Association Mapping

David Evans

University of BRISTOL

Outline

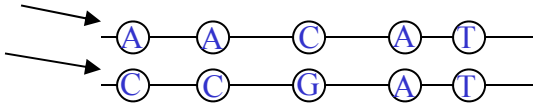
- Definitions / Terminology
- What is (genetic) association?
- How do we test for association?
- When to use association
- HapMap, 1000 genomes and tagging
- Genome-wide Association
- Sequencing and Rare variants

Definitions

Locus: *Location* on the genome

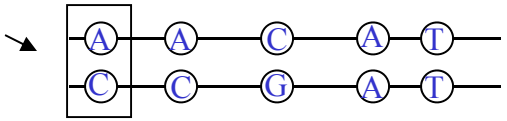
SNP: "Single Nucleotide Polymorphism" a mutation that produces a single base pair change in the DNA sequence

alleles



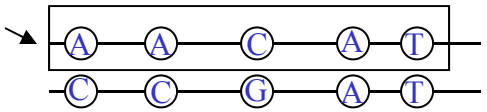
alternate forms of a SNP (mutation)

genotypes



both alleles at a locus form a genotype

haplotypes



the pattern of alleles on a chromosome

QTL: "Quantitative trait locus" a region of the genome that changes the mean value of a quantitative phenotype

What is (genetic) association?

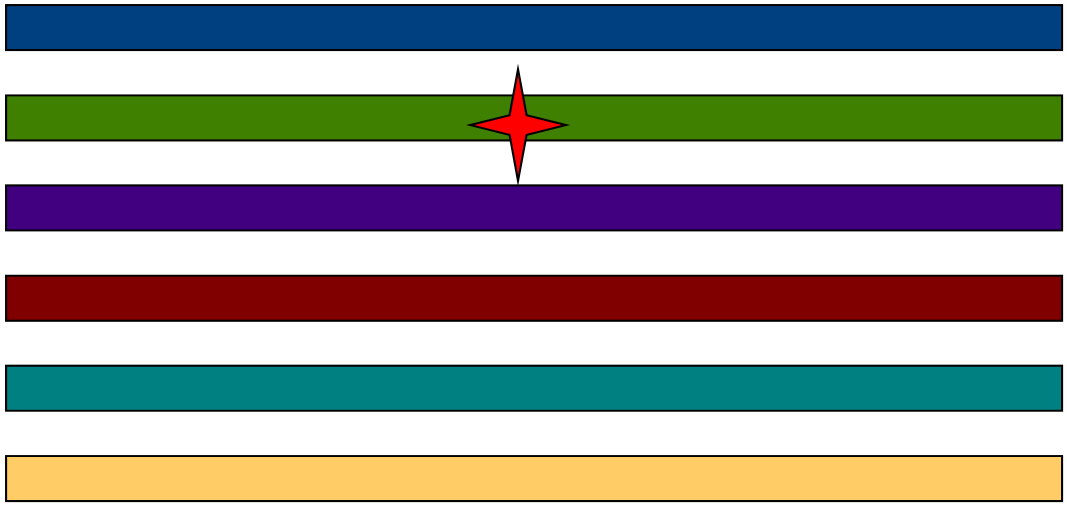
Correlation between an allele/genotype/haplotype and a trait of interest (disease or quantitative phenotype)

Genetic Association
Three Common Forms

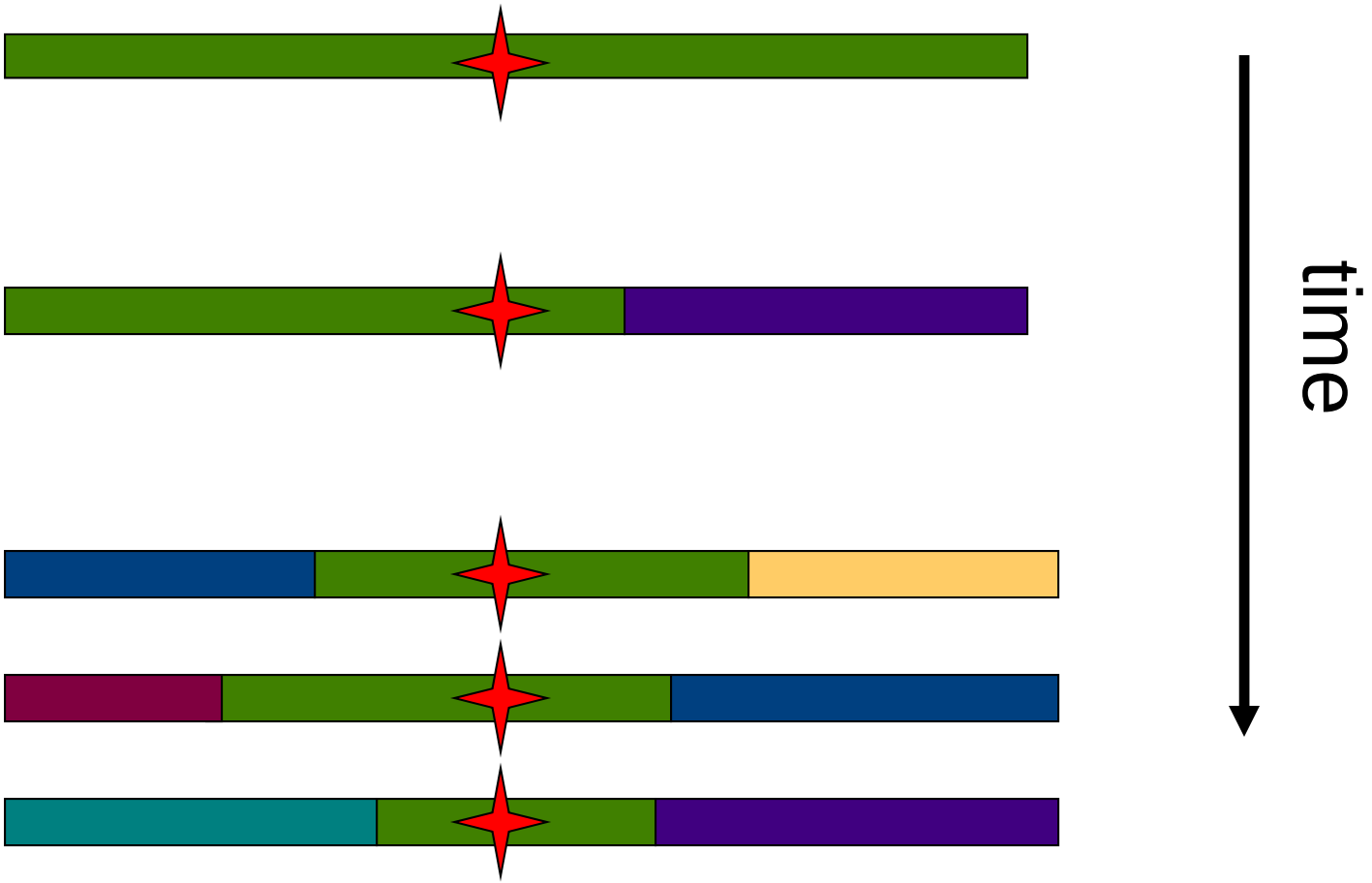
- Direct Association
  - •Polymorphism of interest is functionally involved in the phenotype
  - •~70% of Cystic Fibrosis patients have a deletion of 3 base pairs resulting in the loss of a phenylalanine amino acid at position 508 of the *CFTR* gene

- Indirect Association
  - • Genotyped polymorphism is not involved, but a nearby correlated variant changes the phenotype (linkage disequilibrium)

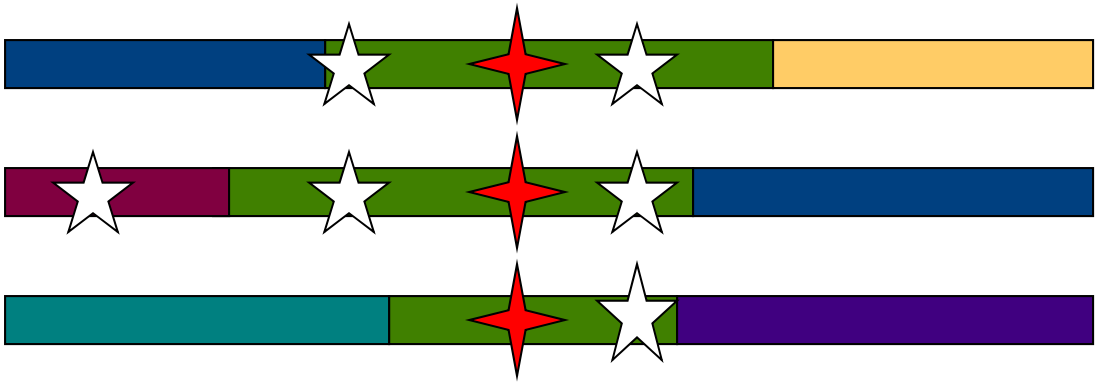
Indirect association and Linkage disequilibrium



Indirect association and Linkage disequilibrium



Linkage Disequilibrium



Variants in close proximity to each other tend to be correlated (i.e. in linkage disequilibrium)

Linkage disequilibrium means that we don’t need to genotype the exact aetiological variant, but only a variant that is correlated with it

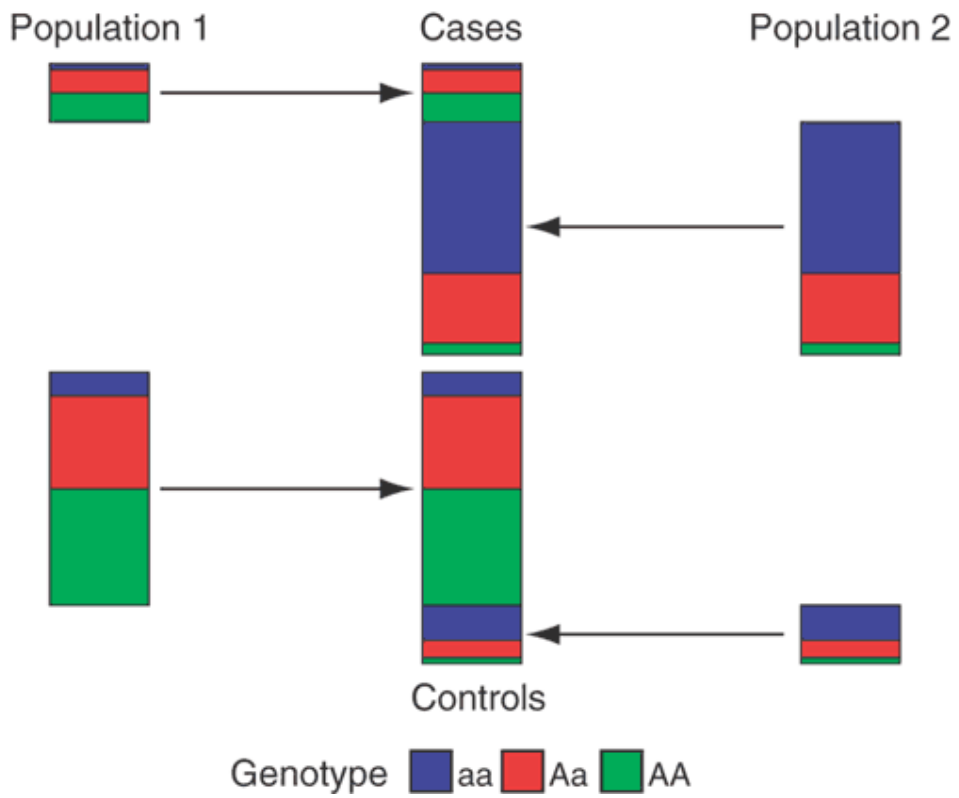
Genetic Association
Three Common Forms

- Direct Association
  - Mutant or ‘susceptible’ polymorphism
  - Allele of interest is itself involved in phenotype

- Indirect Association
  - Allele itself is not involved, but a nearby correlated marker changes phenotype

- Spurious association
  - Apparent association not related to genetic aetiology (e.g. population stratification)

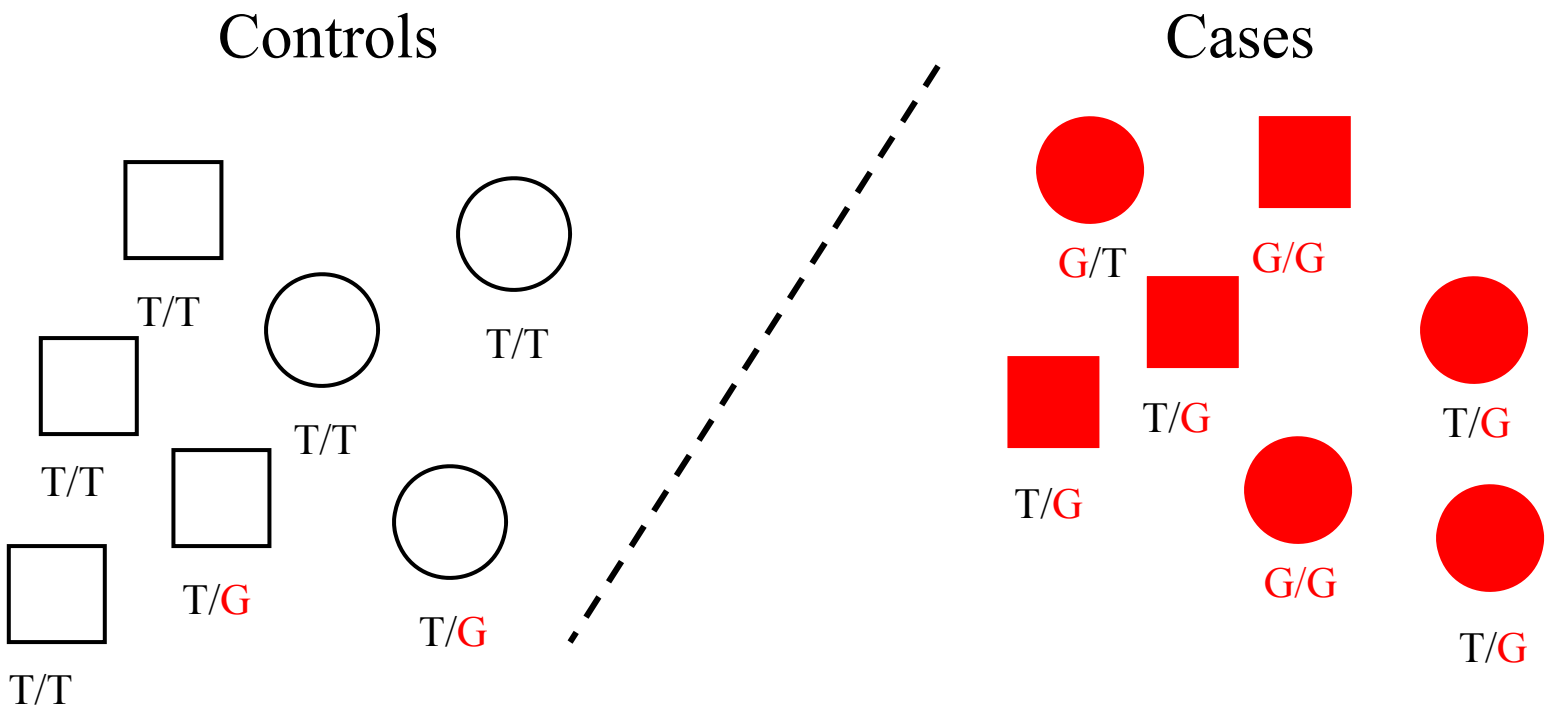
Population Stratification



Marchini, *Nat Genet.* 2004

How do we test for association?

Genetic Case Control Study



Allele G is ‘associated’ with disease

Allele-based tests

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

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

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
G	n_{1A}	n_{1U}	$n_{1\cdot}$
T	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

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
GG	n_{2A}	n_{2U}	$n_{2\cdot}$
GT	n_{1A}	n_{1U}	$n_{1\cdot}$
TT	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

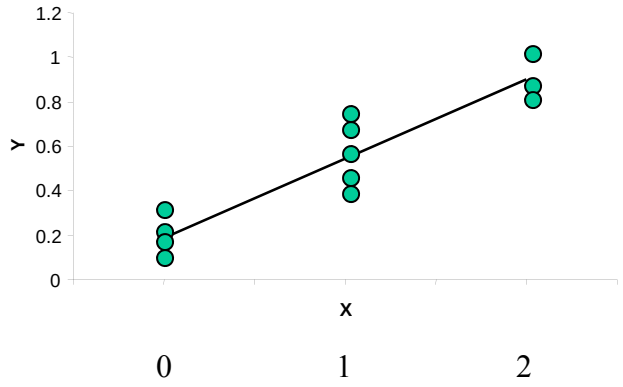
Simple Regression Model of Association (Unrelated individuals)

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

where

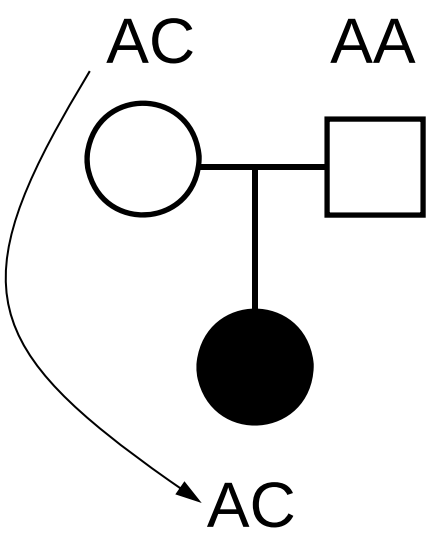
$Y_i =$ trait value for individual i

$X_i =$ number of 'A' alleles an individual has



Association test is whether β is different from 0

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

Transmission Disequilibrium Test

Table 2

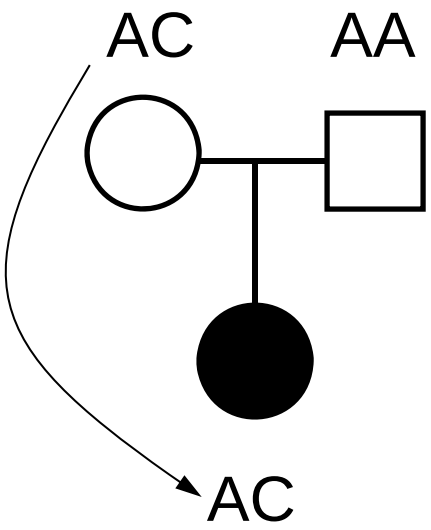
Combinations of Transmitted and Nontransmitted Marker Alleles M_1 and M_2 among $2n$ Parents of n Affected Children

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

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

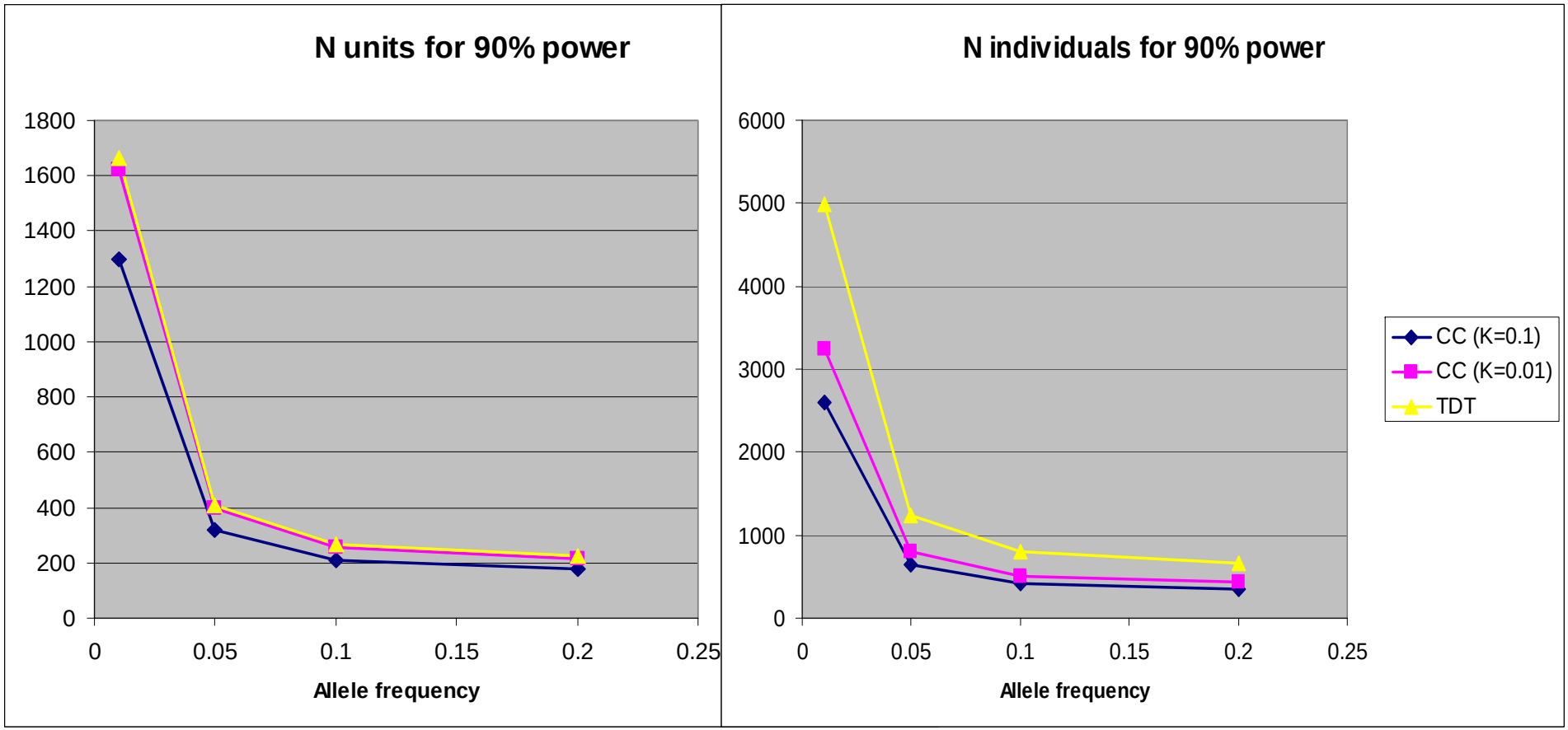
Spielman et al. (1993) AJHG

Transmission Disequilibrium Test



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

Case-control versus TDT



p = 0.1; RAA = RAa = 2

Am. J. Hum. Genet. 64:259–267, 1999

Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits

D. W. Fulker,[1,2] S. S. Cherny,[1,2] P. C. Sham,[2] and J. K. Hewitt[1]

[1]Institute for Behavioral Genetics, University of Colorado, Boulder; and [2]Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, University of London, London

Summary

An extension to current maximum-likelihood variance-components procedures for mapping quantitative-trait loci in sib pairs that allows a simultaneous test of allelic association is proposed. The method involves modeling of the allelic means for a test of association, with simultaneous modeling of the sib-pair covariance structure for a test of linkage. By partitioning of the mean effect of a locus into between- and within-sibship components, the method controls for spurious associations due to population stratification and admixture. The power and efficacy of the method are illustrated through simulation of various models of both real and spurious association.

has been due to their perceived importance within the framework of clinical diagnosis. However, there is increasing recognition that for many traits of clinical interest, such as alcoholism, depression, diabetes, obesity, or hypertension, quantitative phenotypes may be more informative than diagnostic categories for genetic analysis.

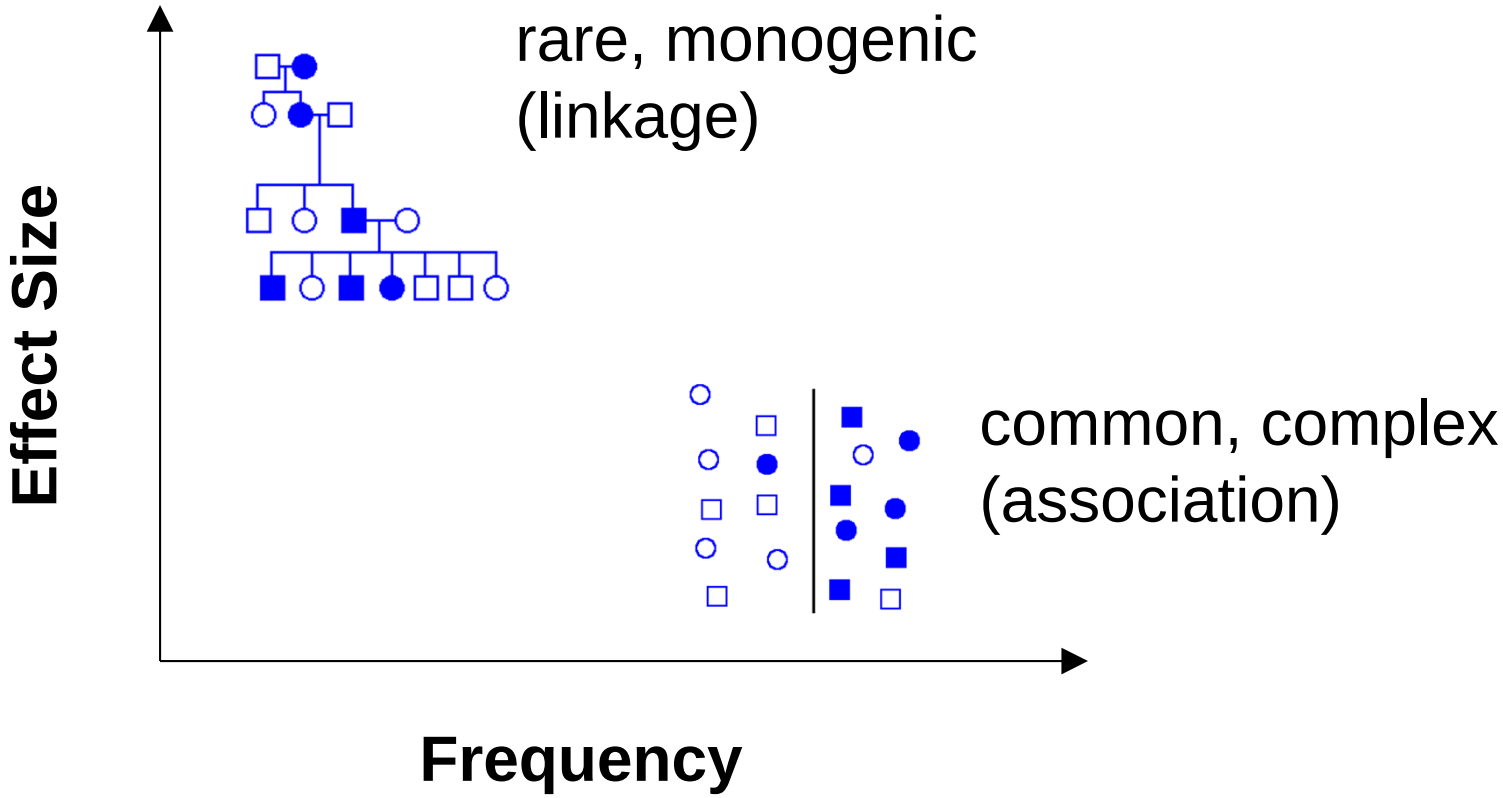
Most notable of the various methodological advances made in the area of association or disequilibrium mapping for qualitative traits are those techniques based on the use of parental control groups, such as the transmission/disequilibrium test (TDT; Spielman et al. 1993), the haplotype–relative risk approach (Terwilliger and Ott 1992), and, more recently, the development of similar procedures that use siblings (Boehnke and Langefeld

$$\hat{y}_{ij} = \mu + \beta_a g_{ij} = \mu + \beta_b b_i + \beta_w w_{ij}$$

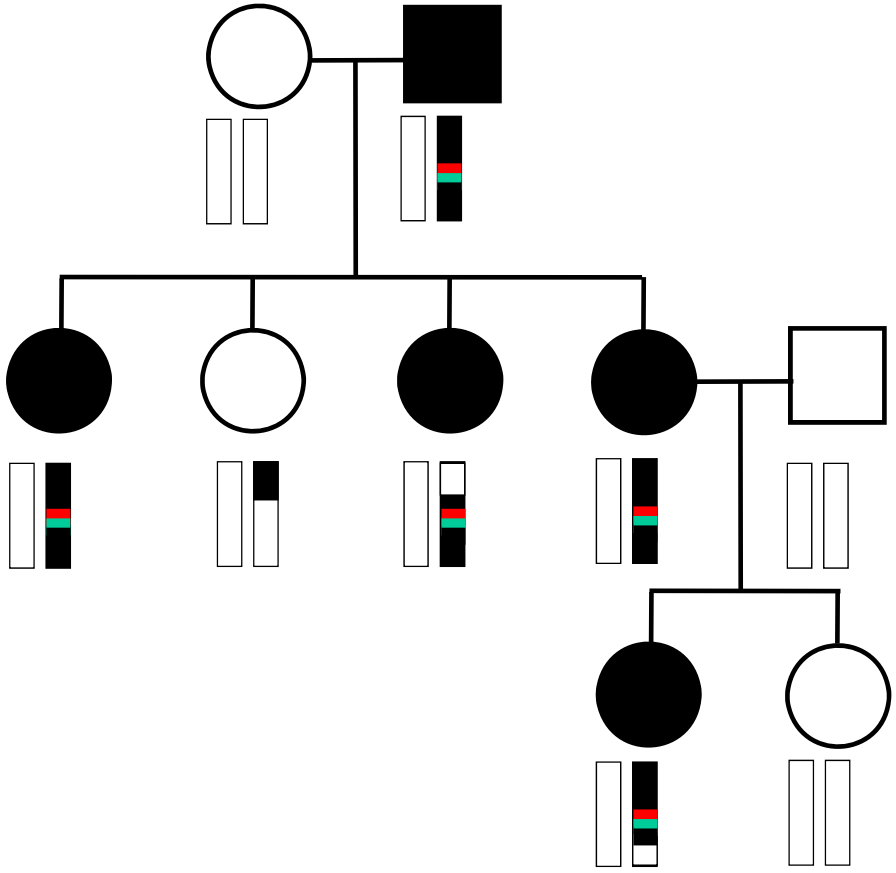
$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_q^2 + \sigma_c^2 + \sigma_e^2 & \hat{\pi}_i \sigma_q^2 + \sigma_c^2 \\ \hat{\pi}_i \sigma_q^2 + \sigma_c^2 & \sigma_q^2 + \sigma_c^2 + \sigma_e^2 \end{pmatrix}$$

When to use association...

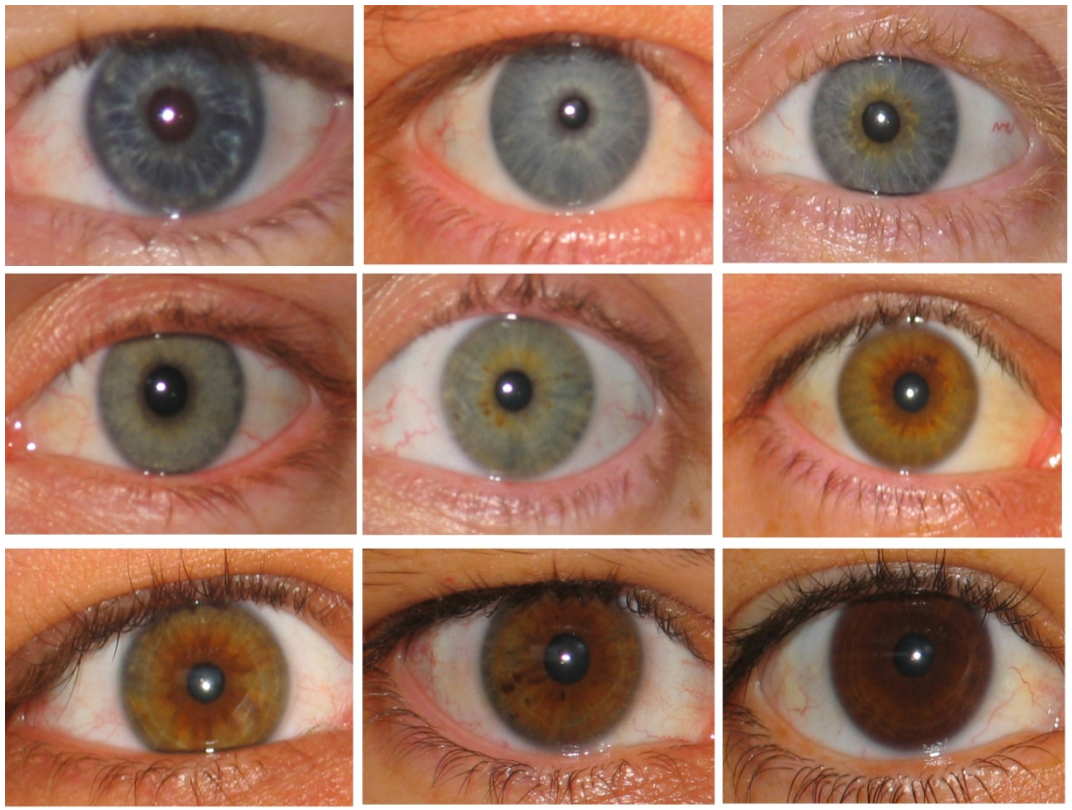
Methods of gene hunting

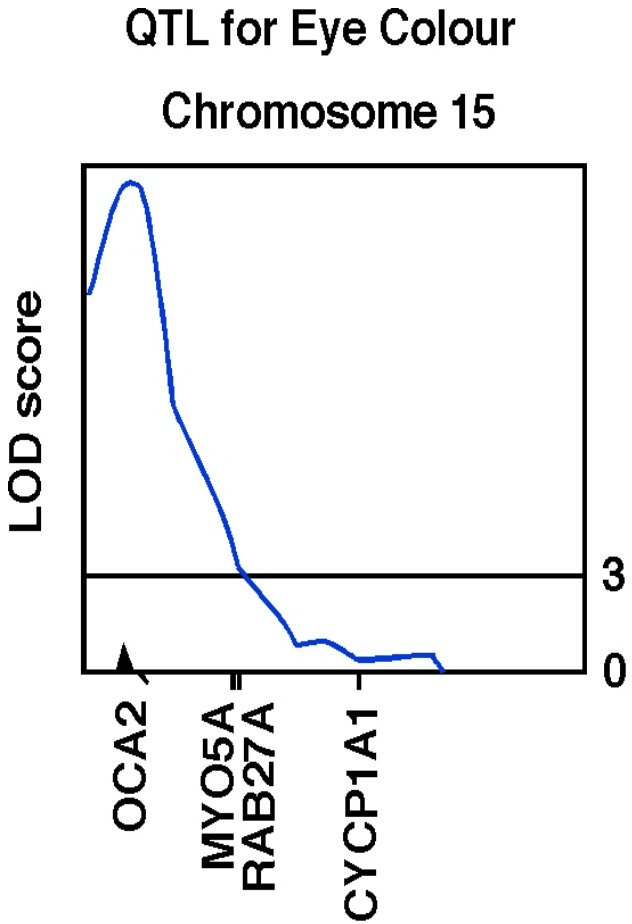


Molecular Genetic Studies- Linkage Analysis



Human OCA2 and eye colour





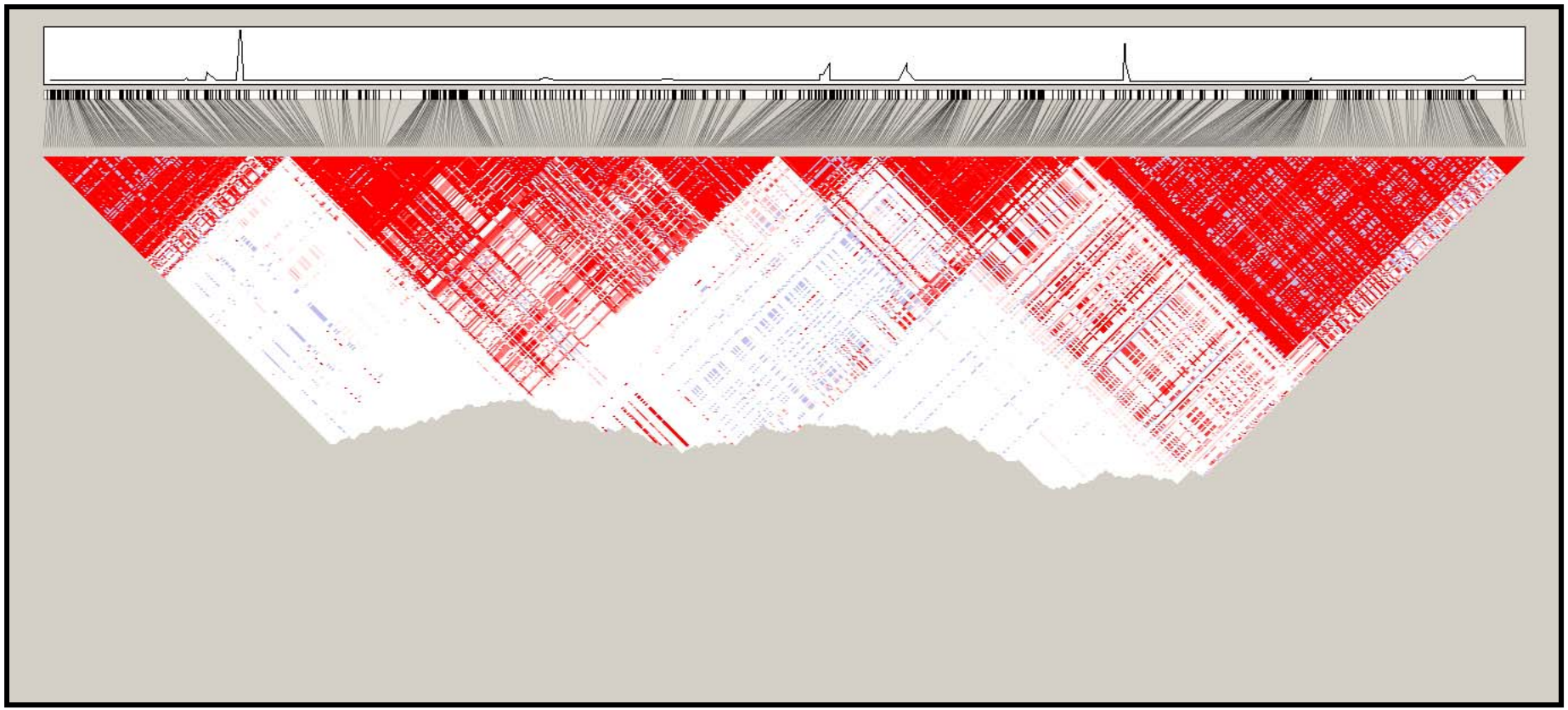
Zhu et al., *Twin Research* 7:197-210 (2004)

Association Summary

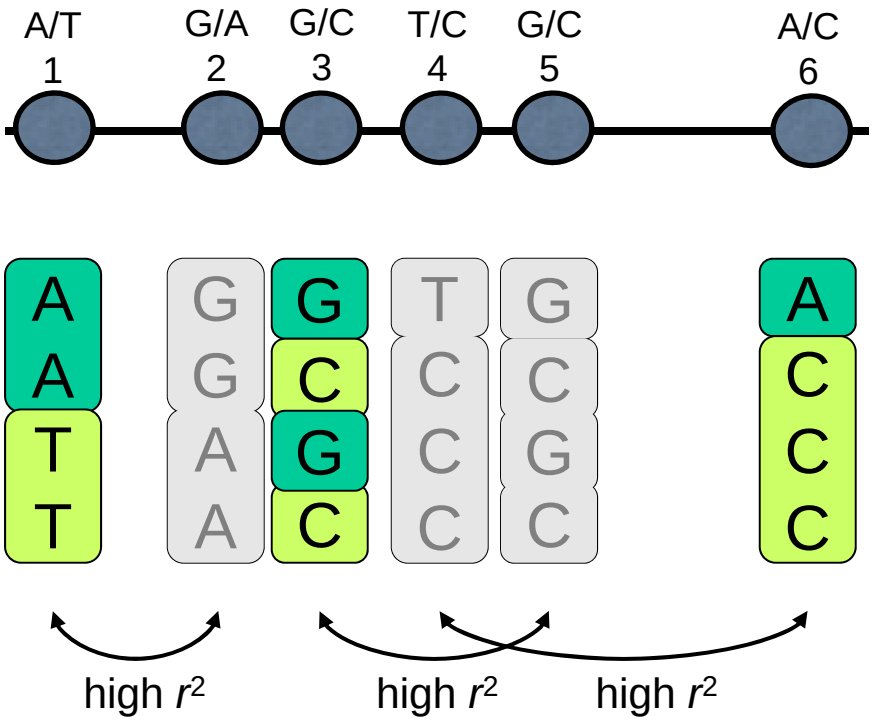
1. Families or unrelateds
2. Matching/ethnicity important
3. Many markers required for genome coverage (10^5 – 10^6 SNPs)
4. Powerful design
5. Ok for initial detection; good for fine-mapping
6. Powerful for common variants; rare variants difficult

HapMap and Tagging

Visualizing empirical LD



Pairwise tagging



Tags:

SNP 1
SNP 3
SNP 6

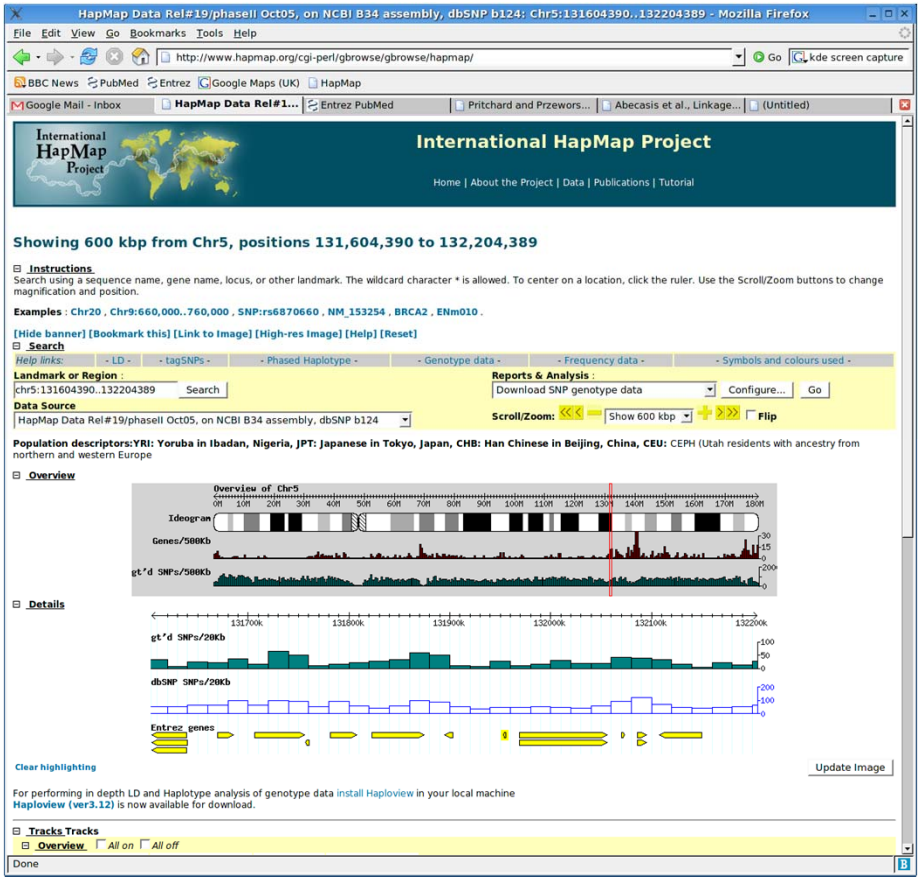
3 in total

Test for association:

SNP 1
SNP 3
SNP 6

Carlson *et al.* (2004) *AJHG* **74**:106

Enabling association studies: HapMap



1000 Genomes - Home - Windows Internet Explorer

http://www.1000genomes.org/page.php

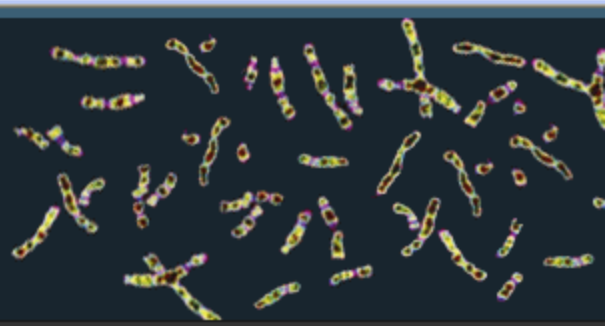
File Edit View Favorites Tools Help

Google housand genomes project Go Bookmarks PageRank 2 blocked Check AutoLink AutoFill Send to one

1000 Genomes - Home

1000 Genomes

A Deep Catalog of Human Genetic Variation



Home About Partners Data Contact Wiki

1000 GENOMES PROJECT DATA RELEASE

SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the EBI FTP site and the NCBI FTP site. The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at http://browser.1000genomes.org. Launch the browser and view a sample region here.

More information about the data release can be found in the data section of this web site.

Download the 1000 Genomes Browser Quick Start Guide

Quick start (pdf)

LOG IN

Username:

Password:

Login (I forgot my password)

LINKS

Download the meeting report

View the participants

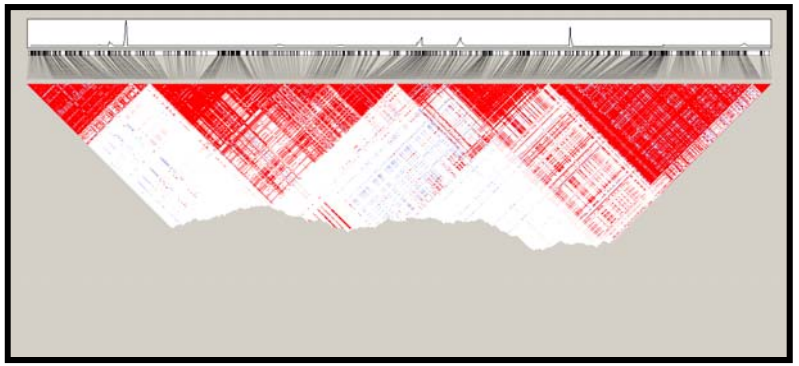
Done

start Missing heritability NG... EpiSlides [Compatibilit... 1000 Genomes - Hom... Boulder 2009

Genome-wide Association Studies

Enabling Genome-wide Association Studies

- HAPlotype MAP



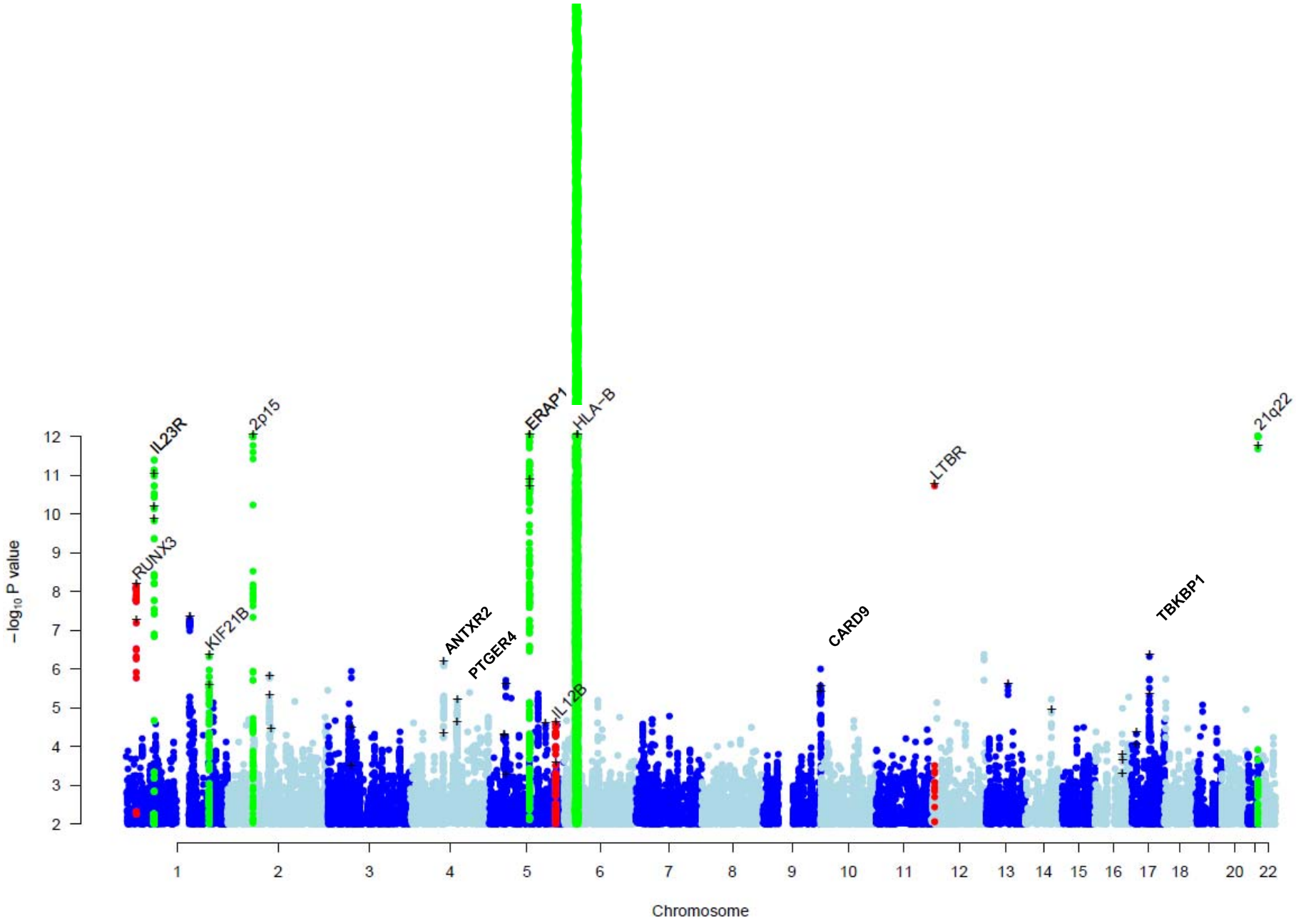
- High throughput genotyping

- Large cohorts

AUSTRALIAN twin REGISTRY

VRIJE UNIVERSITEIT NEDERLANDS TWEELINGEN REGISTER

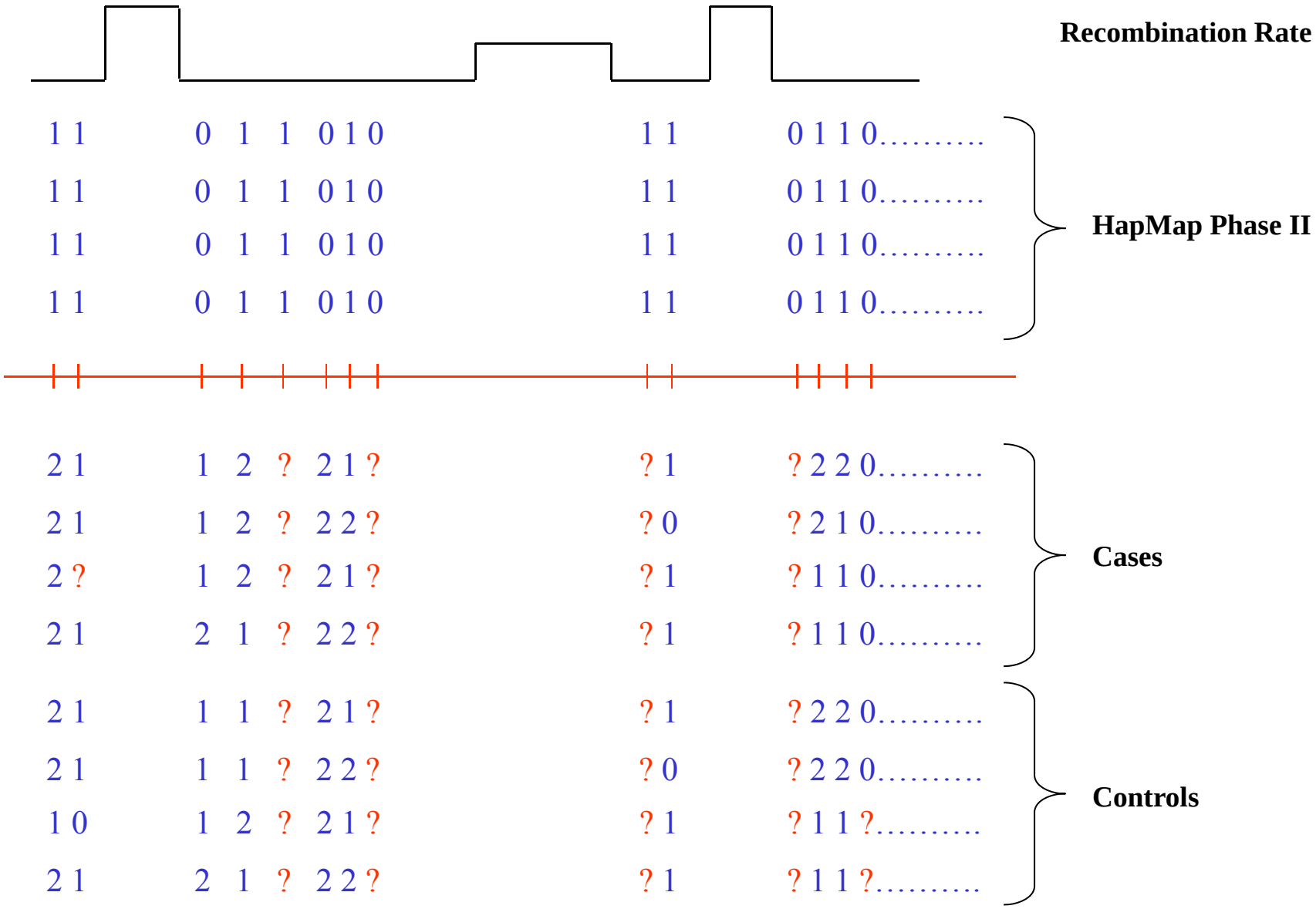
CHILDREN OF THE 90s



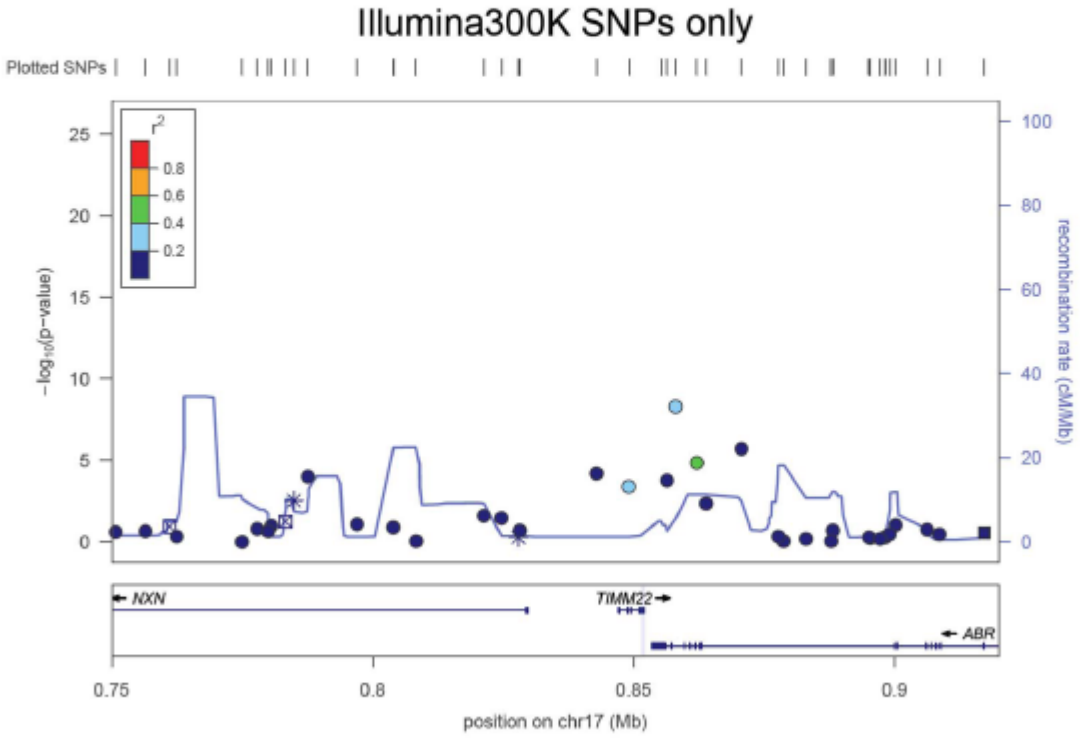
Evans et al. (2011) Nat Genet

Imputation

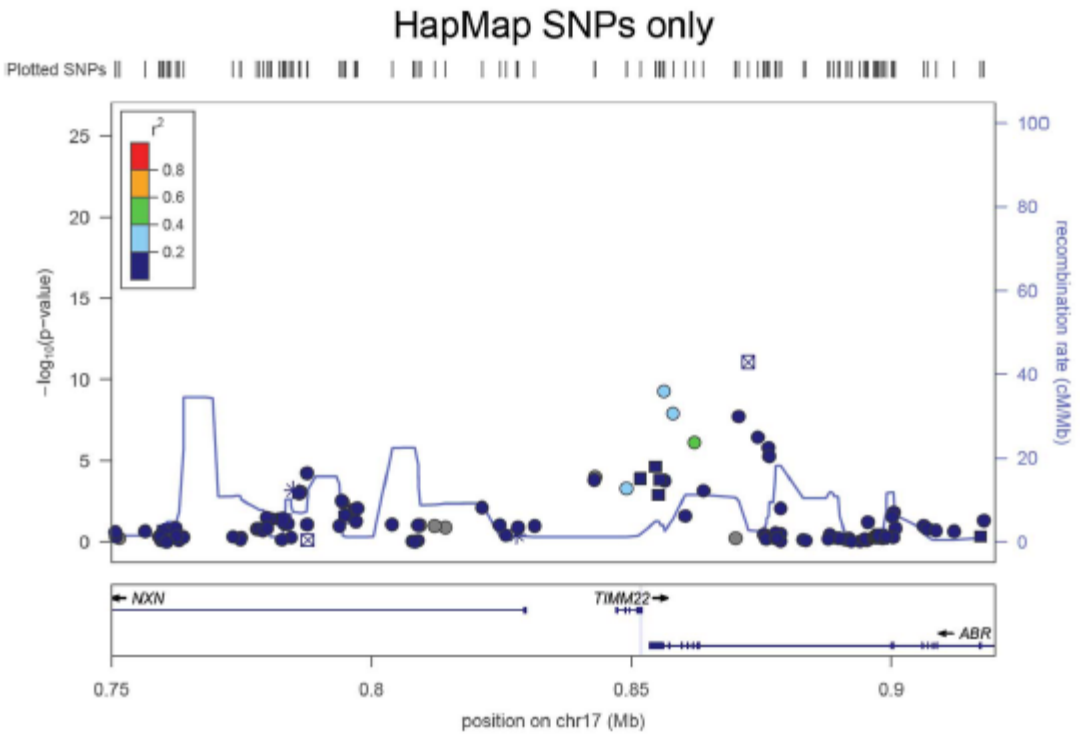
Imputation


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

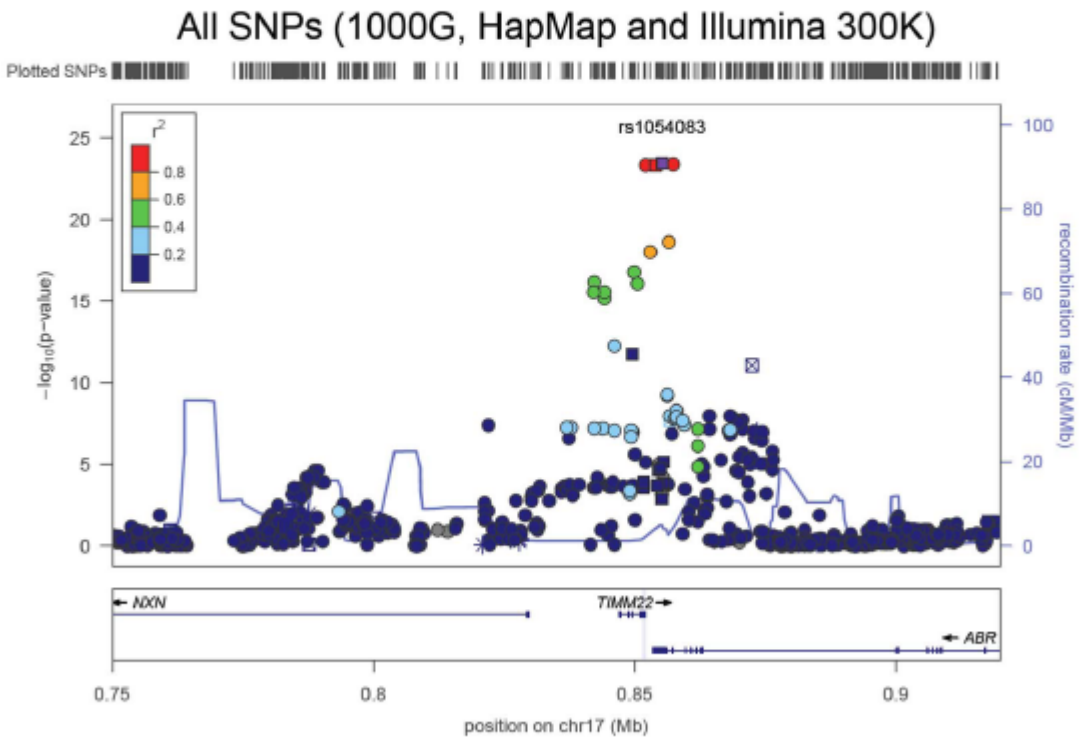
Enhance Association Studies: eQTL Imputation Example



Enhance Association Studies: eQTL Imputation Example

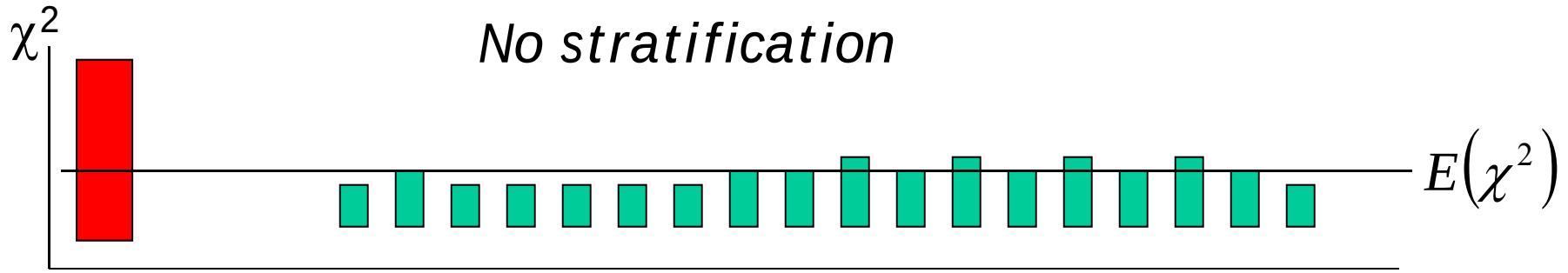


Enhance Association Studies: eQTL Imputation Example



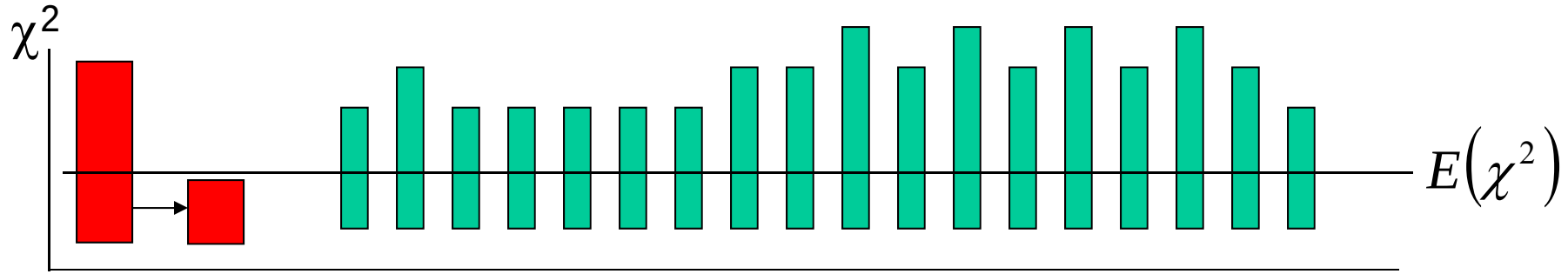
Population Stratification

Genomic control



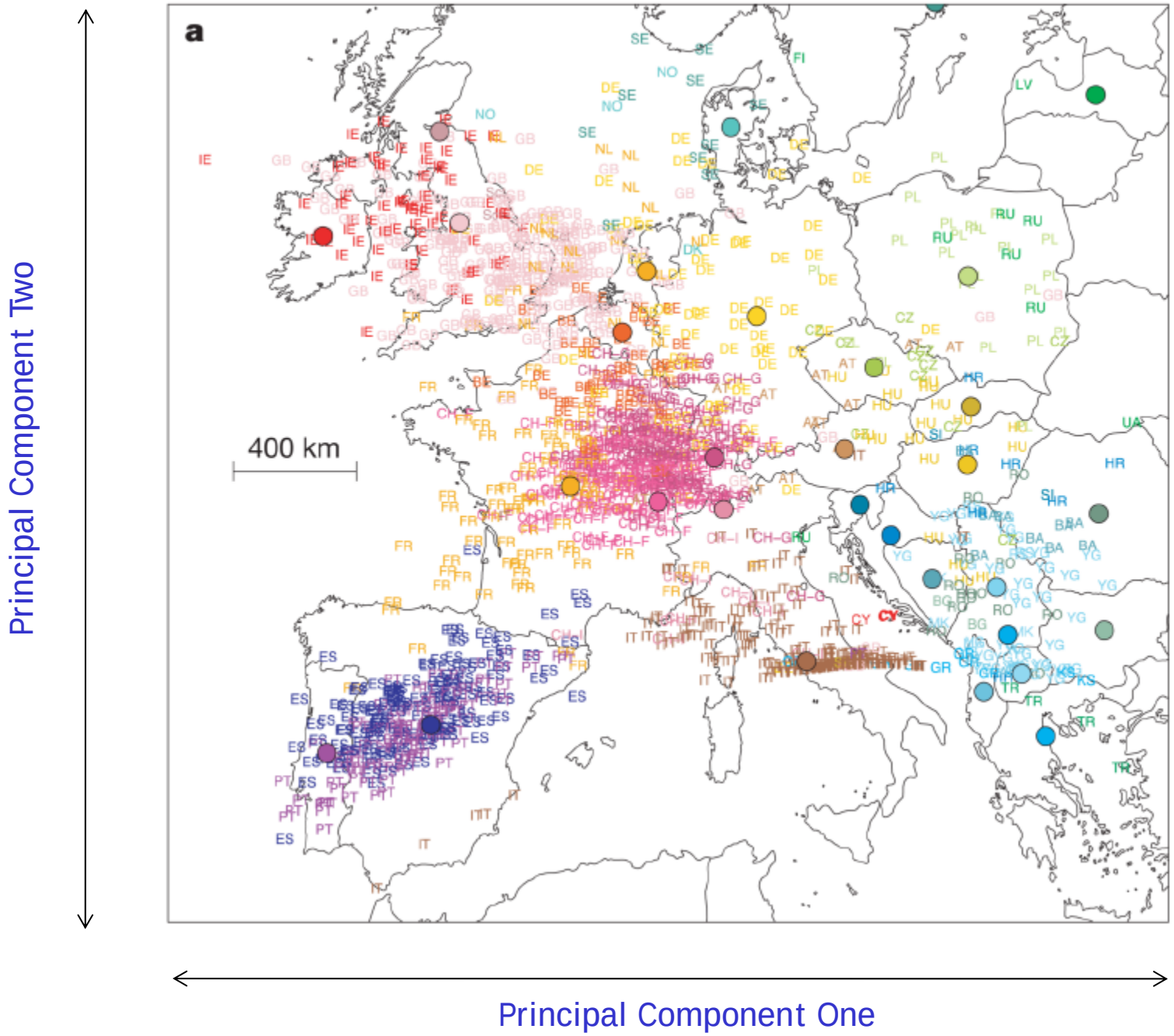
Test locus

Unlinked 'null' markers



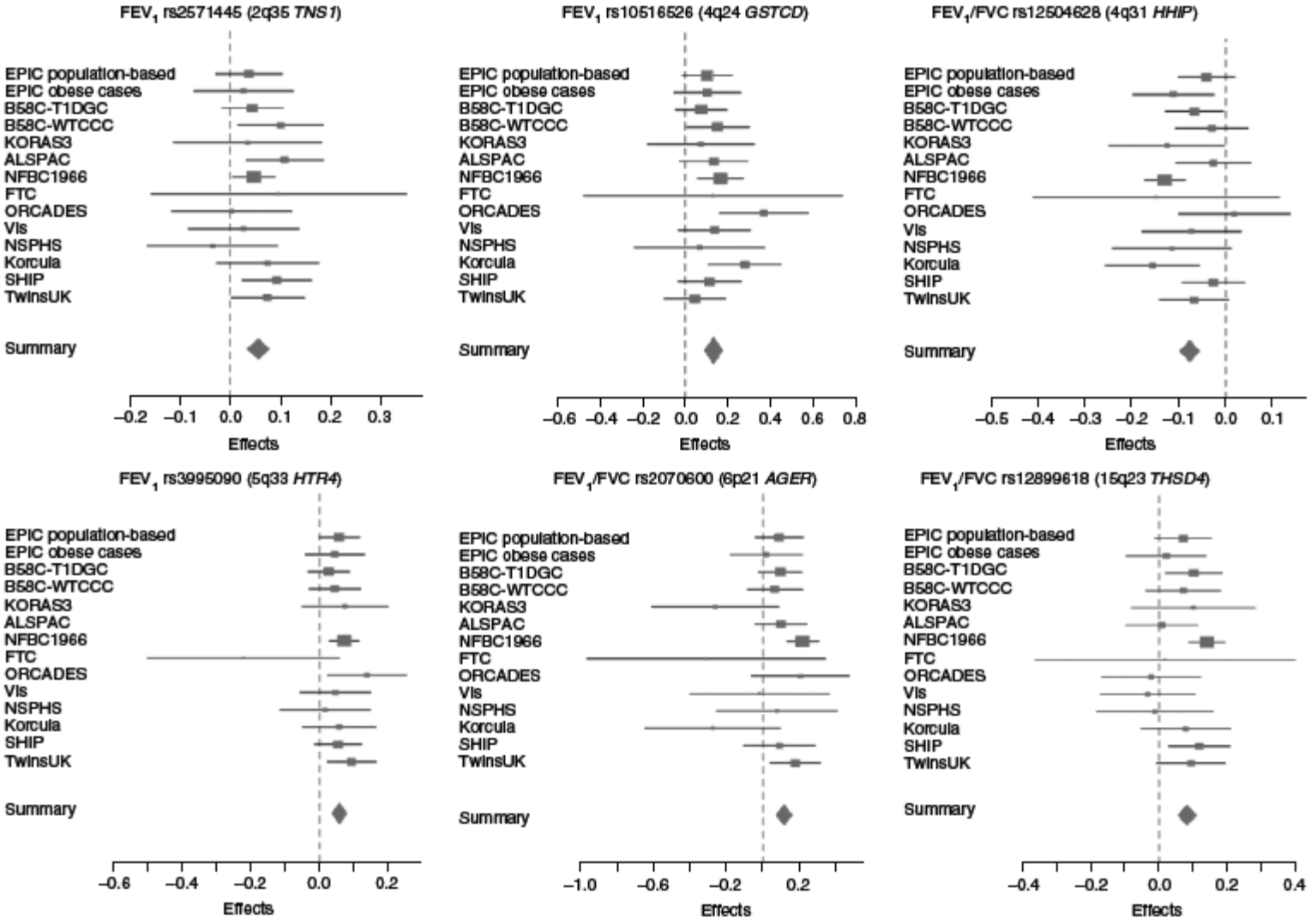
Stratification → adjust test statistic

Novembre et al, Nature (2008)



Meta-analysis

Meta-analysis



Repapi et al. (2009) *Nature Genetics*

Replication

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

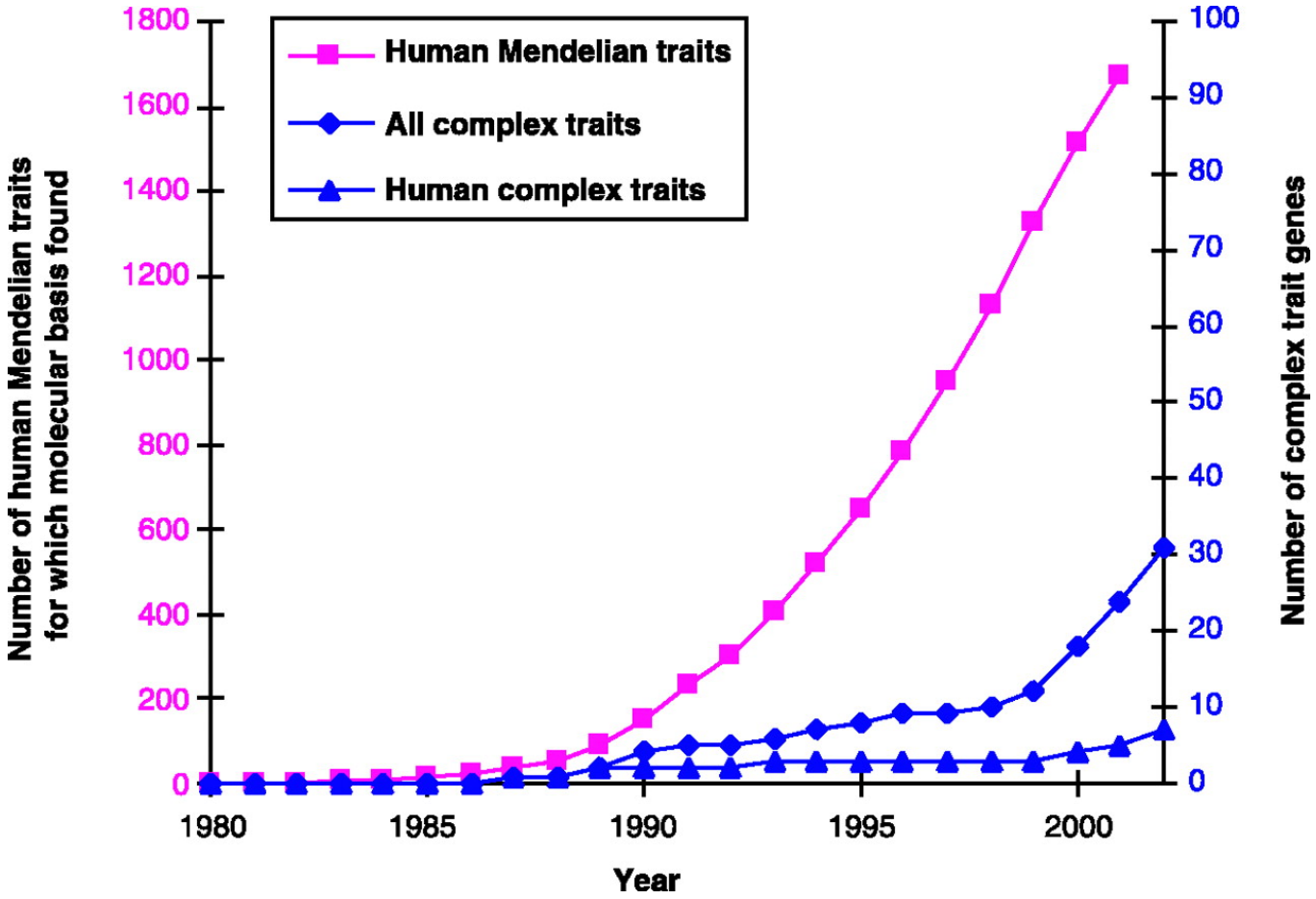
Well designed negative studies are valuable

Programs for performing association analysis

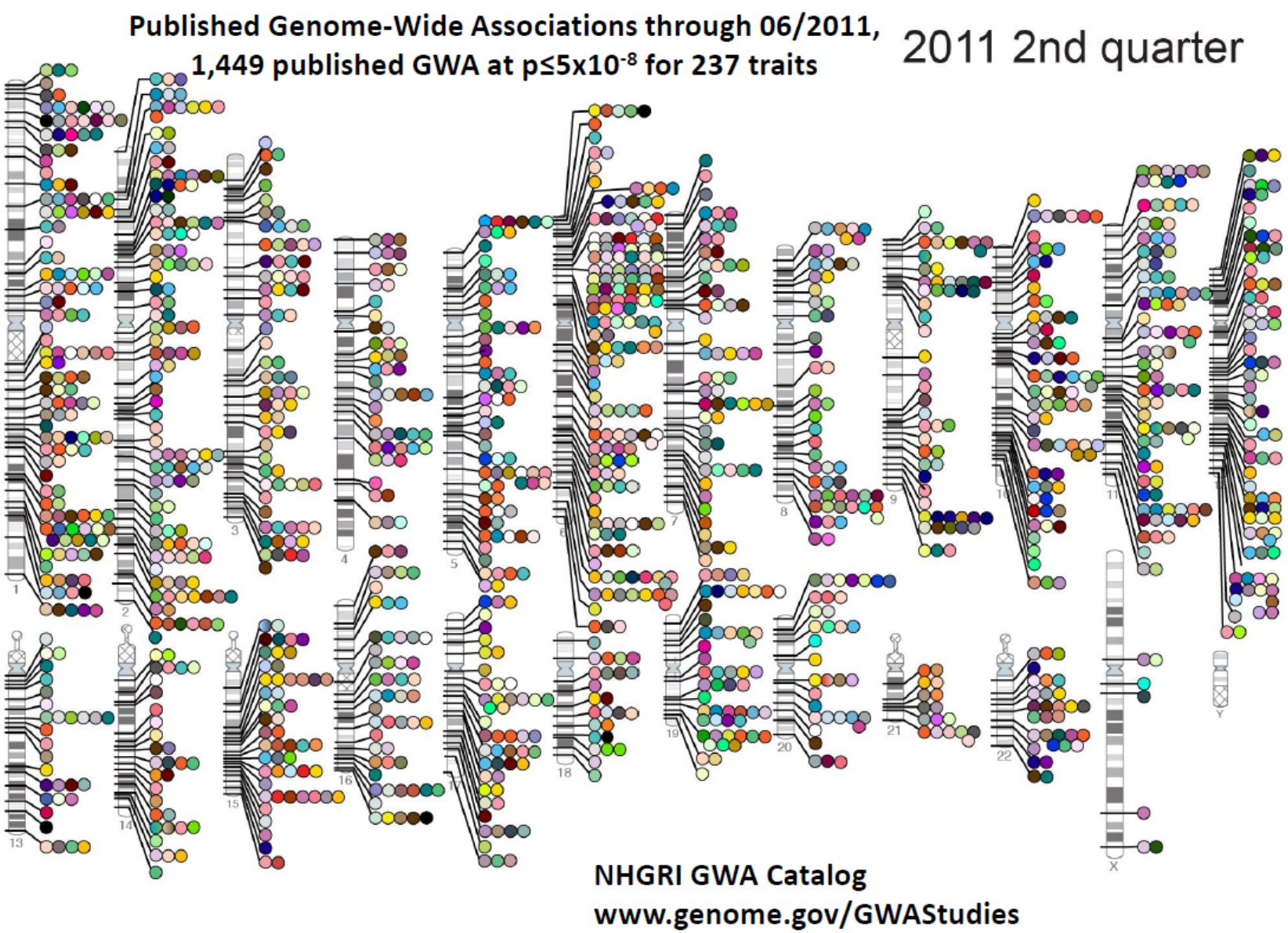
- Mx (Neale)
  - Fully flexible, ordinal data
  - Not ideal for large pedigrees or GWAs
- PLINK (Purcell, Neale, Ferreira)
  - GWA
- Haploview (Barrett)
  - Graphical visualization of LD, tagging, basic tests of association
- MERLIN, QTDT (Abecasis)
  - Association and linkage in families

How have we done?

Historical gene mapping



Glazier et al, *Science* (2002).


Published Genome-Wide Associations through 06/2011,
1,449 published GWA at p≤5x10-8 for 237 traits
2011 2nd quarter
NHGRI GWA Catalog
www.genome.gov/GWAStudies

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.

Sequencing and Rare Variants

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

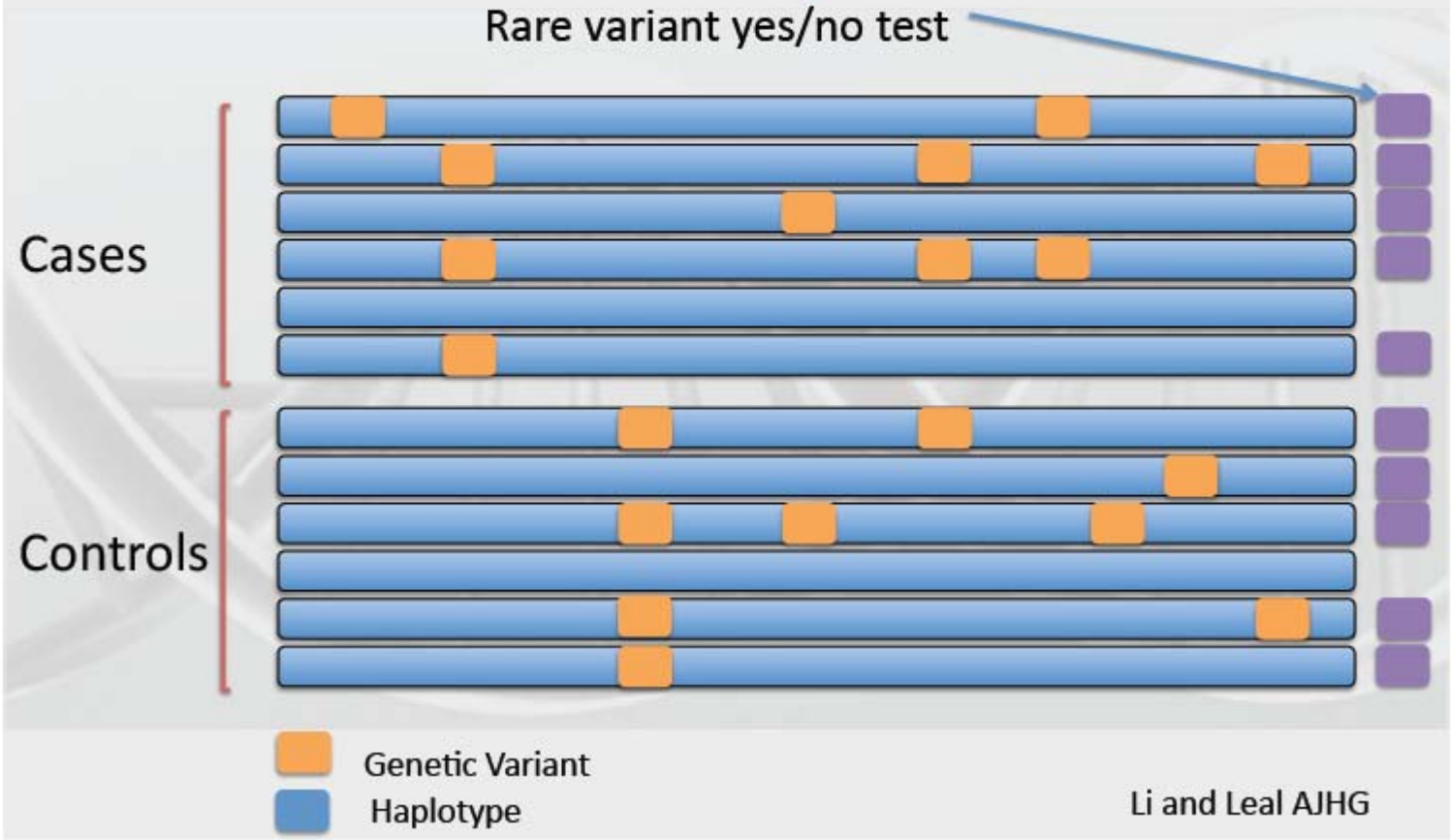
Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

Metzker et al (2010) *Nature Reviews Genetics*

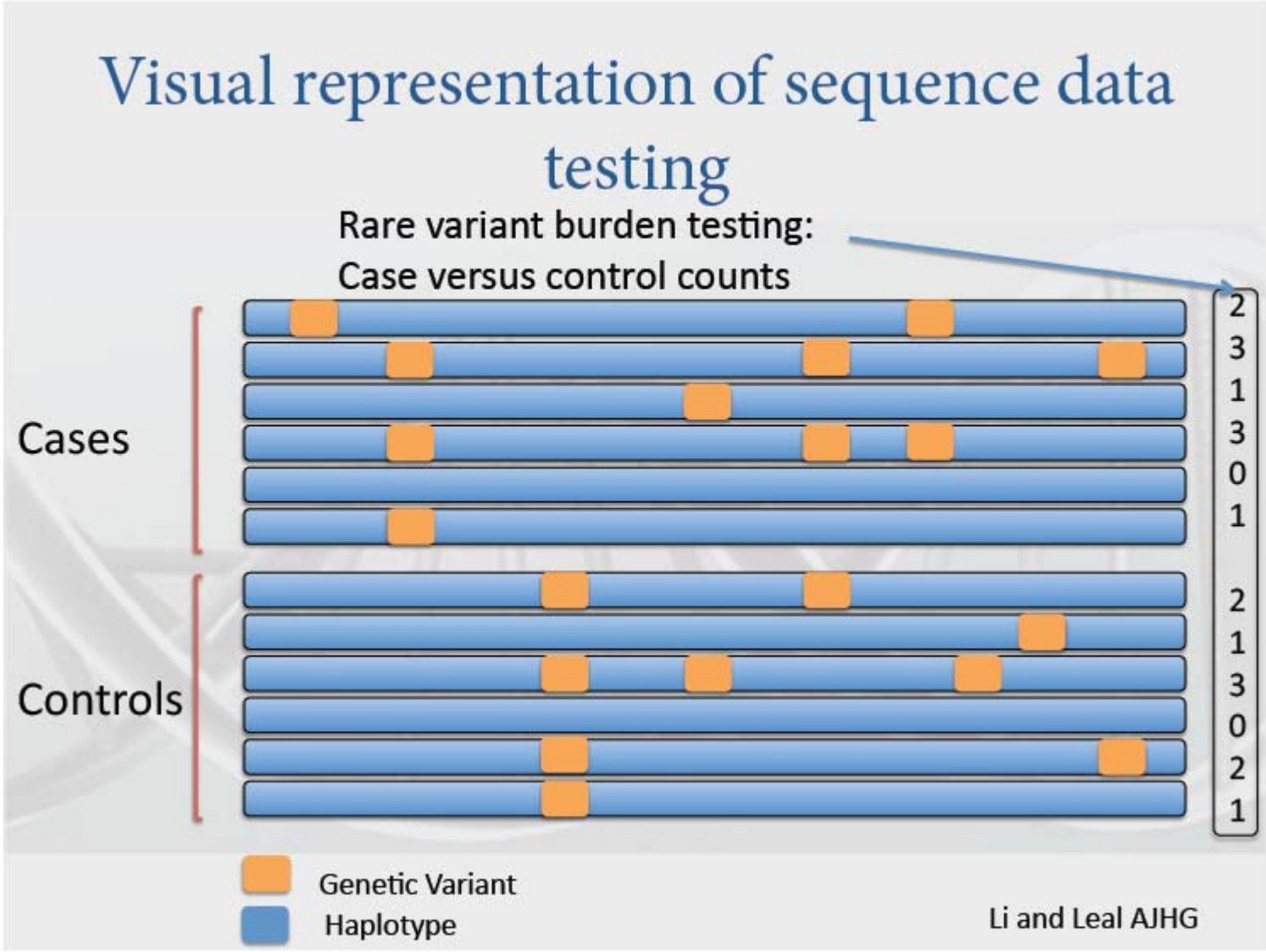
Analysis of Rare Variants

- How to combine rare variants?
  - "Ordinary" tests of association won't work
  - Collapse across all SNPs?
- Which SNPs to include?
  - Frequency?
  - Function?
- How to define a region?

Visual representation of sequence data testing



Visual representation of sequence data testing



Summary

1. Genetic association studies can be used to locate common genetic variants that increase risk of disease/affect quantitative phenotypes

2. Genome-wide association spectacularly successful in identifying common variants underlying complex traits and disease

3. The next challenge is to explain the “missing heritability” in the genome. Genome-wide sequencing and the analysis of rare variants will play a major part in this effort