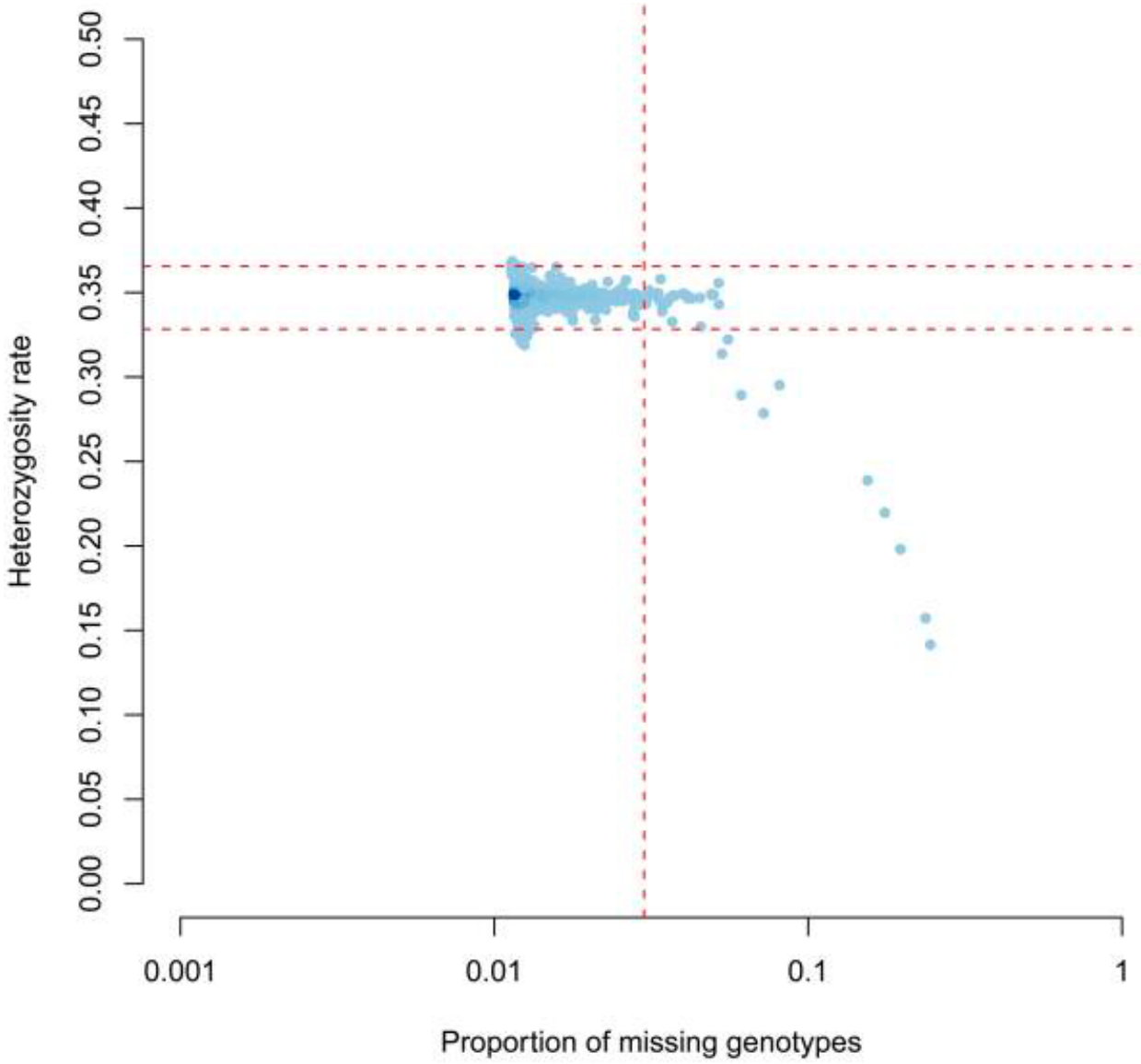
- Sex-check (chr X heterozygosity)

- Genotyping Call Rate (SNPs missing individuals)
- Hardy-Weinberg Equilibrium
- Minor Allele Frequency

- Sample Call Rate (individuals missing genotypes)
- Proportion of Heterozygosity

- Relatedness
- Population Stratification

Heterozygosity vs. Missing



Anderson et al. *Nat Proc.* 2010

Today's Practical

http://labs.med.miami.edu/myers/LFuN/LFUN/DATA.html

LETTERS

nature genetics

doi: 10.1038/ng.207.16

A survey of genetic human cortical gene expression

Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, Victoria L Zismann, Keta Joshipura, Matthew J Huentelman, Diane Hu-Lince4, Keith D Coon, David W Craig, John V Pearson, Peter Holmans, Christopher B Heward, Eric M Reiman, Dietrich Stephan & John Hardy

- 193 individuals
- genotypes from this published data set, but we have made up the phenotype for the purpose of this QC practical.

Files for practical

Data files to clean

- cc.ped
- cc.map

Rscript to plot data

- plink-qc.R

- Text file with unix & plink command

part1_practical_commands.txt

Check the data

```
less -S cc.ped
```

- The ped file contains both phenotype and genotype information.
- Columns: FID IID PID MID SEX Trait SNPs…

```
WGACON 1 0 0 1 2 C T G A A G T T C
WGACON 6 0 0 1 1 C T G A A G 0 0 C
WGACON 7 0 0 1 2 T T A A G G T T C
WGACON 9 0 0 2 1 T T A A G G T T C
```

- Check coding of sex, missing data, case-control status

```
awk
'NR==1{min=max=$6};$6<min{min=$6};$6>max{max=$6}END{
print min,max}' cc.ped
```

Plink resources

https://www.cog-genomics.org/plink/1.9

http://zzz.bwh.harvard.edu/plink/

Default Plink coding

Male = 1, Female = 2;

Case = 1, Control = 2, Missing = -9 or 0

Check the data

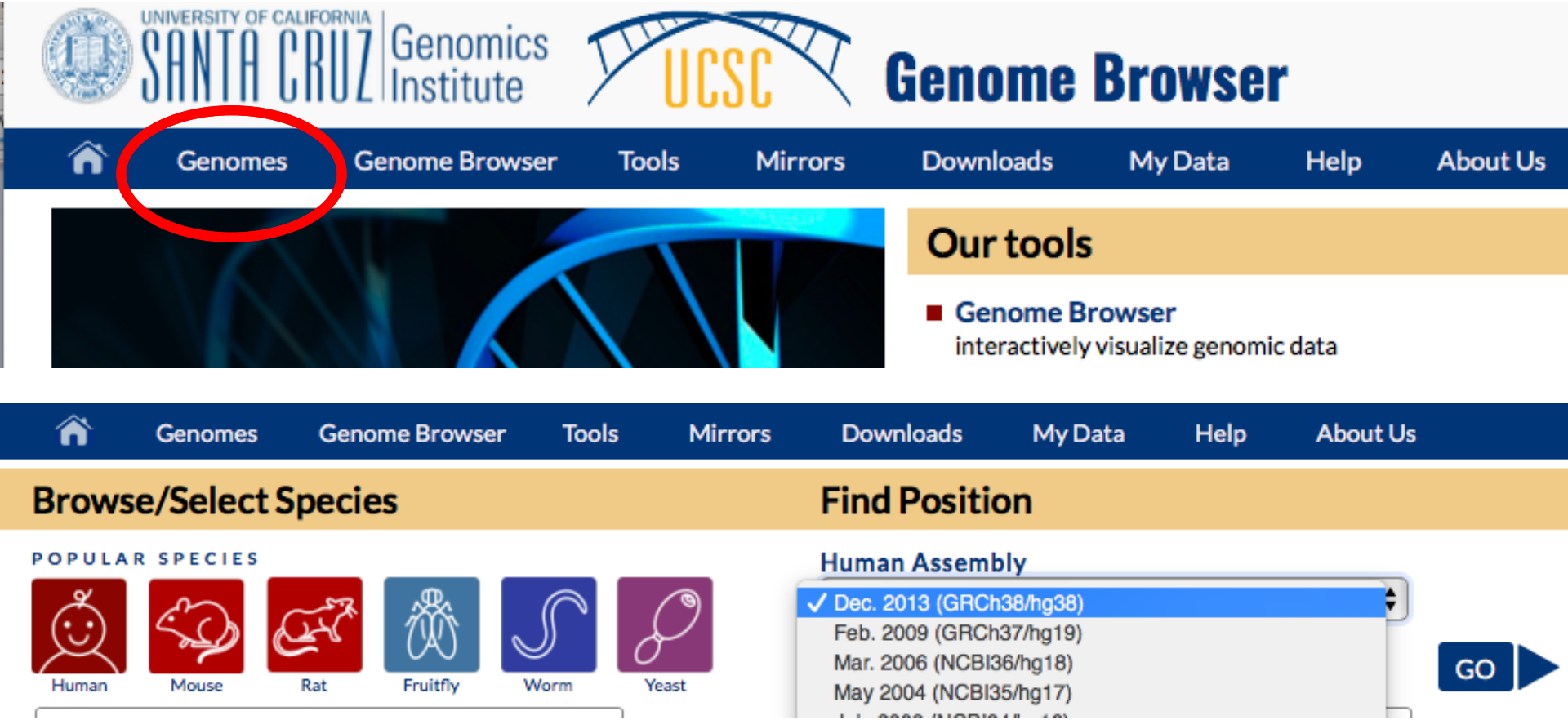
```
less -S cc.map
```

- Map file contains location/position information in the genetic variants
- Columns: CHR SNP centimorgan BP

```
1       rs3094315       0       752566
1       rs4040617       0       779322
1       rs2980300       0       785989
1       rs2905036       0       792480
```

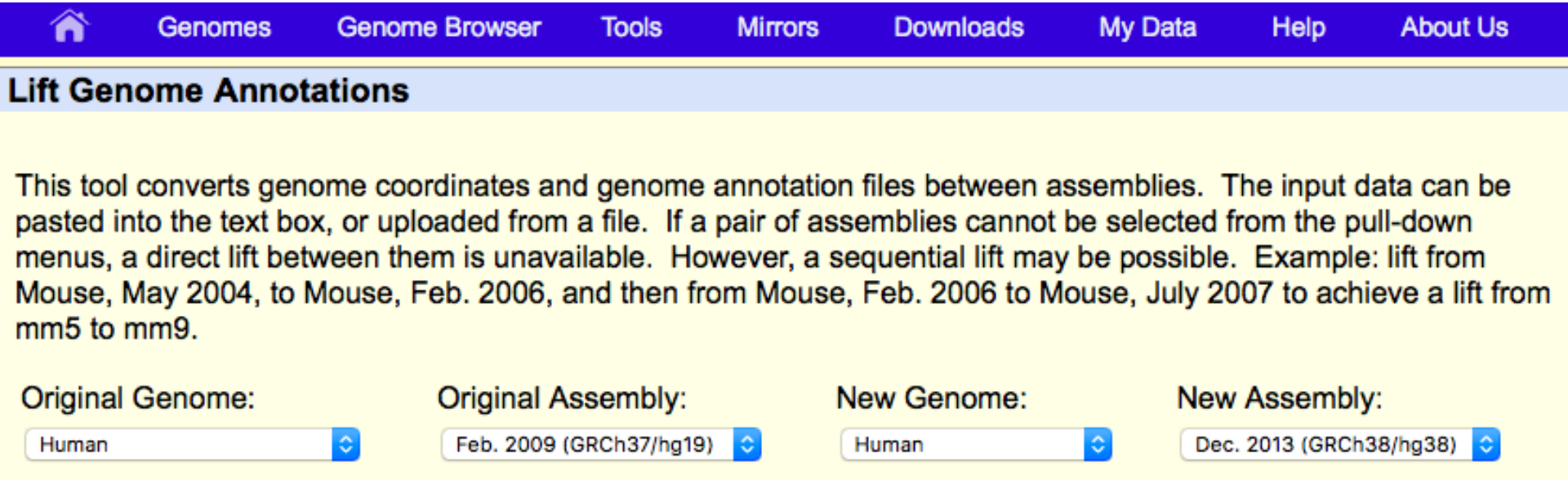
Check the build

https://genome.ucsc.edu



Remap build (if required)

https://genome.ucsc.edu/cgi-bin/hgLiftOver



Plink file format --bfile

- Plink has a binary file format for genotyped data that reduces the size of the ped file.

- Converting to bfile format holds the phenotypic and genetic data in 3 files.
- The phenotype information is contained in the .fam file
- The genotype data is converted into a binary form and saved in the .bed file
- The map file is converted to a .bim file, which in addition to the position information it includes the alleles.

Convert file format to bfile

```
plink --ped cc.ped --map cc.map --make-bed --out
cc.begin
```

- Errors in Plink will stop progress, warnings are for you to make decisions about

- Warning: 2177 het. haploid genotypes present (see cc.begin.hh)
- Heterozygous haploid errors may be pseudo-autosomal region of chr X, if they are can use the split-x fcommand
- First we will check for sex errors in the data.

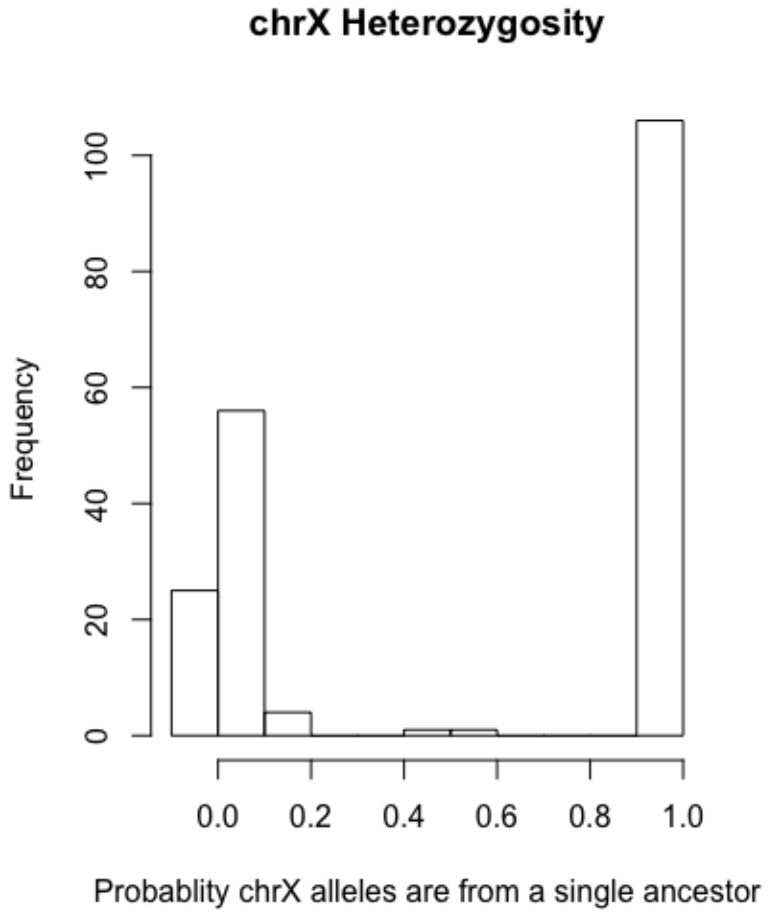
Check sex

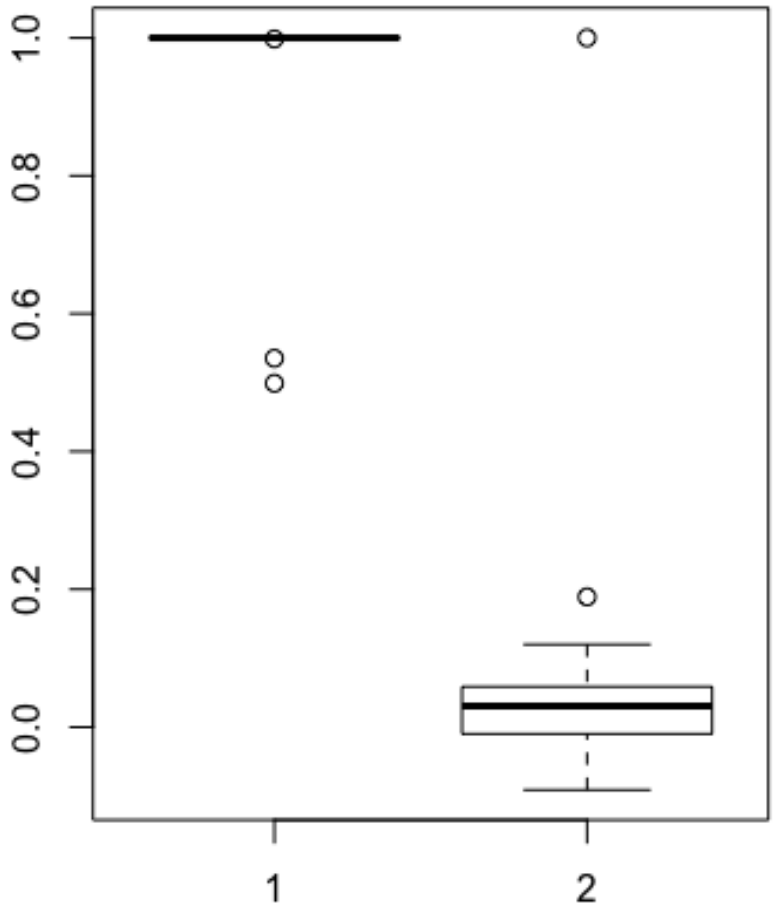
```
plink --bfile cc.begin --check-sex --out sex

less sex.sexcheck
```

- Checking sex early in QC can be a useful to see if there has been a plate rotation in the genotyping.
- check-sex is a heterozygosity check of chr X
- Males have one chr X, females have 2
- The F statistic is the probability an individual inherited two identical alleles from a single ancestor.
- In the case of males this should be close to 1, for females it should be close to 0.

Plot Sex (plink-qc.R)





Remove IDs mismatched on sex

- Select the 3 individuals flagged as not a match between reported and genotyped sex

```
grep PROBLEM sex.sexcheck > sex.drop
```

- Remove those individuals
- plink --bfile cc.begin --remove sex.drop --make-bed --out cc.qc1

```
plink --bfile cc.begin --remove sex.drop --make-bed
--out cc.qc1
```

het haploid

- Use split-x to see if the remaining issues are due to the pseudo-autosomal regions

```
plink --bfile cc.qc1 --split-x b37 no-fail --make-
bed --out cc.qc2
```

- To check chromosome, this awk code will match on the chromosome from the bim file to the remaining het.haploid variants

```
awk 'NR==FNR{a[$2]=$1; next} $3 in a{print
$0,a[$3]}' cc.begin.bim cc.qc2.hh > check.hh
```

het haploid

- Set the remaining het haploid issues to missing

```
plink --bfile cc.qc2 --set-hh-missing --make-bed --
out cc.qc3
```

- These will be excluded from analysis anyway.

Create files for plotting

- Obtain heterozygosity information (this is across chr 1-22)

```
plink --bfile cc.qc3 --het --out het
```

- Create file with the proportion of heterozygosity for each person.

```
echo "FID IID obs_HOM N_SNPs prop_HET" > het.txt

awk 'NR>1{print $1,$2,$3,$5,($5-$3)/$5}' het.het >>
het.txt
```

Call Rate / Missingness

1. Sample = Individuals missing genotypes
2. Genotyping = variants missing data from individuals

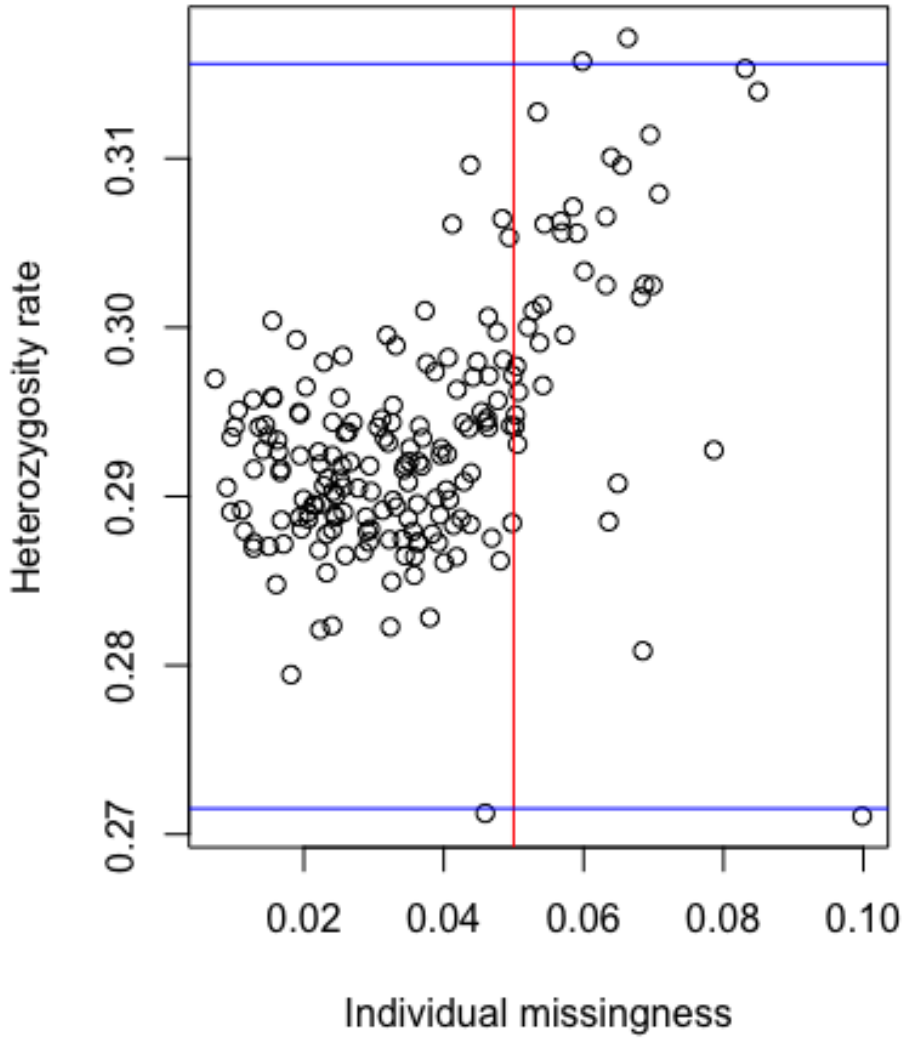
- Obtain missingness on individuals and on SNPs

```
plink --bfile cc.qc3 --missing --out miss
```

- Create a file that holds both proportion of heterozygosity and individual missingness (for plotting)

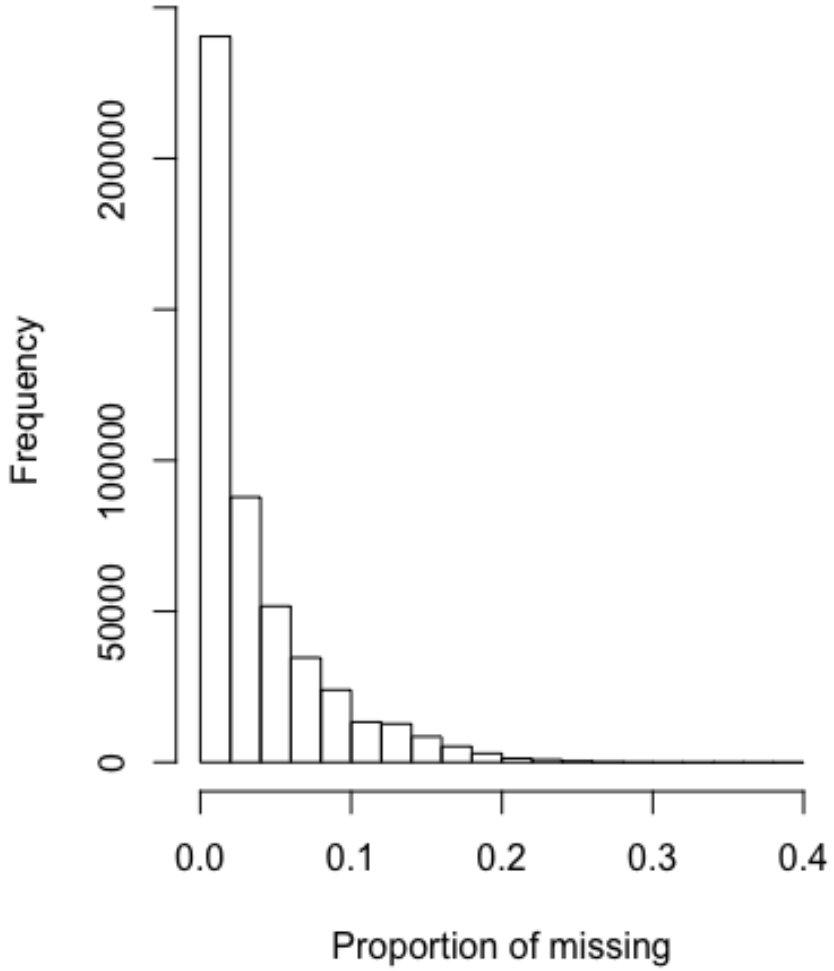
```
awk 'NR==FNR{a[$1,$2]=$5;next}($1,$2) in a{print
$1,$2,$6,a[$1,$2]}' het.txt miss.imiss >
het.imiss.txt
```

Plot Heterozygosity by Sample Call Rate



Genotyping Call Rate

- Plot SNP call rate, variants with proportion of data missing



Genotyping Call Rate

- Remove variants that are missing too much data from individuals

```
plink --bfile cc.qc3 --geno 0.05 --make-bed --out
cc.qc4
```

Genotyping Call Rate by Case-Control

- With Case-Control data check that SNP call rate is not significantly different between cases and controls

```
plink --bfile cc.qc4 --test-missing --out case-control
```

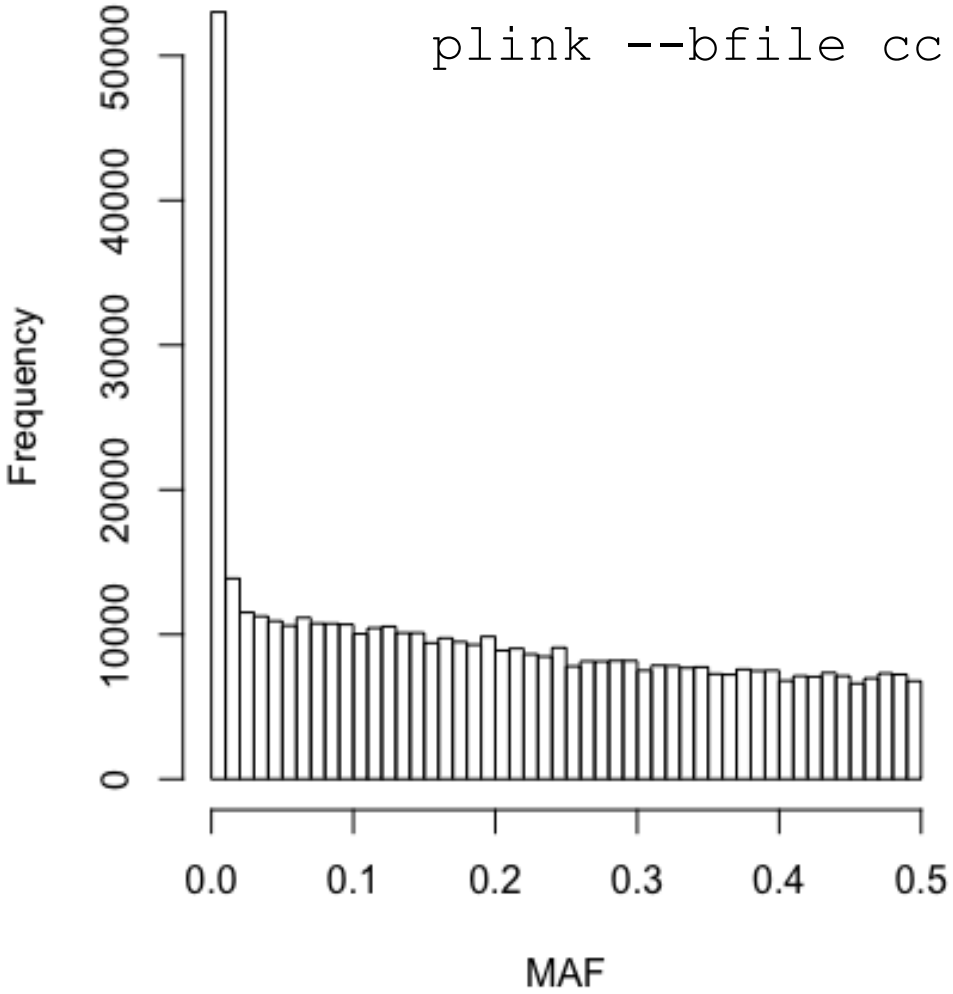
- This awk code will copy any variants that differ in missingness with a p value < .00001 into a file that can be used to exclude them

```
awk '$5<=0.00001' case-control.missing > case-control.missing.drop
```

MAF

- Obtain MAF information

```
plink --bfile cc.qc3  --freq --out maf
```



Hardy-Weinberg Equilibrium

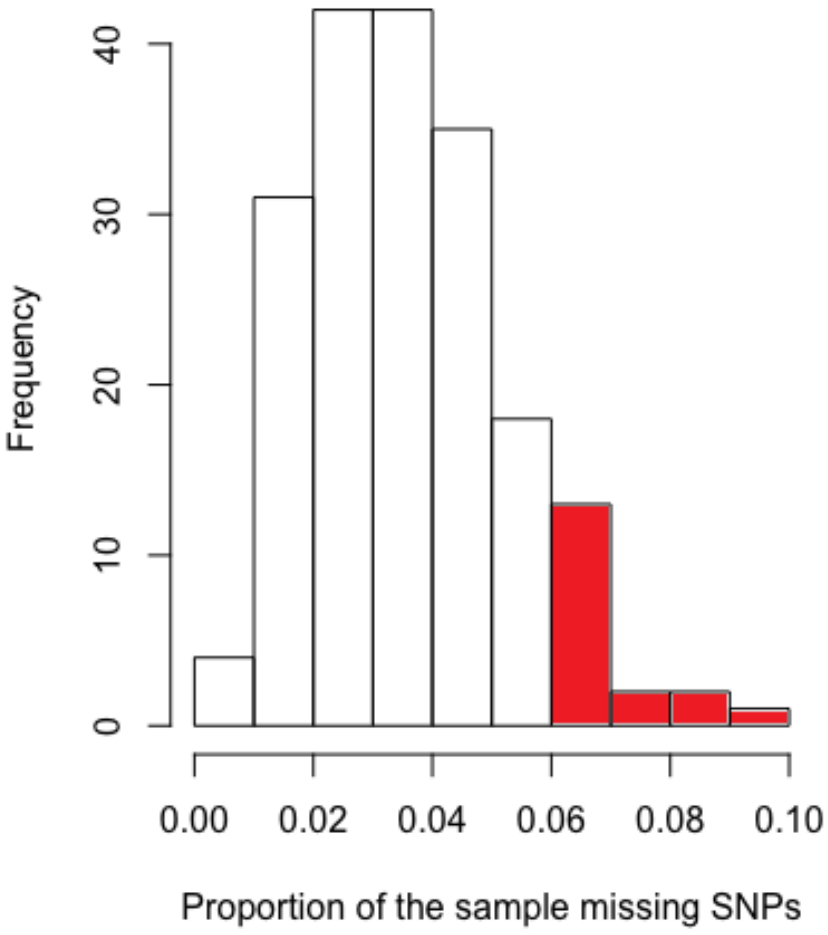
- The default HWE check on case-control data is to only conduct the check on the controls.
- We have overwritten that here with the [include-nonctrl] flag.

```
plink --bfile cc.qc4 --maf 0.01 --hwe 1e-6 include-
nonctrl --make-bed --out cc.qc5
```

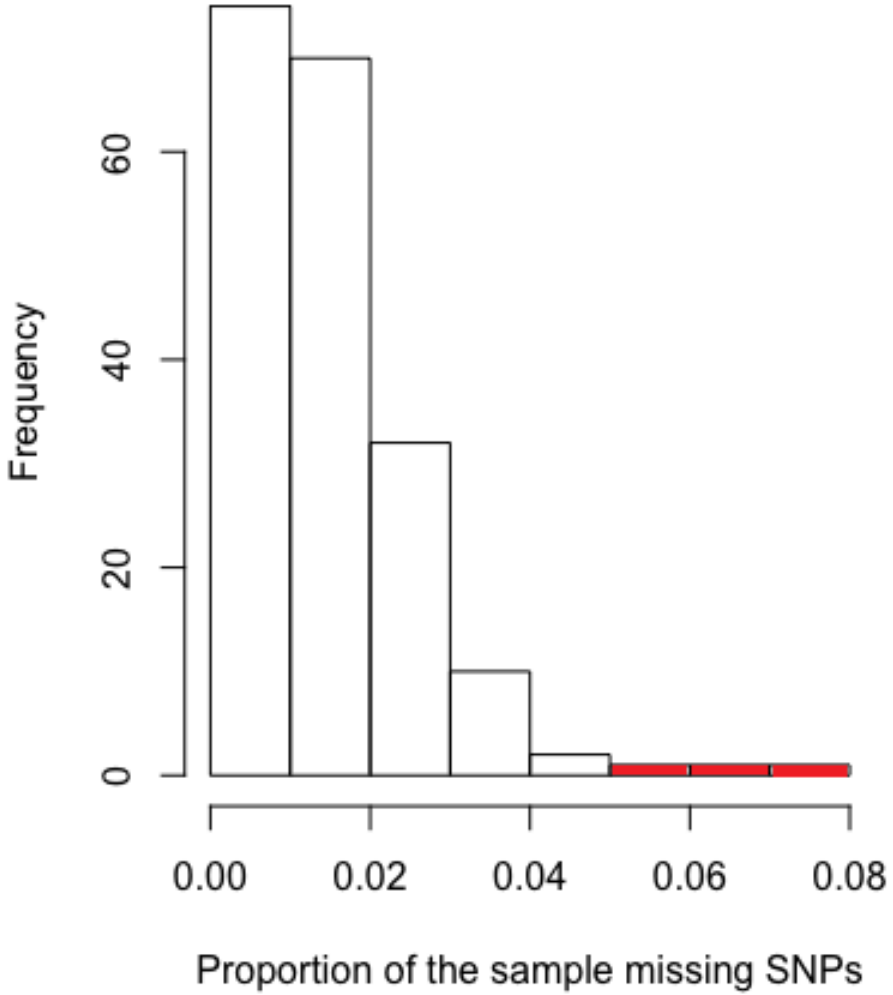
- MAF of .01 in a sample of 190 = MAC of 3.8

….(a bit liberal)

Sample Call Rate



Before QC on genotypes



After QC on genotypes

Sample Call Rate

- Remove individuals that are still missing information on more than 5% of the variants

```
plink --bfile cc.qc5 --mind 0.05 --make-bed --out
cc.qc6
```

- Obtain the heterozygosity scores now that the data variant have been cleaned.

```
plink --bfile cc.qc6 --het --out het2
```

Heterozygosity

- Obtain +/- 3SD cut off for heterozygosity
- First obtain the proportion of heterozygosity per person

```
echo "FID IID obs_HOM N_SNPs prop_HET" > het2.txt
awk 'NR>1{print $1,$2,$3,$5,($5-$3)/$5}' het2.het >>
het2.txt)}' het2.txt
```

- This awk code will give +3SD and -3SD for a file with a header

```
awk 'NR>1{sum+=$5;sq+=$5^2}END{avg=sum/(NR-1);print
avg-3*(sqrt(sq/(NR-2)-2*avg*(sum/(NR-2))+(((NR-
 )*(avg^2))/(NR-2)))),avg+3*(sqrt(sq/(NR-2)-
 *avg*(sum/(NR-2))+(((NR-1)*(avg^2))/(NR-2))))}'
het2.txt
```

Heterozygosity

- Copy the individuals outside 3SD from the mean heterozygosity score into a file

```
awk '$5<=0.29957 || $5>= 0.326526' het2.txt  >
het.drop
```

- Drop those participants

```
plink --bfile cc.qc6 --remove het.drop --make-bed --
out cc.clean
```

Relatedness

- Prune the SNPs and select chr 1-22

```
plink --bfile cc.clean --chr 1-22 --indep-pairwise
1000 5 0.2 --out indep
```

- Calculate identity by descent (IBD) on those pruned SNPs

```
plink --bfile cc.clean --extract indep.prune.in --
genome --out ibd
```

Relatedness

- Have a look at the Identity by descent results

```
less ibd.genome
```

- Z0 Z1 Z2 = the pair of IDs share 0, 1 or 2 alleles by descent
  - 1,0,0 = unrelated (ideallistic)
  - 0,1,0 = parent-child
  - .25,.5,.25 = siblings etc
- pihat = proportion IBD = P(IBD=2)+0.5*P(IBD=1)
  - 0 = unrelated
  - .5 = parent-child or siblings
  - .125 = cousins

Relatedness

- Obtain a list of individuals related more than a pihat of .2

```
plink --bfile cc.clean --extract indep.prune.in --
genome --min 0.2 --out pihat
```

- Check them in the file pihat.genome

```
less pihat.genome
```

Population Stratification

- Systematic differences in allele frequency across populations
- Principal components (PC) is one method to decompose the variation in allele frequency into components that describe the population structure
- PCs can be included as covariates in GWAS as a way to control for population structure as a confounder.
- (more on this Tuesday morning)