H3ABioNet

Pan African Bioinformatics Network for H3Africa

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

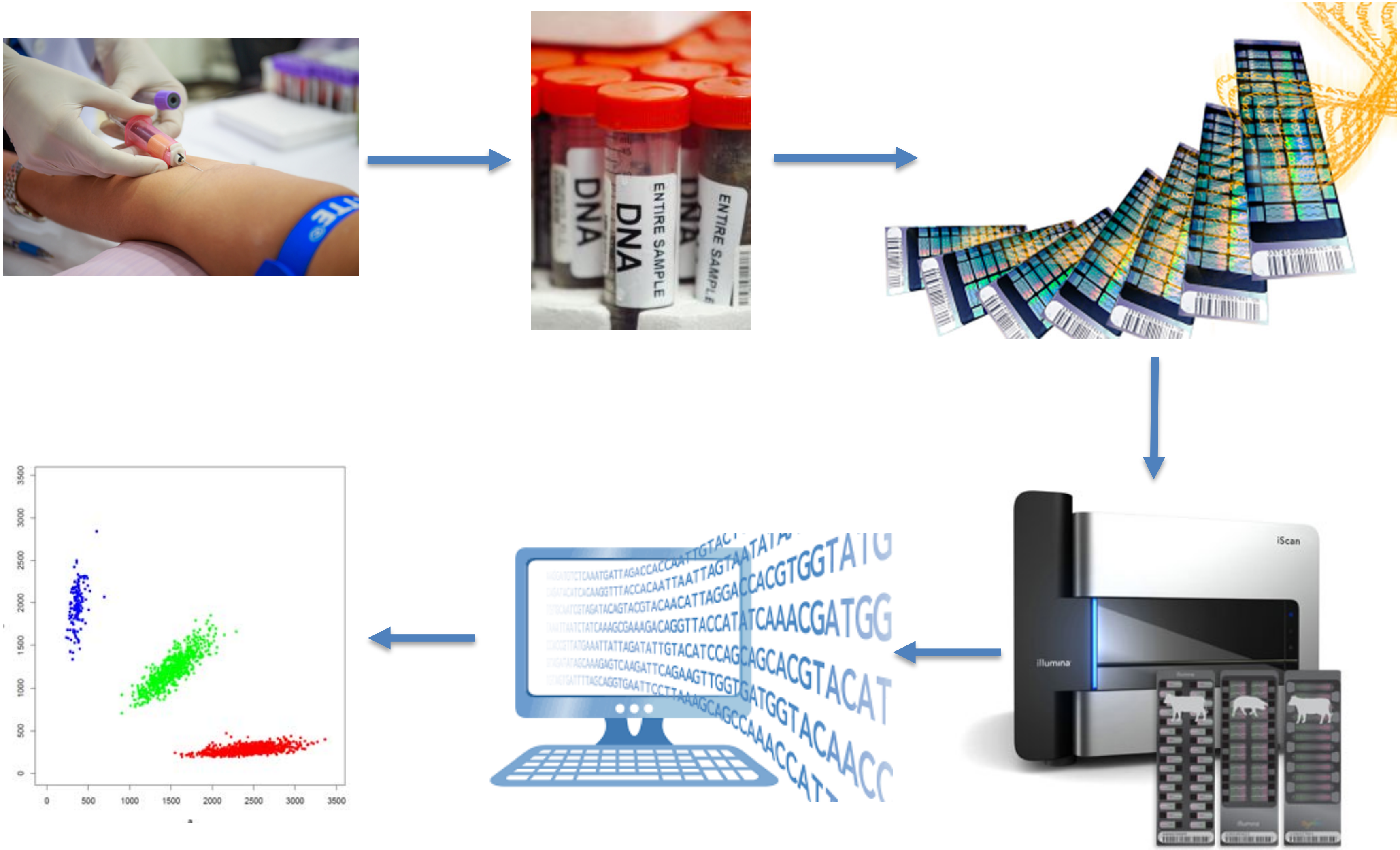
GWAS QC -theory and steps

H3ABioNet Genome Wide Association Studies Lecture Series

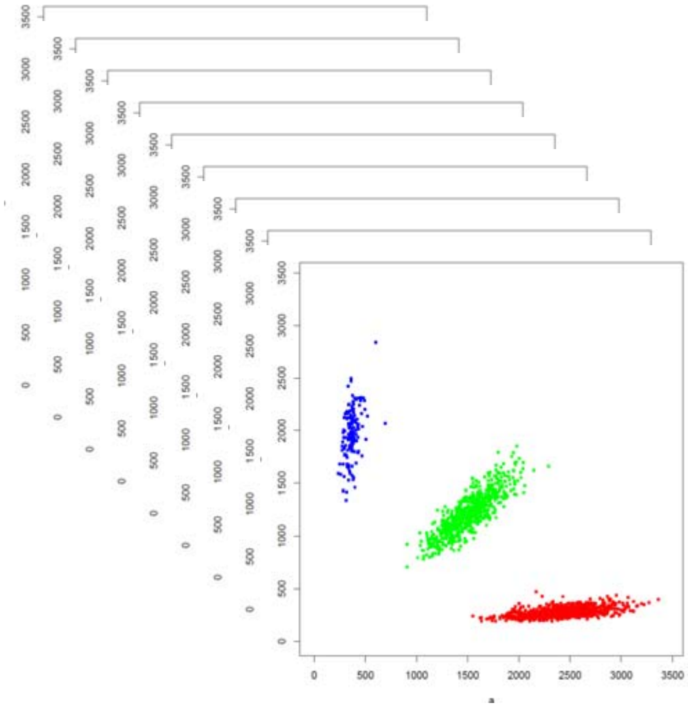
Shaun Aron

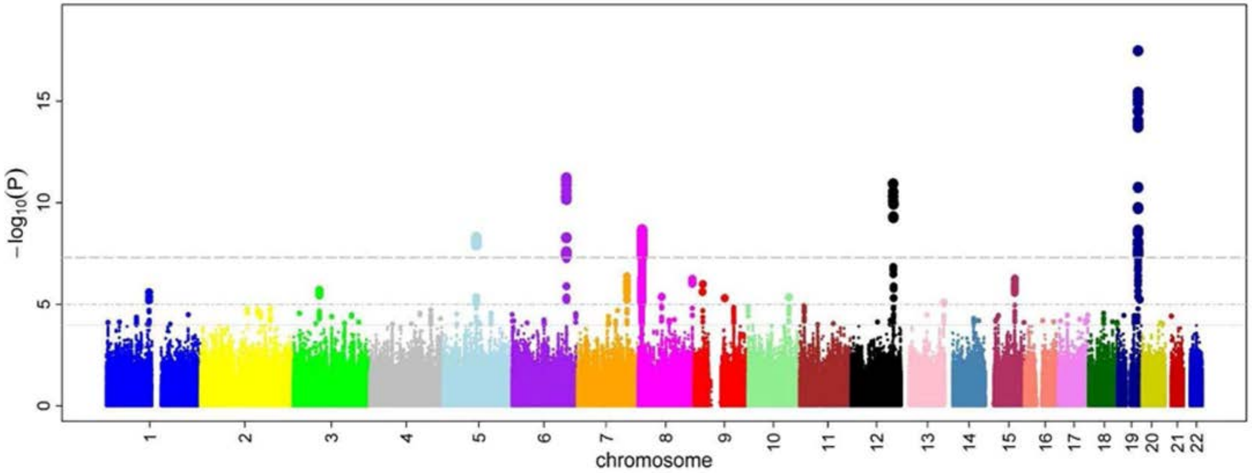
August 2018

GWAS Data Generation



GWAS Data Analysis

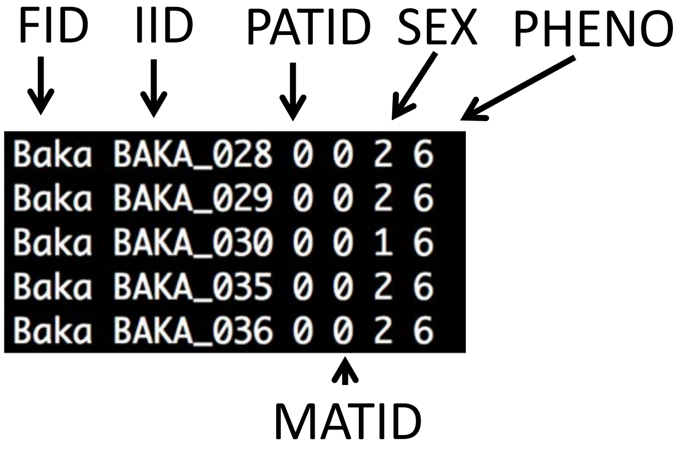




Plink format

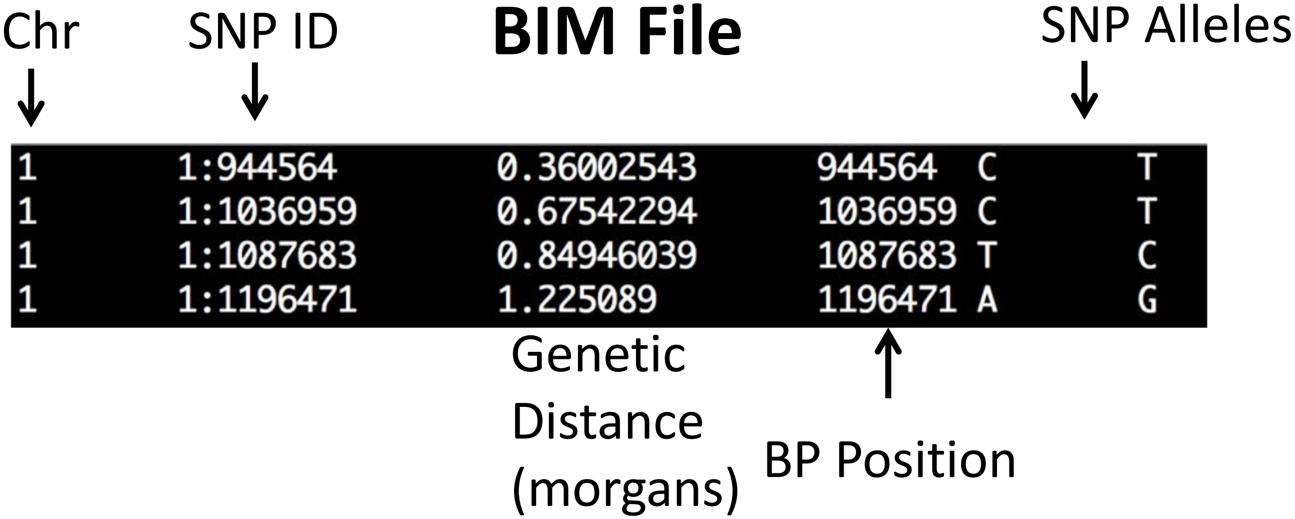
- FAM file – one row per individual
- BIM file – one row per SNP
- BED file – one row per individual – genotype calls for each individual for all SNPs – binary format
- FAM and BIM file are human readable while BED file in not

FAM File



```
Baka BAKA_028 0 0 2 6
Baka BAKA_029 0 0 2 6
Baka BAKA_030 0 0 1 6
Baka BAKA_035 0 0 2 6
Baka BAKA_036 0 0 2 6
```

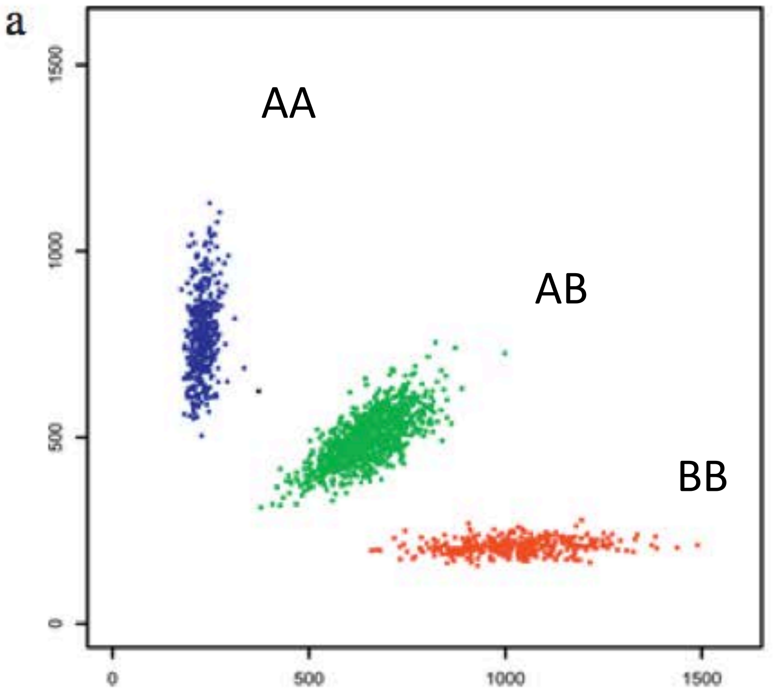
BIM File



```
1         1:944564         0.36002543       944564  C         T
1         1:1036959        0.67542294       1036959 C         T
1         1:1087683        0.84946039       1087683 T         C
1         1:1196471        1.225089         1196471 A         G
```

Why Do We Need Quality Control?

SNP 1



In an ideal world...

our sampling practices would be perfect,

our experiments would run perfectly,

and all our SNP genotypes would look like this.

Why Do We Need Quality Control

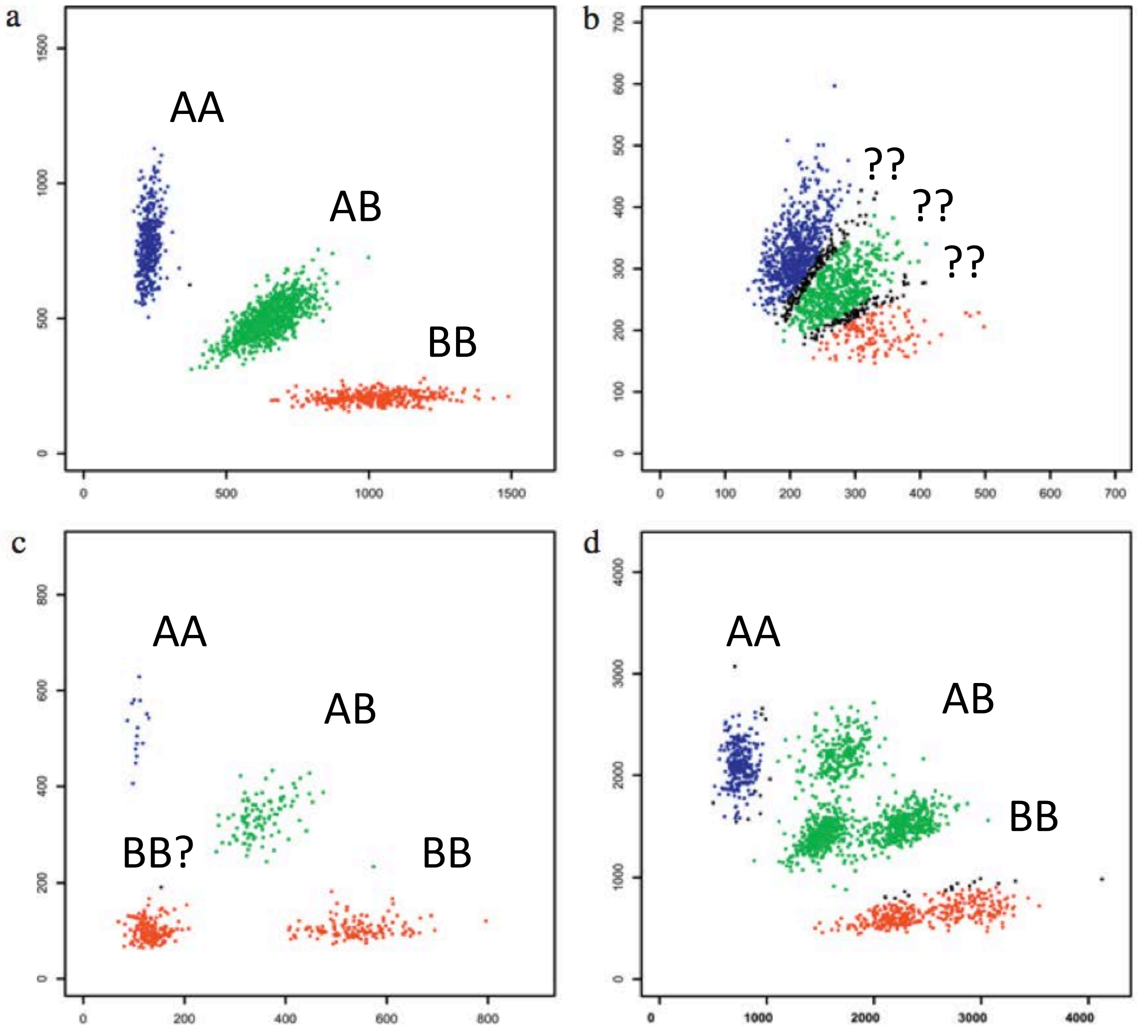
- Large scale experiments generate true results with a certain error rate

- Errors might originate at various steps in in the process:

  - ✓ Sample selection related issues
    - ✓ Cryptic relatedness
    - ✓ Population structure
  - ✓ Sample handling related issues
    - ✓ Labeling/Plating Error
  - ✓ Genotyping array related issues
    - ✓ Genotyping error
  - ✓ Batch effect related issues
    - ✓ Difference in results due to difference in sample processing

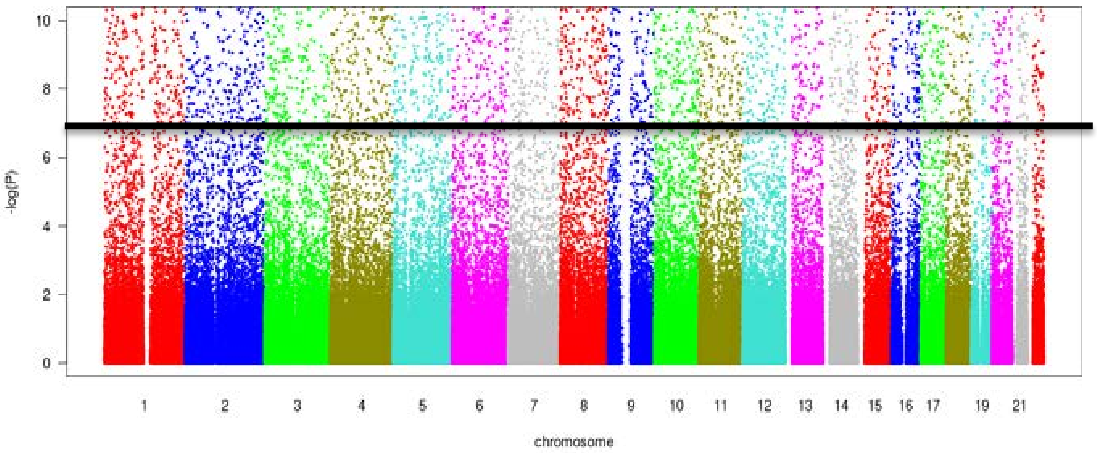
- Not practical to visually assess the genotype plot for every SNP
- Use some biologically relevant metrics as a proxy for quality
- Steps
  - – QC by SNP
  - – QC by sample

Why Do We Need Quality Control

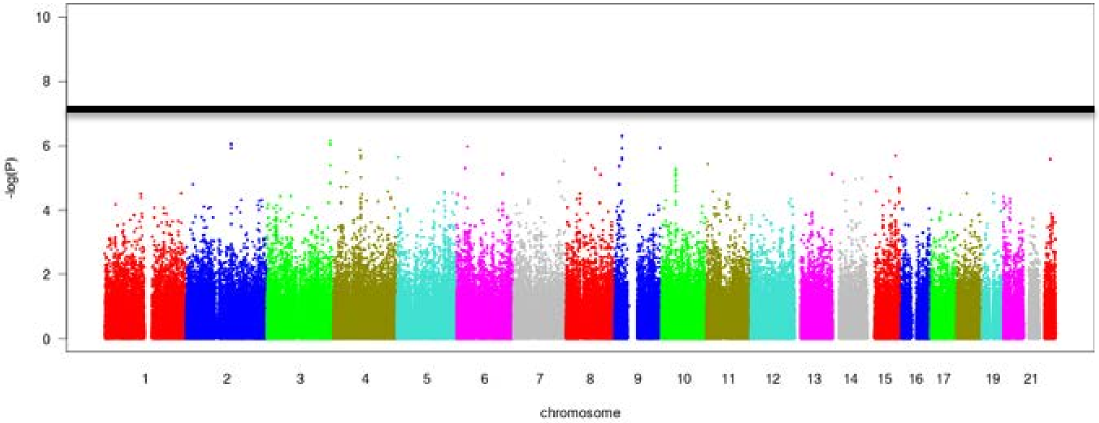


We don’t live in an ideal world…

German MI family study Affymetrix 500K Array Set
SNPs on chips: 493,840



SNPs passing QC: 270,701



Samani et al. 2007 N Engl J Med 357:443-53

QC Roadmap

Sample QC

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

SNP QC

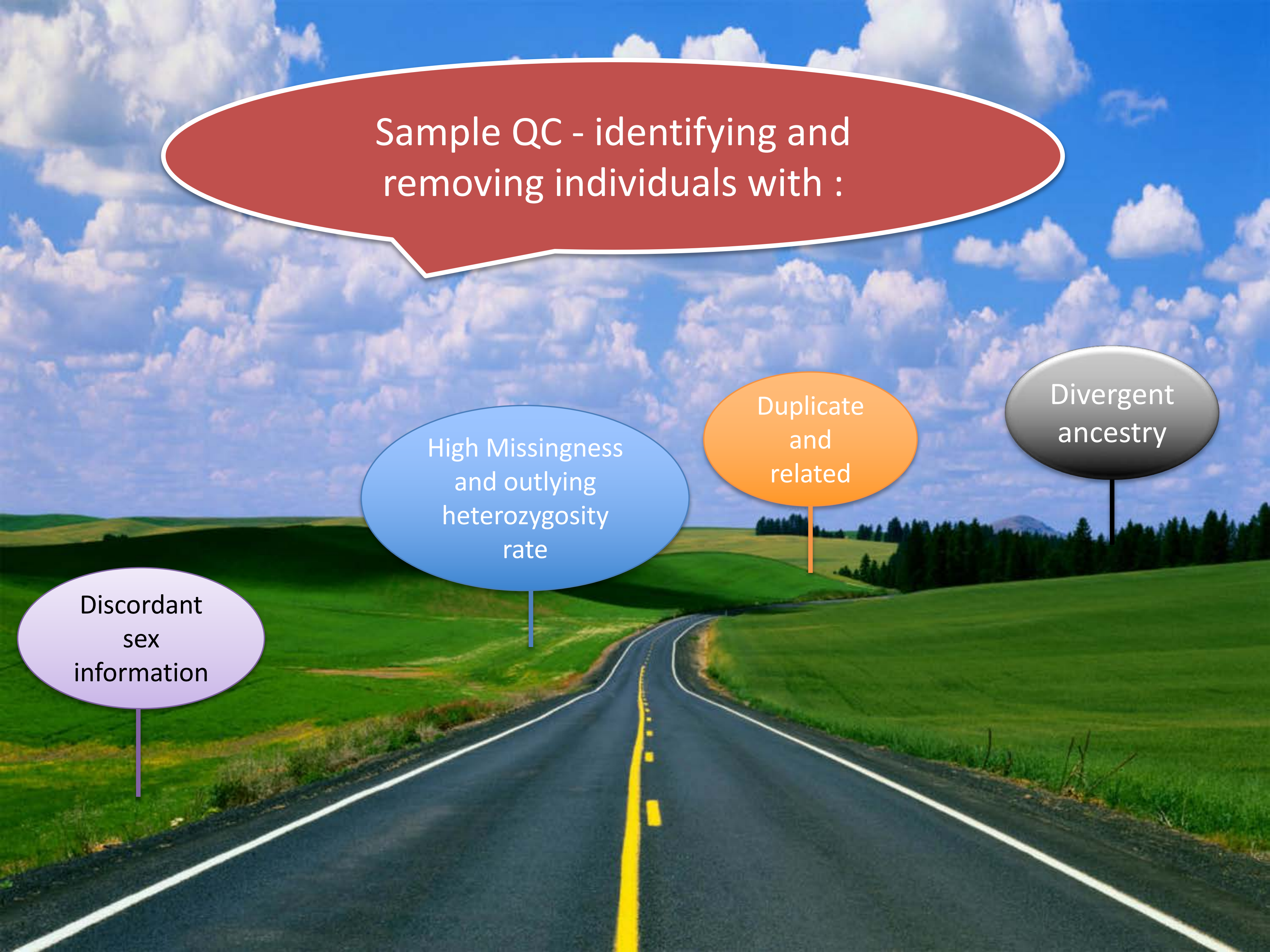
- Low minor allele frequency
- Missingness
- Differential missingness
- Hardy-Weinberg outliers

- For the sake of simplicity I have split the QC into Sample QC and SNP QC
- The order explained here is therefore not always the best way to do they QC
- The H3ABioNet pipeline has been developed using the best generalized order

Software

- **Programs required for QC**
  - PLINK 1.9 (Purcell, 2007)
  - Scripts for processing result files
  - R (Statistical Software) for plotting results
  - The complete process has been built into a pipeline using NextFlow

Sample QC - identifying and removing individuals with :



Discordant sex information

High Missingness and outlying heterozygosity rate

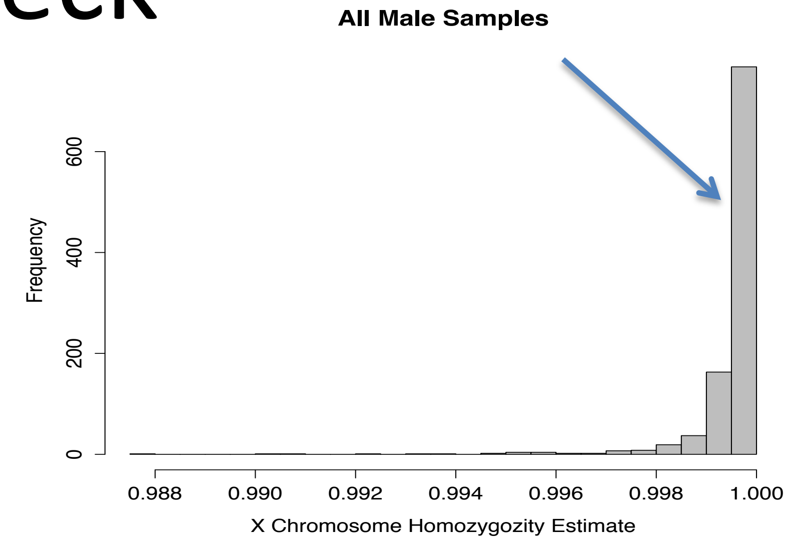
Duplicate and related

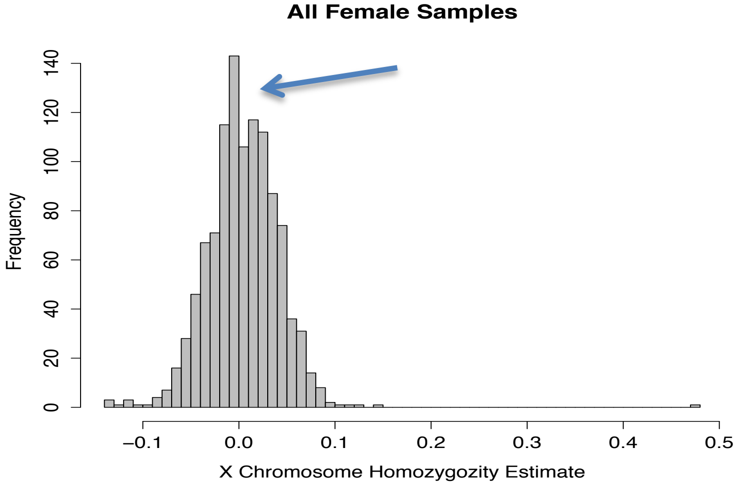
Divergent ancestry

Discordant Sex Check

- **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.





Identify individuals with discordant sex information

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

Creates a file named sexstat.sexcheck

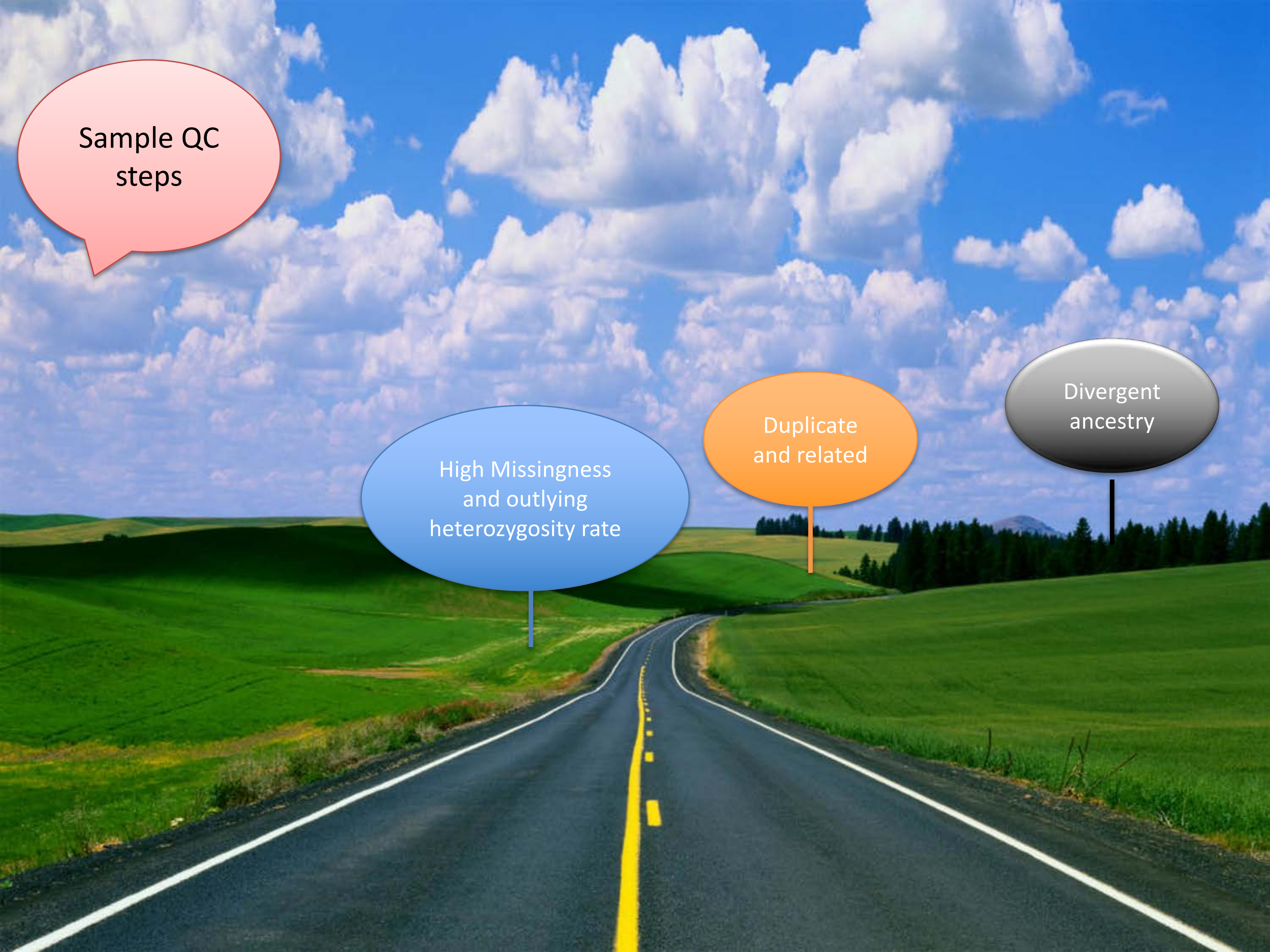
```
FID       IID        PEDSEX        SNPSEX        STATUS          F
   P554      P554              2             2            OK    -0.02654
   P555      P555              1             0       PROBLEM      0.5685
   P557      P557              2             2            OK      0.1264
   P558      P558              2             2            OK  -0.0007684
```

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.txt"

```
grep “PROBLEM” sexstat.sexcheck > fail_sex_check.txt
```

nextflow Parameters

- `sex_info_available = true`
- `f_low_male = 0.8`
- `f_high_female = 0.2`



Sample QC steps

High Missingness and outlying heterozygosity rate

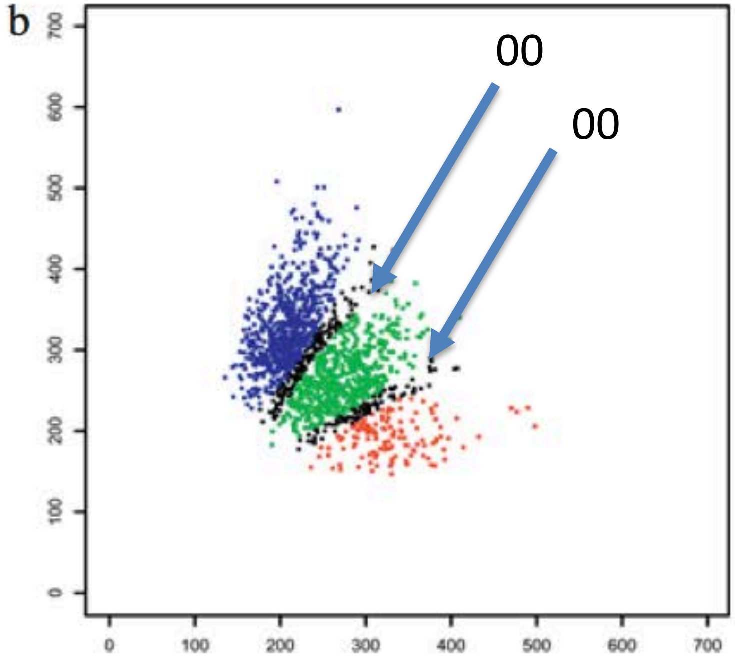
Duplicate and related

Divergent ancestry

Missingness

- Unable to call a genotype for a particular SNP – will be called as missing
- Per sample missingness
  - Percentage of SNPs with missing data per sample
- Per SNP missingness
  - Percentage of missing calls for a SNP

SNP 1



Missingness and Heterozygosity Rate

Genotyping call rate

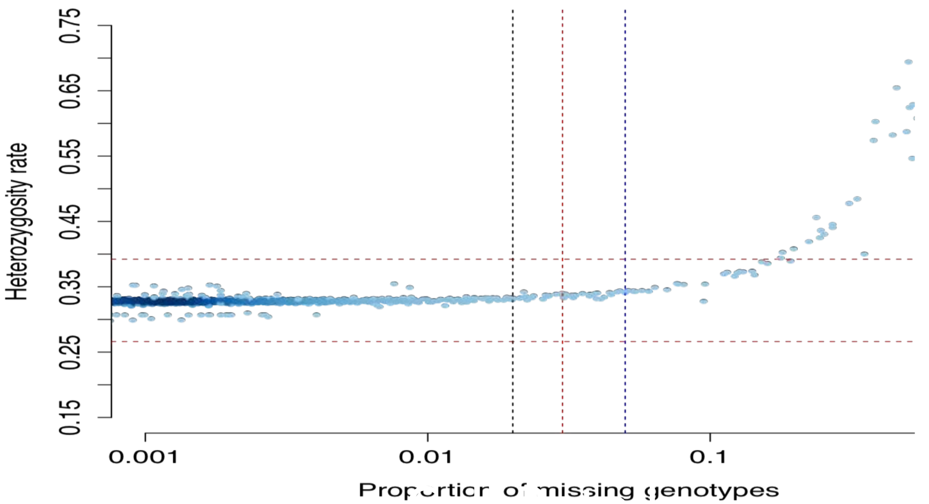
- Per sample (individual) rate
- Number **of non-missing genotypes divided by the total number of genotyped markers.**
- Low genotyping call rate indicate problem with **sample DNA** like low concentration.
- Thresholds used generally vary **between 3% and 7%**

Genotyping call rate and heterozygosity rate are generally plotted together.

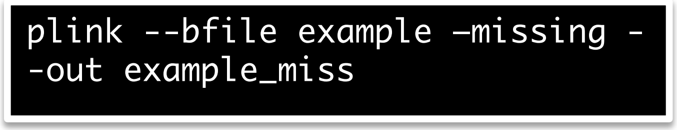
Cutoffs are selected so as to identify outlier individuals based on both the statistics

Heterozygosity Rate

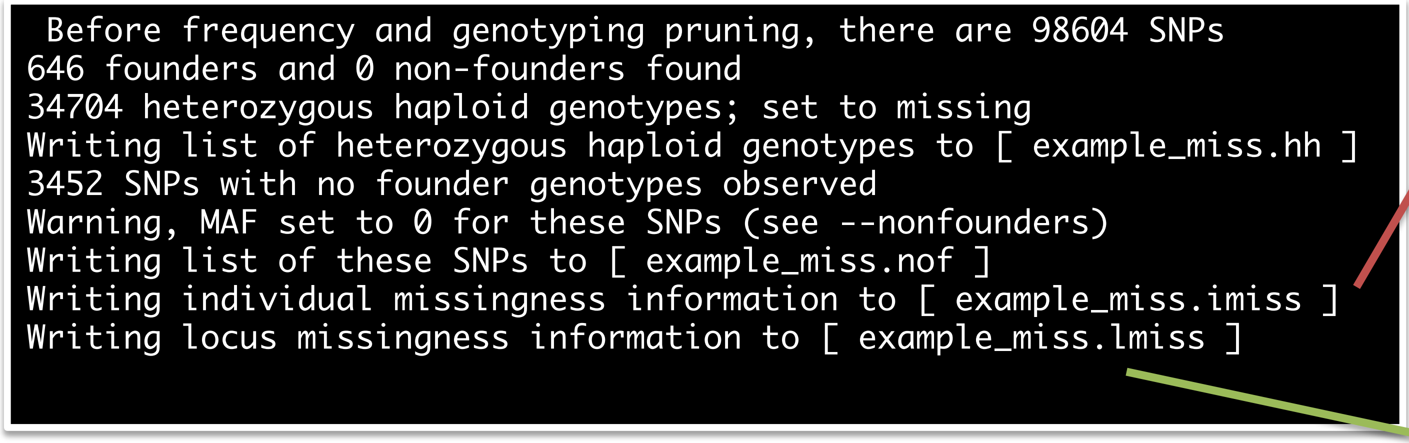
- Per sample (individual) rate
- Number of **(total non-missing genotypes(N) – homozygous(0)) genotypes divided by total non-missing genotypes(N)**
- Excess heterozygosity - Possible sample **contamination**
- Less than expected heterozygosity - Possibly **inbreeding**
- Threshold is usually to remove any individual with +- 3 standard deviations from the mean heterozygosity rate of all samples.



Missingness per Individual and per SNP



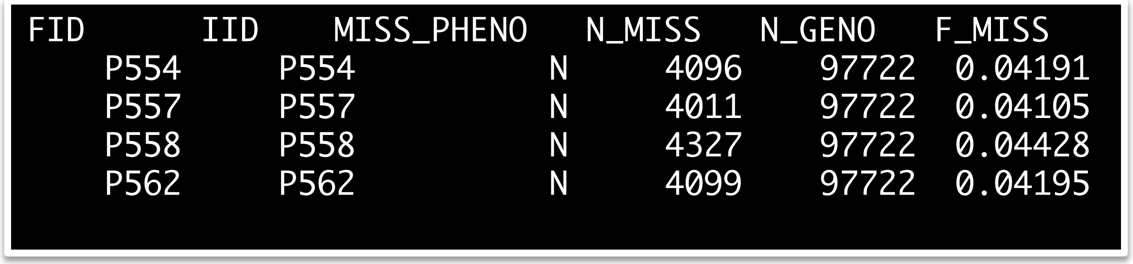
```
plink --bfile example –missing -
-out example_miss
```



```
 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 ]
```

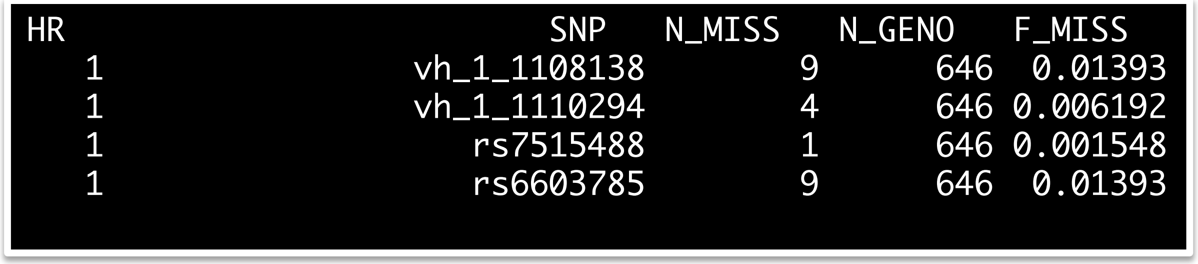
Missing phenotype (Y/NN)

Missingness per individual



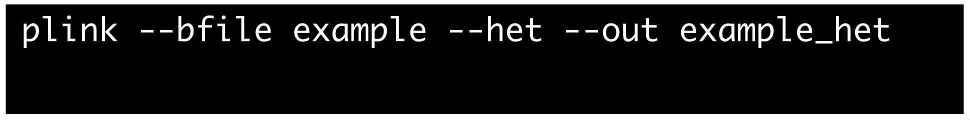
```
FID        IID      MISS_PHENO    N_MISS    N_GENO    F_MISS
    P554       P554            N      4096     97722   0.04191
    P557       P557            N      4011     97722   0.04105
    P558       P558            N      4327     97722   0.04428
    P562       P562            N      4099     97722   0.04195
```

Missingness per SNP (Come back to later)

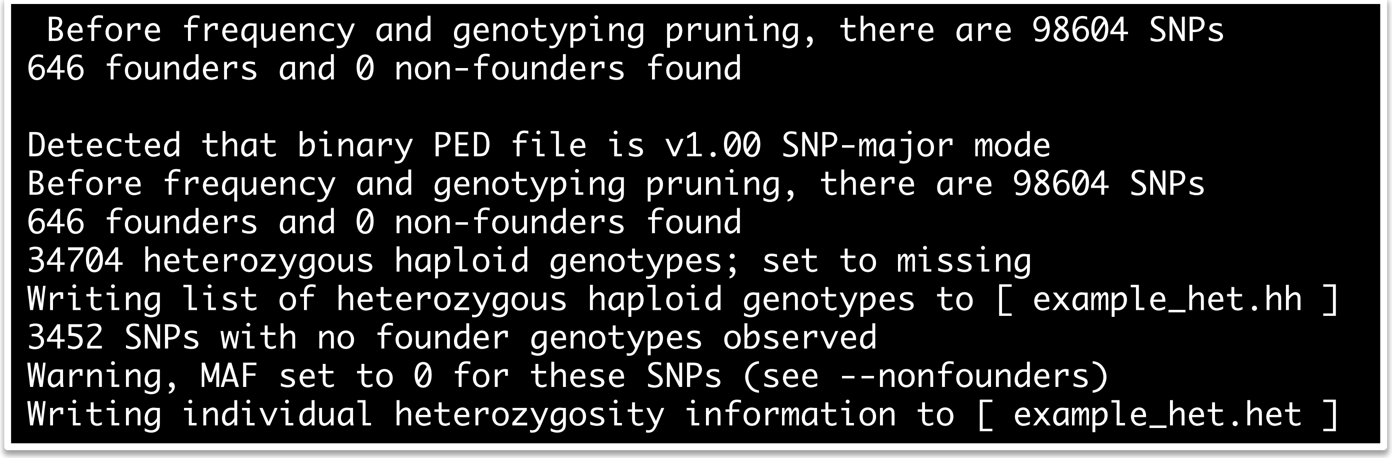


```
HR                          SNP   N_MISS    N_GENO    F_MISS
   1               vh_1_1108138        9       646   0.01393
   1               vh_1_1110294        4       646  0.006192
   1                  rs7515488        1       646  0.001548
   1                  rs6603785        9       646   0.01393
```

Heterozygosity rate per Individual



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



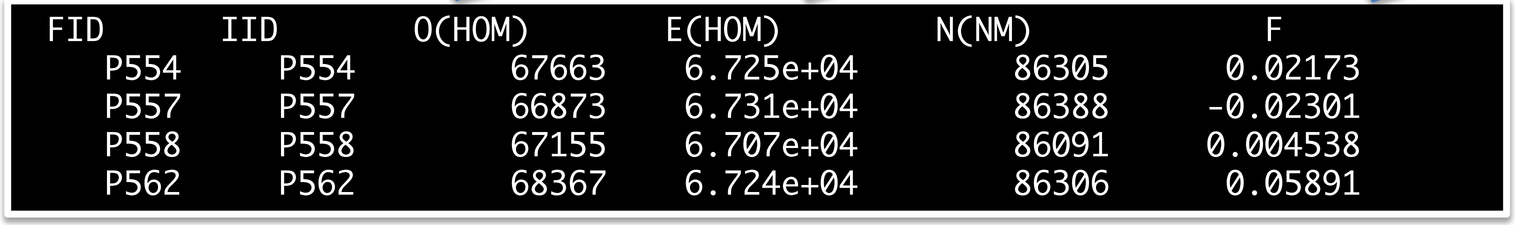
```
 Before frequency and genotyping pruning, there are 98604 SNPs
646 founders and 0 non-founders found

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 ]
```

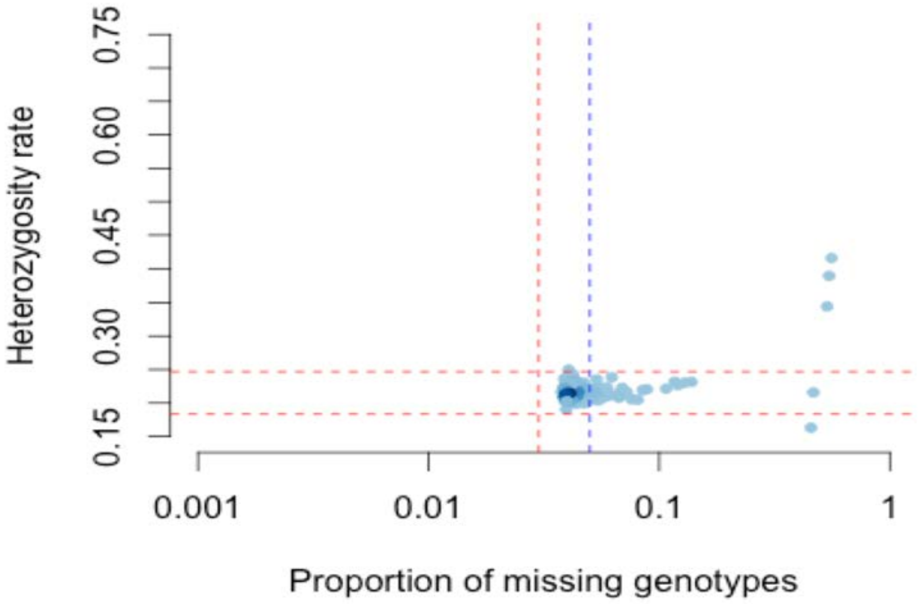
Observed number of homozygous genotypes

Expected number of homozygous genotypes

Inbreeding coefficient estimate



```
  FID        IID        O(HOM)        E(HOM)        N(NM)          F
    P554       P554         67663   6.725e+04        86305    0.02173
    P557       P557         66873   6.731e+04        86388   -0.02301
    P558       P558         67155   6.707e+04        86091   0.004538
    P562       P562         68367   6.724e+04        86306    0.05891
```

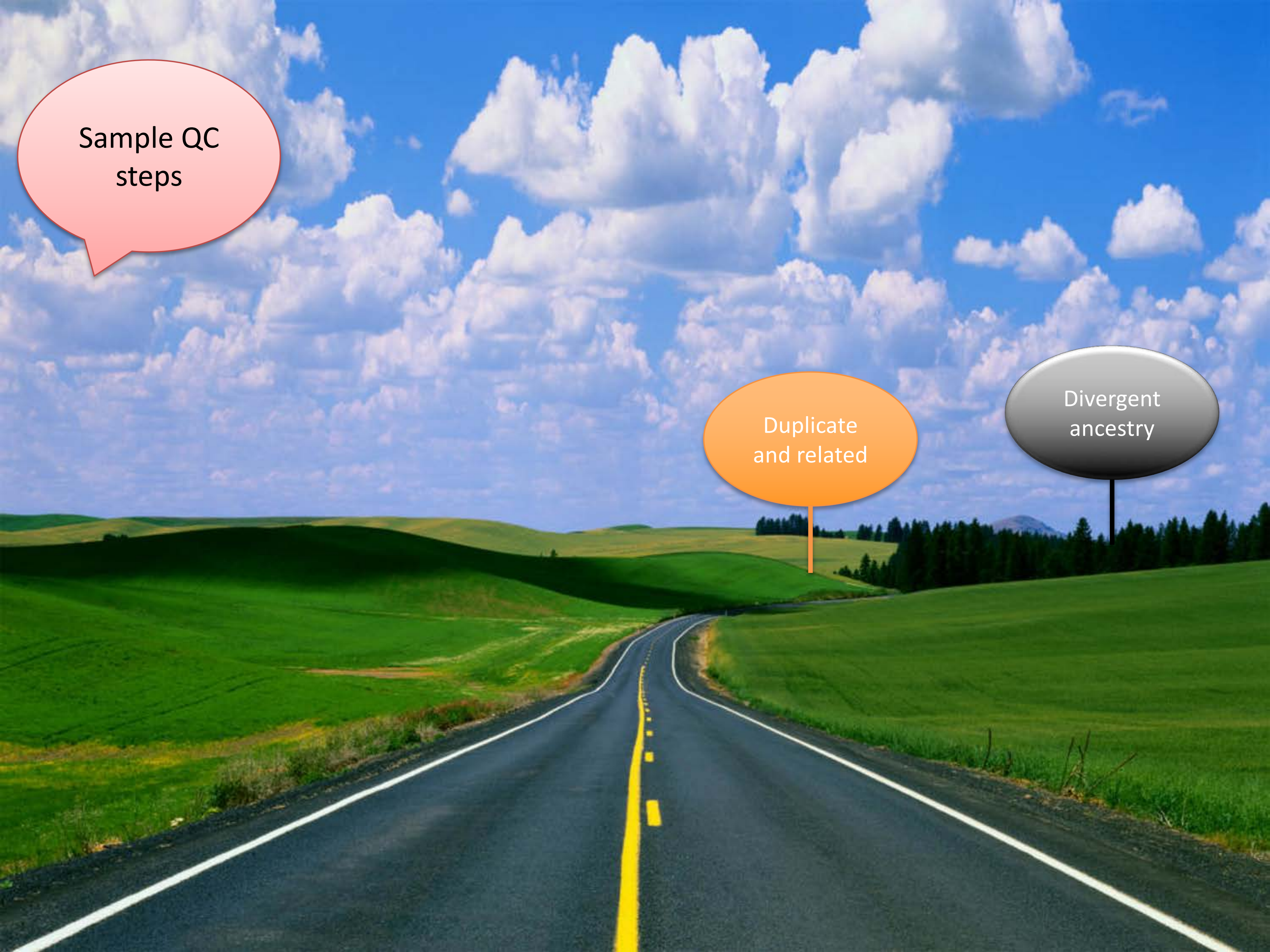


Based on the plot you can decide on reasonable thresholds at which to exclude individuals based on elevated missing or extreme heterozygosity.

We decided to exclude all individuals with a genotype failure rate ≥ 0.06 and heterozygosity rate ± 3 standard deviations from the mean heterozygosity rate of all samples

nextflow Parameters

- cut_mind = 0.02
- cut_het_high = 0.343
- cut_het_low = 0.15



Sample QC steps

Duplicate and related

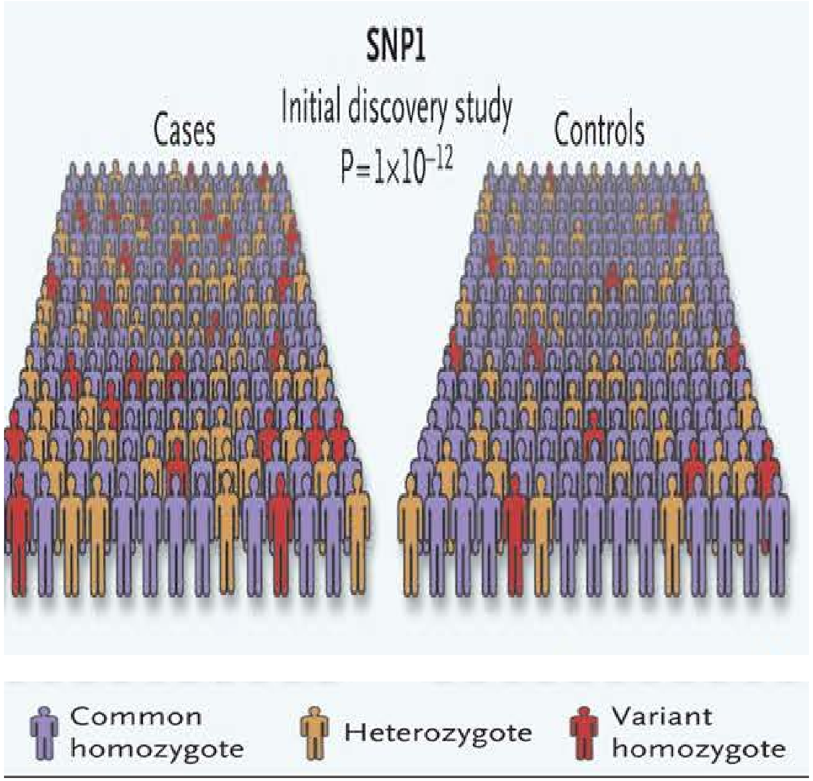
Divergent ancestry

Identify related and duplicate individuals

- A basic assumption of standard population-based 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)



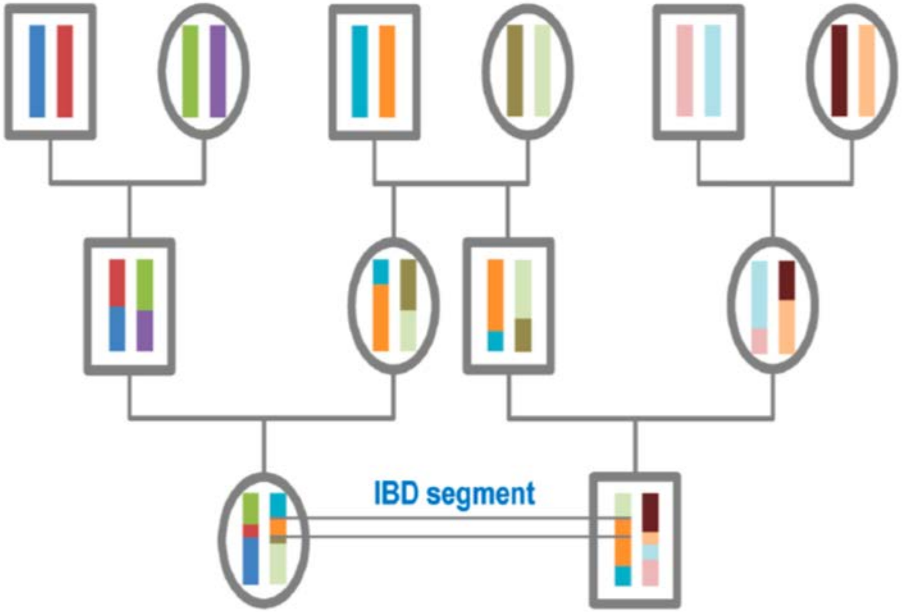
- **Identity by Descent (IBD)** is a **measure** of the **recent shared ancestry** between two individuals based on **genome wide IBS**

- IBD calculations works best when only **independent SNPs** are included in the analysis.

- An independent SNP set for IBD calculations 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.

- IBD is calculated and denoted in Plink as **Pi-hat**

- Convention is to **remove** one individual from a pair with a **Pi-hat > ~0.2**



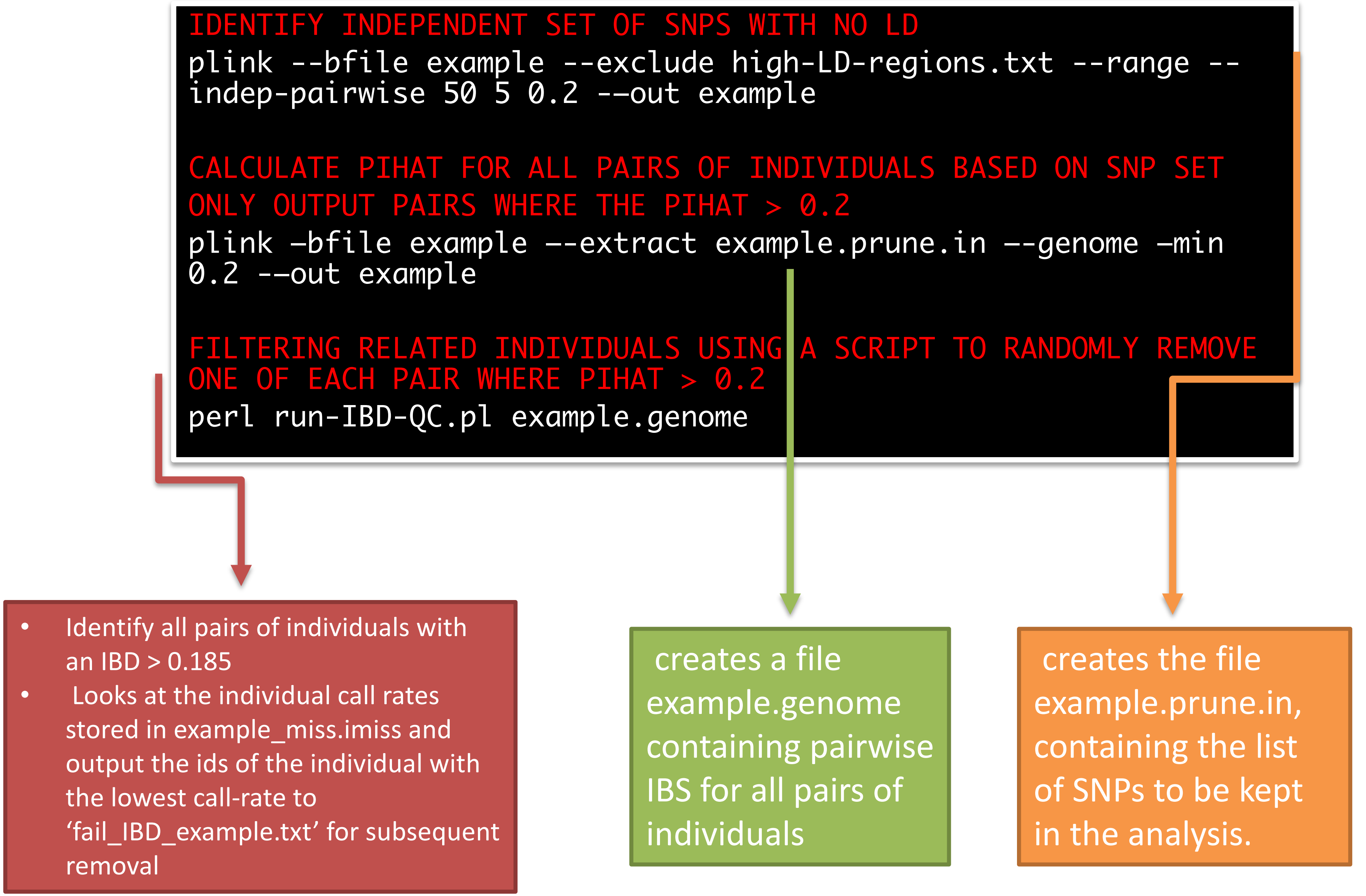
Pi – hat ~ 1 (Duplicate sample or Monozygotic twins)

Pi – hat ~ 0.5 (First degree relative)

Pi – hat ~ 0.25 (Second degree relative)

Pi – hat ~ 0.125 (Third degree relative)

Identification of duplicated or related individuals



```
IDENTIFY INDEPENDENT SET OF SNPS WITH NO LD
plink --bfile example --exclude high-LD-regions.txt --range --
indep-pairwise 50 5 0.2 -–out example

CALCULATE PIHAT FOR ALL PAIRS OF INDIVIDUALS BASED ON SNP SET
ONLY OUTPUT PAIRS WHERE THE PIHAT > 0.2
plink –bfile example --extract example.prune.in --genome –min
0.2 -–out example

FILTERING RELATED INDIVIDUALS USING A SCRIPT TO RANDOMLY REMOVE
ONE OF EACH PAIR WHERE PIHAT > 0.2
perl run-IBD-QC.pl example.genome
```

- 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

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

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

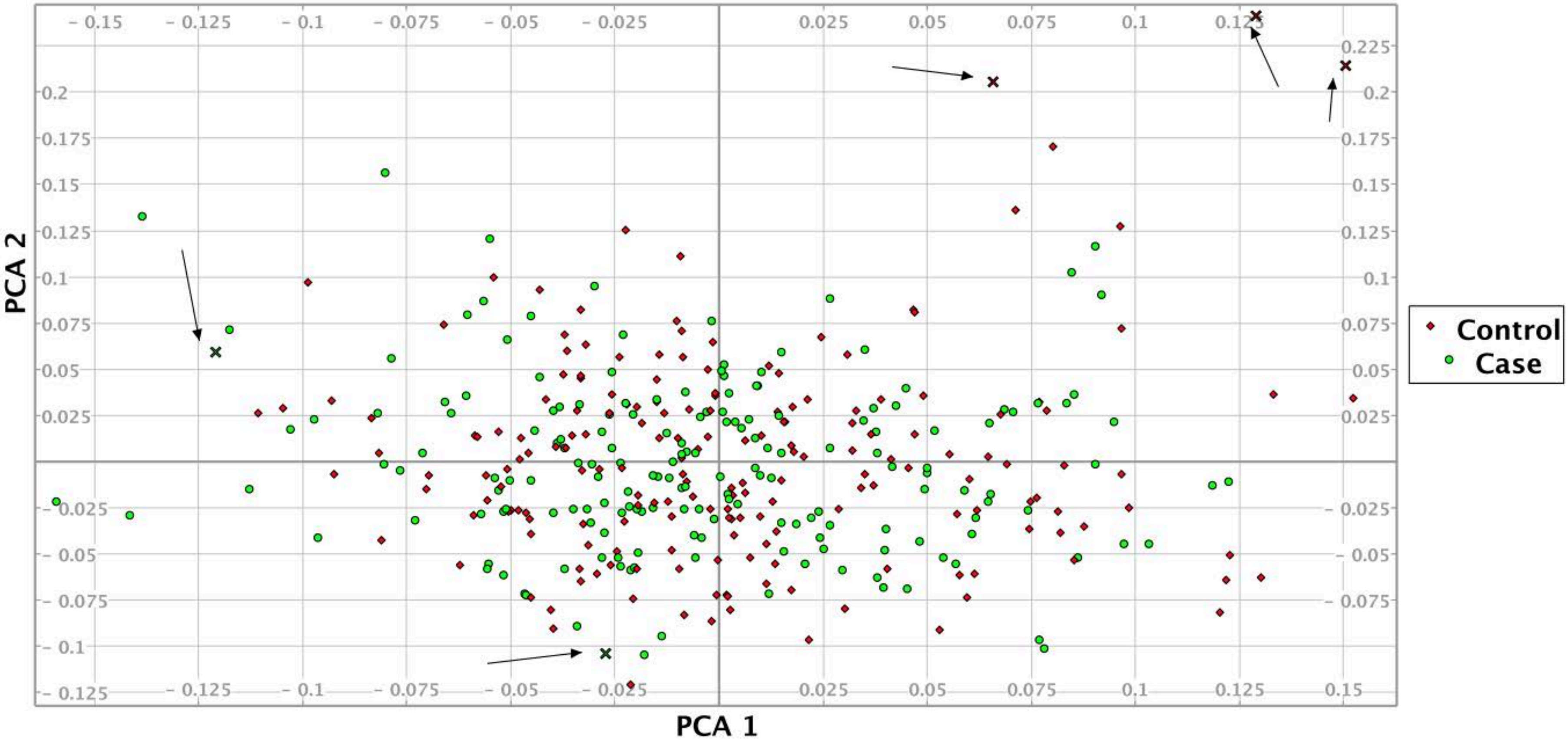
nextflow Parameters

- `pi-hat = 0.11`
- `super_pi_hat = 0.7`

Population structure

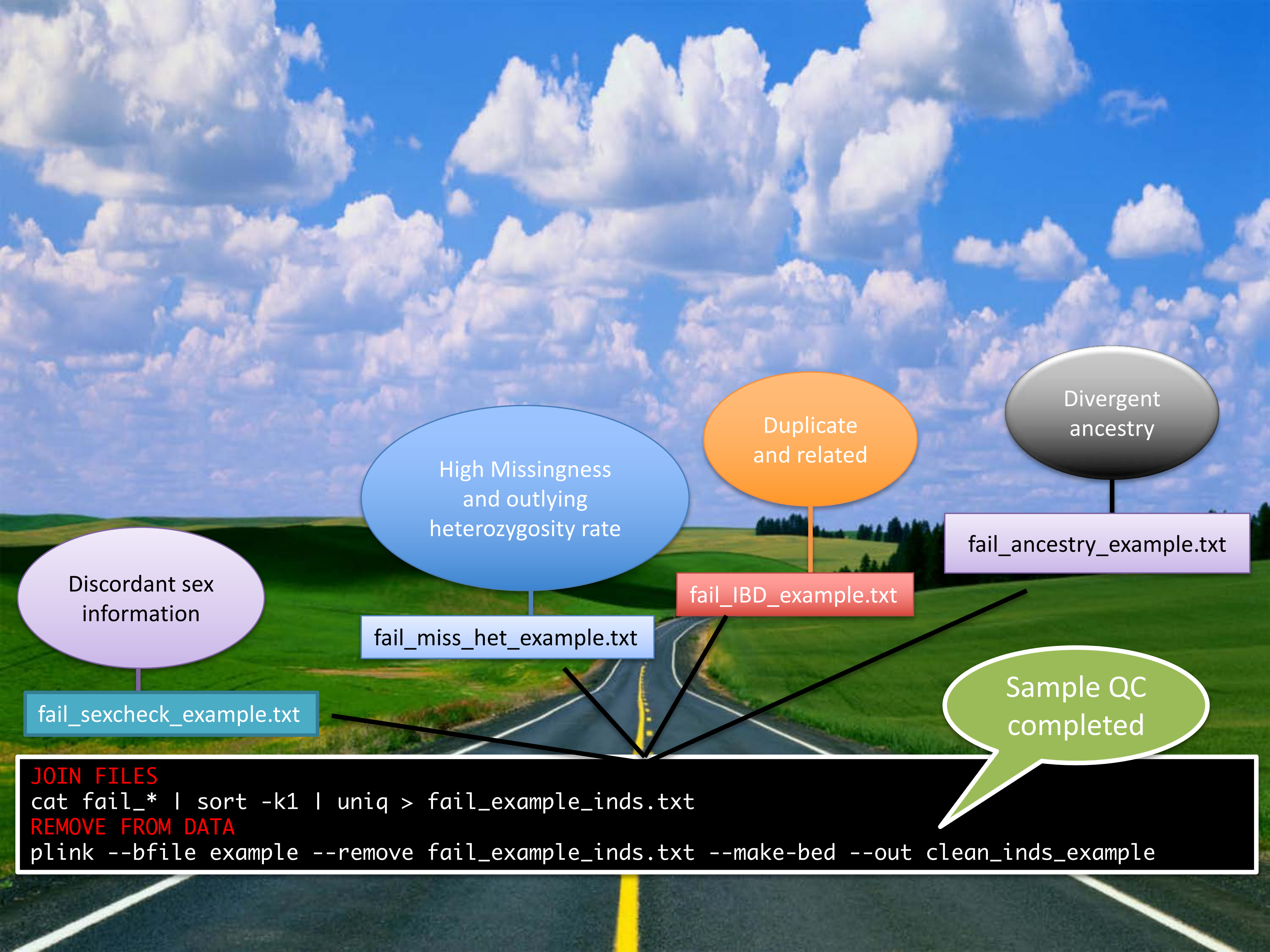
- **Population substructure** or **stratification** occurs when samples have **different genetic ancestries**
- Can lead to **spurious associations** due to **differences in ancestry** rather than true associations
- Imperative to check for population structure within samples
- Can **control** for structure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

Population structure - PCA



nextflow Parameters

- case_control = sample.phe
- case_control_col = PHE
- batch = sample.phe
- batch_col = batch_no



Discordant sex information

fail_sexcheck_example.txt

High Missingness and outlying heterozygosity rate

fail_miss_het_example.txt

Duplicate and related

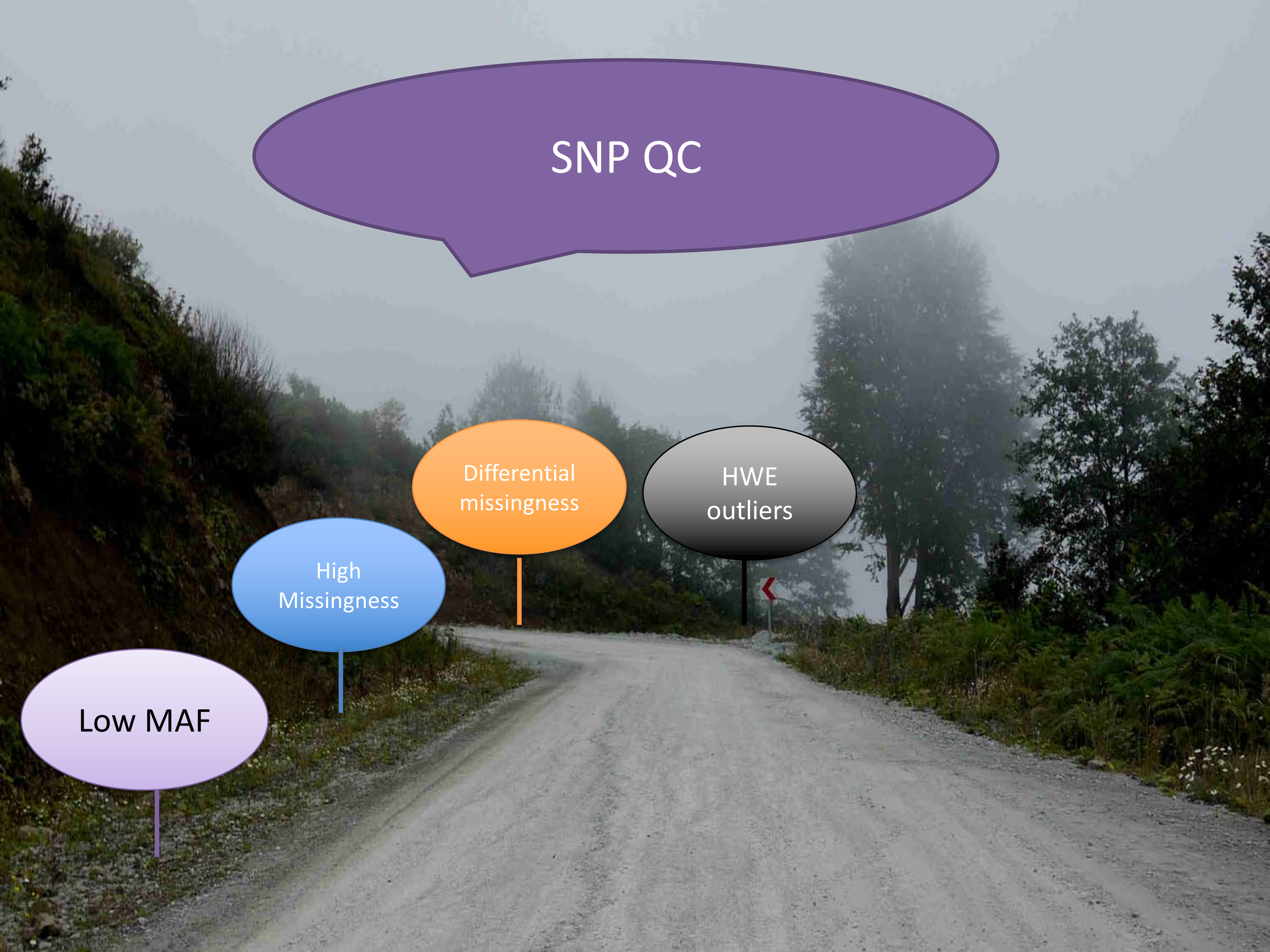
fail_IBD_example.txt

Divergent ancestry

fail_ancestry_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
```

Sample QC completed



SNP QC

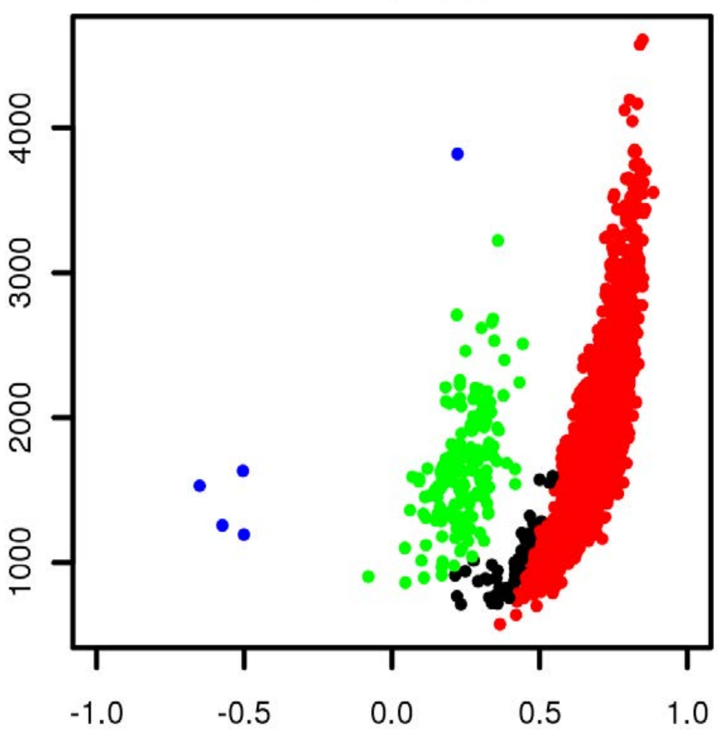
Low MAF

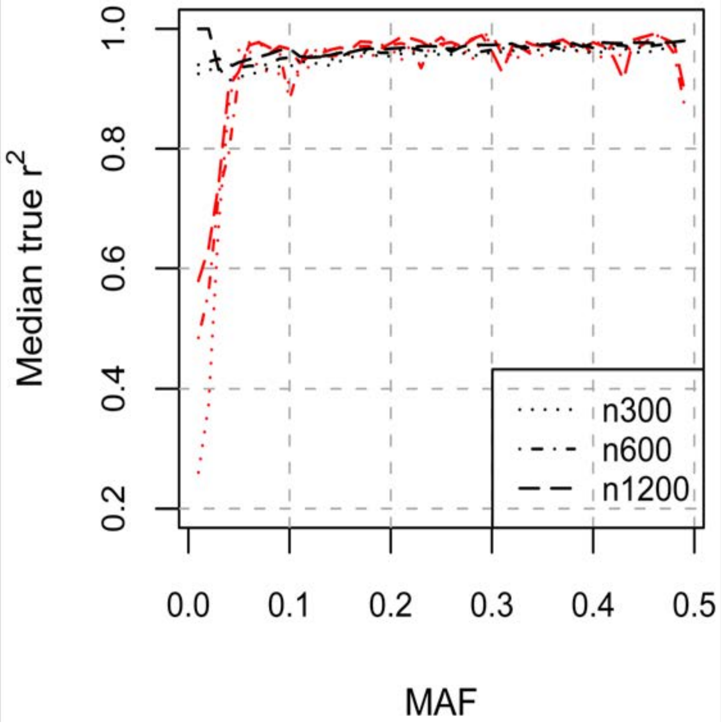
High Missingness

Differential missingness

HWE outliers

Low minor allele frequency SNPs





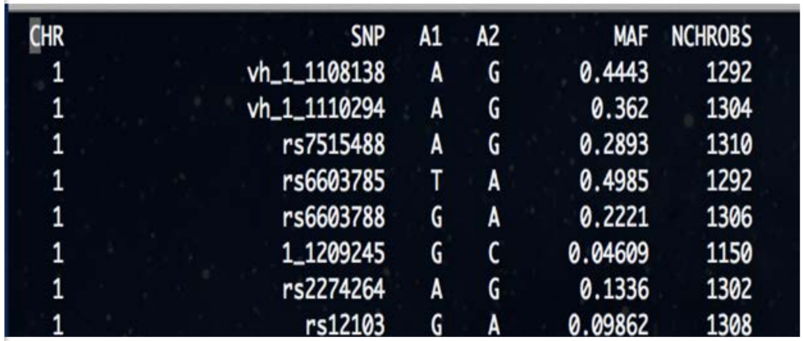
- Genotype calling algorithms **perform poorly** for SNPs with **low MAF** and **low samples sizes.**
- **Power** for detecting associations with SNPs with low MAF is low unless the sample size is very large
- Commonly used exclusion threshold are SNPs with a **MAF 0.01 to 0.05** (dependent on sample size)

Identify low minor allele frequency SNPs

CALCULATE MINOR ALLELE FREQUENCY DISTRIBUTION FOR ALL SNPS

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

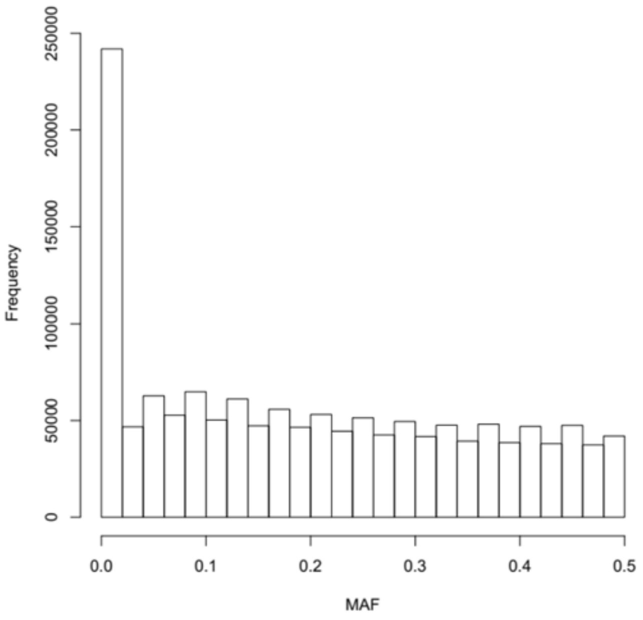
Generates the file “clean_inds_example_freq.frq” containing minor allele frequency of each SNP



```
CHR           SNP   A1   A2       MAF  NCHROBS
  1  vh_1_1108138    A    G    0.4443     1292
  1  vh_1_1110294    A    G     0.362     1304
  1     rs7515488    A    G    0.2893     1310
  1     rs6603785    T    A    0.4985     1292
  1     rs6603788    G    A    0.2221     1306
  1     1_1209245    G    C   0.04609     1150
  1     rs2274264    A    G    0.1336     1302
  1       rs12103    G    A   0.09862     1308
```

GENERATE PLOT OF MAF DISTRIBUTION

PLOT_MAF.R



Choose standard MAF threshold (MAF > 0.01 OR MAF > 0.5) or base on distribution

nextflow Parameters

- `cut_maf = 0.01`

SNP/Marker QC steps

High Missingness

Differential missingness

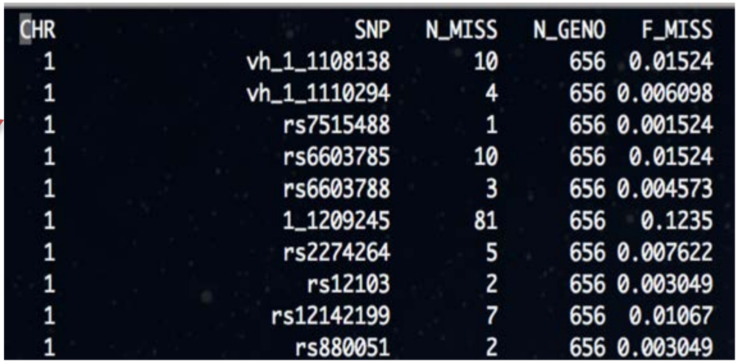
HWE outliers

Identify SNPS with high missingness

CALCULATE MISSINGNESS PER SNP

```
plink --bfile clean_inds_example --missing --out clean_inds_example_missing
```

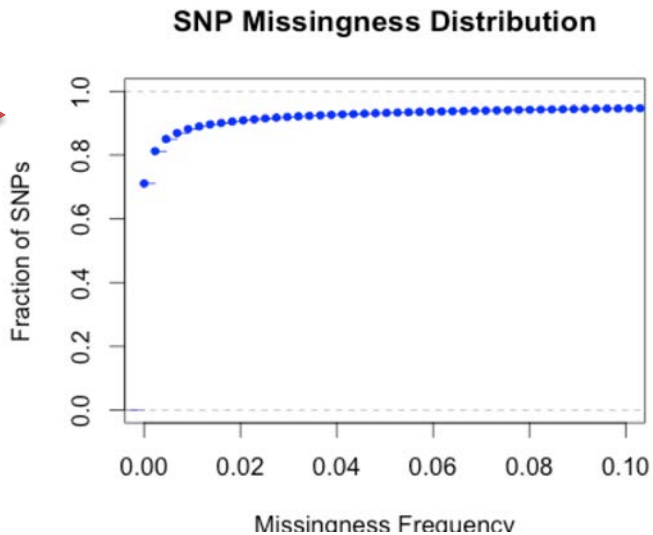
Generates the file “clean_inds_example_mising.lmiss” containing missingness value for each SNP



```
CHR              SNP   N_MISS   N_GENO   F_MISS
  1    vh_1_1108138       10      656  0.01524
  1    vh_1_1110294        4      656 0.006098
  1       rs7515488        1      656 0.001524
  1       rs6603785       10      656  0.01524
  1       rs6603788        3      656 0.004573
  1       1_1209245       81      656   0.1235
  1       rs2274264        5      656 0.007622
  1         rs12103        2      656 0.003049
  1      rs12142199        7      656  0.01067
  1        rs880051        2      656 0.003049
```

GENERATE PLOT OF SNP MISSINGNESS

SNPMISS_PLOT.R



Choose the standard missingness threshold (> 0.05) or choose threshold based on the plot

nextflow Parameters

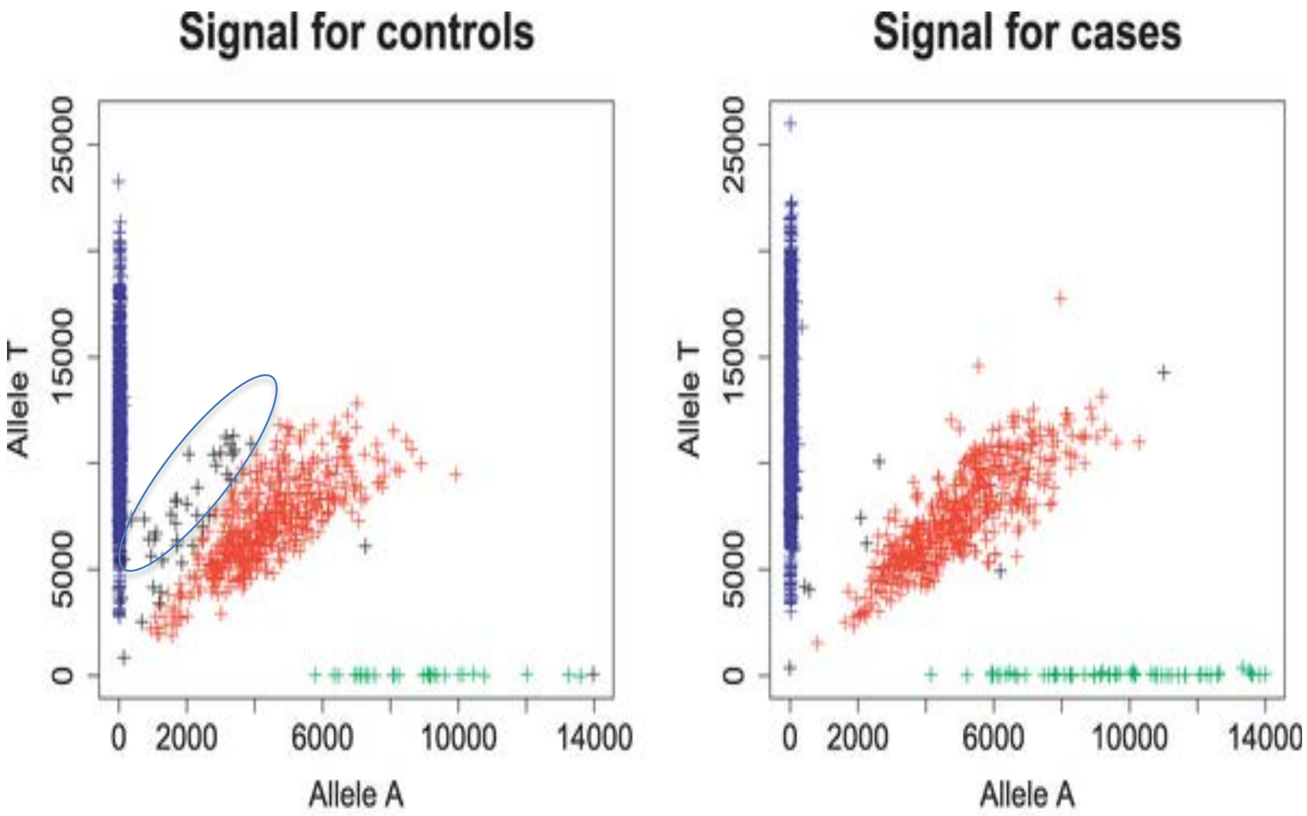
- `cut_geno = 0.01`

SNP/Marker QC steps

Differential missingness

HWE outliers

Differential missingness



- 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 between cases and controls

CALCULATE DIFFERENTIAL MISSINGNESS

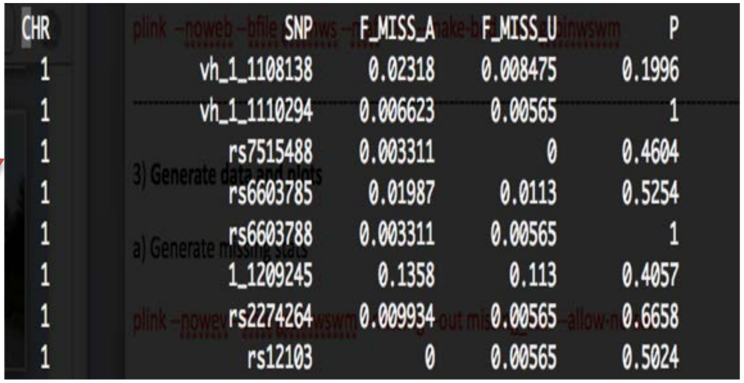
```
plink --bfile clean_inds_example --test-missing --out clean_inds_example_test_missing
```

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

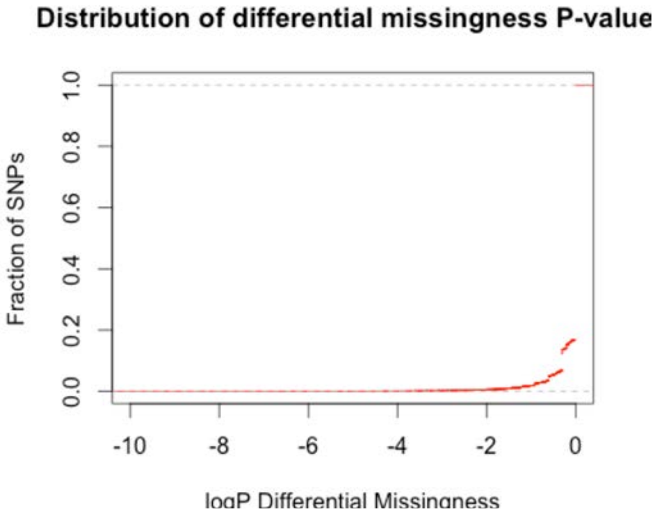
GENERATE PLOT OF P-VALUE DISTRIBUTION

DIFFMISS_PLOT.R

Choose standard differential missingness p-value threshold (0.00001) or choose on the basis of the plot



```
 CHR          SNP   F_MISS_A   F_MISS_U        P
   1  vh_1_1108138    0.02318   0.008475   0.1996
   1  vh_1_1110294   0.006623    0.00565        1
   1    rs7515488   0.003311          0   0.4604
   1    rs6603785    0.01987     0.0113   0.5254
   1    rs6603788   0.003311    0.00565        1
   1    1_1209245     0.1358      0.113   0.4057
   1    rs2274264   0.009934    0.00565   0.6658
   1      rs12103          0    0.00565   0.5024
```



nextflow Parameters

- `cut_diff_miss = 0.05`

SNP/Marker QC steps

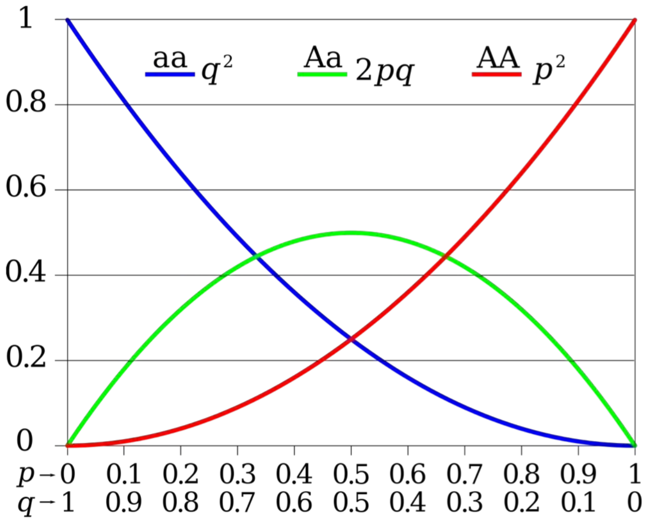
HWE outliers

Hardy Weinberg Equilibrium

- Expected relationship between allele and genotype frequencies under certain assumptions
- Allele frequencies and genotypes remains constant over generations
- Deviations from HWE are used as a proxy for possible genotyping errors

Assumptions

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



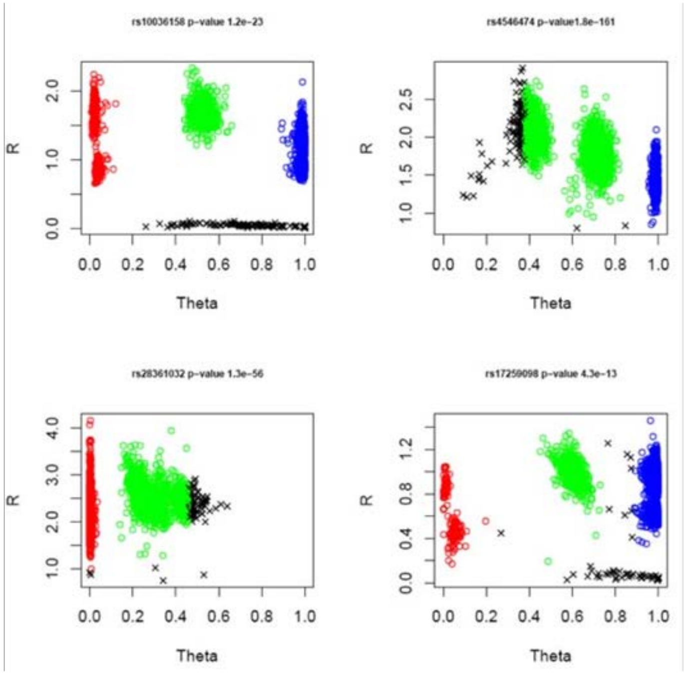
Testing for HWE

- 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
  - χ^2 test

Reasons for HW Deviations

- **Genotyping Error**
- Subdivided Population
- 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}**
- Can **only remove SNPs in controls** which deviate from HWE or **use less stringent HWE threshold** in **SNPs in cases over controls**



GENEVA alcohol-dependence project: Quality control report

Identify SNPS which show extreme HWE deviations

GET DISTRIBUTION OF P-VALUES FOR ALL SNPS

```
plink --bfile clean_inds_example --hardy --out
clean_inds_example_hwe
```

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

SELECT UNAFFECTED

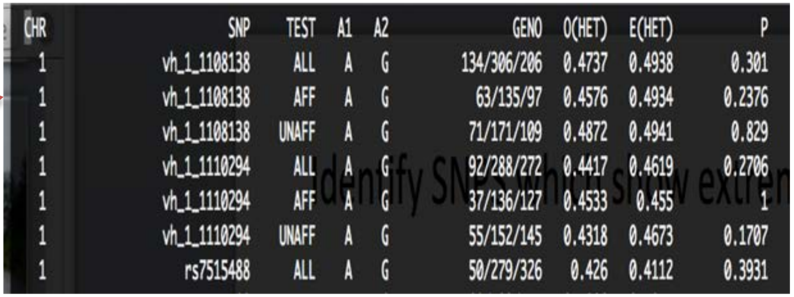
```
head -1 clean_inds_example_hwe.hwe >
example_clean_inds_example_hweu.hwe | grep
"UNAFF" clean_inds_example_hwe.hwe >>
example_clean_inds_example_hweu.hwe
```

GENERATE PLOT USING R SCRIPT

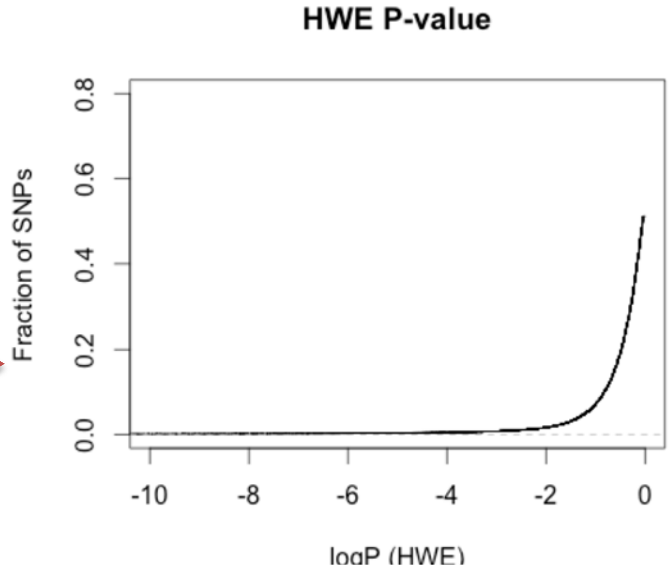
HWE_PLOT.R

(based only on controls)

Choose the standard HWE P-value threshold (0.00001) or select one on the basis of the plot

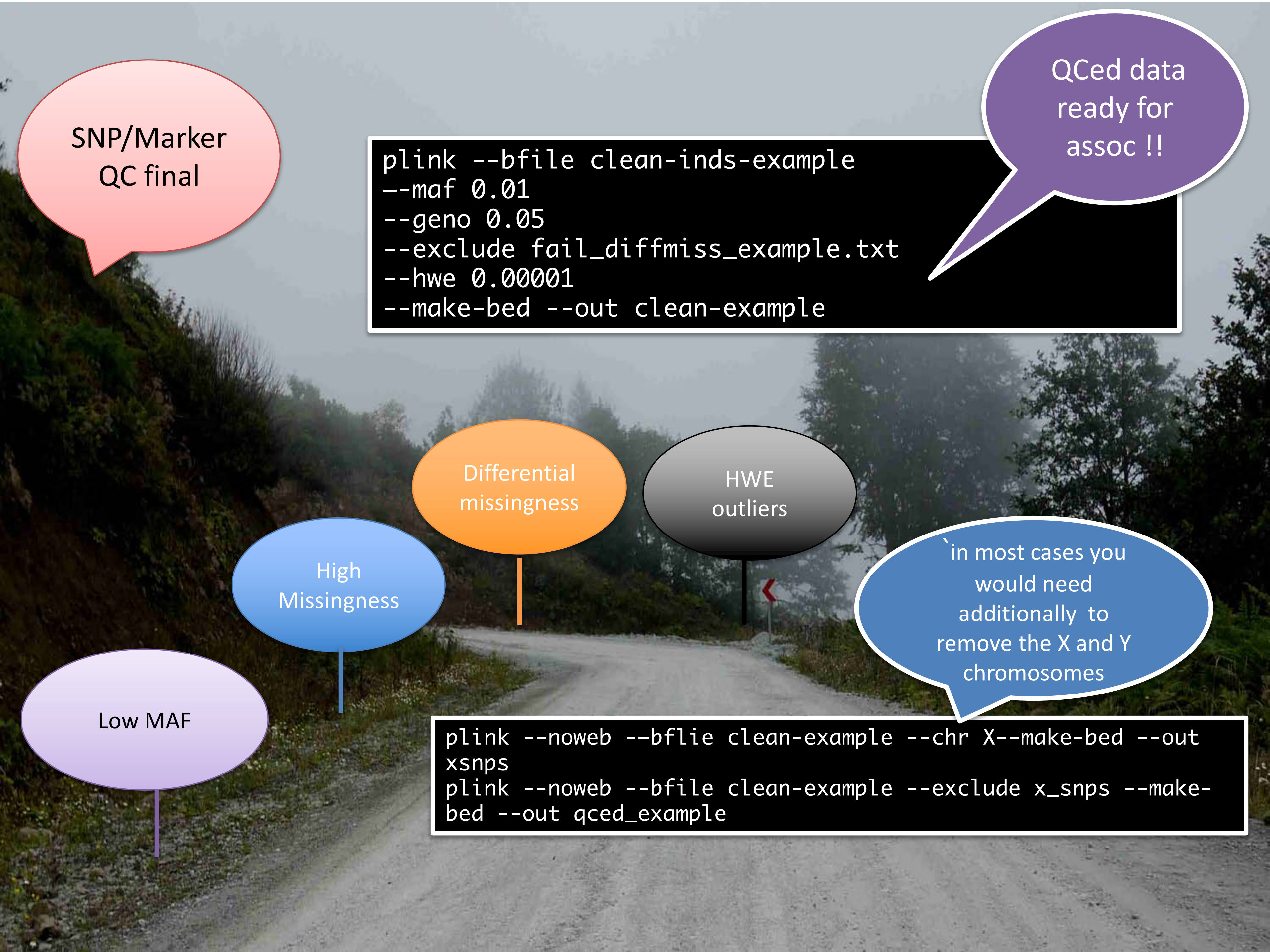


```
CHR          SNP   TEST  A1  A2          GENO   O(HET)  E(HET)       P
  1  vh_1_1108138   ALL   A   G   134/306/206   0.4737  0.4938   0.301
  1  vh_1_1108138   AFF   A   G     63/135/97   0.4576  0.4934  0.2376
  1  vh_1_1108138 UNAFF   A   G    71/171/109   0.4872  0.4941   0.829
  1  vh_1_1110294   ALL   A   G    92/288/272   0.4417  0.4619  0.2706
  1  vh_1_1110294   AFF   A   G    37/136/127   0.4533   0.455       1
  1  vh_1_1110294 UNAFF   A   G    55/152/145   0.4318  0.4673  0.1707
  1     rs7515488   ALL   A   G    50/279/326    0.426  0.4112  0.3931
```



nextflow Parameters

- `cut_hwe = 0.000001`



SNP/Marker QC final

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

QCed data ready for assoc !!

Low MAF

High Missingness

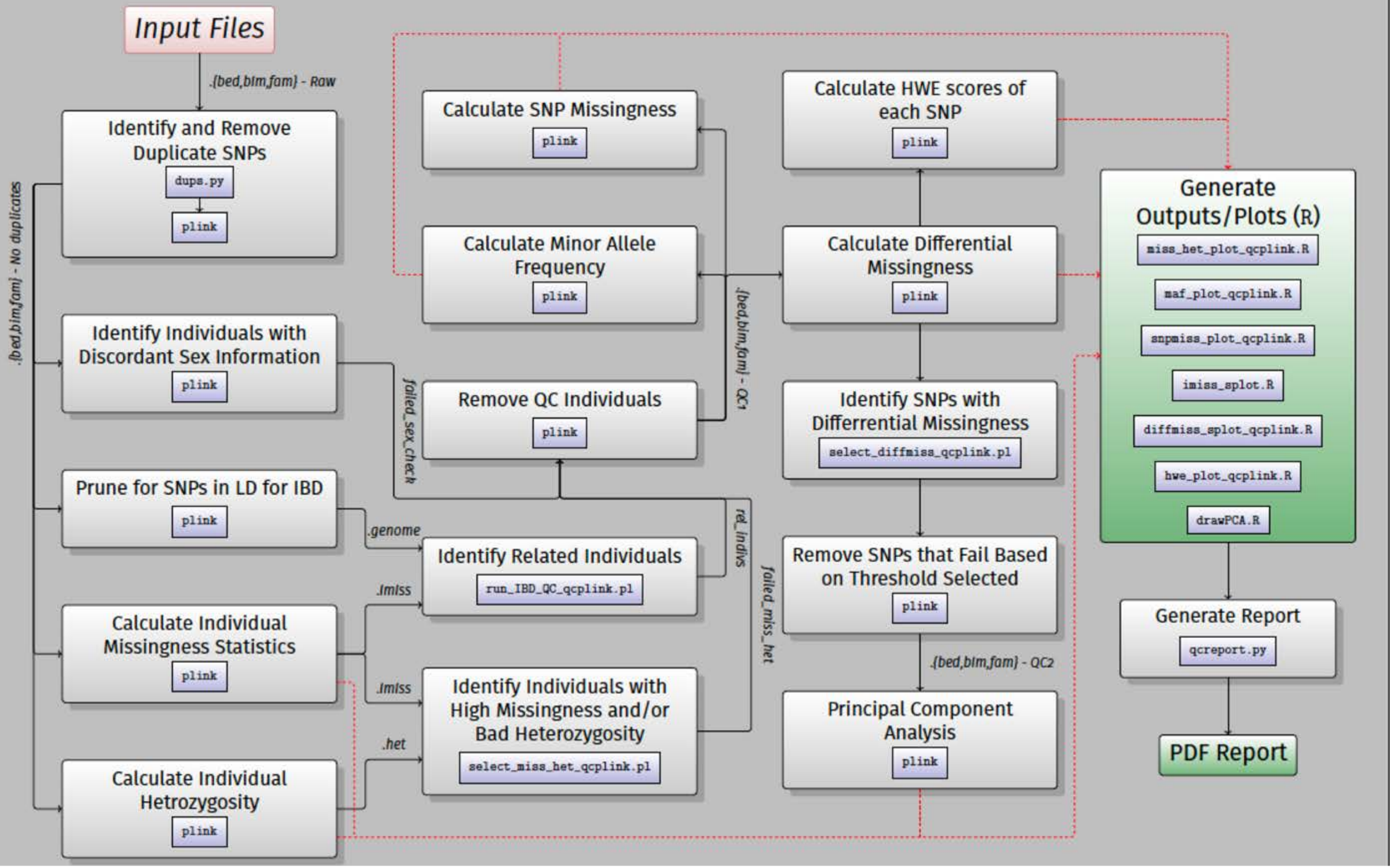
Differential missingness

HWE outliers

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

```
plink --noweb --bflie clean-example --chr X--make-bed --out
xsnps
plink --noweb --bfile clean-example --exclude x_snps --make-
bed --out qced_example
```

Genome Wide-Association Studies



Input Files

Identify and Remove Duplicate SNPs
`dups.py`
`plink`

Identify Individuals with Discordant Sex Information
`plink`

Prune for SNPs in LD for IBD
`plink`

Calculate Individual Missingness Statistics
`plink`

Calculate Individual Hetrozygosity
`plink`

Calculate SNP Missingness
`plink`

Calculate Minor Allele Frequency
`plink`

Remove QC Individuals
`plink`

Identify Related Individuals
`run_IBD_QC_qcplink.pl`

Identify Individuals with High Missingness and/or Bad Heterozygosity
`select_miss_het_qcplink.pl`

Calculate HWE scores of each SNP
`plink`

Calculate Differential Missingness
`plink`

Identify SNPs with Differrential Missingness
`select_diffmiss_qcplink.pl`

Remove SNPs that Fail Based on Threshold Selected
`plink`

Principal Component Analysis
`plink`

Generate Outputs/Plots (R)
`miss_het_plot_qcplink.R`
`maf_plot_qcplink.R`
`snpmiss_plot_qcplink.R`
`imiss_splot.R`
`diffmiss_splot_qcplink.R`
`hwe_plot_qcplink.R`
`drawPCA.R`

Generate Report
`qcreport.py`

PDF Report

Acknowledgments

- Ananyo Choudhury
- H3ABioNet funded by NHGRI grant number U41HG006941