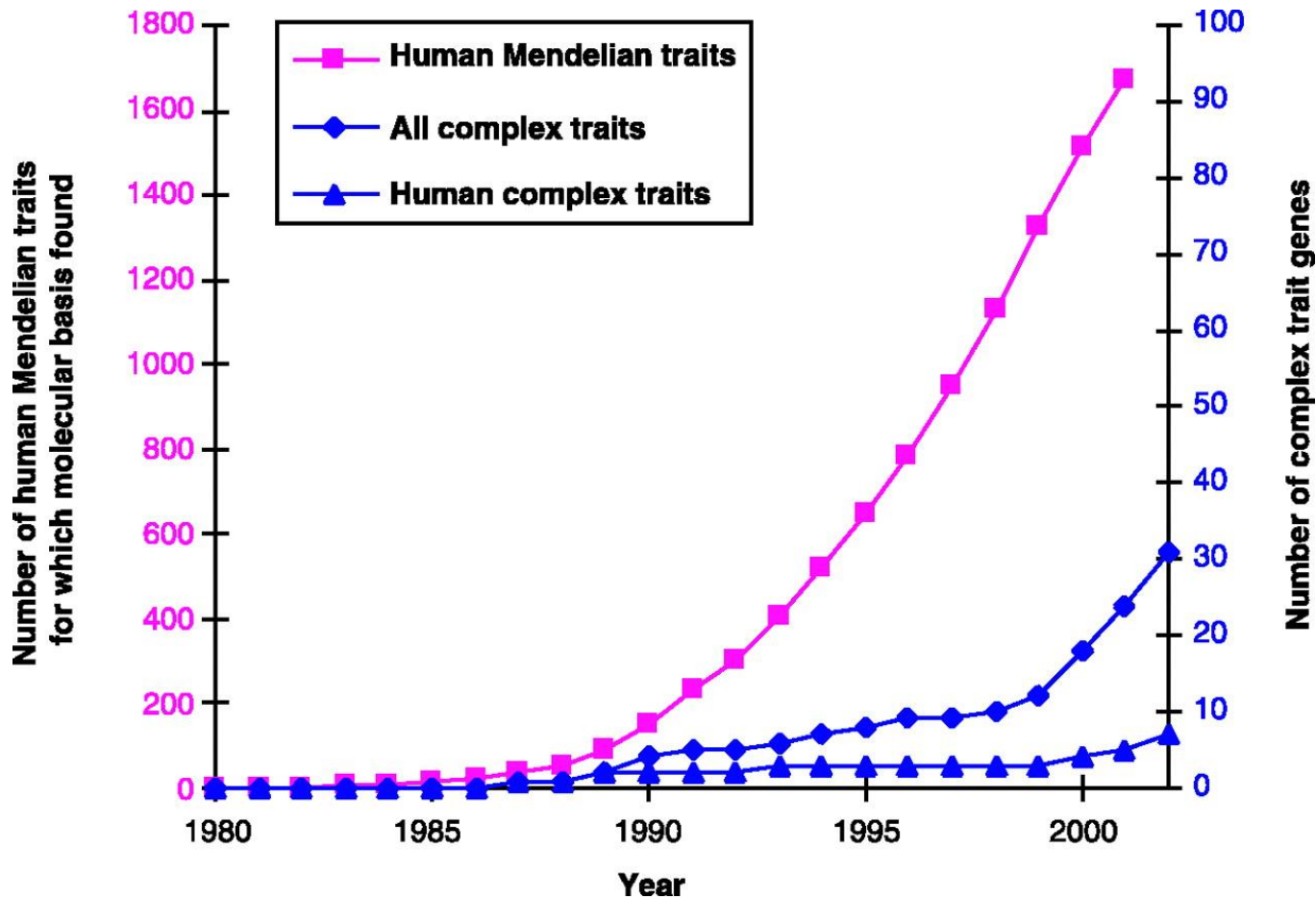


Genome-wide Association

David Evans

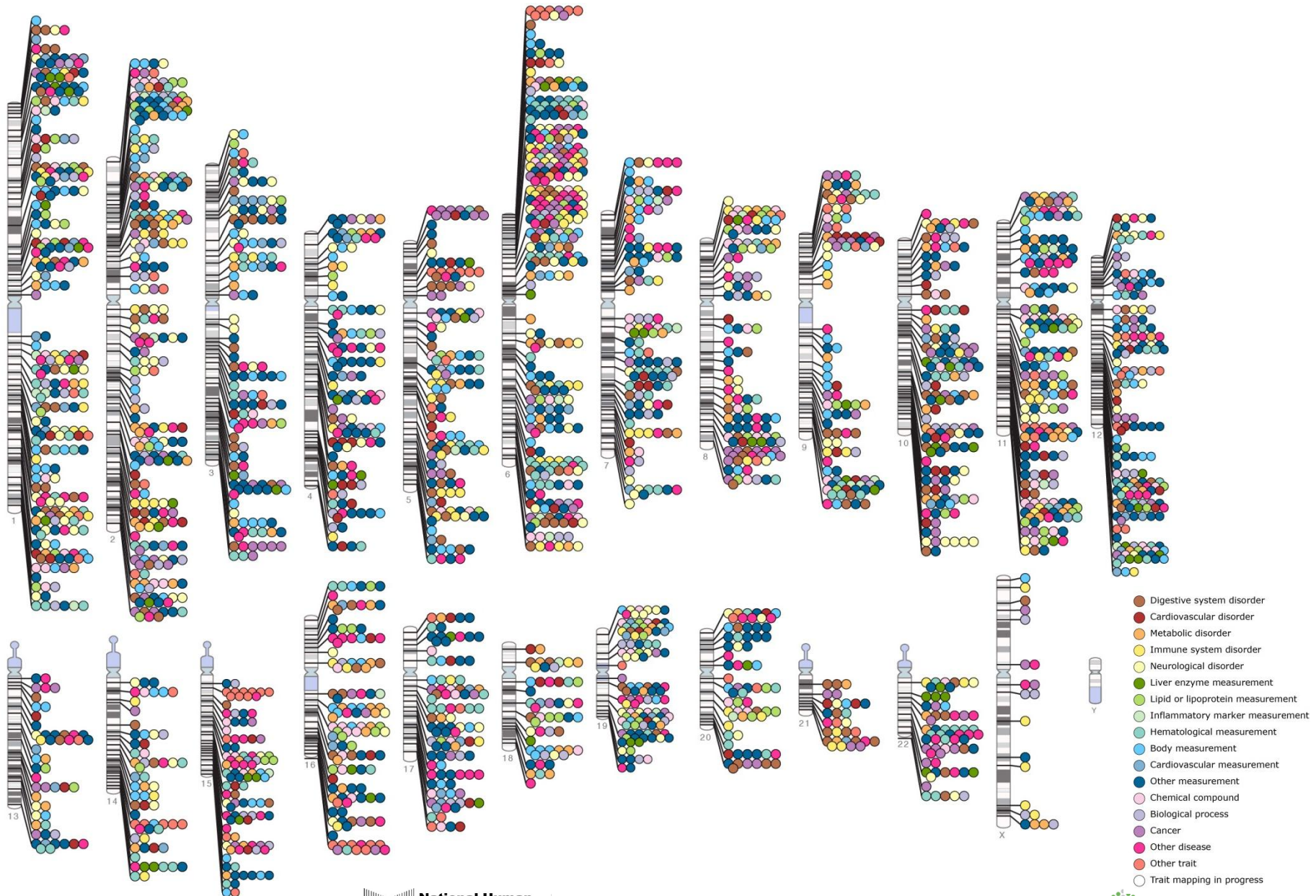


Historical gene mapping



Glazier et al, *Science* (2002).

Published Genome-Wide Associations



The Majority of Heritability for Most Diseases is Yet to Be Explained

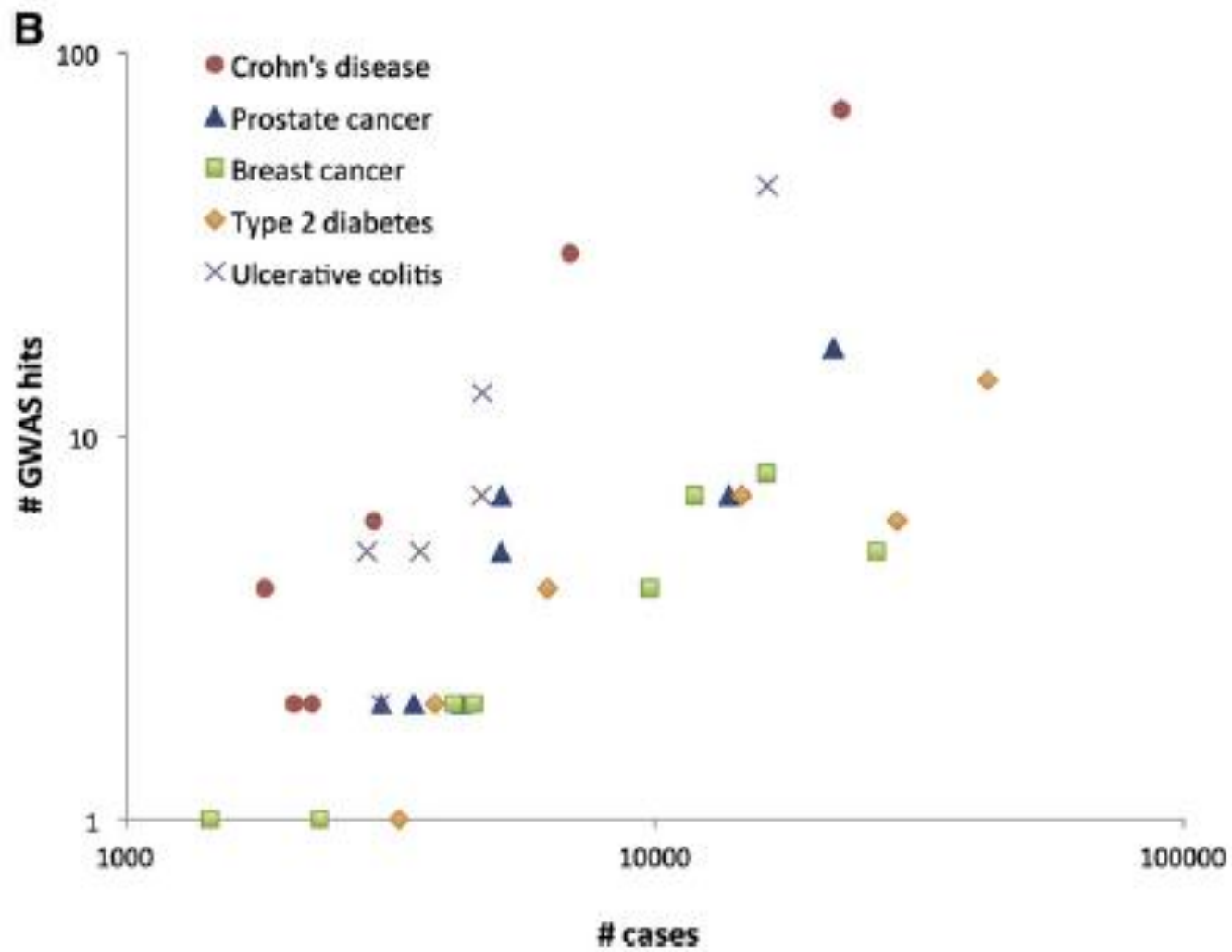
NEWS FEATURE PERSONAL GENOMES

NATURE | Vol 456 | 6 November 2008



The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.



What Has Been The Point?

- Understanding the underlying biology of disease
- The identification of drug targets and the development of new drugs/drug repositioning
- Understanding the basis of individual differences
- Genetic risk prediction?
- Instruments to understand observational epidemiological associations

REVIEW

Five Years of GWAS Discovery

Peter M. Visscher,^{1,2,*} Matthew A. Brown,¹ Mark I. McCarthy,^{3,4} and Jian Yang⁵

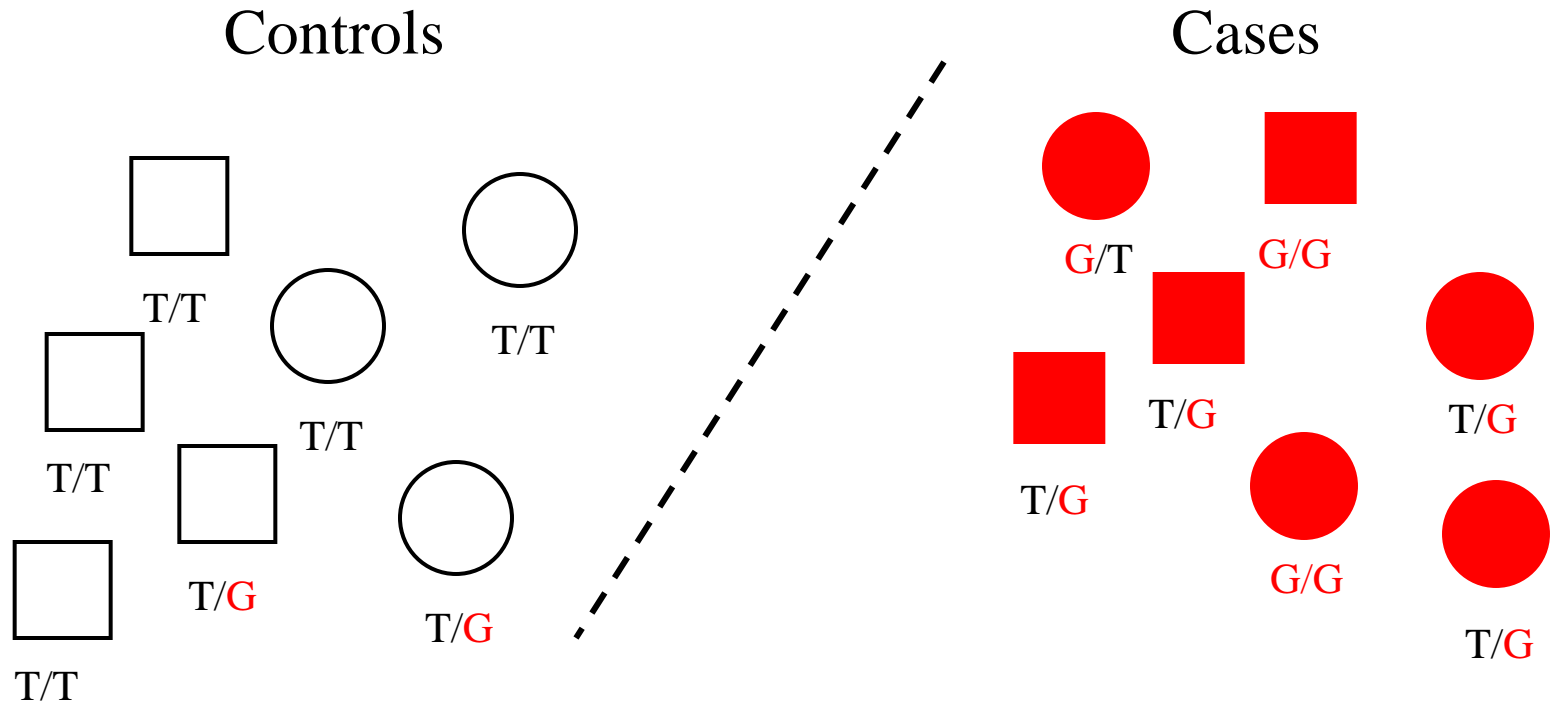
The American Journal of Human Genetics 90, 7–24, January 13, 2012

This Session

- Tests of association in unrelated individuals
- Population Stratification
- Population Stratification Practical
- Assessing significance in genome-wide association
- Replication
- Characterization

Tests of Association in Unrelated Individuals

Genetic Case Control Study



Allele **G** is 'associated' with disease

Allele-based tests

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

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

where

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

- X^2 has χ^2 distribution with 1 degrees of freedom under null hypothesis.
- Armitage trend test preferred (i.e. GG = 0; GT = 1; TT = 2)

	Cases	Controls	Total
G	n_{1A}	n_{1U}	$n_{1.}$
T	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

$$OR = \frac{n_{1A}n_{0U}}{n_{1U}n_{0A}}$$

Genotypic tests

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

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

where

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

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

	Cases	Controls	Total
GG	n_{2A}	n_{2U}	$n_{2.}$
GT	n_{1A}	n_{1U}	$n_{1.}$
TT	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

Dominance Model

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

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

where

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

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

	Cases	Controls	Total
GG/GT	n_{1A}	n_{1U}	$n_{1.}$
TT	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

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

$$\ln(\pi_i / (1 - \pi_i)) = \beta_0 + \beta_M Z_{(M)i} + \beta_{MM} Z_{(MM)i}$$

where

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

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

Indicator variables

- Represent genotypes of each individual by indicator variables:

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

Likelihood calculations

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

$$\ell(\mathbf{y}|\mathbf{G}, \boldsymbol{\beta}) = \sum_i y_i \ln[\pi_i] + (1 - y_i) \ln[1 - \pi_i]$$

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 $\ell(\mathbf{y}|\mathbf{G}, \hat{\boldsymbol{\beta}})$
- 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	$2\left[\ell(\mathbf{y} \mathbf{G}, \hat{\beta}_M, \hat{\beta}_0) - \ell(\mathbf{y} \mathbf{G}, \hat{\beta}_0)\right]$	1
Genotype vs null	$2\left[\ell(\mathbf{y} \mathbf{G}, \hat{\beta}_{MM}, \hat{\beta}_{Mm}, \hat{\beta}_0) - \ell(\mathbf{y} \mathbf{G}, \hat{\beta}_0)\right]$	2

Covariates

- It is straightforward to incorporate covariates in the logistic regression model:
 - age, gender, and other environmental risk factors.
 - Need to be careful
- 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 j th covariate, and γ_j is the corresponding covariate regression coefficient.

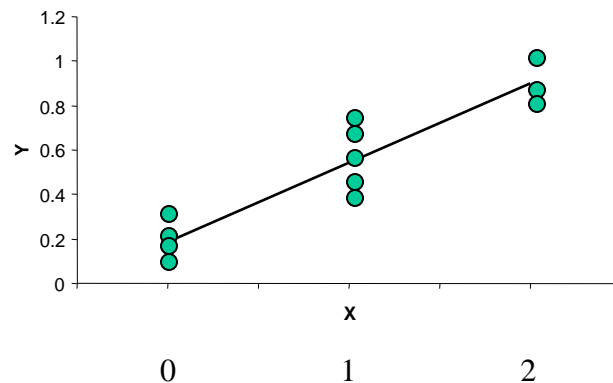
Simple Additive Regression Model of Association (Unrelated individuals)

$$Y_i = \alpha + \beta X_i + e_i$$

where

Y_i = trait value for individual i

X_i = number of 'A' alleles an individual has



Association test is whether $\beta > 0$

Linear Regression Including Dominance

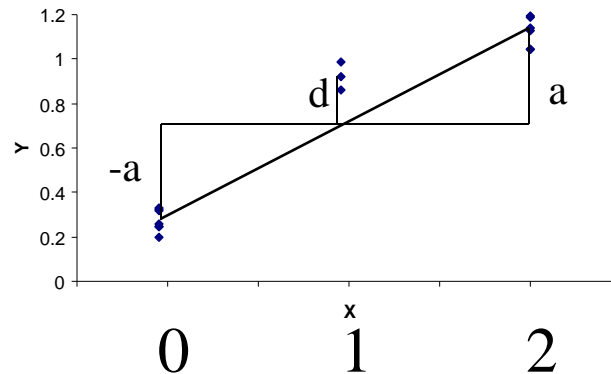
$$Y_i = \alpha + \beta_x X_i + \beta_z Z_i + e_i$$

where

$Y_i =$ trait value for individual i

$X_i =$ 1 if individual i has genotype 'AA'
0 if individual i has genotype 'Aa'
-1 if individual i has genotype 'aa'

$Z_i =$ 0 for 'AA'
1 for 'Aa'
0 for 'aa'



Further extensions

- Can model haplotypes
- Can model imputed genotypes
- Can model interactions

Population Stratification

DEFINITIONS: STRATIFICATION AND ADMIXTURE

1. Stratification / Sub-structure

Refers to the situation where a sample of individuals consists of several discrete subgroups which do not interbreed as a single randomly mating unit

2. Admixture

Implies that subgroups also interbreed. Therefore individuals may be a mixture of different ancestries.

My Samples

Sample 1 Americans

$$\chi^2=0$$

$$p=1$$

Use of Chopsticks

A	Yes	No	Total
A ₁	320	320	640
A ₂	80	80	160
Total	400	400	800

My Samples

Sample 2 Chinese

$$\chi^2=0$$

$$p=1$$

Use of Chopsticks

A	Yes	No	Total
A ₁	320	20	340
A ₂	320	20	340
Total	640	40	680

My Samples

Sample 3 Americans + Chinese

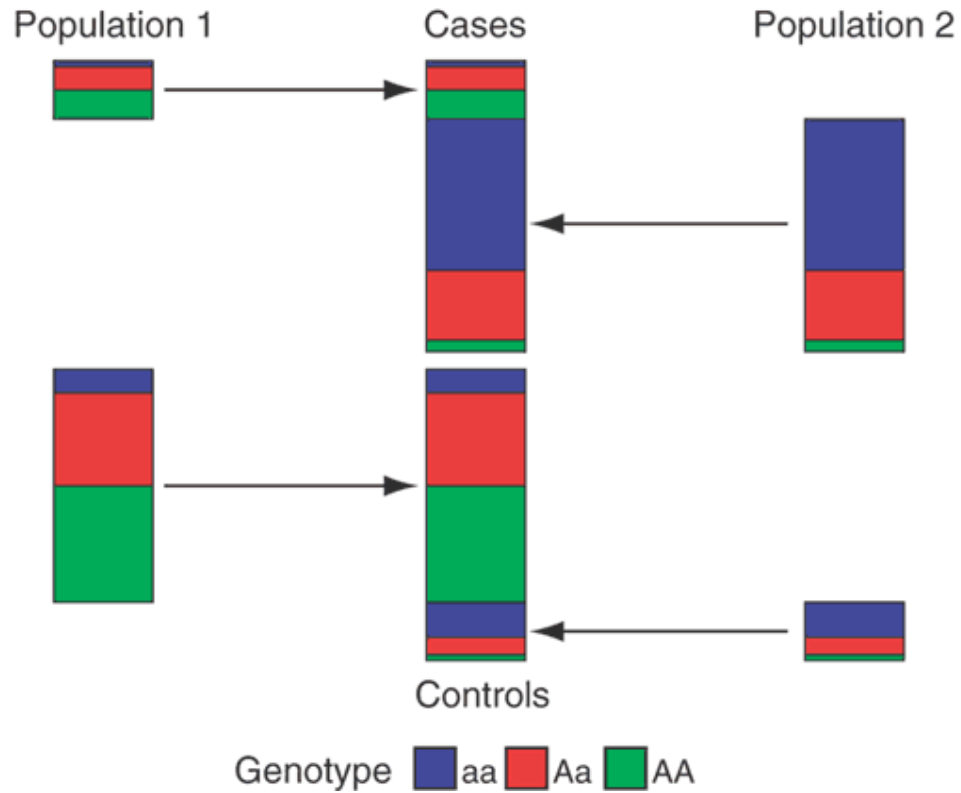
$$\chi^2=34.2$$

$$p=4.9 \times 10^{-9}$$

Use of Chopsticks

A	Yes	No	Total
A ₁	640	340	980
A ₂	400	100	500
Total	1040	440	1480

Population structure



Marchini, *Nat Genet* (2004)

ADMIXTURE: (DIABETES IN AMERICAN INDIANS)

Full heritage American Indian Population

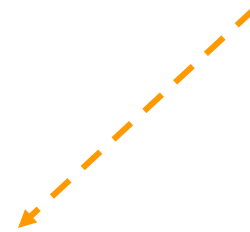
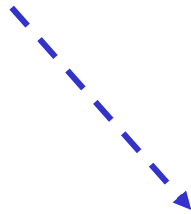
Gm^{3;5,13,14} + -
 ~1% ~99%

(NIDDM Prevalence ≈ 40%)

Caucasian Population

Gm^{3;5,13,14} + -
 ~66% ~34%

(NIDDM Prevalence ≈ 15%)



Study without knowledge of genetic background:

Gm ^{3;5,13,14} haplotype	Cases	Controls
+	7.8%	29.0%
-	92.2%	71.0%

OR=0.27

95%CI = 0.18 - 0.40

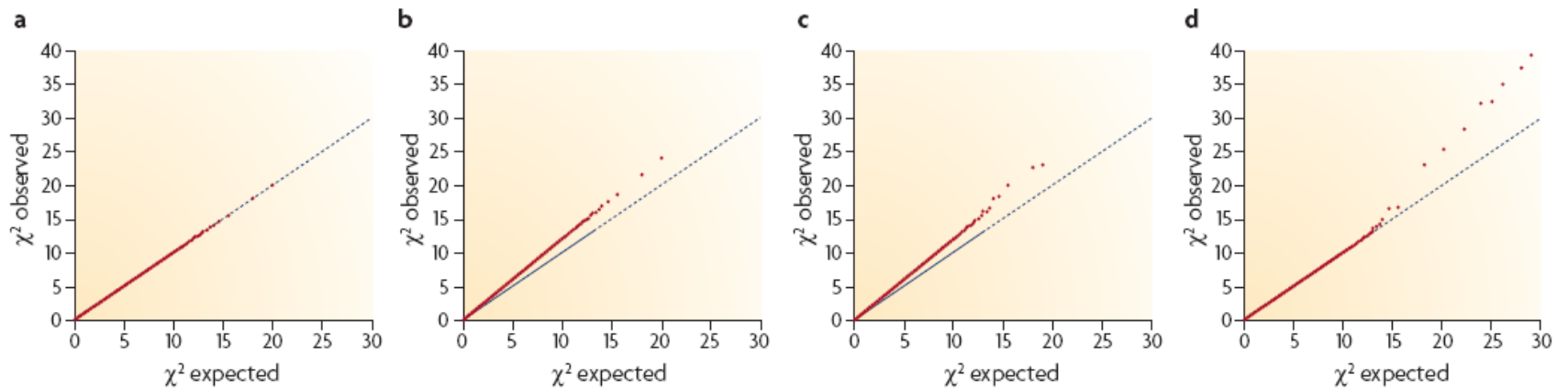
ADMIXTURE: (DIABETES IN AMERICAN INDIANS)

<i>Index of Indian Heritage</i>	Gm^{3;5,13,14}	
	+	-
0	17.8%	19.9%
4	28.3%	28.8%
8	35.9%	39.3%

Gm haplotype serves as a marker for Caucasian admixture

QQ plots

Box 2 | Visualization of genome-wide association data



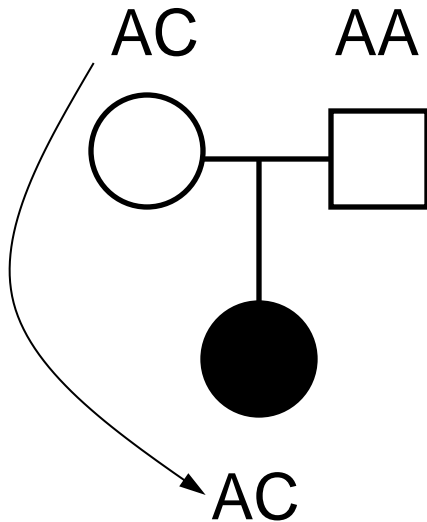
McCarthy et al. (2008) *Nature Genetics*

Solutions

- Family-based Analysis
- Stratified Analysis
 - Analyze Chinese and American samples separately then combine statistically
- Model the confounder
 - Include a term for Chinese or American ancestry in a logistic regression model
 - Principal Components
- Genomic Control
- Linear Mixed Models

Family based Tests of Association

Transmission Disequilibrium Test



- Rationale: Related individuals have to be from the same population
- Compare number of times heterozygous parents transmit “A” vs “C” allele to affected offspring
- Many variations

TDT

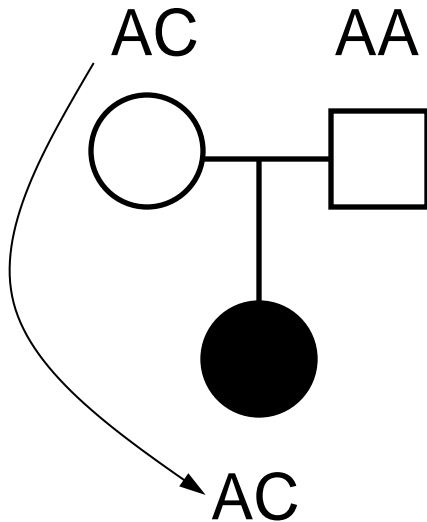
Table 2

Combinations of Transmitted and Nontransmitted Marker Alleles M₁ and M₂ among 2n Parents of n Affected Children

TRANSMITTED ALLELE	NONTRANSMITTED ALLELE		TOTAL
	M ₁	M ₂	
M ₁	<i>a</i>	<i>b</i>	<i>a+b</i>
M ₂	<u><i>c</i></u>	<u><i>d</i></u>	<u><i>c+d</i></u>
Total	<i>a+c</i>	<i>b+d</i>	<i>2n</i>

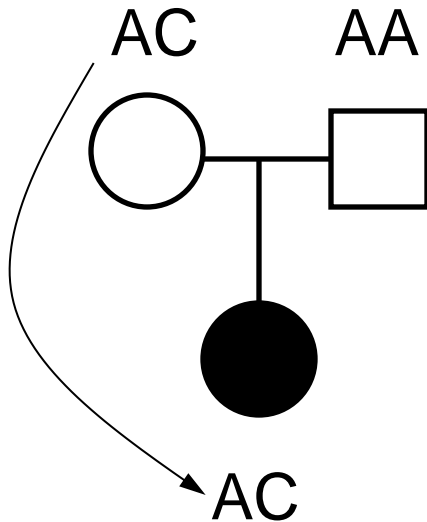
$$\chi^2 = (b-c)^2 / (b+c).$$

TDT Advantages



- Robust to stratification
- Identification of Mendelian Inconsistencies
- Parent of Origin Effects
- More accurate haplotyping

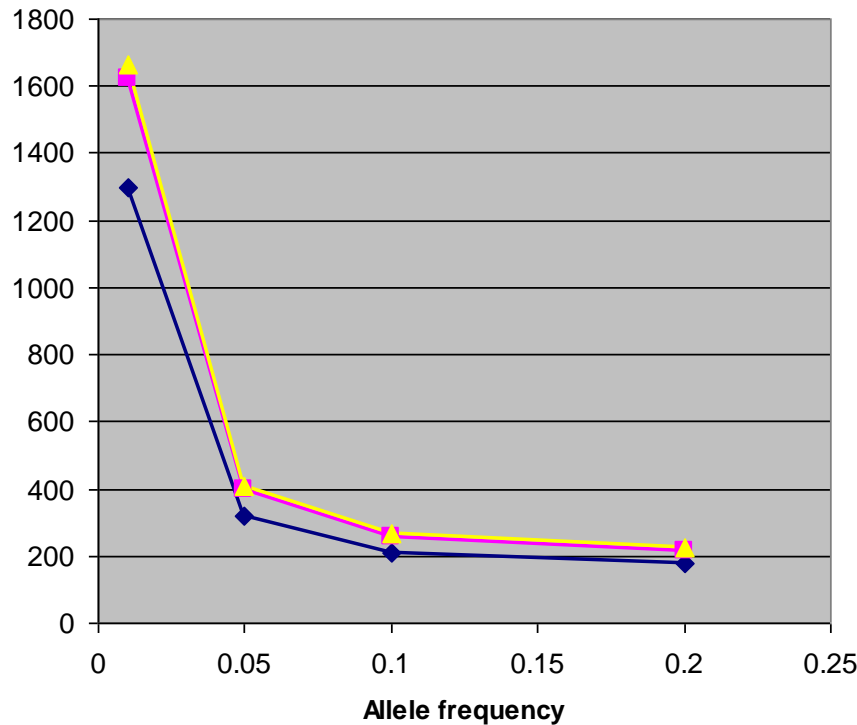
TDT Disadvantages



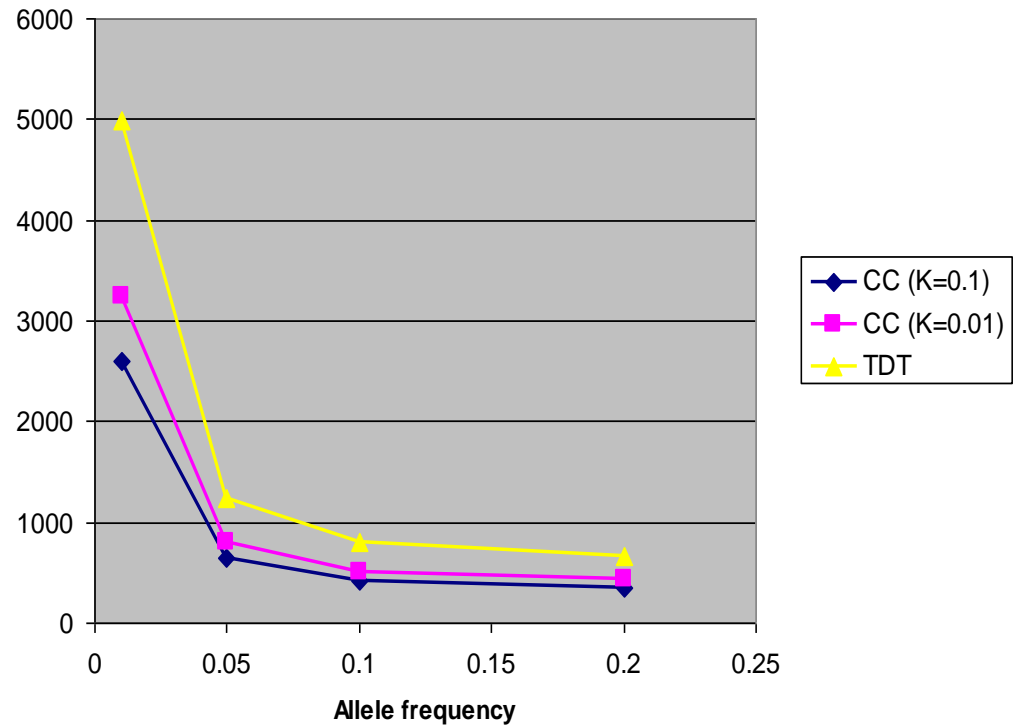
- Difficult to gather families
- Difficult to get parents for late onset / psychiatric conditions
- Genotyping error produces bias
- Inefficient for genotyping (particularly GWA)

Case-control versus TDT

N units for 90% power

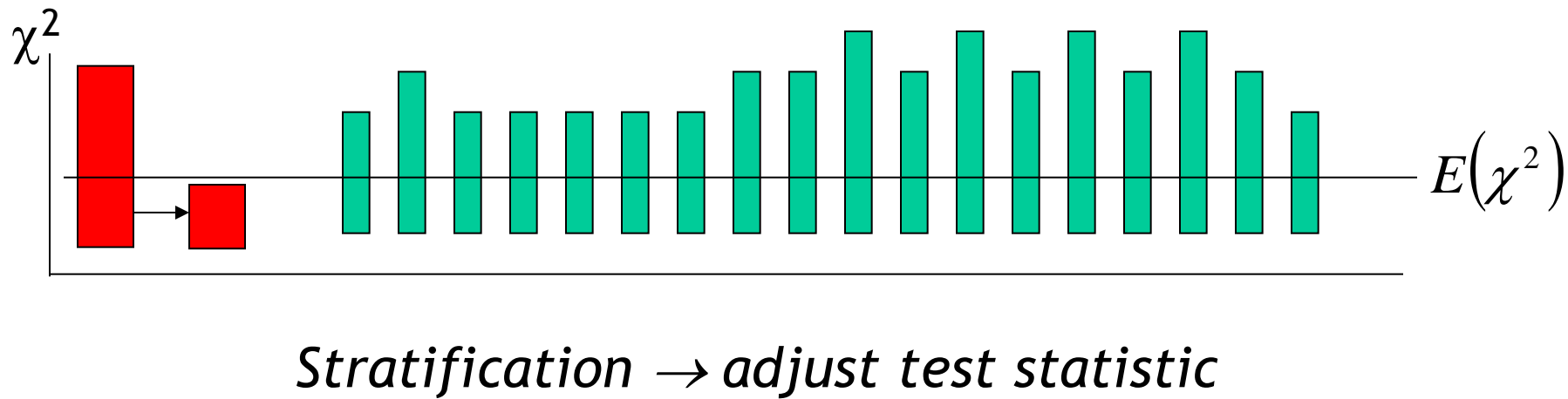
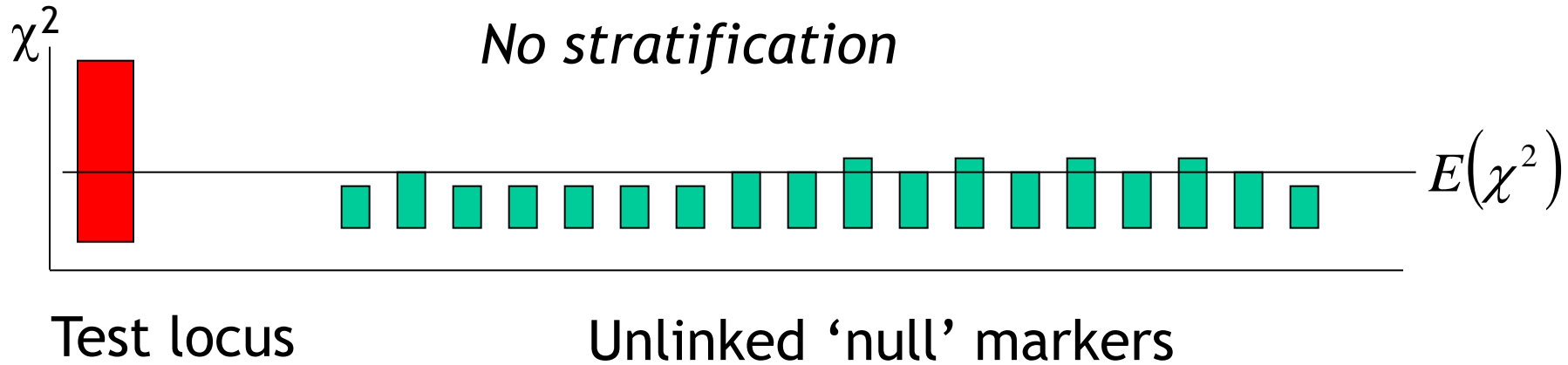


N individuals for 90% power



$p = 0.1; RAA = RAa = 2$

Genomic control



Genomic control

- ▶ “ λ ” is Genome-wide inflation factor

$$\hat{\lambda} = \text{median}\{\chi_1^2, \chi_2^2, \dots, \chi_N^2\} / 0.456$$

- ▶ Test statistic is distributed under the null:

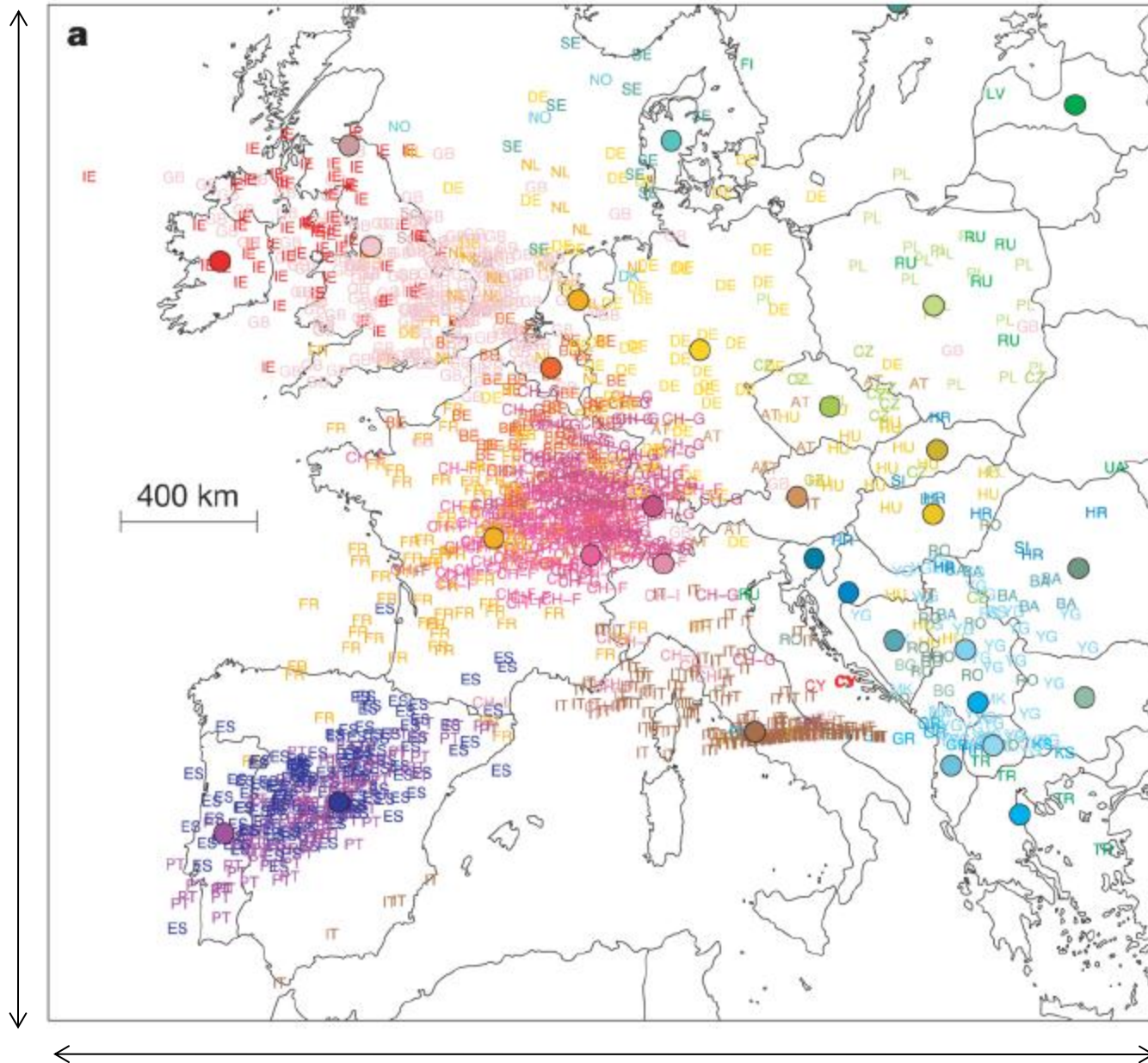
$$T_N / \lambda \sim \chi^2_1$$

- ▶ Problems...

Principal Components Analysis

- Principal Components Analysis is applied to genotype data to infer continuous axes of genetic variation
- Each axis explains as much of the genetic variance in the data as possible with the constraint that each component is orthogonal to the preceding components
- The top principal Components tend to describe population ancestry
- Include principal components in regression analysis => correct for the effects of stratification
- EIGENSTRAT, SHELLFISH

Principal Component Two




Principal Component One

Wellcome Trust Case Control Consortium

(Ireland, €1) 70p
Thursday 7 June 2007
www.independent.co.uk
NUMBER 442

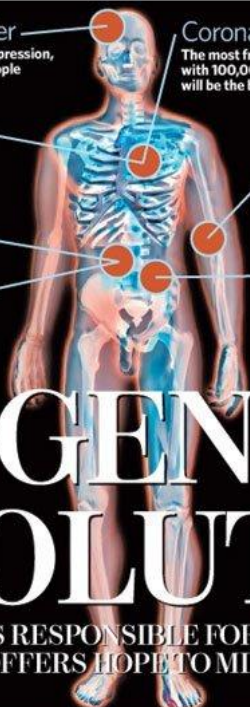
THE INDEPENDENT



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

Population structure - λ

	Disease	
Genomic control - λ genome-wide inflation of median test statistic	1	1.15
	2	1.08
	3	1.09
	4	1.26
	5	1.06
	6	1.07
	7	1.10

Disease collection center

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

Center 3: $\lambda = 1.77$

All others: $\lambda = 1.09$

Multi-dimensional Scaling



- WTCCC
- + Excluded samples
- YRI
- CEU
- CHB+JPT



Linear Mixed Models

- The test of association is performed in the model for the means
- “Relatedness” between individuals (due to both population structure and cryptic relatedness) is captured in the modelling of the covariance between individuals
- Requires genome-wide data
- Often more effective than other approaches
- Variety of software packages (e.g. GEMMA)

Example

LETTER

doi:10.1038/nature10251

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*

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 disability¹. 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 risk⁴. Modestly powered genome-wide association studies (GWAS)⁵⁻¹⁹ have enabled more than 20 additional risk loci to be identified and have shown that multiple variants exerting modest individual effects have a key role in disease susceptibility¹¹. Most of the genetic architecture underlying susceptibility to the disease remains to be defined and is anticipated to require the analysis of sample sizes that are beyond the numbers currently available to individual research groups. In a collaborative GWAS involving 9,772 cases of European descent collected by 23 research groups working in 15 different countries, we have replicated almost all of the previously suggested associations and identified at least a further 29 novel susceptibility loci. Within the MHC we have refined the identity of the *HLA-DRB1* risk alleles and confirmed that variation in the *HLA-A* gene underlies the independent protective effect attributable to the class I region. Immunologically relevant genes are significantly overrepresented among those mapping close to the identified loci and particularly implicate T-helper-cell differentiation in the pathogenesis of multiple sclerosis.

We performed a large GWAS as part of the Wellcome Trust Case Control Consortium 2 (WTCCC2) project. Cases were recruited through the International Multiple Sclerosis Genetics Consortium (IMSGC) and compared with the WTCCC2 common control set^{2,13} supplemented by data from the control arms of existing GWAS. We introduced a number of novel quality control methods for processing these data sets (see Supplementary Information), which ultimately provided reliable information from 9,772 cases and 17,376 controls (Fig. 1a). After single nucleotide polymorphism (SNP)-based quality controls, data from 465,434 autosomal SNPs, common to all internally and externally generated data sets, were available for analysis.

The multi-population nature of our study (Fig. 1a, b) afforded an opportunity to assess various published approaches for controlling the potential confounding effects of population structure, several of which (in the event) proved unhelpful (see Supplementary Information). Although not common in primary GWAS undertaken to date, the challenge of combining data across populations, in contexts where not all case samples have controls available from the same population (thus precluding standard meta-analytical techniques), may become more routine as study sizes increase.

We attempted analyses of the non-United Kingdom (UK) data with the now widespread technique of using principal components as covariates to correct for structure. However, even use of all seven top principal components that captured genome-wide effects in our data

resulted in an unacceptably high genomic inflation: for example, the genomic control factor²⁰ (λ) was $\lambda = 1.2$. We tried to reduce the genomic inflation by discarding the case samples that seemed least well matched to control sets. Removal of half the available cases in this fashion only reduced λ to 1.1. In another approach to handling structure, statistical clustering algorithms were successful in identifying subgroups of the data within which cases and controls seemed well matched for ancestry (see Supplementary Fig. 17). However, tests within these subgroups combined via fixed-effects meta-analysis also yielded unacceptably high genomic inflation ($\lambda > 1.4$) in an analysis with seven matched subgroups of cases and controls. Lastly, we applied a novel variance components method (similar to one described previously²¹), separately to the UK and non-UK data sets, which explicitly accounts for correlations among the

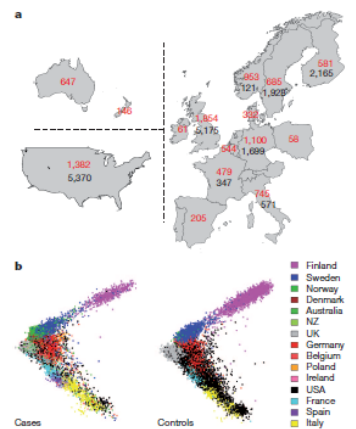
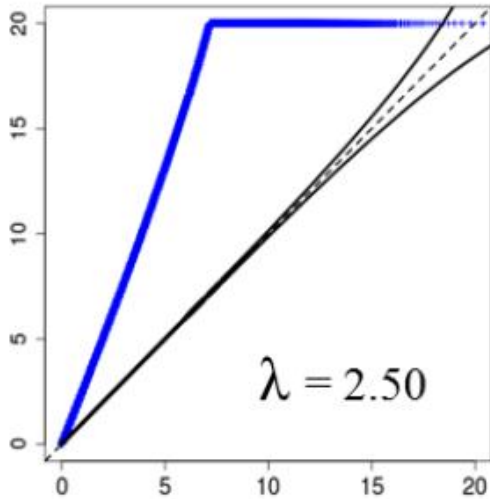


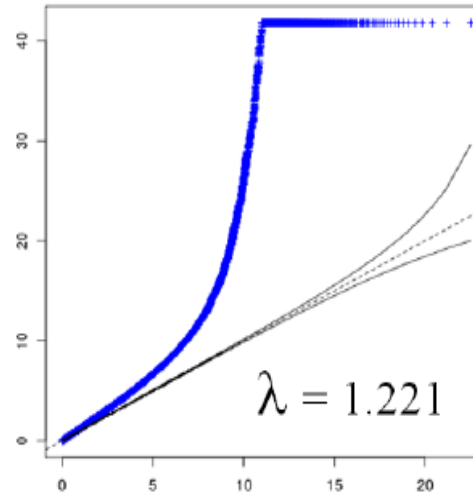
Figure 1 | Distribution of cases and controls. a, All cases and controls were drawn from populations with European ancestry: cases from 15 countries and controls from 8. a, Numbers of case (red) and control (black) samples from each country. b, The projection of samples onto the first two principal components of genetic variation, with cases shown on the left and controls on the right. The axes are orientated to approximate the geography, and samples are colour coded as indicated in the legend. NZ, New Zealand. We genotyped the cases (9,772) and some Swedish controls (527) using the Illumina Human 660-Quad platform, and the UK controls (5,175, the WTCCC2 common control set^{2,13}) using the Illumina 1.2M platform. All other controls were genotyped externally using various Illumina genotyping systems (see Supplementary Information).

*A list of authors and their affiliations appears at the end of the paper; membership of both consortia is listed in Supplementary Information.

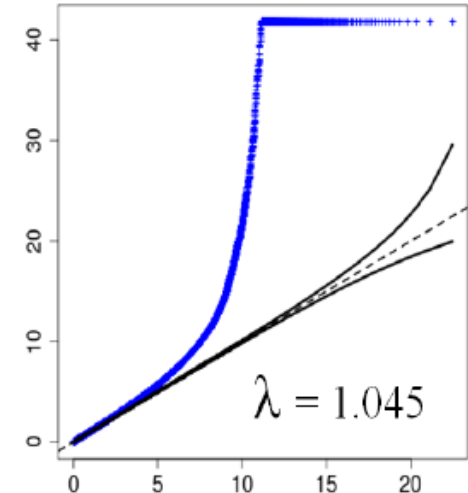
Comparison of Approaches in Sawcer et al.



No correction



PCA correction
(top 100 PCs)



Mixed-model
correction

Practical

Assessing
“Significance” in
Genome-wide
Association Studies

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

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 $\alpha_{\text{FWER}} = .05$
- Bonferroni correction for m tests
 - Set significance level to $\alpha = .05/m$
- “Effective number of statistical tests
- “Genome-wide Significance” suggested at around $\alpha = 5 \times 10^{-8}$ for European populations

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

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 be 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

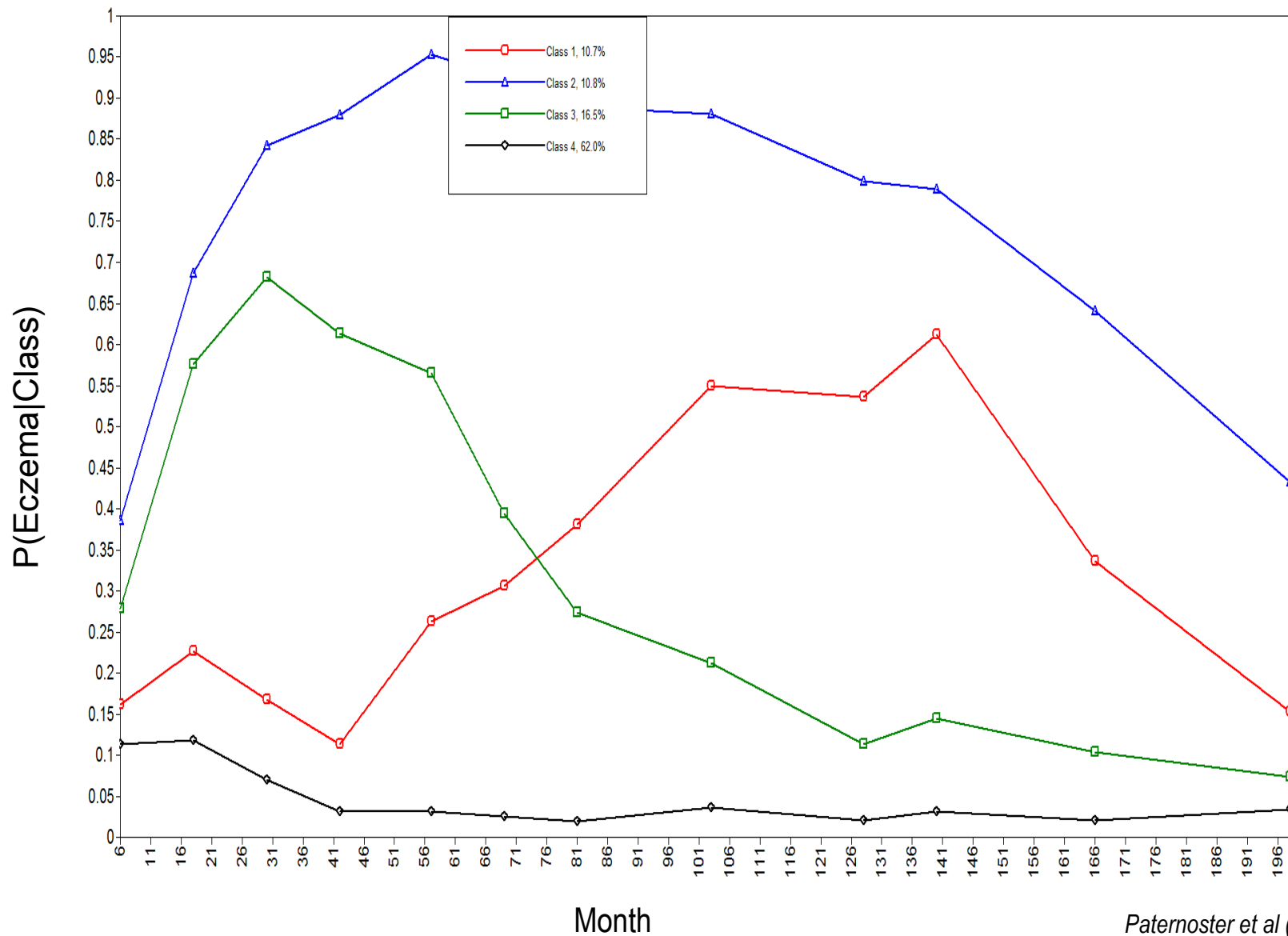
Well designed negative studies are valuable

Characterization

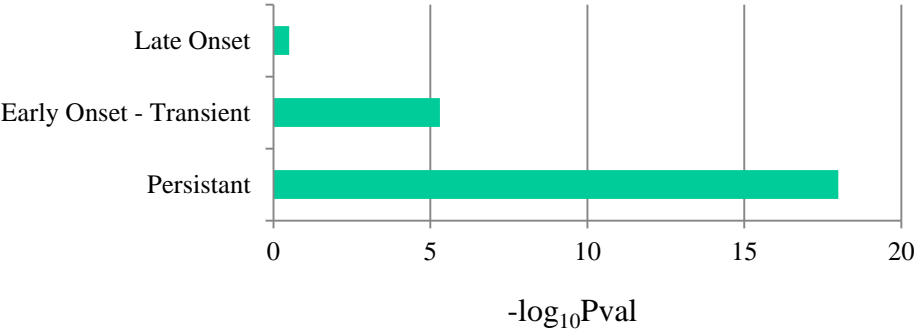
Characterization

- Functional assays
 - Gene expression
 - Mouse/animal models
 - In vitro models
- Conditional analyses and fine mapping
- Phenotypic refinement

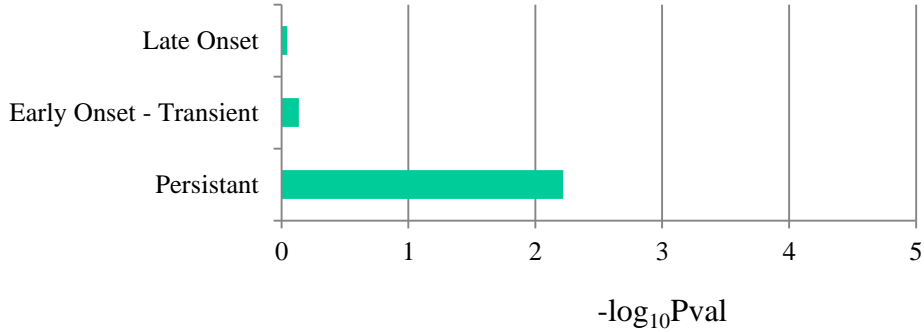
Eczema (Latent Class Analysis)



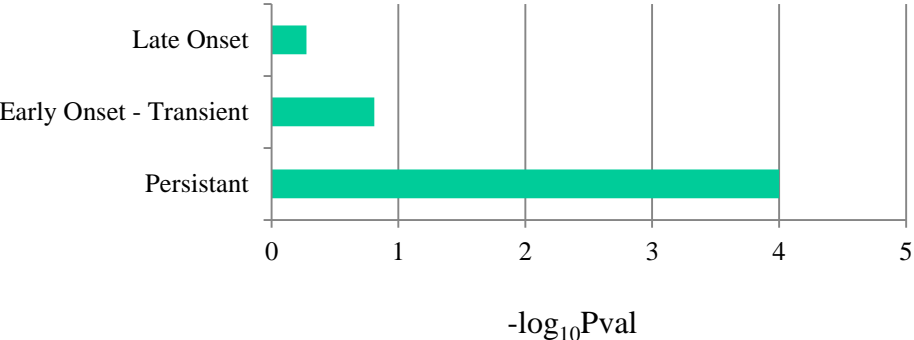
FLG



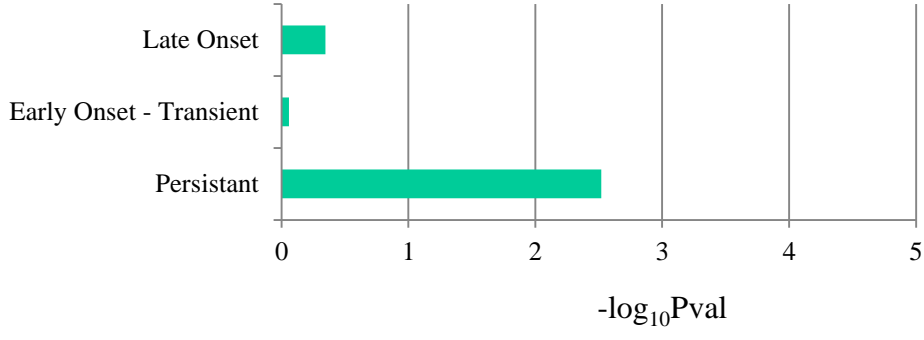
KIF3A



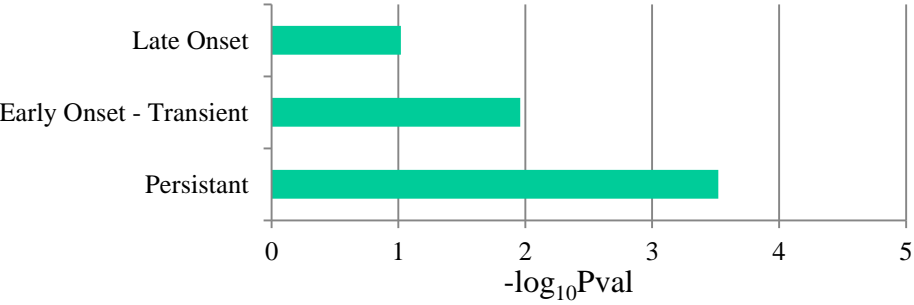
OVOL1



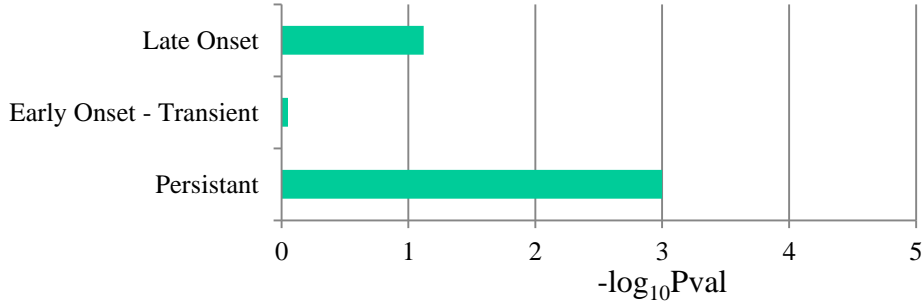
C11orf30



ACTL9



RTEL1



Useful References

Useful References

- ▶ [Wellcome Trust Case Control Consortium \(2007\). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447, 661-678](#)
- ▶ [NCI-NHGRI Working Group on Replication in Association Studies \(2007\). Replicating genotype-phenotype associations. Nature, 447, 655-660](#)
- ▶ [Purcell, Neale, ..., Ferreira, ..., Sham \(2007\). PLINK: a tool set for whole genome association and population based linkage analysis. Am J Hum Genet, 81, 559-75](#)
- ▶ [McCarthy MI et al. \(2008\). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Genetics, 9, 356-369](#)