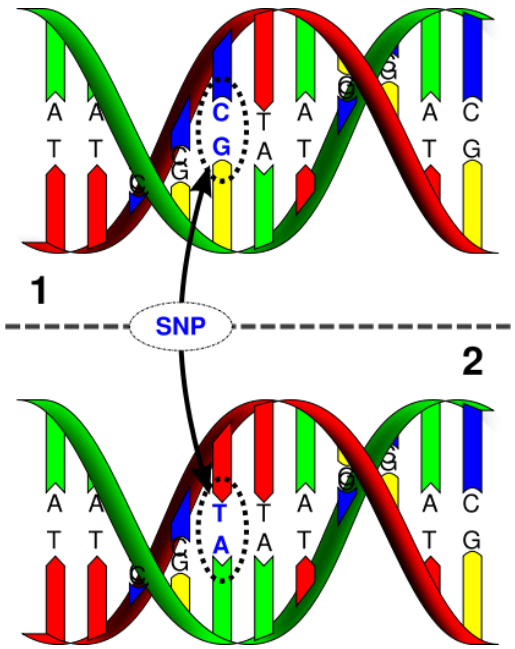
Module 2: Introduction to PLINK and Quality Control

1. Introduction to PLINK
2. Quality Control

Single Nucleotide Polymorphism (SNP)

A SNP (pronounced snip) is a single basepair at which more than one nucleotide is observed. E.g., if Basepair 1 000 000 is nucleotide C for many genomes in the population, but some genomes have nucleotide T, then this position is a SNP, with alleles C and T. Because mutations occur rarely, only two alleles are observed at the majority of SNPs, but three or even all four possible alleles (A,C,G,T) are possible



The Minor Allele Fraction (MAF) is the relative frequency in a relevant population of the minor (2nd most common) allele. People often talk of common and rare SNPs, although the distinction is not well defined, and likely depends on sample size: some consider common as MAF >0.05, others MAF >0.01

Single Nucleotide Polymorphisms

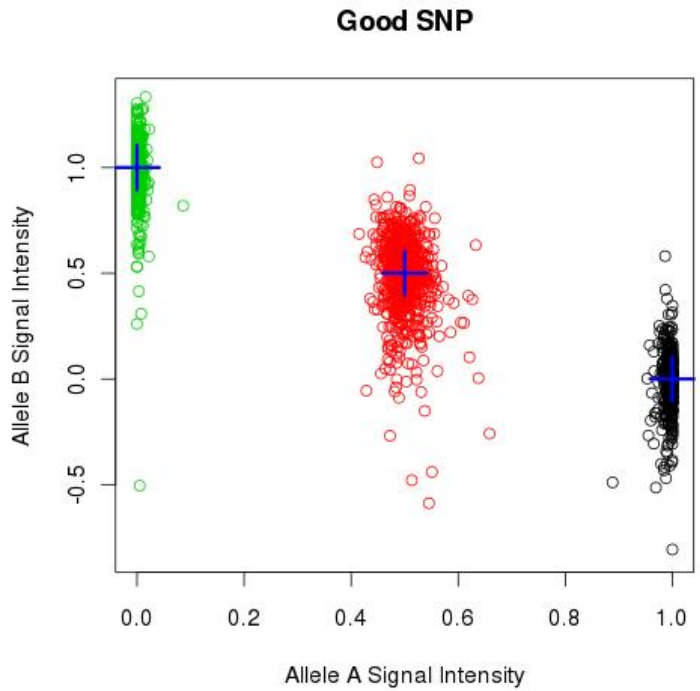
SNP genotypes are usually encoded as 0, 1 or 2, based on the number of copies of non-reference alleles. E.g., suppose a SNP has reference allele G, and A is the only variant allele observed, then

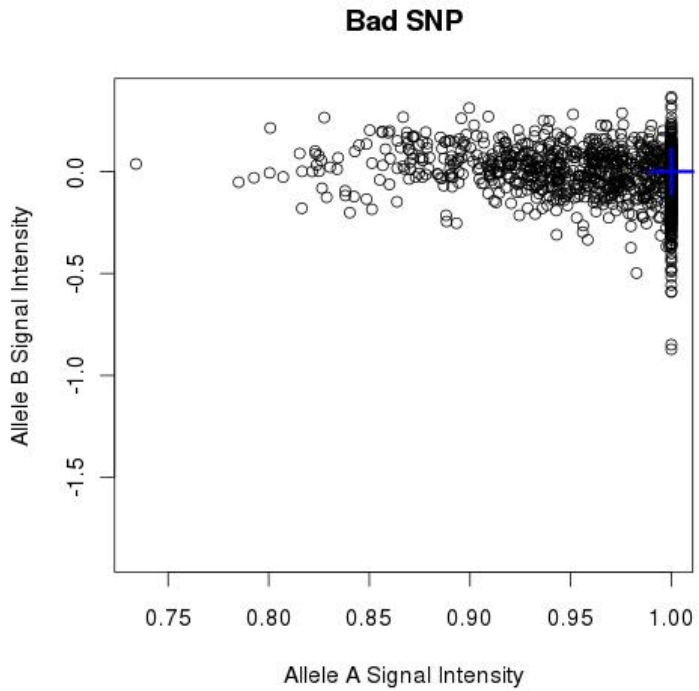
- genotype CC is coded as 0 (homozygous non-reference)
- genotype CT is coded as 1 (heterozygous)
- genotype TT is coded as 2 (homozygous reference)

The reference allele is sometimes referred to as the "wildtype" allele, while the non-reference is referred to as the alternative or "mutant" allele

Usually the reference allele is the most common one, but this is not universally true, and the most common allele in one population may not be most common in another. Care must be taken when combining datasets to ensure that the reference alleles are the same.

Genotype calling is a major challenge





Each circle represents an individual's genotype at the SNP. We want to see tight clusters, one for each genotype, otherwise the SNP may be rejected. Individual genotypes can be assigned probabilities based on their positions relative to the clusters. Genotypes below a threshold (e.g. 0.9 or 0.95) can be coded as missing, or genotype probabilities can be incorporated into some analyses.

Storing SNP Data in PLINK

http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml

Suppose we have n individuals genotyped for N SNPs

$$\mathbf{X} = \begin{bmatrix} AA & CG & TT & \dots & GG \\ AG & CG & AT & \dots & CG \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ GG & CG & 00 & \dots & CC \end{bmatrix} \begin{matrix} \leftarrow \text{Individual 1} \\ \leftarrow \text{Individual 2} \\ \vdots \\ \leftarrow \text{Individual n} \end{matrix}$$

SNP 1 SNP 2 SNP 3 SNP N

The genotypes correspond to a matrix $\mathbf{X}$ of size $n \times p$, where N is usually $\gg n$ (i.e. the matrix is short and very fat)

Pedigree Files

```
doug@doug-laptop:~/Desktop/workingdir/wtccc$ head test.ped -n 9
IND1 IND1 0 0 2 1 A A G A A T
IND2 IND2 0 0 1 1 A A A A A T
IND3 IND3 0 0 2 1 A A A A A T
IND4 IND4 0 0 2 1 A A 0 0 A T
IND5 IND5 0 0 1 1 A A A A T T
IND6 IND6 0 0 2 1 A A A A A T
```

Ped files have $6 + 2n$ columns, providing:

1. Family ID,
2. Individual ID,
3. Paternal ID (0 if father not in dataset),
4. Maternal ID (0 if mother not in dataset),
5. Sex (1=Male, 2=Female, Other=Unknown),
6. Phenotype (here 1 or 2, corresponding to case and control),
7. 2 alleles for each SNP (0 = missing)

Pedigree Files

```
doug@doug-laptop:~/Desktop/workingdir/wtccc$ head test.map
1       rs12025928      0       536560
1       rs6421779       0       555887
1       rs1780604       0       15449253
```

Map files have 4 columns, providing:

1. Chromosome,
2. SNP Name,
3. Genetic Distance from previous SNP (0 = unknown),
4. Basepair

Bed (Binary Pedigree) Files

PLINK also offers a (much) more efficient binary format, which uses fam(ily), bim (binary mapping) and bed (binary pedigree) files.

Fam files are first six columns of ped file

```
doug@doug-laptop:~/Desktop/workingdir/wtccc$ head -n 9 test.fam
IND1 IND1 0 0 2 1
IND2 IND2 0 0 1 1
IND3 IND3 0 0 2 1
IND4 IND4 0 0 2 1
IND5 IND5 0 0 1 1
IND6 IND6 0 0 2 1
IND7 IND7 0 0 1 1
IND8 IND8 0 0 1 1
IND9 IND9 0 0 2 1
```

Bed (Binary Pedigree) Files

PLINK also offers a (much) more efficient binary format, which uses bed (binary pedigree, next slide), bim (binary mapping) and fam(ily) files.

Bim files are map files plus two columns, providing the A1 and A2 alleles

```
doug@doug-laptop:~/Desktop/workingdir/wtccc$ head test.bim
1       rs12025928      0       536560  G       A
1       rs6421779       0       555887  G       A
1       rs1780604       0       15449253        A       T
```

In PLINK, the SNP genotype (0/1/2) represents the number of copies of the A1 allele; when performing association analysis (see later modules), effect sizes / odds ratios will be reported w.r.t. the A1 allele

However, this is not universally true, so be careful!

The Bed File

The bed file is a matrix of 0s, 1s, 2s or NAs stored in binary format

In computer lingo, a bit is one pieice of information (represented by 0 or 1)

A byte is 8 bits (e.g., 01101011) - can store $2^8 = 256$ bits of information

To save the four possible genotypes (0/1/2/NA) requires two bits

so each byte can store 4 genotypes

`PLINK uses the following two-bit coding of genotypes:`

- `00 = A1/A1 (Homozygous non-reference)`
- `01 = A1/A2 (Heterozygous)`
- `11 = A2/A2 (Homozygous reference)`
- `10 = 0/0 (Missing)`

The Bed File

The first three bytes of the bed file are headers:

- magic number -		mode
01101100 00011011		0000001

The remainder each contain 4† genotypes, which are "read backwards":

0 1 1 0 1 1 0 0
HGFEDCBA

AB	$00 = A1/A1$
CD	$11 = A2/A2$
EF	$01 = A1/A2$
GH	$10 = 0/0$

† SNPs are "padded-out" to ensure they have a multiple of four genotypes. So suppose there were 998 individuals, bits HGFE of the final byte would not be used

The Bed File

Each bed file has size 3 + p x ceiling(n/4) bytes

e.g., to store 100 SNPs for 10 individuals takes 3+100*3=303 bytes

It might actually be more efficient to compress the ped file. However, a key feature of bed files is that genotypes can be accessed immediately

Suppose we have 10 individuals; to obtain the genotype for individual 5, SNP 10, read the byte located at position 3+(9x3)+1

(this will contain genotypes for SNP 10 for individuals 5 to 8)

See http://pngu.mgh.harvard.edu/~purcell/plink/binary.shtml for more information

Other file formats include VCF and Oxford (a.k.a. Chiamo)

C code for reading SNPs from bedfile

```
//int num_samples / num_preds contain total numbers of individuals / SNPs in bed file
//will read from SNP (start+1) to SNP end and put into matrix data (num_samples x num_preds)
//in most implementations, sizeof(char)=1

int read_bed_fly(char *bedfile, double *data, int num_samples, int num_preds, \
int start, int end, double missingvalue)
{
int i, j, count, current, rowlength, gen, gen2;

unsigned char check[3], *rowbytes;
double conv[4], *datatemp;

FILE *input;

//four values stored per byte, each SNP will take up ceiling(num_samples/4) bits
rowlength=(int)((num_samples-1)/4)+1;
rowbytes=malloc(sizeof(unsigned char)*rowlength);

//the codes 0, 1, 2, 3 map to genotypes 2, NA, 1, 0
conv[0]=2.0;conv[1]=missingvalue;conv[2]=1.0;conv[3]=0.0;

//open the file
if((input=fopen(bedfile,"rb"))==NULL)
{printf("Error opening %s\n\n",bedfile);exit(1);}

//check the file has length (3+num_preds*rowlength) bytes
fseeko(input, 0, SEEK_END);
if(ftello(input)!=(off_t)sizeof(unsigned char)*rowlength*num_preds+sizeof(unsigned char)*3)
{printf("Error reading %s; should have size %jd (%d ind x %d predictors), but instead has \
size %jd\n\n", bedfile, (off_t)sizeof(unsigned char)*rowlength*num_preds+sizeof(unsigned char)*3,\
num_samples, num_preds, ftello(input));exit(1);}
```

C code for reading SNPs from bedfile

```
//check the file has length (3+num_preds*rowlength) bytes
fseeko(input, 0, SEEK_END);
if(ftello(input)!=(off_t)sizeof(unsigned char)*rowlength*num_preds+sizeof(unsigned char)*3)
{printf("Error reading %s; should have size %jd (%d ind x %d predictors), but instead has size \
%jd\n\n", bedfile, (off_t)sizeof(unsigned char)*rowlength*num_preds+sizeof(unsigned char)*3, \
num_samples, num_preds, ftello(input));exit(1);}

//check the first three bytes have expected values
fseeko(input, 0, SEEK_SET);
if(fread(check, sizeof(unsigned char), 3, input)!=3)
if(check[0]!=108||check[1]!=27)
{printf("Error reading %s; does not appear to be a binary PLINK file\n\n", bedfile);exit(1);}
if(check[2]!=1)
{printf("Error reading %s; can only read in SNP-major mode\n\n", bedfile);exit(1);}

//read each SNP in turn
for(j=start;j<end;j++)
{
//move to the correct SNP in file
if(fseeko(input, (off_t)sizeof(unsigned char)*rowlength*j+sizeof(unsigned char)*3, SEEK_SET)!=0)
{printf("Error reading %s; unable to find Predictor %d\n\n", bedfile, j+1);exit(1);}

//read the SNPs values
if(fread(rowbytes, sizeof(unsigned char), rowlength, input)!=rowlength)
{printf("Error reading values for Predictor %d from %s\n\n", j+1, bedfile);exit(1);}
```

C code for reading SNPs from bedfile

```
//convert to basepairs - convert each byte to (up to) 4 SNPs
i=0;
for(count=0;count<rowlength;count++)
{
//gen contains all 4 SNPs; gen2 examines one SNP at a time
gen=(int)rowbits[count];
gen2=gen%4;
data[i+(j-start)*num_samples]=conv[gen2];i++;
if(i==num_samples){break;}
gen2=(gen>>2)%4;
data[i+(j-start)*num_samples]=conv[gen2];i++;
if(i==num_samples){break;}
gen2=(gen>>4)%4;
data[i+(j-start)*num_samples]=conv[gen2];i++;
if(i==num_samples){break;}
gen2=(gen>>6)%4;
data[i+(j-start)*num_samples]=conv[gen2];i++;
if(i==num_samples){break;}
}

}
fclose(input);

free(rowbytes);
return(0);
} //end of read_bed_fly
```

Phenotypes and Covariates

```
doug@doug-laptop:~/Desktop/workingdir/wtccc$ head phen.pheno
IND1 IND1 NA
IND2 IND2 1.544696
IND3 IND3 -0.219270
IND4 IND4 1.360046
IND5 IND5 -1.721269
IND6 IND6 2.212904
IND7 IND7 -1.026232
IND8 IND8 -1.306353
IND9 IND9 -0.929093
IND10 IND10 1.928810
```

Phenotype files have $2 + M$ columns, providing Family ID, Individual ID, then value for each of M phenotypes

when analysing, use --mpheno to specify the phenotype

Covariate files take same format (do not include covariate for mean)

PLINK commands

Consist of the command `plink` followed by at least 3 options specifying:

- the function to be executed,
- an input datafile,
- an output filestem.

Examples:

```
plink --make-bed --file data_ped --out data_bed
plink --hardy --bfile data_bed --out stats
plink --linear --bfile data_bed --out assoc
```

Note: options are preceded by two dashes

For a full list of options see

http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

Profile Scoring

```
plink --score <profile> --bfile <data> --out <output>
```

The profile has three columns: SNP Name, Test Allele, Effect. E.g.

```
SNPA    A     1.95
SNPB    C     2.04
SNPC    C    -0.98
SNPD    C    -0.24
```

Suppose Individual 1 has genotypes AA (2), GG (0), AC (1), 00 (NA)

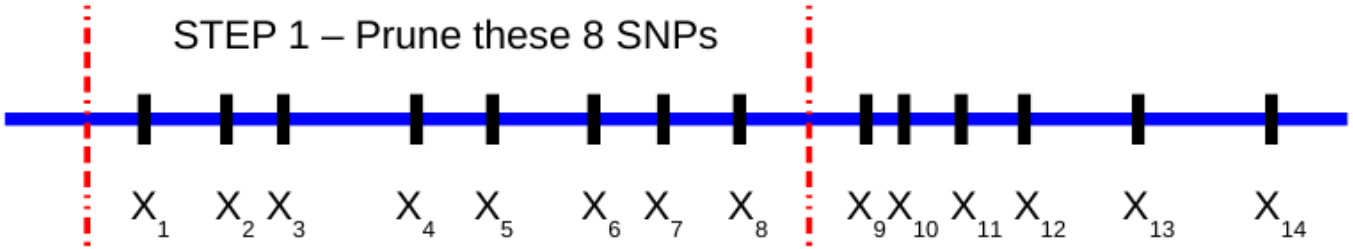
```
Variant(1/2)          A/T          C/G          A/C          C/G
Freq. of allele 1     0.20         0.43         0.02         0.38

Ind 1 genotype        A/A          G/G          A/C          0/0
# ref alleles          2            0            1          2*0.38 (=expectatio

Score           (  2*1.95    +   0*2.04  +  1*(-0.98) +  2*0.38*(-0.24) ) /
                =    2.74 / 4   =  0.68
```

SNP Pruning

```
plink --indep-pairwise <window> <step> <rsq> --bfile <data> --out <output>
```

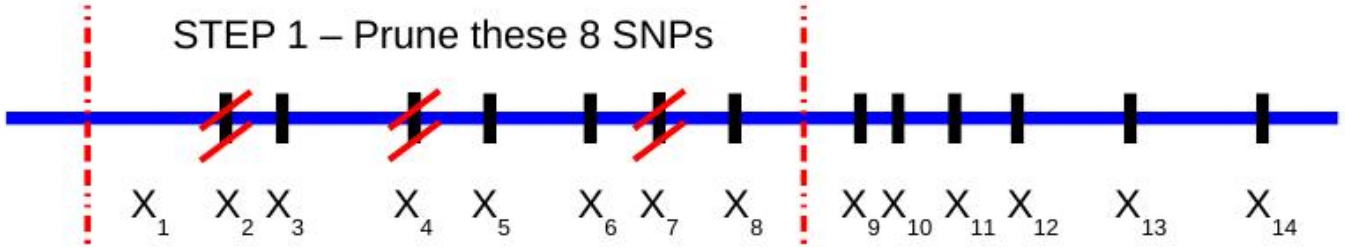
SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



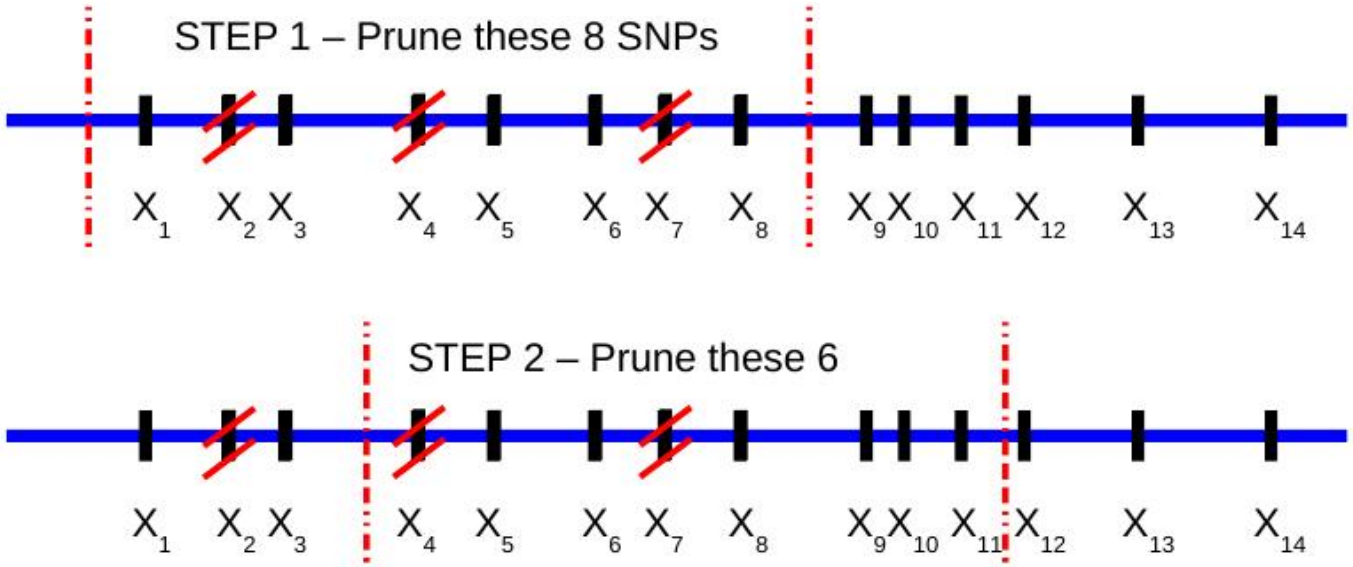
SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



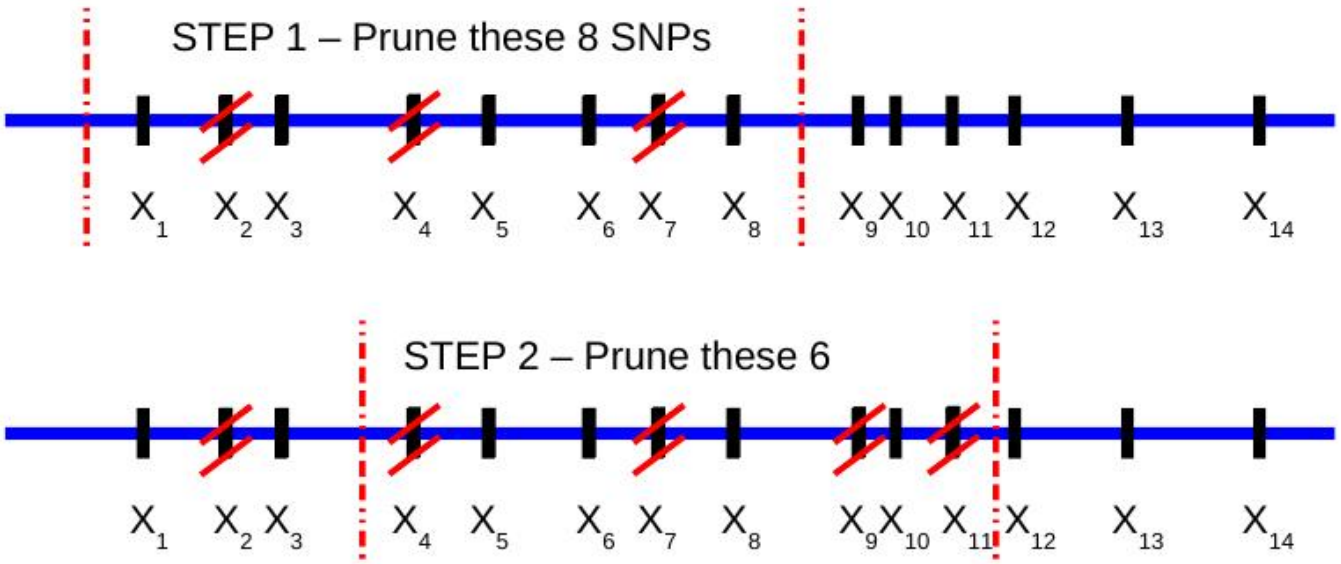
SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



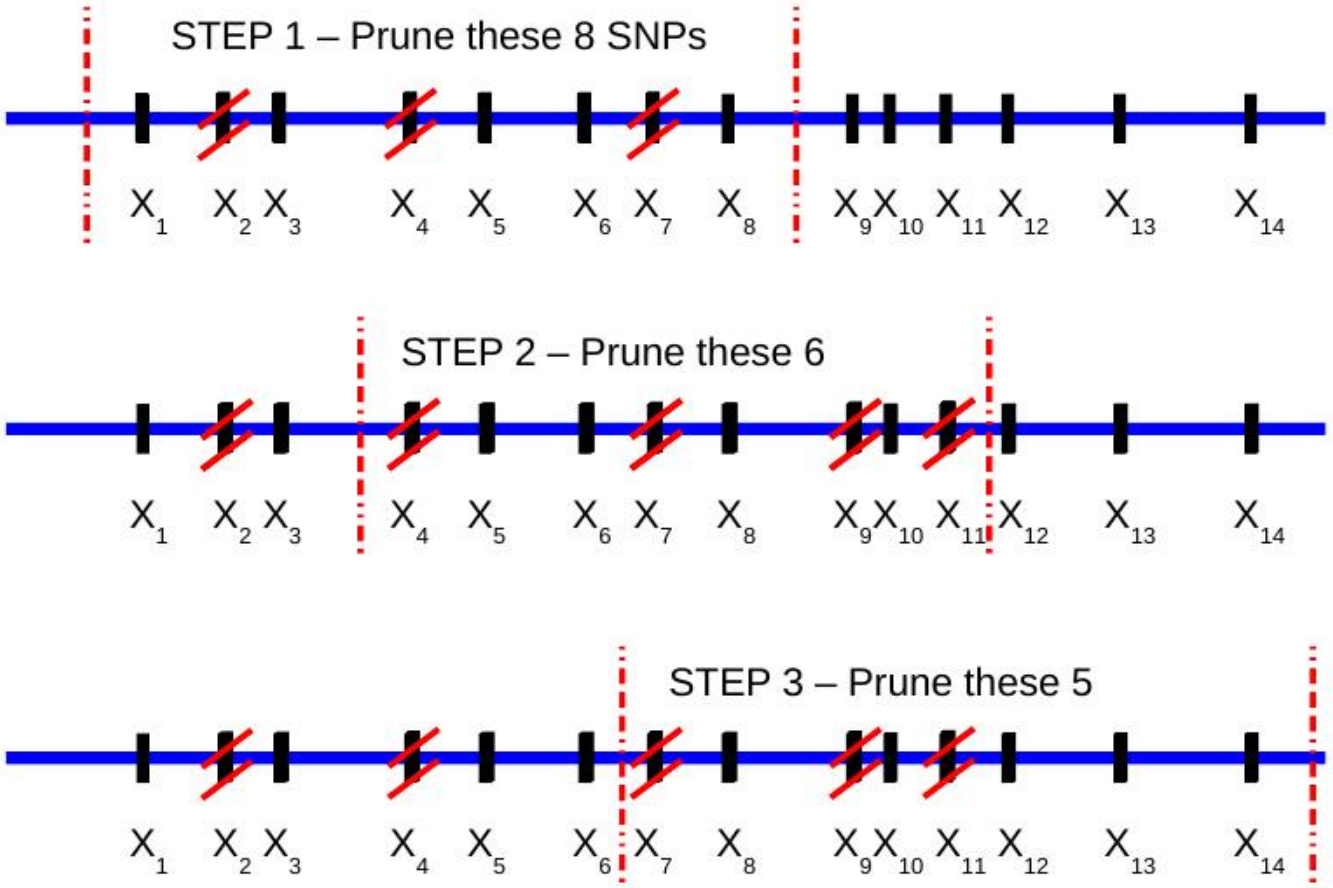
SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



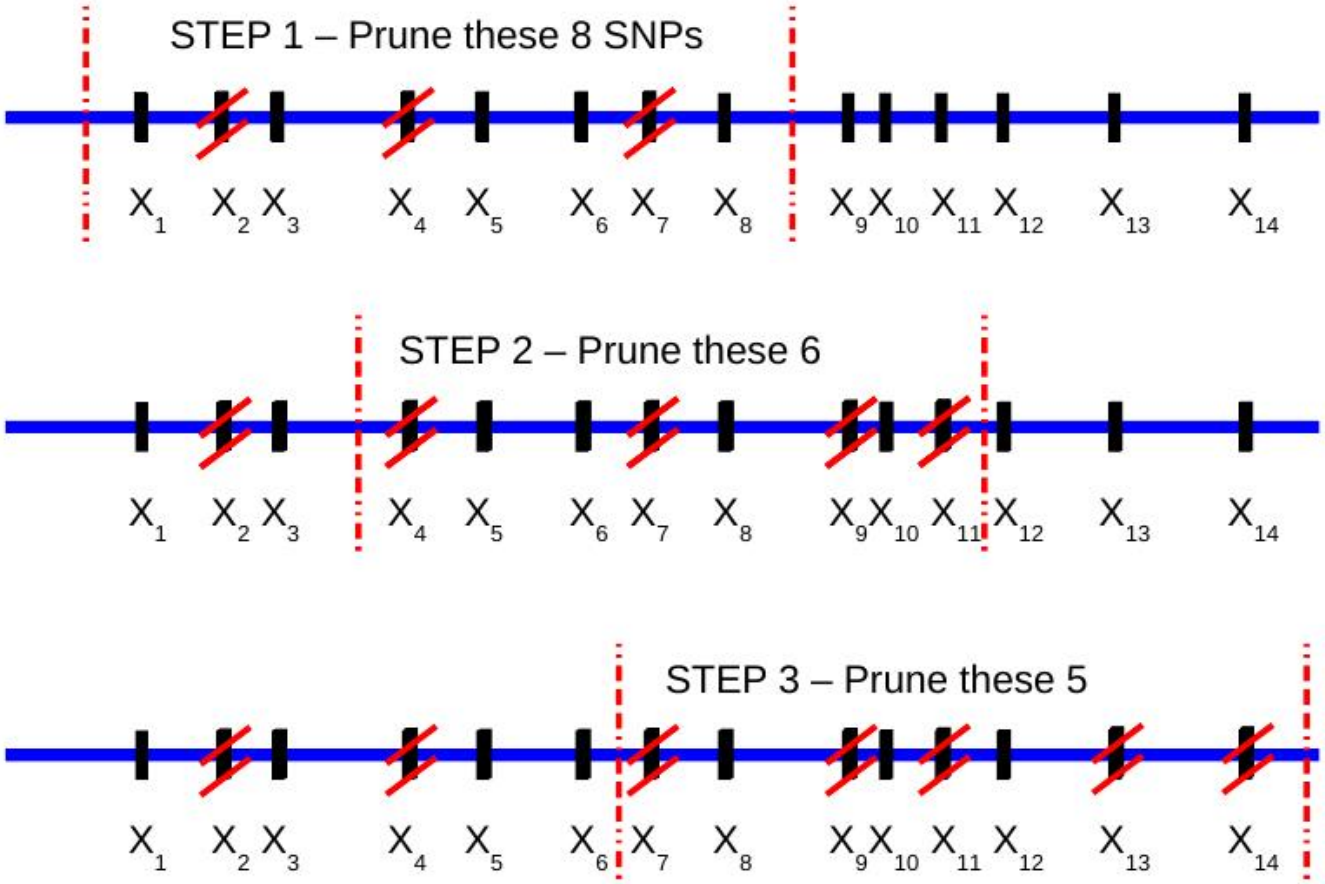
SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



SNP Pruning

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



Problems: assymetric; windows are fixed no. of SNP (not genetic distance)

Sources of bias in GWAS

Compared with other epidemiological studies, GWAS are subject to very few sources of bias or confounding

- genotypes rarely change in an individual's lifetime, and therefore there are few possibilities for a variable to affect both genotype and phenotype (i.e. confounding variables).

The two principle classes of confounders that can affect GWAS are:

1. Population structure and cryptic relatedness - to be discussed in Module 4 this afternoon.
2. Technical artefacts arising from sample handling or lab processing that are correlated with phenotype; confounding can arise in case-control studies if the case samples have been stored or analysed differently from the control samples.

Quality Control

Stringent quality control (QC) of both individuals and SNPs (possibly multiple rounds of each) can ensure high-quality genotypes and eliminate much of the effect, and statistical analyses can look for systematic differences between cases and controls that might reflect technical artefacts.

QC should be dataset-specific AND analysis-specific

In my opinion, QC should be agnostic of phenotype

It's usually best to err in the direction of being cautious: if we have a lot of data, it is better to lose some than to be misled by erroneous data

Can check impact of QC by repeating analyses with alternative thresholds

In some analyses, it is possible to flag QC outliers rather than remove them

e.g., in single-SNP association testing, can include dodgy SNPs

Quality Control

I tend to perform individual QC first, then SNP QC.

For individuals, common metrics are

- missing rate (--missing),
- heterozygosity (--het) and
- check of recorded with genotyped sex (--check-sex)

  I tend to compute these using pruned / decent-quality SNPs

For SNPs, common metrics are

- MAF (--freq),
- callrate (--missing) and
- Hardy-Weinberg Equilibrium (--hwe).

For binary traits, can also test for differential missingness (--test-missing)

Heterozygosity/HWE checks

Individuals: High homozygosity can reflect poor genotyping due to a poor-quality sample, or sample contamination generating additional variation. Extreme outliers for heterozygosity should be discarded.

SNPs: testing for HWE is routinely performed at each SNP because poor-quality genotyping can result in heterozygotes being called as homozygotes, generating more homozygotes than expected. However, deviation from HWE may also be due to processes related to disease, and in general it is best not to discard SNPs that are only mildly discordant with HWE, but only to flag them for extra checking if they do show association with phenotype. SNPs extremely discordant with HWE are usually removed: setting $P < 10^{-6}$ as the threshold implies one SNP per million will be removed when HWE holds.

Missingness and MAF

In most genotyping systems a quality score is assigned to the genotype calls, and below a threshold no call is made and the genotype is reported as "missing". Overall rates of missingness need to be interpreted with care:

- lower quality thresholds lead to a lower rate of missingness but can allow more doubtful calls to be included in the analyses;
- higher thresholds increase the rate of missingness but ensure higher quality of the analysed SNPs.

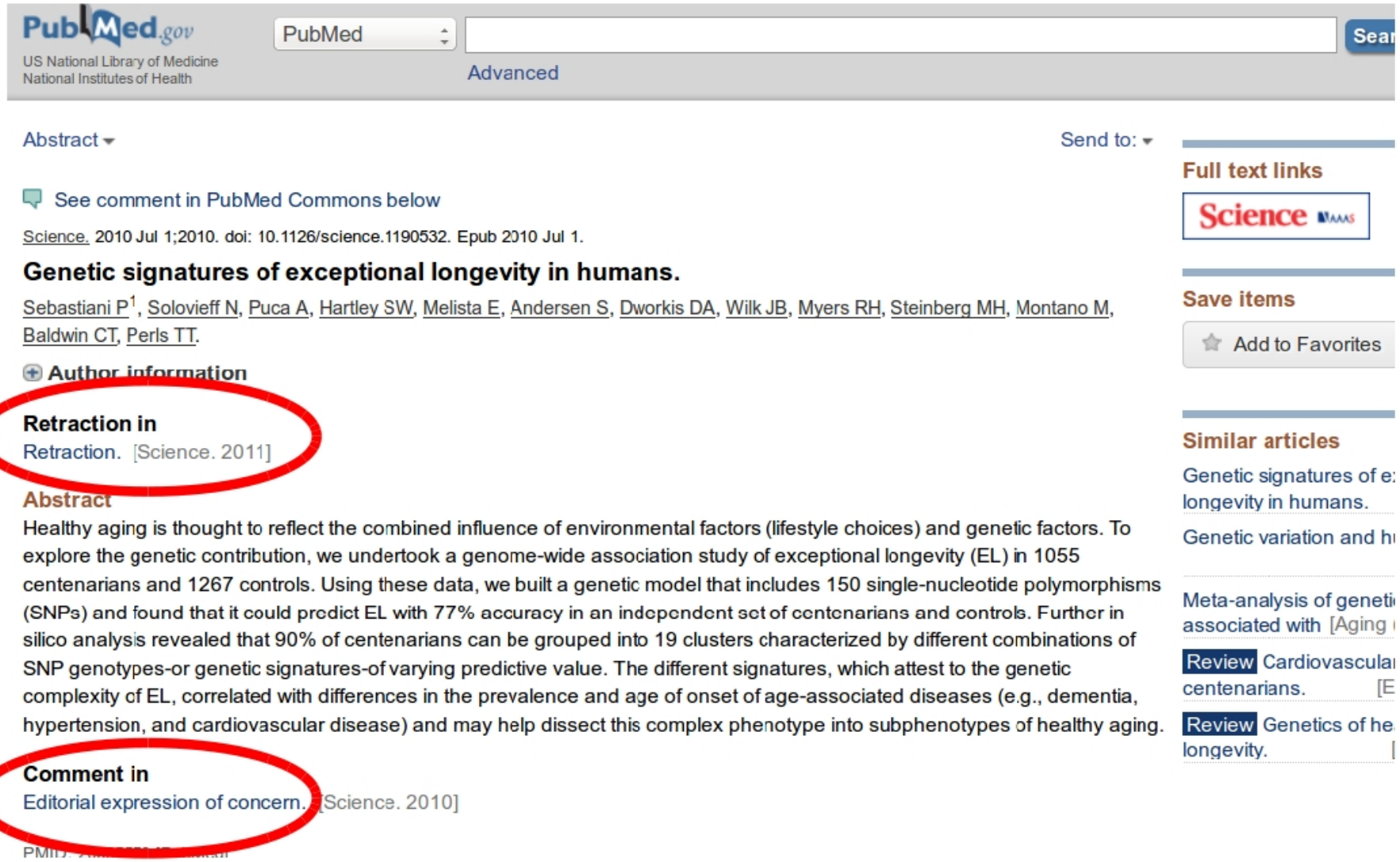
Extreme outliers for missingness in either SNPs or individuals should be investigated and/or discarded:

SNPs: excessive missingness can reflect a problem assay,

Individuals: excessive missingness can reflect poor sample quality.

Screening on **minor allele frequency (MAF)** removes SNPs below a threshold, often around 1%. Low MAF SNPs are more susceptible to genotyping errors, and also they have low power to detect any association for a given effect size.

A Cautionary Tale: Markers for Longevity



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Advanced | Sear

Abstract | Send to:

See comment in PubMed Commons below

Science. 2010 Jul 1;2010. doi: 10.1126/science.1190532. Epub 2010 Jul 1.

Genetic signatures of exceptional longevity in humans.

Sebastiani P[1], Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH, Montano M, Baldwin CT, Perls TT.

Author information

Retraction in
Retraction. [Science. 2011]

Abstract

Healthy aging is thought to reflect the combined influence of environmental factors (lifestyle choices) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity (EL) in 1055 centenarians and 1267 controls. Using these data, we built a genetic model that includes 150 single-nucleotide polymorphisms (SNPs) and found that it could predict EL with 77% accuracy in an independent set of centenarians and controls. Further in silico analysis revealed that 90% of centenarians can be grouped into 19 clusters characterized by different combinations of SNP genotypes-or genetic signatures-of varying predictive value. The different signatures, which attest to the genetic complexity of EL, correlated with differences in the prevalence and age of onset of age-associated diseases (e.g., dementia, hypertension, and cardiovascular disease) and may help dissect this complex phenotype into subphenotypes of healthy aging.

Comment in
Editorial expression of concern. [Science. 2010]

PMID:

Full text links
Science

Save items
Add to Favorites

Similar articles
Genetic signatures of e: longevity in humans.
Genetic variation and h
Meta-analysis of genetic associated with [Aging
Review Cardiovascular centenarians. [E
Review Genetics of he longevity. [

http://classic.sciencemag.org/content/early/2011/07/20/science.1190532.long

A Cautionary Tale: Markers for Longevity

Science | Science Translational Medicine | Science Signaling | Science Advances

SHARE

LETTERS

Retraction

Paola Sebastiani[1,*], Nadia Solovieff[1], Annibale Puca[2], Stephen W. Hartley[1], Efthymia Melista[3], Stacy Andersen[4], Daniel A. Dworkis[3], Jemma B. Wilk[5], Richard H. Myers[5], Martin H. Steinberg[6], Monty Montano[3], Clinton T. Baldwin[6,7], Thomas T. Perls[4,*]

+ Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.T.P.)

Science 22 Jul 2011:
Vol. 333, Issue 6041, pp. 404
DOI: 10.1126/science.333.6041.404-a

Article | Info & Metrics | eLetters | PDF

After online publication of our Report "Genetic signatures of exceptional longevity in humans" (*1*), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated

A Cautionary Tale: Markers for Longevity

nature International weekly journal of science

Login

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject
- Genetics
- Health and medicine
- Lab life

Stories by keywords
- Longevity
- Retraction
- Genome-wide association studies
- GWAS

This article elsewhere
- Blogs linking to this article
- Add to Digg
- Add to Facebook
- Add to Newsvine
- Add to Del.icio.us
- Add to Twitter

Published online 21 July 2011 | Nature | doi:10.1038/news.2011.429

News

Paper on genetics of longevity retracted

Technical problems mar study of centenarians.

Heidi Ledford

A prominent paper that claimed to reveal the genetic factors that help people live to 100 or older has been retracted[1], a year after it was first released.

Whether someone lives to their 100th birthday is determined by more complicated genetics than the retracted study suggested.

Tom Benitez/ MCT via Getty Images

The study, published in *Science*[2], reported 150 genetic variations that could be used to predict whether a person was genetically inclined to see their 100th birthday. The results were based on a search through the genomes of more than 1,000 centenarians.

But shortly after the paper was published, a host of criticisms arose. In particular, geneticists noted that the control samples and the samples from centenarians were analysed in slightly different ways. Last November, *Science* editor Bruce Alberts published an editorial expression of concern[3] and noted that the authors were working to address the issue.

Related stories
- Dwarfism may stymie diseases of old age 16 February 2011
- Ageing cells lose protein pumps 25 July 2010
- Genetic variations offer longer life 01 July 2010
- A pill for longer life? 08 July 2009

Naturejobs

Research and Teaching Positions Available in Estuarine, Coastal and Marine Research at the East China Normal University, Shanghai, China
State Key Laboratory of Estuarine and Coastal Research (SKLEC), East China Normal University

Research and Teaching Positions Available in Estuarine, Coastal and Marine Research at the East China Normal University, Shanghai, China
State Key Laboratory of Estuarine and Coastal Research (SKLEC), East China Normal University

More science jobs

Post a job

Resources

Send to a Friend

Reprints & Permissions

A Cautionary Tale: Markers for Longevity

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani, Nadia Solovieff, Andrew T. DeWan, Kyle M. Walsh, Annibale Puca, Stephen W. Hartley, Efthymia Melista, Stacy Andersen, Daniel A. Dworkis, Jemma B. Wilk, Richard H. Myers, Martin H. Steinberg, Monty Montano, [...], Thomas T. Perls

[view all]

Published: January 18, 2012 • DOI: 10.1371/journal.pone.0029848

Article metrics are unavailable for recently published articles.

Article | About the Authors | Metrics | Comments | Related Content

Download PDF | Print | Share

Abstract

Like most complex phenotypes, exceptional longevity is thought to reflect a combined influence of environmental (e.g., lifestyle choices, where we live) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity in 801 centenarians (median age at death 104 years) and 914 genetically matched healthy controls. Using these data, we built a genetic model that includes 281 single nucleotide polymorphisms (SNPs) and discriminated between cases and controls of the discovery set with 89% sensitivity and specificity, and with 58% specificity and 60% sensitivity in an independent cohort of 341 controls and 253 genetically matched nonagenarians and centenarians (median age 100 years). Consistent with the hypothesis that the genetic contribution is largest with the oldest ages, the sensitivity of the model increased in the independent cohort with older and older ages (71% to classify subjects with an age at death>102 and 85% to classify subjects with an age at death>105). For further validation, we applied the model to an additional, unmatched 60 centenarians (median age 107 years) resulting in 78% sensitivity, and 2863 unmatched controls with 61% specificity. The 281 SNPs include the SNP rs2075650 in *TOMM40/APOE* that reached irrefutable genome wide significance (posterior probability of association=1) and replicated in the independent cohort. Removal of this SNP from the model reduced the accuracy by only 1%. Further in-silico analysis suggests that 90% of centenarians can be grouped into clusters characterized by different "genetic signatures" of varying predictive values for exceptional longevity. The correlation between 3 signatures and 3 different life spans was replicated in the combined replication sets. The different signatures may help dissect this complex phenotype into sub-phenotypes of exceptional longevity.

CrossMark

Subject Areas

- Genetics of disease
- Variant genotypes
- Genome-wide assoc...
- Alleles
- Alzheimer disease
- Genotyping
- Population genetics
- Aging

A Cautionary Tale: Markers for Longevity

Scholar

About 797 results (**0.04** sec)

Articles

Case law

My library

Any time
Since 2016
Since 2015
Since 2012
Custom range...

2010 —

Search

Sort by relevance

User profiles for **paola sebastiani**

Paola Sebastiani
Boston University
Verified email at bu.edu
Cited by 6605

[HTML] Genetic signatures of exceptional longevity in humans

P **Sebastiani**, N Solovieff, AT DeWan, KM Walsh... - PloS one, 2012 - dx.plos.org

Abstract Like most complex phenotypes, exceptional longevity is thought to reflect a combined influence of environmental (eg, lifestyle choices, where we live) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association ...

Cited by 167 Related articles All 32 versions Web of Science: 79 Cite Save More

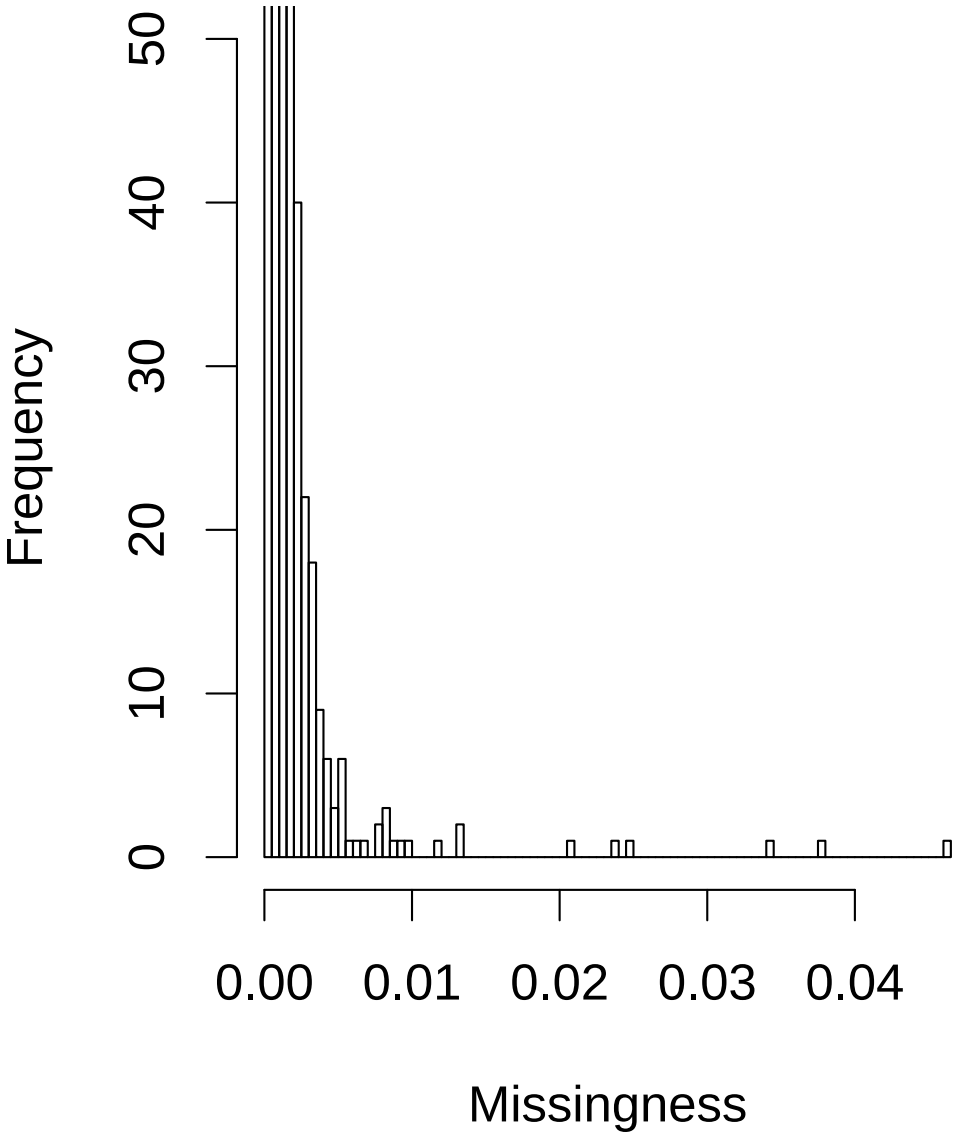
[HTML] Genetic signatures of exceptional longevity in humans

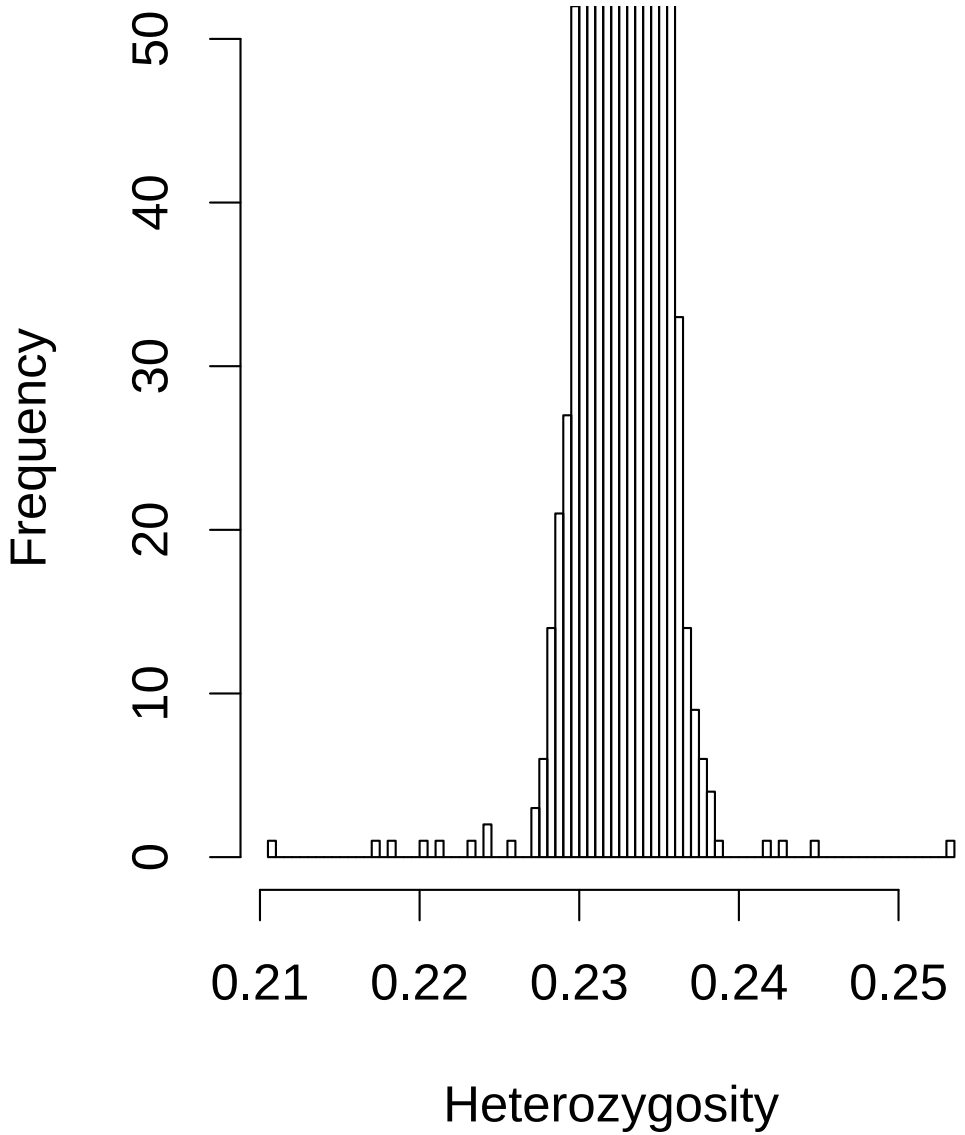
P **Sebastiani**, N Solovieff, A Puca, SW Hartley... - Science, 2010 - sciencemag.org

Abstract Healthy aging is thought to reflect the combined influence of environmental factors (lifestyle choices) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity (EL) in 1055 centenarians and ...

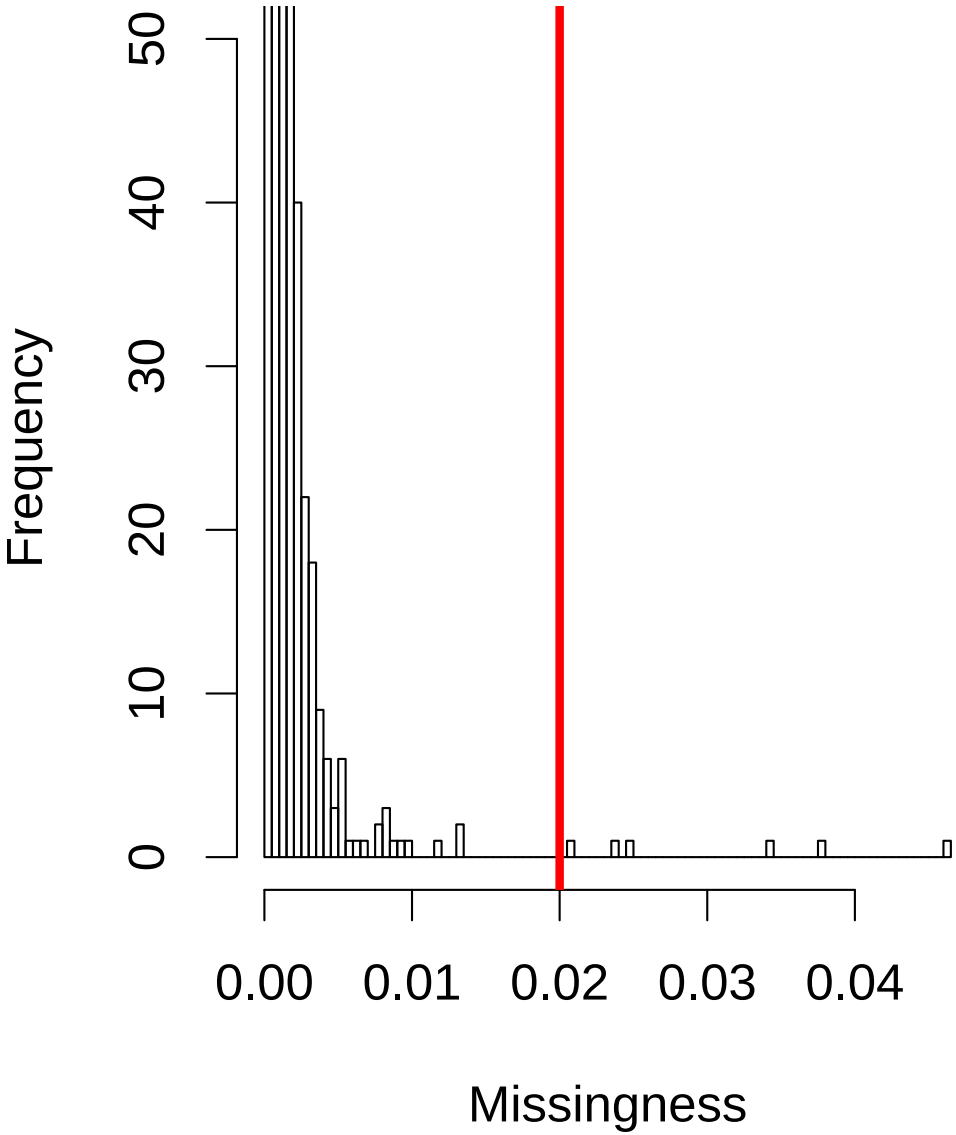
Cited by 108 Related articles All 11 versions Web of Science: 14 Cite Save

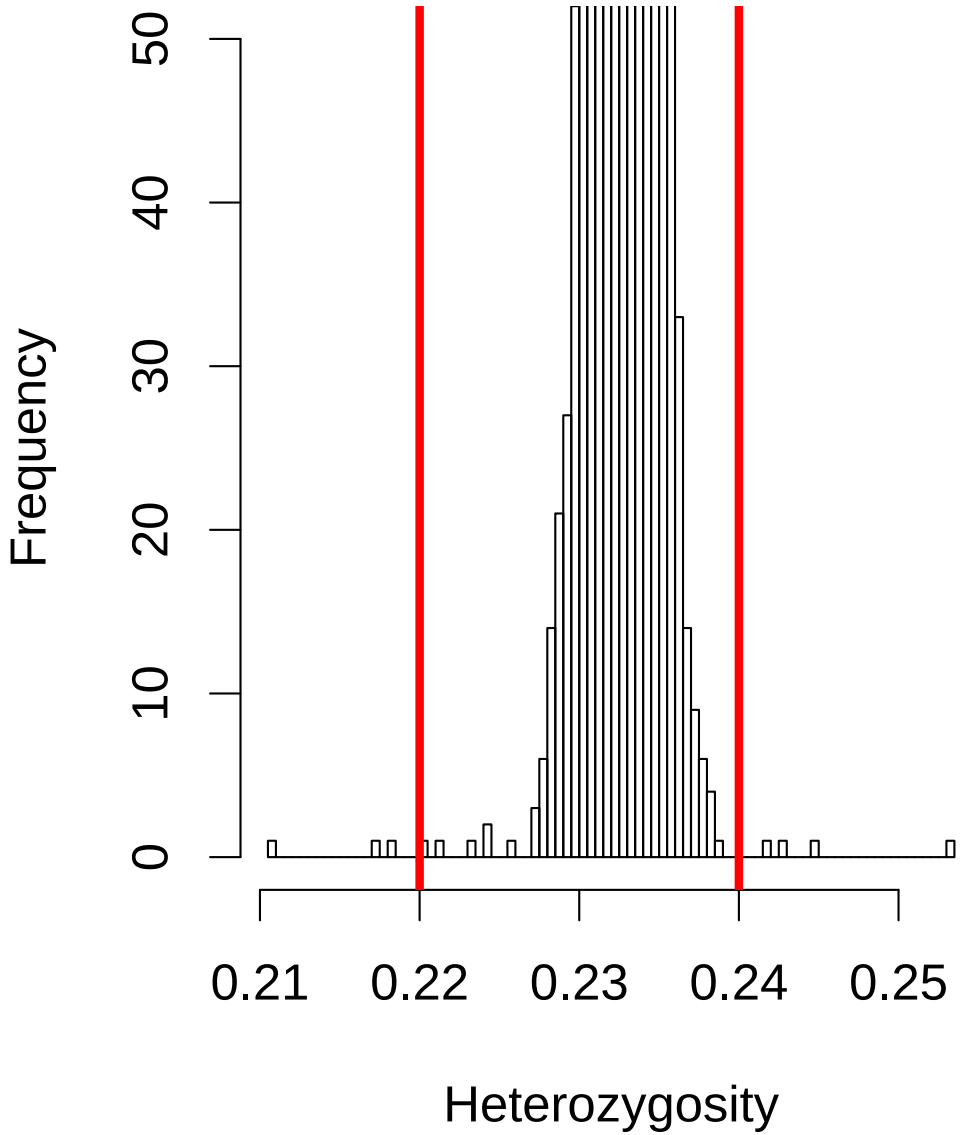
Individual QC: where to draw the line?





Individual QC: where to draw the line?





Individual QC

No hard rules - can try and model distribution

- Watch out for merged datasets

E.g., missingness $\sim$ Binomial(n,p) or heterozygosity $\sim$ Normal(μ,σ^2)

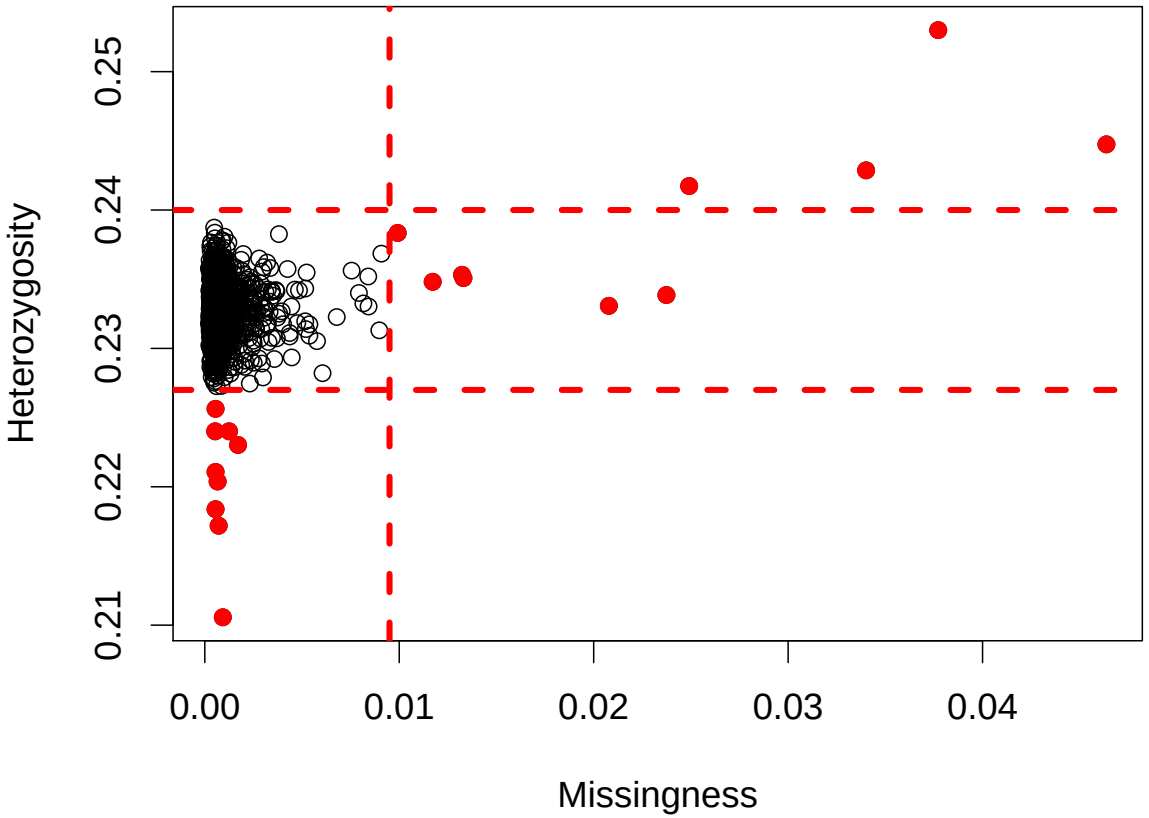
- estimate parameters
- compute p-value
- exclude individuals with (say) $P < 0.0001$

For more ideas, see Mike Weale's Chapter on Quality Control:

link.springer.com/content/pdf/10.1007%2F978-1-60327-367-1_19.pdf

Individual QC

I decide by eye:

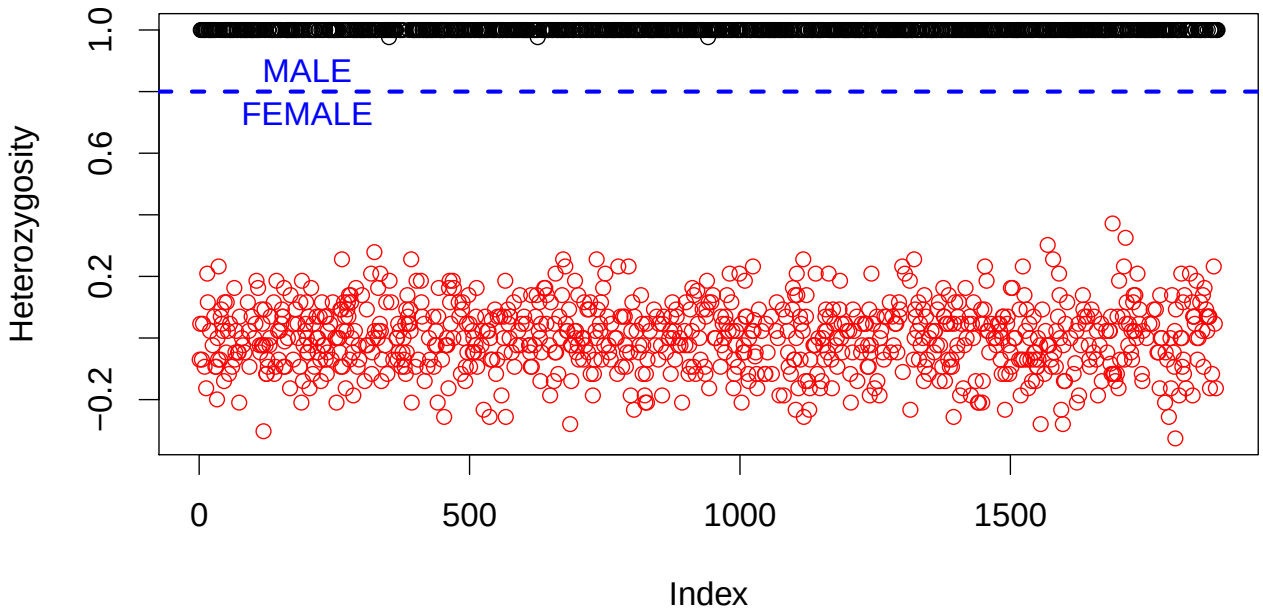


On a 2-D plot, can also exclude based on diagonal lines / ellipses

I DO NOT remove population outliers at this point

Individual QC

```
plink --check-sex --bfile <data> --out <output>
```

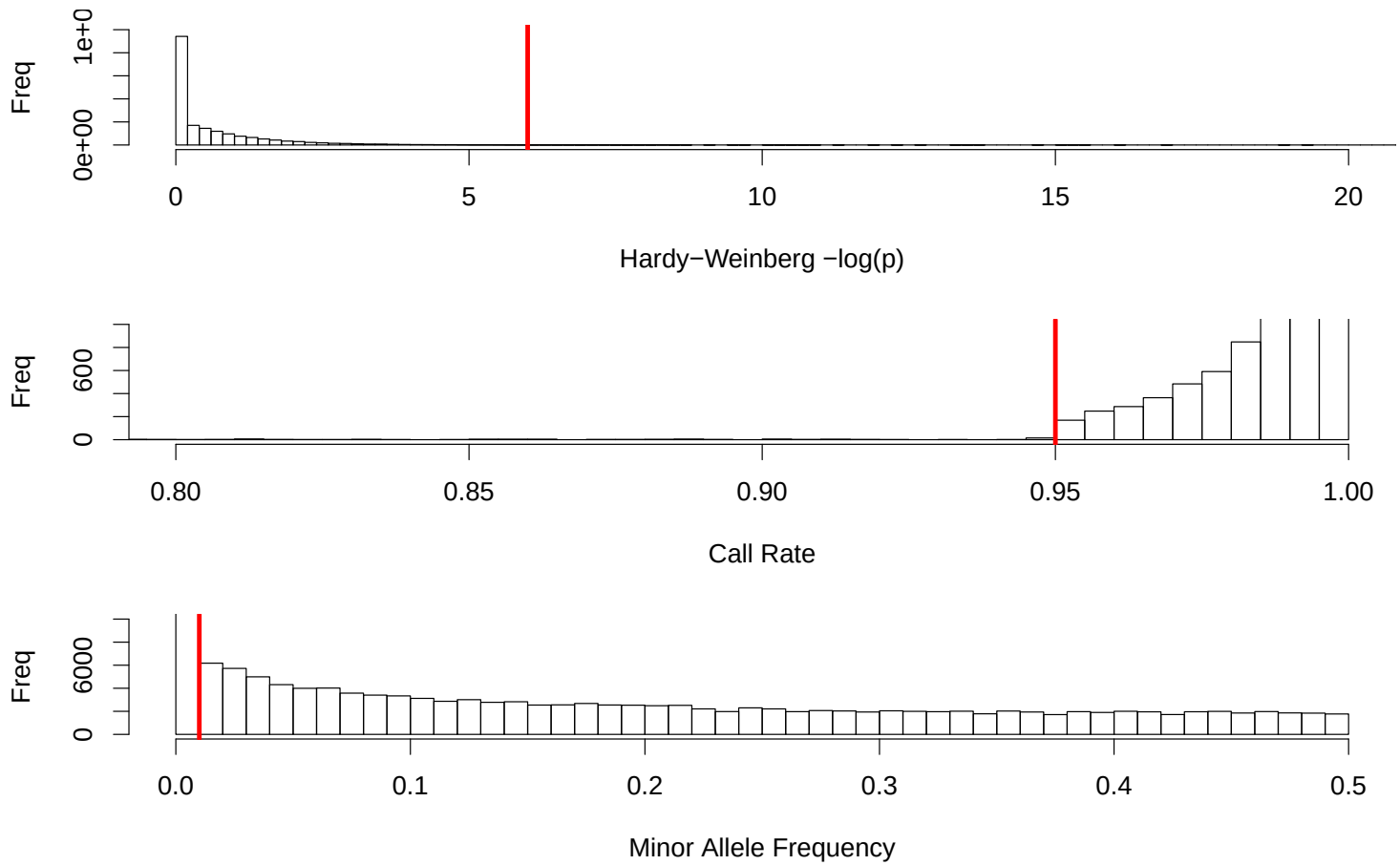


y-axis provides (estimate of) heterozygosity across X chromosome. Females (XX) will have (expected) heterozygosity 0, males (XY) 1

I use a threshold of 0.8. `PLINK` stresses importance of pruning

I DO NOT exclude sex miss-matches

SNP QC



Typical thresholds: HWE $P > 10^{-6}$, CR$>$ 0.95, MAF$>$ 0.01

At this point I remove population outliers (see later module)