H3ABioNet
Pan African Bioinformatics Network for H3Africa

GWAS Pipelines

Scott Hazelhurst

University of the Witwatersrand, Johannesburg

8th H3A Consortium Meeting

H3ABioNet
Pan African Bioinformatics Network for H3Africa

GWAS pipelines overview

This talk gives only a flavour of steps that need to be done

- See SOP from H3A Bionet

Genome-Wide Association Studies

Overall goal: Know a disease has genetic factors, but don't know where

- Search across the entire genome (3×10^9 positions) for which variations where differences in cases/control

Genome-Wide Association Studies

Overall goal: Know a disease has genetic factors, but don't know where

- Search across the entire genome (3×10^9 positions) for which variations where differences in cases/control
- But sequencing a genome expensive (≈$1k per genome)

Genome-Wide Association Studies

Overall goal: Know a disease has genetic factors, but don't know where

- Search across the entire genome (3×10^9 positions) for which variations where differences in cases/control
- But sequencing a genome expensive (≈$1k per genome)
- SNP-chip: sample many positions (e.g. 2.5×10^6) across the genome (SNPs)
  Cost about $85 per person


Genome-Wide Association Studies

Overall goal: Know a disease has genetic factors, but don't know where

- Search across the entire genome (3×10^9 positions) for which variations where differences in cases/control
- But sequencing a genome expensive (≈\$1k per genome)
- SNP-chip: sample many positions (e.g. 2.5×10^6) across the genome (SNPs)
  Cost about \$85 per person
- Do statistical association tests




H3ABioNet

H3ABioNet

Issues:

- Which? – Choose ones where variation in humans – but this is population dependant.
- How many? NB: cost trade-off between number of people sampled, and number of positions sampled

Simple overview of genotyping

- At each sample position, each person has two copies (one from mother, one from father)
- Typically two possible *alleles*, e.g. most people have A, some people have G.
- SNP chip has a *probe*, which can detect which, for each SNP
- So three possible cases: e.g., AA, AG, GG

- DNA from each person progressively exposed to each probe

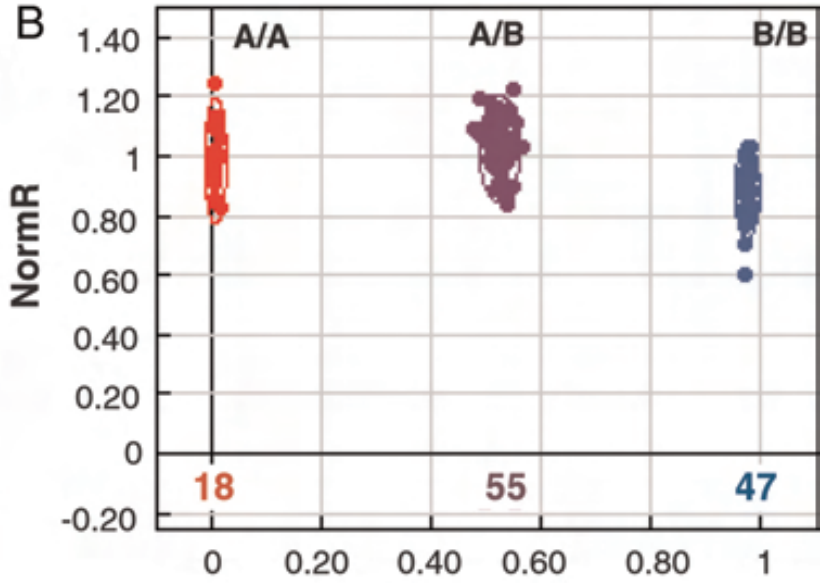
- DNA from each person progressively exposed to each probe
- “Red” light if probe sees one option, “Green” light if it sees the other

- DNA from each person progressively exposed to each probe
- “Red” light if probe sees one option, “Green” light if it sees the other
- Can get double red, green/red, or double green

- DNA from each person progressively exposed to each probe
- "Red" light if probe sees one option, "Green" light if it sees the other
- Can get double red, green/red, or double green
- For each person, SNP measure how much redness, greenness


- DNA from each person progressively exposed to each probe
- "Red" light if probe sees one option, "Green" light if it sees the other
- Can get double red, green/red, or double green
- For each person, SNP measure how much redness, greenness
- Noisy, needs calibration

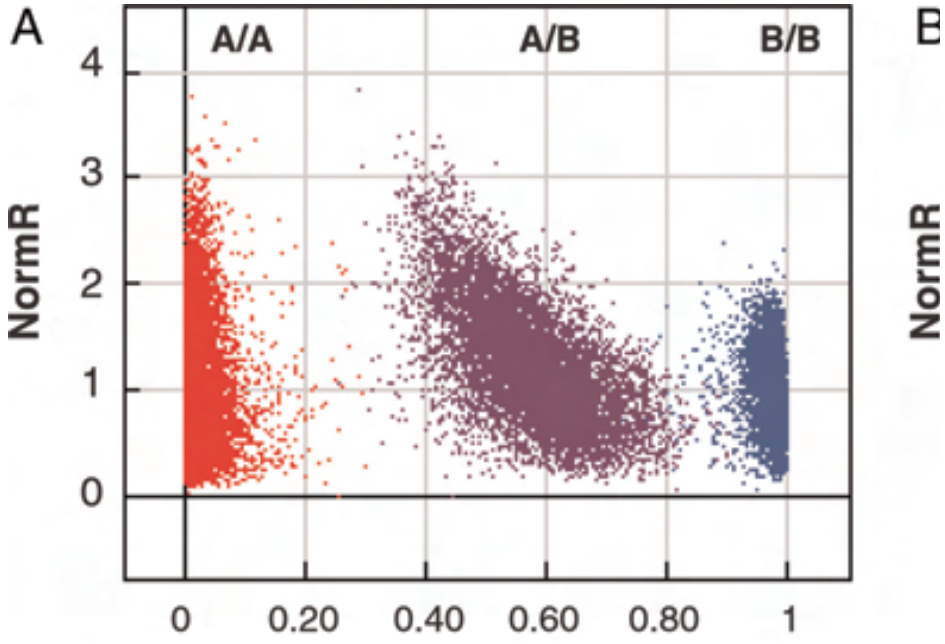
One SNP: each dot a person, measuring result from probe



Credit: Lamy et al, *Human Genetics*, 2011

Which cluster dot belongs to tells us what the reading is for that person

Image processing problem:




Variants are called

Converted from images into data files that can be analysed: PLINK is a standard tool, but there are others

- FAM file: describes the people in the study
- BIM file: which SNPs are captured and what the choices/alleles are
- BED file: actual data – what each person has for each SNP

Need for pipeline

GWAS is complex, computationally expensive, takes human time

- requires multiple steps, computers, techniques

Pipelines provide two big advantages:

- Must be reproducible
- Allow quick turn-around time

Overview of process

1. Genotype calling (computationally expensive):
   - QC: eg. batch effects
   - Convert from image to text (e.g. PLINK format)
2. QC on PLINK files
3. Population structure analysis
4. Imputation
5. Statistical testing


QC

Many things go wrong, data is noisy, QC essential

- Batch effects
- Replicates?
- Problems with SNPs and individuals
  - High missingness
  - Hardy-Weinberg Equilibrium
  - Minor Allele Frequency
  - Sample mix-up (check known sex)
  - Serious genetic problems or errors?
  - Relatedness?

Population Structure

Apparent genetic diversity within sample may be a big issue

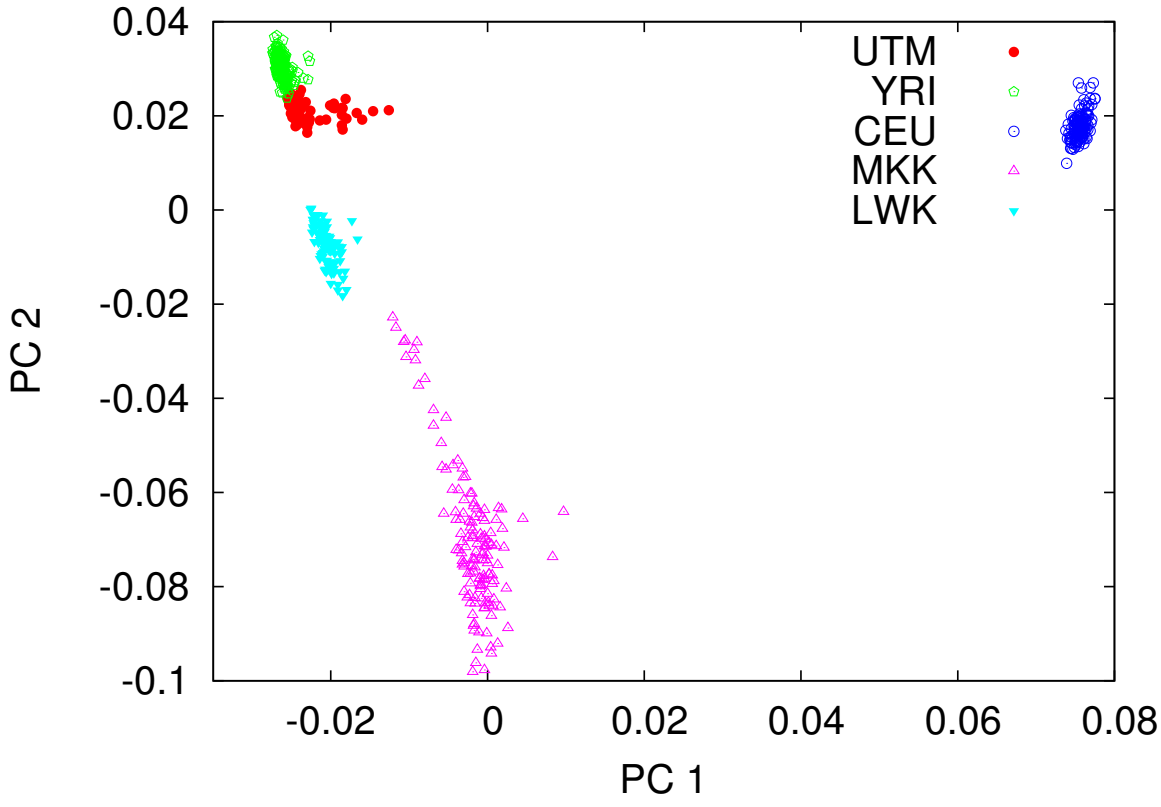
- real genetic diversity in population
- poor choice of cases/controls
- artefactual – e.g., batch effect

If not managed you will get false positives.

Analysing population structure

- principal component analysis
- structure/admixture

PC analysis is particularly used for GWAS


0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
PC 2
-0.02
0
0.02
0.04
0.06
0.08
PC 1
UTM
YRI
CEU
MKK
LWK

How you handle depends on structure

- tight cluster
- looser uniform cluster
- very variable cluster/poorly define cluster
- multiple clusters
- ...

Imputation

Limitation of SNP-Chip is that it only samples a small proportion of genome

- May miss many SNPs, other types of variation
- Associated SNP may not be close to cause

Imputation

Limitation of SNP-Chip is that it only samples a small proportion of genome

- May miss many SNPs, other types of variation
- Associated SNP may not be close to cause

With good reference genomes for the pops can impute

- statistically predict what the intermediate SNPs are.

Imputation

Limitation of SNP-Chip is that it only samples a small proportion of genome

- May miss many SNPs, other types of variation
- Associated SNP may not be close to cause

With good reference genomes for the pops can impute

- statistically predict what the intermediate SNPs are.

Imputation

Limitation of SNP-Chip is that it only samples a small proportion of genome

- May miss many SNPs, other types of variation
- Associated SNP may not be close to cause

With good reference genomes for the pops can impute

- statistically predict what the intermediate SNPs are.

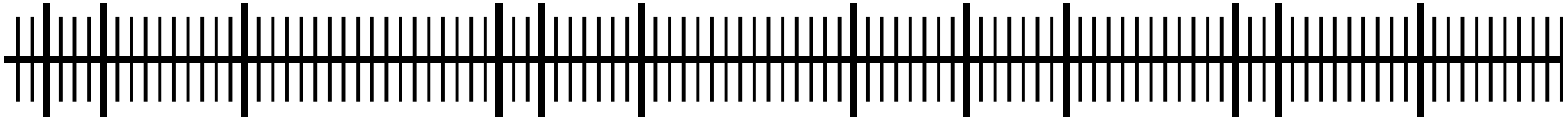
Imputation

Limitation of SNP-Chip is that it only samples a small proportion of genome

- May miss many SNPs, other types of variation
- Associated SNP may not be close to cause

With good reference genomes for the pops can impute

- statistically predict what the intermediate SNPs are.



Covariates

Are there environmental factors that might affect disorder?

- age
- sex
- smoking
- other lifestyle issues
- other phenotype: e.g. with T2D might include BMI ...

Statistical testing

Requires expert intervention

- What question?
- How do genetic factors manifest (e.g., recessive, dominant)
- Degree of relatedness
- What covariates
- Population structure
- Interactions

Statistical testing gives two results

- p-value: how statistically likely the result is.
  Need to take into account multiple testing, so will be very strict with cut off
- Odds ratio/Effect size: $(0, \infty)$
  How big an effect does the SNP have – often small

Post-association test analysis

Now we have some matching results, is there a story – what effect do the SNPs have

- identify genes or other parts of the genome where the SNPs can be found
- identify metabolic pathways – set of processes determined or regulated by a set of genes etc that perform some important function
- Need insight into biology


Archiving

Need to make a choice of what data must be kept

- Raw image files: ≈3GB per person (30TB for 10k people)
- Intermediate analyses: YMMV but probably similar
- PLINKed format: probably 2MB per person (30GB for 10k people); may want several copies
- Meta data, subsidiary analyses
- May be multiple versions, different parameters
  Need to go back to the data and know how produced.

H3ABioNet funded by NHGRI grant number U41HG006941