(Genome-wide) association analysis

Key concepts

- Mapping QTL by association relies on linkage disequilibrium in the population;
- LD can be caused by close linkage between a QTL and marker (= good) or by confounding between a marker and other effects (= usually bad);
- The power of QTL detection by LD depends on the proportion of phenotypic variance explained at a marker;
- Mixed models are good for performing GWAS
- Genetic (co)variance can be estimated from GWAS summary statistics

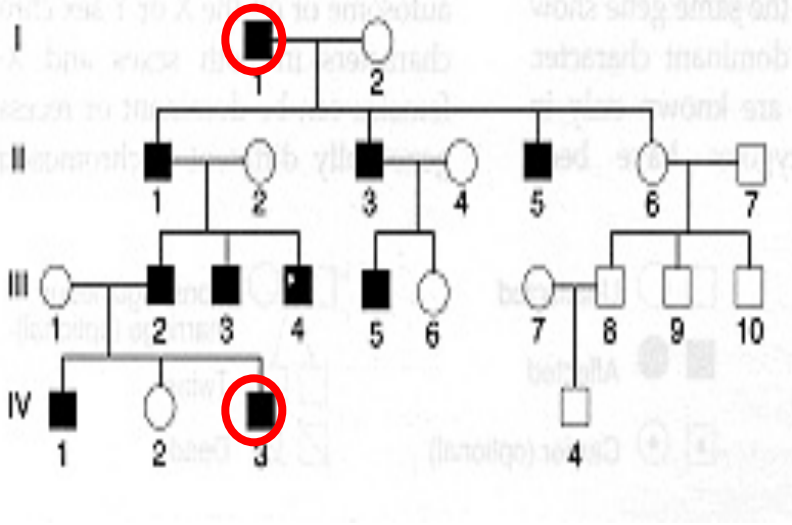
Outline

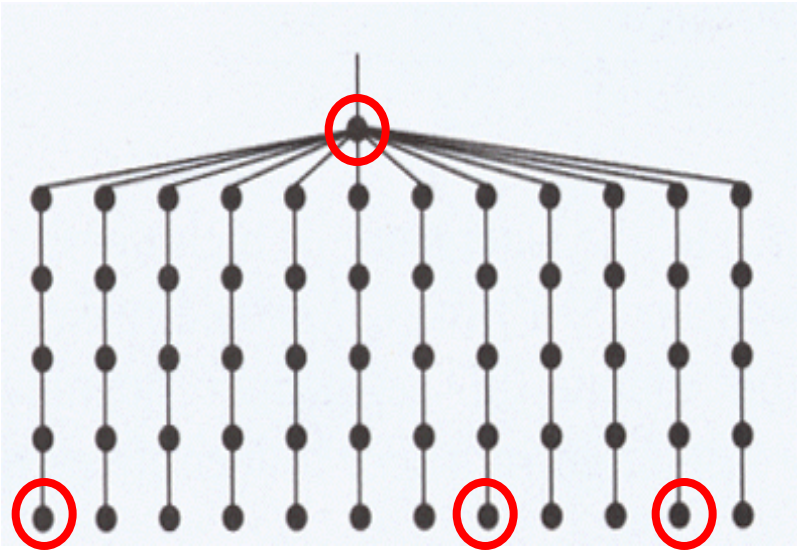
- Association vs linkage
- Linkage disequilibrium
- Analysis: single SNP

- GWAS: design, power
- GWAS: analysis

Linkage

Association

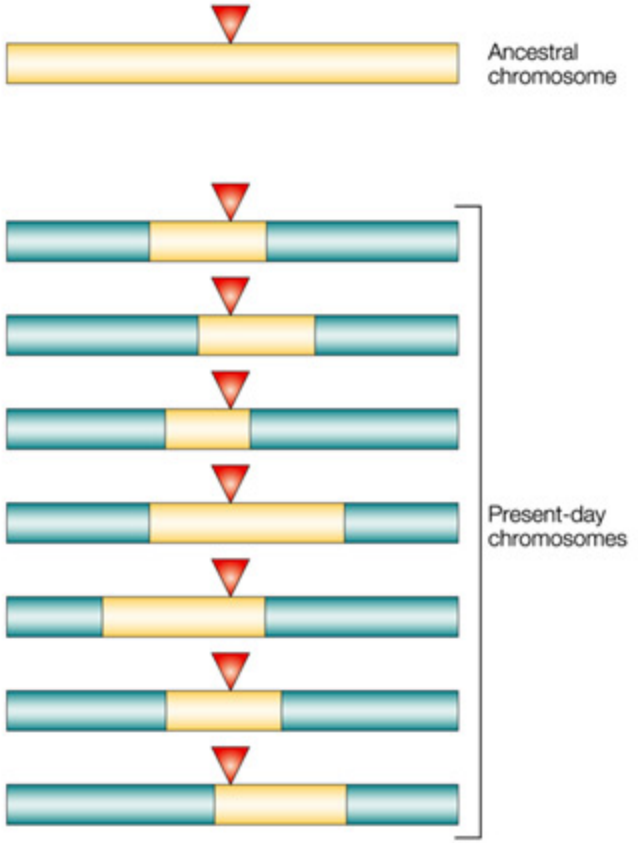




Families

Populations

Linkage disequilibrium around an ancestral mutation



LD

- Non-random association between alleles at different loci
- Many possible causes
  - mutation
  - drift / inbreeding / founder effects
  - population stratification
  - selection
- Broken down by recombination

Definition of D

- 2 bi-allelic loci
  - Locus 1, alleles A & a, with freq. p and (1-p)
  - Locus 2, alleles B & b with freq. q and (1-q)
  - Haplotype frequencies p_{AB}, p_{Ab}, p_{aB}, p_{ab}

$D = p_{AB} - pq$

r^2

$r^2 = D^2 / [pq(1-p)(1-q)]$

- Squared correlation between presence and absence of the alleles in the population
- 'Nice' statistical properties

[Hill and Robertson 1968]

Properties of r and r^2

- Population in 'equilibrium'

  $E(r) = 0$

  $E(r^2) = var(r) \approx 1/[1 + 4Nc] + 1/n$

  N = effective population size

  n = sample size (haplotypes)

  c = recombination rate

- $nr^2 \sim \chi_{(1)}^2$
- Human population is NOT in equilibrium

LD depends on population size and recombination distance

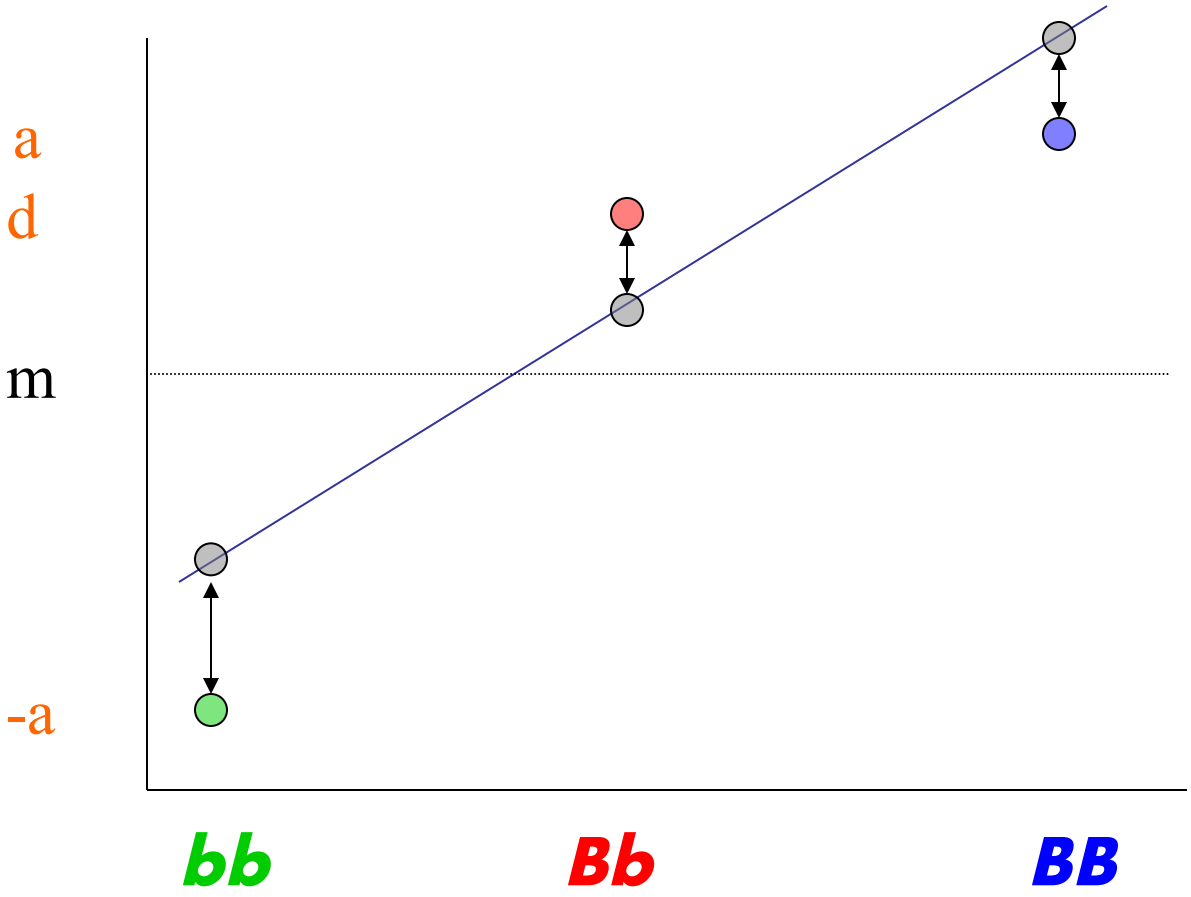
[Sved 1971; Weir and Hill 1980]

Analysis

- Single locus association
- GWAS

- Least squares
- ML
- Bayesian methods

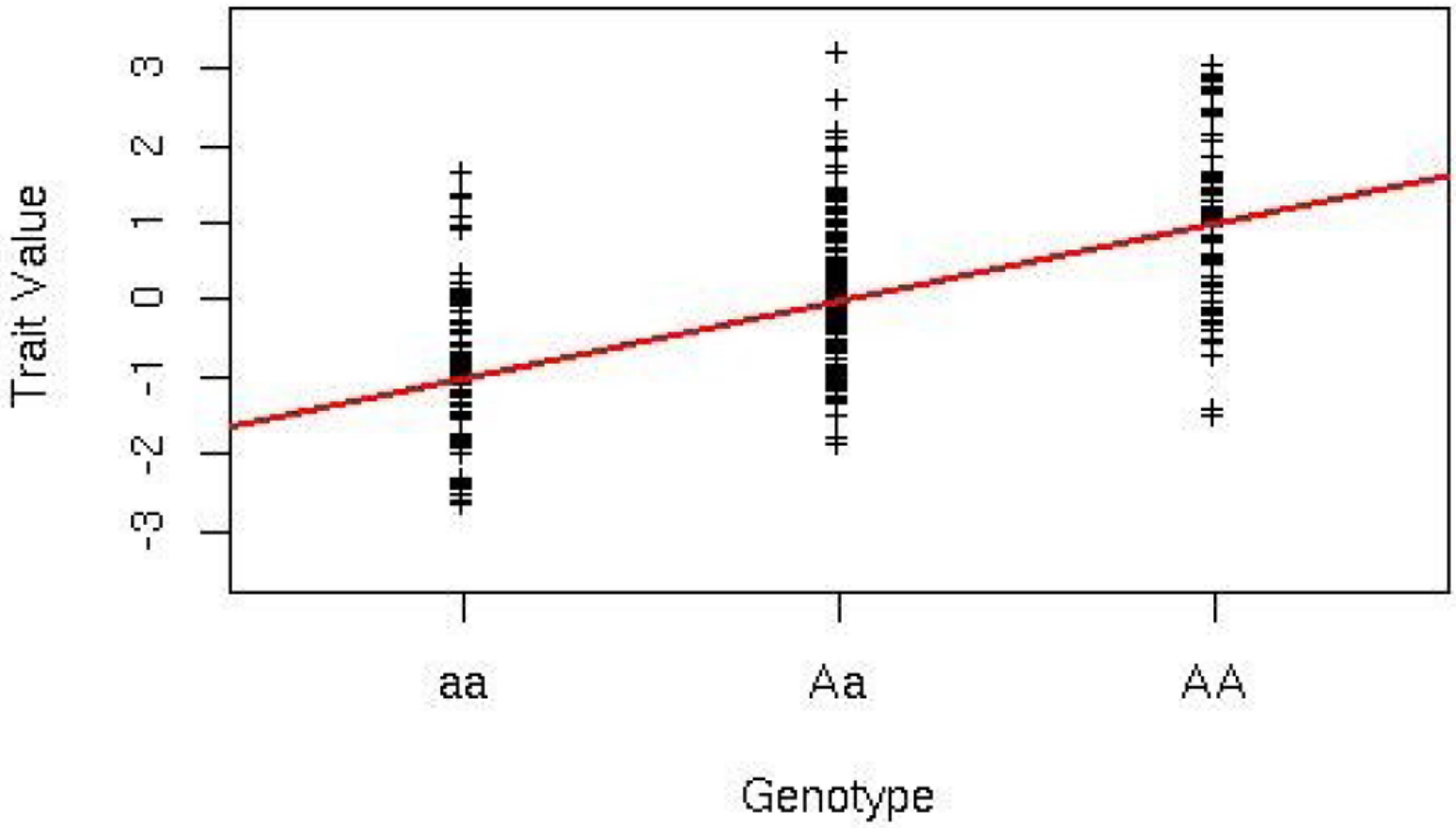
Falconer model for single biallelic QTL



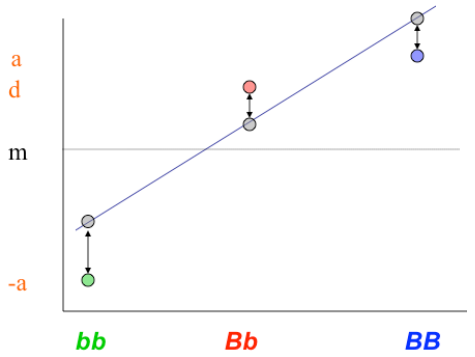
Var (*X*) = Regression Variance + Residual Variance
= Additive Variance + Dominance Variance

Unrelated Samples

$$\hat{y}_i = \mu + \hat{\beta}\, x_i$$



Statistical power (linear regression)



$y = \mu + \beta x + e, \quad x = 0, 1, 2$

$\sigma_y^2 = \sigma_q^2 + \sigma_e^2$ regression + residual

$\sigma_x^2 = 2p(1-p)$ p = allele frequency for indicator x

{HWE: note x is usually considered fixed in regression}

$\sigma_q^2 = \beta^2\sigma_x^2 = [a + d(1-2p)]^2 * 2p(1-p)$

$q^2 = \sigma_q^2 / \sigma_y^2$ {QTL heritability}

Statistical Power

χ^2 test with 1 df:

$E(X^2) = 1 + n R^2 / (1 - R^2)$

$= 1 + nq^2/(1-q^2)$

$= 1 + NCP$

NCP = non-centrality parameter

Power of association proportional to q^2
(Power of linkage proportional to q^4)

Statistical Power (R)

```
alpha= 5e-8
threshold= qchisq(1-alpha,1)
q2= 0.005
n= 10000
ncp= n*q2/(1-q2)
power= 1-pchisq(threshold,1,ncp)
threshold
ncp
power
```

```
> alpha= 5e-8
> threshold= qchisq(1-alpha,1)
> q2= 0.005
> n= 10000
> ncp= n*q2/(1-q2)
> power= 1-pchisq(threshold,1,ncp)
> threshold
[1] 29.71679
> ncp
[1] 50.25126
> power
[1] 0.9492371
```

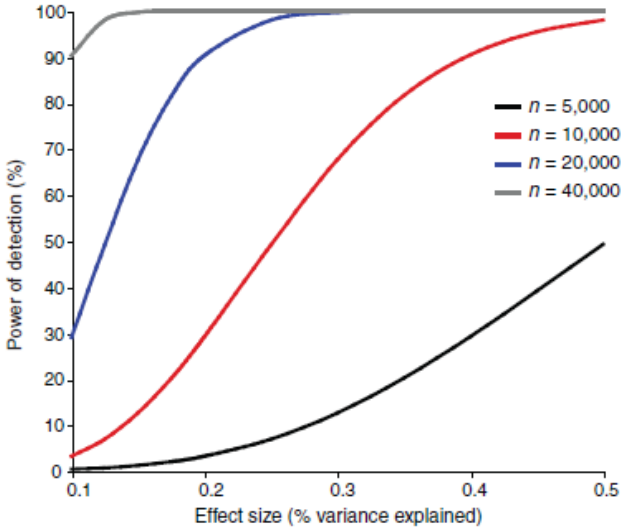


Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical for a GWAS with a sample size of n = 5,000–40,000.

In-class demo

Power by association with SNP

(small effect; HWE)

$NCP(SNP) = n\, r^2\, q^2$

$= r^2 * NCP(\text{causal variant})$

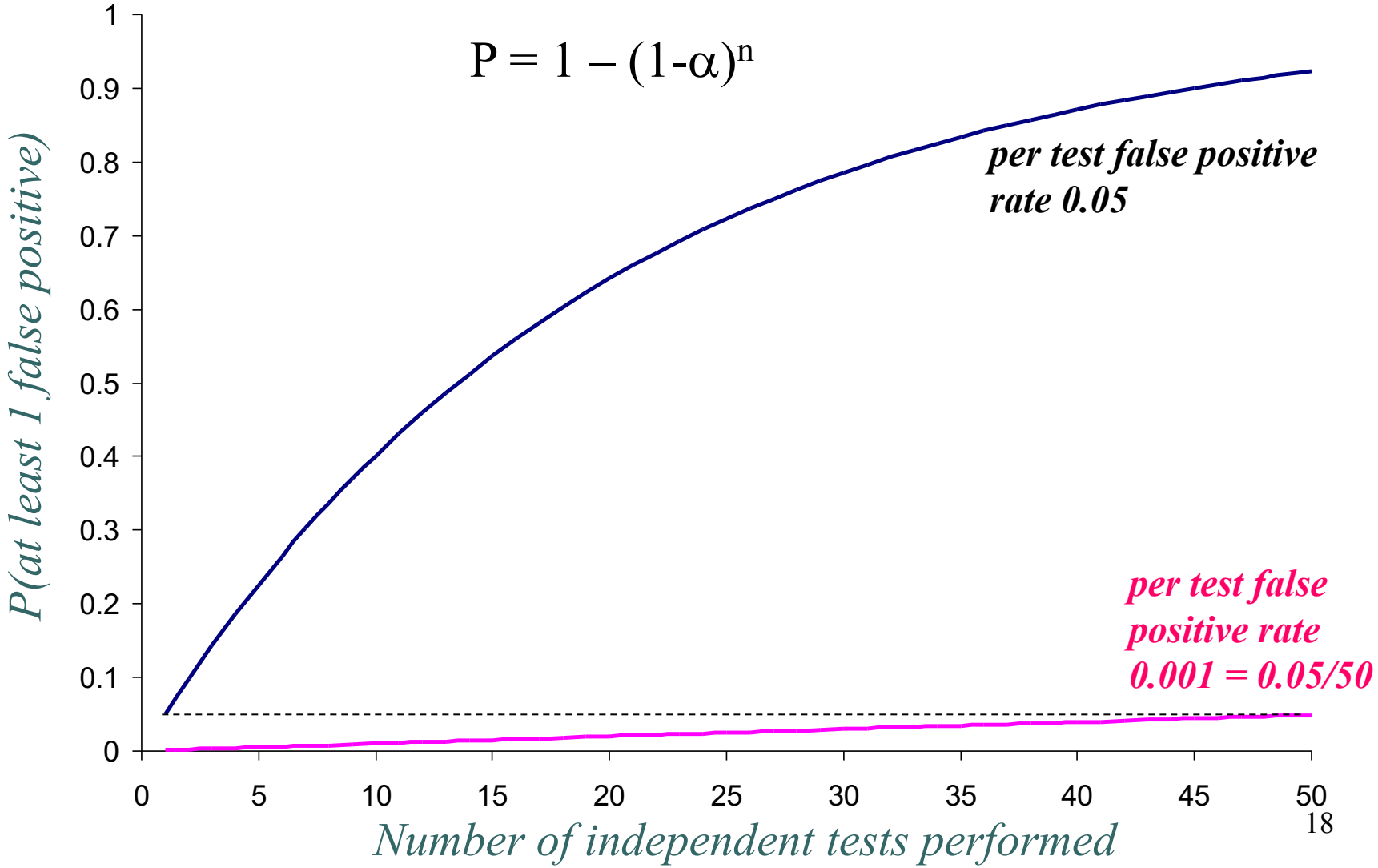
$= n * \{r^2\, q^2\} = n * (\text{variance explained by SNP})$

Power of LD mapping depends on the experimental sample size, variance explained by the causal variant and LD with a genotyped SNP

GWAS

- Same principle as single locus association, but additional information
  - QC
    - Duplications, sample swaps, contamination
  - Power of multi-locus data
    - Unbiased genome-wide association
    - Relatedness
    - Population structure
    - Ancestry
    - More powerful statistical analyses

The multiple testing burden



Population stratification (association unlinked genes)

Both populations are in linkage equilibrium; genes unlinked

	Allele frequency		Haplotype frequency			
	p_{A1}	p_{B1}	p_{A1B1}	p_{A1B2}	p_{A2B1}	p_{A2B2}
Pop. 1	0.9	0.9	0.81	0.09	0.09	0.01
Pop. 2	0.1	0.1	0.01	0.09	0.09	0.81
Average	0.5	0.5	0.41	0.09	0.09	0.41

Combined population: $D = 0.16$ and $r^2 = 0.41$

Population stratification (genes and phenotypes)

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to *Molecular Psychiatry* heralding the discovery of the 'successful-use-of-selected-hand-instruments gene' (SUSHI).

[Hamer & Sirota 2010 Mol Psych]

Population stratification (genes and phenotypes)

It took another 2 years to discover that SUSHI is a histocompatibility antigen gene that has nothing to do with chopstick use but just happens to have different allele frequencies in Asians and Caucasians, who of course differ in chopstick use for purely cultural rather than biological reasons. Even though the association data were highly significant and readily replicated, they were biologically meaningless.

[Hamer & Sirota 2010 Mol Psych]

Population stratification (genes and phenotypes)

The source of confounding in the chopstick example is better thought of as the environment. The problem arises because different subgroups have different levels of exposure to chopsticks. This type of confounding is extremely familiar to genetic epidemiologists, but it is unimportant in settings where the environment can be experimentally controlled or randomized (as is routinely done in plant breeding, for example).

There is another source of confounding, however, and that is the genetic background. The estimate of the effect of a particular locus can be confounded by the other causal loci in the genome. This genetic background effect will always be present to some extent, even

[Vilhjalmsson & Nordborg 2013 Nature Reviews Genetics]

Demonstrating stratification in a European American population

Catarina D Campbell[1,2], Elizabeth L Ogburn[1], Kathryn L Lunetta[3,8], Helen N Lyon[1,2], Matthew L Freedman[4–6], Leif C Groop[7], David Altshuler[2,4,5], Kristin G Ardlie[3] & Joel N Hirschhorn[1,2,4]

Table 2 No evidence for stratification using standard methods

		χ^2 values[a]		Estimates of stratification parameters[b]		
	SNPs	Median	Mean	λ_{max}	λ	*P*
Random SNPs	111	0.37	0.96	3.21	1	0.61
AIMs	67	0.58	0.95	–	–	0.61
Total	178	0.49	0.95	–	–	0.66

Table 3 A strong association of *LCT* –13910C→T and height is reduced by rematching subjects on the basis of ancestry

			Origin of grandparents[a]			
		All	Four US-born	Southeastern	Northwestern	Combined[b]
N	Total	2,179	1,282	354	543	–
	Tall	1,123	645	127	351	–
	Short	1,056	637	227	192	–
LCT –13910 genotype counts[c]	Total	392:918:869	142:543:596	182:141:31	68:233:243	–
	Tall	161:474:489	66:265:314	54:55:18	41:154:157	–
	Short	231:444:380	76:278:282	128:86:13	27:79:86	–
Hardy-Weinberg *P*	Total	5.6×10^{-7}	0.57	0.89	0.89	–
	Tall	0.03	0.66	0.81	0.92	–
	Short	2.5×10^{-5}	0.86	0.96	0.45	–
Association *P*		3.6×10^{-7}	0.098	0.0016	0.71	0.0074
OR (95% c.i.)[d]		1.37 (1.22–1.54)	1.15 (0.97–1.36)	1.70 (1.22–2.38)	1.05 (0.81–1.37)	1.19 (1.05–1.36)

Table 4 No association of *LCT* –13910C/T and height in other European populations

		Polish	Scandinavian	Combined
Genotypes (CC:CT:TT)	Tall	166:251:86	–	–
	Short	174:235:96	–	–
Transmissions of T allele (T:U)[a]	Tall	–	65:68	–
	Short	–	76:66	–
P		0.92	0.43	0.58
OR (95% c.i.)[b]		0.99 (0.83–1.18)	0.91 (0.72–1.15)	0.96 (0.83–1.11)

Stratification

$y = \Sigma g_i + \Sigma e_i$

$r(y,g_i)$ due to

- causal association with g_i
- correlation g_i and g_j and causal association with g_j (LTC and height)
- correlation g_i and environmental factor e_j (chopsticks)

How to deal with structure?

- Detect and discard ‘outliers’
- Detect, analysis and adjustment
  - E.g. genomic control
- Account for structure during analysis
  - Fit a few principal components as covariates
  - Fit GRM

GWAS using mixed linear models

$y = \mathbf{Xb} + \beta x + \mathbf{g} + \mathbf{e}$

$var(\mathbf{g}) = \mathbf{G}\sigma_g^2$

G = genetic relationship matrix (GRM)

Model conditions on effects of all other variants

Power depends on whether x is included (MLMi) or excluded (MLMe) from the construction of **G**.

[Yang et al. 2014 Nature Genetics]

GWAS using mixed linear models: statistical power

For linear regression (LR), the expected mean of χ^2 association statistics (λ_{mean}) is

$$\lambda_{mean}\,(\mathrm{LR}) = 1 + Nh_g^2 / M \quad (1)$$

regardless of the genetic architecture of the trait[24].

For MLMi, the λ_{mean} value at markers used to construct the GRM is

$$\lambda_{mean}\,(\mathrm{MLMi}) = 1 \quad (2)$$

Equation (2) highlights the dangers of using λ_{mean} (or λ_{median}) to assess the presence of population stratification or other artifacts. A researcher who observes lower λ_{mean} (or λ_{median}) values for MLMi than for linear regression might conclude that this difference is due to correction for confounding, but this result is in fact expected, even in the absence of any confounding.

Finally, for MLMe,

$$\lambda_{mean}\,(\mathrm{MLMe}) \approx 1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$$

r^2 here is the squared correlation between **g**-hat and **g**

[Yang et al. 2014 Nature Genetics]

Exploiting GWAS summary statistics

- Summary stats are typically in the form of
  - SNP, SNP allele, effect size, SE (or p-value)
- 100s of datasets in the public domain
- Uses:
  - Prediction
  - Conditional analysis (GCTA-COJO)
  - Estimation of genetic parameters (LD scoring)

Exploiting known LD between SNPs

N = sample size, M = number of markers

- Single SNP, assume SNP scores are standardised ($w = (x\text{-}2p)/\sqrt{2p(1\text{-}p)}$)

$y = wb + e$

$b = \Sigma wy / \Sigma w^2$

$E(\Sigma w^2) \approx N \text{ var}(w) = N$

$E(\Sigma wy) \approx Nb$

Multiple regression

$\mathbf{y} = \mathbf{Wb} + \mathbf{e}$

$\mathbf{b} = (\mathbf{W'W})^{-1}\mathbf{W'y}$

Diagonal element of $\mathbf{W'W}$ are Σw^2 with

$E(\Sigma w^2) \approx N$

Off-diagonal elements are $\Sigma w_i w_j$ with

$E(\Sigma w_i w_j) \approx N r_{ij}$

→ $(\mathbf{W'W})/N$ is an LD correlation matrix

LD correlation matrix from reference

- Let **R** be an LD correlation matrix estimated from a reference sample with individual-level data
- Then E(**W'W**) ~ N**R**

Approximate joint analaysis using summary statistics

$\mathbf{b}_{GWAS} = \text{diag}(\mathbf{W'W})^{-1}\mathbf{W'y}$

We want $\mathbf{b}_{joint} = (\mathbf{W'W})^{-1}\mathbf{W'y}$

Approximate solution:

$E(\mathbf{W'W}) \sim N\mathbf{R}$

$\mathbf{W'y} = \text{diag}(\mathbf{W'W})\mathbf{b}_{GWAS} = N\mathbf{b}_{GWAS}$

$\mathbf{b}_{joint} = (N\mathbf{R})^{-1}N\mathbf{b}_{GWAS} = \mathbf{R}^{-1}\mathbf{b}_{GWAS}$

- Allows joint fitting and conditional analysis

[Yang et al. 2012, Nature Genetics]

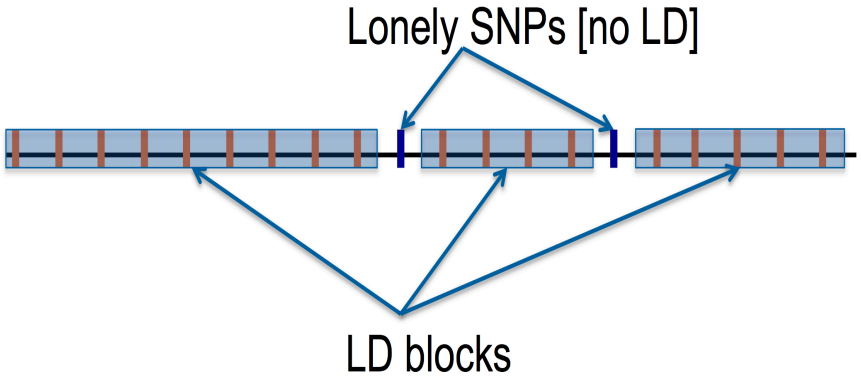
LD score regression

- Exploit summary statistics to estimate genetic parameters and detect evidence of population stratification
- Principle:
  - $E(\chi^2) = Nq^2$ for single causal variant
  - $E(\chi^2) = Nq^2(1 + r^2)$ for causal variant correlated with another causal variant

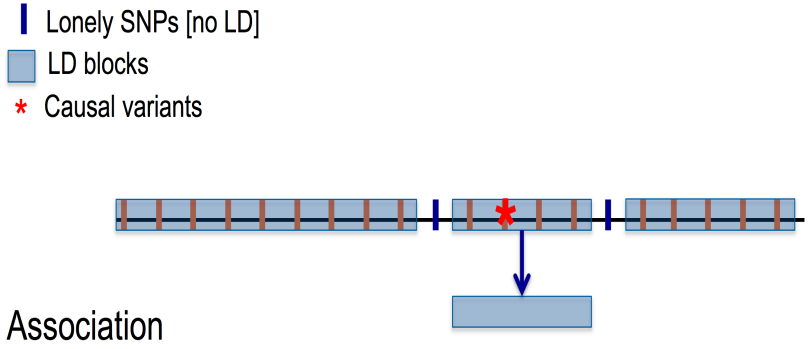
How does LD shape association

A set of markers along a chromosome region:

Superimpose LD between markers



Consider causal SNPs

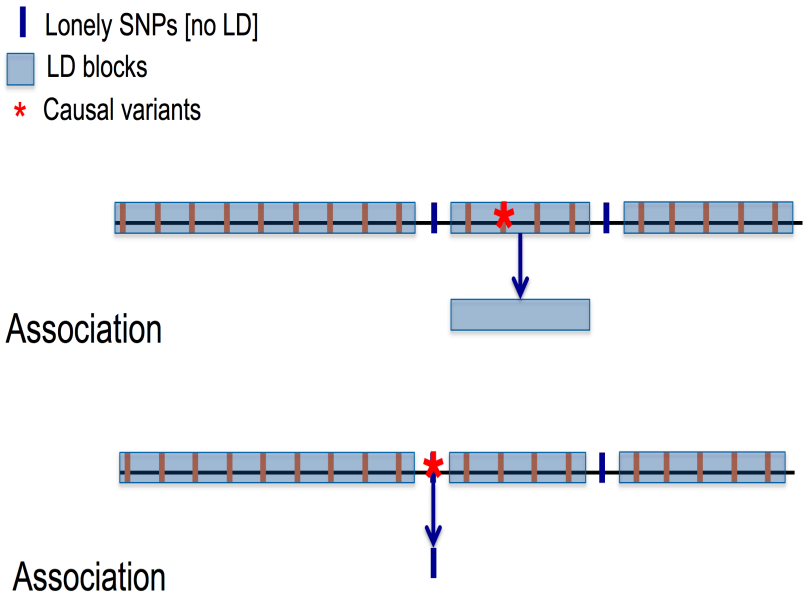


All markers correlated with a causal variant show association

Slide from B Neale

How does LD shape association

Consider causal SNPs



All markers correlated with a causal variant show association.
Lonely SNPs only show association if they are causal
The more you tag the more likely you are to tag a causal variant

Assuming all SNPs gave an equal probability of association given LD status,
we expect to see more association for SNPs with more LD friends.
This is a reasonable assumption under a polygenic genetic architecture

Slide from B Neale

LD score regression

$$l_i := \sum_{j=1}^{M} r_{ij}^2$$

Quantifies local LD for SNP *i*

$$\mathrm{E}[\chi^2 | l_i] = N h^2 \, l_i / M + N a + 1$$

Test statistic is linear in LD score

→ regression of test statistic on LD score provides an estimate of SNP heritability

Use GWAS summary statistics and reference sample for LD score estimation

[Yang 2011 EJHG; Bulik-Sullivan 2015 NG]

Same principle for genetic covariance

$$\mathrm{E}\left[z_{i,1}z_{i,2}|l_i\right] = \frac{\sqrt{N_1N_2}\rho_{\mathrm{g}}}{M}l_i + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

N_s is the number of overlapping samples

z = test statistics from GWAS summary statistics
N = sample size
M = number of markers
ρ_{g} = genetic covariance between traits
ρ = phenotypic correlation between traits

[Bulik-Sullivan 2015 Nature Genetics]

Key concepts

- Mapping QTL by association relies on linkage disequilibrium in the population;
- LD can be caused by close linkage between a QTL and marker (= good) or by confounding between a marker and other effects (= usually bad);
- The power of QTL detection by LD depends on the proportion of phenotypic variance explained at a marker;
- Mixed models are good for performing GWAS
- Genetic (co)variance can be estimated from GWAS summary statistics