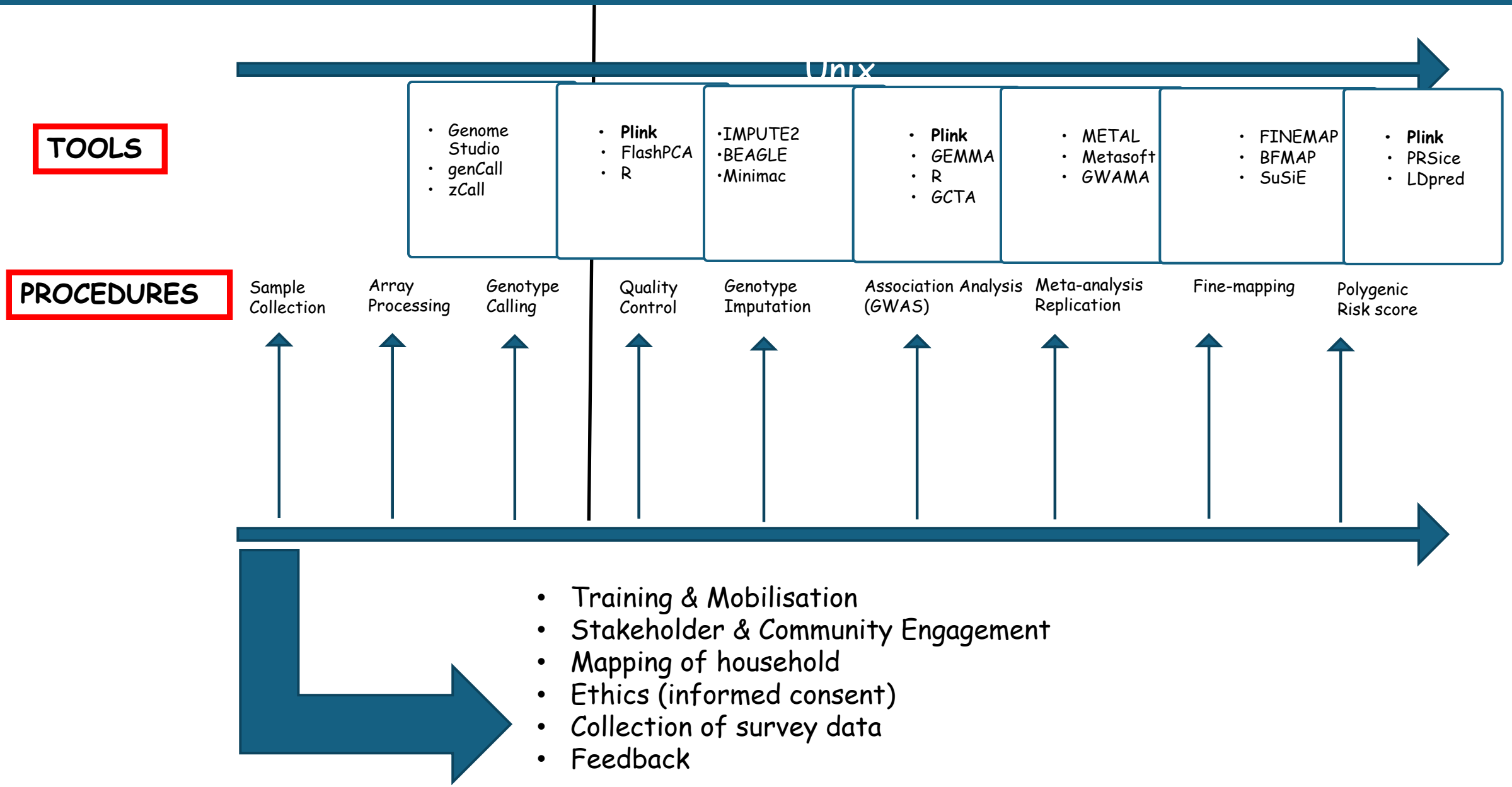
GWAS data preparation and Quality control

Sola Ojewunmi, PhD
Senior Research Fellow, Queen Mry University of London
Assistant Professor, MRC/UVRI & LSHTM Uganda Unit

GWAS pipeline and Computational tools

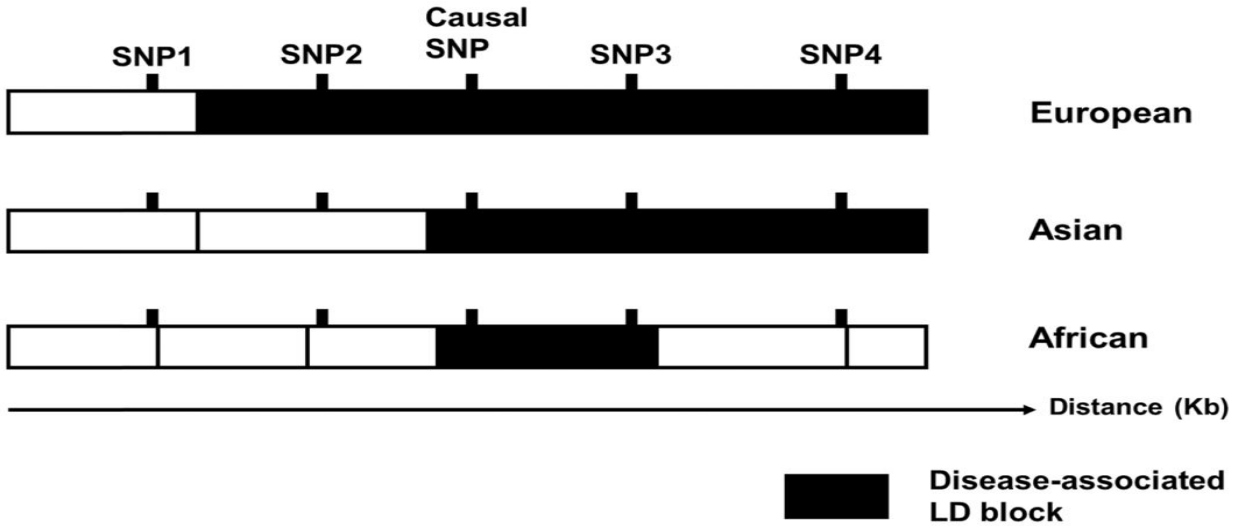


Genotyping platforms

Illumina Microarray Technology

Affymetrix Technology

Infinium™ H3Africa Consortium Array



Gabriel et al., 2002; Li and Meyre, 2014

.

- The custom content was selected by including specific single nucleotide polymorphisms (SNPs) with known association with diseases.

- The custom SNPs were selected to improve coverage, imputation accuracy, and enrichment in novel but common variants in African populations.

Feature	Description
Species	Human
Total number of markers	2,271,503
Number of samples per BeadChip	8
DNA input requirement	200 ng
Assay chemistry	Infinium LCG
Instrument support	iScan™ System
Sample throughput	~1728 samples/week
Scan time per sample	35 minutes

An overview of H3Africa Illumina SNP genotyping

Day 1

1 **Quantify DNA (Optional)**
Hands-on: 30 minutes/plate
Fluorometer: 5 minutes/plate
Reagents: PicoGreen dsDNA quantification
Output: Sample DNA plate with quantitated DNA

Safe Stopping Point

2 **Amplify DNA**
Hands-on: ~1 hour
Reagents: 0.1 N NaOH, MA1, RPM, MSM
Output: MSA6 plate

3 **Incubate DNA**
Incubation: 20–24 hours
Output: MSA6 plate with amplified DNA

Overnight Incubation

Day 2

4 **Fragment DNA**
Hands-on: ~30 minutes
Incubation: 1 hour
Reagents: FMS
Output: MSA6 plate

Safe Stopping Point

5 **Precipitate DNA**
Hands-on: 30 minutes
Incubation/dry time: ~2 hours
Reagents: 2-propanol, PM1
Output: MSA6 plate

Safe Stopping Point

6 **Resuspend DNA**
Hands-on: ~30 minutes
Incubation: 1 hour
Reagents: RA1
Output: MSA6 plate

Safe Stopping Point

7 **Hybridize DNA to BeadChip**
Hands-on: ~30 minutes
Incubation: 20 minutes
Cool down: 30 minutes
Incubation: 16–24 hours
Reagents: 100% EtOH, PB2, XC4

Overnight Incubation

Day 3

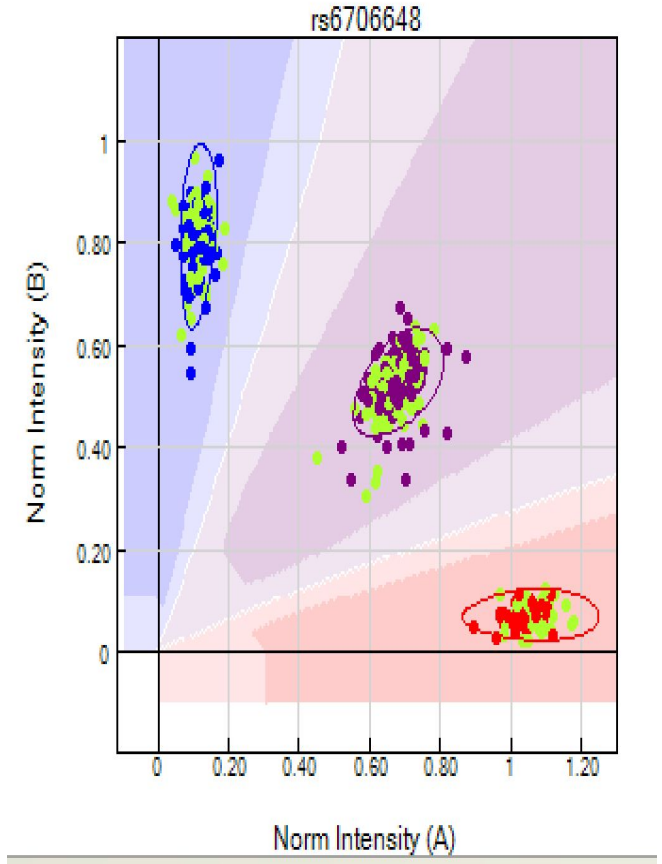
8 **Wash BeadChip**
Hands-on: ~20 minutes/4 BeadChips
Cool down: 25 minutes
Reagents: PB1
Output: BeadChip

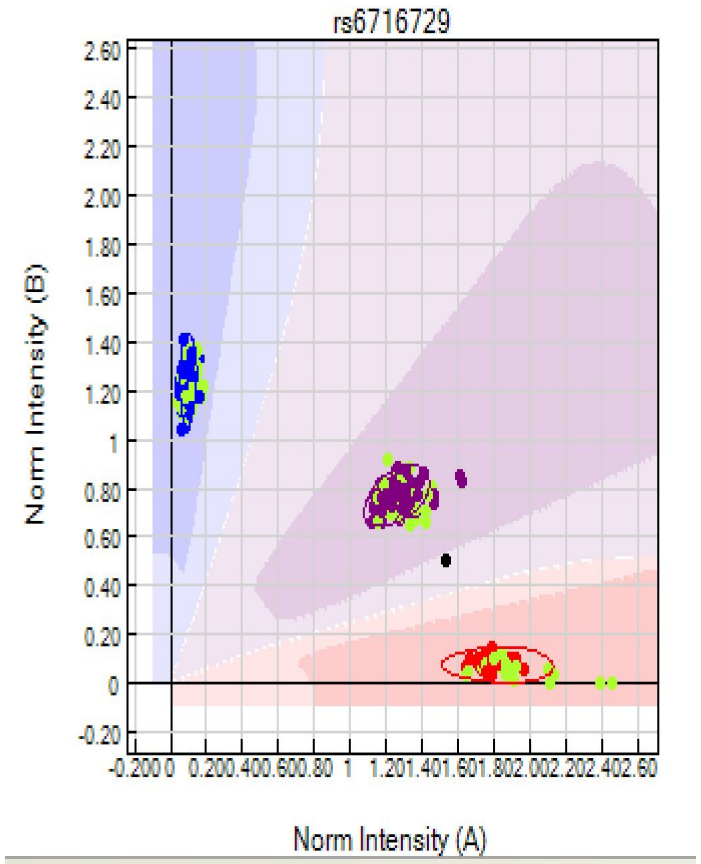
9 **Extend and Stain BeadChip**
Hands-on: ~2.5 hours
Cool down: 25 minutes
Reagents: 75% EtOH, 95% form/EDTA, ATM, EML, LX1, LX2, PB1,RA1, SML, XC3, XC4
Output: BeadChip

10 **Image BeadChip**
Scanning: 52 minutes/BeadChip
Output: Image and data files

Pre-Amp Post-Amp

Genotype calling in Genome Studio





Data Format

Text PLINK files

***.ped**

FID	ID	F	M	S	P	-GENETIC INFO-							
CH18526	NA18526	0	0	2	1	G	G	C	C	T	T	A	A
CH18524	NA18524	0	0	1	1	G	G	C	C	T	T	A	A
CH18529	NA18529	0	0	2	1	C	G	C	C	T	T	C	A
CH18558	NA18558	0	0	1	1	G	G	C	C	G	T	A	A
CH18532	NA18532	0	0	2	1	G	G	C	C	T	T	A	A

***.map**

Chr	SNP	SNP Position	Base-Pair Coordinate
8	rs17121574	12.7991	12799052
8	rs754238	12.8481	12848056
8	rs11203962	12.8484	12848438
8	rs6999231	12.8623	12862253
8	rs17178729	12.867	12867001

Minor allele

Major allele

***.fam**

FID	ID	F	M	S	P
CH18526					
CH18524					
CH18529					
CH18558					
CH18532					

***.bed**

Binary version

***.bim**

Chr	SNP	SNP	Base-Pair	Allele1	Allele2
					G
					G
					G
					G
					G

For instance, rs1427407:
GRCh37:chromosome 2:60718043
GRCh38: chromosome 2:60490908

Covariates

FID	ID	Sex	Cohort	PC1	PC2	etc…
CH18526	NA18526	2	1	0.00542	-0.00876	
CH18524	NA18524	1	1	0.04517	-0.00761	
CH18529	NA18529	2	4	0.07776	-0.00231	
CH18558	NA18558	1	3	0.00125	-0.00356	
CH18532	NA18532	2	2	0.00456	-0.00651	

Always state the human reference genome version for your genetic data.

Why QA/QC?

Don't throw the baby out with the bathwater

Quality assurance: QA = Good practice for *generating* quality data

Quality control: QC = Throw bad data away to conform to quality metrics

Why QC in GWAS?

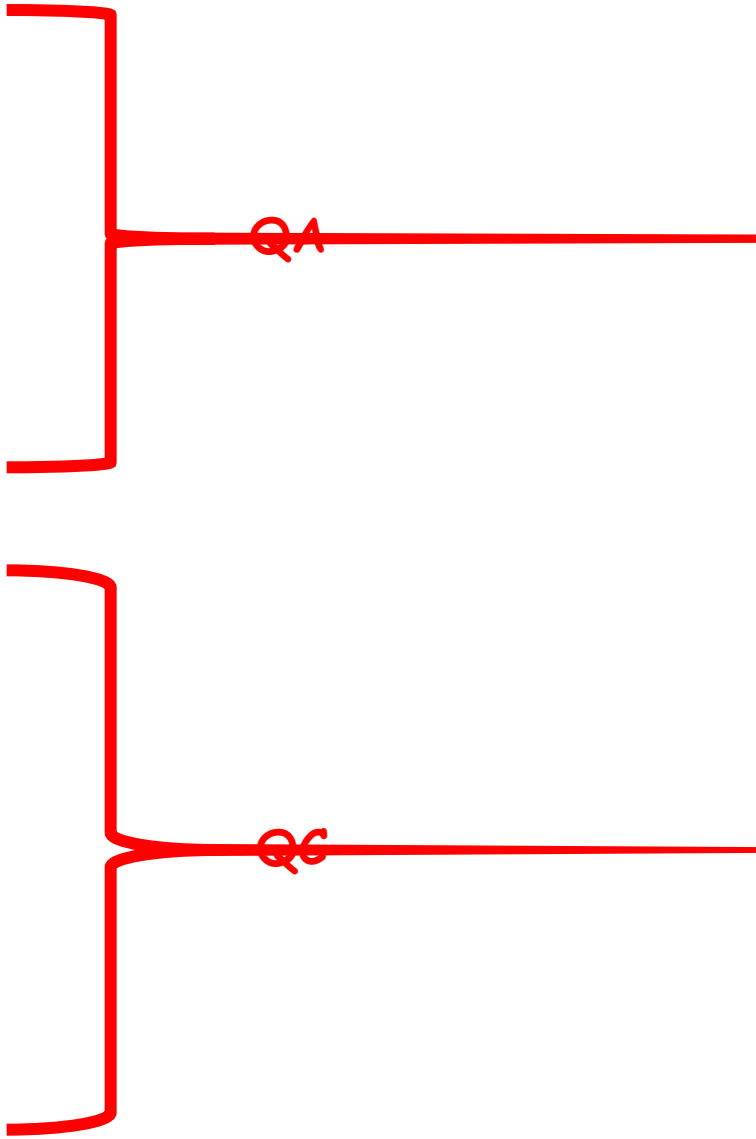
- GWA studies involve a large number of SNPs, and errors in genotyping can be detrimental.
- To avoid false positive association signals.

Overview of QA/QC checks

- **Pre-genotyping checks**
  - DNA preparation & quantification
  - Equal treatment of cases and controls
- **Genotype calling**
  - Re-run after removing bad samples?
  - Run on cases + controls together?

Overview of QA/QC checks

- **Post-genotyping checks**
  - Individual QC
  - SNP/Marker QC
  - Choosing QC thresholds
- **Post-association checks**
  - Re-examine SNP cluster plots
  - Re-examine QC metrics



Structure of PLINK command line in a Linux terminal

```
root@KCLJ6DWDB3:~# ./plink --bfile HapMap_3_r3_1 --freq --out HapMap_3_r3_1_freq
```

path to the directory containing your files

Specifies software PLINK

Input genetic file

specifies the option

specifies the output file name

Quality Assurance and Control – 'Individuals'

WHAT: QC	WHY	HOW (QC filters)
Individual Missingness	Indicates poor quality DNA. Informative missingness.	Filter out individuals with very high levels of missingness. Exclude sample call rate that is < 95%. Plink option: --missing : --mind
Sex Check	Indicates data recording problem	Chrom X/Y Data Inbreeding coeff (F): F=0 for females; F=1 for males Use F estimate >0.8 for males and females <0.2. Plink option: --sex-check
Relatedness: the possibility that individuals in your study may be close relatives.	Violates association testing assumptions; To ensure that the observed associations between genetic variants and traits are not confounded by shared ancestry.	Prune SNPs for LD. Calculate IBS and IBD IBD=1: duplicate/monozygotic twins IBD=0.5: 1st degree (full siblings; parent-offspring) IBD=0.25: 2nd degree (half-siblings; grandparent-grandchild; uncle/aunt -nephew/niece) IBD=0.125: 3rd degree (first cousins) Use a pi-hat threshold = 0.1875 threshold to exclude related samples, except you will use software that will account for relatedness during assoc. analysis. Plink option: --genome
Population stratification: Differences in allele frequencies between cases and controls resulting from systematic differences in ancestry rather than association of genes with disease.	False positives	PCA: Use independent SNPs (**pruning**) for this analysis and limit it to autosomal chromosomes only. Plink option: --pca *admixture
Heterozygosity: the proportion of heterozygous genotypes. -low heterozygosity - high heterozygosity	Sample contamination: Inbreeding (consanguinity) or individual belonging to a sub-population (the Wahlund effect)	Check heterozygosity: Remove individuals who deviate ±3 SD from the samples' heterozygosity rate mean. Plink option: --het

Calculating Sample/SNP Call Rate

	SNP1	SNP2	SNP3	SNP4	SNP5	Individual Missingness (%)
SAMPLE 1	**00**	AG	GG	GA	**00**	40%
SAMPLE 2	**00**	GG	GG	AA	CC	20%
SAMPLE 3	AC	**00**	GG	AA	CC	20%
SAMPLE 4	AA	AG	GC	AA	CC	0%
SAMPLE 5	AC	AA	**00**	AA	CA	20%
SNP Missingness (%)	40%	20%	20%	0%	20%	

plink.imiss

plink.lmiss

NOTE. In general, 1 to 5% of missing data per individual and SNP are allowed.

Quality Assurance and Control – 'SNPs'

WHAT:QC	WHY	HOW (QC filters)
SNP Missingness	Informative missingness.	Check rate of SNP missingness and filter out SNPs with very high levels of missingness. Plink option: --missing : --geno
Minor Allele Frequency (MAF)	Low MAF SNPs lack the power to detect real associations and are prone to genotyping errors.	Includes only SNPs above the set MAF threshold. The MAF threshold should depend on your sample size; larger samples can use lower MAF thresholds. *N* = 100,000 vs. moderate samples (*N* = 10,000), 0.01 and 0.05, respectively are commonly used as MAF thresholds. Plink option: --freq (distribution) Plink option for maf filter: --maf
Hardy-Weinberg Equilibrium (HWE): states that allele and genotype frequencies are constant over generations in the absence of evolutionary forces such as migration, natural selection or mutation.	Departure indicates that the genotype frequencies are significantly different from expectations, which may be due to genotyping errors or evolutionary selection.	Calculate p-value (null = HWE) and exclude SNPs that deviate from HWE. For binary traits we suggest to exclude: HWE *p*-value <1e−10 in cases and <1e−6 in controls. For quantitative traits, apply HWE *p* value <1e-6. Plink option: --hwe

Population Structure

Association signals observed in GWAS can be false positives (spurious association) as a consequence of population structure.

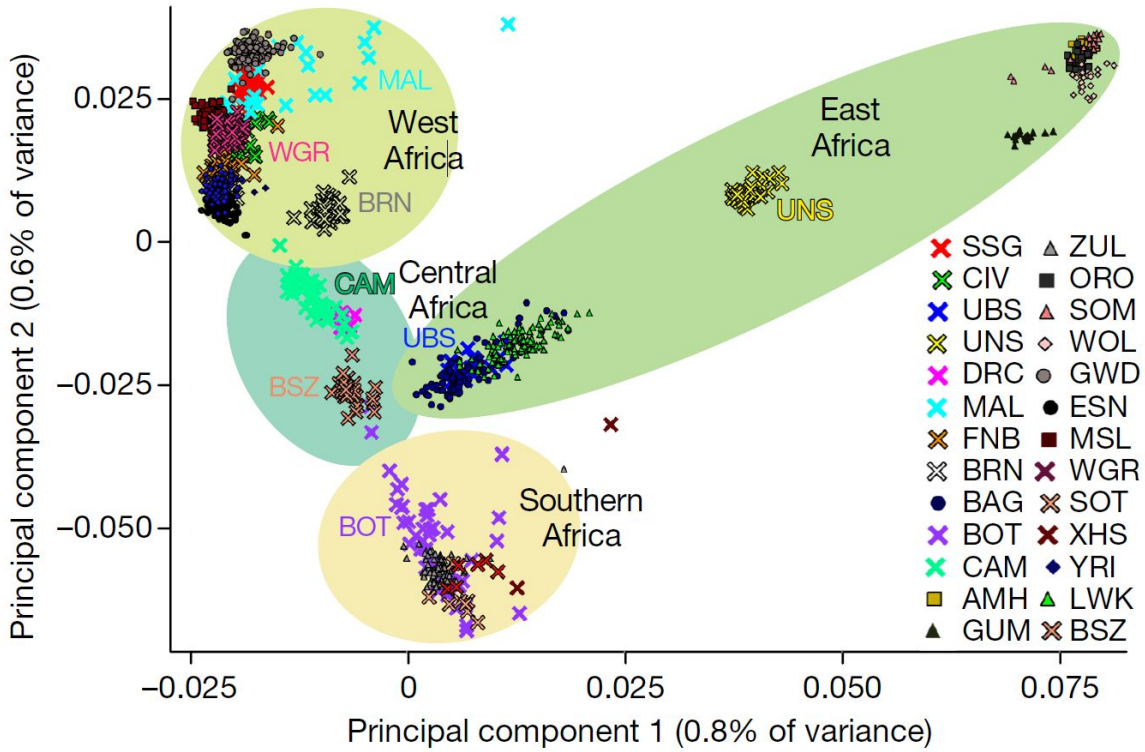
Population structure or population stratification refers to the presence of subgroups (e.g., individuals with different ethnic background) in a study.

Why does population structure matter?

- Allele frequencies vary by ancestry.
- Traits could also vary by ancestry due to environment, diet, etc.

How to handle population structure:

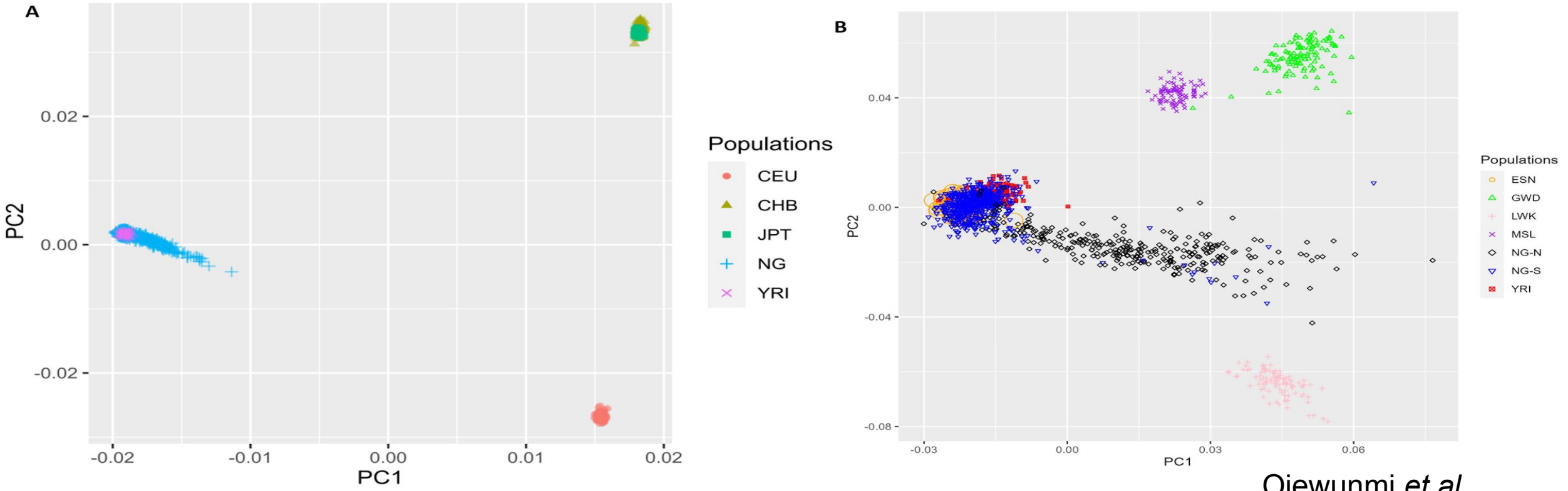
- Principal component analysis (PCA): This is a widely used method for identifying and correcting ancestry difference.
- Linear mixed model
- Stratified or ancestry-specific analysis



PCA of African WGS data

Choudhury *et al.*, 2020

Investigating Population stratification with PCA



Ojewunmi *et al.*, 2024

(A) **The principal component analysis of study participants with the global populations**. **CEU**: Utah residents with Northern and Western European ancestry representing Europe; **CHB:** Han Chinese in Beijing, China and **JPT**: Japanese in Tokyo, Japan representing East Asia; **NG**: Nigeria (This study); **YRI**: Yoruba in Ibadan representing Africa.

(B)**The principal component analysis of study participants** with the African continental populations. **ESN**: Esan in Nigeria; **GWD**: Gambian in Western Division; **LWK**: Luhya in Webuye, Kenya; **MSL**: Mende in Sierra Leone; **NG-S**: study participants enrolled from the Nigeria South-west recruitment site; **NG-N**: study participants enrolled from the North-central and North-west recruitment sites; **YRI**: Yoruba in Ibadan, Nigeria.

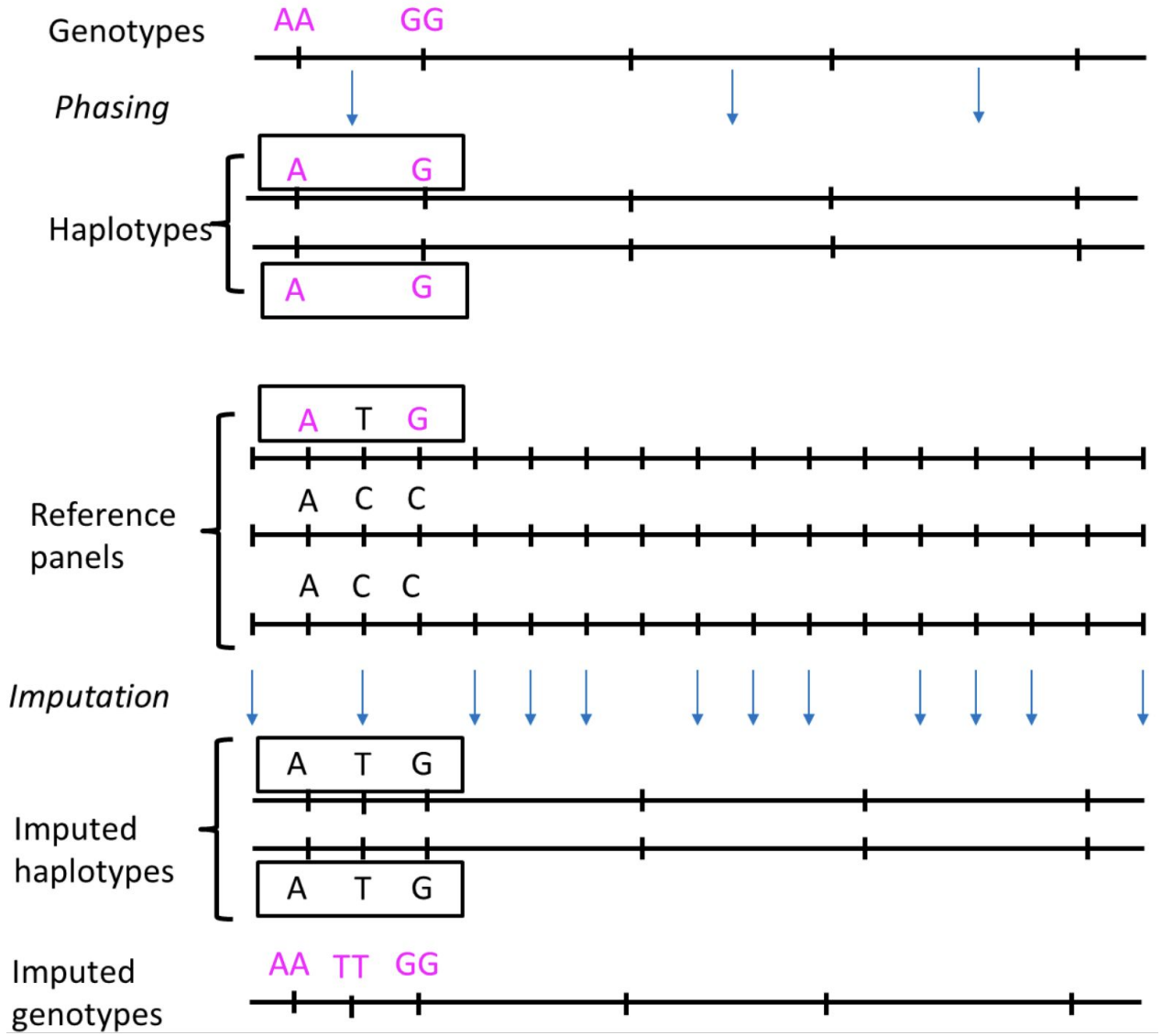
Imputation

Genotype imputation refers to inferring unobserved genotypes in a sample of individuals, i.e., predicting genotypes that have not been directly typed, based on LD patterns, to increase the coverage of genotyping arrays.

It is an essential step before a genome-wide association analysis.

Imputation can **increase power** to detect causal SNP, which may be present in the reference panel, enhance **fine mapping** to get a high-resolution view in identifying causal variants and **facilitate meta-analysis**.

Genotype imputation is affected by the quality of genotyped data, **reference panel** [Number of individuals, SNP density/sequencing coverage, and similarity of ancestry with study sample], and MAF.



Imputation QC

Pre-imputation:

- It is essential that you exclude poor-quality variants.
- Common to exclude MAF<1% variants.

Post imputation:

- Imputation quality is assessed by "information measures" in the range of 0-1.
  - Typical to filter SNPs by excluding r square <0.8, **<0.3**.

Commonly used Reference panels for imputation

Reference panel	Samples	Sites (millions)	Ancestry distribution	Panel content	Phasing and imputation
AGR	4,956	93	predominantly African[a]	chr1-22 and X; biallelic SNPs only	EAGLE2+ PBWT
TOPMed	97,256	308	multi-ethnic	chr1-22 and X; SNPs and indels	EAGLE2+ Minimac4
KGP_S	2,504	85	multi-ethnic	chr1-22 and X; SNPs and indels	EAGLE2+ PBWT
KGP_M	2,504	49	multi-ethnic	chr1-22 and X; SNPs and indels	EAGLE2+ Minimac4
HRC	32,470	40	predominantly European	chr1-22 and X; SNPs only	EAGLE2+ PBWT

Panel codes: AGR, African Genome Resource hosted at the Sanger Imputation Server (SIS); KGP_S, 1000 Genomes Project hosted at the SIS; HRC, Haplo-type Reference Consortium hosted at the SIS; KGP_M, 1000 Genomes Project hosted at the Michigan Imputation Server; TOPMed, hosted at the TOPMed Imputation Server.

[a]Populations from Uganda, Ethiopia, Egypt, and Nama from South Africa and the 1000 Genomes Project.

Sengupta *et al.*, 2023

Summary

- African populations have high genetic diversity and short LD blocks.
- H3Africa SNP-array is the ideal array for African ancestry genotyping.
- Ensure quality control steps are carried out to remove poorly genotyped samples and variants during quality control.
- Investigate population structure and include PCs in association analysis.
- Genotyping arrays assay only a subset of genetic variants.
- Imputation increases genomic coverage using ancestry-matched reference panels.
- African-enriched reference panels improve imputation accuracy.

Materials for further reading

- H3A SNP array: https://h3africa.org/index.php/2019/12/12/h3africa-chip-faq/
- Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499-511. doi: 10.1038/nrg2796. PMID: 20517342.
- Deng T, Zhang P, Garrick D, Gao H, Wang L, Zhao F. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. Front Genet. 2022 Jan 3;12:704118. doi: 10.3389/fgene.2021.704118. PMID: 35046990; PMCID: PMC8762119.
- Sengupta D, Botha G, Meintjes A, Mbiyavanga M; AWI-Gen Study; H3Africa Consortium; Hazelhurst S, Mulder N, Ramsay M, Choudhury A. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. Cell Genom. 2023 May 23;3(6):100332. doi: 10.1016/j.xgen.2023.100332.
- Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608. doi: 10.1002/mpr.1608.
- Pre-imputation site, tools and tutorials:

https://imputationserver.readthedocs.io/en/latest/prepare-your-data/

- TOPMed imputation server:

https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home

Q & A