Center for Genomics Research and Innovation

H3ABioNet

Genome-Wide Association Study (GWAS):

Quality Control

Segun Fatumo *PhD*

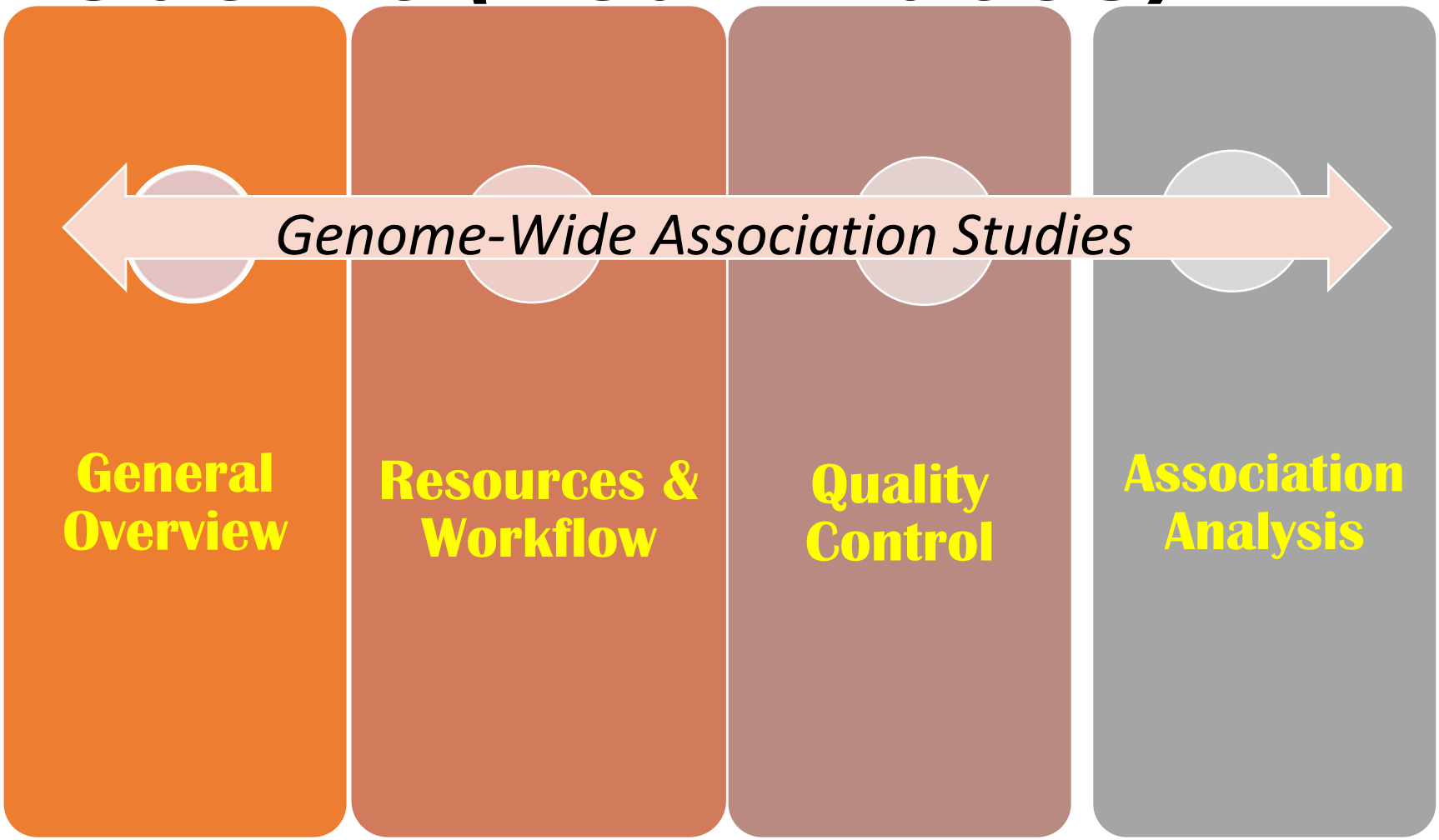
Associate Professor, Department of Non-Communicable Disease Epidemiology, LSHTM, UK

Group Leader, The African Computational Genomics Group, MRC/UVRI and LSHTM, Uganda

@SFatumo

segun.fatumo@lshtm.ac.uk

Outline (Four Videos)



Intended Learning Outcomes

By the end of the session, students will be able to

- Explain the reasons for performing quality control for genome-wide association studies (GWAS)
- Discuss relevant quality control steps.
- Recognise poor quality control in genetics association research papers.

Why Quality Control ?

- 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

- To identify samples and Single nucleotide polymorphisms (SNPs) of poor quality or questionable identity

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)

Genotype data

	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

00 = missing data

Missing call rate - 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)

Calculating Sample 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

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)

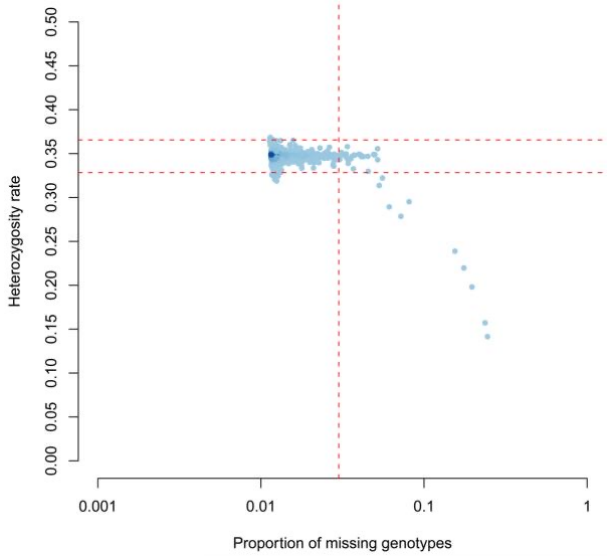
Heterozygosity Rate

homozygous

heterozygous

heterozygosity = 2/8 = 0.25

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



Quiz

What are the factors that might contribute to excessive or reduced proportion of heterozygote genotypes ?

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 coefficient F <0.2 female)

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.
- 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.

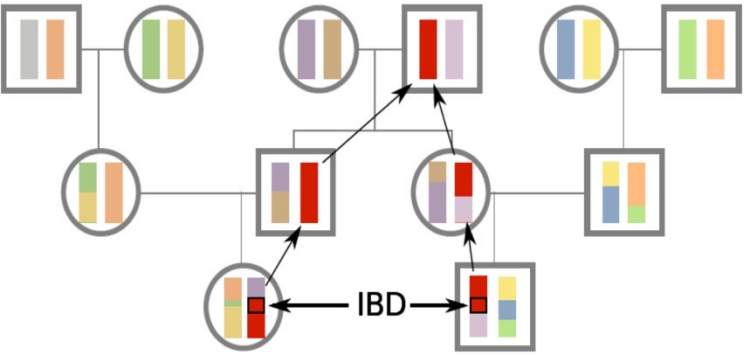
Relatedness / IBD

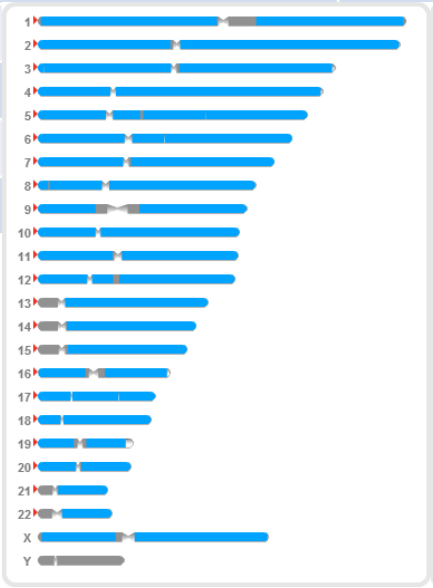
Completely identical

Half-identical

Not identical

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





Me and my mom

Me and my sister

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)

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:**

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

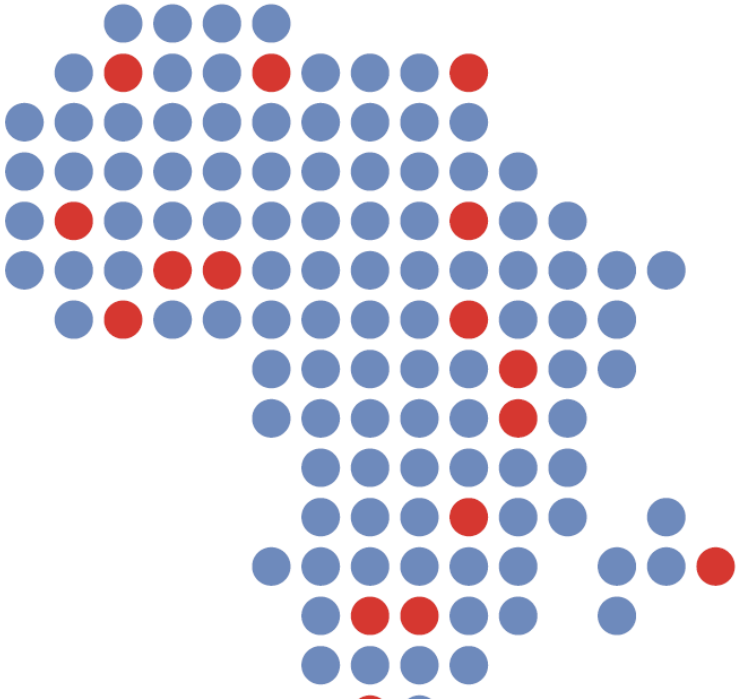
Summary

- 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.

Further Reading

Anderson, Carl A., Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. "**Data quality control in genetic case-control association studies**." ***Nature protocols*** 5, no. 9 (2010): 1564-1573. https://www.nature.com/articles/nprot.2010.116

Marees, Andries T., Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. "**A tutorial on conducting genome-wide association studies: Quality control and statistical analysis**." *International journal of methods in psychiatric research* 27, no. 2 (2018): e1608. https://onlinelibrary.wiley.com/doi/full/10.1002/mpr.1608



H3ABioNet