OUTLINE

- Overview of genome-wide association studies (GWAS)
- GWAS Quality Control
- Association Testing
- Data Format
- PRACTICAL (Itunuoluwa & Melek)

QUALITY CONTROL STEPS

Sample QC

SNP QC

Sample Call Rate

Heterozygosity

Sexy

IBD

PCA

SNP Call Rate

Hardy Weinberg Equilibrium

SNP CALL RATE/PROPORTION

	SNP		SNP	SNP	SNP	SNP	SNP
Ones	00	Sample	00	AG	GG	GA	00
Palwende	00	Sample	00	GG	GG	AA	CC
Itunu	AC	Sample	AC	00	GG	AA	CC
Conrad	AA	Sample	AA	AG	GC	AA	CC
Melek	AC	Sample	AC	AA	00	AA	CA
SNP Call Rate	60%						

GENOTYPE DATA

	SNP1	SNP2	SNP3	SNP4	SNP5
Jones	**00**	AG	GG	GA	**00**
Palwende	**C0**	TC	CT	TT	CC
Itunu	AC	**00**	CC	CA	AA
Conrad	AT	TA	TT	**00**	AA
Melek	CG	CC	**00**	GC	GG

00 = missing data

MISSING CALL RATE - SAMPLE

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

	SNP1	SNP2	SNP3	SNP4	SNP5
Jones	**00**	AG	GG	GA	**00**
Palwende	**C0**	TC	CT	TT	CC
Itunu	AC	**00**	CC	CA	AA
Conrad	AT	TA	TT	**00**	AA
Melek	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	**Sample Call Rate**
Jones	**00**	AG	GG	GA	**00**	60%
Palwende	**00**	GG	GG	AA	CC	80%
Itunu	AC	**00**	GG	AA	CC	80%
Conrad	AA	AG	GC	AA	CC	100%
Melek	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)

HETEROZYGOSITY RATE

	Copy 1	Copy 2
SNP1	A	A
SNP2	C	C
SNP3	A	T
SNP4	C	C
SNP5	C	G
SNP6	A	A
SNP7	A	A
SNP8	T	T

chromosome 1

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

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

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		SNP1	SNP2	SNP3	SNP4	SNP5
Jones	**00**	Sample1	**00**	AG	GG	GA	**00**
Palwende	**00**	Sample2	**00**	GG	GG	AA	CC
Itunu	AC	Sample3	AC	**00**	GG	AA	CC
Conrad	AA	Sample4	AA	AG	GC	AA	CC
Melek	AC	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:

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

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

ASSOCIATION TESTING

INTENDED LEARNING OUTCOMES

- Describe the statistical models for identifying phenotype-genotype associations.

- Identify appropriate model for adjusting for potential confounding when performing genetic association study.

- Interpret plots from genetics association study

GENOTYPE-PHENOTYPE ASSOCIATIONS

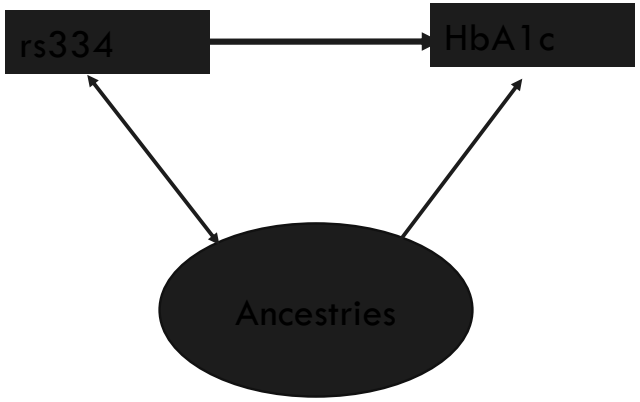
Case-control GWAS

- Identify SNPs where one allele is significantly more common in cases than controls
- The SNP is *associated* with disease
- **Logistic regression**

Continuous trait GWAS

- Fit **linear regression** model for each (SNP, phenotype) pair

MIXED MODEL



- A major concern in GWAS is the need to account for the complicated dependence-structure of the data both between loci as well as between individuals. (Adjust for the confounding variable)

- For example, ethnic groups (and subgroups) often share distinct dietary habits and other lifestyle characteristics that leads to many traits of interest being correlated with ancestry and/or ethnicity

- The linear mixed model (LMM) is a powerful way to control covariation arising from complex correlation

TESTING FOR ASSOCIATION

- All methods produce a and a indicating how significant the association/correlation appears to be
  - i.e. how likely it was to have occurred by chance

- In GWAS, we require (e.g. $p = 5 \times 10^{-8}$) to overcome the multiple testing problem incurred when we test many SNPs throughout the genome
  - If testing 1 million SNPs using $p = 0.05$, we would obtain 50,000 'significant' results just by chance!

  - Bonferroni correction

USEFUL SOFTWARE

PLINK (Standard tool for QC and association analysis)

http://pngu.mgh.harvard.edu/~purcell/plink

GEMMA (Association) – Implemented Mixed Model

http://home.uchicago.edu/xz7/software

SNPTEST (Association)

https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

GWAS OUTPUT

risk allele: in the context of a disease, this is the allele that confers a risk of developing the disease. Most of the time, risk allele = minor allele, as most people will not carry the risk allele. However, in some case, the risk allele can in fact be the major allele.

- Output from GWAS analysis is a Summary Statistics
- Most important columns are:
  - Chr: Chromosome
  - SNP: SNP Id
  - BP: Basepair
  - EA: Effect Allele
  - NEA: Non Effect Allele or Reference Allele
  - EAF: Effect Allele Frequency
  - Effect: Odd ratio or Beta
  - SE: Standard Er
  - P: Pvalue

MANHATTAN PLOT

CLOSE-UP OF HIT REGION

QUANTILE-QUANTILE (QQ) PLOT

FIRST GWAS OF KIDNEY FUNCTION IN AFRICA

We identified and validated two eGFR loci.

At ***GATM*** locus, the association signal (lead SNP rs2433603, $P = 1.0 \times 10^{-8}$) in the Uganda GPC GWAS was distinct from previously reported signals at this locus.

The lead SNP at the *HBB* locus accounted for 88% of the posterior probability of causality after fine-mapping.

CONCLUSIONS

In designing GWAS, the selection of the appropriate association test depends on a number of factors:

- type of phenotype,
- control for covariates eg population structure.
- etc

Visualisation of GWAS output is key

- Manhattan
- QQ Plot
- Regional plot

FURTHER READING

Pearson, T.A. and Manolio, T.A., 2008. How to interpret a genome-wide association study. *Jama*, *299*(11), pp.1335-1344.
http://www.biostat.jhsph.edu/~iruczins/teaching/misc/gwas/papers/pearson2008.pdf

DATA FORMAT AND PLINK

PLINK FLAT FILES (MAP/PED)

PLINK is a very widely used application for analyzing genotypic data. It can be considered the “de-facto” standard of the field, although newer formats are starting to be widespread as well.

Standard PLINK format: MAP/PED

Binary format - BED/BIM/FAM

MAP files

The **MAP** file describes the SNPs.

The fields in a MAP file are:

- Chromosome
- Marker ID
- Genetic distance
- Physical position

MAP FILE

Chr ↓

SNP ID ↓

BP Position ↓

↑ Genetic Distance (morgans)

PED FILE

PED format

- PED file – Sample/Individual information

Family ID

Individual ID

Paternal ID

Maternal ID

Sex (1=male, 2=female, other)

Phenotype (missing -9, control 1, case 2 or quantitative trait)

Followed by allele calls for the variant in a pairwise fashion – Different encodings

PED FORMAT

FID IID PATID MATID PHENO SEX Alleles for SNP 1 Alleles for SNP 2

DATA STORAGE

Plain text format for thousands of samples for millions of SNPs would require a large amount of space for storage

PLINK rather works with Binary versions of the PED files

Method to compress and reduce the size of the PED and MAP files

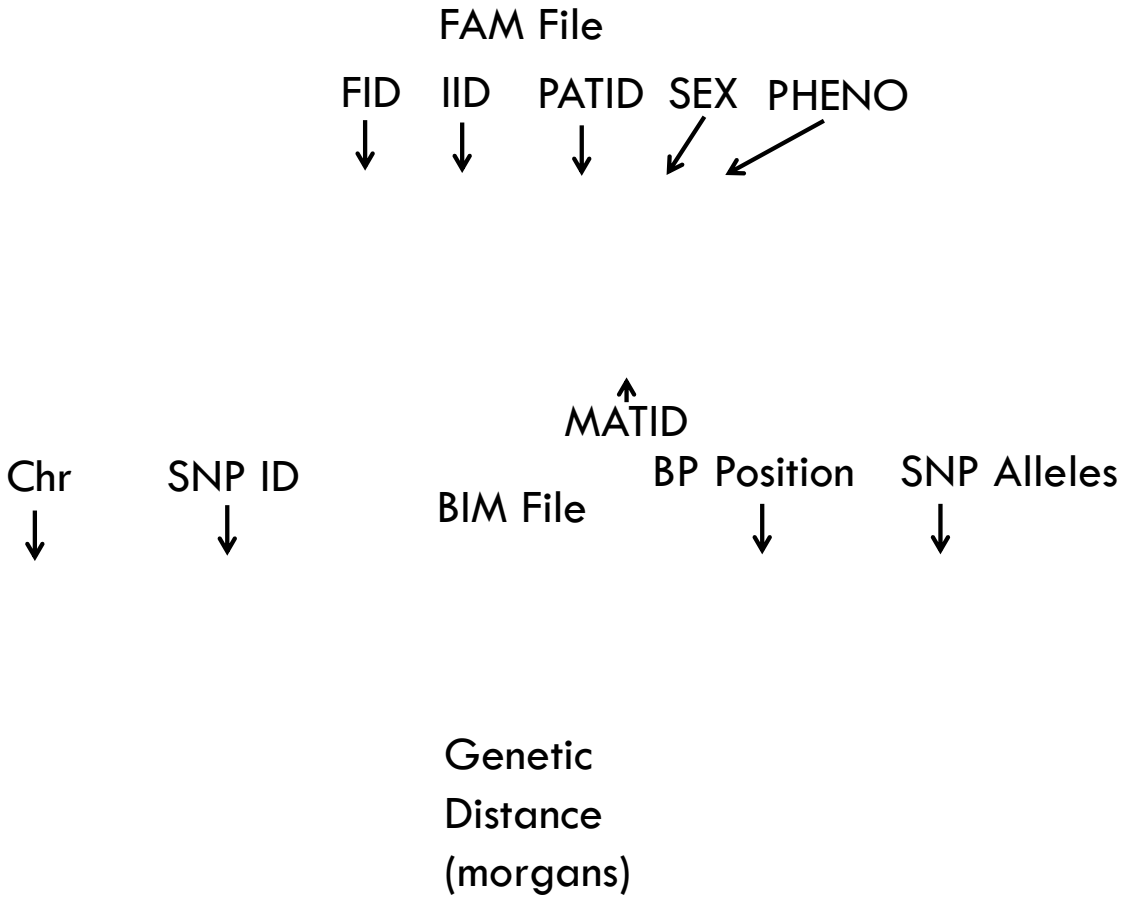
BINARY PED FORMAT

FAM file – one row per individual – first 6 columns of PED file

BIM file – one row per SNP – MAP file + two alleles for that SNP

BED file – one row per individual – genotype calls for each individual for all SNPs – rest of PED file in binary format

FAM and BIM file are human readable while BED file in not


FAM File
FID
IID
PATID
SEX
PHENO
MATID
Chr
SNP ID
BIM File
BP Position
SNP Alleles
Genetic
Distance
(morgans)

PRACTICAL
(ITUNUOLUWA & MELEK)

We established KidneyGenAfrica to facilitate multi-centre research in Africa, enhance large-scale genomic study and deliver research and training excellence in genomics of kidney function in Africa

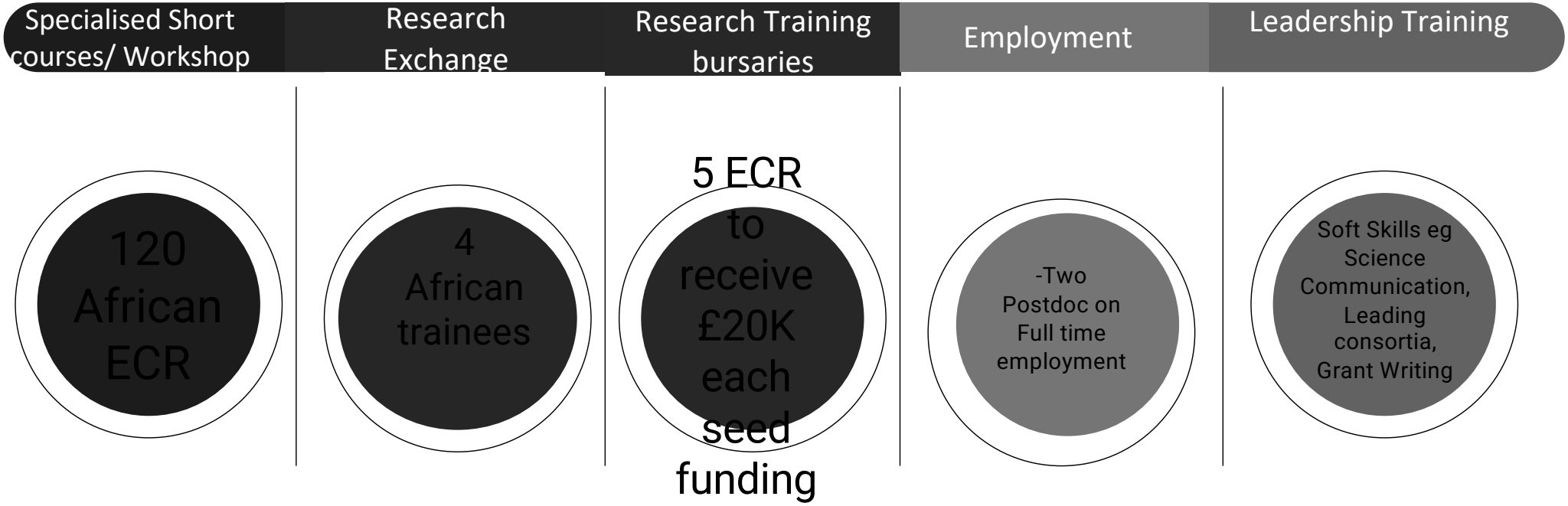
Funding: UKRI/MRC (2024-2029)

https://www.kidneygenafrica.org/

Multi-centre research and large-scale genomic studies

KIDNEYGENAFRICA: A PARTNERSHIP TO DELIVER RESEARCH AND TRAINING EXCELLENCE IN GENOMICS OF KIDNEY DISEASE IN AFRICA

KidneyGenAfrica Envisioned Excellence in Training and Capacity Building



THANK YOU and Welcome to Uganda