

Lecture 2 : Basic tools and formats in bioinformatics



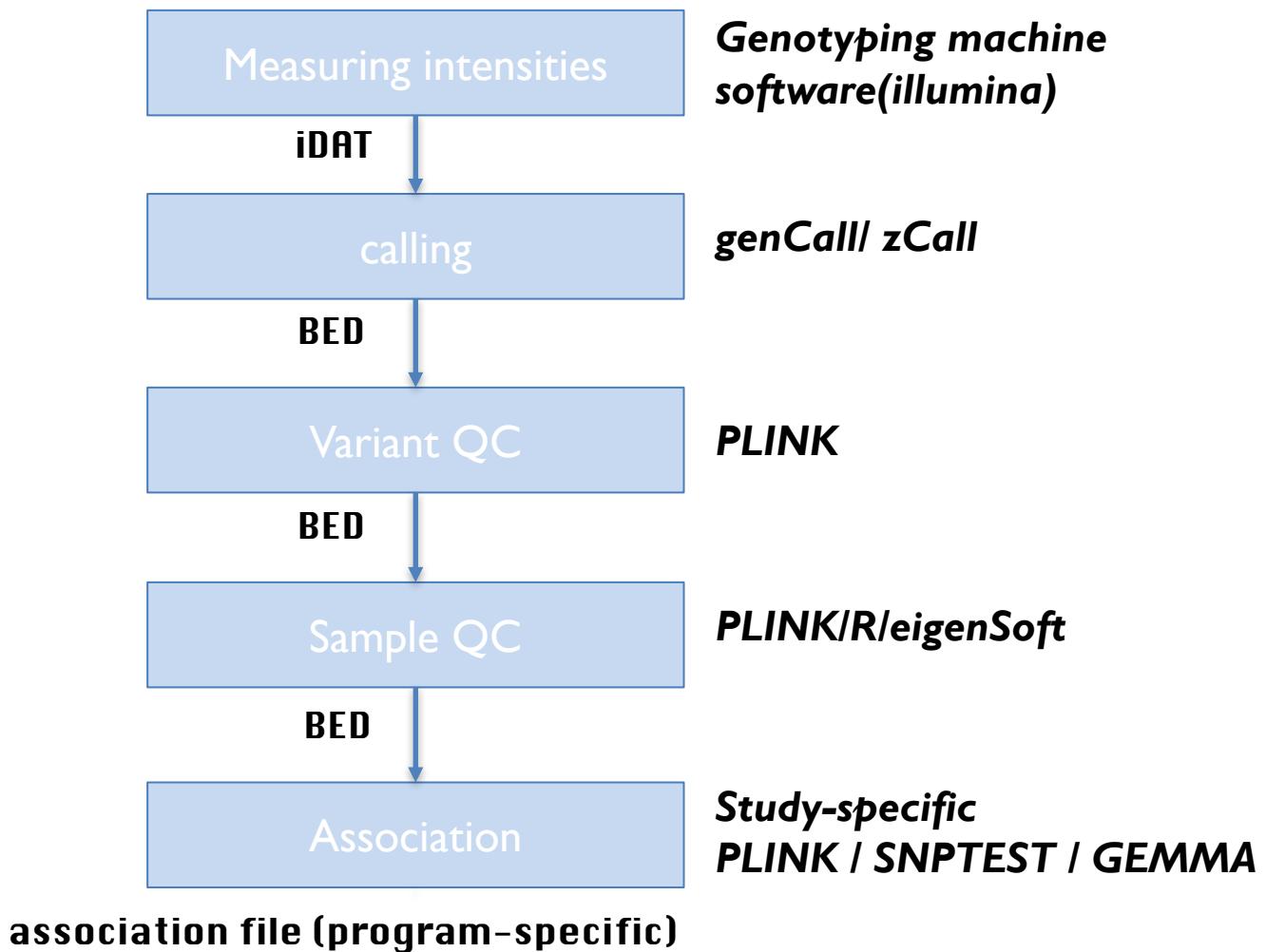
Volos Summer School

30 / 05 / 2017

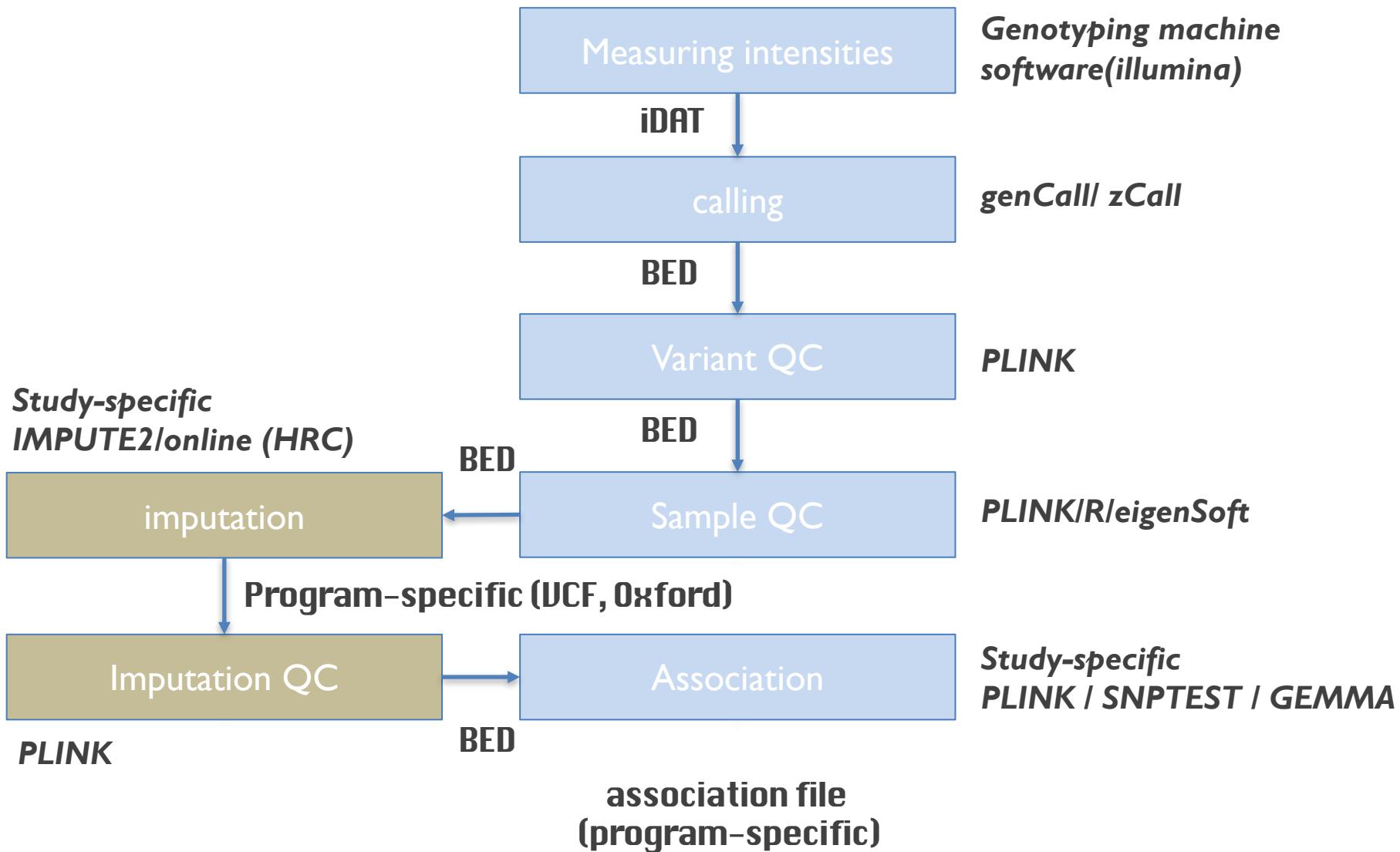
Arthur Gilly

The GWAS analysis pipeline

Tools

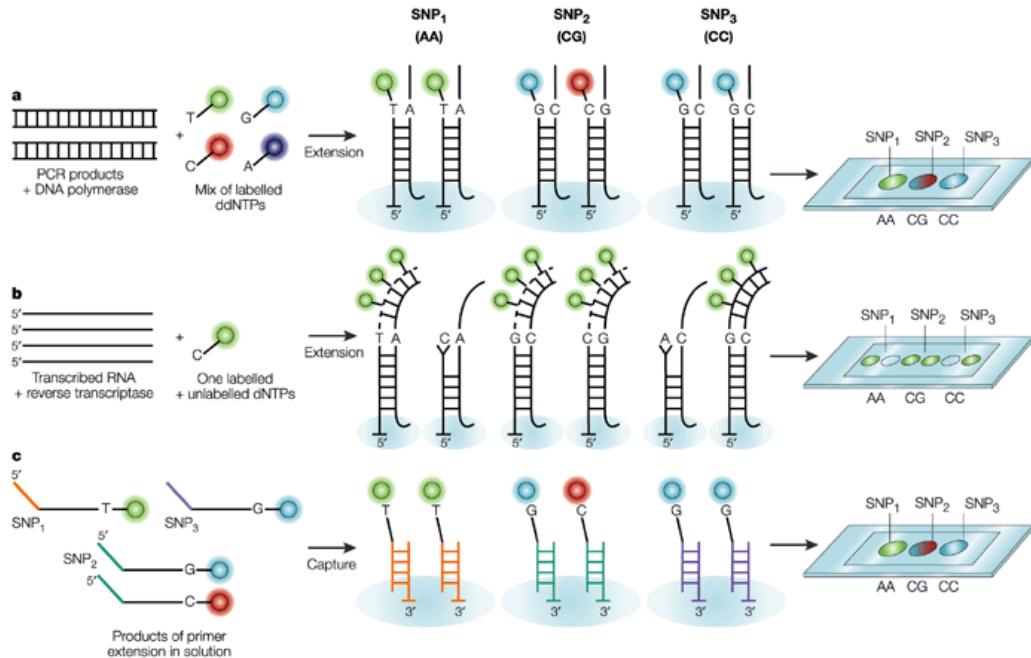


The (imputed) GWAS analysis pipeline

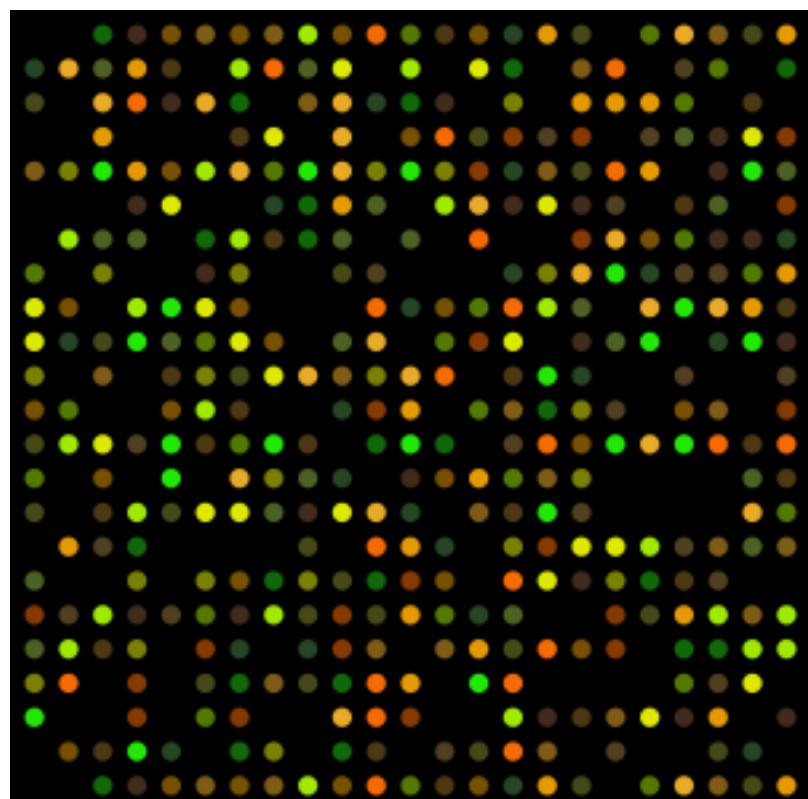


Genotyping data calling

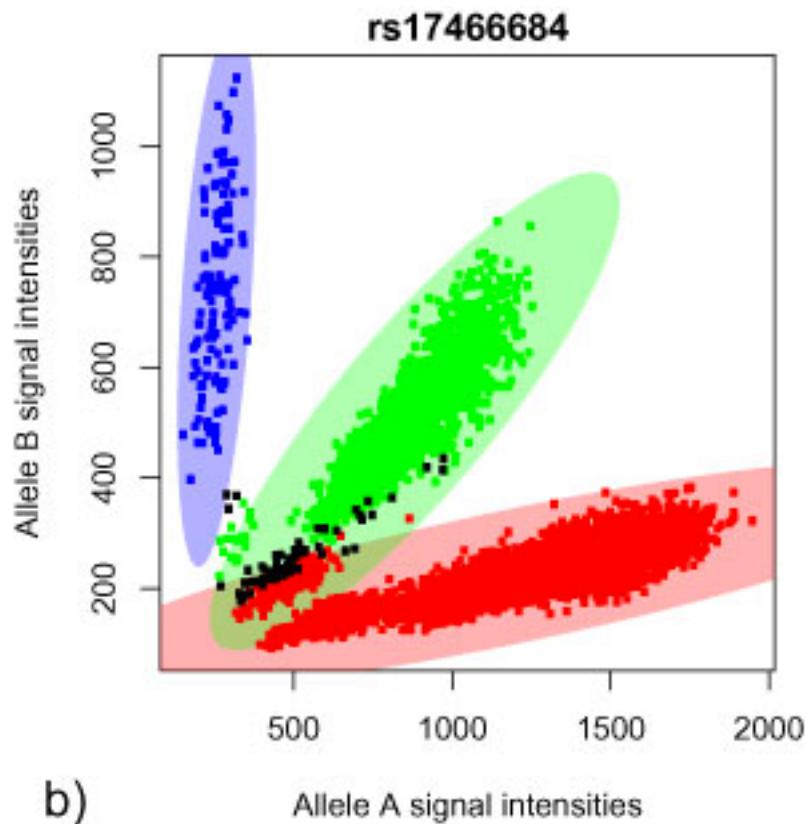
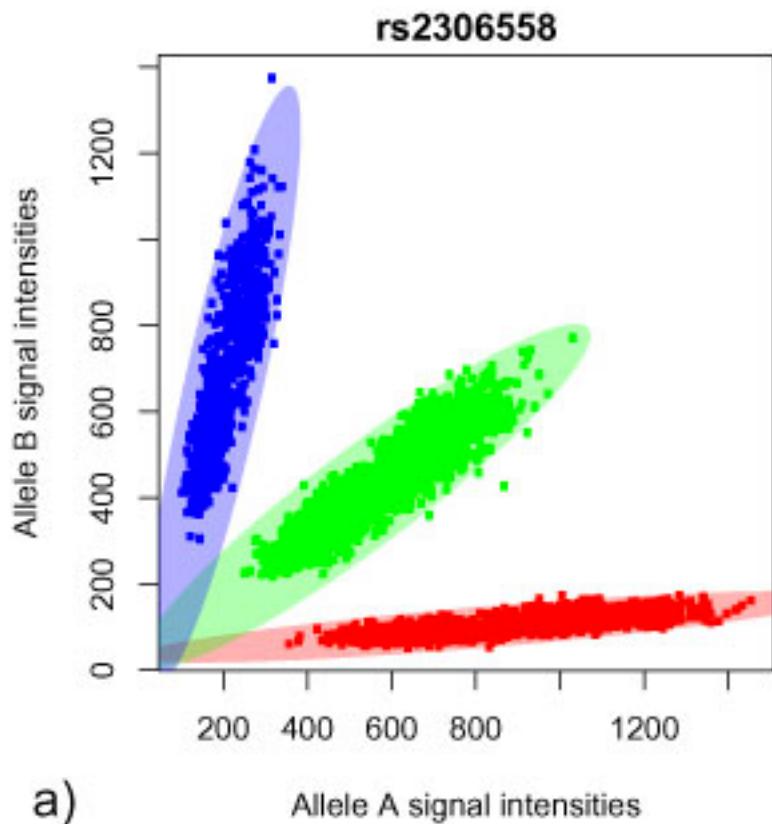
Intensities: what intensities?



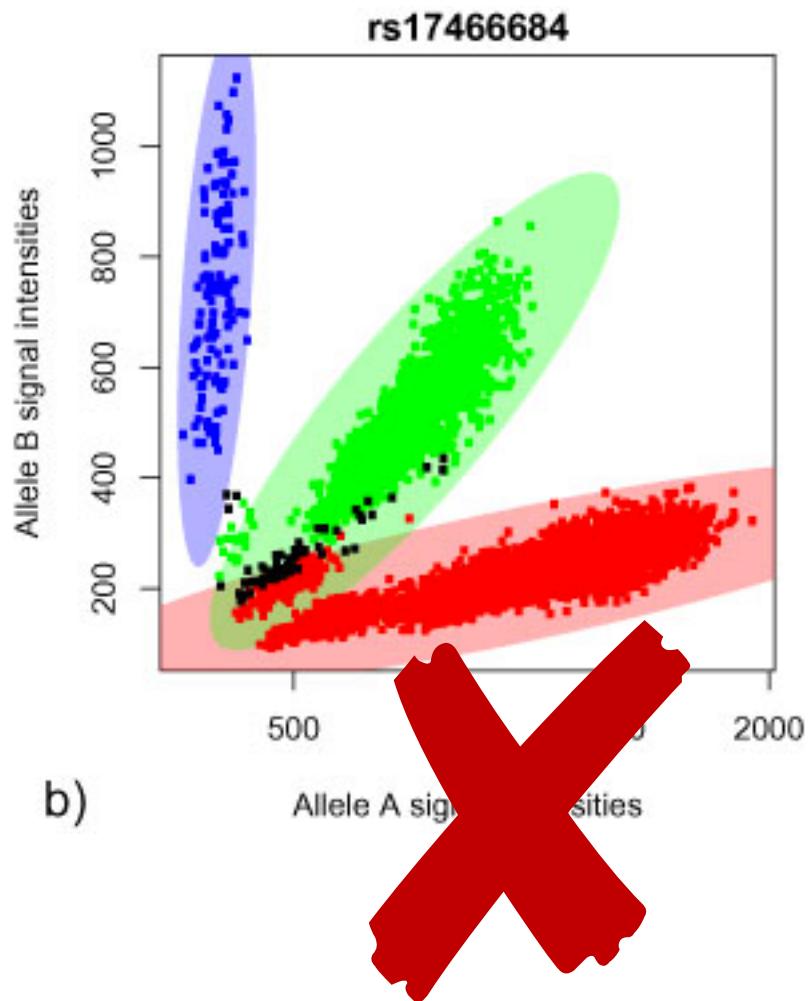
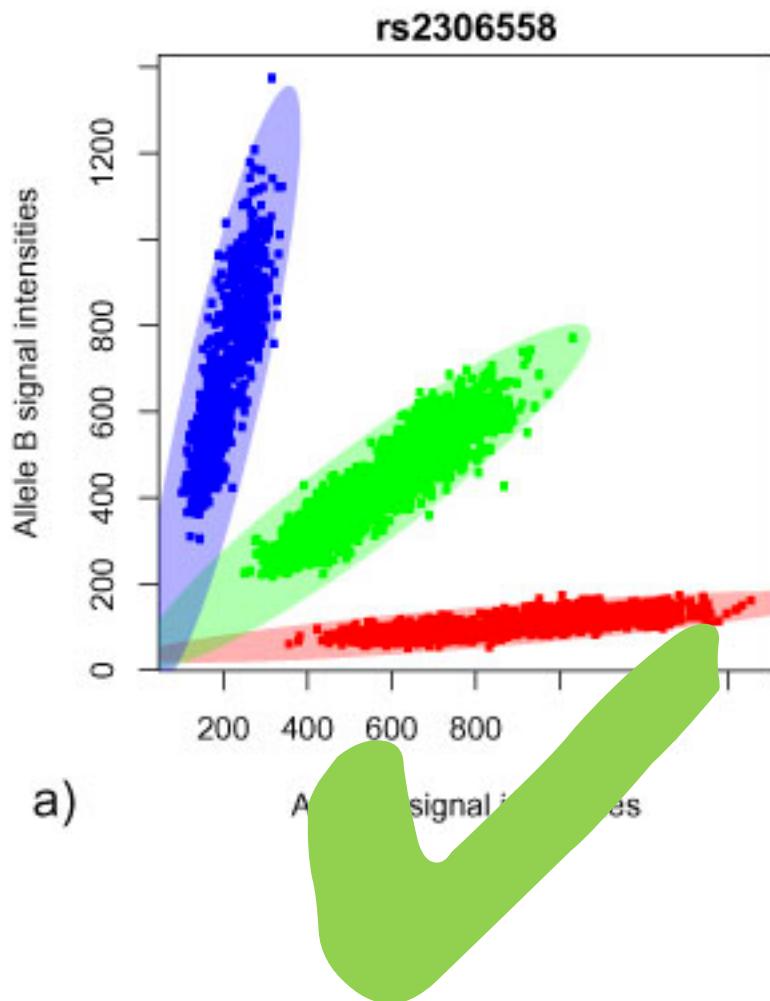
Nature Reviews | Genetics



Intensities: the good and the bad



Intensities: the good and the bad



Genotyping data storage

Which data types do we need?

Which data types do we need?

phenotype ~

Which data types do we need?

phenotype ~ genotype

Which data types do we need?

phenotype ~ genotype + covariates

Which data types do we need?

phenotype ~ genotype + covariates + structure

Which data types do we need?

phenotype ~ *genotype* + *covariates* + *structure* + ϵ

Which data types do we need?

phenotype ~ $\beta \times$ genotype + covariates + structure + ϵ

Which data types do we need?

phenotype $\sim \beta \times$ *genotype* + *covariates* + *structure* + ϵ

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \quad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \quad \begin{bmatrix} male \\ \vdots \\ female \end{bmatrix} \begin{bmatrix} 22 \text{ years} \\ \vdots \\ 65 \text{ years} \end{bmatrix} \quad \begin{bmatrix} r_{00} & \dots & r_{0n} \\ \vdots & & \vdots \\ r_{ij} & \dots & r_{nn} \end{bmatrix}$$

Which data types do we need?

phenotype $\sim \beta \times$ genotype + covariates + structure + ϵ

As we go from variant to variant...

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \quad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \quad \begin{bmatrix} male \\ \vdots \\ female \end{bmatrix} \begin{bmatrix} 22\ years \\ \vdots \\ 65\ years \end{bmatrix} \quad \begin{bmatrix} r_{00} & \dots & r_{0n} \\ \vdots & & \vdots \\ r_{ij} & \dots & r_{nn} \end{bmatrix}$$

Which data types do we need?

$$\text{phenotype} \sim \beta \times \text{genotype} + \text{covariates} + \text{structure} + \epsilon$$

As we go from variant to variant...

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \quad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \quad \begin{bmatrix} male \\ \vdots \\ female \end{bmatrix} \quad \begin{bmatrix} 22 \text{ years} \\ \vdots \\ 65 \text{ years} \end{bmatrix} \quad \begin{bmatrix} r_{00} & \dots & r_{0n} \\ r_{ij} & \dots & r_{nn} \end{bmatrix}$$


These stay constant (they describe the samples)

Which data types do we need?

$$\text{phenotype} \sim \beta \times \text{genotype} + \text{covariates} + \text{structure} + \epsilon$$

As we go from variant to variant...

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix}$$

$$\begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix}$$

$$\begin{bmatrix} male \\ \vdots \\ female \end{bmatrix}$$

$$\begin{bmatrix} 22 \text{ years} \\ \vdots \\ 65 \text{ years} \end{bmatrix}$$

$$\begin{bmatrix} r_{00} \\ \vdots \\ r_{n0} \end{bmatrix}$$

$$\begin{bmatrix} \dots & r_{0n} \\ r_{ij} & \vdots \\ \dots & r_{nn} \end{bmatrix}$$



These stay constant (they describe the samples)

This one changes

Our first format: TPED

$$\begin{bmatrix} id_0 & pheno_0 & male & 22 \text{ years} \\ \vdots & \vdots & \vdots & \vdots \\ id_n & pheno_n & female & 65 \text{ years} \end{bmatrix}$$

FAM/TFAM file

$$\begin{bmatrix} r_{00} & \dots & r_{0j} \\ \vdots & \ddots & \vdots \\ r_{n0} & \dots & r_{nn} \end{bmatrix}$$

Matrix file
(program-specific)

all variants |
$$\begin{bmatrix} A/T \dots G/C \\ \vdots & \vdots & \vdots \\ T/T \dots G/G \end{bmatrix}$$

all individuals

TPED file

Our first format: TPED

$$\begin{bmatrix} id_0 & pheno_0 & male & 22 \text{ years} \\ \vdots & \vdots & \vdots & \vdots \\ id_n & pheno_n & female & 65 \text{ years} \end{bmatrix}$$

- One of PLINK's traditional formats
 - Not used in practice
 - Convenient for looping over SNPs
 - Input `--tfile`
 - Output `--recode transpose`

all variants |
$$\begin{bmatrix} A/T \dots G/C \\ \vdots \quad \vdots \quad \vdots \\ T/T \dots G/G \end{bmatrix}$$

all individuals

TPED file

| | | | | | | |
|---------|---------|---|---|---|----|-----|
| FAMILY1 | SAMPLE1 | 0 | 0 | 1 | 22 | 1.5 |
| FAMILY2 | SAMPLE2 | 0 | 0 | 2 | 65 | 2.1 |

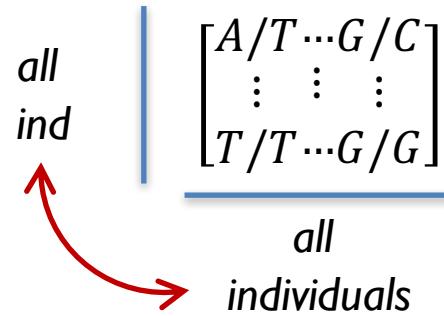
| | | | | | | | |
|---|----------|---|--------|---|---|---|---|
| 1 | rs15933 | 0 | 752721 | A | G | G | G |
| 1 | 1:846808 | 0 | 846808 | C | C | T | C |

Another format: PED/MAP

```
FAMILY1 SAMPLE1 0 0 1 1.5 A G C C
FAMILY2 SAMPLE2 0 0 2 2.1 T T A A
```

FAM/TFAM file

PED file



- One of PLINK's traditional formats
 - Not used in practice
 - Convenient for looping over samples
 - Input --file
 - Output --recode

```
1 rs15933 0 752721
1 1:846808 0 846808
```

MAP file

Exercise 1 : Format conversion

- In /Workshop_data/Lecture2/Exercise1
 - Convert cohort1.tped/tfam to cohort1.ped/map
 - Use the transpose.sh script provided or try to d.i.y

```
FAMILY1 SAMPLE1 0 0 1 1.5 A G C C
FAMILY2 SAMPLE2 0 0 2 2.1 G G T C
```

- Convert the file to PED using PLINK
- Compare both files using diff

Exercise 1 : Solution

- `cut -d' ' -f1-4 cohort1.tped | tr ' ' '\t'> cohort1.map`
- `paste -d' ' cohort1.tfam <(./transpose.sh <(cut -d' ' -f5- cohort1.tped)) > cohort1.ped`
- `plink --tfile cohort1 --recode --out fortest`
- `diff cohort1.ped fortest.ped`

Exercise 2 : Storage

- Consider 3 different genotyping chips
 - 500,000 SNPs (Illumina OmniExpress)
 - 1,000,000 SNPs (ExomeChip)
 - 2,500,000 SNPs (Illumina Onmi 2.5)
- How large is a PED file containing genetic information for 10,000 samples on each of these chips?

Exercise 2 : Storage

- Consider 3 different genotyping chips
 - 500,000 SNPs (Illumina OmniExpress)
 - 1,000,000 SNPs (ExomeChip)
 - 2,500,000 SNPs (Illumina Onmi 2.5)
- 1 character = 1 byte
- Each genotype = 2 alleles + 2 spaces = 4 characters

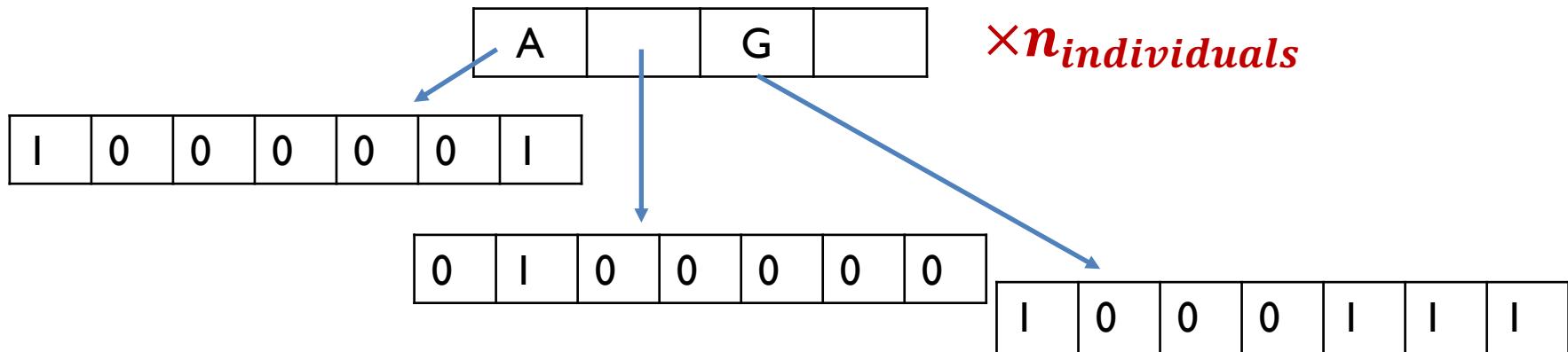
$$n_{SNPs} \times 4 \times n_{individuals} = 19\text{ Gb}, 38\text{ Gb}, 95\text{ Gb}$$

Binary formats

- 1 character = 1 byte
- Each genotype = 2 alleles + 2 spaces = 4 characters
- **Can we make this better?**

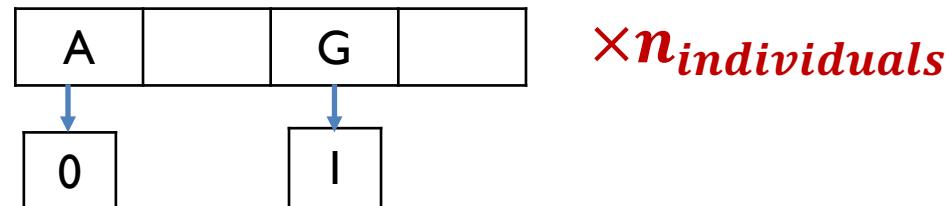
Binary formats

- 1 character = 1 byte
- Each genotype = 2 alleles + 2 spaces = 4 characters
- **Can we make this better?**
- 2 solutions
 - Compress using ZIP/GZIP
 - Use binary formats



Binary formats

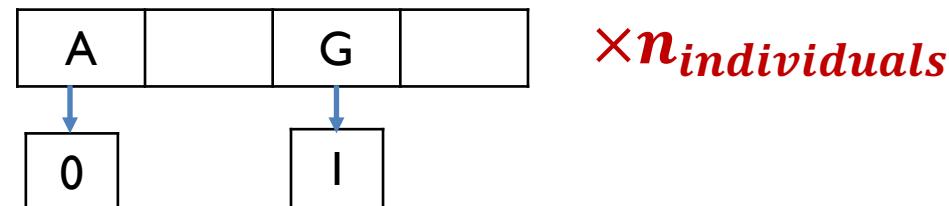
- 1 character = 1 byte
- Each genotype = 2 alleles + 2 spaces = 4 characters
- **Can we make this better?**
- 2 solutions
 - Compress using ZIP/GZIP
 - Use binary formats



- Question: how smaller is the size now?

Binary formats

- 1 character = 1 byte
- Each genotype = 2 alleles + 2 spaces = 4 characters
- **Can we make this better?**
- 2 solutions
 - Compress using ZIP/GZIP
 - Use binary formats



- Question: how smaller is the size now?

$$\frac{n_{SNPs} \times n_{individuals}}{4} = 1.1Gb$$

One (binary) format to rule them all : BED/BIM/FAM

```
FAMILY1 SAMPLE1 0 0 1 22 1.5
FAMILY2 SAMPLE2 0 0 2 65 2.1
```

FAM/TFAM file

```
1 rs15933 0 564862 C T
1 1:752566 0 752566 G A
```

BIM file

```
10101111 10101111 10100010 10111011 10101000 10000000
00101011 00100000 10101000 10001011 00000011 11111111
11111111 11111111 11111111 11111110 11111111 11111111
11111111 11111110 11111110 11111110 11101111 11111111
```

BED file

- Input: --bfile
- Output: --make-bed
- Do not open a BED file with less, cat, head, or tail !
- If you absolutely want to look, xxd -b or od -c

Genotyping data : common operations

Sample management

| | |
|-----------------|------------------------|
| --keep [file] | Keep samples in file |
| --remove [file] | Remove samples in file |

SNP management

| | |
|------------------|---------------------|
| --extract [file] | Keep SNPs in file |
| --exclude [file] | Remove SNPs in file |

Extracting regions

| | |
|-----------------|--------------------------------------|
| --chr [name] | Extract data on specified chromosome |
| --from-bp [pos] | From specified position |
| --to-bp [pos] | To specified position |

Variant QC

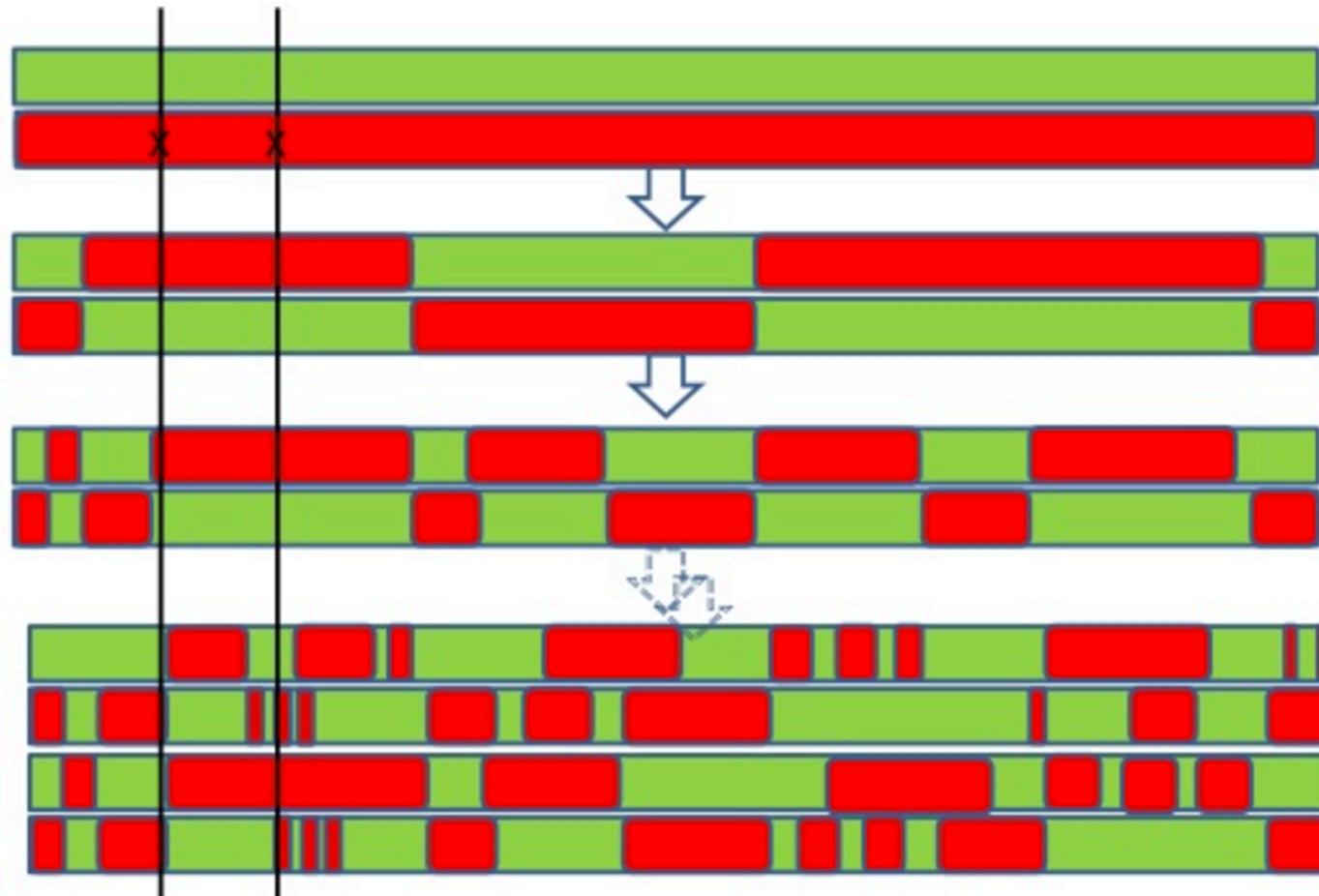
| | |
|---------------------------|---------------------------------------|
| --maf [threshold] | Keep variants with MAF>threshold |
| --hwe midp [threshold] | Keep variants with HWE p>threshold |

Sample QC

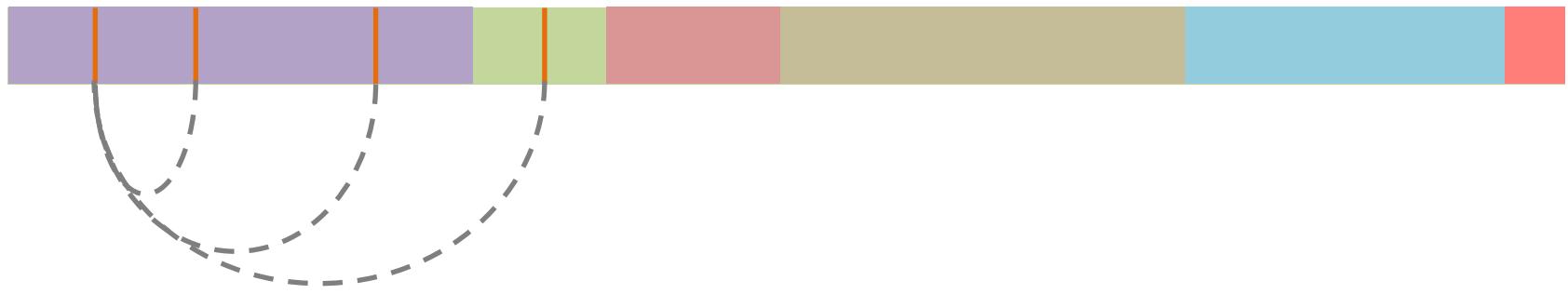
| | |
|-------------|--|
| --missing | Compute per-sample and per-variant missingness |
| --check-sex | Check sexes by looking at chrX |
| --genome | Compute relatedness, check for duplicates |

Linkage disequilibrium

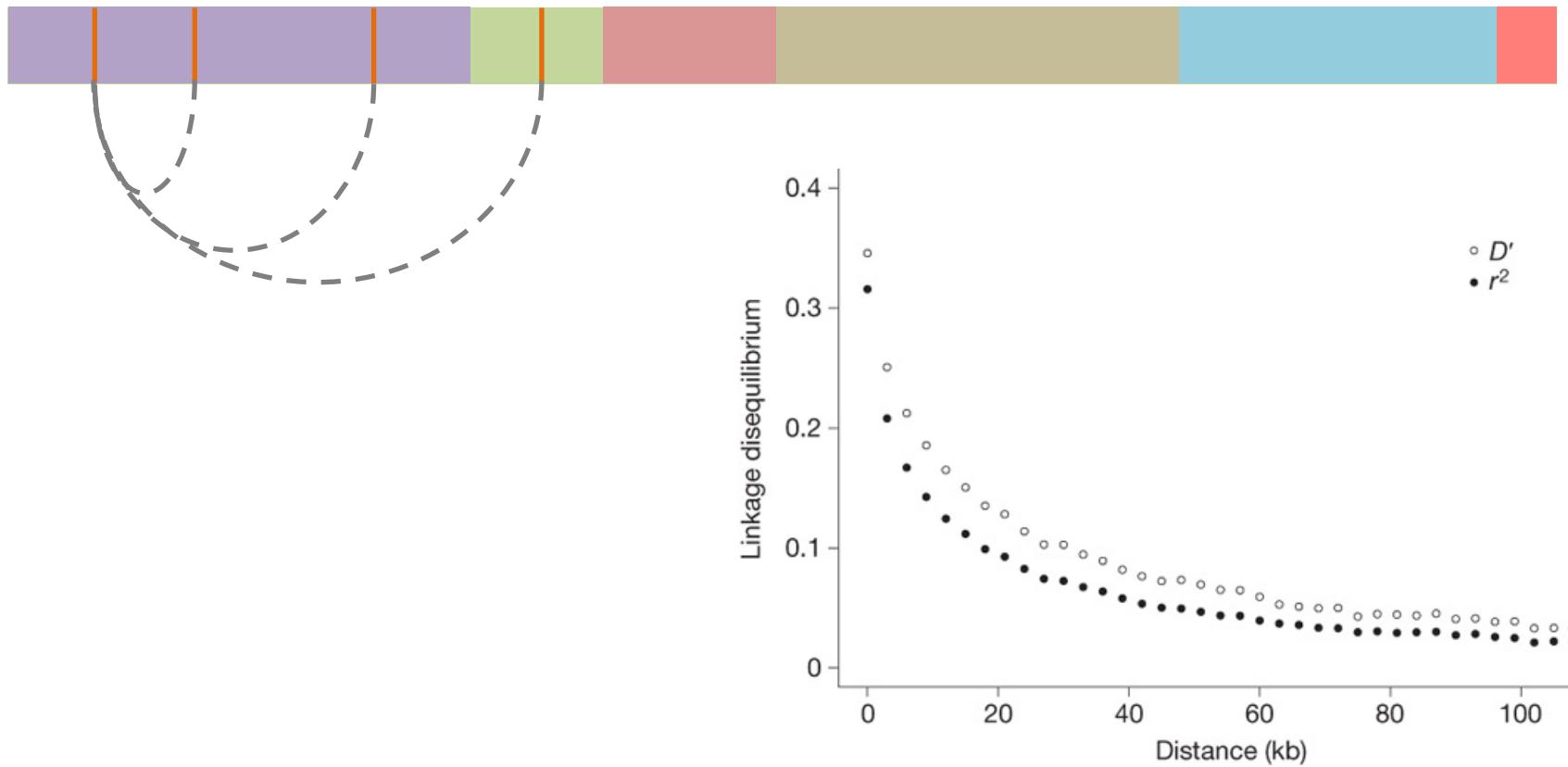
The strange beautiful world of linkage disequilibrium



The strange beautiful world of linkage disequilibrium



The strange beautiful world of linkage disequilibrium



LD between 2 or more SNPs

```
plink --r2 --ld-snps rs1234,rs4567
```

```
plink --r2 --ld-snp-list [file]
```

Pairwise LD in a region

```
plink --r2 --ld-window 10 --ld-window-kb 1000 -  
-ld-window-r2 0.2
```

LD-pruning (only independent SNPs)

```
plink --indep 50 5 2
```

```
plink --indep-pairwise 50 5 0.2
```

```
Plink --indep-pairphase 50 5 0.2
```

Exercise 3 : Stretching the PLINK muscle

- In `/Workshop_data/Lecture2/Exercise3`
 - How many common ($\text{MAF} > 5\%$) variants are there on chromosome 11 in the `cohort1` dataset?
 - How many variants are in LD ($r^2 > 0.4$) with 21:28759840 on chromosome 21 in a 1Mbp window?

Exercise 3 : Stretching the PLINK muscle

```
plink --bfile cohort1 --maf 0.05 --chr 11 --out  
chr11
```

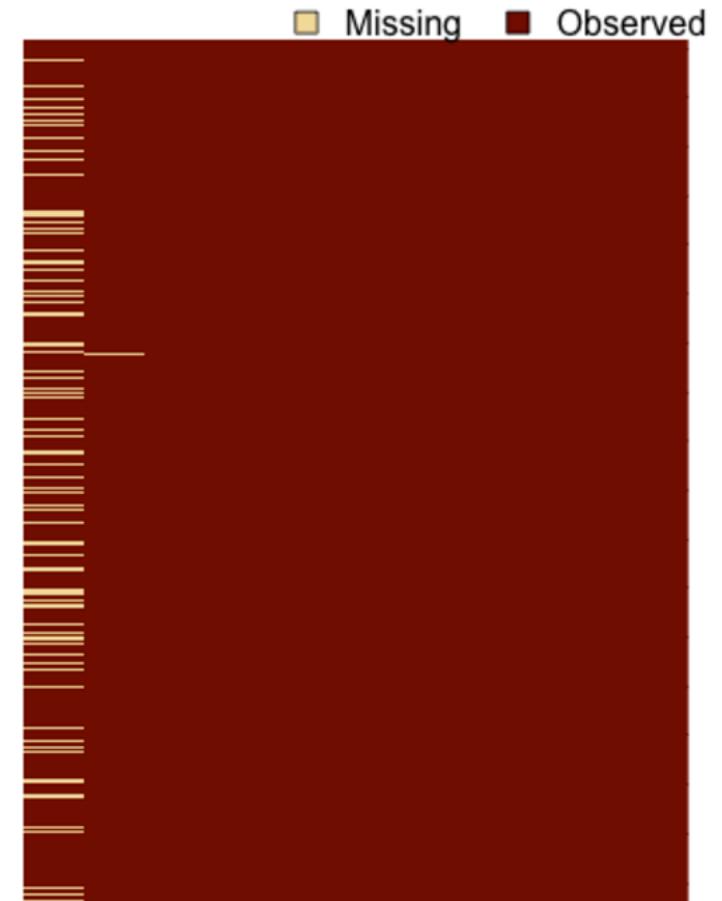
```
wc -l chr11.bim
```

```
plink --bed cohort1.bed --bim cohort1.bim --fam  
cohort1.fam --r2 --ld-snp 21:28759840 --ld-window-kb  
1000000 --ld-window 1000000 -ld-window-r2 0.4
```

```
wc -l plink.ld
```

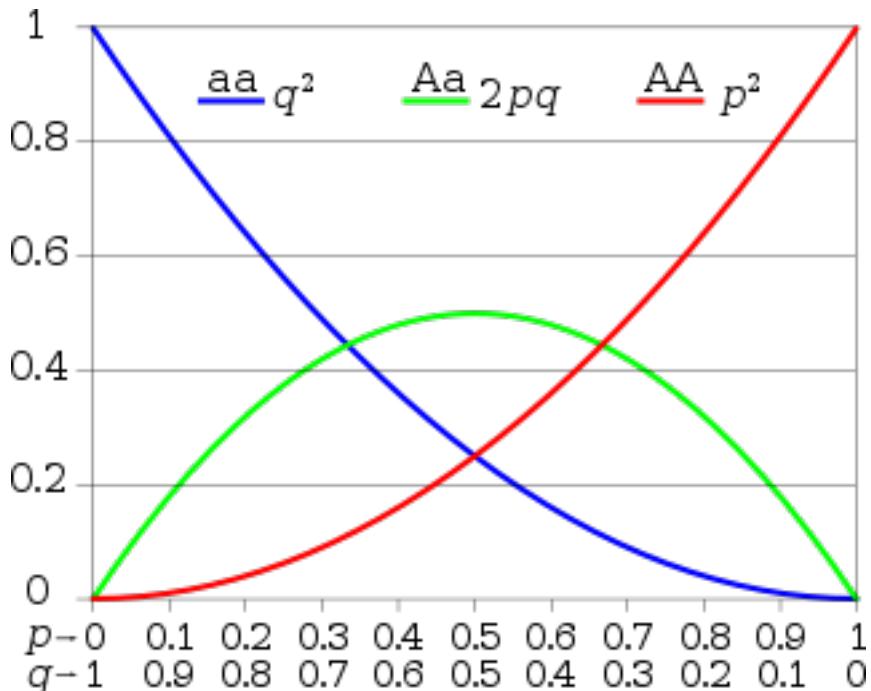
QC steps

Variant QC: which variants do we want to remove?



Variant QC: which variants do we want to remove?

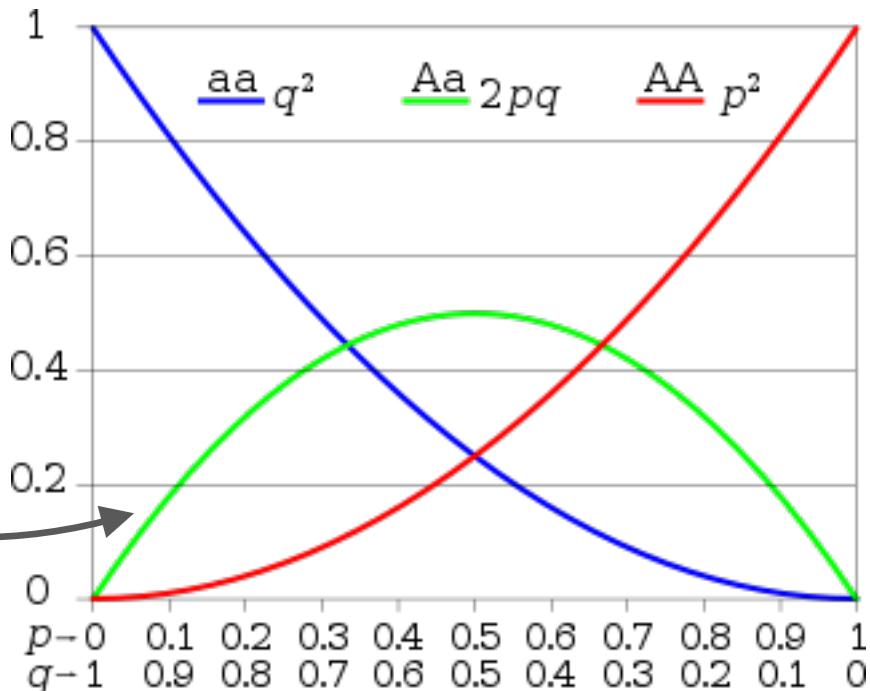
- Calling is not perfect: some genotypes are missing



Variant QC: which variants do we want to remove?

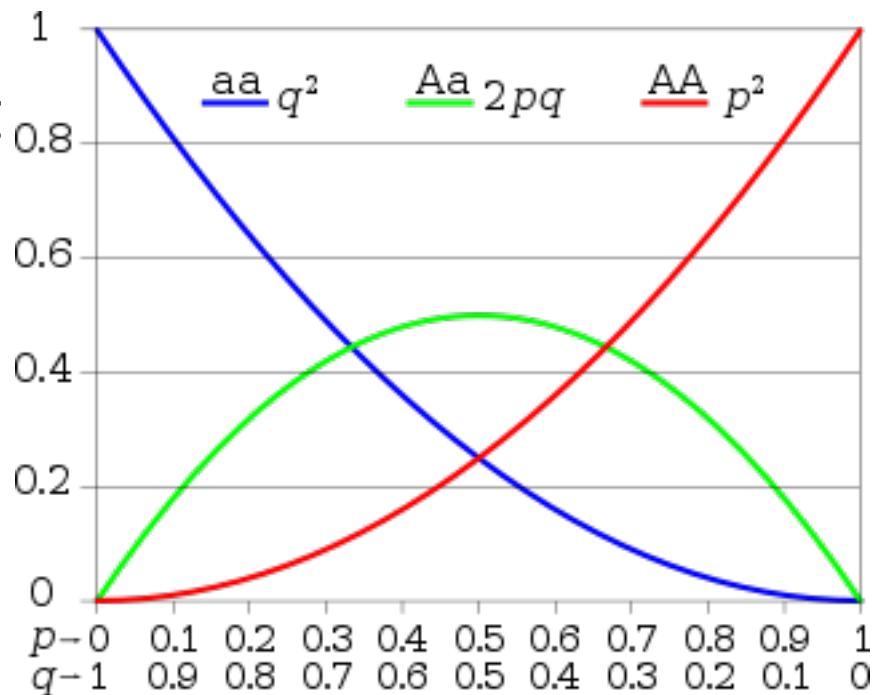
- Calling is not perfect: some genotypes are missing

If we see that 40% of all alleles are a , what is the proportion of aa , Aa , AA ?



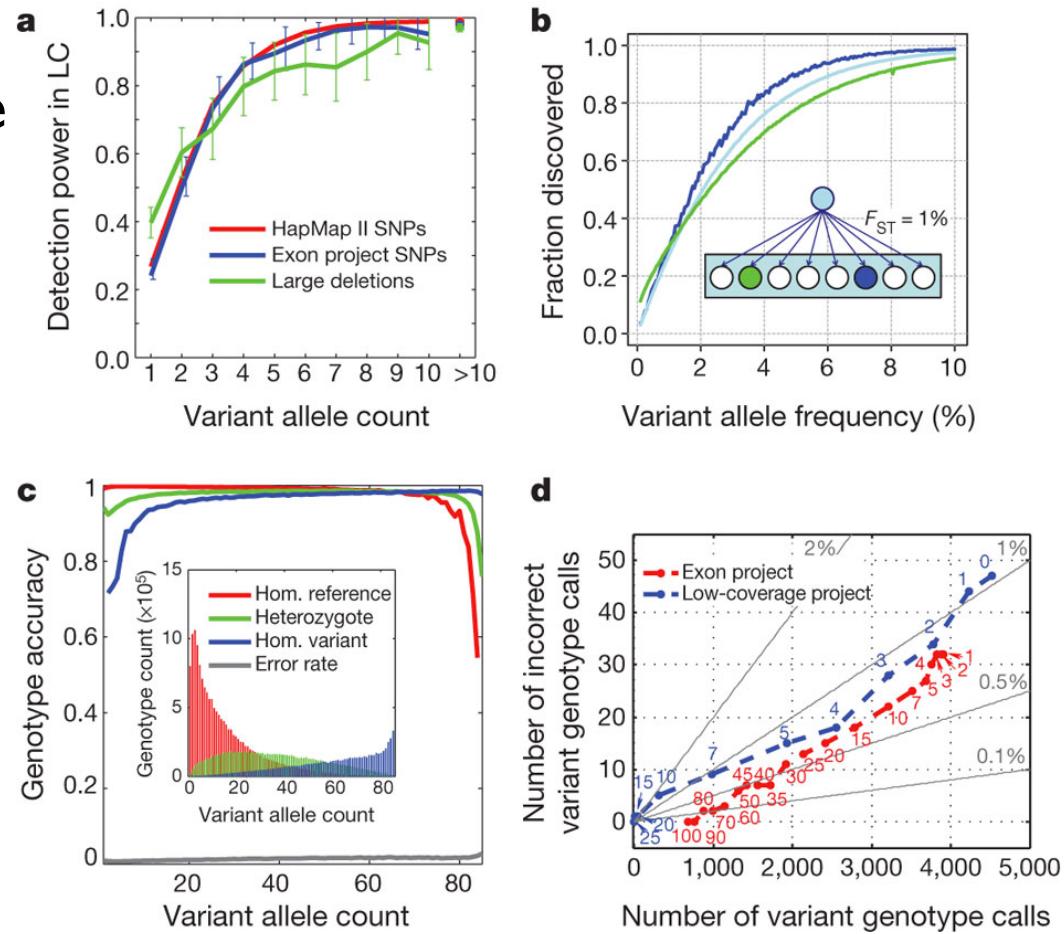
Variant QC: which variants do we want to remove?

- Calling is not perfect: some genotypes are missing
- Variants violating Hardy-Weinberg equilibrium are improbable



Variant QC: which variants do we want to remove?

- Calling is not perfect: some genotypes are missing
- Variants violating Hardy-Weinberg equilibrium are improbable



Variant QC: which variants do we want to remove?

- Calling is not perfect: some genotypes are missing
- Variants violating Hardy-Weinberg equilibrium are improbable
- Rare variants are difficult to call



Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

What are some defining sample characteristics?

Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

What are some defining sample characteristics?



Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

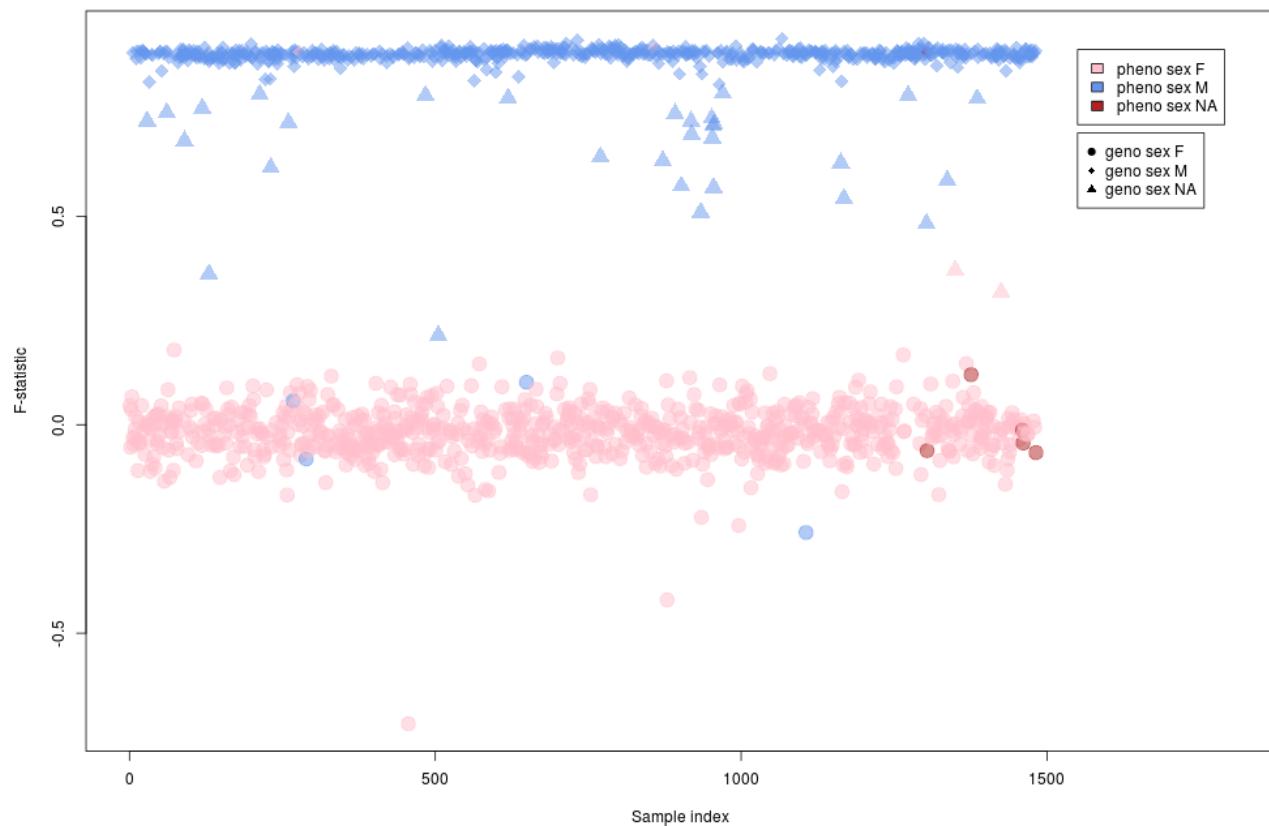
What are some defining sample characteristics?



Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

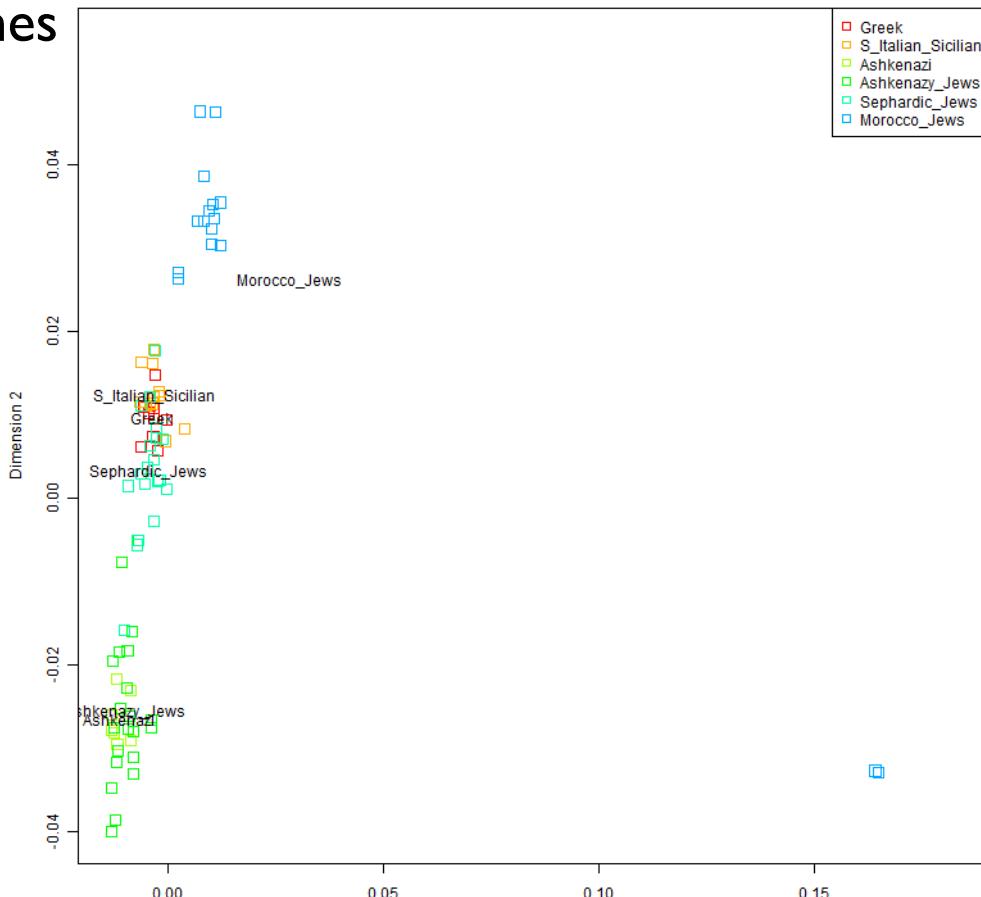
- Sex checks
- Ethnicity checks



Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

- Sex checks
- Ethnicity checks



Sample QC: which individuals do we want to remove?

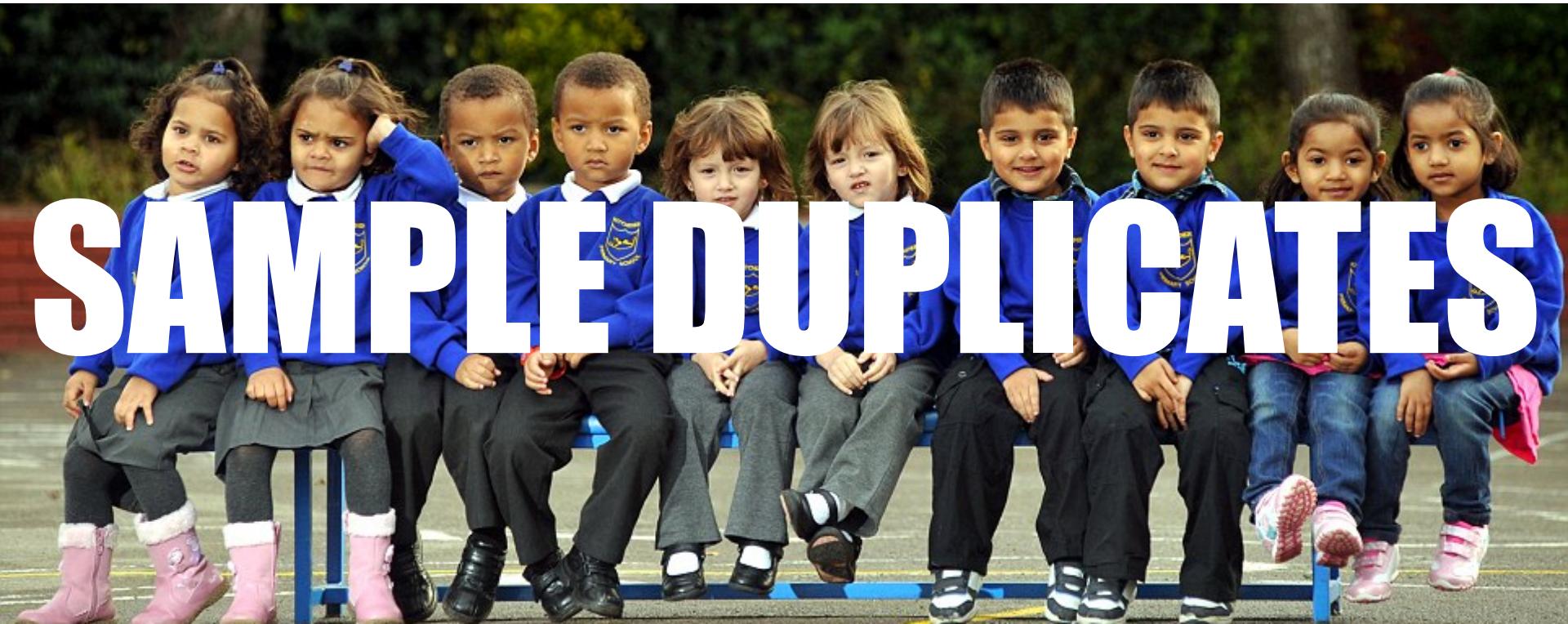
All the different ways in which our samples could be the wrong ones



Sample QC: which individuals do we want to remove?

All the different ways in which our samples could be the wrong ones

SAMPLE DUPLICATES



Sample QC: which individuals do we want to remove?



HETEROZYGOATS