

Data Quality Control (QC) in Association Studies

Svetlana (Sarah) Cherlin

Populatin Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

`svetlana.cherlin@newcastle.ac.uk`





- Poor study design and errors in genotype calling can introduce **systematic bias** in association studies





- Poor study design and errors in genotype calling can introduce **systematic bias** in association studies
 - ▶ increase in **false positive error rate**
 - ▶ decrease in **power**





- Poor study design and errors in genotype calling can introduce **systematic bias** in association studies
 - ▶ increase in **false positive error rate**
 - ▶ decrease in **power**
- Assess data quality **to remove sub-standard genotypes, samples and SNPs** from subsequent association analysis

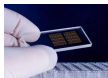




- Poor study design and errors in genotype calling can introduce **systematic bias** in association studies
 - ▶ increase in **false positive error rate**
 - ▶ decrease in **power**
- Assess data quality **to remove sub-standard genotypes, samples and SNPs** from subsequent association analysis
- **Tutorials**
 - ▶ Anderson et al. Nature Protocols 2010, doi:10.1038/nprot.2010.116
 - ▶ Turner et al. Curr Protoc Hum Genet. 2011. doi:10.1002/0471142905.hg0119s68
 - ▶ Marees et al. Int J Methods Psychiatr Res. 2018. doi: 10.1002/mpr.1608

Genotype Calling

illumina

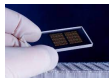


affymetrix

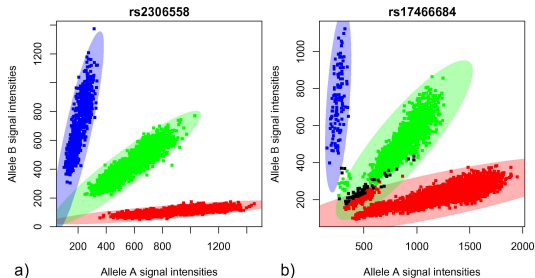


Genotype Calling

illumina



affymetrix



- Examples of cluster plots for two SNPs. One spot corresponds to one sample.
- Samples with genotypes **AA** and **BB** are red and blue, respectively. **Heterozygous** samples are shown in green; samples with **missing genotypes** are black. The ellipses represent the cluster boundaries as computed by ACPA.
- **a)** No samples in overlapping ellipses; **b)** Red samples lie in the green ellipse. At the bottom of the green ellipse, samples have been erroneously classified as red samples.

Genotype Calling

- For large-scale GWA studies, automated genotype calling algorithms have been developed



Genotype Calling

- For large-scale GWA studies, **automated genotype calling algorithms** have been developed
 - ▶ often specific to genotype calling technology
 - ▶ estimate probability or confidence that any specific genotype is AA, AB or BB
 - ▶ apply threshold to probabilities or confidence in order to call genotype, otherwise treated as missing



Genotype Calling

- For large-scale GWA studies, **automated genotype calling algorithms** have been developed
 - ▶ often specific to genotype calling technology
 - ▶ estimate probability or confidence that any specific genotype is AA, AB or BB
 - ▶ apply threshold to probabilities or confidence in order to call genotype, otherwise treated as missing
- Choice of calling **threshold** will impact results



Genotype Calling

- For large-scale GWA studies, **automated genotype calling algorithms** have been developed
 - ▶ often specific to genotype calling technology
 - ▶ estimate probability or confidence that any specific genotype is AA, AB or BB
 - ▶ apply threshold to probabilities or confidence in order to call genotype, otherwise treated as missing
- Choice of calling **threshold** will impact results
 - ▶ too low: include poor quality genotypes
 - ▶ too high: unnecessarily remove high quality genotypes, or may introduce bias by preferentially calling specific genotypes (e.g. rare homozygotes)

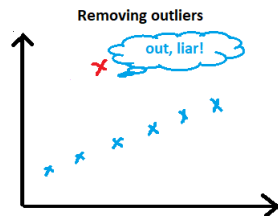


Genotype Calling

- For large-scale GWA studies, **automated genotype calling algorithms** have been developed
 - ▶ often specific to genotype calling technology
 - ▶ estimate probability or confidence that any specific genotype is AA, AB or BB
 - ▶ apply threshold to probabilities or confidence in order to call genotype, otherwise treated as missing
- Choice of calling **threshold** will impact results
 - ▶ too low: include poor quality genotypes
 - ▶ too high: unnecessarily remove high quality genotypes, or may introduce bias by preferentially calling specific genotypes (e.g. rare homozygotes)
- **Missing call rate** is not only a measure of data completeness, but is also a measure of **genotype quality**

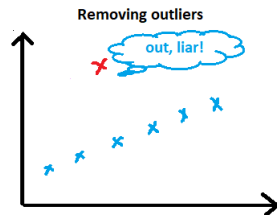


Sample Quality Control



Sample Quality Control

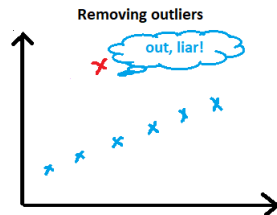
Remove samples on the basis of:



Sample Quality Control

Remove samples on the basis of:

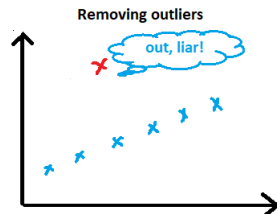
- **Low call rate**
 - ▶ poor DNA quality



Sample Quality Control

Remove samples on the basis of:

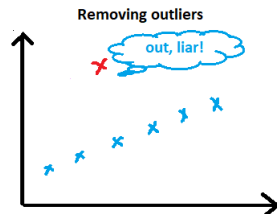
- Low **call rate**
 - ▶ poor DNA quality
- Outlying **heterozygosity** across autosomes
 - ▶ DNA sample contamination or inbreeding



Sample Quality Control

Remove samples on the basis of:

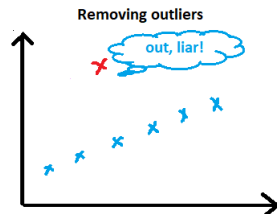
- Low **call rate**
 - ▶ poor DNA quality
- Outlying **heterozygosity** across autosomes
 - ▶ DNA sample contamination or inbreeding
- **Duplication or relatedness** based on identity-by-state or identity-by-descent
 - ▶ if not taken account of in the analysis



Sample Quality Control

Remove samples on the basis of:

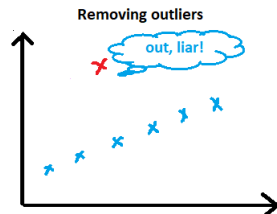
- Low **call rate**
 - ▶ poor DNA quality
- Outlying **heterozygosity** across autosomes
 - ▶ DNA sample contamination or inbreeding
- **Duplication or relatedness** based on identity-by-state or identity-by-descent
 - ▶ if not taken account of in the analysis
- **Mismatches with external information**, i.e. sex discrepancy
 - ▶ sample mix-up



Sample Quality Control

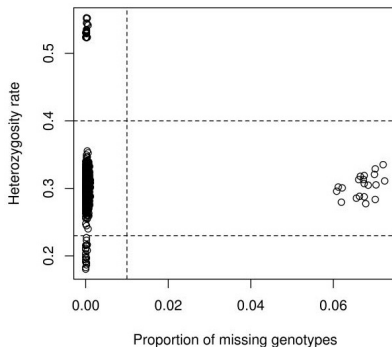
Remove samples on the basis of:

- Low **call rate**
 - ▶ poor DNA quality
- Outlying **heterozygosity** across autosomes
 - ▶ DNA sample contamination or inbreeding
- **Duplication or relatedness** based on identity-by-state or identity-by-descent
 - ▶ if not taken account of in the analysis
- **Mismatches with external information**, i.e. sex discrepancy
 - ▶ sample mix-up
- Outlying population **ancestry**
 - ▶ confounding due to population structure



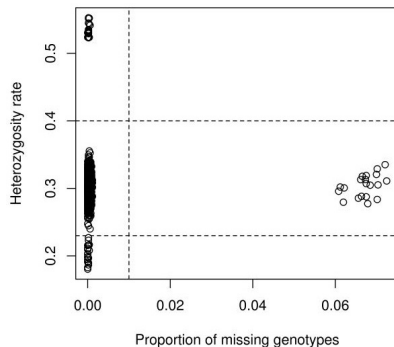
Call Rate and Heterozygosity

- There are samples with high levels of missing data and samples with unusually high and low heterozygosity



Call Rate and Heterozygosity

- There are samples with high levels of **missing data** and samples with unusually high and low **heterozygosity**

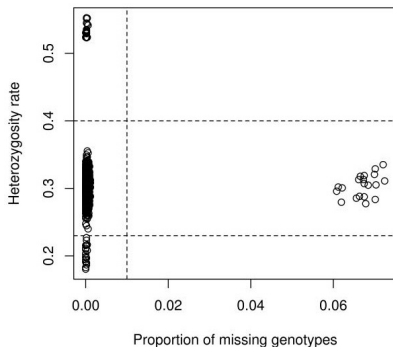


- Decide upon **thresholds** for removing individuals based on the plot



Call Rate and Heterozygosity

- There are samples with high levels of **missing data** and samples with unusually high and low **heterozygosity**

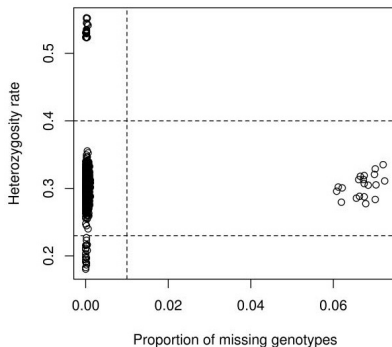


- Decide upon **thresholds** for removing individuals based on the plot
- Dashed lines denote QC thresholds (exclude samples with missing call rate > 0.1 , and samples with heterozygosity rate < 0.23 and > 0.4)



Call Rate and Heterozygosity

- There are samples with high levels of **missing data** and samples with unusually high and low **heterozygosity**



- Decide upon **thresholds** for removing individuals based on the plot
- Dashed lines denote QC thresholds (exclude samples with missing call rate > 0.1 , and samples with heterozygosity rate < 0.23 and > 0.4)
- Rule of thumb:** remove individuals who deviate ± 3 SD from the samples' heterozygosity rate mean



Identity-by-state (IBS)

- Two alleles are **identical by state (IBS)** if they have the same nucleotide sequence



Identity-by-state (IBS)

- Two alleles are **identical by state (IBS)** if they have **the same nucleotide sequence**
- Over M markers, **the IBS** between the individuals i and j is

$$\text{IBS}_{ij} = 1 - \frac{1}{2M} \sum_k |G_{ik} - G_{jk}|,$$

G_{ik} and G_{jk} : the number of minor alleles (0, 1 or 2) carried by the individuals i and j at SNP k



Identity-by-state (IBS)

- Two alleles are **identical by state (IBS)** if they have **the same nucleotide sequence**
- Over M markers, **the IBS** between the individuals i and j is

$$\text{IBS}_{ij} = 1 - \frac{1}{2M} \sum_k |G_{ik} - G_{jk}|,$$

G_{ik} and G_{jk} : the number of minor alleles (0, 1 or 2) carried by the individuals i and j at SNP k

- **Identical** samples will share IBS **near to 100%**
 - ▶ allowing for genotyping errors



Identity-by-state (IBS)

- Two alleles are **identical by state (IBS)** if they have **the same nucleotide sequence**
- Over M markers, **the IBS** between the individuals i and j is

$$\text{IBS}_{ij} = 1 - \frac{1}{2M} \sum_k |G_{ik} - G_{jk}|,$$

G_{ik} and G_{jk} : the number of minor alleles (0, 1 or 2) carried by the individuals i and j at SNP k

- **Identical** samples will share IBS **near to 100%**
 - ▶ allowing for genotyping errors
- **Related** individuals will share **higher IBS** than unrelated individuals



Identity-by-state (IBS)

- Two alleles are **identical by state (IBS)** if they have **the same nucleotide sequence**
- Over M markers, **the IBS** between the individuals i and j is

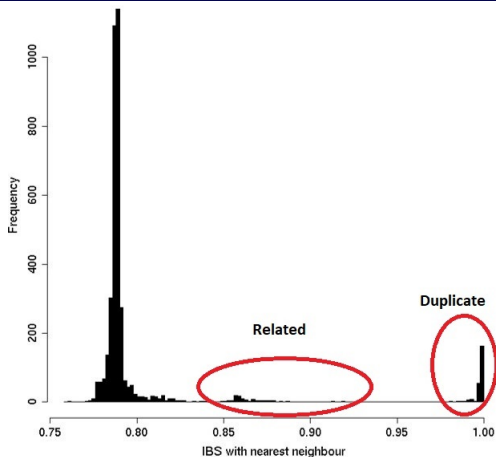
$$\text{IBS}_{ij} = 1 - \frac{1}{2M} \sum_k |G_{ik} - G_{jk}|,$$

G_{ik} and G_{jk} : the number of minor alleles (0, 1 or 2) carried by the individuals i and j at SNP k

- **Identical** samples will share IBS **near to 100%**
 - ▶ allowing for genotyping errors
- **Related** individuals will share **higher IBS** than unrelated individuals
- Common to **plot histogram** of IBS of each individual with “nearest neighbour”



IBS Distribution



- For each individual, the **distance** to its nearest neighbour is calculated
- Remove one sample from each **duplicate or related pair** (usually one with lowest call rate)
- **Alternative:** take account of relatedness in analysis
- The **absolute** amount of IBS sharing depends on allele frequencies in the population
- Methods that estimate kinship or relatedness coefficients typically aim for estimating **identity-by-descent (IBD)**

Alternative Measure of Kinship: Identity-by-descent (IBD)

- The degree of **recent shared ancestry** for a pair of individuals



Alternative Measure of Kinship: Identity-by-descent (IBD)

- The degree of **recent shared ancestry** for a pair of individuals
- The alleles are **identical by descent (IBD)** if they have been inherited from a **common ancestor**



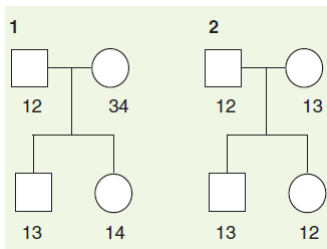
Alternative Measure of Kinship: Identity-by-descent (IBD)

- The degree of **recent shared ancestry** for a pair of individuals
- The alleles are **identical by descent (IBD)** if they have been inherited from a **common ancestor**
- **IBD must also be IBS**, the converse of this statement is not true



Alternative Measure of Kinship: Identity-by-descent (IBD)

- The degree of **recent shared ancestry** for a pair of individuals
- The alleles are **identical by descent (IBD)** if they have been inherited from a **common ancestor**
- **IBD must also be IBS**, the converse of this statement is not true



- Pedigree 1: Siblings share **allele 1 IBD** (inherited from the father)
- Pedigree 2: Siblings share **allele 1 IBS** (inherited from different parents)

Forabosco et al. Expert Rev. Mol. Diagn. 5(5), (2005). doi: 10.1586/14737159.5.5.781



Identity-by-descent (IBD)

- Proportion of the genome at which a pair of individuals share 0, 1 or 2 alleles (Z_0 , Z_1 and Z_2), or **probabilities of sharing 0, 1 and 2 alleles**



Identity-by-descent (IBD)

- Proportion of the genome at which a pair of individuals share 0, 1 or 2 alleles (Z_0 , Z_1 and Z_2), or **probabilities of sharing 0, 1 and 2 alleles**
- **Mean IBD sharing** $\hat{\pi} = 0 \times Z_0 + 1 \times Z_1 + 2 \times Z_2$



Identity-by-descent (IBD)

- Proportion of the genome at which a pair of individuals share 0, 1 or 2 alleles (Z_0 , Z_1 and Z_2), or **probabilities of sharing 0, 1 and 2 alleles**
- **Mean IBD sharing** $\hat{\pi} = 0 \times Z_0 + 1 \times Z_1 + 2 \times Z_2$
- **Expected patterns of mean IBD** for known related pairs
 - ▶ IBD $\hat{\pi} = 1$ for **duplicates or monozygotic twins**
 - in practice, use $\hat{\pi} > 0.98$
 - ▶ IBD $\hat{\pi} = 0.5$ for **first-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.25$ for **second-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.125$ for **third-degree relatives**



Identity-by-descent (IBD)

- Proportion of the genome at which a pair of individuals share 0, 1 or 2 alleles (Z_0 , Z_1 and Z_2), or **probabilities of sharing 0, 1 and 2 alleles**
- **Mean IBD sharing** $\hat{\pi} = 0 \times Z_0 + 1 \times Z_1 + 2 \times Z_2$
- **Expected patterns of mean IBD** for known related pairs
 - ▶ IBD $\hat{\pi} = 1$ for **duplicates or monozygotic twins**
 - in practice, use $\hat{\pi} > 0.98$
 - ▶ IBD $\hat{\pi} = 0.5$ for **first-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.25$ for **second-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.125$ for **third-degree relatives**
- **Remove one from each pair** with $\hat{\pi} > 0.185$
 - ▶ halfway between the expected IBD for third- and second-degree relatives

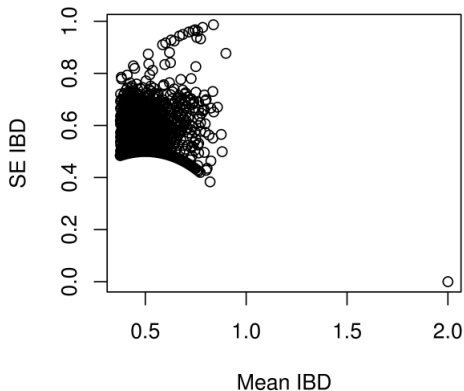


Identity-by-descent (IBD)

- Proportion of the genome at which a pair of individuals share 0, 1 or 2 alleles (Z_0 , Z_1 and Z_2), or **probabilities of sharing 0, 1 and 2 alleles**
- **Mean IBD sharing** $\hat{\pi} = 0 \times Z_0 + 1 \times Z_1 + 2 \times Z_2$
- **Expected patterns of mean IBD** for known related pairs
 - ▶ IBD $\hat{\pi} = 1$ for **duplicates or monozygotic twins**
 - in practice, use $\hat{\pi} > 0.98$
 - ▶ IBD $\hat{\pi} = 0.5$ for **first-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.25$ for **second-degree relatives**
 - ▶ IBD $\hat{\pi} = 0.125$ for **third-degree relatives**
- **Remove one from each pair** with $\hat{\pi} > 0.185$
 - ▶ halfway between the expected IBD for third- and second-degree relatives
- **Prune the data for LD** before assessing IBD
 - ▶ shared region of high LD results in more shared variants than one of low LD, even if the two regions are the same size



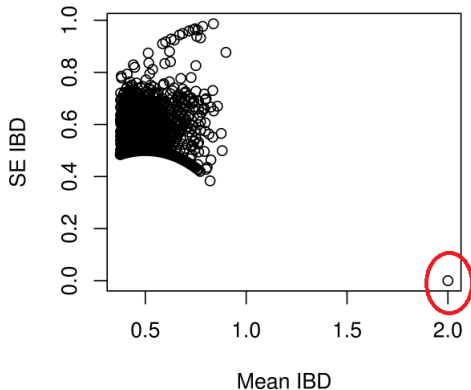
IBD Plot



Spot the duplicates...



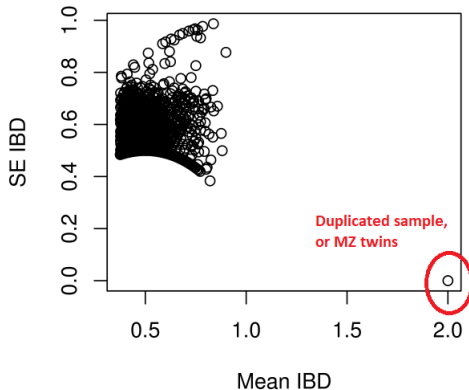
IBD Plot



Spot the duplicates...



IBD Plot



Spot the duplicates...



More on Relatedness

- Traditional approaches to association analysis assume that individuals are unrelated to each other



More on Relatedness

- **Traditional approaches** to association analysis assume that individuals are **unrelated** to each other
- In practice ...

MZ twins



Duplicated samples



DZ twins



Cryptic relatedness



More on Relatedness

- **Traditional approaches** to association analysis assume that individuals are **unrelated** to each other
- In practice ...

MZ twins



Duplicated samples



DZ twins



Cryptic relatedness



- Including related individuals in the analysis, without accounting for these relationships, can increase **false positive error rates** and reduce **power**



More on Relatedness

- **Traditional approaches** to association analysis assume that individuals are **unrelated** to each other
- In practice ...

MZ twins



Duplicated samples



DZ twins



Cryptic relatedness



- Including related individuals in the analysis, without accounting for these relationships, can increase **false positive error rates** and reduce **power**
- **Mixed modelling approaches** account for “relatedness” between individuals (families, cryptic relatedness, population structure) by allowing for kinship



Gender Check - X Chromosome



- Calculate the **homozygosity rate F** across all X-chromosome SNPs for each individual in the sample and compare these to the expected rate



Gender Check - X Chromosome



- Calculate the **homozygosity rate F** across all X-chromosome SNPs for each individual in the sample and compare these to the expected rate
- **Expected homozygosity rates**
 - ▶ $F > 0.8$ for male samples and $F < 0.2$ for female samples
 - ▶ males have only one X - cannot be heterozygous
 - expect some genotyping errors



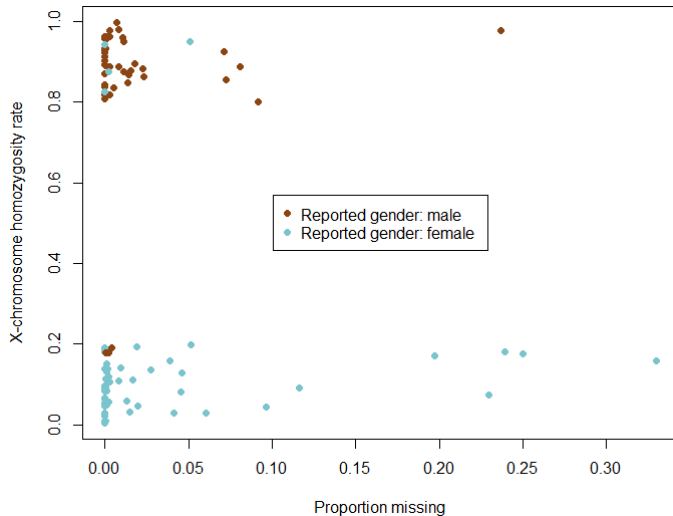
Gender Check - X Chromosome



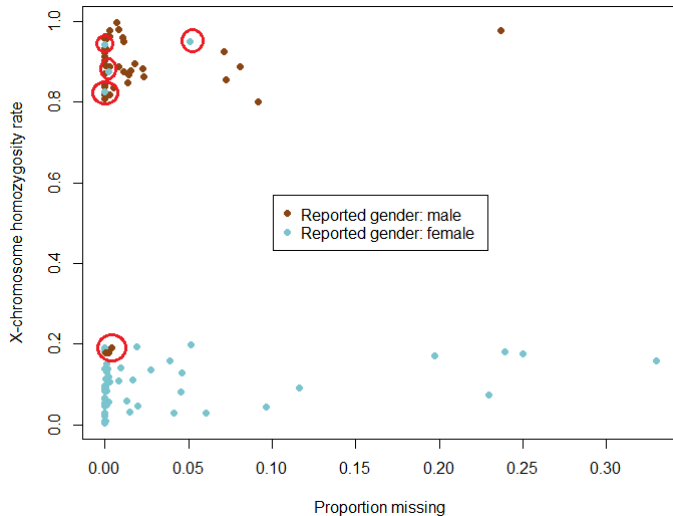
- Calculate the **homozygosity rate F** across all **X-chromosome SNPs** for each individual in the sample and compare these to the expected rate
- **Expected homozygosity rates**
 - ▶ $F > 0.8$ for male samples and $F < 0.2$ for female samples
 - ▶ males have only one X - cannot be heterozygous
 - expect some genotyping errors
- Gender error reported for **mismatch in reported and genetic sex**
- **Discrepancies** with external gender information may reflect:
 - ▶ **errors** in external data
 - ▶ sample **mix-up**



Gender Check - Plot



Gender Check - Plot



Spot the discrepancies...



Gender Check - Examples

Reported gender	Homozygosity rate	Gender according to SNPs
Male	0.98	Male
Female	0.03	Female
Female	0.99	Male
Female	0.28	Unknown*
Female	0.35	Unknown**

* Likely a female with sex chromosome anomaly (e.g. XX/XO mosaic, loss-of-heterozygosity on X)

** Likely a male with sex chromosome anomaly (e.g. XXY or XX/XY mosaic)

Adapted from Turner et al. Curr Protoc Hum Genet, (2011). doi:10.1002/0471142905.hg0119s68



Remove SNPs on the basis of:



Remove SNPs on the basis of:

- Low **call rate** (95% or 99%), variable by MAF
 - ▶ poor quality SNP



Remove SNPs on the basis of:

- Low **call rate** (95% or 99%), variable by MAF
 - ▶ poor quality SNP
- Extreme deviation from **Hardy-Weinberg equilibrium**
 - ▶ in cases, controls or both for autosomes
 - ▶ incompatible with the assumption of random mating
 - ▶ possible genotyping error



Remove SNPs on the basis of:

- Low **call rate** (95% or 99%), variable by MAF
 - ▶ poor quality SNP
- Extreme deviation from **Hardy-Weinberg equilibrium**
 - ▶ in cases, controls or both for autosomes
 - ▶ incompatible with the assumption of random mating
 - ▶ possible genotyping error
- Low **frequency** SNPs (MAF 1-5%)
 - ▶ difficult to call using current genotype calling algorithms
 - ▶ possible genotyping error
 - ▶ association signals seen at these rare SNPs are underpowered



Remove SNPs on the basis of:

- Low **call rate** (95% or 99%), variable by MAF
 - ▶ poor quality SNP
- Extreme deviation from **Hardy-Weinberg equilibrium**
 - ▶ in cases, controls or both for autosomes
 - ▶ incompatible with the assumption of random mating
 - ▶ possible genotyping error
- Low **frequency** SNPs (MAF 1-5%)
 - ▶ difficult to call using current genotype calling algorithms
 - ▶ possible genotyping error
 - ▶ association signals seen at these rare SNPs are underpowered
- **Study specific** SNP QC filters
 - ▶ differences in allele frequencies between multiple control cohorts



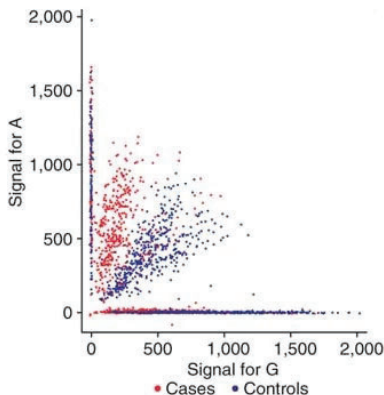
Remove SNPs on the basis of:

- Low **call rate** (95% or 99%), variable by MAF
 - ▶ poor quality SNP
- Extreme deviation from **Hardy-Weinberg equilibrium**
 - ▶ in cases, controls or both for autosomes
 - ▶ incompatible with the assumption of random mating
 - ▶ possible genotyping error
- Low **frequency** SNPs (MAF 1-5%)
 - ▶ difficult to call using current genotype calling algorithms
 - ▶ possible genotyping error
 - ▶ association signals seen at these rare SNPs are underpowered
- **Study specific** SNP QC filters
 - ▶ differences in allele frequencies between multiple control cohorts
- Extreme **differential call rates between cases and controls**



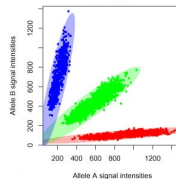
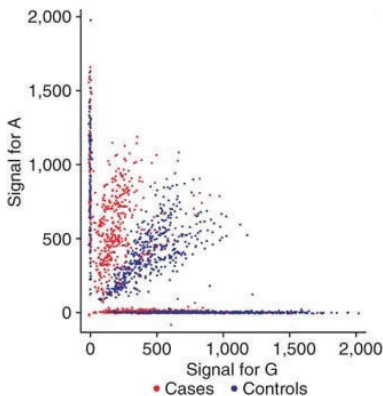
Effect of Differential Call Rate

Clayton et al. Nature Genetics, (2005). doi:10.1038/ng1653



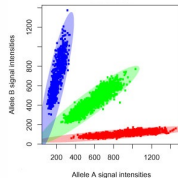
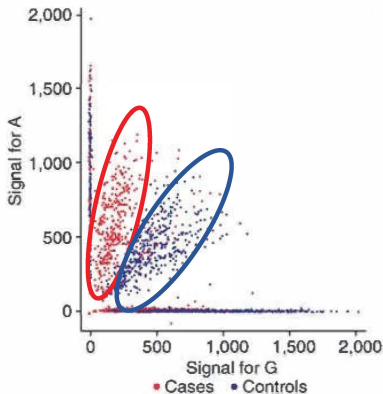
Effect of Differential Call Rate

Clayton et al. Nature Genetics, (2005). doi:10.1038/ng1653



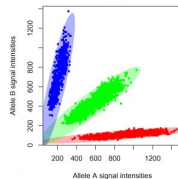
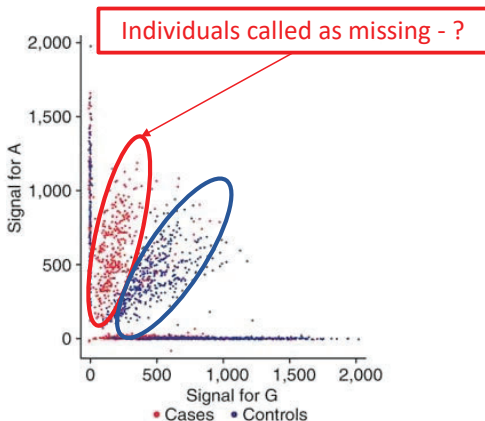
Effect of Differential Call Rate

Clayton et al. Nature Genetics, (2005). doi:10.1038/ng1653



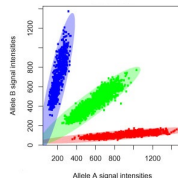
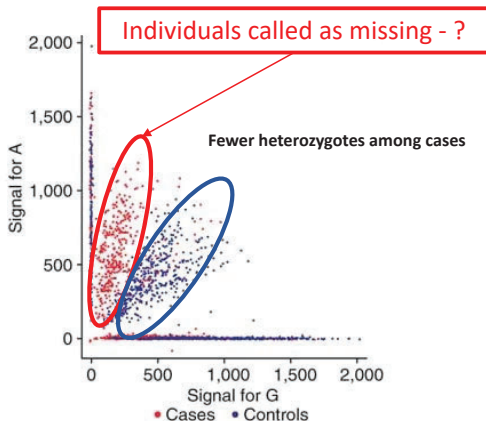
Effect of Differential Call Rate

Clayton et al. Nature Genetics, (2005). doi:10.1038/ng1653



Effect of Differential Call Rate

Clayton et al. Nature Genetics, (2005). doi:10.1038/ng1653



Software for Quality Control



- Specialised **software** for quality control of genome-wide association studies that can handle **scale and complexity of data**



Software for Quality Control



- Specialised **software** for quality control of genome-wide association studies that can handle **scale and complexity of data**
- **QCTOOL**: flexible command line software with range of filtering options



Software for Quality Control



- Specialised **software** for quality control of genome-wide association studies that can handle **scale and complexity of data**
- **QCTOOL**: flexible command line software with range of filtering options
- **PLINK**: whole-genome association analysis toolset



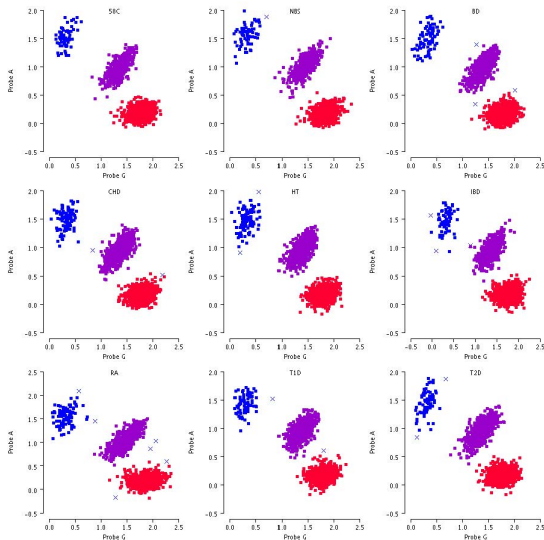
Software for Quality Control



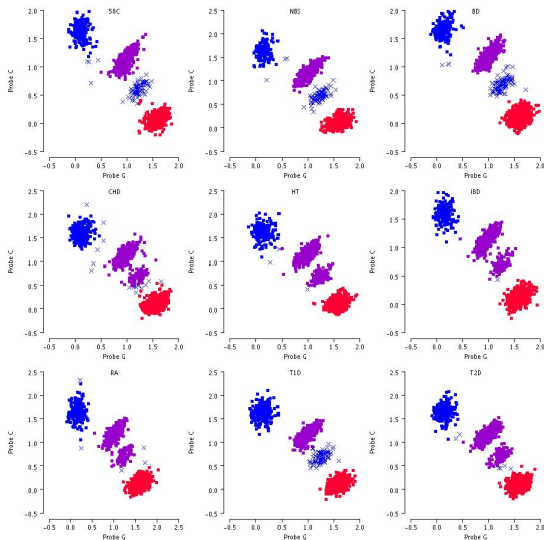
- Specialised **software** for quality control of genome-wide association studies that can handle **scale and complexity of data**
- **QCTOOL**: flexible command line software with range of filtering options
- **PLINK**: whole-genome association analysis toolset
 - ▶ **Spoiler alert**: PLINK will be used in practicals



Visual Inspection of Cluster Plots - Good SNP



Visual Inspection of Cluster Plots - Bad SNP





© CanStockPhoto.com

- QC is an **essential step** of the analysis
- QC criteria are **subjective** and vary from one study to another
- Sample QC filters should **not be so stringent** as to remove the majority of the analysis cohort
- SNP QC filters should **eliminate the worst quality markers**
- All SNPs demonstrating evidence for association should be followed up with **visual inspection** of cluster plots

