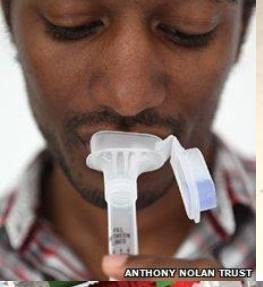
Preparing data for GWAS analysis: Quality Control

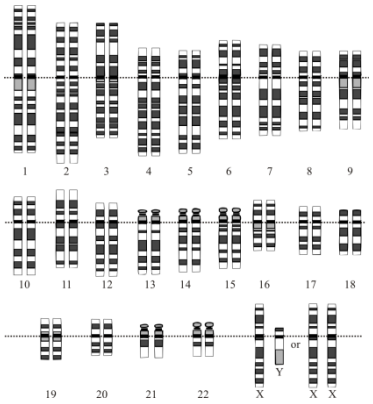
Why QC ?

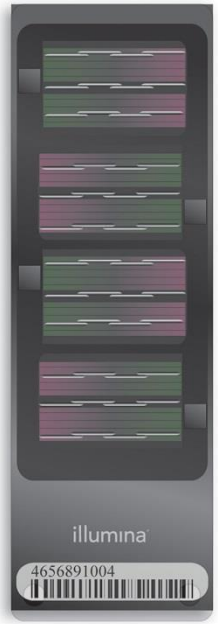
- The capability of GWAS to identify true genetic association depends upon the overall quality of the data.

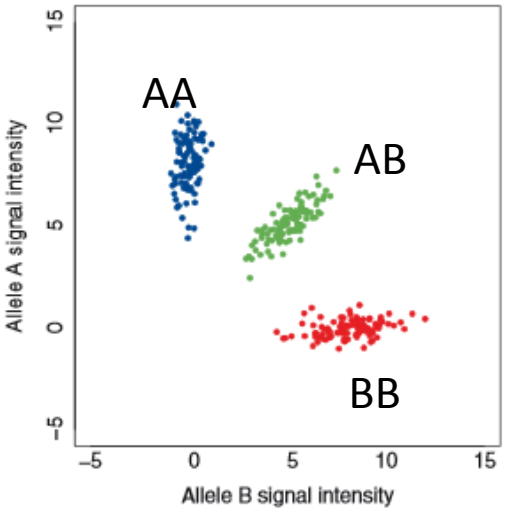
- The ultimate purpose is to minimize potential bias and error in GWAS results

- Our Objective of carrying our QC procedure: To identify samples and SNPs of poor quality or questionable identity





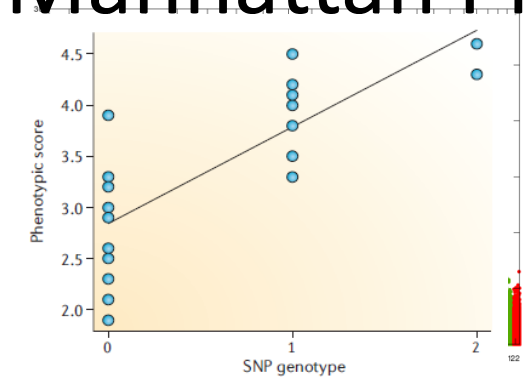




?

Quality Control
QC

Manhattan Plot



Genotype data

	SNP A	SNP B	SNP C	SNP D	SNP E
Female 1	**00**	AG	GG	GA	**00**
Male 1	**00**	GG	GG	AA	CC
Female 2	AC	**00**	GG	AA	CC
Female 3	AA	AG	GC	AA	CC
Male 2	AC	AA	**00**	AA	CA

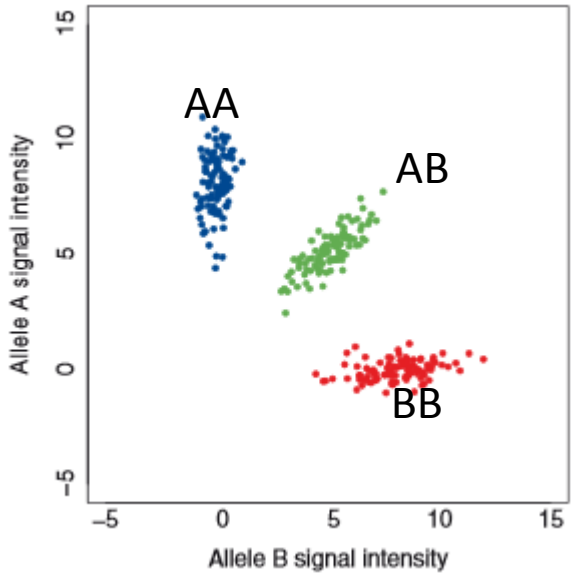
00 = missing data

Adopted Sequential steps

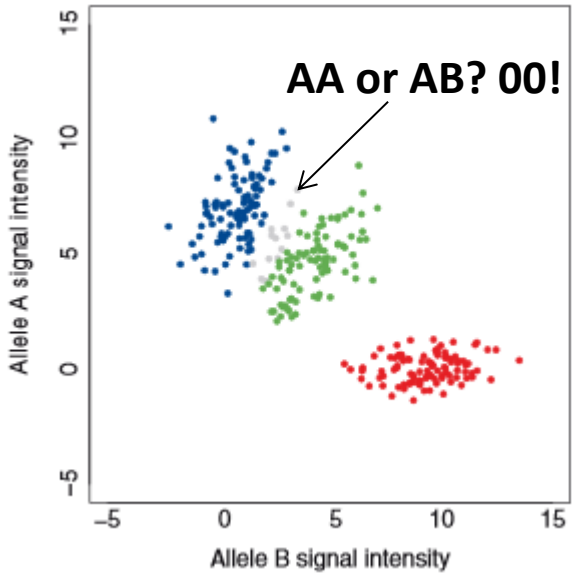
Steps	Sample QC	SNPs QC
1	Sample Call Rate	-
2	Heterozygosity	-
3	Sex	-
4	-	SNP call rate
5	IBD (Identity-by-descent)	-
6		HWE(Hardy-Weinberg equilibrium)
7		
8		

How did we get the genotype data? Genotype Calling

Good data
SNP1



Bad data
SNP2



Sample QC and SNP QC

SNP QC

sample QC

	SNP A	SNP B	SNP C	SNP D	SNP E
Female 1	**00**	AG	GG	GA	**00**
Male 1	**00**	GG	GG	AA	CC
Female 2	AC	**00**	GG	AA	CC
Female 3	AA	AG	GC	AA	CC
Male 2	AC	AA	**00**	AA	CA

00 = missing data

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

Hardy Weinberg Equilibrium (HWE)

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

Hardy Weinberg Equilibrium (HWE)

Missing call rate - Sample

- Missing call rate is the fraction of missing calls per SNP over sample

	SNP1	SNP2	SNP3	SNP4	SNP5
Sample1	**00**	AG	GG	GA	**00**
Sample2	**C0**	TC	CT	TT	CC
Sample3	AC	**00**	CC	CA	AA
Sample4	AT	TA	TT	**00**	AA
Sample5	CG	CC	**00**	GC	GG

How much samples with Missing call rate should be used

-97% call rate was used in (WTCCC (2007)

- 95% was used in(Laurie *et al.,* 2010)

Sample Call Rate/Proportion

	SNP1	SNP2	SNP3	SNP4	SNP5	**Sample Call Rate**
Sample1	**00**	AG	GG	GA	**00**	60%
Sample2	**00**	GG	GG	AA	CC	80%
Sample3	AC	**00**	GG	AA	CC	80%
Sample4	AA	AG	GC	AA	CC	100%
Sample5	AC	AA	**00**	AA	CA	80%

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

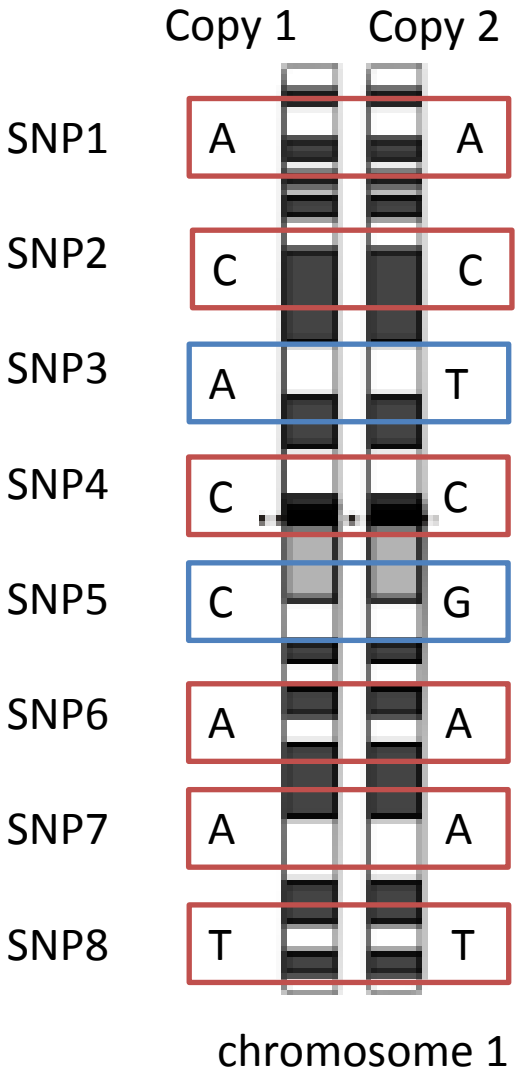
Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

Hardy Weinberg Equilibrium (HWE)

Definition of Heterozygosity Rate



homozygous

heterozygous

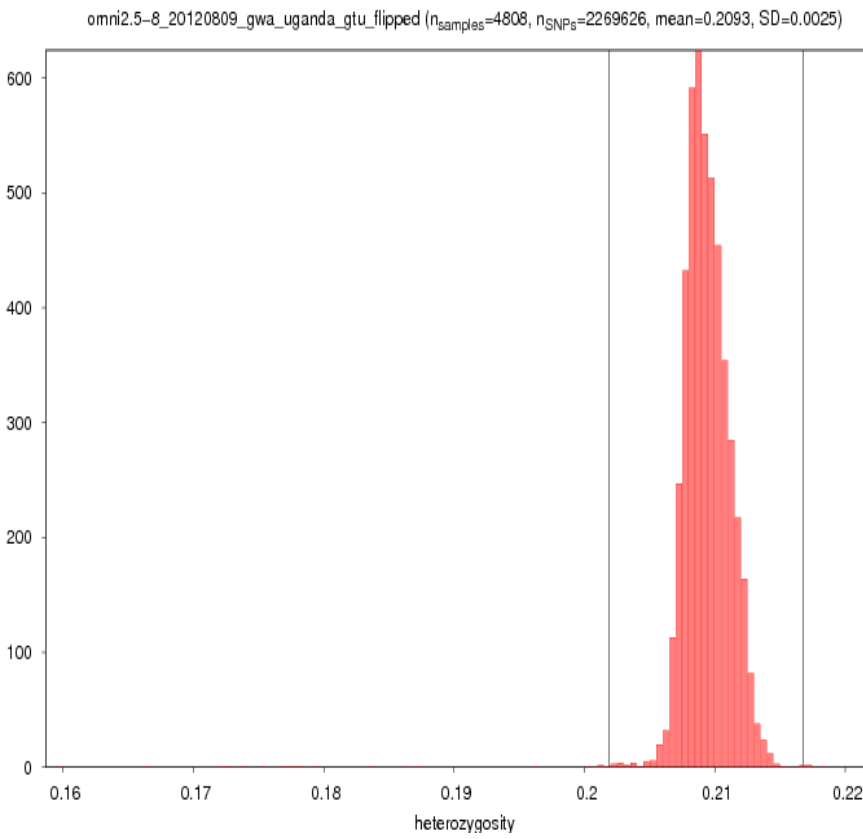
heterozygosity = 2/8 = 0.25

Heterozygosity

- The heterozygosity for each samples is computed as the ratio of the number of heterozygote genotype calls to the total number of non-missing calls.

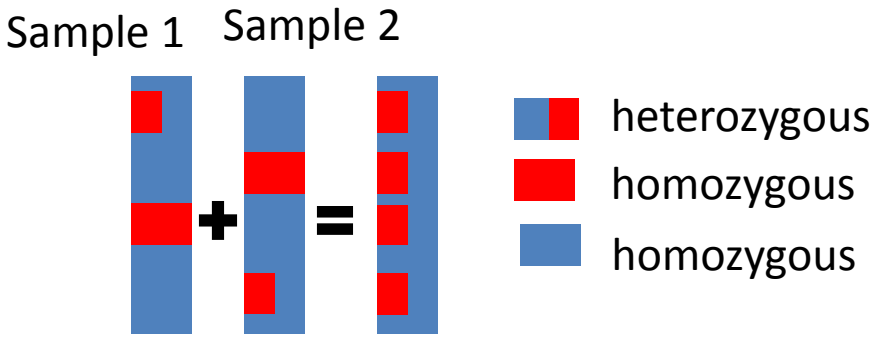
- Heterozygosity = [N(NM)-O(HOM)]/N(NM

- The samples are flagged if their heterozygosity is too low or too high, both on the absolute and the relative scale



Heterozygosity

- Remove samples deviating from average
- Deviations could arise due to several reasons
  - Contamination of samples (high heterozygosity)
  - Inbreeding (low heterozygosity)
  - Ancestral differences
  - Data quality / Poor genotype calling
    - Heterozygotes more likely to be missing



Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

Hardy Weinberg Equilibrium (HWE)

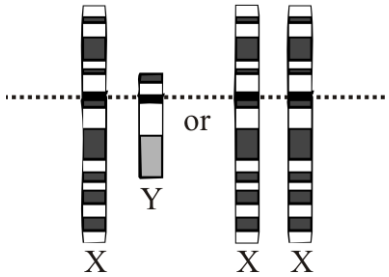
Gender Checks

- Using X and Y Chromosome, it is easy to spot individual who are genetically male but are phenotypically labelled as female or vice versa

- Carry out gender checks on X chromosome (F>0.8 male, Inbreeding coeffient F <0.2 female)

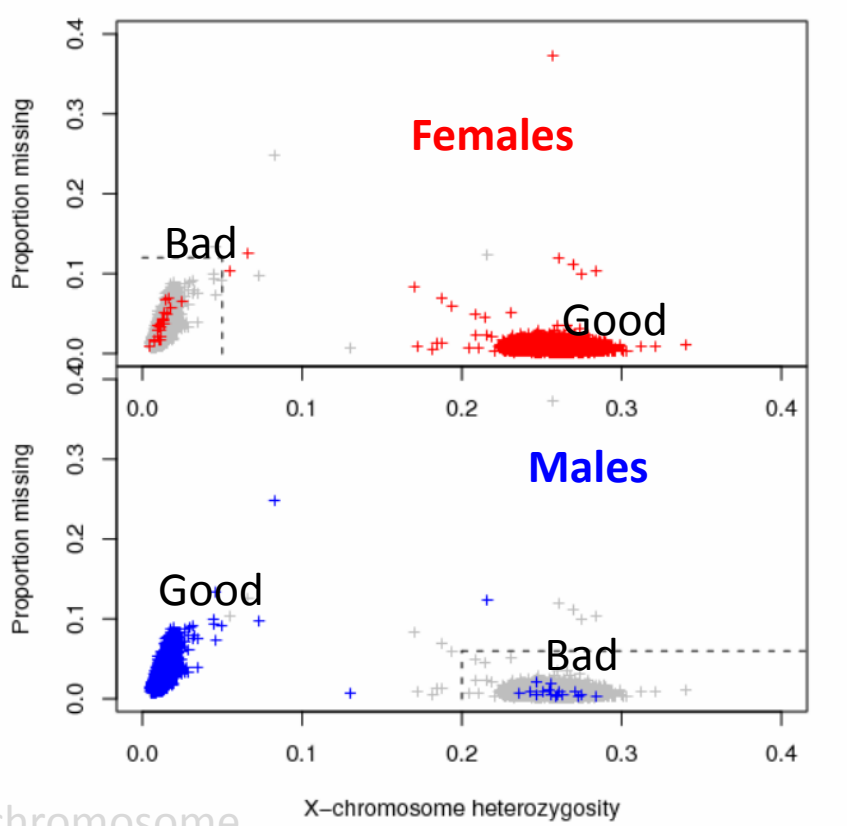
FID	Supplied gender	Inferred gender	F
216767_H11_APP5211886	M	F	0.0567
216768_E11_APP5212097	M	0	0.7943
232626_A02_APP5292757	M	F	0.1661
232628_A09_APP5292795	F	0	0.5810

Sex check
Looking for mislabelled samples



Male
1 allele

Female
2 alleles



Crossover between the X and Y chromosome happens between pseudoautosomal regions. SNPs in PARs are thus excluded from analysis.

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

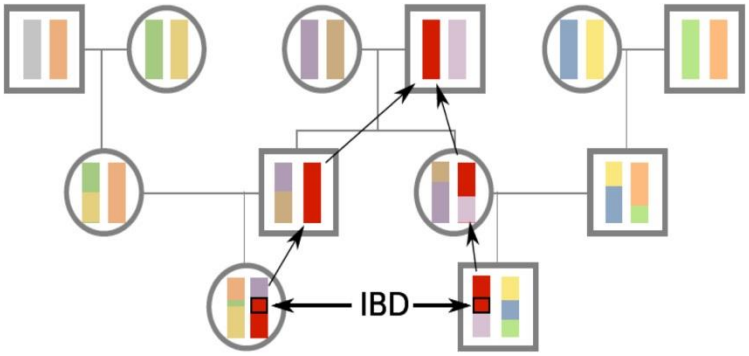
Hardy Weinberg Equilibrium (HWE)

Relatedness

- Relatedness is a problem because of overrepresentation of selected alleles, which will bias any multivariate analysis (correlated data!); e.g. PCA or multivariate regression
- Related samples need to be excluded or taken into account during subsequent analyses
- One metric of relatedness is Identity By Descent (**IBD**), which involves calculation of proportion of common alleles between two individuals.
- Prior to the calculation of IBD, SNPs with a low call rate are permanently excluded and rare SNPs (MAF<5%) and SNPs in Linkage Disequilibrium (**LD**) are temporarily excluded.

Relatedness / IBD

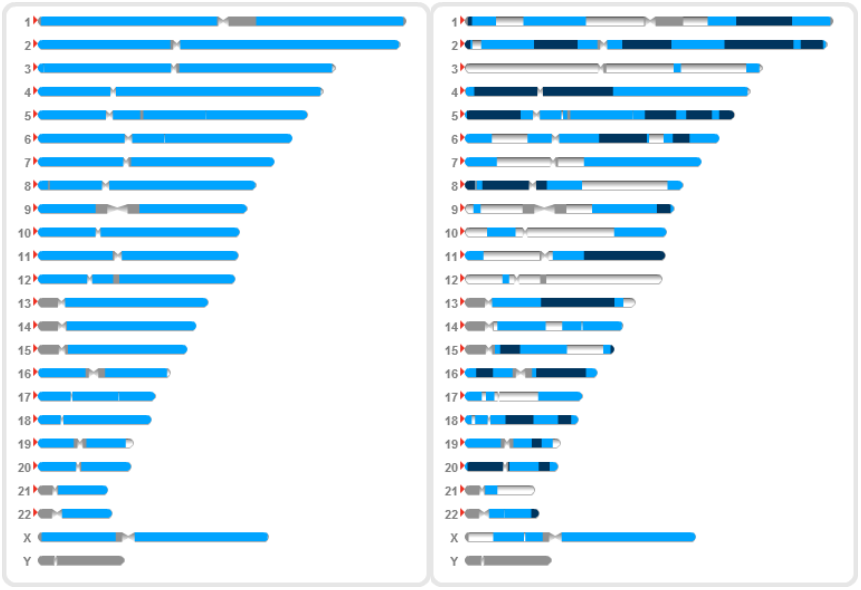
Relationship category	Relatedness
Monozygotic twins	1
Parent-Offspring	**1/2**
Full siblings	**1/2**
Grandparent-grandchild	**1/4**
Uncle/Aunt-Nephew/Niece	**1/4**
First cousins	**1/8**
Unrelated	0



Completely identical

Half-identical

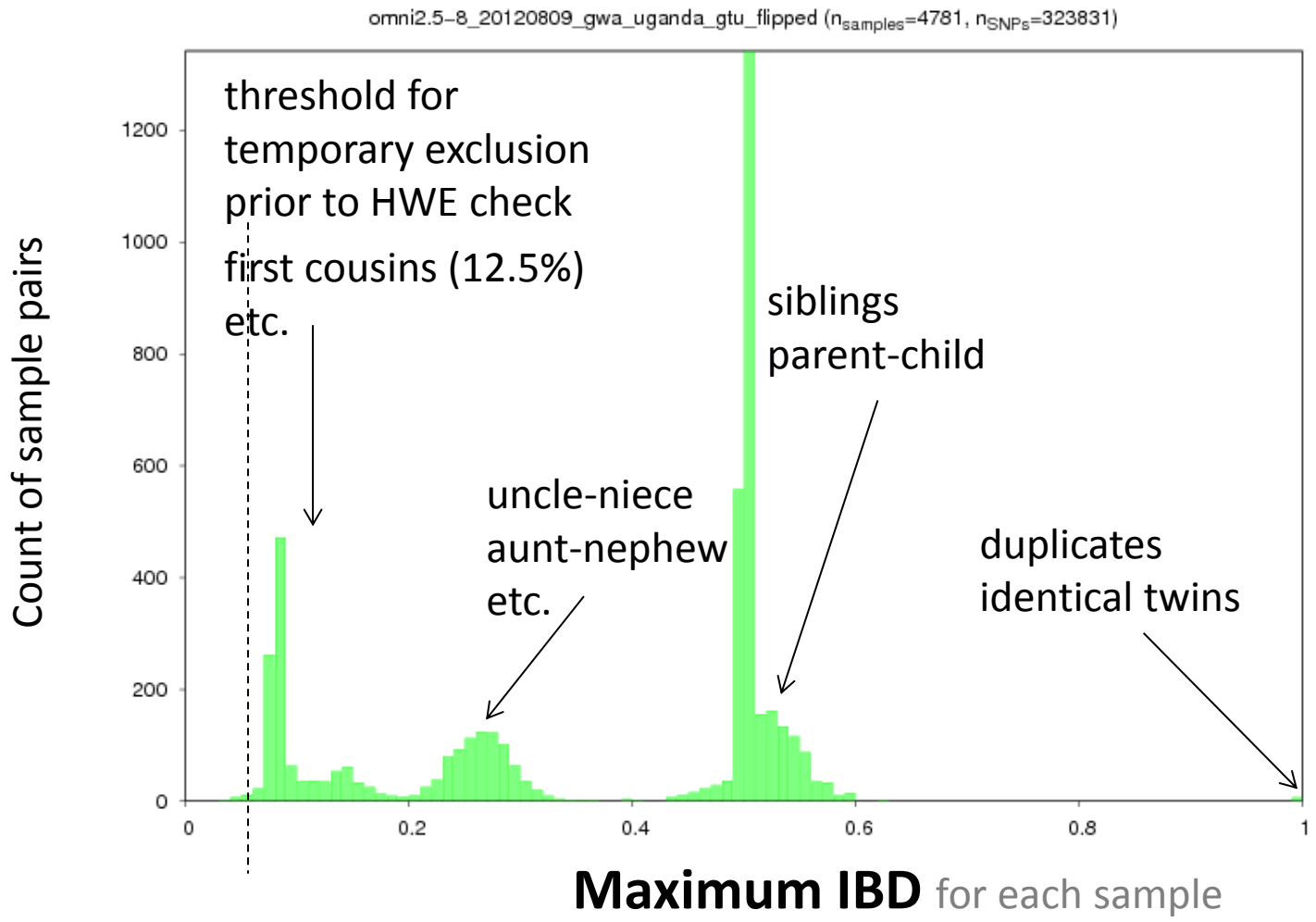
Not identical



Me and my mom

Me and my sister

Relatedness / IBD
A Ugandan cohort study as an example



Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

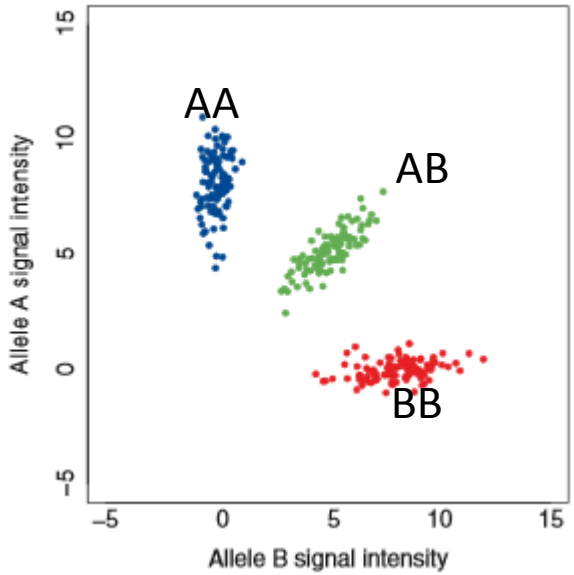
SNP QC

SNP Call Rate/Proportion

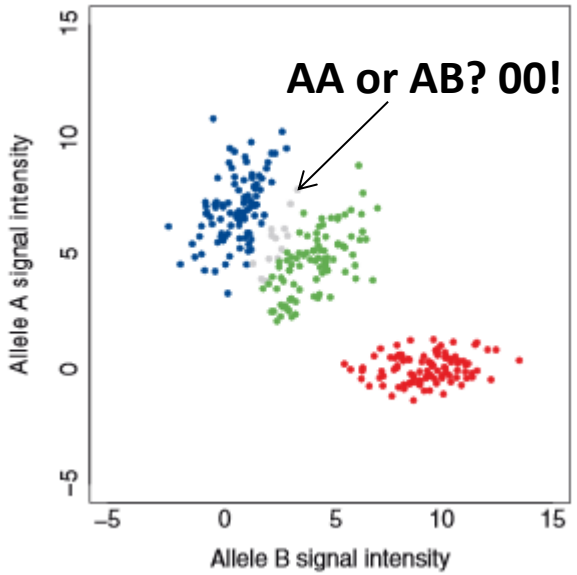
Hardy Weinberg Equilibrium (HWE)

How did we get the data? Genotype Calling

Good data
SNP1



Bad data
SNP2



SNP Call Rate/Proportion

	SNP1	SNP2	SNP3	SNP4	SNP5
Sample1	**00**	AG	GG	GA	**00**
Sample2	**00**	GG	GG	AA	CC
Sample3	AC	**00**	GG	AA	CC
Sample4	AA	AG	GC	AA	CC
Sample5	AC	AA	**00**	AA	CA
SNP Call Rate	**60%**	80%	80%	100%	80%

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

Too Little Relatedness
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

Hardy Weinberg Equilibrium (HWE)

Hardy–Weinberg principle

- The **Hardy–Weinberg principle** (also known as the **Hardy–Weinberg equilibrium, model, theorem**, or **law**) states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.

- **The Hardy-Weinberg principle can be illustrated mathematically with the equation:**

$$p^2+2pq+q^2 = 1$$

Hardy Weinberg Equilibrium

random mating

		Females	
		A (p)	C (q)
Males	A (p)	AA (p^2)	AC (pq)
	C (q)	AC (pq)	CC (q^2)

	SNP1	SNP2	SNP3	
Sample1	AC	AA	AC	
Sample2	AA	AA	AC	
Sample3	AC	CC	AC	
Sample4	CC	CC	AC	
f(A)=p	4/8	4/8	4/8	allele frequencies
f(C)=q=1-p	4/8	4/8	4/8	
f_e(AA)=p^2	1/4	1/4	1/4	expected genotype frequencies
f_e(AC)=2pq	2/4	2/4	2/4	
f_e(CC)=q^2	1/4	1/4	1/4	
f_o(AA)	1/4	2/4	0/4	observed genotype frequencies
f_o(AC)	2/4	0/4	4/4	
f_o(CC)	1/4	2/4	0/4	

When HWE does not apply

- Non-random mating
- Selection forces
- Alleles in disease causing loci
  - Apply HWE only to controls in a case-control study
- Migration
- Data quality

Quality Control Steps

Sample QC

Sample Call Rate/Proportion

Autosomal **Heterozygosity**

Sex / Gender
X Chromosome Heterozygosity

Too Much Relatedness
Identity By Descent (**IBD**)

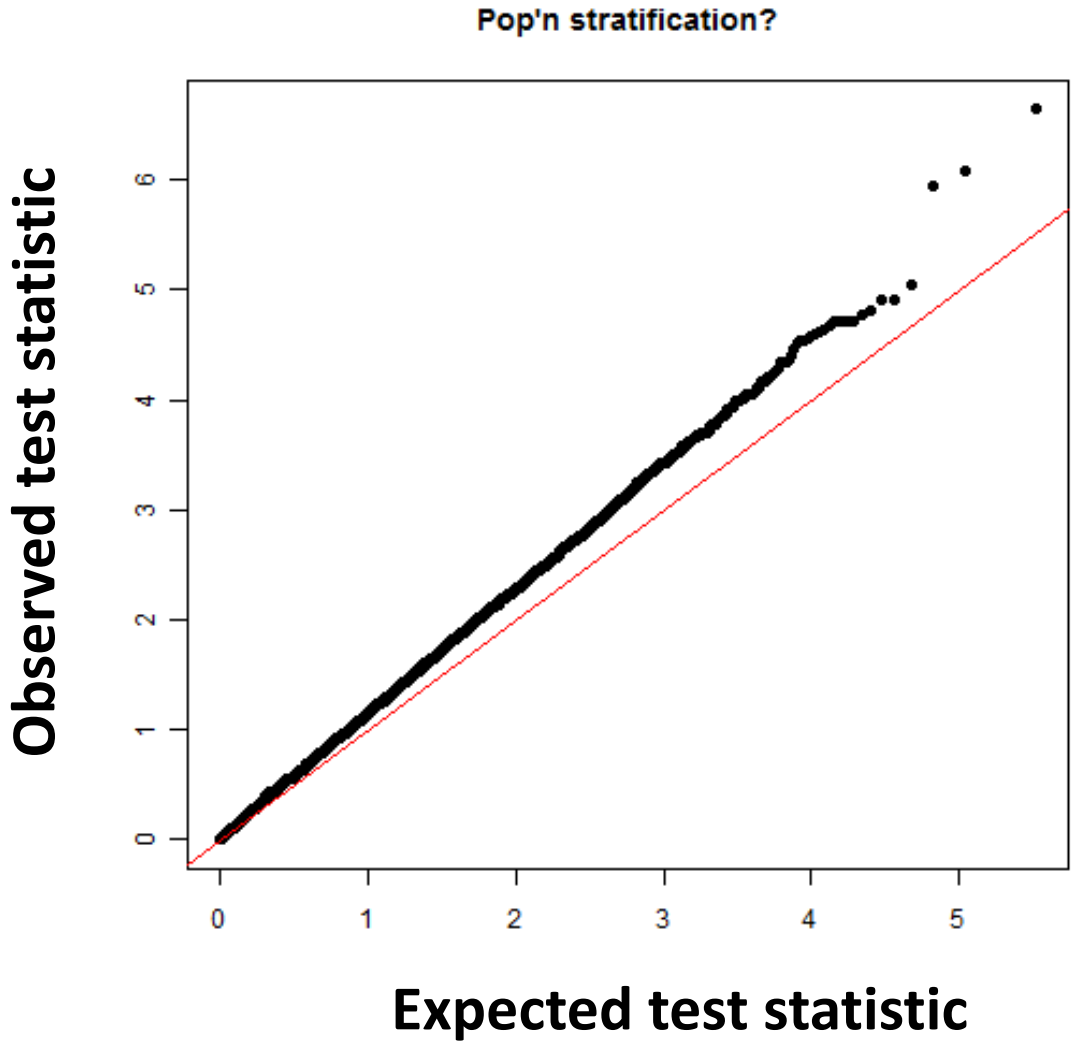
Too Little Relatedness / Confounding
Principal Component Analysis (**PCA**)

SNP QC

SNP Call Rate/Proportion

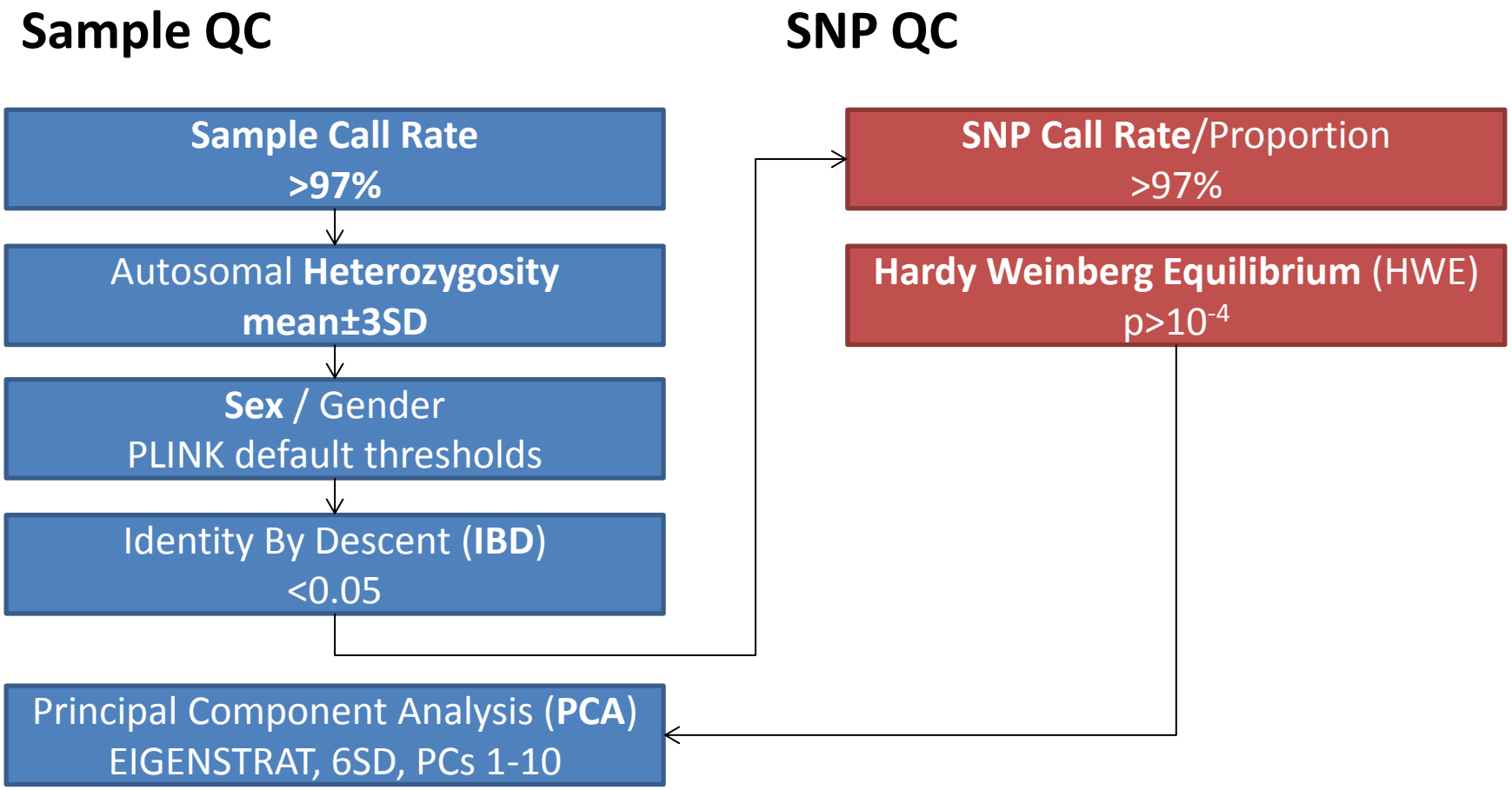
Hardy Weinberg Equilibrium (HWE)

Population structure - λ
Inflated QQ-plot



λ is a measure of the deviation from the diagonal

Simple But Effective QC
Common Thresholds



Summary

- 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 without “throwing the baby out with the bathwater”.
- All SNPs demonstrating evidence for association should be followed up with visual inspection of cluster plots.

Useful references

PROTOCOL

Data quality control in genetic case-control association studies

Carl A Anderson[1,2], Fredrik H Pettersson[1], Geraldine M Clarke[1], Lon R Cardon[3], Andrew P Morris[1] & Krina T Zondervan[1]

[1]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]GlaxoSmithKline, King of Prussia, Pennsylvania, USA. Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk) or K.T.Z. (krinaz@well.ox.ac.uk).

Published online 26 August 2010; doi:10.1038/nprot.2010.116

This protocol details the steps for data quality assessment and control that are typically carried out during case-control association studies. The steps described involve the identification and removal of DNA samples and markers that introduce bias. These critical steps are paramount to the success of a case-control study and are necessary before statistically testing for association. We describe how to use PLINK, a tool for handling SNP data, to perform assessments of failure rate per individual and per SNP and to assess the degree of relatedness between individuals. We also detail other quality-control procedures, including the use of SMARTPCA software for the identification of ancestral outliers. These platforms were selected because they are user-friendly, widely used and computationally efficient. Steps needed to detect and establish a disease association using case-control data are not discussed here. Issues concerning study design and marker selection in case-control studies have been discussed in our earlier protocols. This protocol, which is routinely used in our labs, should take approximately 8 h to complete.

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point P values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.

Useful references

A tutorial on statistical methods for population association studies

David J. Balding

Abstract | Although genetic association studies have been with us for many years, even for the simplest analyses there is little consensus on the most appropriate statistical procedures. Here I give an overview of statistical approaches to population association studies, including preliminary analyses (Hardy–Weinberg equilibrium testing, inference of phase and missing data, and SNP tagging), and single-SNP and multipoint tests for association. My goal is to outline the key methods with a brief discussion of problems (population structure and multiple testing), avenues for solutions and some ongoing developments.

NATURE REVIEWS | GENETICS VOLUME 7 | OCTOBER 2006 | 781

COMMENT

The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] and Magnus Nordborg[3,4]

The authors argue that population structure per se is not a problem in genome-wide association studies — the true sources are the environment and the genetic background, and the latter is greatly underappreciated. They conclude that mixed models effectively address this issue.

NATURE REVIEWS | GENETICS VOLUME 14 | JANUARY 2013 | 1

GENOME-WIDE ASSOCIATION STUDIES

New approaches to population stratification in genome-wide association studies

Alkes L. Price, Noah A. Zaitlen, David Reich and Nick Patterson

Abstract | Genome-wide association (GWA) studies are an effective approach for identifying genetic variants associated with disease risk. GWA studies can be confounded by population stratification — systematic ancestry differences between cases and controls — which has previously been addressed by methods that infer genetic ancestry. Those methods perform well in data sets in which population structure is the only kind of structure present but are inadequate in data sets that also contain family structure or cryptic relatedness. Here, we review recent progress on methods that correct for stratification while accounting for these additional complexities.

Useful software

- PLINK (QC)
- http://pngu.mgh.harvard.edu/~purcell/plink

- EIGENSTRAT (PCA)
- http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

- GEMMA (Association)
- http://home.uchicago.edu/xz7/software

- SNPTEST (Association)
- https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

- shellfish : Parallel PCA and data processing for genome-wide SNP data
- http://www.stats.ox.ac.uk/~davidson/software/shellfish/shellfish.php

Thank you for your attention!