Association Testing for GWAS

Scott Hazelhurst

June 2014

Huge growth in studies

- Feb 2013: $>$ 1500 GWAS studies in the NIH catalogue
  `http://www.genome.gov/gwastudies`

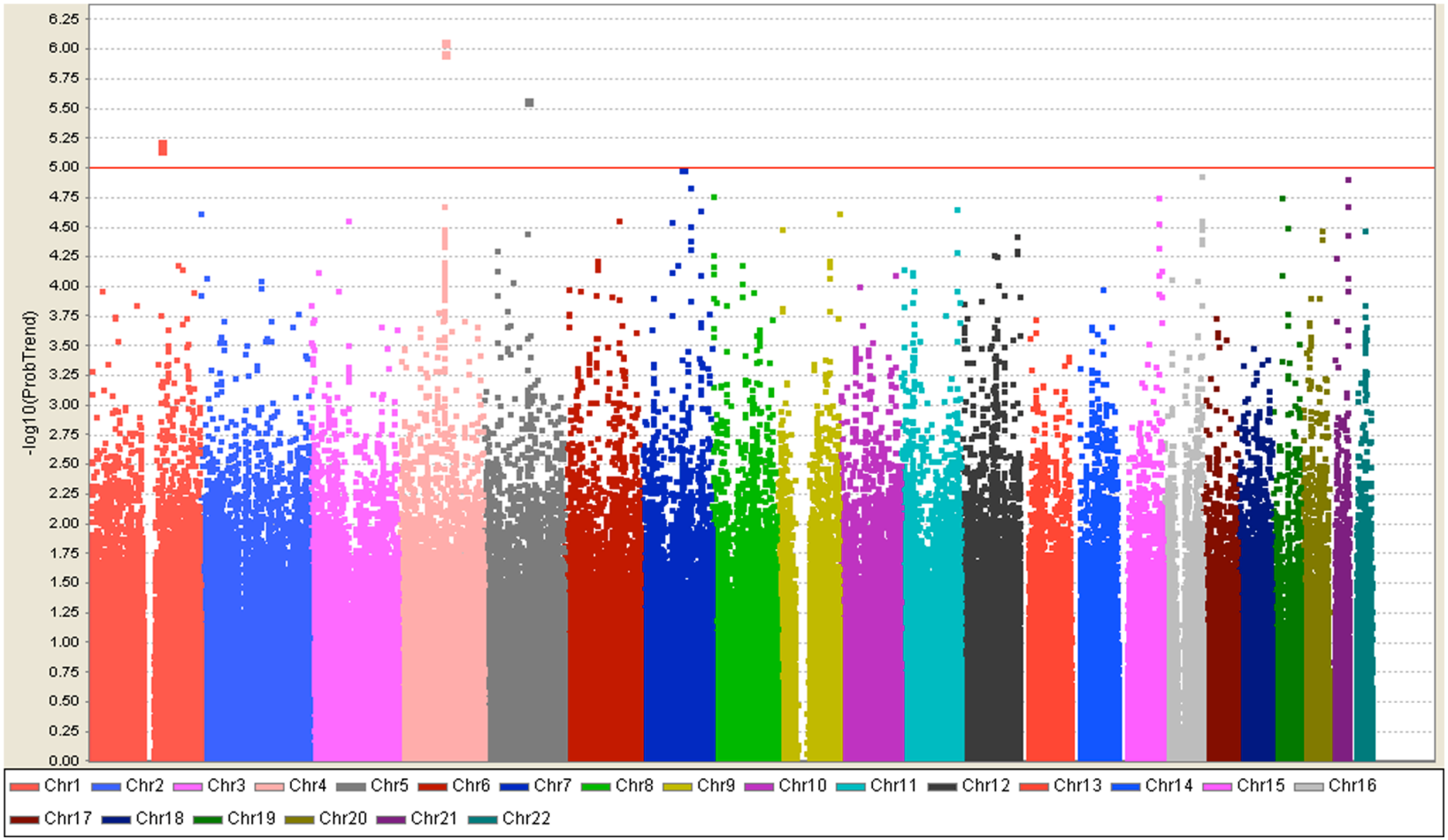
Huge growth in studies

- Feb 2013: $>$ 1500 GWAS studies in the NIH catalogue
  `http://www.genome.gov/gwastudies`
- Feb 2014 $>$ 2000 GWAS studies
- Provides new tools for insights into complex diseases
- Understanding of population history

Genome Wide Association Studies Overview

The goal of GWAS is to "seek strong associations between phenotype and genetic variations that represent genomically proximal causal genetic effects" [Zhang et al.]

- Don't need *a priori* knowledge of responsible SNPs;
  Typically use $> 10^5$ SNPs
- Can be used for complex disorders – multiple contributing components
- Can detect small effects


-log10(ProbTrend)
6.25
6.00
5.75
5.50
5.25
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16
Chr17 Chr18 Chr19 Chr20 Chr21 Chr22

GWAS very powerful tool

. . . but some fundamental limitations . . .

Hek *et al*, *Biological Psychiatry*, 2013.

ARCHIVAL REPORT

A Genome-Wide Association Study of Depressive Symptoms

Karin Hek, Ayse Demirkan, Jari Lahti, Antonio Terracciano, Alexander Teumer, Marilyn C. Cornelis, Najaf Amin, Erin Bakshis, Jens Baumert, Jingzhong Ding, Yongmei Liu, Kristin Marciante, Osorio Meirelles, Michael A. Nalls, Yan V. Sun, Nicole Vogelzangs, Lei Yu, Stefania Bandinelli, Emelia J. Benjamin, David A. Bennett, Dorret Boomsma, Alessandra Cannas, Laura H. Coker, Eco de Geus, Philip L. De Jager, Ana V. Diez-Roux, Shaun Purcell, Frank B. Hu, Eric B. Rimm, David J. Hunter, Majken K. Jensen, Gary Curhan, Kenneth Rice, Alan D. Penman, Jerome I. Rotter, Nona Sotoodehnia, Rebecca Emeny, Johan G. Eriksson, Denis A. Evans, Luigi Ferrucci, Myriam Fornage, Vilmundur Gudnason, Albert Hofman, Thomas Illig, Sharon Kardia, Margaret Kelly-Hayes, Karestan Koenen, Peter Kraft, Maris Kuningas, Joseph M. Massaro, David Melzer, Antonella Mulas, Cornelis L. Mulder, Anna Murray, Ben A. Oostra, Aarno Palotie, Brenda Penninx, Astrid Petersmann, Luke C. Pilling, Bruce Psaty, Rajesh Rawal, Eric M. Reiman, Andrea Schulz, Joshua M. Shulman, Andrew B. Singleton, Albert V. Smith, Angelina R. Sutin, André G. Uitterlinden, Henry Völzke, Elisabeth Widen, Kristine Yaffe, Alan B. Zonderman, Francesco Cucca, Tamara Harris, Karl-Heinz Ladwig, David J. Llewellyn, Katri Räikkönen, Toshiko Tanaka, Cornelia M. van Duijn, Hans J. Grabe, Lenore J. Launer, Kathryn L. Lunetta, Thomas H. Mosley Jr., Anne B. Newman, Henning Tiemeier, and Joanne Murabito

Background: Depression is a heritable trait that exists on a continuum of varying severity and duration. Yet, the search for genetic variants associated with depression has had few successes. We exploit the entire continuum of depression to find common variants for depressive symptoms.

Methods: In this genome-wide association study, we combined the results of 17 population-based studies assessing depressive symptoms with the Center for Epidemiological Studies Depression Scale. Replication of the independent top hits ($p < 1 \times 10^{-5}$) was performed in five studies assessing depressive symptoms with other instruments. In addition, we performed a combined meta-analysis of all 22 discovery and replication studies.

Results: The discovery sample comprised 34,549 individuals (mean age of 66.5) and no loci reached genome-wide significance (lowest $p = 1.05 \times 10^{-7}$). Seven independent single nucleotide polymorphisms were considered for replication. In the replication set ($n =$

- Starting point: previous GWASs with 34k individuals
- Selected top 7 SNPs in a meta-analysis
- Replicate in new studies (17k individuals)
- Only 1 replicated
- Also: of 17 SNPs identified by other work, none replicate

Message more complex than this, but highlights the challenges of GWAS

A review paper of 2012, starts off

- Introduction: Have GWASs been a failure?

  There comes a point at which the genetic skeptic can be pardoned the suggestion that if the genes are so small and so multiple, what they are hardly matters, the dividing line between polygenes and no genes is of little practical consequence. Have we reached this point? [Crow, 2011]

We'll come back to this . . .

GWAS study approach...

For each individual, each SNP

- record what variant individual has

Is it:

- Major, minor allele?

Assume only biallelic.

- Minor allele frequency (MAF) is important measure

Power of GWAS

Design of GWAS must take into account power:

Effect size

What is the relative risk for a SNP?

Power of GWAS

Design of GWAS must take into account power:

Effect size

What is the relative risk for a SNP?

Power

What proportion of SNPs with this effect size do you wish to capture?

Depends on:

- Number of SNPs chosen
- LD-structure given population
- Ability to impute on population
- Number of samples
- MAFs of SNPs.
- Structure of the population
- Relatedness

Absolutely critical to success of project – pre-knowledge of population very beneficial.

Measuring phenotype

Correctly measuring phenotype a big issue:

- Categorical/quantitative
- Objectively/subjectively measured
- Consistency is crucial — difficult for large projects
- Can be measured specifically for a project, or retrospectively (e.g., from medical records)

Measuring effect

Expect MAF ratio to be same in cases/controls.
Deviation measured by

- Odds ratio: a generalised measure of the proportion of MAF in cases to controls.
  Deviation from 1 may be significant.
  Increase of probability of being a case.
- Statistical significance: χ^2 test $\Longrightarrow$ p-value (unadjusted)
- Absolute MAF difference
  Reality check: within error margin of sample?
  what evolutionary significance?

Basic statistical test

For case/control tests, most common test is χ^2 test.

	A1	A2
Cases	u_1	u_2
Control	v_1	v_2

- Generate p value from this
- $OR = \frac{u_2/u_1}{v_2/v_1}$

Starting point should be biological model...

What is the effect?

- Recessive?
- Dominant?
- Additive?

Correct statistical test proceeds from this.

- Many options available

Quantitative traits

For QT studies, linear regression is used.

- For each SNP j, build a linear model $y = \beta x_j + \mu + e$

Key parameters are:

- y phenotype
- x_j : SNP value at position j.
- β – slope
- μ – intercept
- e – error

Solved using a least-squares method

- use known data to build model
- can then evaluate model against data to get test statistic – a p value
- can extend to handle covariates
- does support different inheritance models
- what if there is a non-linear relationship?

Multiple testing

To recap basic stats:

- null hypothesis: no association
- p value is probability true
- So rejecting at $p = 0.05$ says 95% of the time, null hypothesis is correctly rejected as false, but 5% of time actually true

Interlude – sample versus data size

Consider recent study in *New England Journal of Medicine*

- Tiotropium versus Salmeterol for the Prevention of Exacerbations of COPD (doi10.1056/NEJMoa100837)

Interlude – sample versus data size

Consider recent study in *New England Journal of Medicine*

- Tiotropium versus Salmeterol for the Prevention of Exacerbations of COPD (`doi10.1056/NEJMoa100837`)

Experimental design

- tiotropium (3707 patients);
- salmeterol (3669 patients)

For each, measure

- effectiveness of drug in controlling disease.

GWAS

10000 individuals

1 million data points each

"Conventional"

7300 individuals

1 data point (which drug) each

Multiple testing a key issue

Family-wise error rate

Given a family of tests:

- What is the probability of incorrectly rejecting a null hypothesis

FWER

Bonferroni correction

Strictest multiple-testing correction is Bonferroni. If the desired FWER is α', then choose $\alpha = \frac{\alpha'}{n}$

- If you want all your tests to be correct at the 0.05 level,
- And there are a million SNPs
- Then, your unadjusted p-value must be less than 5×10^{-8}

Problem with BF is that it is very strict

- Assumes each test is independent
- But: LD means that they are not
- So effectively the number of SNPs is much smaller

Lots of alternative methods: e.g., FDR (Benjamini-Hochberg), permutation testing

Covariates

Need to handle other factors, e.g.,:

- Environmental (e.g., does the person smoke?)
- Age, gender
- Population structure

Confounding

Covariant not independent of genotype

- e.g., population structure
- Ignoring yields false positives
- Must take into account in analysis

Confounding

Covariant not independent of genotype

- e.g., population structure
- Ignoring yields false positives
- Must take into account in analysis

Non-confounding

- e.g., smoking, . . .
- Tension between increasing power, reduces precision of estimator
- Rough guide: when disease prevalence high include; otherwise not.
- Do you sample according to covariate?
- Complex issues. . .

Population structure

Huge genetic diversity in human populations

- great diversity in African populations

Interesting in its own right

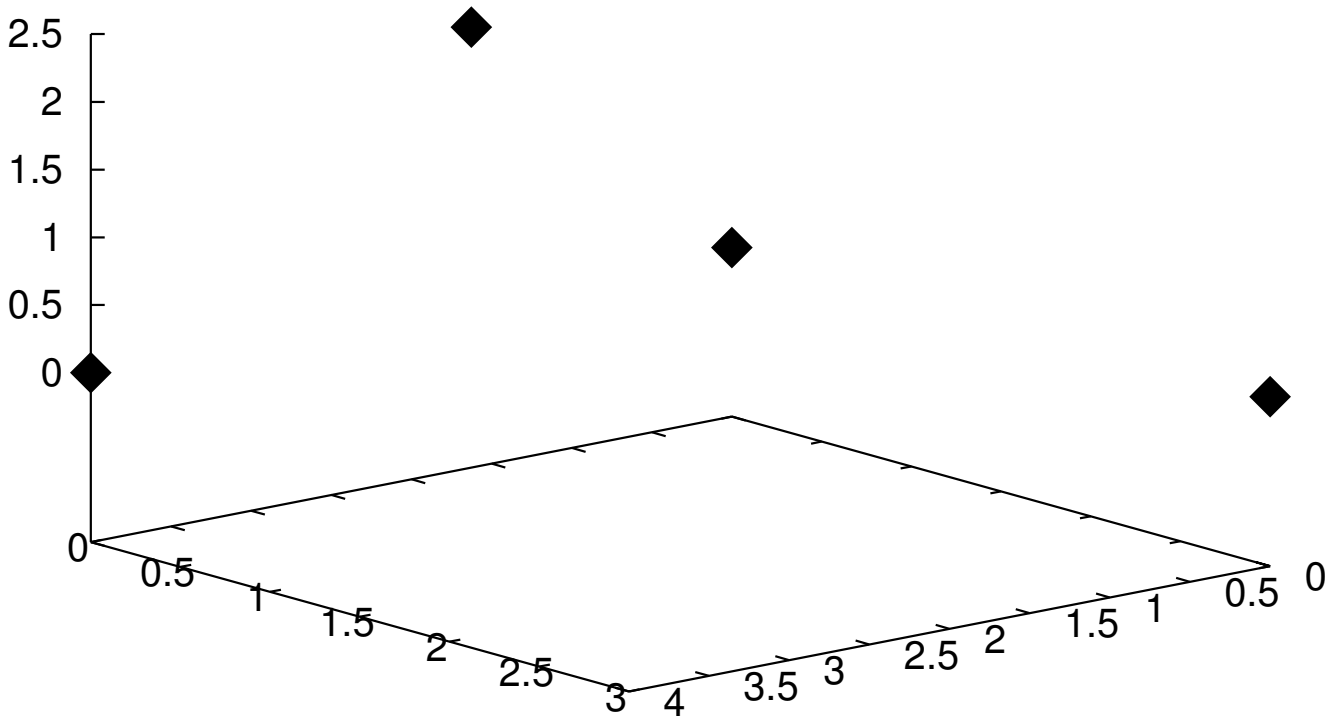
- Has implications for GWAS

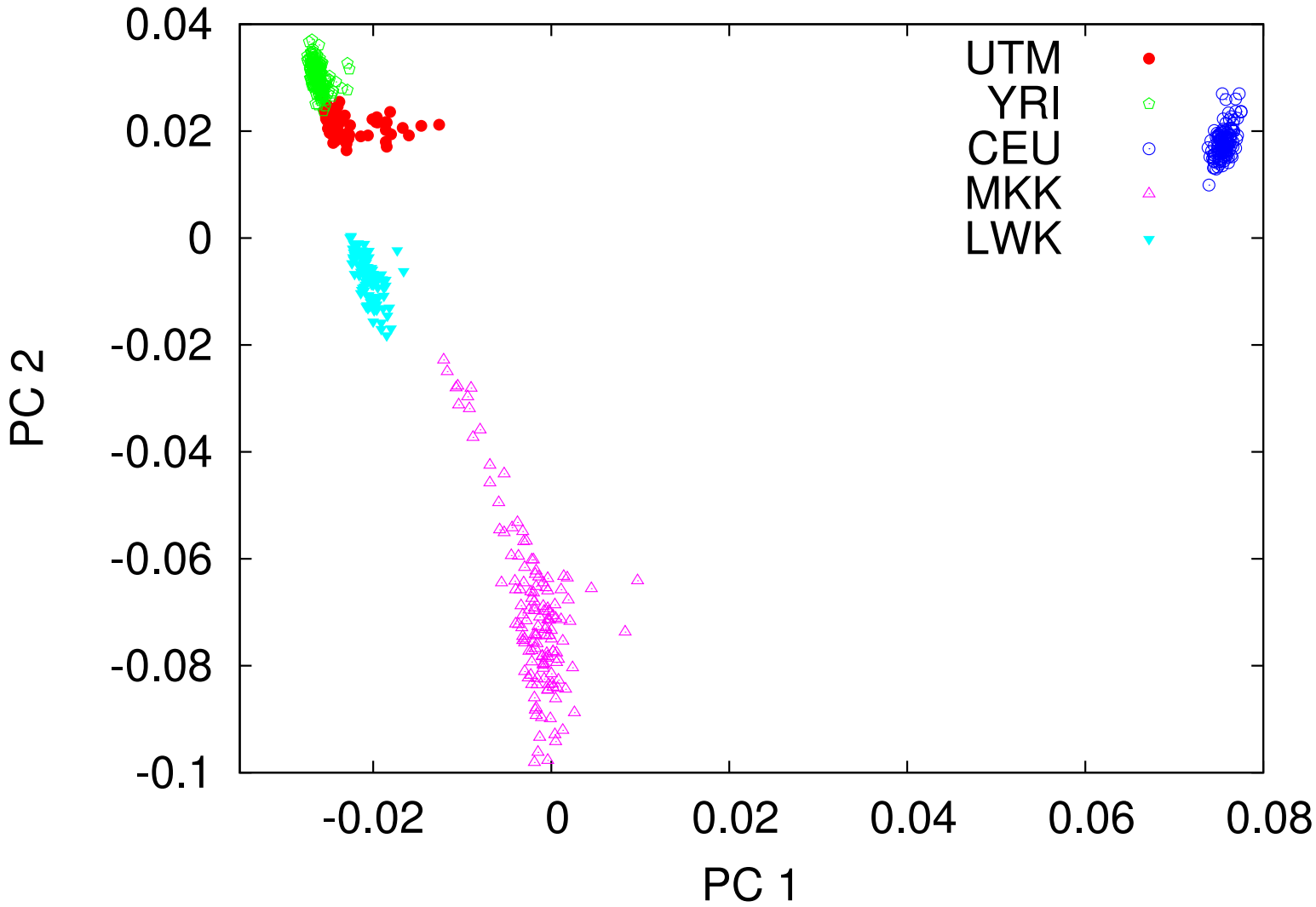
A major impact of H3A projects will be to generate novel African data.

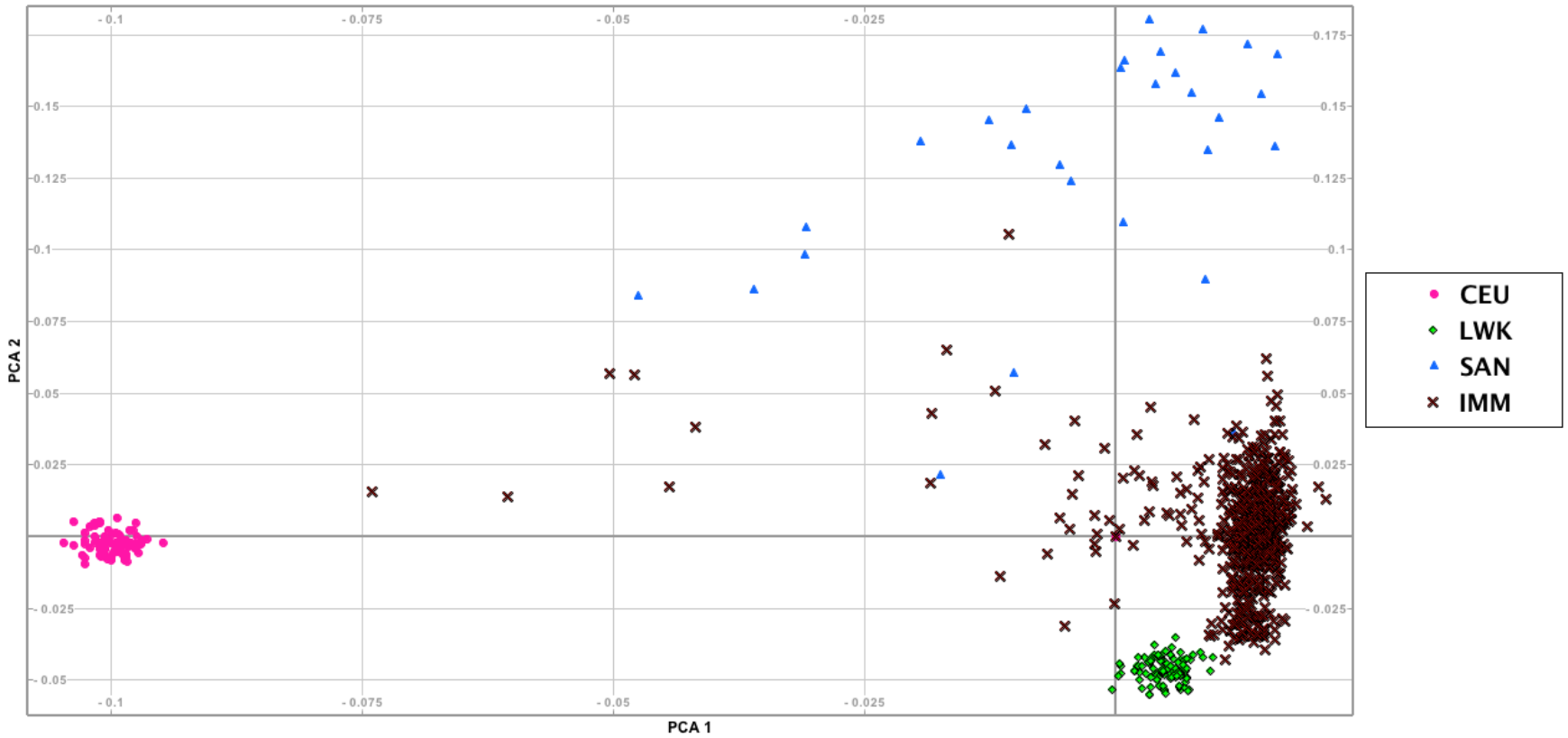
From genotype, sequence data can build mathematical models to visualise

- differences between populations;
- homogoneity/heterogeneity within population

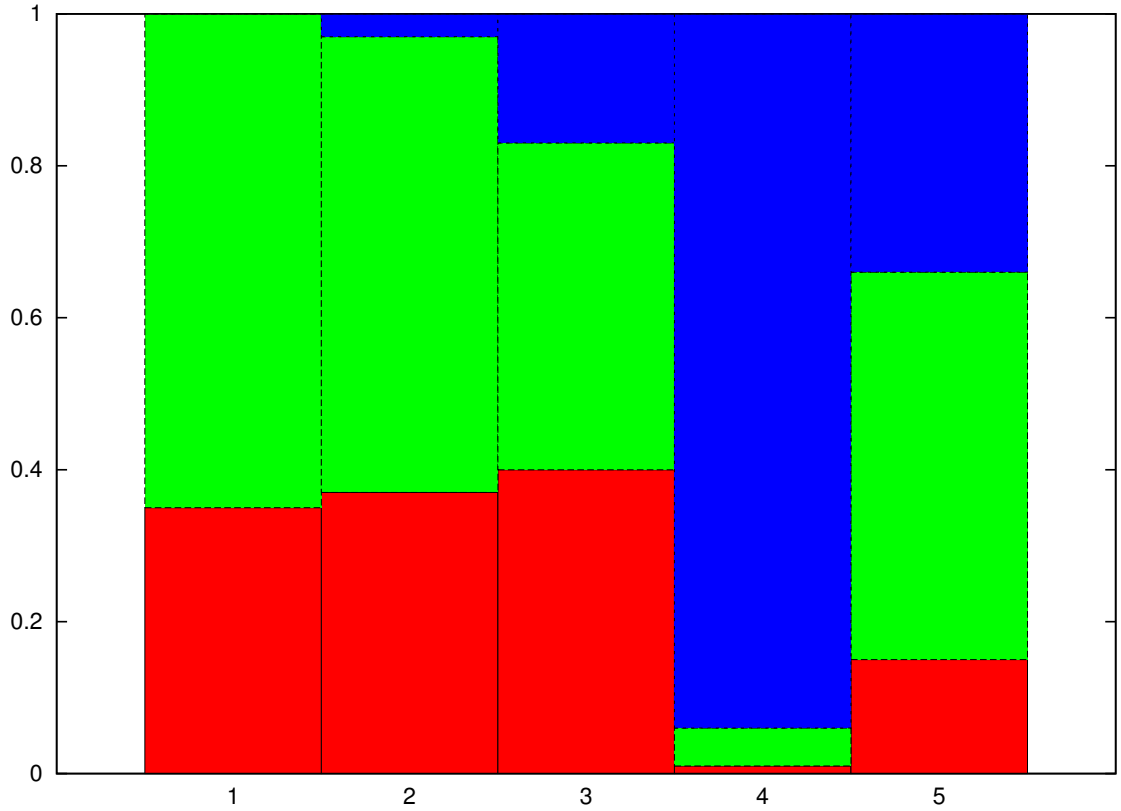
Principal component analysis

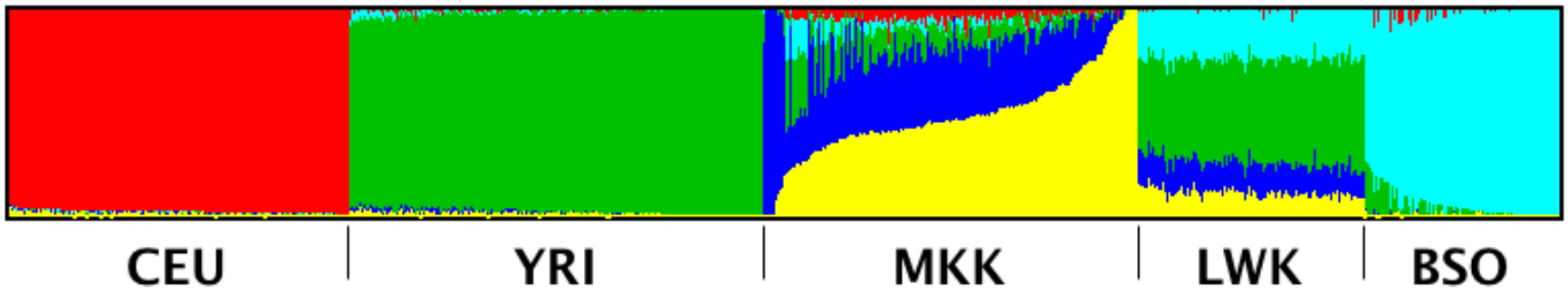



0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
PC 2
-0.02
0
0.02
0.04
0.06
0.08
PC 1
UTM
YRI
CEU
MKK
LWK


- 0.1
- 0.075
- 0.05
- 0.025
0.175
0.15
0.125
0.1
0.075
0.05
0.025
PCA 2
PCA 1
CEU
LWK
SAN
IMM

Structure chart




CEU
YRI
MKK
LWK
BSO

Managing population structure

Population structure a confounding factor

- May be completely false – sampling error
- May be linked but not causal via other genetic factors.

Suppose sample brown bears and black bears and more brown bears are cases

- GWAS will pick up Ancestry Informative Markers (AIMs)
- "Real" signal way down list.

Designing your experiment

- Matching cases/controls

Is structure a problem?

Perform PCA, check association with axes

e.g, eigenstrat tells you if there is association with case/control against any axes.

Dealing with structure

- Remove outliers
- Adjust for genomic control
- Perform meta-analysis
- Use structure as covariate
- Mixed-model method

There's lots more

- Imputation
- Multiple locus tests
- Epistatis
- Copy number variation
- Replication studies!!!
- Functional studies!!!

Computational resources

GWAS needs decent computational resources

- Access to a cluster very very helpful (but not absolutely necessary)

Computational cost a small part of the cost of the whole study

- If it takes a week or two longer, not the bottleneck
- Relatively small investment compared to other costs

Planning and sharing

Need a plan for computational resources

- not just about one project/analysis
- plan to grow sensibly

Planning and sharing

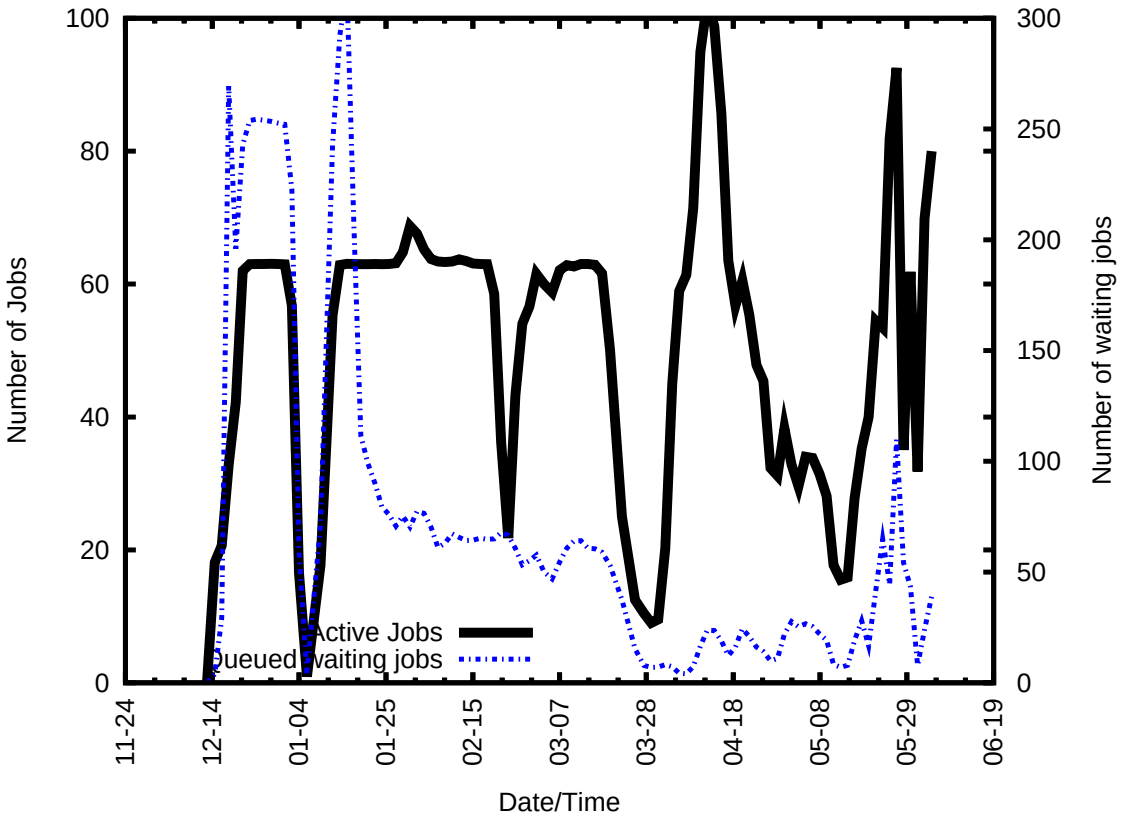
Need a plan for computational resources

- not just about one project/analysis
- plan to grow sensibly

Key issues:

- operating system
- computing power
- RAM
- disk space
- skills and capacity

Can you share?



- with others
- with your-self. . . virtualisa

Operating system

Use Linux:

- Can get good value from Windows and Apple machines, but central resource should be Linux
- We recommend Scientific Linux 6 or Ubuntu

Invest in training:

- system admin
- command line

Computing power

For GWAS and NGS, easiest component, easiest to scale

- Get several quotes and don't pay premium for 10% extra power.

RAM

Critical issue: if your CPU is under-powered you can take a few days longer. If you don't have enough RAM you can't do it.

- Rough guide: 16GB of RAM for an average size GWAS
- But: if doing population studies you may need much more.
- If you are wanting to parallelise, key issue is RAM/core
- For NGS you may need much more depending

Plan for future upgrade – even if it costs more

Computer will support maximum number of DIMMs – ask your supplier how many

- 8GB DIMM: USD120 – USD15/GB
- 16GB DIMM: USD310 – USD20/GB

Rather get 16GB (or even 32GB, 64GB)

- Not all computers need large RAM.

Disk space

Long term planning an issue

- Disaster recovery plan?
- How much do you want to grow?
- Disaster recovery plan?
- Security
- Disaster recovery plan?
- Meta-data/catalogue
- Disaster recovery

Issue is file system rather than cost of individual disk

Sample system:

Server, 12 cores, 8 DIMM slots	4200
32GB RAM (2 DIMMs)	600
4TB of disk	1000
	6000

Plus some backup?

H3ABionet

- Provides training (system, application)
- Has SOPs
- Can help advise