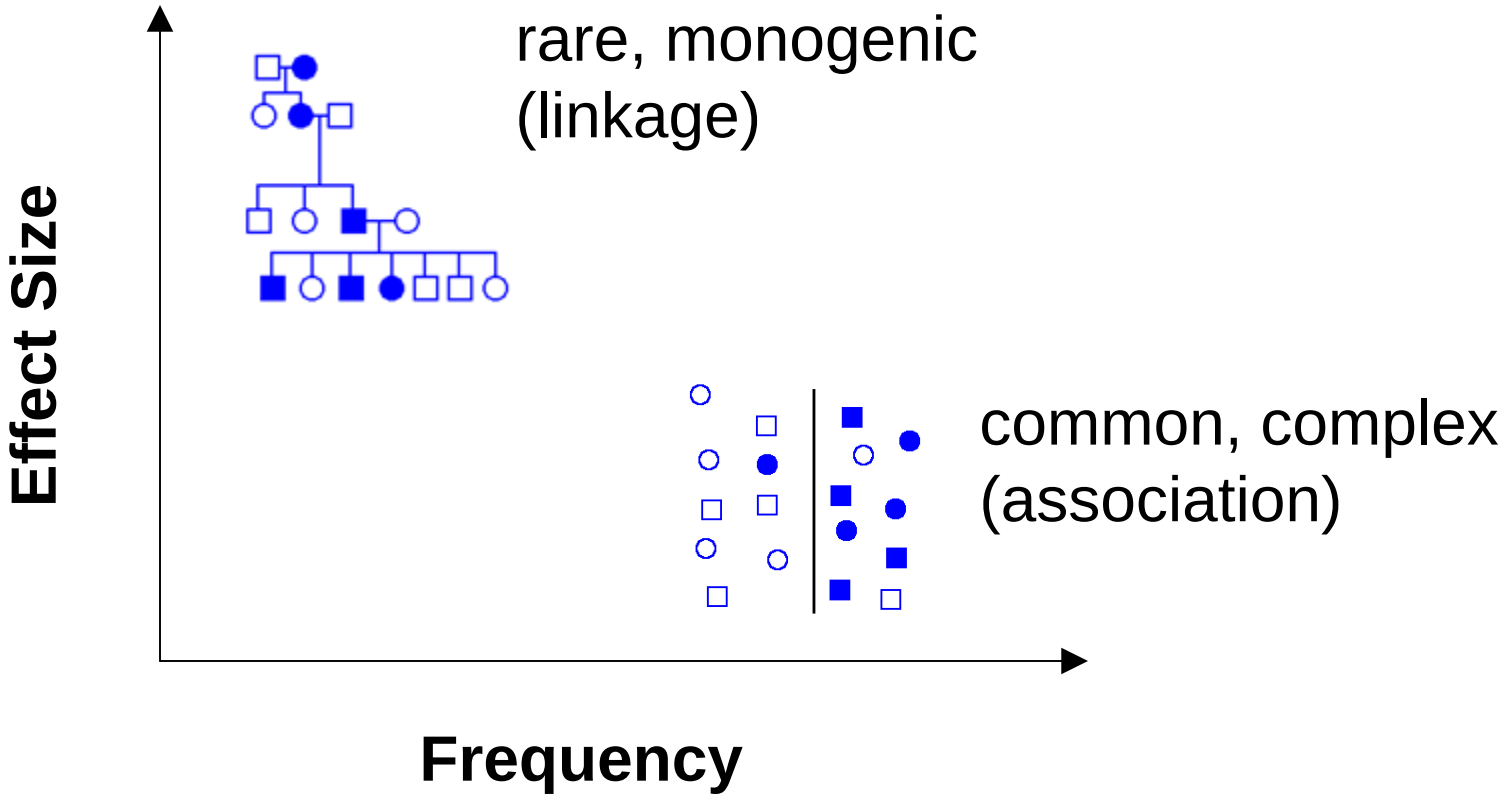
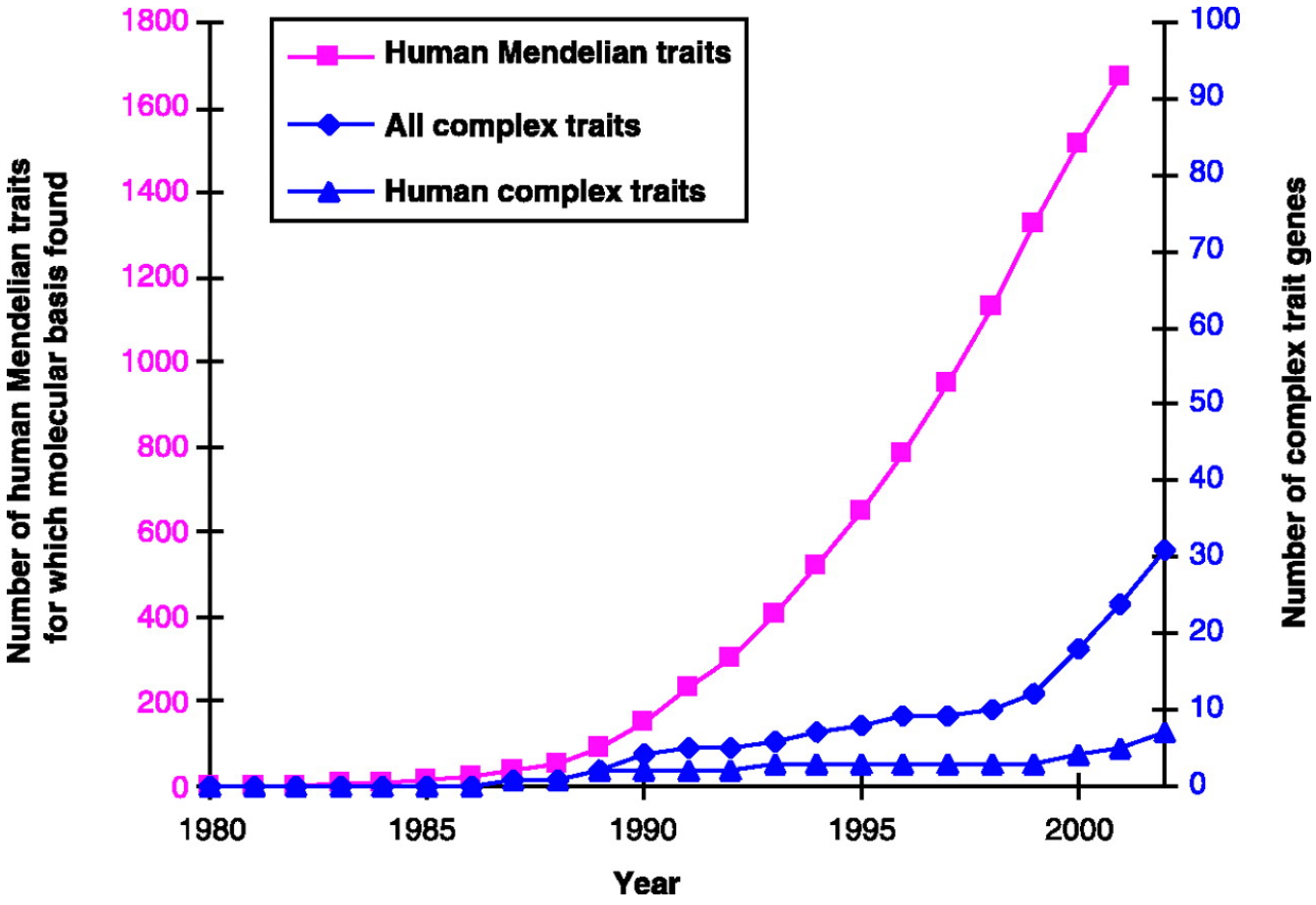
Design and Analysis of Genome-wide Association Studies

David Evans

Methods of gene hunting



Historical gene mapping

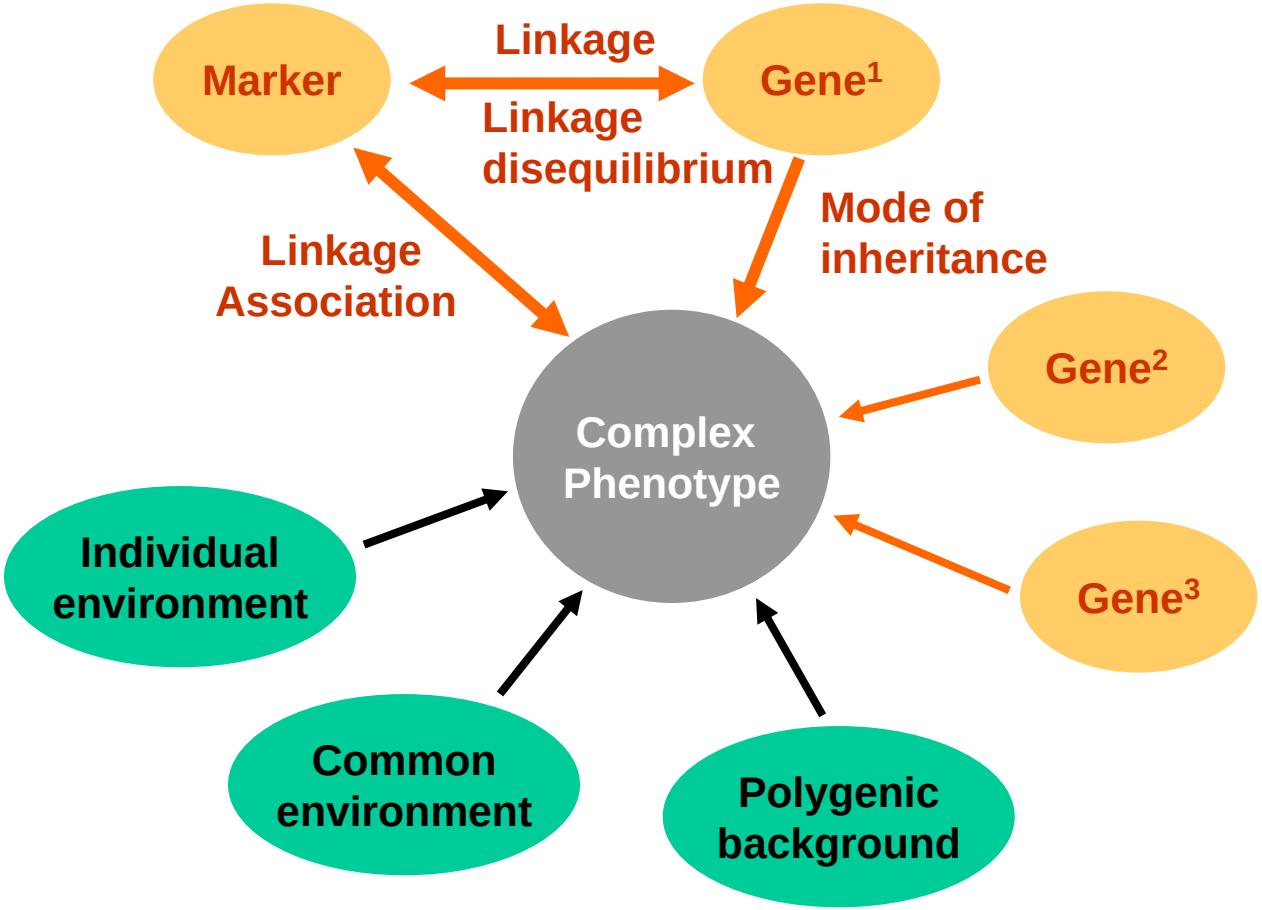


Glazier et al, *Science* (2002).

Reasons for Failure

- Linkage not powerful enough!
- Inadequate Marker Coverage (Candidate gene studies)
- Too optimistic about sample size

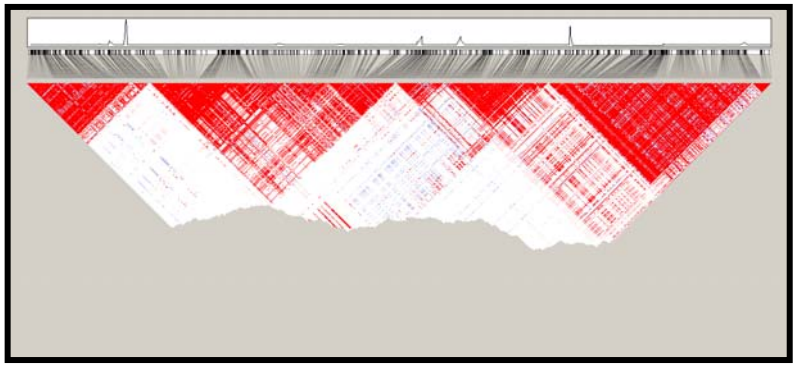
Reasons for Failure?



Weiss & Terwilliger (2000) *Nat Genet*

Enabling Genome-wide Association Studies

- HAPlotype MAP



- High throughput genotyping

- Large cohorts

Wellcome Trust Case Control Consortium

THE INDEPENDENT

(Ireland, €1) 70p

Thursday 7 June 2007

www.independent.co.uk

Tracey Emin

Exclusive: How I created the show of my life

PLUS YOUR CHANCE TO OWN A LIMITED-EDITION ARTWORK IN EXTRA

Bipolar disorder
Also known as manic depression, it affects 100 million people around the world

Coronary heart disease
The most frequent cause of death in Britain, with 100,000 victims every year. By 2020, it will be the biggest killer in the world

Hypertension
High blood pressure affects 16 million people in Britain. Can lead to stroke, heart disease and kidney failure

Rheumatoid arthritis
Nearly 400,000 people in Britain are afflicted with this auto-immune disease of the joints

Type 1 diabetes
Diabetic condition in which sufferers have to inject insulin. Affects 350,000 people in UK

Crohn's disease
Up to 60,000 people are affected by this debilitating bowel condition which can cause distress and pain for a lifetime

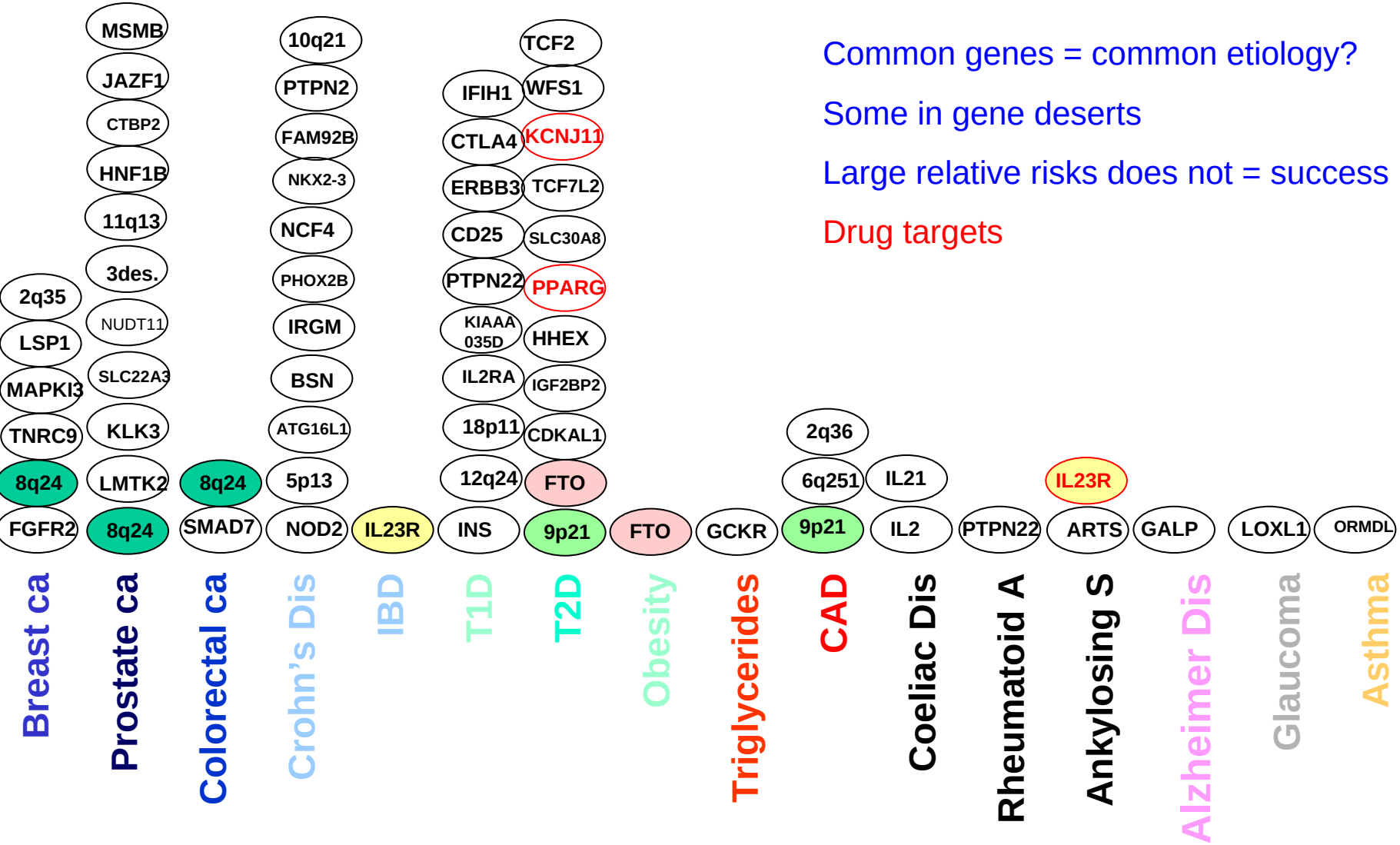
Type 2 diabetes
Almost 2 million Britons are affected by this late-onset disease, which is linked with the growing obesity epidemic

THE GENETIC REVOLUTION

DISCOVERY OF GENES RESPONSIBLE FOR SEVEN OF THE MOST COMMON ILLNESSES OFFERS HOPE TO MILLIONS OF SUFFERERS

FULL STORY, PAGE 2

Successes…



Common genes = common etiology?

Some in gene deserts

Large relative risks does not = success

Drug targets

Study Design

Genetic Power Calculator

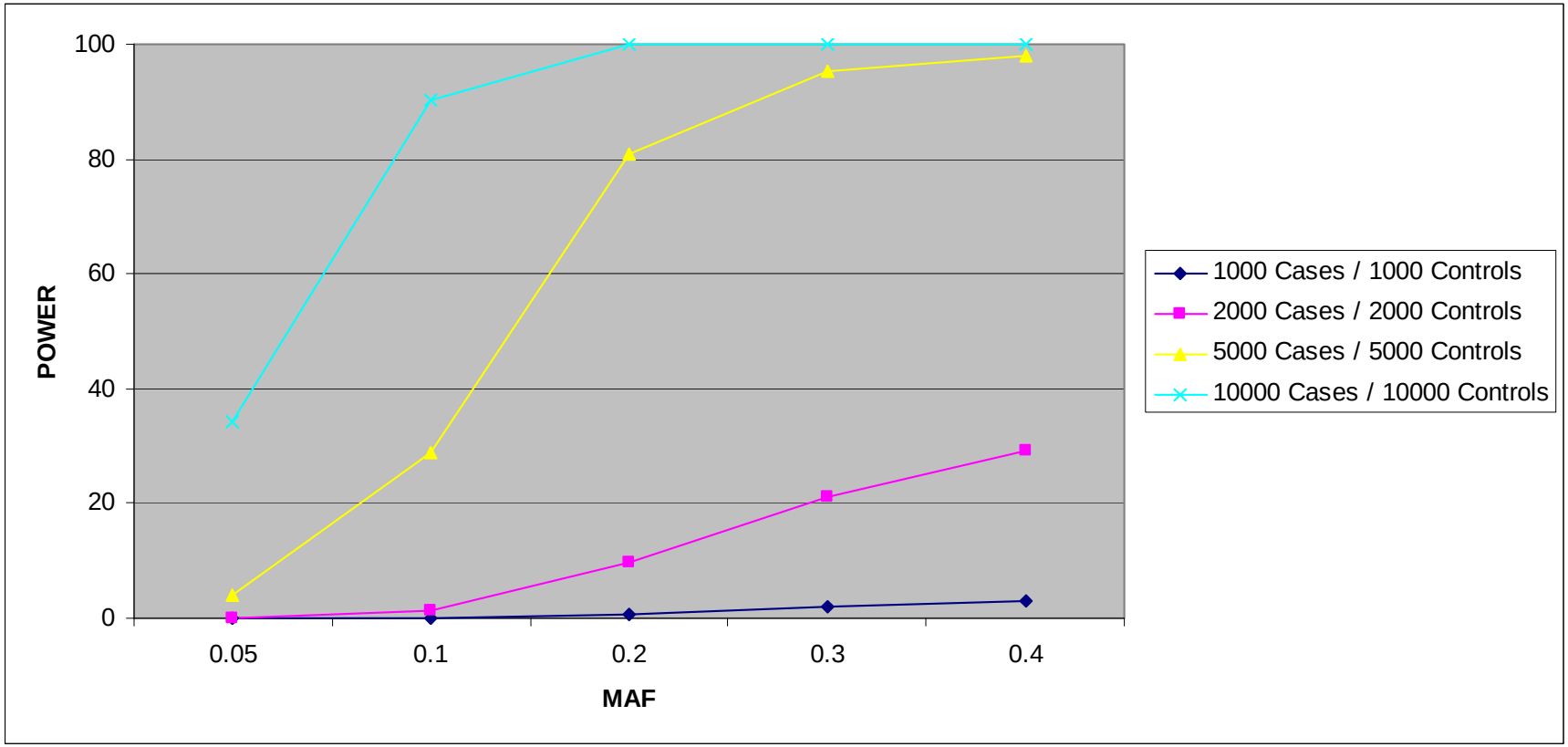
Case - control for discrete traits

High risk allele frequency (A)	:	(0 - 1)
Prevalence	:	(0.0001 - 0.9999)
Genotype relative risk Aa	:	(>1)
Genotype relative risk AA	:	(>1)
D-prime	:	(0 - 1)
Marker allele frequency (B)	:	(0 - 1)
Number of cases	:	(0 - 10000000)
Control : case ratio	:	(>0) (1 = equal number of cases and controls)
User-defined type I error rate	: 0.05	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Process Reset

Created by Shaun Purcell 6.12.2001

Case- Control Studies



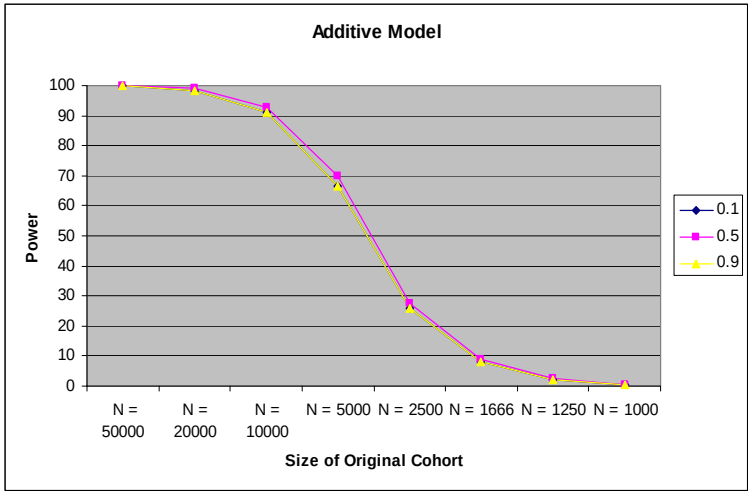
(Multiplicative model;r^2 =1; RR_{Aa} = 1.2; $\alpha = 5 \times 10^{-7}$)

Case to Control Ratio

- Most efficient ratio is 1:1
- Sometimes difficult to recruit cases, in this situation power can still be increased by ascertaining controls
- In the hypothetical situation of an infinite number of controls, only half the number of cases would be required
- Most increase in power occurs when the number of controls is 3 - 5 times the number of cases

Other Strategies to Increase Power

- Minimize phenotypic heterogeneity
- Early age of onset
- Family cases
- Quantitative traits- Extreme cases
- BUT must be careful…

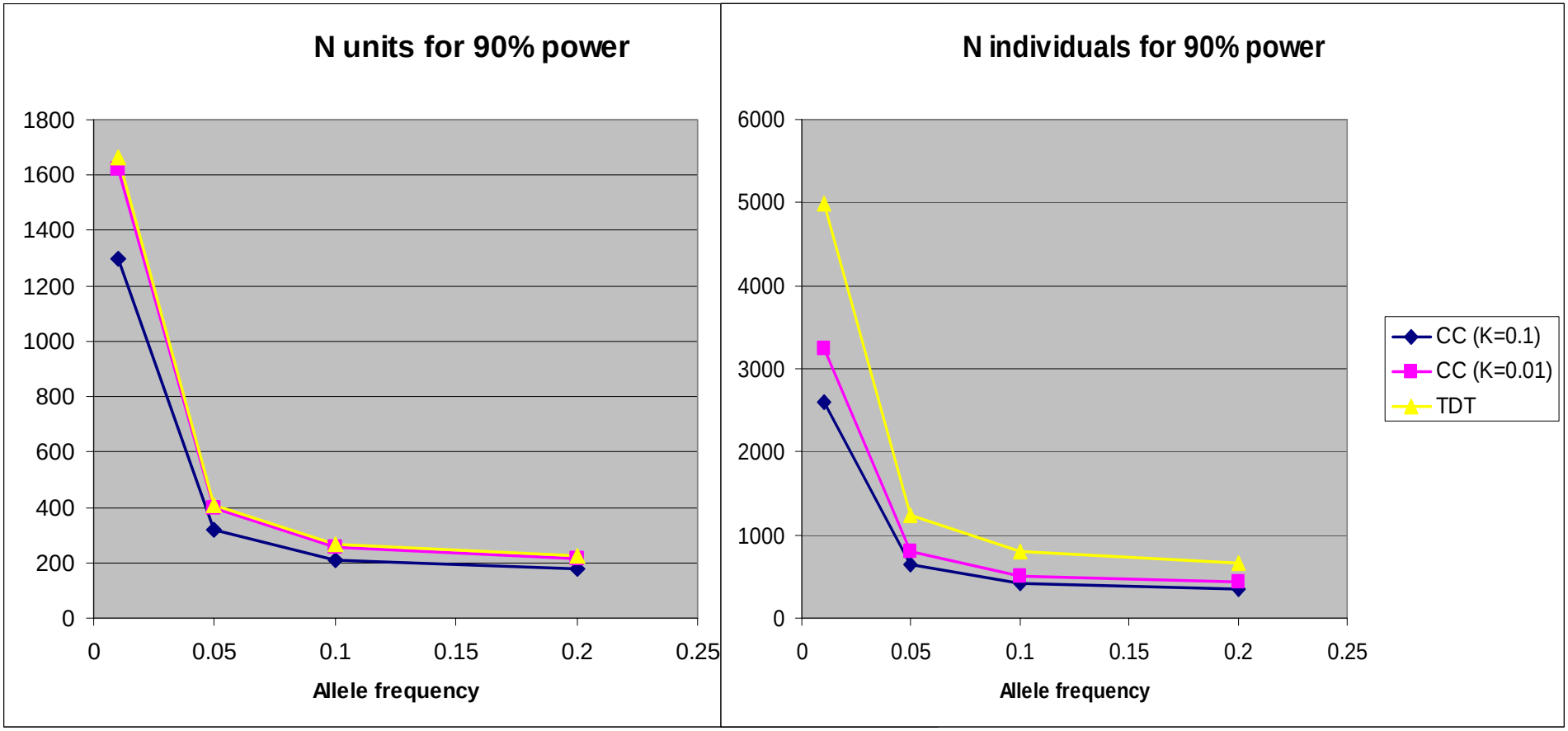


(500 individuals taken from top and bottom; $\alpha = 5 \times 10^{-7}$)

Phenotypic Misclassification

- Misclassification in psychiatric genetics
- Random misclassification should not affect type I error but will decrease power
- Misclassifying cases is not the same as misclassifying controls. The effect of each depends on the prevalence of disease
- For example, for diseases where prevalence less than 10% much more important to ensure cases are truly affected than controls are really unaffected
- Use of historic controls (but note stratification; batch effects; platform differences)

TDT vs Case Control



p = 0.1; RAA = RAa = 2

Number of units similar for each => 2/3 Number of individuals for TDT

Prevalence affects CC power but not TDT power

Quantitative Traits

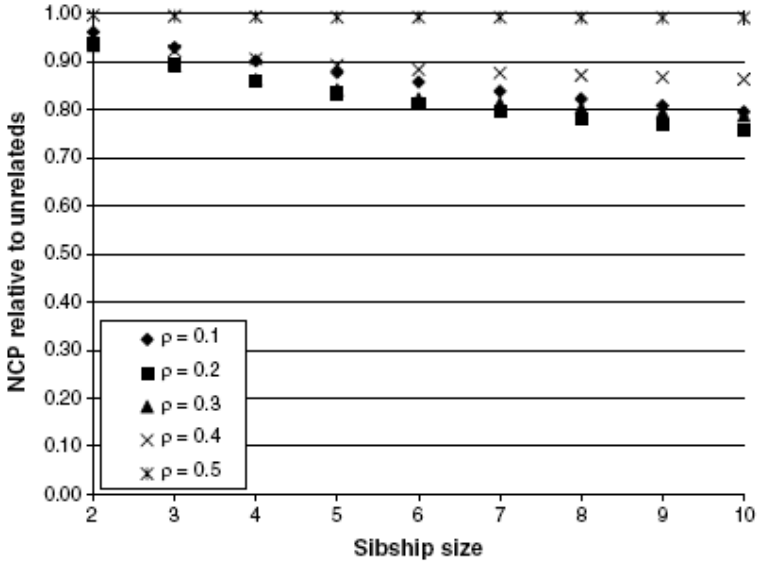


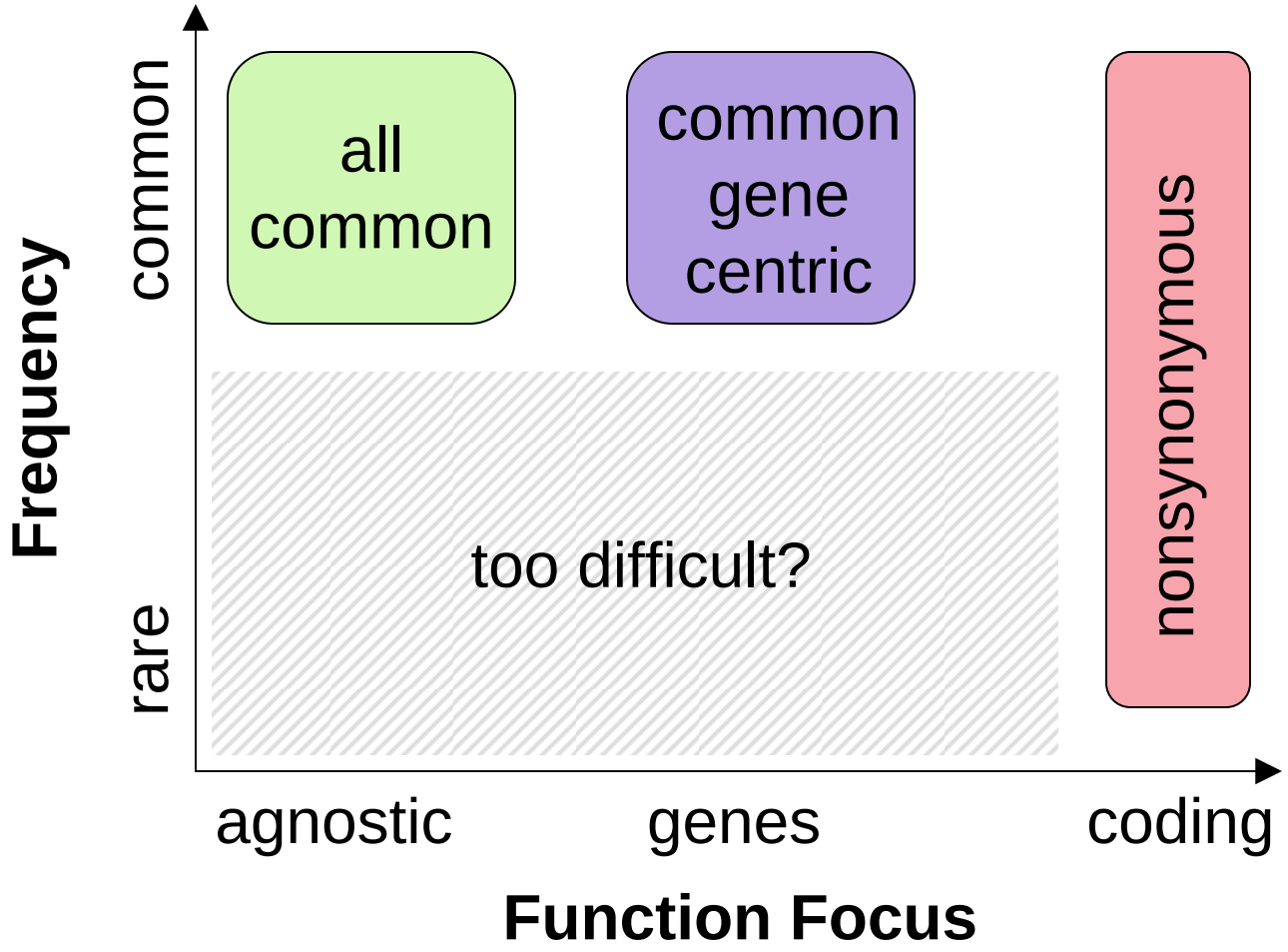
Figure 1 Relative power of GWAS for sibships *versus* unrelated individuals, for the same cost of genotyping. ρ is the phenotypic correlation between siblings.

Visscher et al. (2008) *EJHG*

- Little power lost by analyzing families relative to singletons
- It may be efficient to genotype only some individuals in larger pedigrees
- Pedigrees allow error checking, within family tests, parent-of-origin analyses, joint linkage and association etc

Genotyping Platform

Selecting Markers: Strategies



Some Commercial Alternatives…

Affymetrix SNP array 5.0 (500K)

Affymetrix SNP array 6.0 (1.8M)

Illumina 317K -> Illumina 370K

Illumina 550K -> Illumina 610K

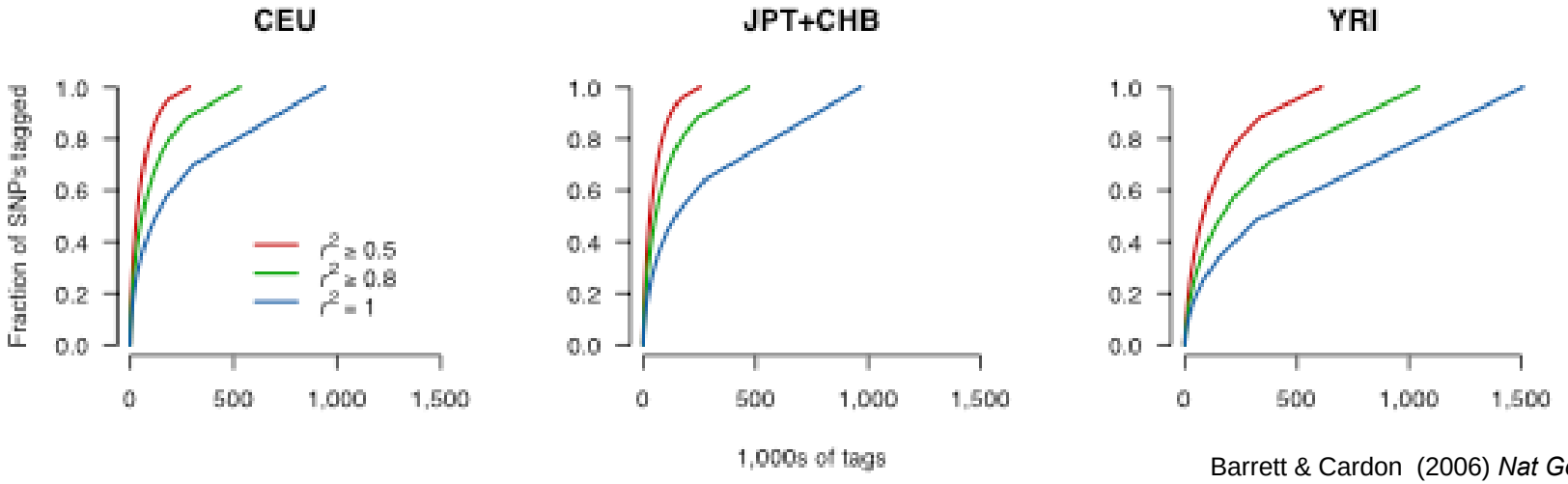
Illumina 1M

Illumina Human Exon 510S

Illumina Human NS_12 Beadchip (15K)

How many SNPs to tag the genome?

- Ideal tag sets



Barrett & Cardon (2006) *Nat Genet*

- 500,000 tags SNPs to tag all common variation in CEU at $r^2 > 0.8$
- Diminishing returns as coverage increases (e.g 250K tags 85% of genome)
- Linear relationship for “singleton” SNPs

How Do The Chips Do?

Table 1. Estimates of Genomic Coverage for Currently Available Genome-wide SNP Platforms Alone and after Imputation

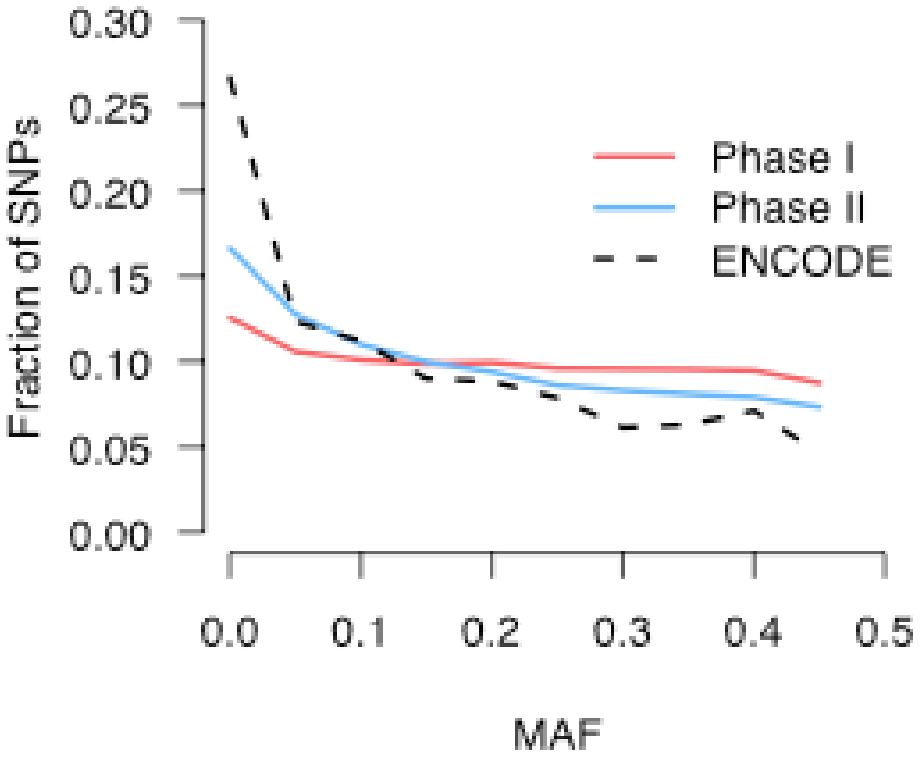
	Percentage of Genomic Coverage at $r^2 \geq 0.8$	Percentage of Genomic Coverage at $r^2 = 1$
Affymetrix SNP Array 5.0	65	43
Affymetrix SNP Array 5.0 plus imputed SNPs	73	54
Affymetrix SNP Array 6.0	80	59
Illumina HumanHap 300	77	42
Illumina HumanHap 300 plus imputed SNPs	81	50
Illumina HumanHap 550	87	57
Illumina HumanHap 1M	91	68

Estimates evaluated with Phase II HapMap data from the CEU population. Coverage estimates for Illumina HumanHap 1M and Affymetrix SNP-array-6.0 are likely to be biased downward because the genotypes at approximately 10% of the SNPs on each platform are not currently publicly available for the CEU HapMap individuals. Where imputations are included, all SNPs passing imputation-filter thresholds and with an $r^2 \geq 0.8$ between known and imputed genotypes are included along with the SNPs on the genome-wide SNP chip.

Anderson et al. (2008) Nature Genetics

- Some of the difference in coverage can be recovered through imputation
- If sample size limited, but funding not, use chip with best coverage
- If cost limited but sample size not use Illumina 300K? (Cost efficiency)

Most SNPs are Rare



- Hapmap and SNP chips biased towards common variants
- Rare SNPs are not tagged well by common SNPs!

What about nsSNP chips?

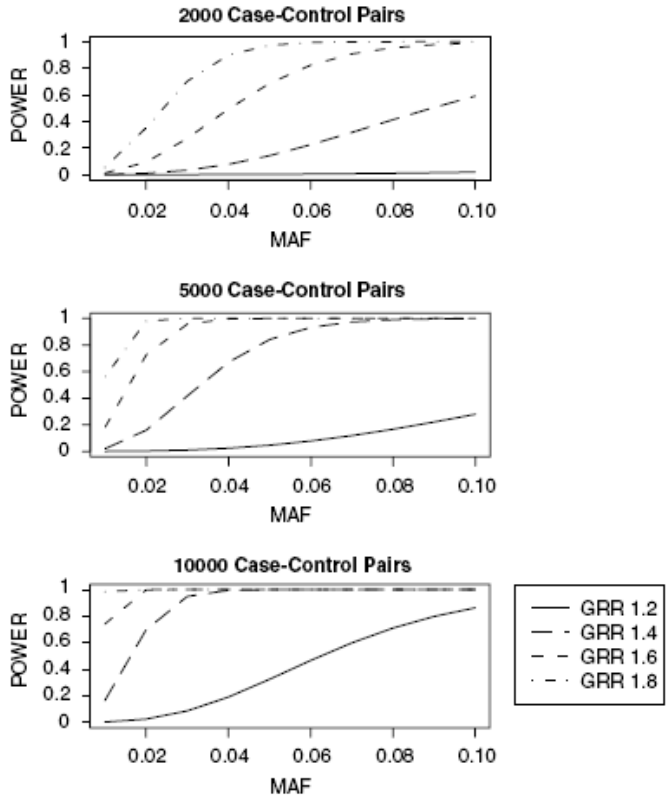


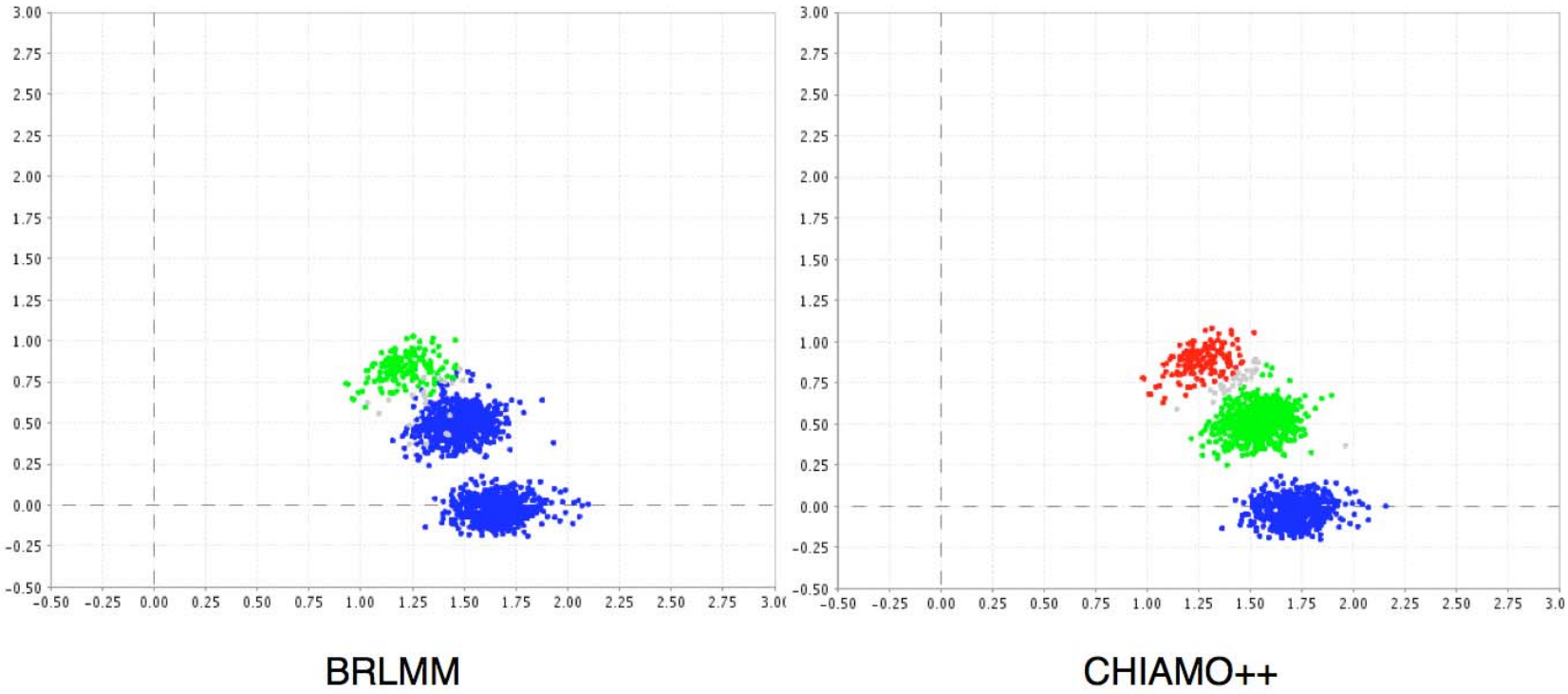
Figure 2 Relationship between MAF, heterozygote GRR, and power to detect association assuming a multiplicative disease model. Results are shown for 2000, 5000, and 10 000 case–control pairs assuming a disease prevalence of 1% and a type I error rate of $\alpha = 3.6 \times 10^{-6}$. The figure illustrates that it is possible to detect rare variants of intermediate penetrance using current sample sizes of 2000 case–control pairs. To detect rare alleles of smaller effect, far larger sample sizes will need to be employed.

- Non-synonymous SNPs produce changes in amino acid sequence
- Most common nsSNPs tagged by existing genome-wide products
- Little to add to genome-wide chips in terms of identifying common variants
- May help identify rare variants of intermediate penetrance

Evans et al. (2008) *EJHG*

“Cleaning” Data

Genotypes are not raw data

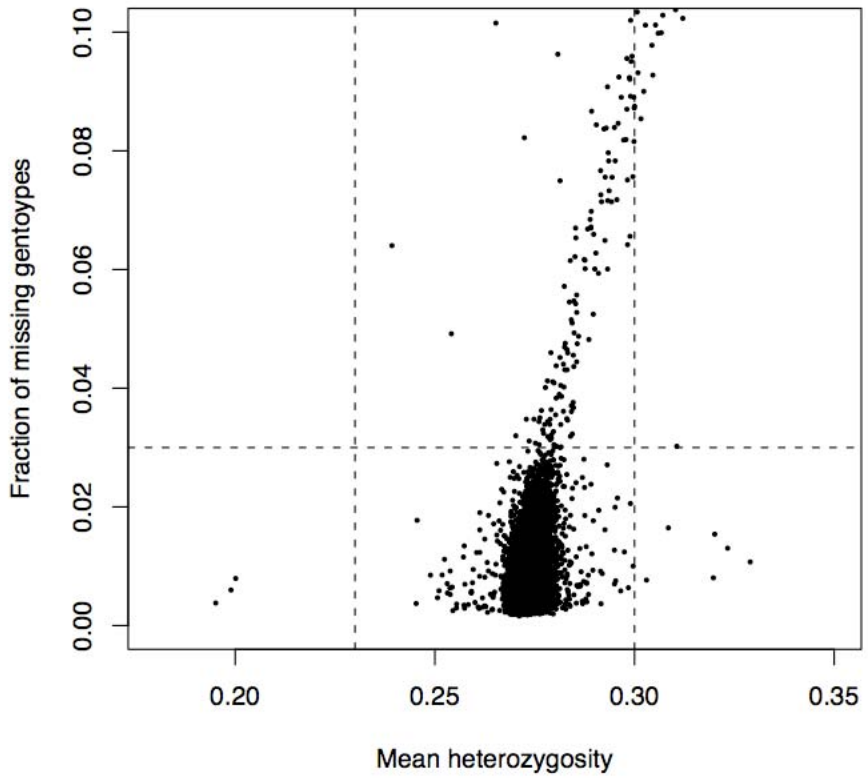


- Trade off between stringency and call rate (no universal value)
- Raw intensities of ALL putative associations should be checked!

SNP Quality Control

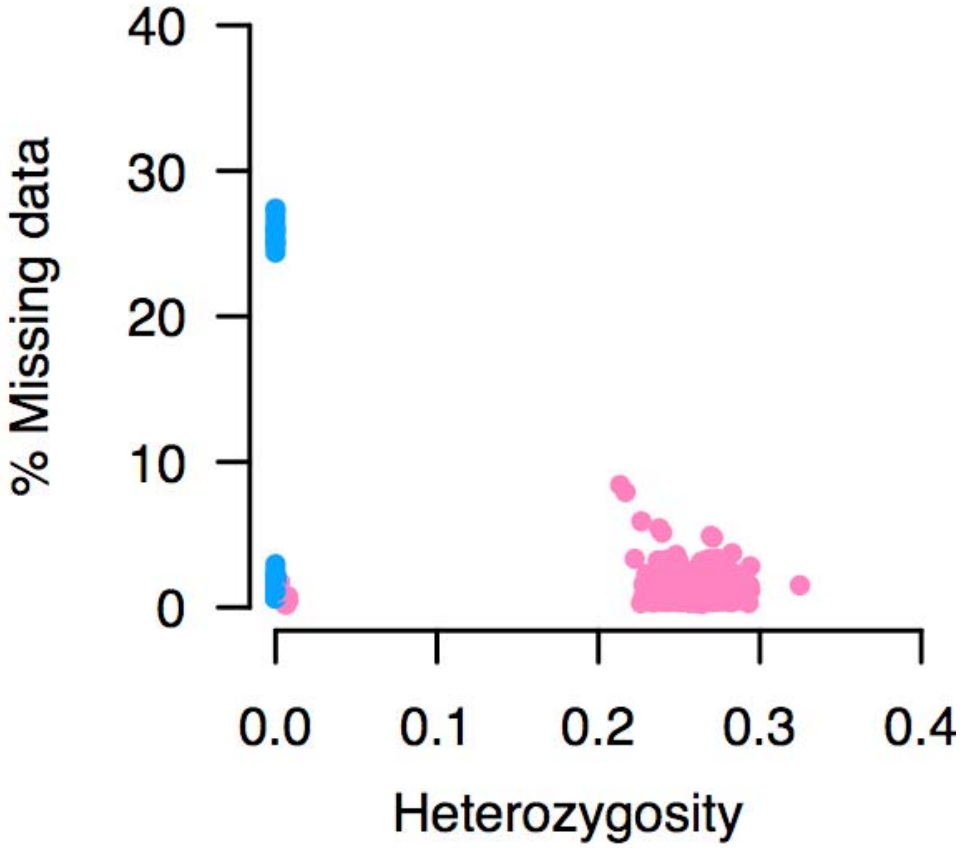
- Missing Data Rate (SNPs, Individuals, cases vs controls)
- Hardy Weinberg Equilibrium
- Allele frequency
- Mendelian Inconsistencies

Sample Heterozygosity



$$\text{het.} = \frac{N_{het}}{N_{het} + N_{hom}}$$

Sample Gender



Association Analysis

Genotypic tests

- SNP marker data can be represented in 2x3 table.
- Test of association

$$X^2 = \sum_{i=0,1,2} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

where

$$E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

- X^2 has χ^2 distribution with 2 degrees of freedom under null hypothesis.

	Cases	Controls	Total
MM	n_{2A}	n_{2U}	$n_{2\cdot}$
Mm	n_{1A}	n_{1U}	$n_{1\cdot}$
mm	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Sensitive to genotyping error
- Often not as powerful as trend test

Allele-based tests

- Each individual contributes two counts to 2x2 table.
- Test of association

$$X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

where

$$E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

- X^2 has χ^2 distribution with 1 degrees of freedom under null hypothesis.

	Cases	Controls	Total
M	n_{1A}	n_{1U}	$n_{1\cdot}$
m	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Assumes cases and controls in HWE
- Assumes multiplicative disease model

Logistic regression framework

- Model case/control status within a logistic regression framework.
- Let π_i denote the probability that individual i is a case, given their genotype G_i.
- Logit link function

where

$$\pi_i = \Pr\left(\text{i is case} \mid G_i, \boldsymbol{\beta}\right) = \frac{\exp\left[\eta_i\right]}{1+\exp\left[\eta_i\right]}$$

$$\eta_i = \begin{cases} \beta_0 & \text{null model} \\ \beta_0 + \beta_M Z_{(M)i} & \text{additive model} \\ \beta_0 + \beta_{Mm} Z_{(Mm)i} + \beta_{MM} Z_{(MM)i} & \text{genotype-based model} \end{cases}$$

Indicator variables

- Represent genotypes of each individual by indicator variables:

	Additive model	Genotype model	
Genotype	$Z_{(M)i}$	$Z_{(Mm)i}$	$Z_{(MM)i}$
mm	0	0	0
Mm	1	1	0
MM	2	0	1

Likelihood calculations

- Log-likelihood of case-control data given marker genotypes

$$\ell\left(\mathbf{y}\middle|\mathbf{G},\boldsymbol{\beta}\right)=\sum_{i} y_i \ln\left[\pi_i\right]+\left(1-y_i\right)\ln\left[1-\pi_i\right]$$

  where $y_i = 1$ if individual i is a case, and $y_i = 0$ if individual i is a control.
- Maximise log-likelihood over $\boldsymbol{\beta}$ parameters, denoted $\ell\left(\mathbf{y}\middle|\mathbf{G},\hat{\boldsymbol{\beta}}\right)$
- Models fitted using PLINK.
- Additive model equivalent to Armitage test for trend

Model comparison

- Compare models via deviance, having a χ^2 distribution with degrees of freedom given by the difference in the number of model parameters.

Models	Deviance	df
Additive vs null	$2\left[\ell\left(\mathbf{y}\middle\vert\mathbf{G},\hat{\beta}_M,\hat{\beta}_0\right)-\ell\left(\mathbf{y}\middle\vert\mathbf{G},\hat{\beta}_0\right)\right]$	1
Genotype vs null	$2\left[\ell\left(\mathbf{y}\middle\vert\mathbf{G},\hat{\beta}_{MM},\hat{\beta}_{Mm},\hat{\beta}_0\right)-\ell\left(\mathbf{y}\middle\vert\mathbf{G},\hat{\beta}_0\right)\right]$	2

Covariates

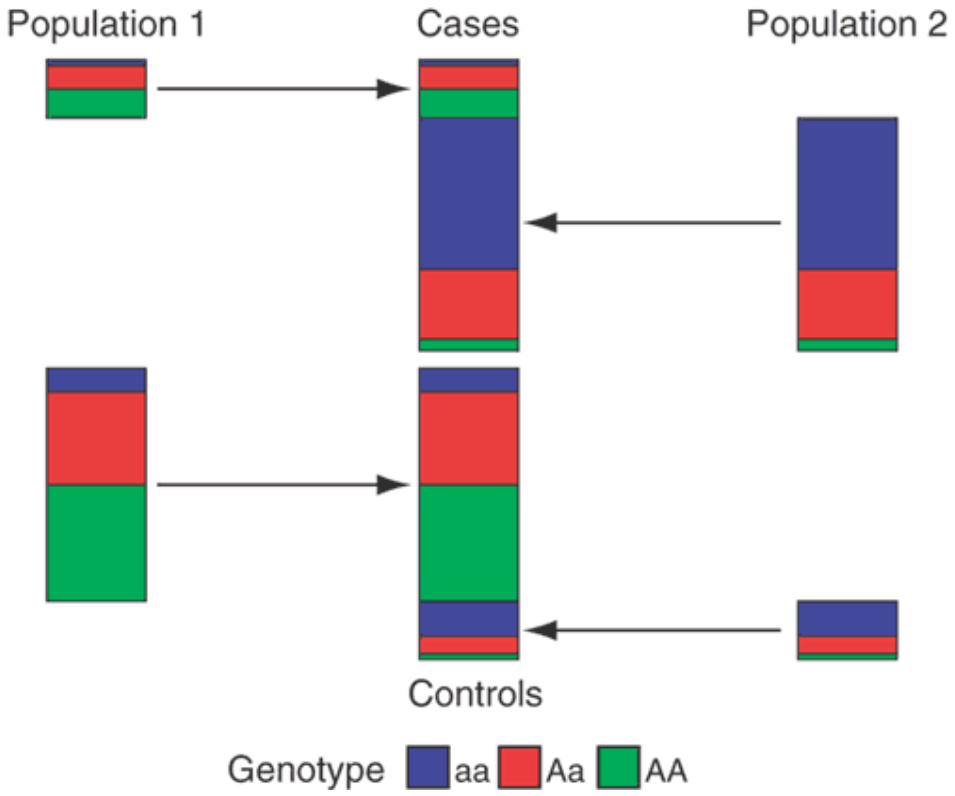
- It is straightforward to incorporate covariates in the logistic regression model:
  - age, gender, and other environmental risk factors.
  - genotypes at unlinked markers to control for population stratification.
- Generalisation of link function, e.g. for additive model:

$$\eta_i = \beta_0 + \beta_M Z_{(M)i} + \sum_j \gamma_j X_{ij}$$

where X_{ij} is the response of individual i to the jth covariate, and γ_j is the corresponding covariate regression coefficient.

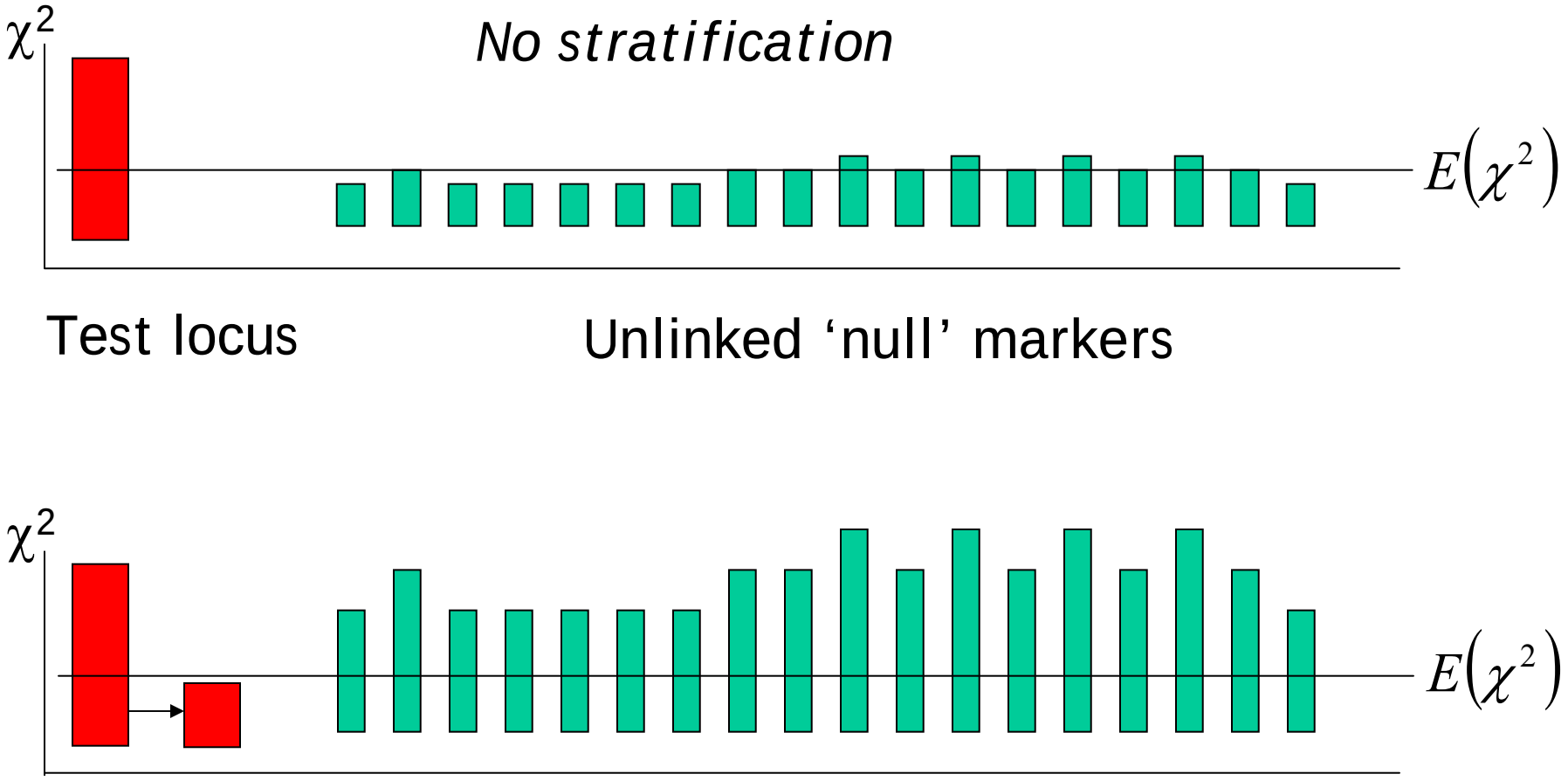
Controlling for Population Stratification

Population structure



Marchini, *Nat Genet* (2004)

Genomic control

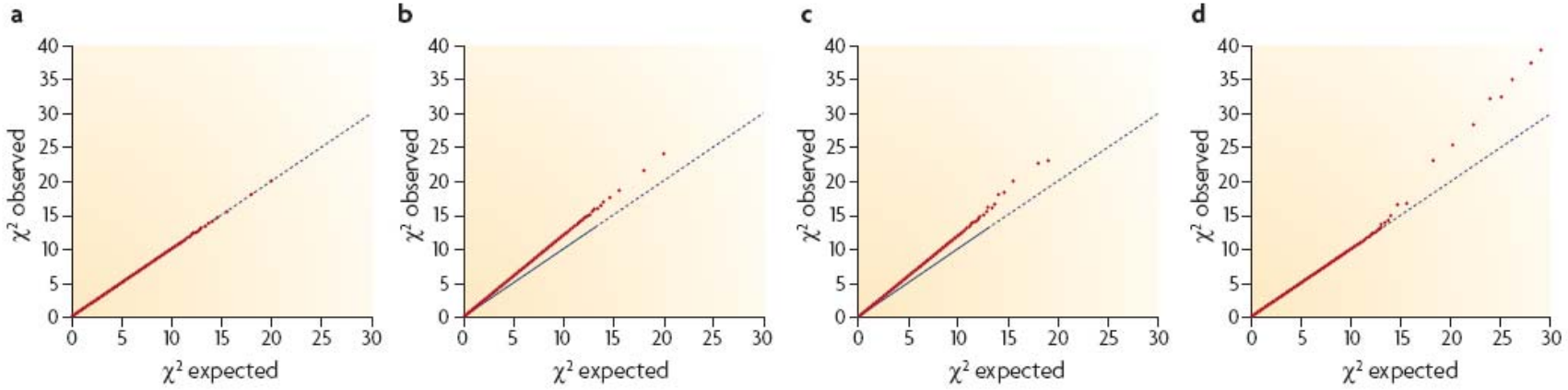


Stratification → adjust test statistic

'λ' is inflation factor (=1 if no inflation)

QQ plots

Box 2 | **Visualization of genome-wide association data**



McCarthy et al. (2008) Nature Genetics

Population structure - λ

Genomic control - λ

genome-wide inflation of median test statistic

BD	1.15
CAD	1.08
HT	1.09
CD	1.26
RA	1.06
T1D	1.07
T2D	1.10

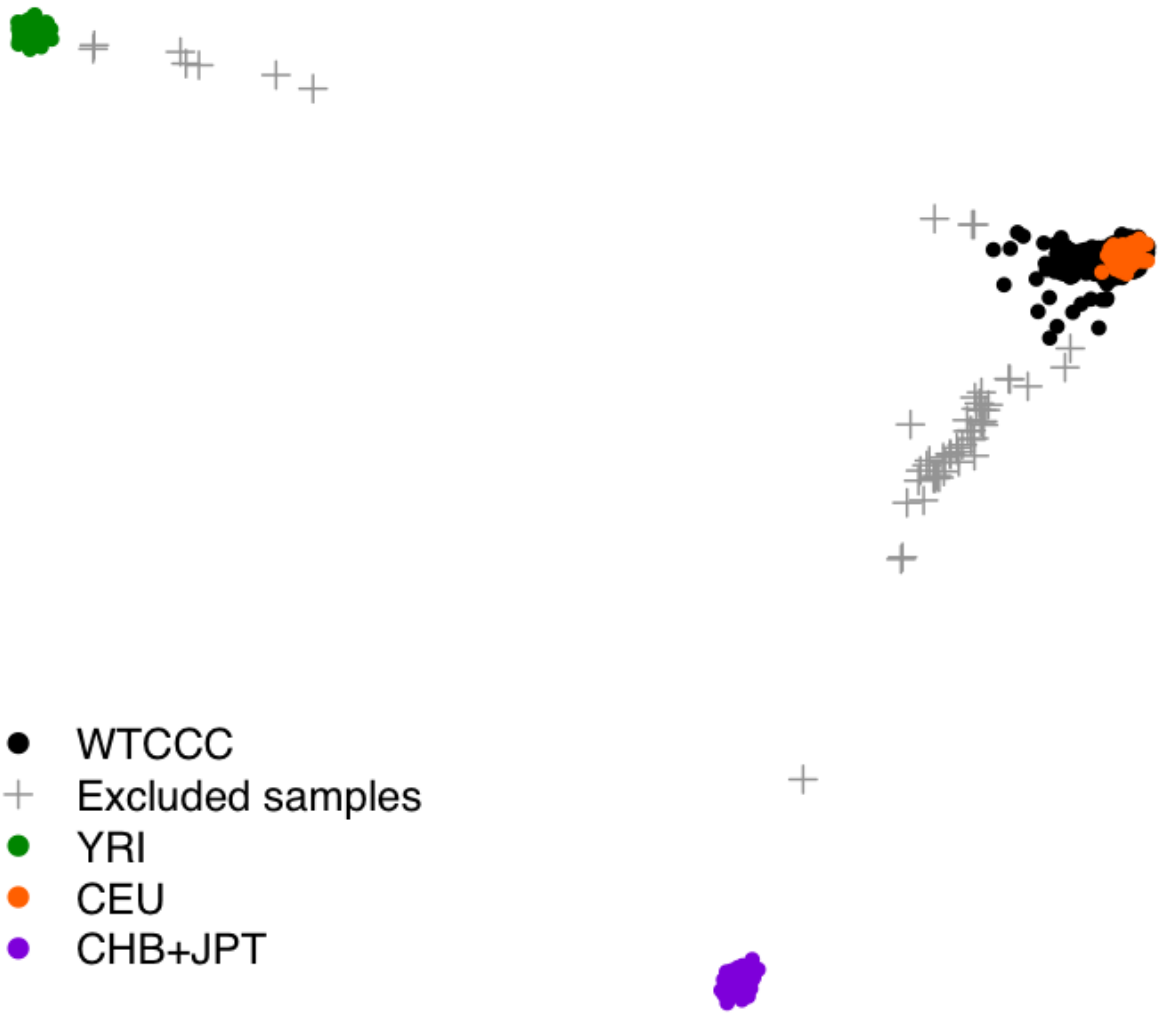
Crohn's collection center

Center	No. of samples
1	524
2	271
3	439
4	465
5	301

Center 3: $\lambda = 1.77$

All others: $\lambda = 1.09$

Crohn's Multidimensional Scaling



Principal Components Analysis

- Principal Components Analysis is a data reduction technique where many variables are reduced to a few “principal components”:
  - Each component describes as much variability as possible
  - Components are orthogonal and describe consecutively smaller proportions of the variance
  - First few components reflect population ancestry

- Genotypes and phenotypes are adjusted by amounts attributable to ancestry along each component by computing residuals of linear regressions

- Association statistics are computed using ancestry adjusted genotypes and phenotypes

Geographic Interpretation

South-east Europe

North-west Europe

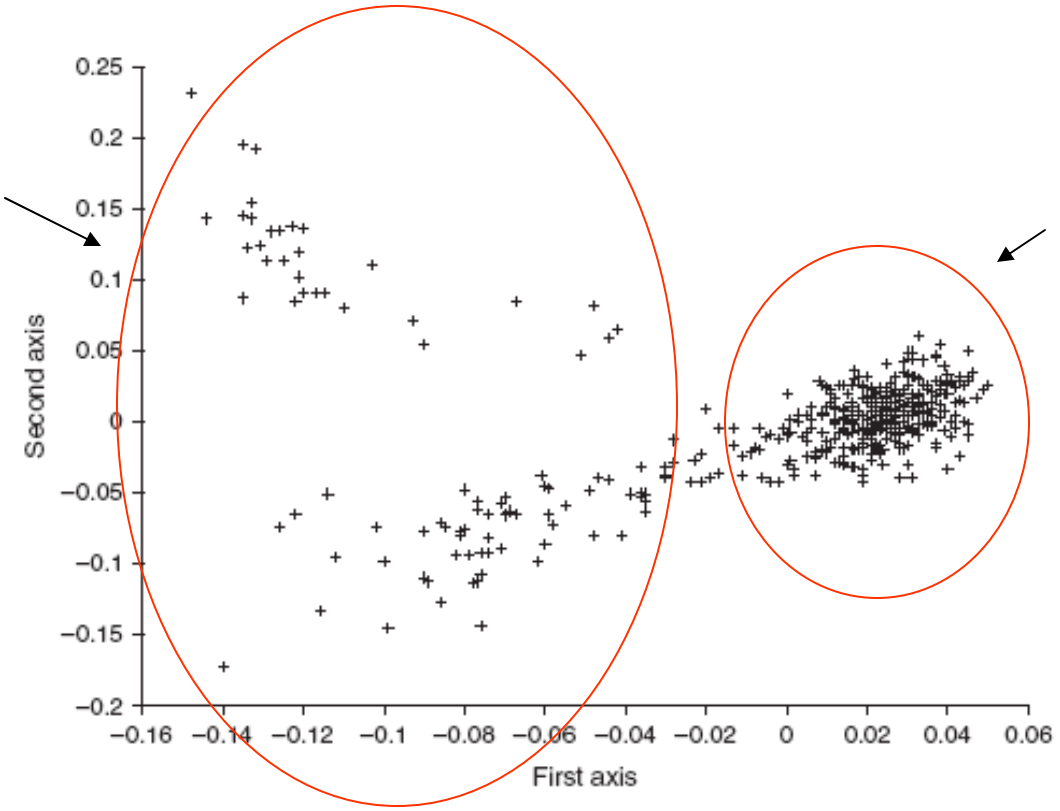
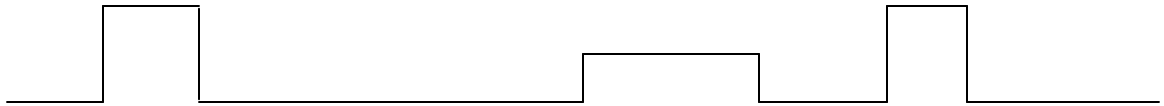


Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

Imputation

Imputation



Recombination Rate

1 1	0 1 1 0 1 0	1 1	0 1 1 0……
1 1	0 1 1 0 1 0	1 1	0 1 1 0……
1 1	0 1 1 0 1 0	1 1	0 1 1 0……
1 1	0 1 1 0 1 0	1 1	0 1 1 0……

HapMap Phase II

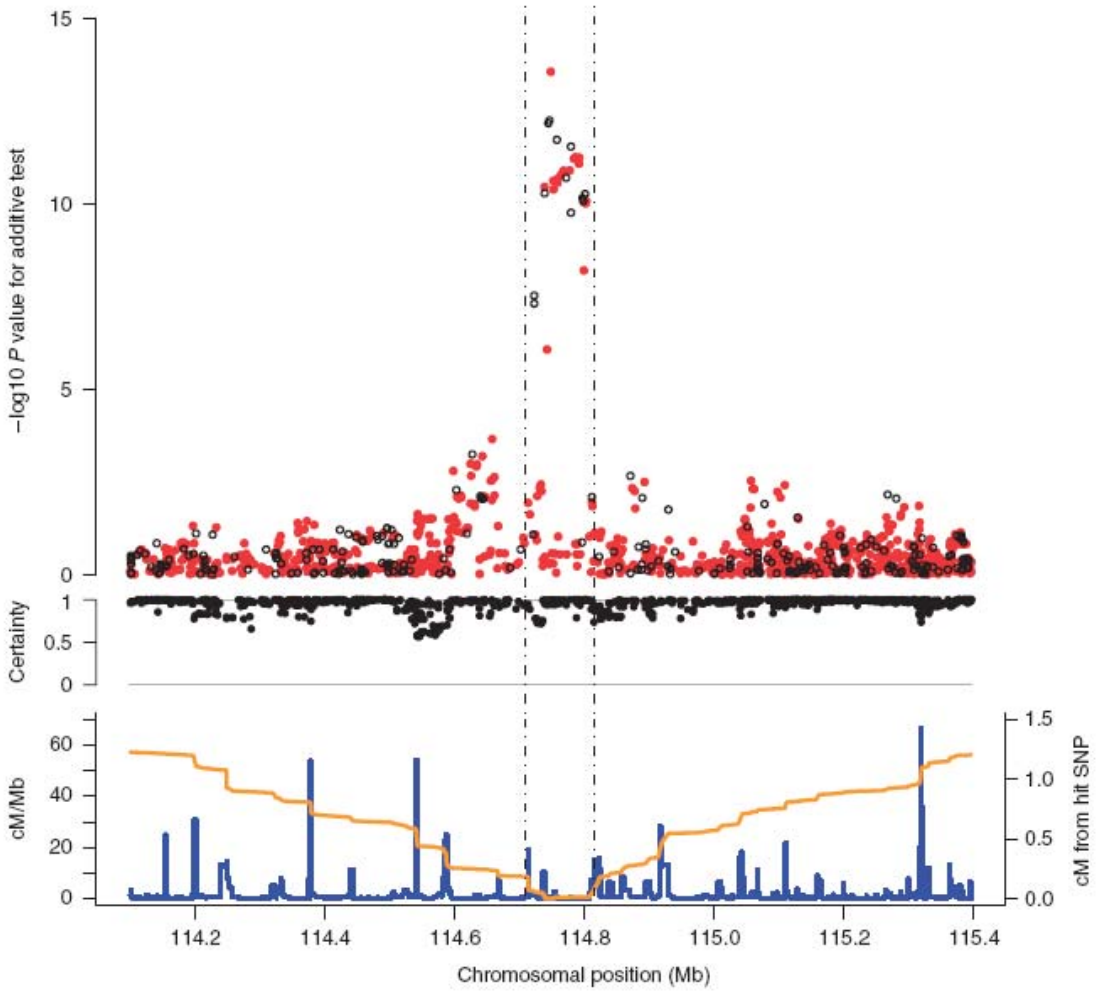
2 1	1 2 ? 2 1 ?	? 1	? 2 2 0……
2 1	1 2 ? 2 2 ?	? 0	? 2 1 0……
2 ?	1 2 ? 2 1 ?	? 1	? 1 1 0……
2 1	2 1 ? 2 2 ?	? 1	? 1 1 0……

Cases

2 1	1 1 ? 2 1 ?	? 1	? 2 2 0……
2 1	1 1 ? 2 2 ?	? 0	? 2 2 0……
1 0	1 2 ? 2 1 ?	? 1	? 1 1 ?……
2 1	2 1 ? 2 2 ?	? 1	? 1 1 ?……

Controls

Imputation


15
10
5
0
−log10 P value for additive test
Certainty
1
0.5
0
60
40
20
0
cM/Mb
1.5
1.0
0.5
0.0
cM from hit SNP
114.2
114.4
114.6
114.8
115.0
115.2
115.4
Chromosomal position (Mb)

Interpretation and Prioritizing SNPs

Asymptotic P values

- "The probability of observing the test result or a more extreme value than the test result under the null hypothesis"
- The p value is NOT the probability that the null hypothesis is true
- The probability that the null/alternate hypothesis is true is a function of the evidence contained in the data (p value), the power of the test, and the prior probability that the association is true/false
- The p value is a fluid measure of the strength of evidence against the null hypothesis that was designed to be interpreted in conjunction with other (pre-existing) evidence

Interpreting p values

STRONGER EVIDENCE	WEAKER EVIDENCE
Genotyping error unlikely	"Suspicious" SNP
Stratification unlikely	Stratification possible
Low p value	Borderline p value
Powerful Study	Weak Study
High MAF	Low MAF
Candidate Gene	Intergenic region
Previous Association	No previous evidence

Criticisms of p values

- Doesn't formally incorporate prior information
- Discards information on the power of the test
- Does not take into account the size of the observed effect
- Ranking SNPs by p value is problematic!!!

Multiple Testing

- Multiple Testing Problem: The probability of observing a "significant" result purely by chance increases with the number of statistical tests performed

- For testing 500,000 SNPs
- 5,000 expected to be significant at $\alpha < .01$
- 500 expected to be significant at $\alpha < .001$
- …
- 0.05 expected to be significant at $\alpha < 10^{-7}$

- One solution is to maintain $\alpha_{FWER} = .05$

- Bonferroni correction for m tests
- Set significance level to $\alpha = .05/m$

- "Genome-wide Significance" suggested at around $\alpha = 5 \times 10^{-7}$

Problems with Bonferroni Adjustments

- Bonferroni adjustments are conservative when statistical tests are not independent
- Bonferroni adjustments control the error rate associated with the omnibus null hypothesis
- The interpretation of a finding depends on how many statistical tests were performed
- What tests should be included?
- Bonferroni adjustments decrease power

Permutation Testing

- The distribution of the test statistic under the null hypothesis can be derived by shuffling case-control status relative to the genotypes, and performing the test of association many times

- Permutation breaks down the relationship between genotype and phenotype but maintains the pattern of linkage disequilibrium in the data

- Appropriate for rare genotypes, small studies, non-normal phenotypes etc.

Replication

- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect
- It is unlikely that a single study will unequivocally establish an association without the need for replication

Guidelines for Replication

Replication studies should be of sufficient size to demonstrate the effect

Replication studies should conducted in independent datasets

Replication should involve the same phenotype

Replication should be conducted in a similar population

The same SNP should be tested

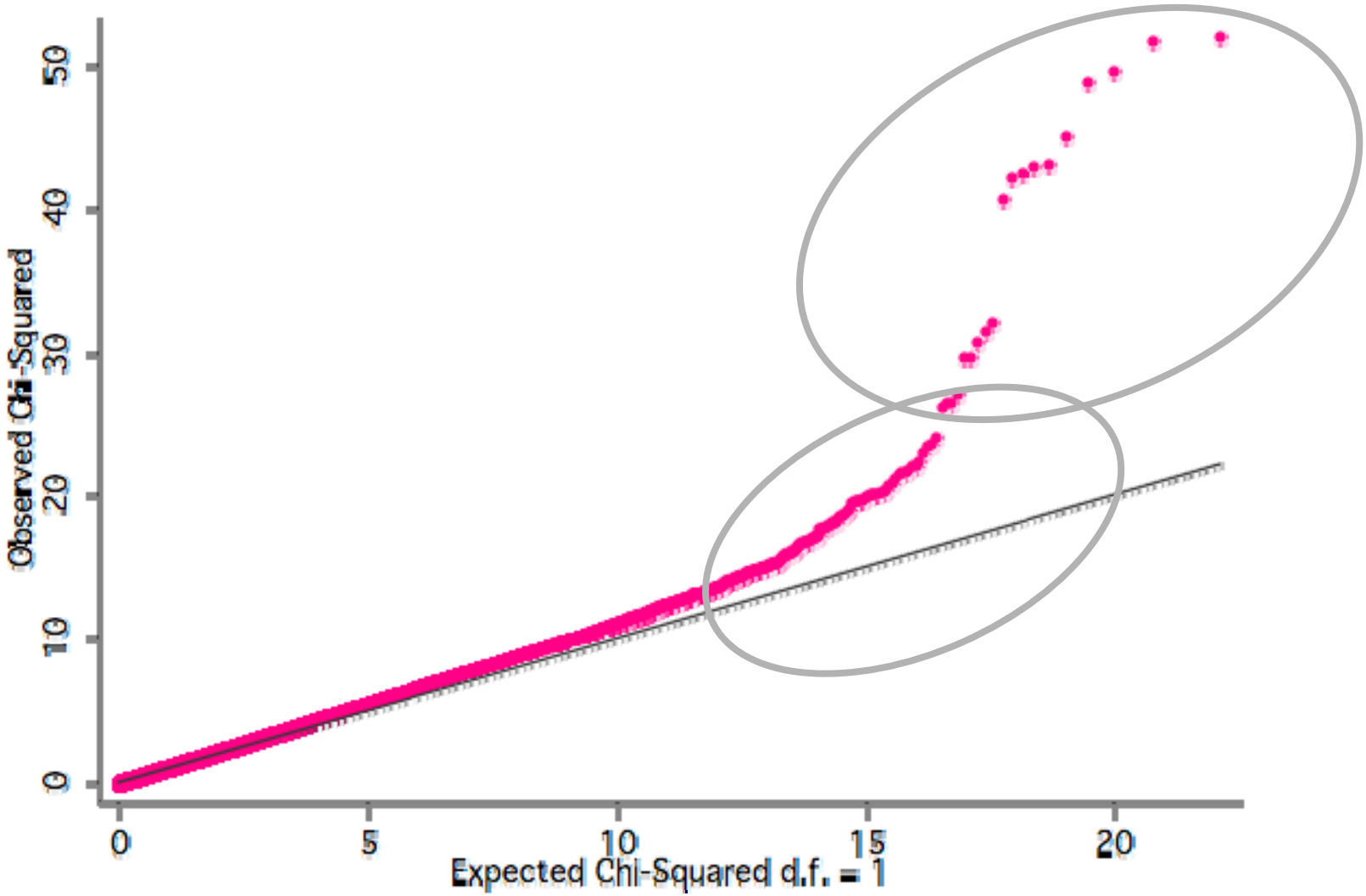
The replicated signal should be in the same direction

Joint analysis should lead to a lower p value than the original report

Well designed negative studies are valuable

Meta-analysis

Meta-analysis



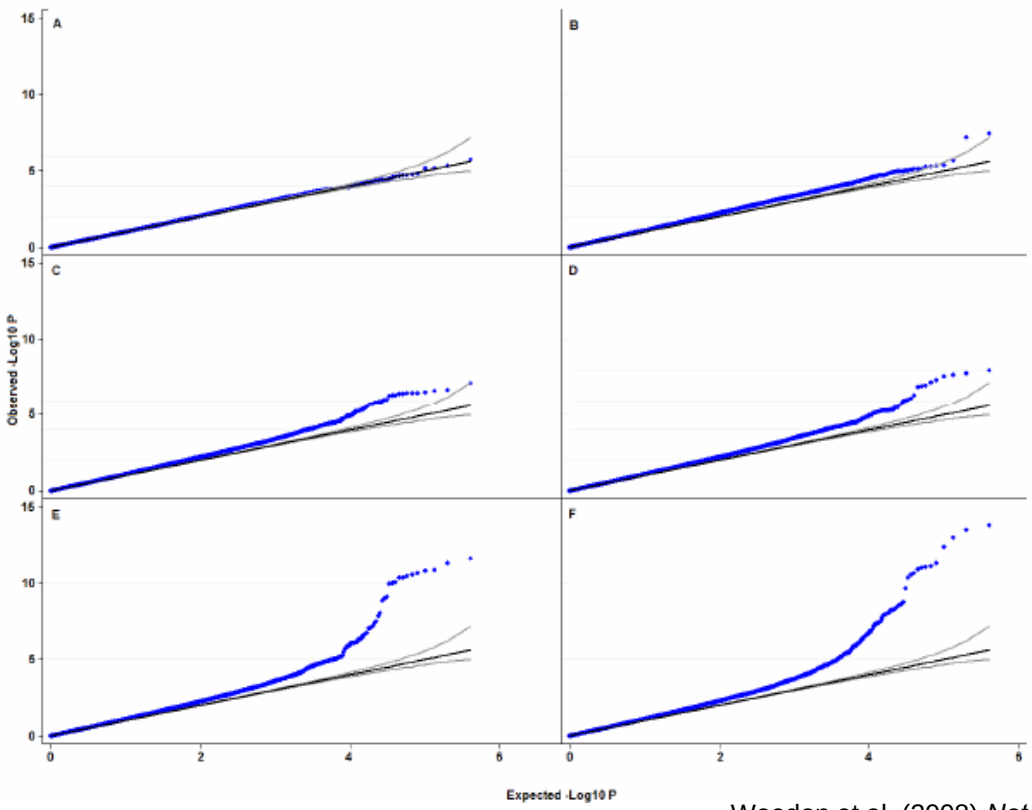
Meta-analysis

- Aims to combine statistical evidence from different studies
- Aims to provide a better estimate of the underlying effect size
- In the context of GWA used to identify polymorphisms that contribute to variation but are located lower down the distribution

Meta-analysis

- Larger studies carry more weight
- Fixed versus Random Effects
- Assessment of Heterogeneity

Example: Meta-analysis of Height



A- 1914 Cases (WTCCC T2D)

B- 4892 Cases (DGI)

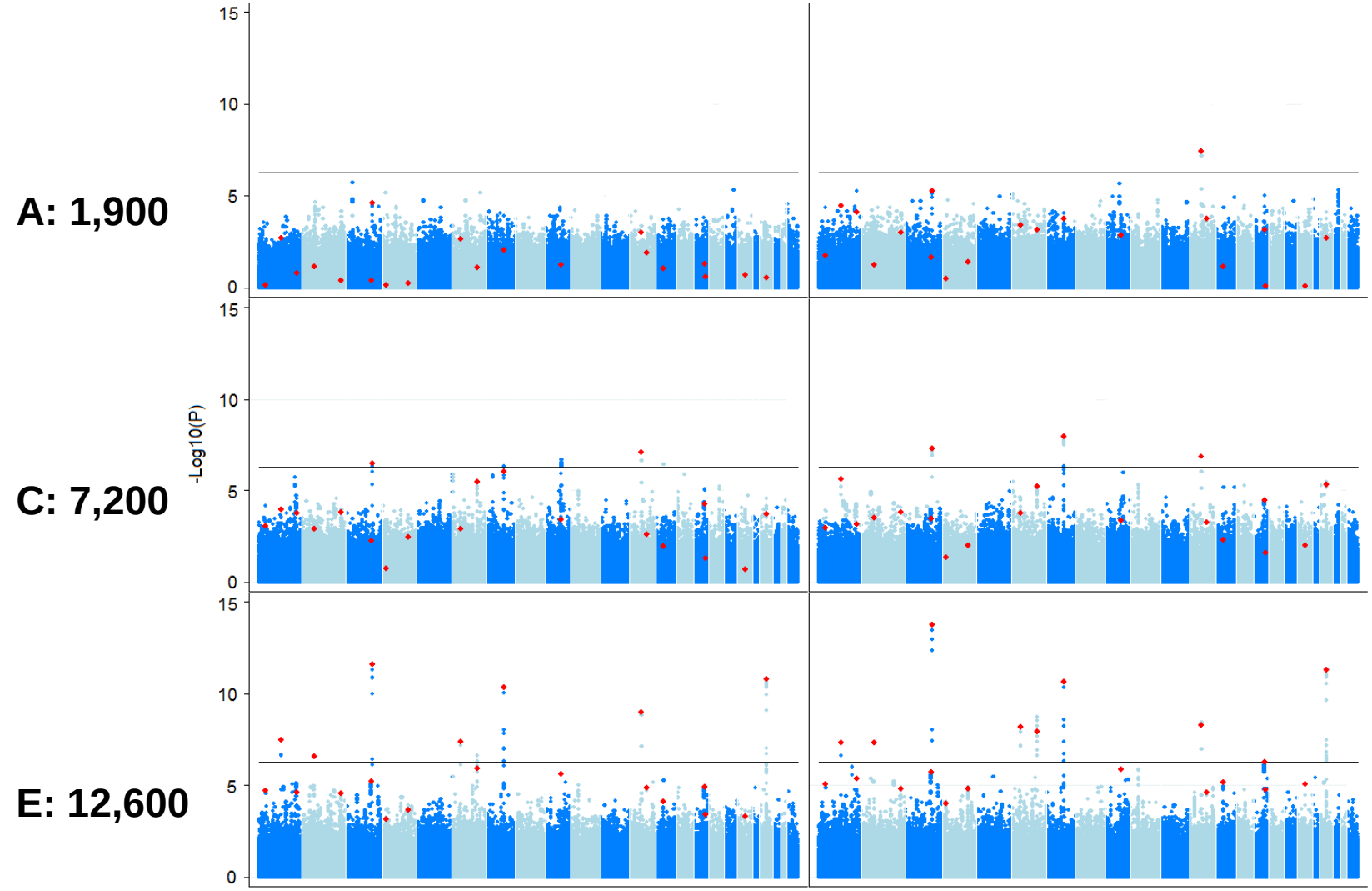
C- 6788 Cases (WTCCC HT)

D- 8668 Cases (WTCCC CAD)

E- 12228 Cases (EPIC)

F- 13665 Cases (WTCCC UKBS)

Weedon et al. (2008) *Nat Genet*



Weedon et al. (2008) *Nat Genet*

- Some real hits sit in the bottom of the distribution
- Some hits initially look interesting but then go away

Statistical Methods for the Analysis of Genome-wide Association Studies

HENRY STEWART TALKS

"This is an outstanding collection. Alongside journals and books no self-respecting library in institutions hosting research in Biomedicine and the Life Sciences should be without access to these talks."

PROFESSOR ROGER KORNBERG, NOBEL LAUREATE
STANFORD UNIVERSITY SCHOOL OF MEDICINE

Go to: http://www.hstalks.com