Genome-wide and family-based association studies

Dmitry Prokopenko

dprokopenko@mgh.harvard.edu

Last lecture

- Basic terminology
- What is linkage
- Linkage methods
- What is association

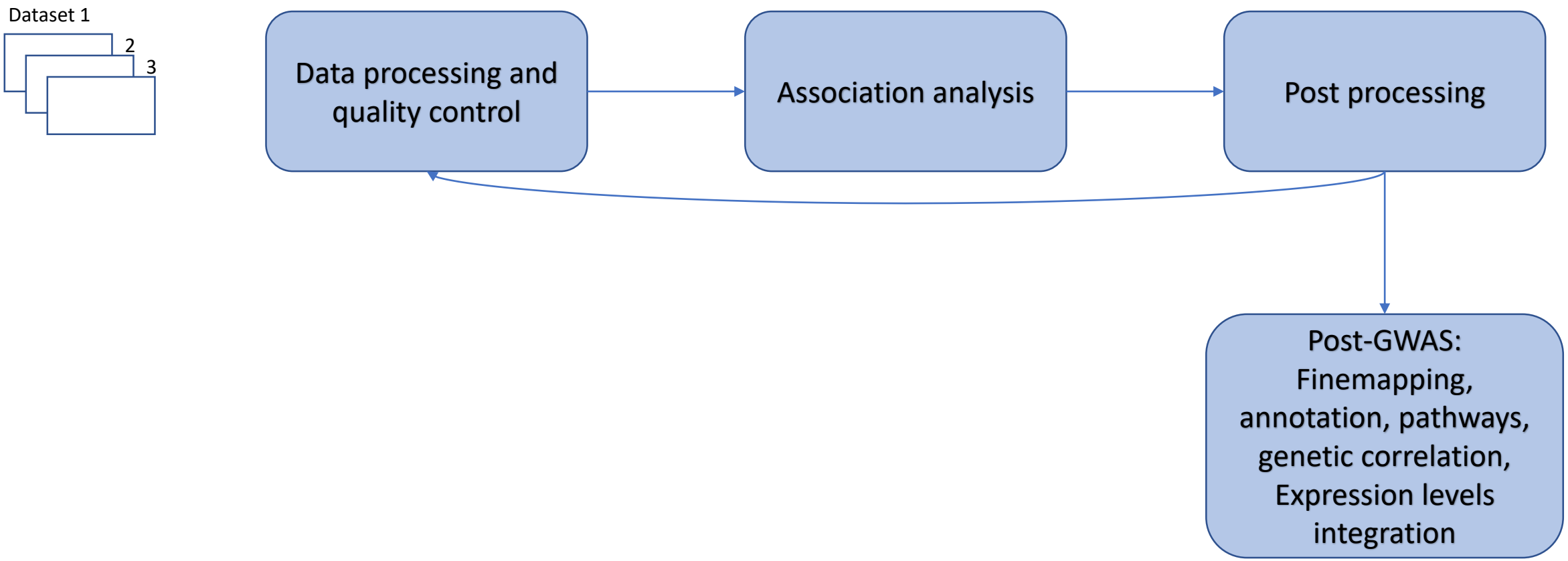
This lecture

- Association methods
  - Data quality
    - Variant-based quality
    - Sample-based quality
  - Population-based association studies, aka GWAS, aka common variant association studies
  - Family-based association studies
  - Rare variant association studies

Software?

- Plink 1.9, 2alpha (https://www.cog-genomics.org/plink2) - universal
- Bcftools – fast for vcf/bcf format (https://samtools.github.io/bcftools/bcftools.html)
- Oxford set of tools (gen/bgen format) (https://www.well.ox.ac.uk/~gav/bgen_format/software.html)
- BOLT-LMM – fast LMM models
- GCTA – originally for heritability estimation, now pretty universal
- Good old R / python

GWAS workflow



Quality control (QC)

- Variant-based
  - Calling quality
  - Variant missingness rate
  - Deviance from HWE
  - Mendelian consistency

- Sample-based
  - Cryptic relatedness
  - Population structure
  - Inbreeding coefficient
  - Wrong pedigree information
  - Sex verification (based on X)

Anderson et al. 2011

GWAS / WGAS

- Number of samples (n)
  - 500-500,000 (UK Biobank)
  - Larger n -> more statistical power and more computational burden
- Number of SNVs (m)
  - Genotyped
    - 500,000 -1,700,000
    - Illumina MEGA Ex array ~1.7M (Multi-Ethnic Global)
  - Imputed
    - 8M common variants based on HRC imputation panel (n=65k)

GWAS / WGAS 2

- Whole Exome Sequence (WES)
  - 1-2M variants in exome regions only (coding regions)

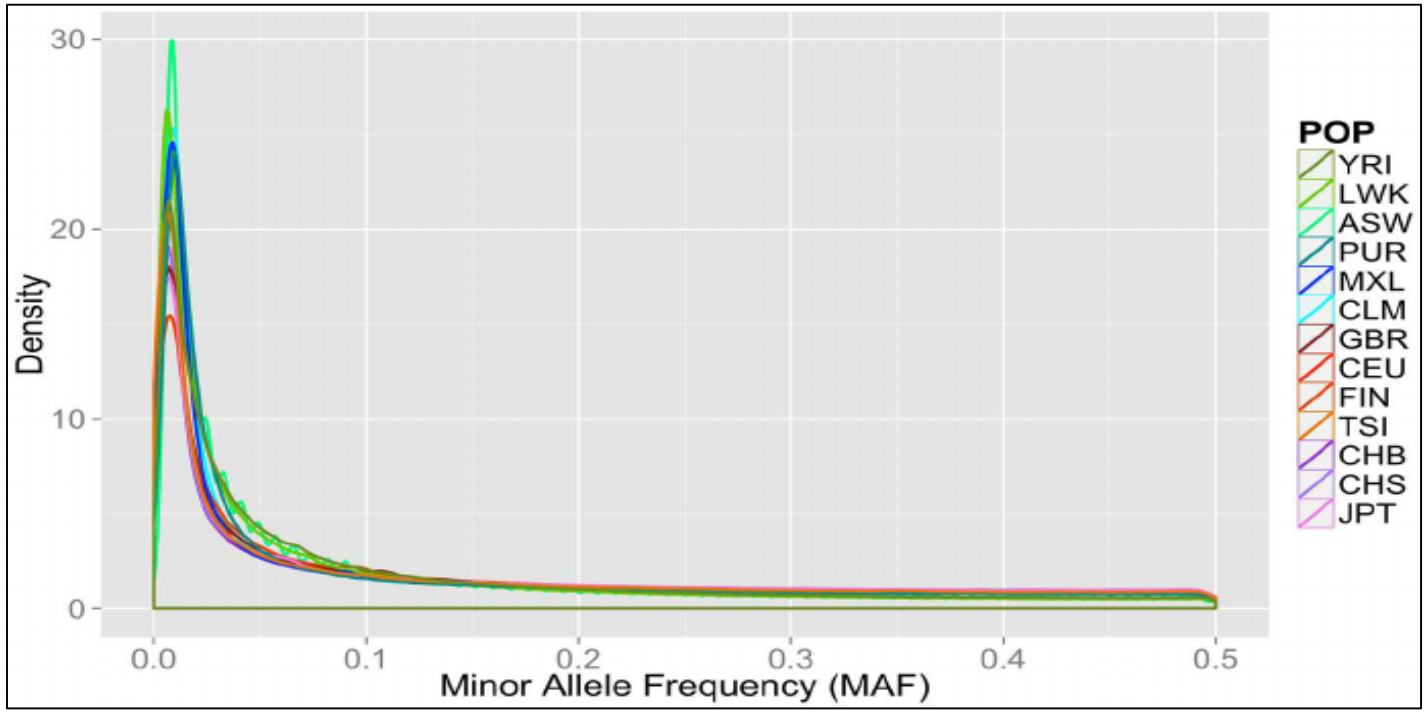
- Whole Genome Sequence (WGS)

N samples	M variants
4300	87M
10600	141M
19k	219M
65k	582M
122k	721M

https://www.nhlbiwgs.org/genetic

MAF spectrum

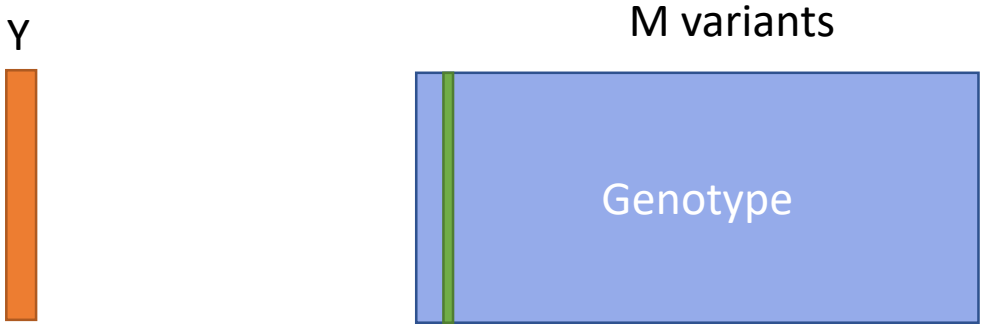
- Mathematically described by the Ewen's sampling formula
- Rule of thumb: 70% of variants below 5% (Visscher, Goddard, Derks, Wray 2012)



WJ Ewens. Mathematical population genetics

Population-based association studies

- Find variants that are statistically associated with phenotype
  1. Dichotomous phenotype
     - Case versus control
  2. Quantitative phenotype
     - Height
     - Blood markers
     - Gene expression measures (QTL)
  3. Time-to-event
     - Survival
     - Time to onset



Case/control

```
test<-mutate(ADSP,affected=ifelse(affected==0,NA,affected-1))
tab<-table(test$affected,test$rs429358)
print(tab)
```

```
##
##        0    1    2
##   0 3907  527   15
##   1 3137 1986  163
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 989.65, df = 2, p-value < 2.2e-16
```

- H0: Genotype frequencies are same for cases and controls
- Alternatives include allelic test (1df), Fisher exact test, Cochran-Armitage trend

Linear regression / logistic regression

- Can account for covariates and better correct for population stratificiation

$$\mathbf{Y} \sim \boldsymbol{\beta G} + \boldsymbol{\gamma Z} + \boldsymbol{\varepsilon}$$

- Identity link function for linear regression
- Logit link function for logistic regression

$$P(y_i = 1) = \frac{\exp(\alpha + \beta g_i + \gamma z_i)}{1 + \exp(\alpha + \beta g_i + \gamma z_i)}$$

Example

```
fit<-glm(affected~rs429358+sex+PC1+PC2,family = binomial(),data=test)
summary(fit)
```

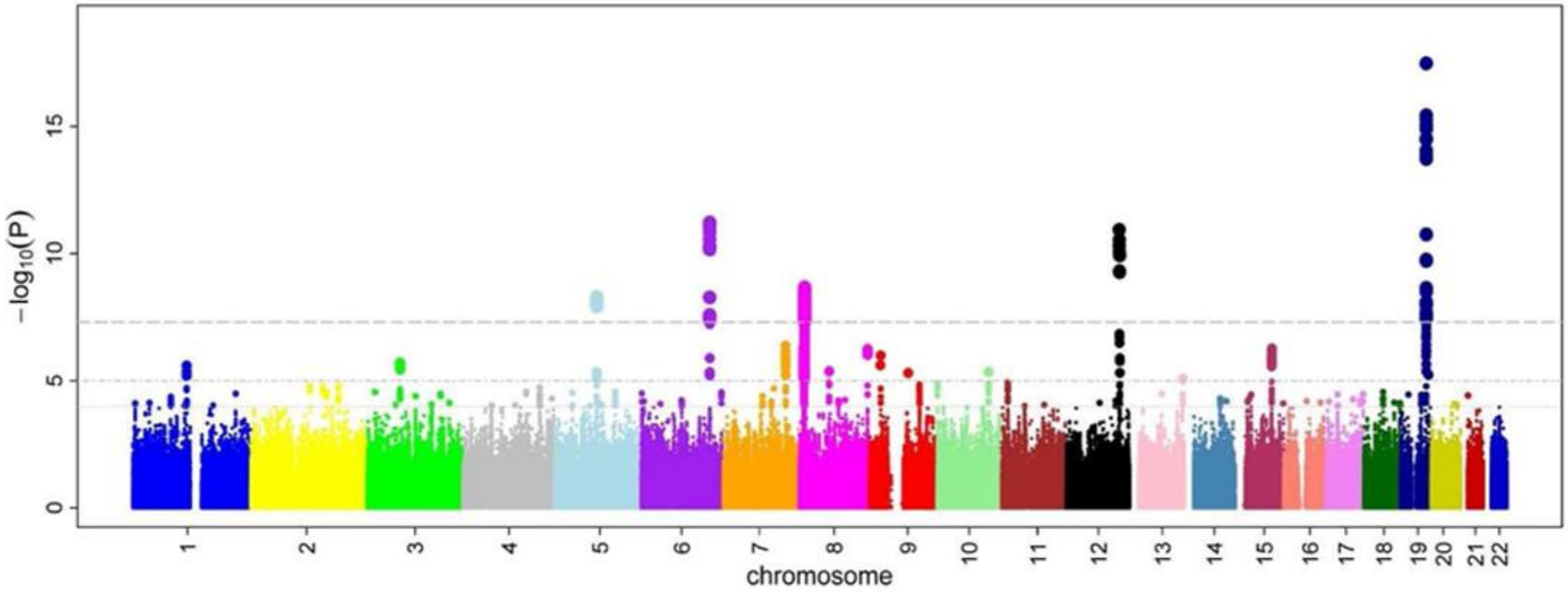
```
##
## Call:
## glm(formula = affected ~ rs429358 + sex + PC1 + PC2, family = binomial(),
##     data = test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4075  -1.0843   0.6856   1.2627   1.3097
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.27128    0.07371  -3.681 0.000233 ***
## rs429358     1.52019    0.05240  29.009  < 2e-16 ***
## sex          0.03440    0.04365   0.788 0.430584
## PC1          3.13185    2.16073   1.449 0.147214
## PC2         -1.87293    2.18200  -0.858 0.390697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13424  on 9734  degrees of freedom
## Residual deviance: 12364  on 9730  degrees of freedom
##   (648 observations deleted due to missingness)
## AIC: 12374
##
## Number of Fisher Scoring iterations: 3
```

Example plink

- plink19 --bfile filename --pheno phenofile --pheno-name Affection.Status --covar covarfile --covar-name PC1-PC5,sex --logistic --out outname

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	1:762159:T:C	762159	C	ADD	10154	0.9459	-0.07175	0.9428
1	1:762159:T:C	762159	C	sex	10154	0.957	-1.08	0.2801
1	1:762159:T:C	762159	C	PC1	10154	0.8619	-0.0715	0.943
1	1:762159:T:C	762159	C	PC2	10154	0.001005	-3.339	0.0008419
1	1:762159:T:C	762159	C	PC3	10154	0.07598	-1.265	0.2059
1	1:762159:T:C	762159	C	PC4	10154	2.857e-10	-8.558	1.146e-17
1	1:762159:T:C	762159	C	PC5	10154	0.2254	-0.4586	0.6465
1	1:861368:C:T	861368	T	ADD	10020	0.7031	-0.2489	0.8034
1	1:861368:C:T	861368	T	sex	10020	0.9481	-1.3	0.1936
1	1:861368:C:T	861368	T	PC1	10020	0.6602	-0.1993	0.8421
1	1:861368:C:T	861368	T	PC2	10020	0.001068	-3.288	0.001009
1	1:861368:C:T	861368	T	PC3	10020	0.09682	-1.137	0.2556
1	1:861368:C:T	861368	T	PC4	10020	3.386e-10	-8.463	2.609e-17
1	1:861368:C:T	861368	T	PC5	10020	0.2167	-0.4703	0.6381
1	1:865628:G:A	865628	A	ADD	10154	0.762	-1.504	0.1326
1	1:865628:G:A	865628	A	sex	10154	0.9564	-1.095	0.2734
1	1:865628:G:A	865628	A	PC1	10154	0.8359	-0.08621	0.9313
1	1:865628:G:A	865628	A	PC2	10154	0.001047	-3.318	0.0009056
1	1:865628:G:A	865628	A	PC3	10154	0.07233	-1.289	0.1974

Manhattan plot

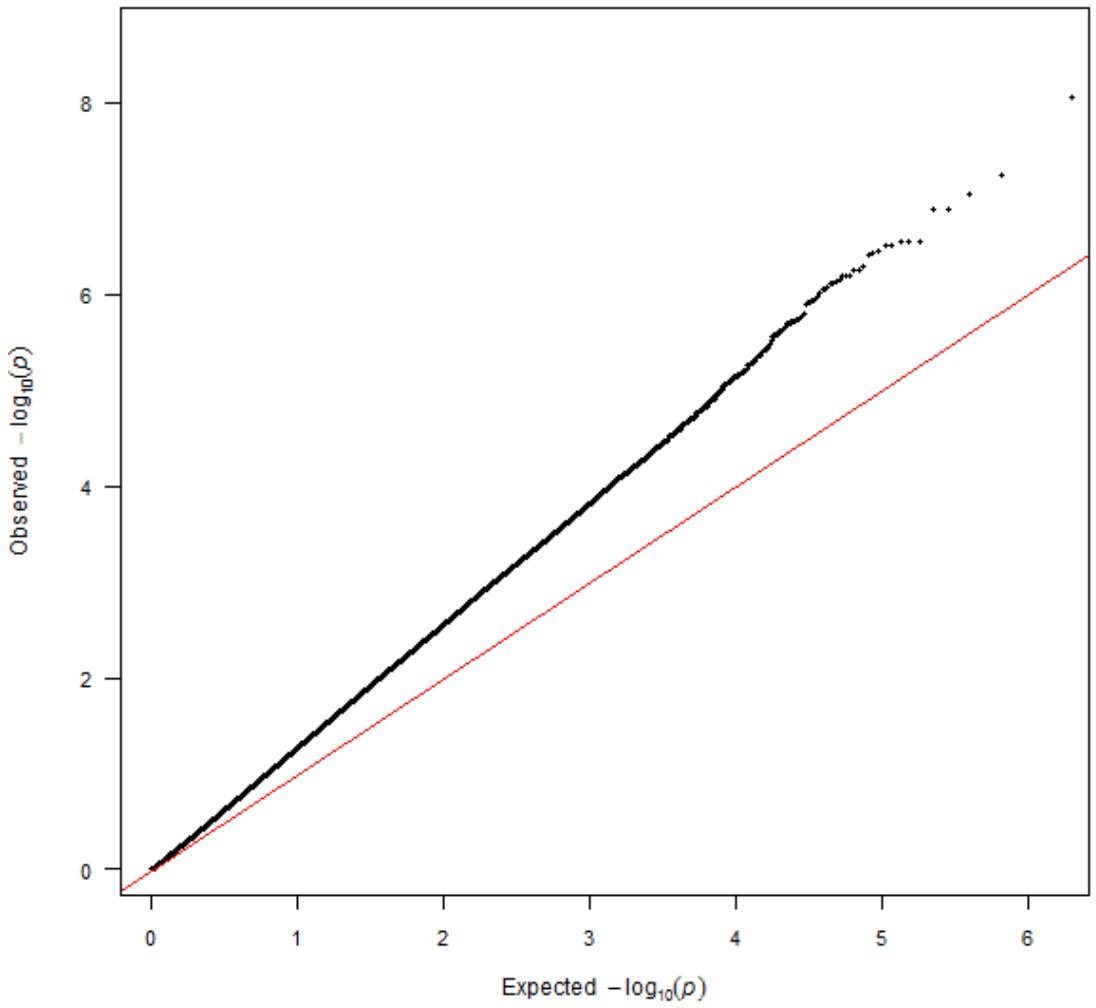


Ikram et al. PLoS Genetics 2010

Multiple testing problem

- If $\alpha=0.05$, then $m*\alpha$ are expected to be found just by chance
- Use multiple testing correction (Bonferroni)
- GWAS significance level ~5e-08, widely accepted
- For WGAS with rare variants should be even smaller (Fadista et al. 2016)

QQ plot

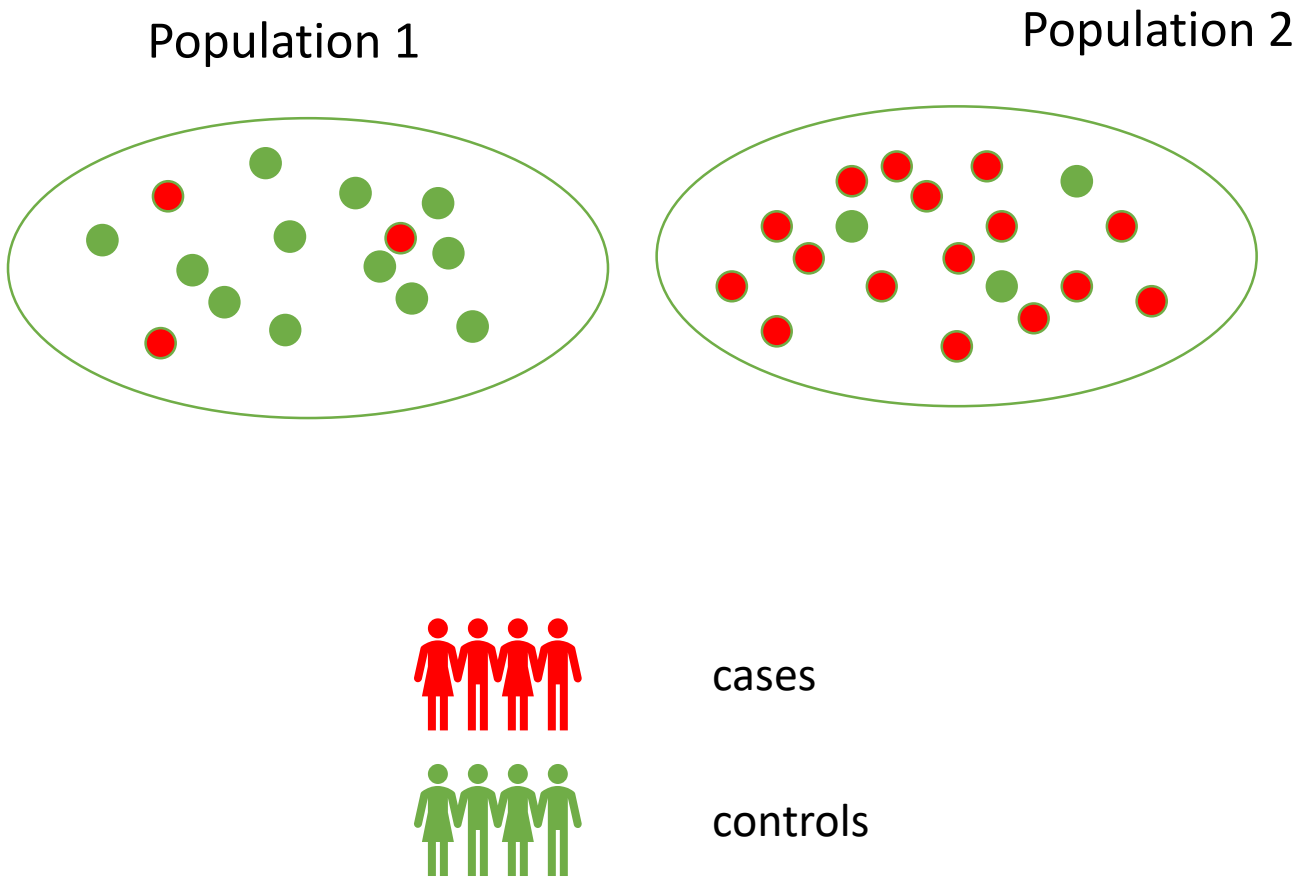


Bias due to population structure

- allele frequency differences between populations due to genetic drift and gene flow

- Since we campare allele frequencies sampling from different populations can lead to false-positive association findings
- Suppose cases are over-sampled from group 2, relative to controls

- Then any allele which is more common (higher minor allele frequency) in group 2 will appear to be associated with the trait



Example (simulation)

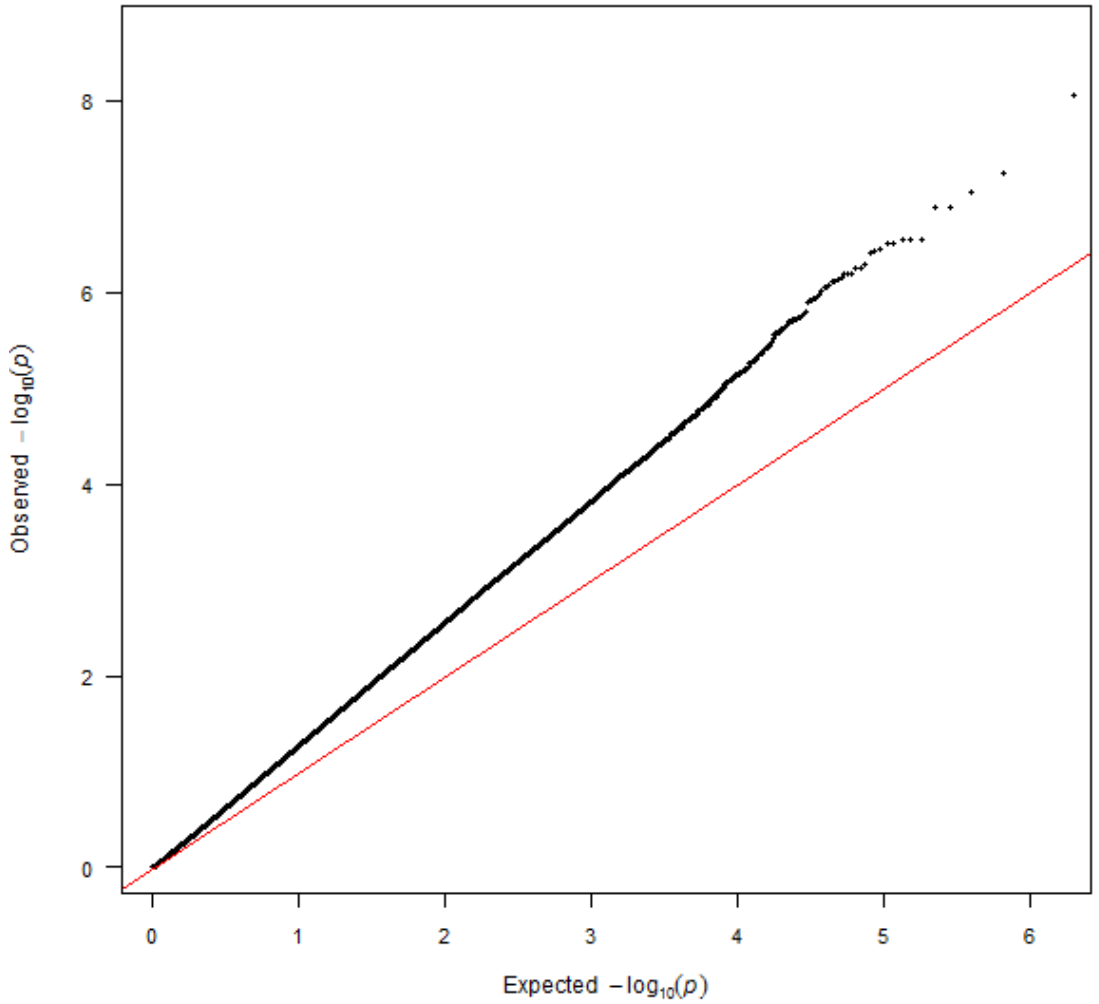
- 1000 cases / 1000 controls
- 100k null SNPs

Population	cases	controls
Population 1	400	600
Population 2	600	400

Genotype	cases	controls
AA	421	319
AB	469	505
BB	110	176

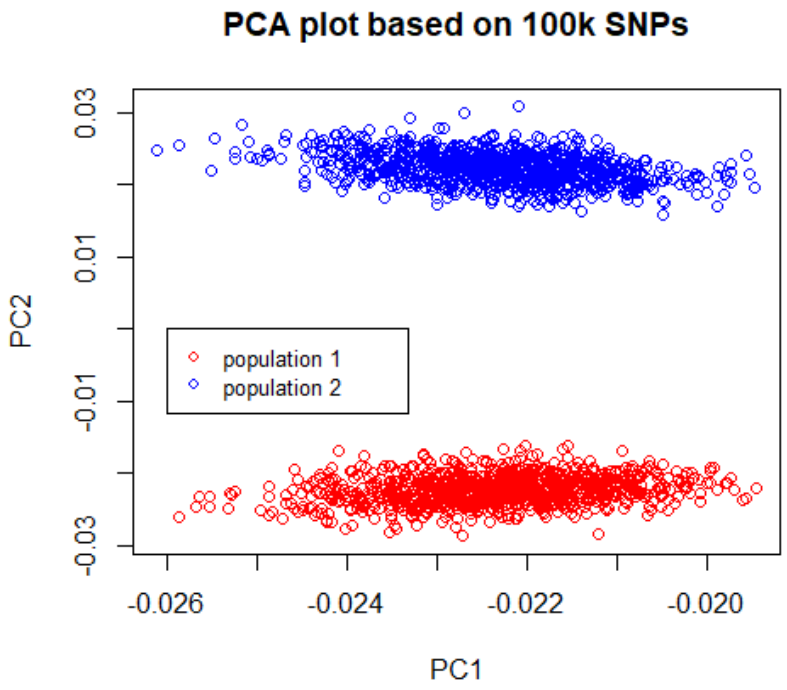
P=3.8e-08

Inflated QQ plot



Identification and correction: PCA

- Identification of stratification based on systematic patterns along the entire genome
- Approriate similarity measure between two individuals
- Most popular: GRM (EIGENSTRAT, Price et al. 2010)
- Incorporation of information into association test for single variant (principal components)



GRM

- Calculate genetic covariance matrix (genetic relationship matrix)
- N individuals, m markers:

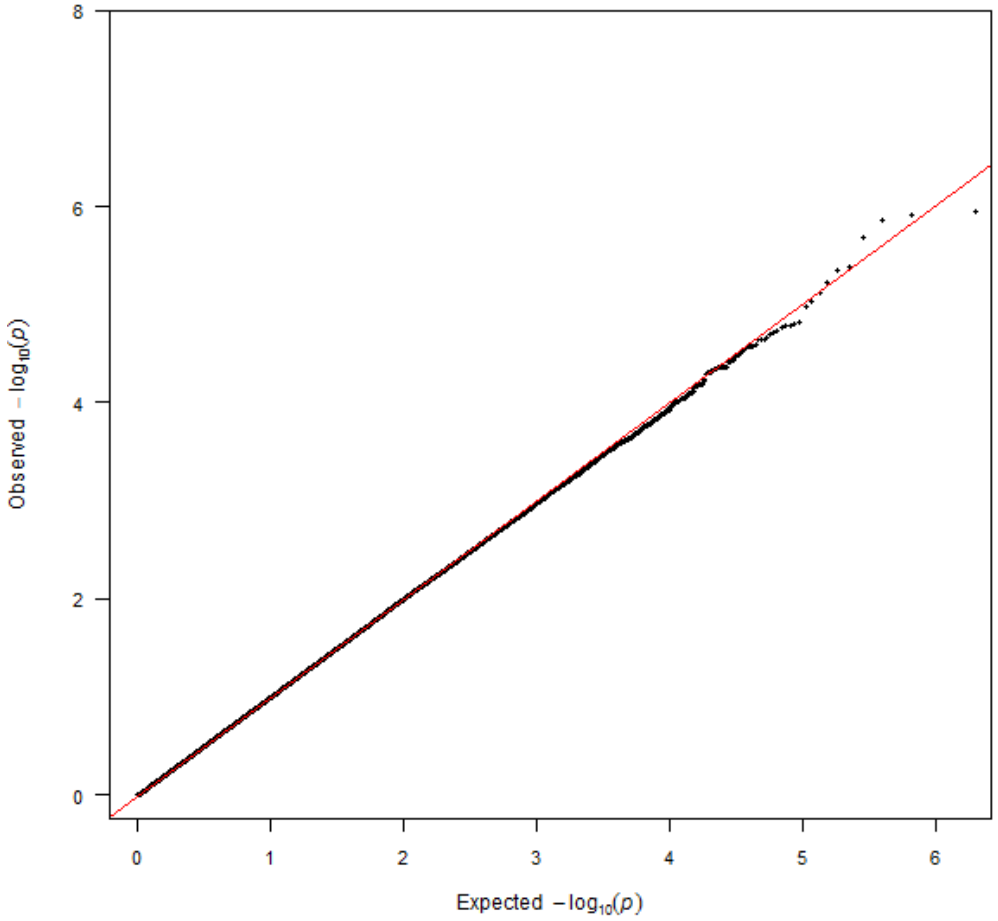
$$G = \begin{pmatrix} g_{11} & \cdots & g_{1m} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{nm} \end{pmatrix}$$

$$\hat{\Psi}_{ij} = \frac{1}{m} \sum_{s=1}^{m} \frac{(G_{is} - 2\hat{p}_s)(G_{js} - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}$$

- Perform an eigenvalue decomposition and use top principal components in a regression as covariates

Price A.L. et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies.

QQ plot after PCA correction



Pcorrected=2.2e-05

Mixed models

- The test of association is performed in the fixed effects part of the model
- Implicitly captures population structure and cryptic relatedness by modelling the covariance matrix.

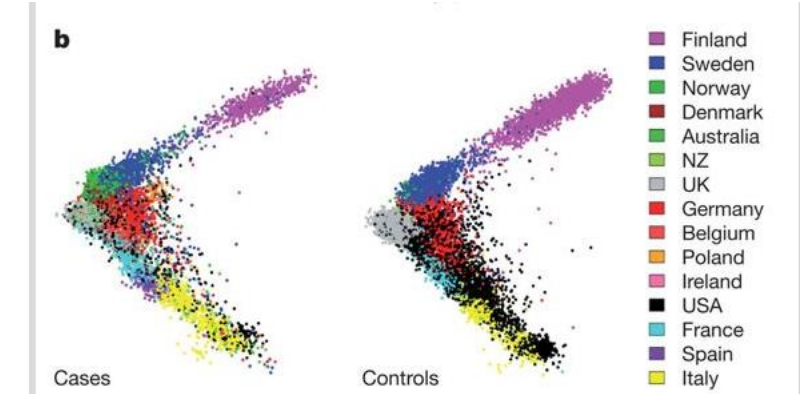
- Can increase power by implicitly conditioning on associated loci other than the candidate locus and by larger sample sizes (related+unrelated)
- software packages (e.g. EMMAX, GCTA, GEMMA, LMM-BOLT, GMMAT, SAIGE)

Mixed models

- $\mathbf{Y} \sim \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\varepsilon}$
- Y – phenotype
- X – vector of covariates (fixed effects)
- β – vector of fixed effects coefficients
- $g \sim (0, K\sigma_g^2)$ – total genetics effects per ind, $\varepsilon \sim (0, I\sigma_e^2)$
- K – relationship matrix, often GRM is taken.

Example mixed models

- Large GWAS, several populations
- Compared several approaches



a, **b**, All cases and controls were drawn from populations with European ancestry; cases from 15 countries and controls from 8. **a**, Numbers of case (red) and control (black) samples from each country. **b**, The projection of samples onto the first two principal components of genetic variation, with cases shown on the left and controls on the right. The axes are

MENU
nature
International journal of science

Letter | Published: 10 August 2011

Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis

The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2

Nature **476**, 214–219 (11 August 2011) | Download Citation

Abstract

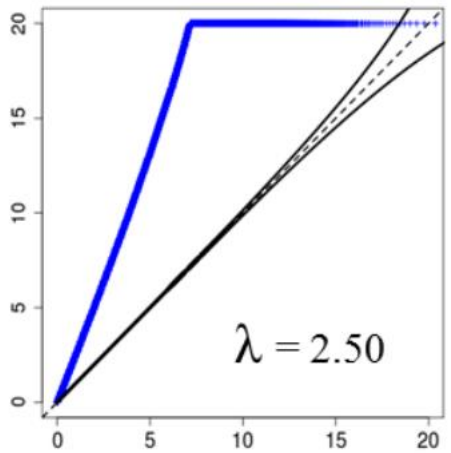
Multiple sclerosis is a common disease of the central nervous system in which the interplay between inflammatory and neurodegenerative processes typically results in intermittent neurological disturbance followed by progressive accumulation of disability[1]. Epidemiological studies have shown that genetic factors are primarily responsible for the substantially increased frequency of the disease seen in the relatives of affected individuals[2,3], and systematic attempts to identify linkage in multiplex families have confirmed that variation within the major histocompatibility complex (MHC) exerts the greatest individual effect on risk[4]. Modestly powered genome-wide association studies (GWAS)[5,6,7,8,9,10] have enabled more than 20 additional risk loci to be

Sawcer et al. Nature 2011

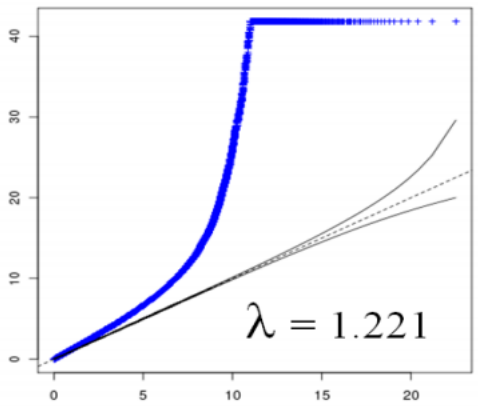
Example mixed models

Lambda – genomic inflation factor, median inflation of test statistics
$\lambda = median(\chi_1^2, \chi_2^2, \dots, \chi_n^2)/0.455$

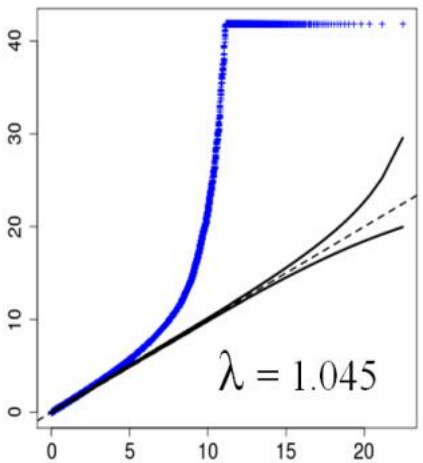
No correction



PCA correction, 100 PCs



Mixed model approach

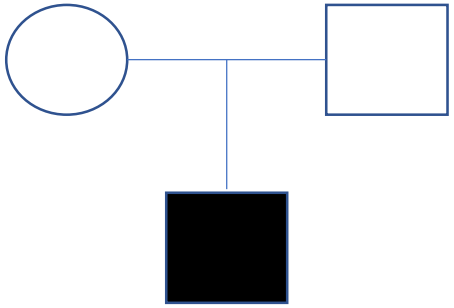


Association methods (family-based)

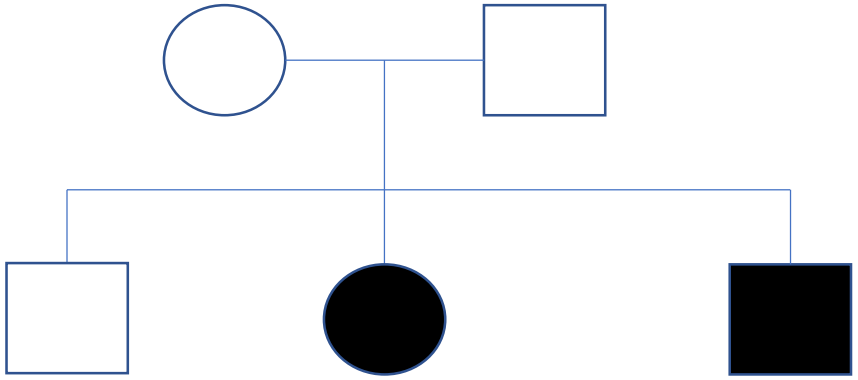
- Test for both linkage and association
- Robust to population substructure: different environments, admixture, stratification, failure of HWE
- Requires genotyped families (parent-child, or siblings)
- TDT test for trio design (affected offspring)
- FBAT - generalization to general phenotypes, general pedigrees, missing parental genotypes, and multiple variants (Lake and Laird 2001, Laird and Lange 2006,…)

Family-based designs

- Trio
  - both parental genotypes observed

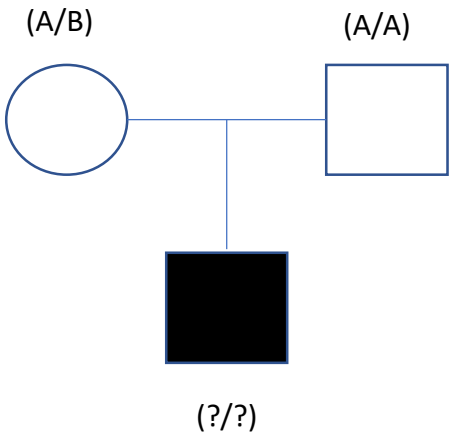


- Affected siblings
  - Usually no parent genotypes



TDT

- classical trio design : affected offspring
- implies outcome-based sampling
- if variant is associated, observed transmission rates should deviate from mendelian
- compares transmissions from heterozygous parents to offspring with expectation under Mendel‘s laws



TDT

Table 2

Combinations of Transmitted and Nontransmitted Marker Alleles M_1 and M_2 among $2n$ Parents of n Affected Children

Transmitted Allele	Nontransmitted Allele		Total
	M_1	M_2	
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

$$\chi^2 = (b-c)^2/(b+c).$$

Spielman et al. 1993

Generalizations

- FBAT: generalization of TDT to general phenotypes, general pedigrees, missing parental genotypes, and multiple variants (Lake and Laird 2001, Laird and Lange 2006,…)
- GDT: incorporates parental phenotypes (Chen et al. 2009)

Received: 12 June 2018 | Revised: 26 September 2018 | Accepted: 26 November 2018

DOI: 10.1002/gepi.22181

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

A comparison of popular TDT-generalizations for family-based association analysis

Julian Hecker | **Nan Laird** | **Christoph Lange**

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

Abstract

The transmission disequilibrium test (TDT) is the gold standard for testing the

FBAT general framework

$$U = \sum T(X - E(X|P))$$

$$U = \sum (Y - \mu)(X - E(X|P))$$

$$Z = U/sqrt(Var(U))$$

- T – trait, based on phenotype Y and offset
- X – genotype
- P – parental genotypes
- Sum over all offspring
- E(X|P) is the expected marker score computed under H0, conditional on P
- Equivalent to TDT, when trio design and no missing data
- FBAT toolkit

Variance explained by common variants

- Schizophrenia
  - Estimated heritability from twin studies: 65-80%
  - Proportion of heritability explained by common SNPs: 25-31%
- Biploar Disorder
  - Estimated heritability from twin studies: 75-85%
  - Proportion of heritability explained by common SNPs: 25-31%

Lee et al. „Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs." *Nature Reviews Genetics 44.3 (2012): 247-250.*
Kieseppa et al. „High concordance of bipolar I disorder in a nationwide sample of twins.“ *American Journal of Psychiatry 161 (2004): 1814-1821.*

Where is the missing heritability? Theories:

- Lack of Power: weak effects
- Rare variants
- Epistasis: combinations of SNPs
- Epigenetics: external and environmental factors

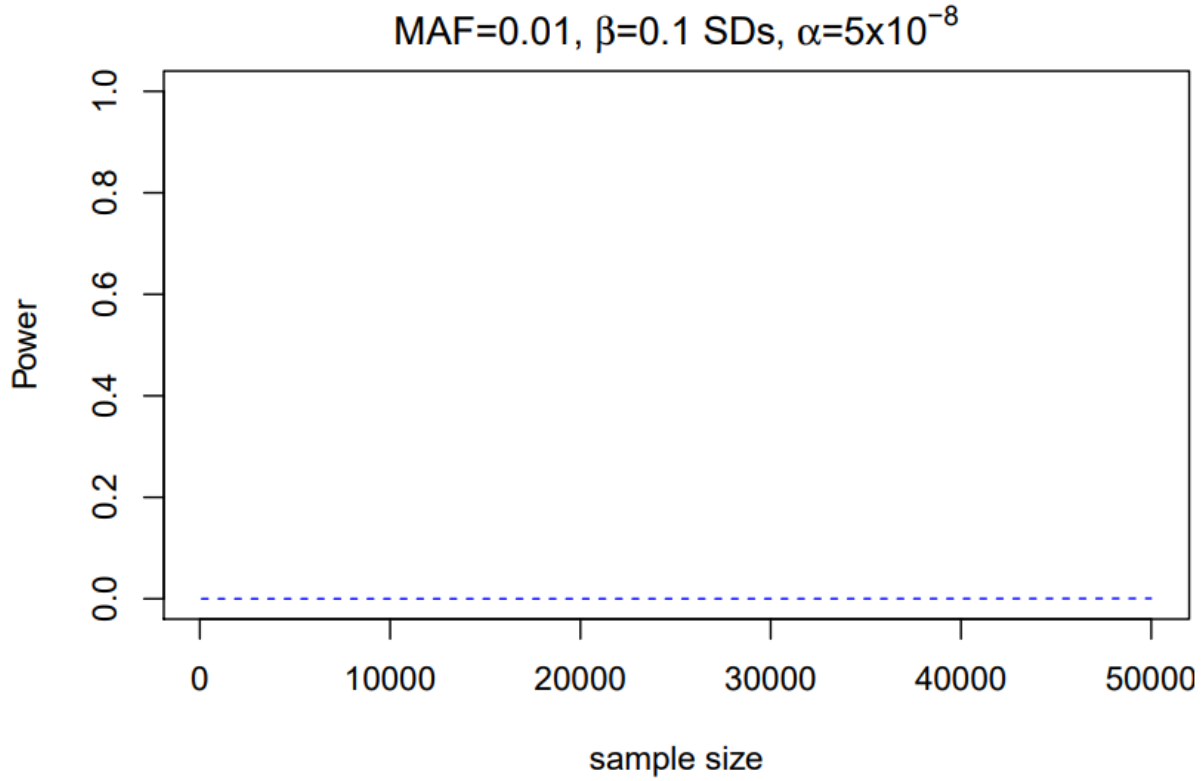
Where is the missing heritability? Theories:

- Lack of Power: weak effects
- Rare variants
- Epistasis: combinations of SNPs
- Epigenetics: external and environmental factors

Rare variants

- characterized by small minor allele frequencies (i.e. below 5% or 1%)
- due to small allele frequencies a weaker LD-structure compared to common variants
- Singletons: rare variants with an allele count of 1, private mutations, unclear role

Major problem - power



3.8

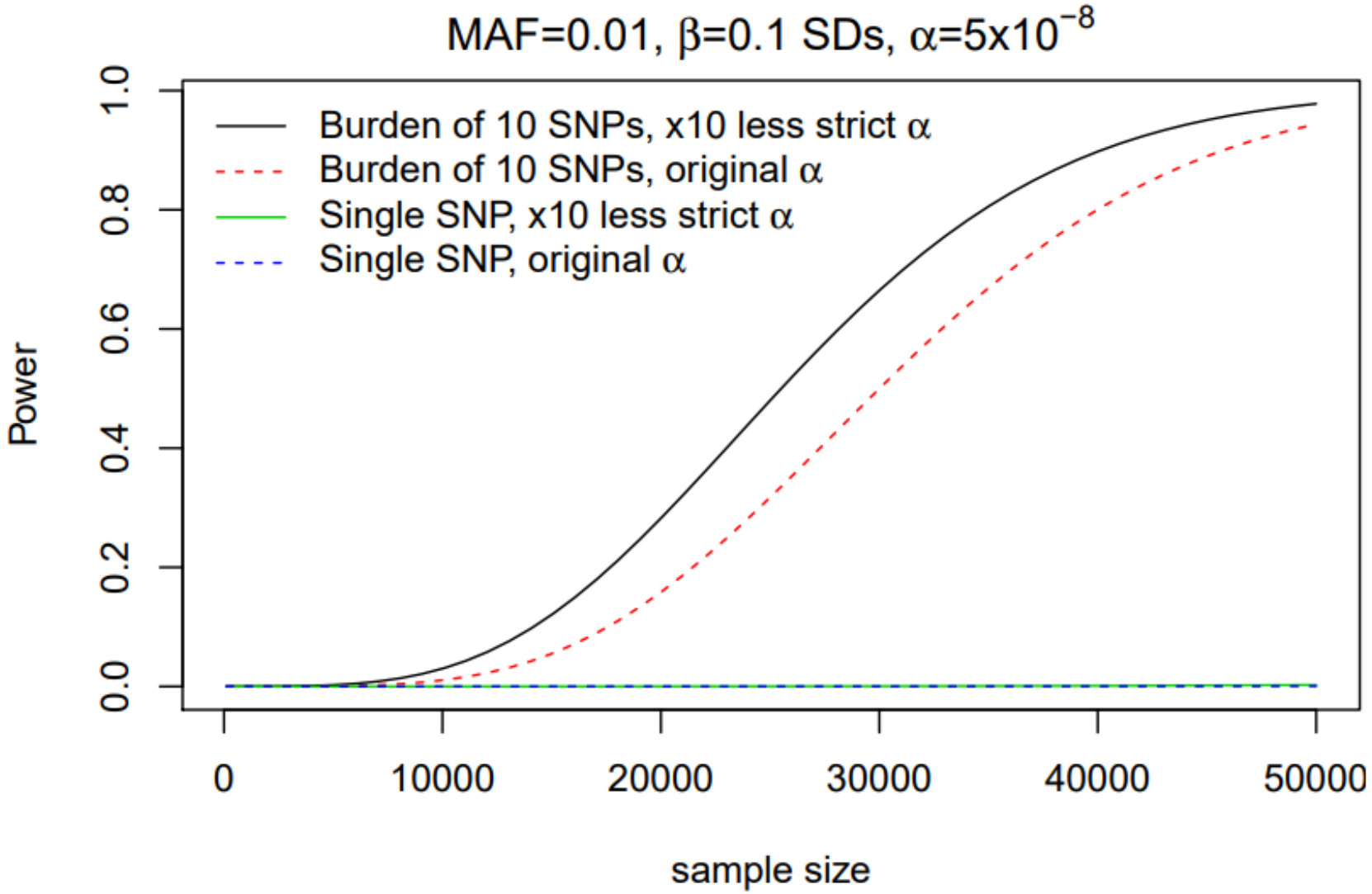
Slide courtesy: Ken Rice

Rare variant analysis

- When the sample size is limited try to combine the signal from multiple rare variants.
- How to test multiple rare variants?
  - Combine signals

- How to group rare variants for testing?
  - Functional annotations
  - Sliding windows

Power



Slide courtesy: Ken Rice

Rare variant analysis

- Burden (CAST,CMC,WSS)
  - Testing the combined effect of multiple rare variants
  - Work well for signals in one direction

$$Q_B = \left[\sum_{i=1}^{n} (y_i - \hat{\pi}_i)(\sum_{j=1}^{m} w_j g_{ij}) \right]^2$$

- Variance-component (SKAT)
  - Jointly test individual variant-score test statistics
  - Robust to effect direction

$$Q_S = \sum_{j=1}^{m} w_j^2 S_j^2 = \sum_{j=1}^{m} w_j^2 \left\{ \sum_{i=1}^{n} g_{ij} (y_i - \hat{\pi}_i) \right\}^2$$

- SKAT-O - weighted average of SKAT and burden-test statistics

$$Q_p = pQ_B + (1-p)Q_S$$

Morgenthaler and Thilly. Mut. Res. 2007
Li and Leal. AJHG 2008
Madsen and Browning. PLOS Genet. 2009
Wu et al. AJHG 2011
Lee et al. AJHG 2012

Rare variants explain the missing heritability?

New Results

1 comment

Recovery of trait heritability from whole genome sequence data

Pierrick Wainschtein, Deepti P. Jain, Loic Yengo, Zhili Zheng,
TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium,
L. Adrienne Cupples, Aladdin H. Shadyab, Barbara McKnight, Benjamin M. Shoemaker, Braxton D. Mitchell,
Bruce M. Psaty, Charles Kooperberg, Dan Roden, Dawood Darbar, Donna K. Arnett, Elizabeth A. Regan,
Eric Boerwinkle, Jerome I. Rotter, Matthew A. Allison, Merry-Lynn N. McDonald, Mina K Chung,
Nicholas L. Smith, Patrick T. Ellinor, Ramachandran S. Vasan, Rasika A. Mathias, Stephen S. Rich,
Susan R. Heckbert, Susan Redline, Xiuqing Guo, Y.-D Ida Chen, Ching-Ti Liu, Mariza de Andrade, Lisa R. Yanek,
Christine M. Albert, Ryan D. Hernandez, Stephen T. McGarvey, Kari E. North, Leslie A. Lange, Bruce S. Weir,
Cathy C. Laurie, Jian Yang, Peter M. Visscher

doi: https://doi.org/10.1101/588020

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

Heritability, the proportion of phenotypic variance explained by genetic factors, can be estimated from pedigree data [1], but such estimates are uninformative with respect to the underlying genetic architecture. Analyses of data from genome-wide association studies (GWAS) on unrelated individuals have shown that for human traits and disease, approximately one-third to two-thirds of heritability is captured by common SNPs [2-5]. It is not known whether the remaining heritability is due to the imperfect tagging of causal variants by common SNPs, in particular if the causal variants are rare, or other reasons such as over-estimation of heritability from pedigree data. Here we show that pedigree heritability for height and body mass index (BMI) appears to be fully recovered from whole-genome sequence (WGS) data on 21,620 unrelated individuals of European ancestry. We assigned 47.1 million genetic variants to groups based upon their minor allele frequencies (MAF) and linkage disequilibrium (LD) with variants nearby, and estimated and partitioned variation accordingly. The estimated heritability

Previous | Next

Posted March 25, 2019.

Download PDF
Supplementary Material
Email
Share
Citation Tools

Tweet | Нравится 23

Subject Area

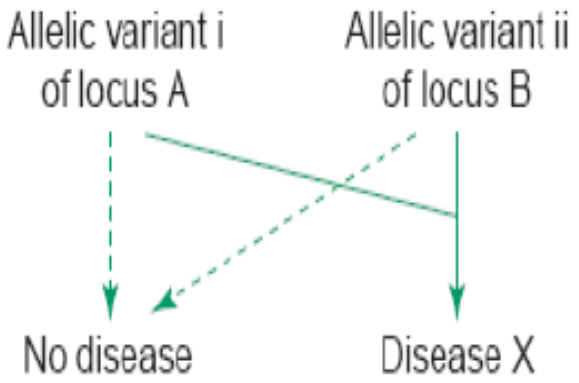
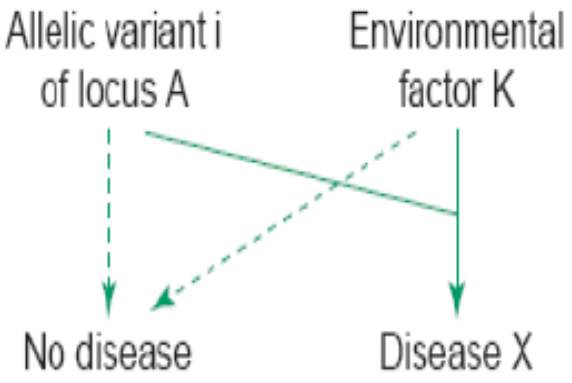
Genetics

Subject Areas

All Articles

- Animal Behavior and Cognition
- Biochemistry
- Bioengineering
- Bioinformatics
- Biophysics
- Cancer Biology
- Cell Biology
- Clinical Trials
- Developmental Biology
- Ecology
- Epidemiology
- Evolutionary Biology
- Genetics
- Genomics

Beyond main effects

	Gene–gene interaction	Gene–environment interaction
Definition	When two or more DNA variations interact either directly (DNA–DNA or DNA–mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects	When a DNA variation interacts with an environmental factor, such that their combined effect is distinct from their independent effects
Diagram		

Thornton-Wells et al. 2004

Summary

- Association methods
  - Data quality
  - GWAS
  - Multiple testing problem
  - Population stratification
  - Mixed models
  - Family-based association studies
  - Rare variant association studies