Concepts in complex diseases: from Fisher to GWAS

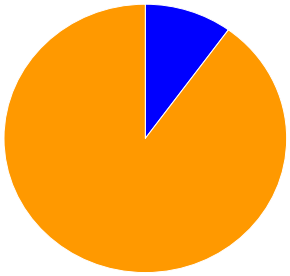
Guy Van Camp

Department of Medical Genetics
University of Antwerp

Approaches to study complex diseases

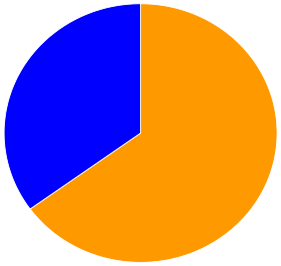
- Polygenic theory of Fisher
- Linkage disequilibrium
- How to identify genes for complex diseases?
- What has been accomplished today?
- Pitfalls of genetic association studies
  - Multiple testing
  - Missing heritability
- How do genetic variants exert an effect

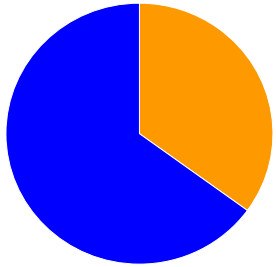
Genes and Disease



Monogenic Diseases

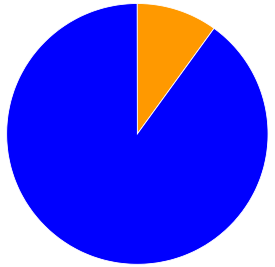
- **Huntington Disease**
- **Cystic fibrosis**
- **Fragile X syndrome**





Complex Diseases

- **Myocardial infarction**
- **Hypertension**
- **Alzheimer disease**



Environmental Diseases

- **Pathogens**
- **Poisoning**

Environment **Genes**

Mendelian disorder

Complex disorder

Some differences …

- Mutation in a gene is sufficient to cause the disorder
- Recognizable inheritance patterns
- One gene per family
- Less common diseases

- Mutation in a gene confers an increased risk, but is not sufficient to cause the disorder
- No clear inheritance pattern
- Involves many genes or genes and environment
- Many are common diseases

Complex traits: polygenic theory

Sir Ronald Aylmer Fisher (1890-1962)

- Created the foundations for modern statistical science
- Reconciled the discontinuous nature of Mendelian inheritance with continuous variation

248

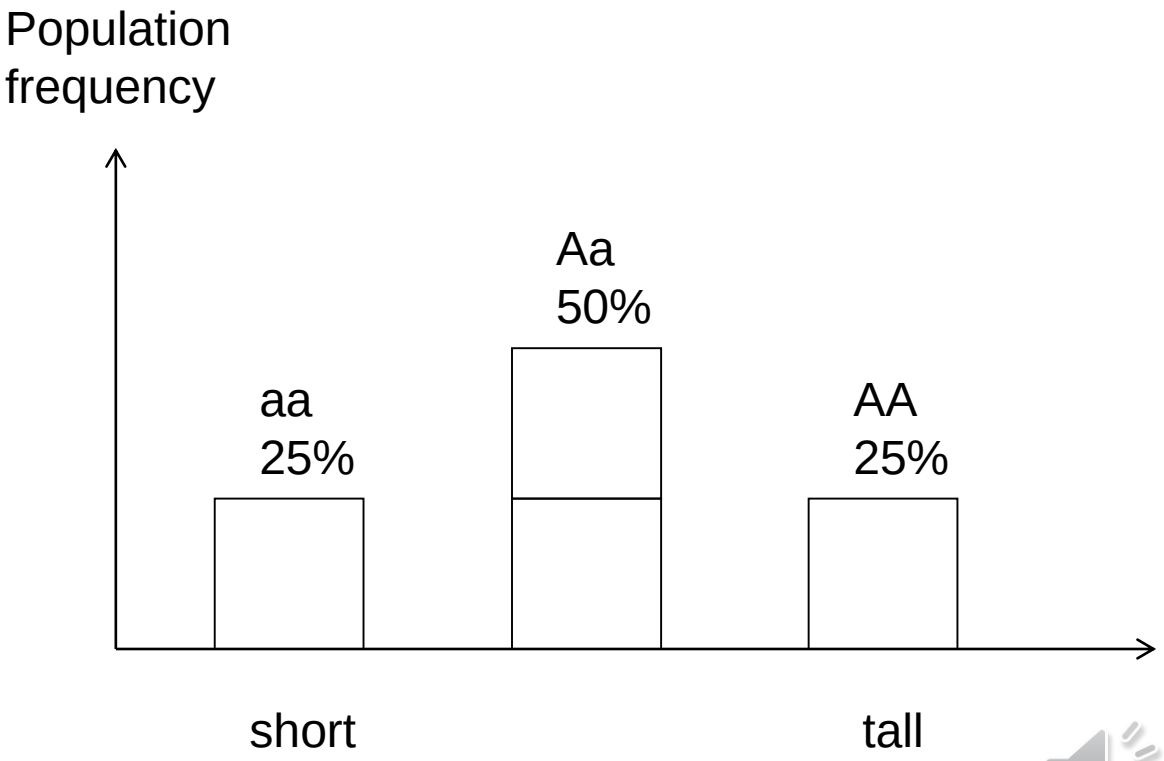
STATISTICAL METHODS IN GENETICS

R. A. FISHER

*Being the Bateson Lecture delivered at the John Innes Horticultural Institution on Friday, 6th July 1951 **

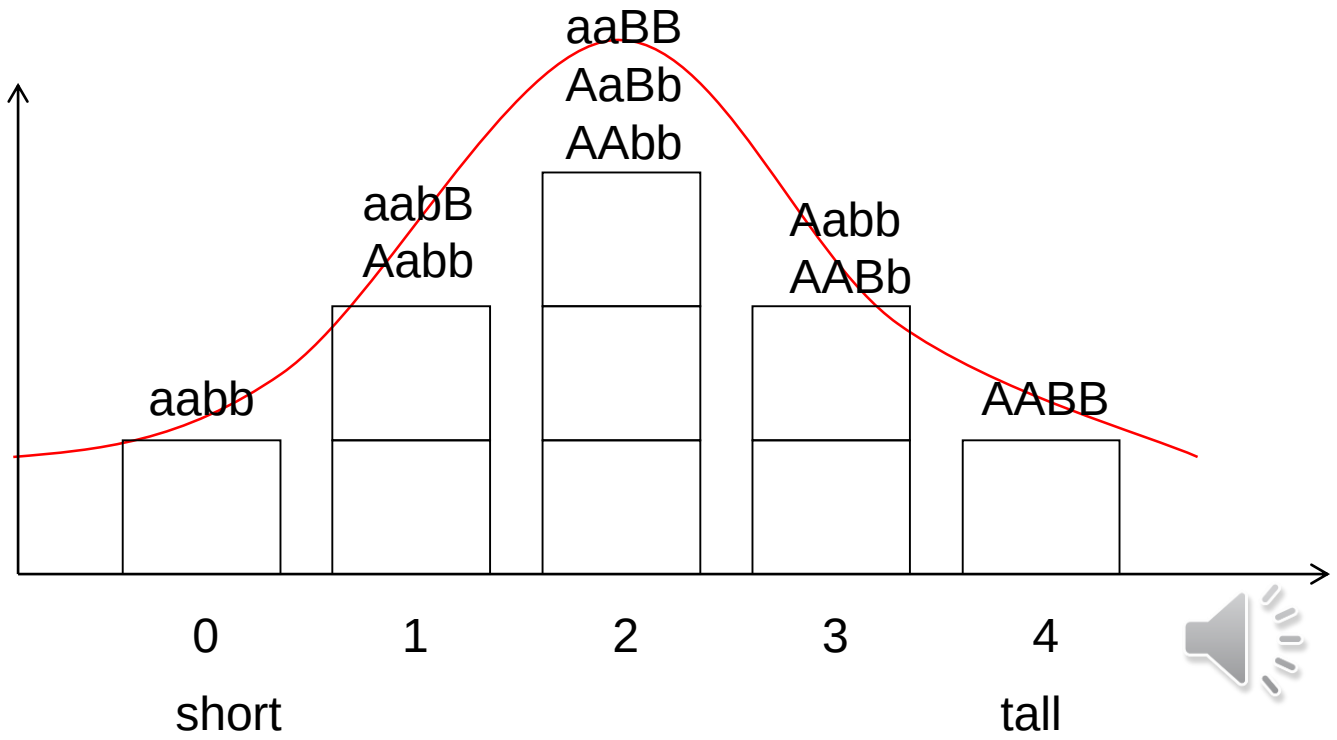
Complex traits: polygenic theory

Example: body length
Suppose simple monogenic trait with **one gene A**, two alleles



Suppose simple monogenic trait with **two genes A and B**, 4 alleles

aabb	aabB Aabb	aaBB AaBb AAbb	AaBB AABb	AABB
0	1	2	3	4

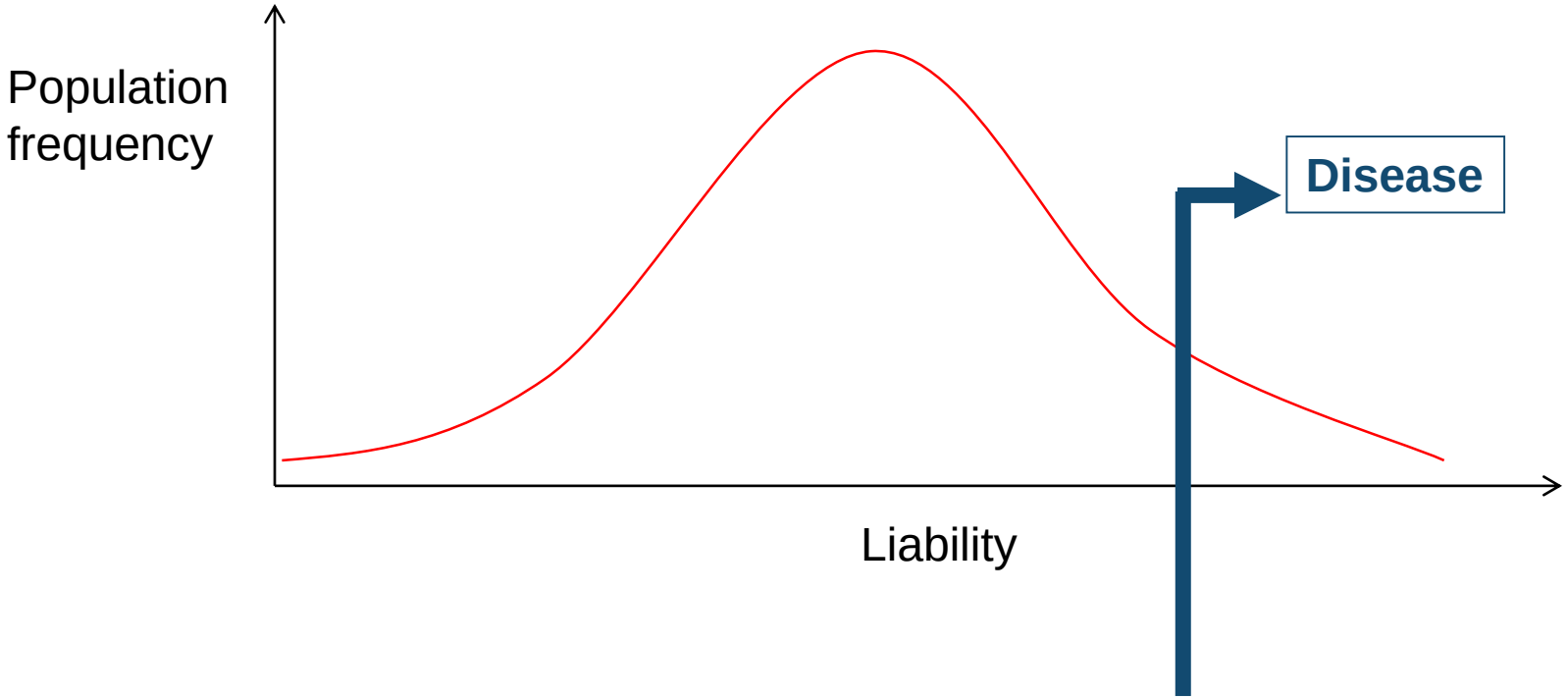


Binary traits (health-disease)

Some traits are binary, not continuous

e.g. Disease or health

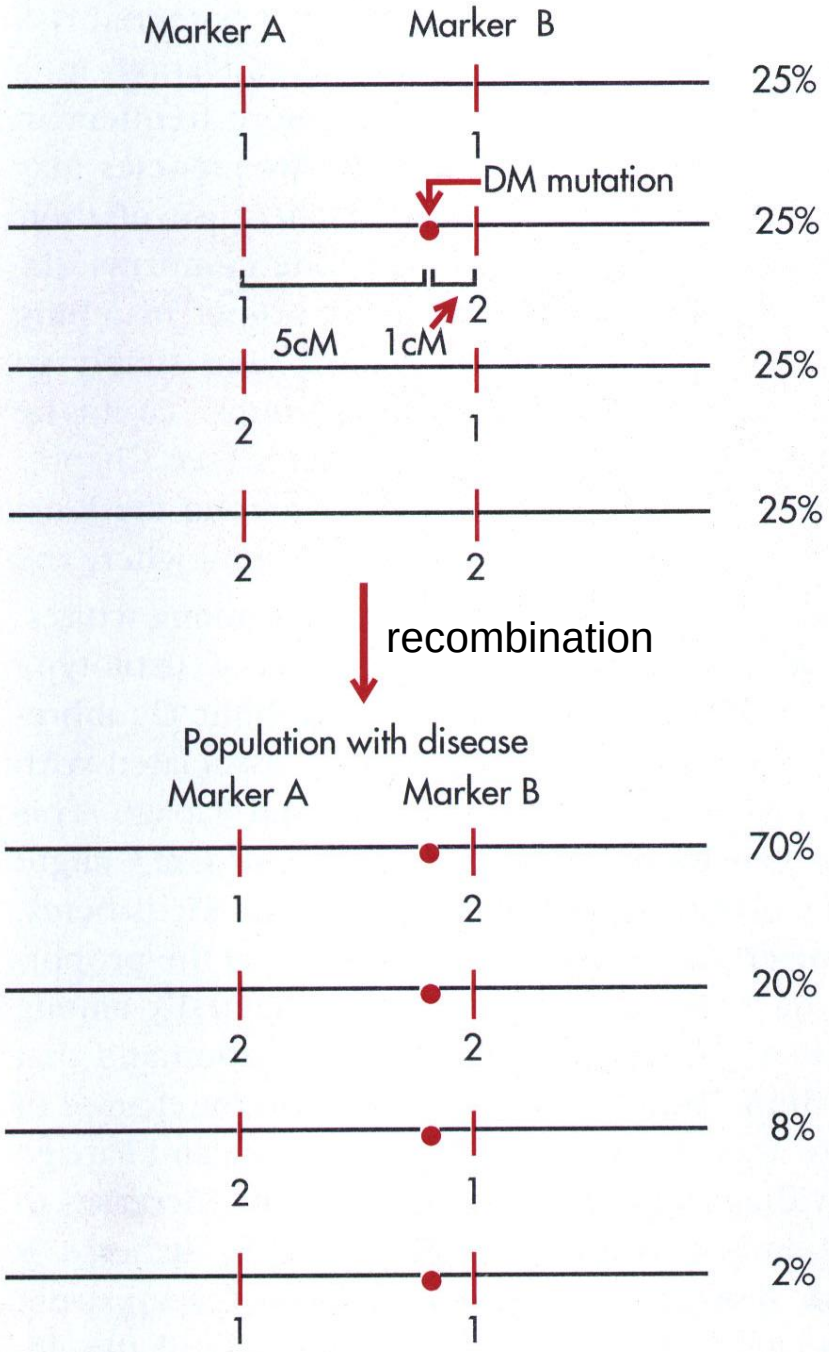
Liability distribution, threshold model



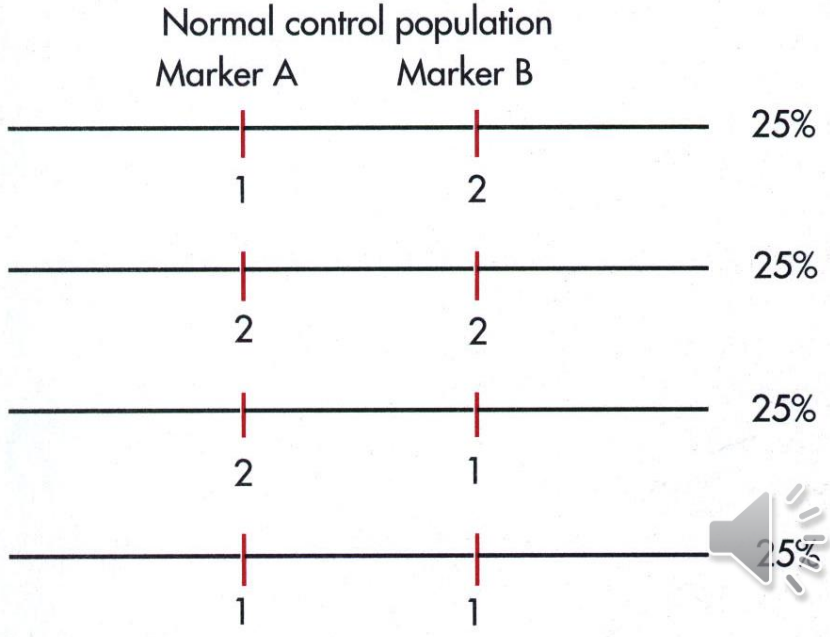
Linkage disequilibrium (LD)

- Linkage Disequilibrium is the non-random association of alleles at two or more loci
- Some haplotypes occur more or less frequently than would be expected on the basis of their allele frequencies
- Can occur between a disease mutation and markers
  - Monogenic diseases (e.g. myotonic dystrophy, cystic fibrosis)
  - Complex diseases
  - Due to common ancestor
- Can occur between DNA variants

LD between a mutation and markers

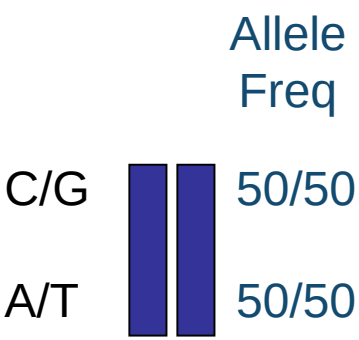


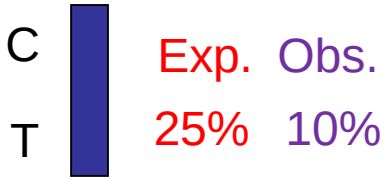
- Without recombination 100% of the mutations is on the original haplotype
- Mutations on another haplotype originate by recombination

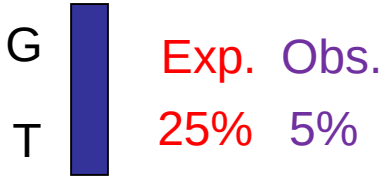


LD between SNPs

2 SNPs closely together: expected and observed haplotype frequencies







4 haplotypes

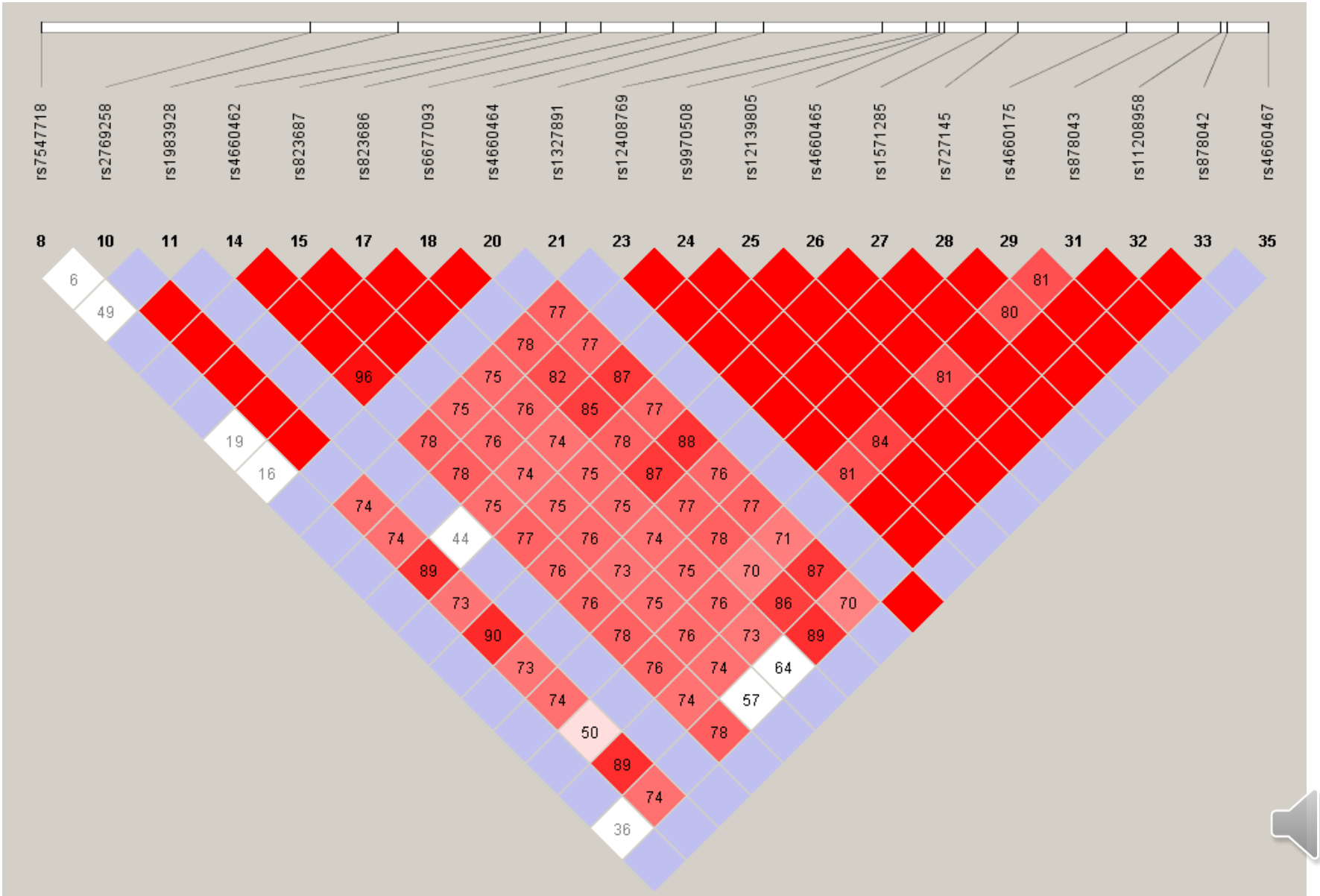
LD decay

- A new SNP allele that arises by mutation is in LD with all surrounding alleles of the haplotype on which it arose
- LD breaks down by recombination
- Remaining LD is due to lack of historic recombination between adjacent markers
  - On average, pairwise LD decays with distance between SNPs
  - Over short distances, this decay is not a smooth function, rather stepwise

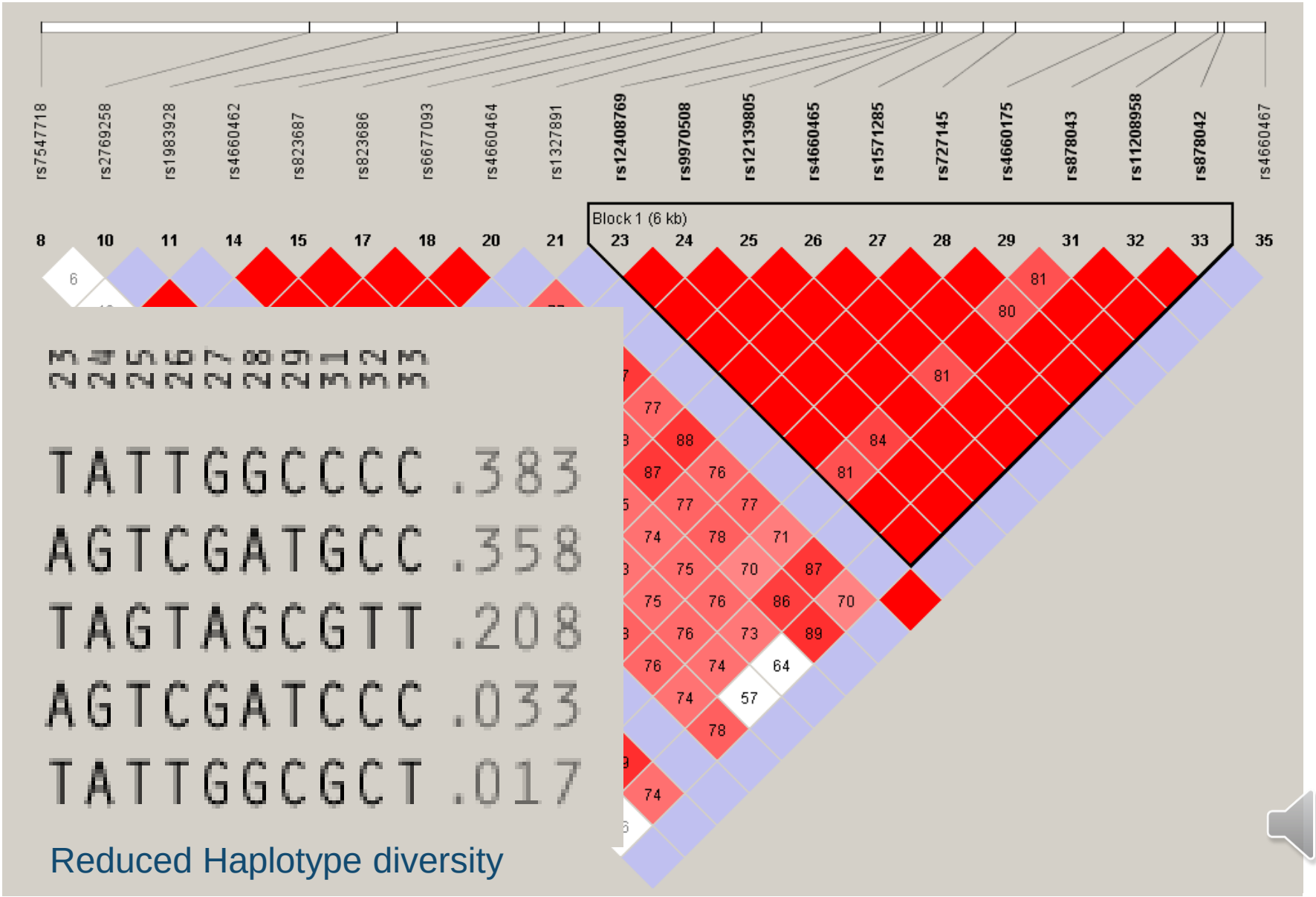
LD blocks

- Long stretch of markers in LD followed by recombination hotspot
- LD block:
  - region of high LD between adjacent SNPs
  - region of limited haplotype diversity
- Blocks are found over entire genome, but boundaries not always clear

LD structure in Haploview

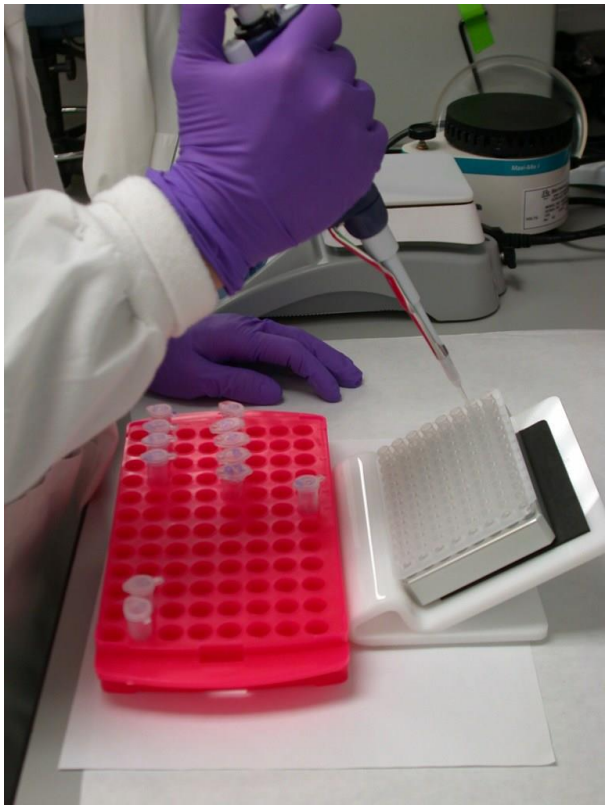


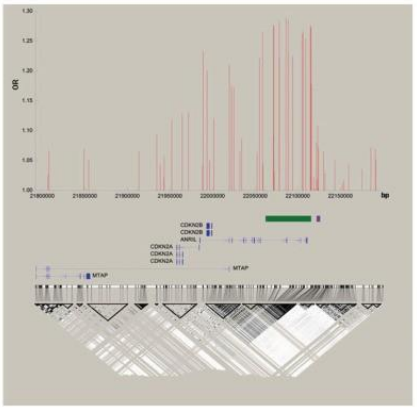
LD structure in Haploview



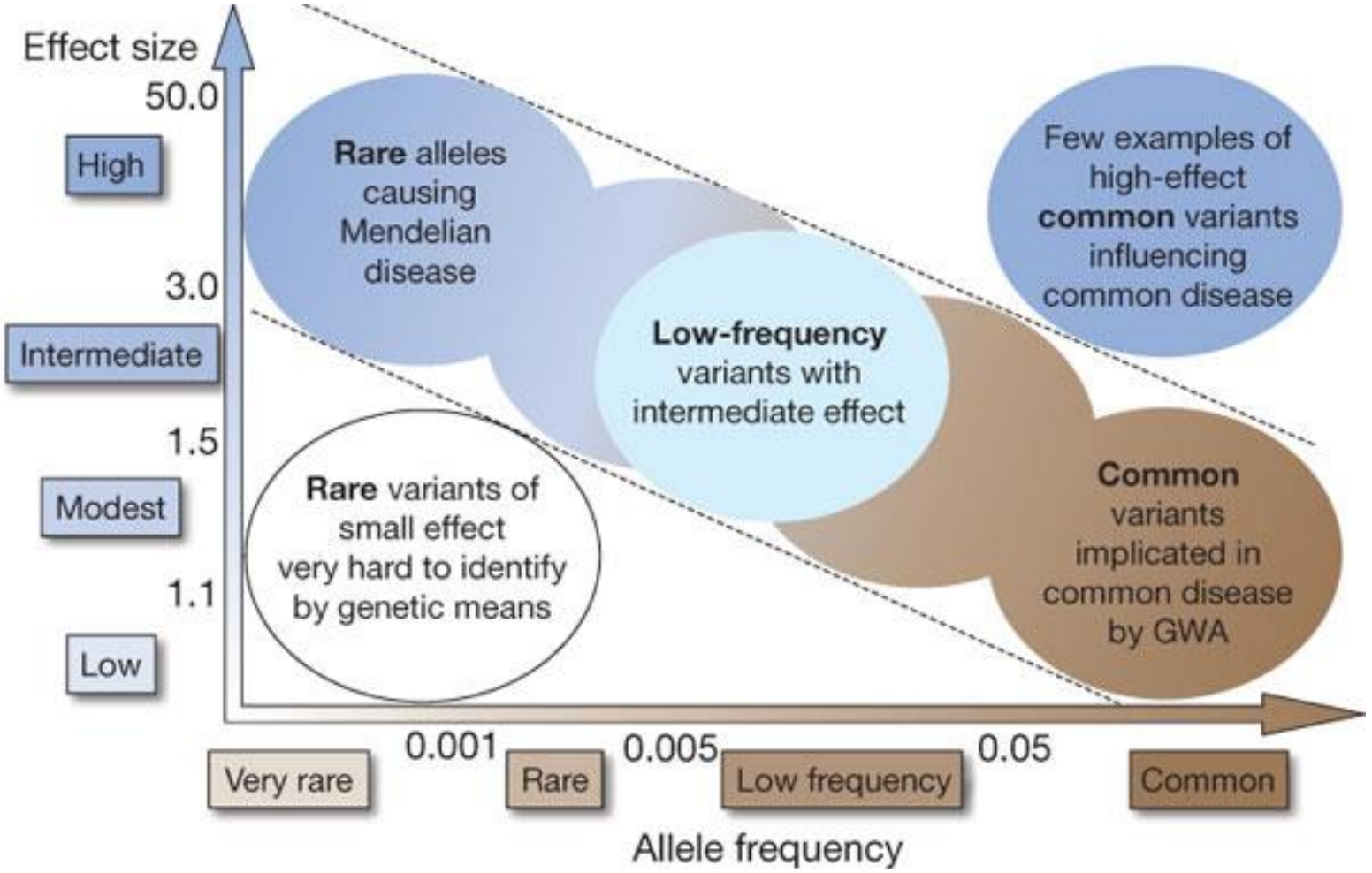
Reduced Haplotype diversity

How to identify genes for complex phenotypes?



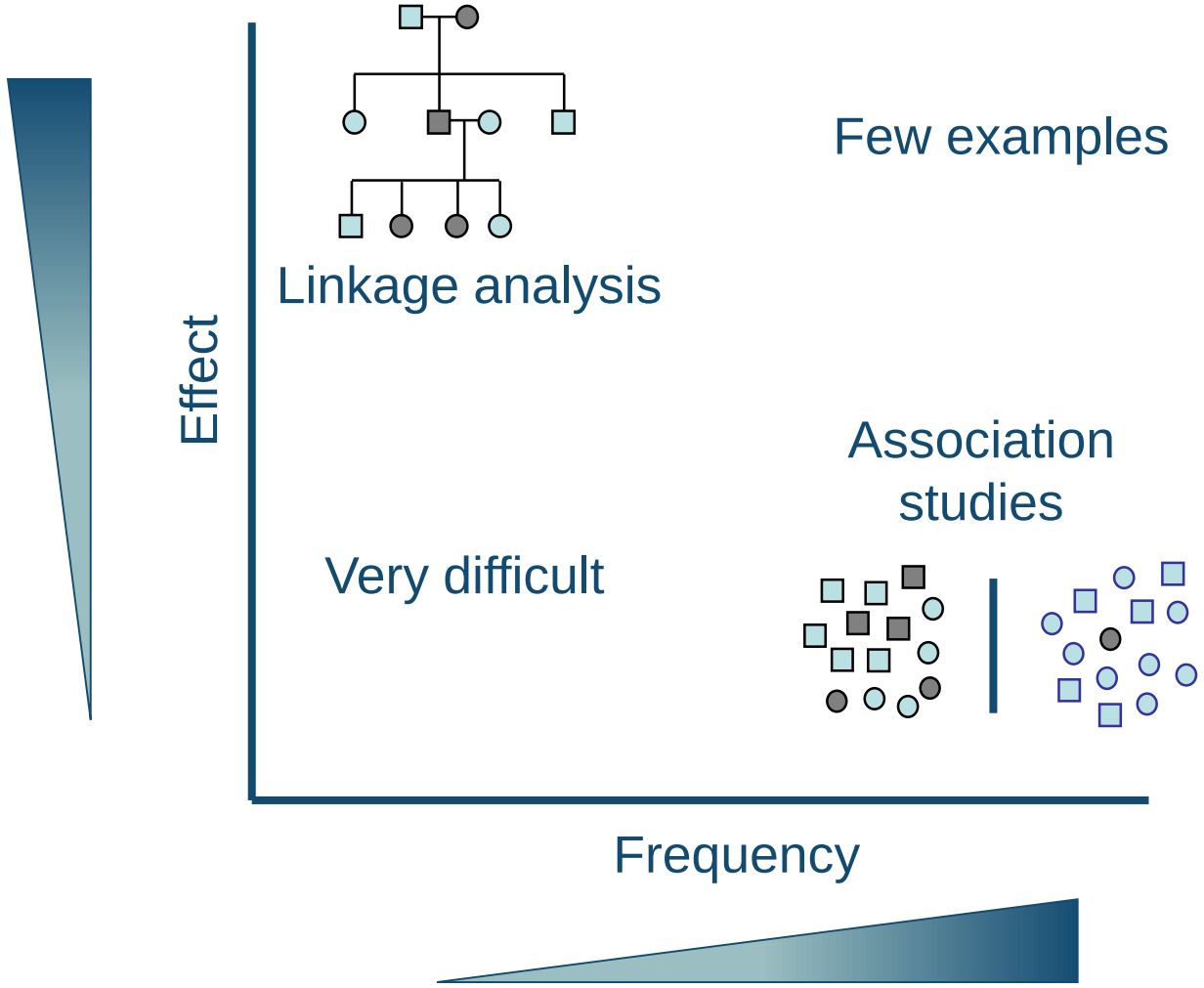


Feasibility of identifying disease genes



Manolio *et al. Nature* **461**, 747-753 (2009)

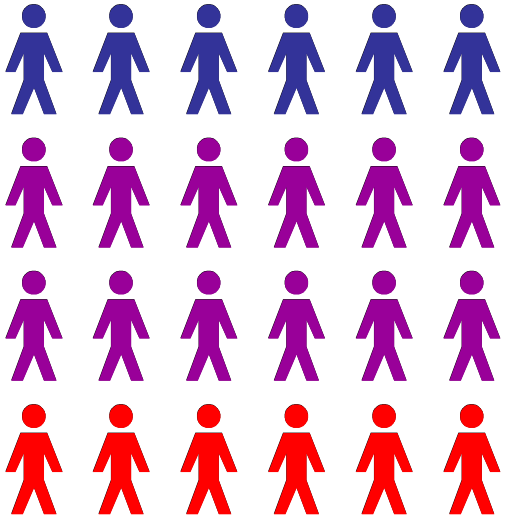
Popular methods for disease gene identification



Protective

… A T G C A A T G A C …
… A T G C A A T G A C …
… A T G C A A T G A C …
… A T G C A T T G A C …
… A T G C A T T G A C …
… A T G C A T T G A C …

Risk

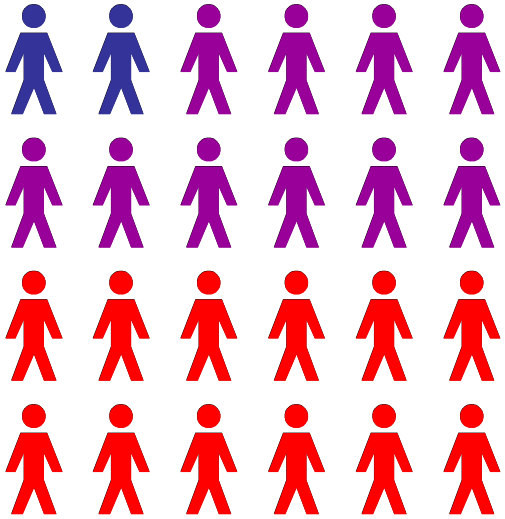


General population

SNP: AA

SNP: TT

SNP: AT



Patients

Genome-wide Association Studies (GWAS)

- Genotype using SNP arrays
- Due to LD: no need to type all SNPs
  - tagSNPs on array give info on non-typed SNPs: imputation of non-typed SNPs is possible
  - Mostly 500,000 to 10^6 SNPs on an array
  - Illumina 550 K using tagSNPs: 89% coverage ($r^2 > 0.8$)

LD : strength or weakness ?

- Pro : Can pick up association through surrounding markers in LD

- Con : If you find an associated SNP, you can’t be sure it’s the causative one

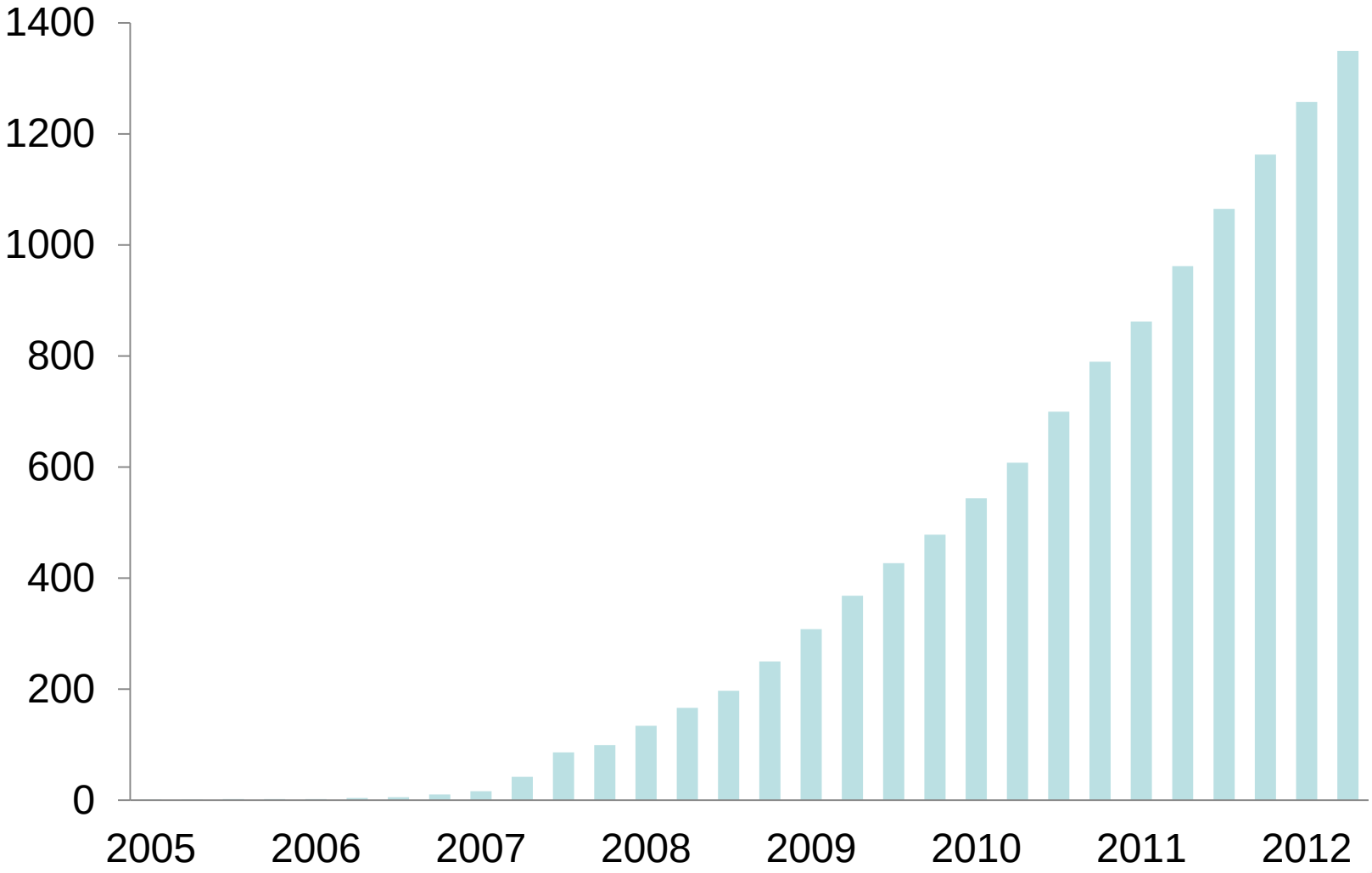
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

Nature, June 2007

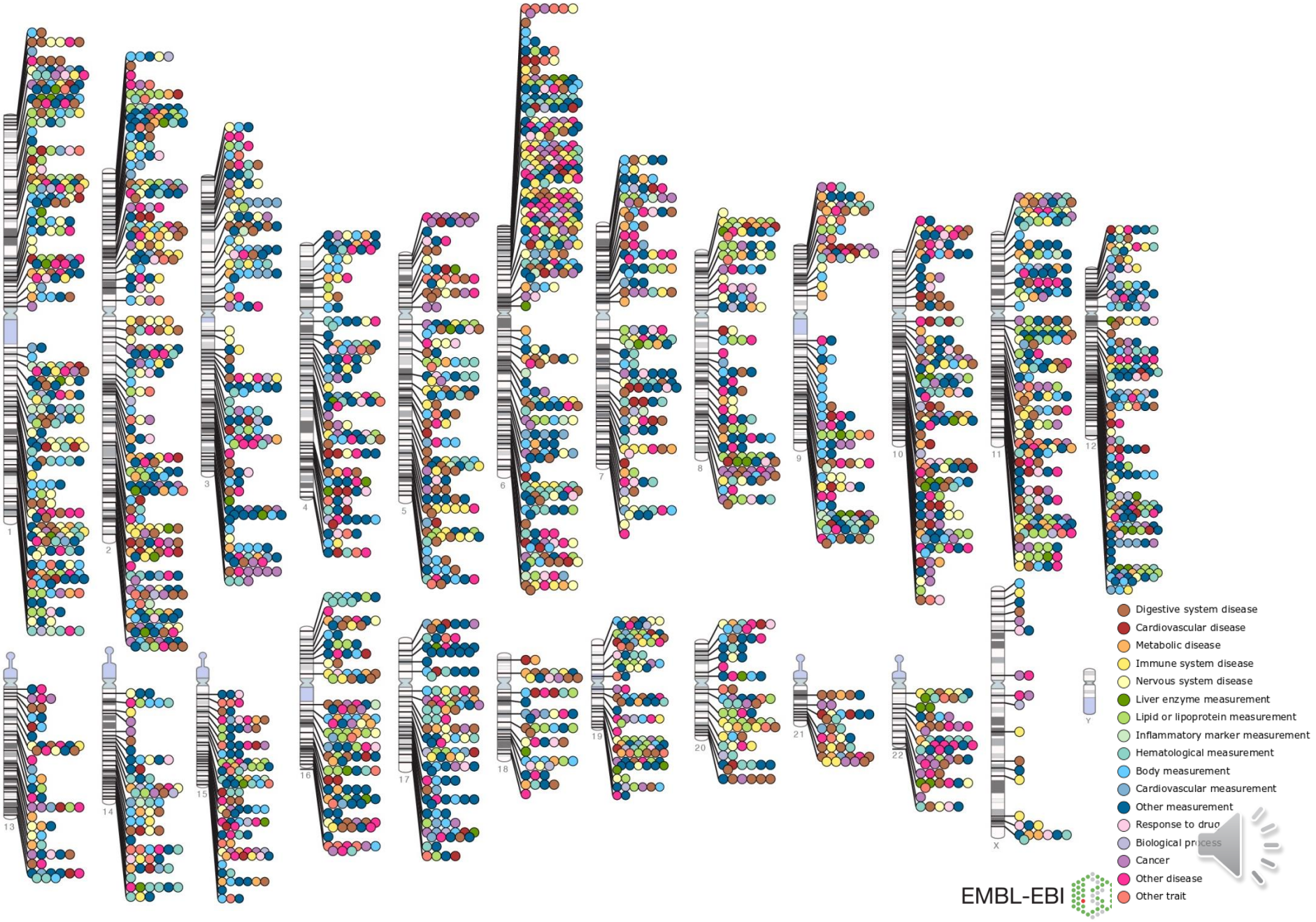
- 2000 patients for 7 diseases, 3000 controls
- 500,000 SNPs analysed
- 24 clear signals
- Small effects
- Replication = gold standard

Published GWA Reports, 2005 – 6/2012

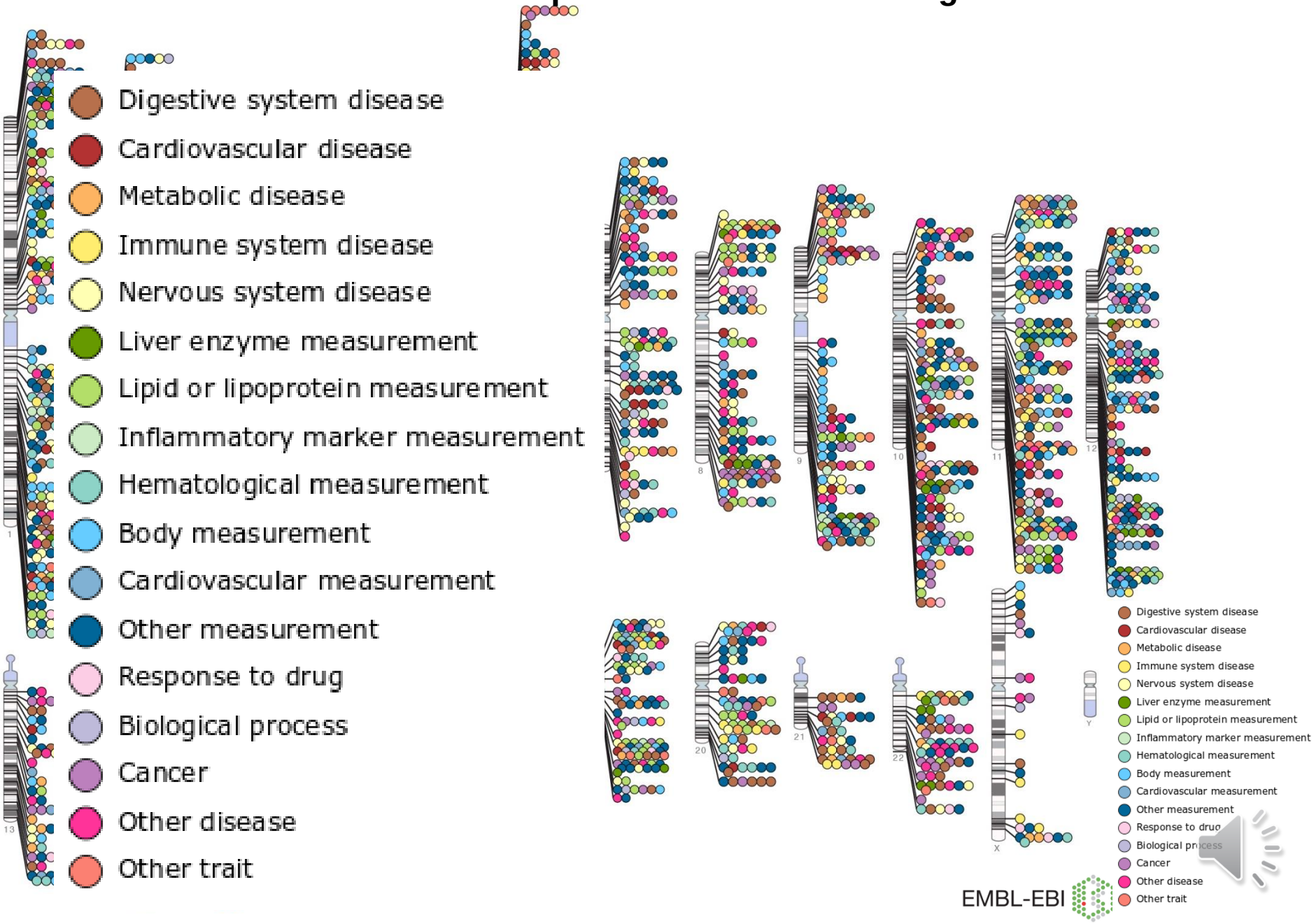


NHGRI GWA Catalog
www.genome.gov/GWAStudies

GWAS catalog (http://www.ebi.ac.uk/gwas)



Published Genome-Wide Associations
Published GWA at p≤5X10^{-8} for 17 trait categories



Body length: 90 years after Fisher

Nature Genetics **40**, 575 - 583 (2008)

Genome-wide association analysis identifies 20 loci that influence adult height

Michael N Weedon, Hana Lango, Cecilia M Lindgren, Chris Wallace, David M Evans, Massimo Mangino, Rachel M Freathy, John R B Perry, Suzanne Stevens, Alistair S Hall, Nilesh J Samani, Beverly Shields, Inga Prokopenko, Martin Farrall, Anna Dominiczak, Diabetes Genetics Initiative, The Wellcome Trust Case Control Consortium, Toby Johnson, Sven Bergmann, Jacques S Beckmann, Peter Vollenweider, Dawn M Waterworth, Vincent Mooser, Colin N A Palmer, Andrew D Morris, Willem H Ouwehand, Cambridge GEM Consortium, Mark Caulfield, Patricia B Munroe, Andrew T Hattersley, Mark I McCarthy & Timothy M Frayling

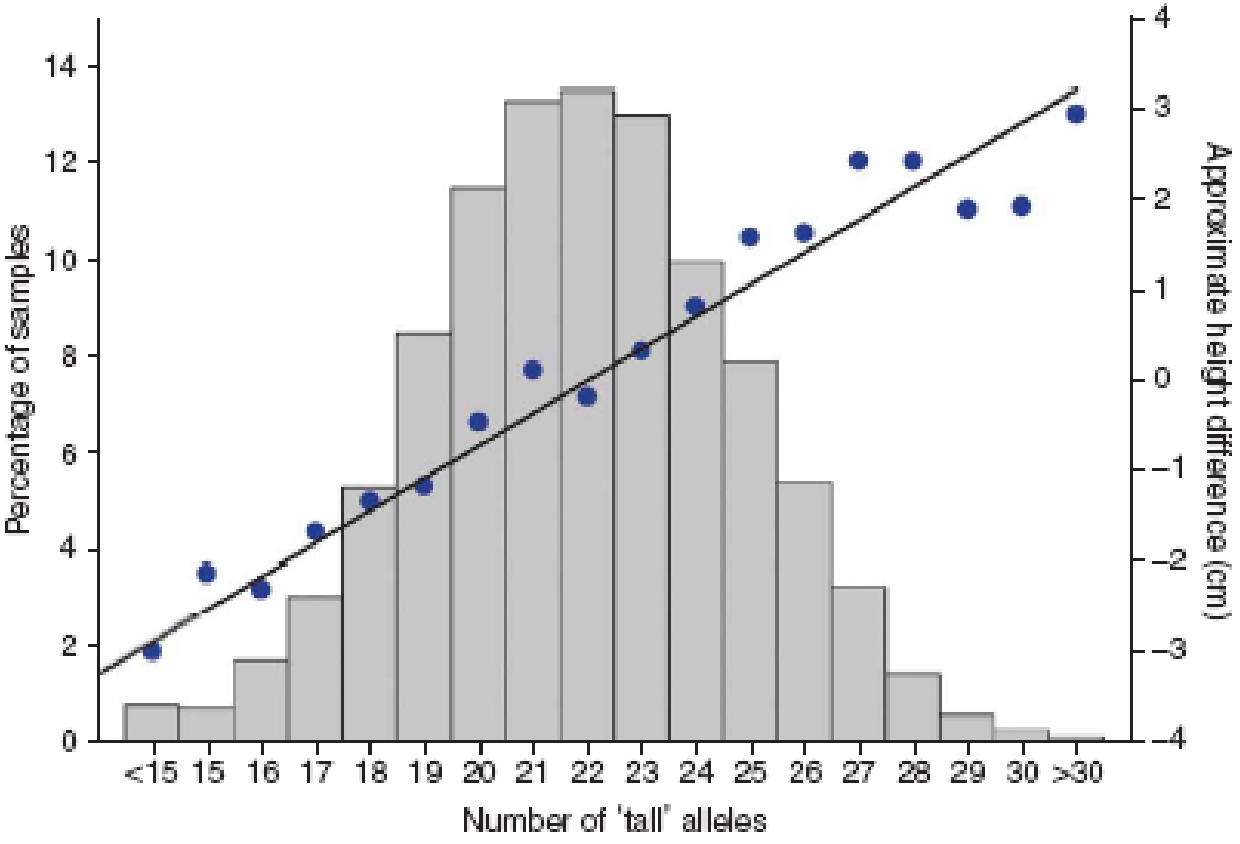
Body length: 90 years after Fisher

- Heritability close to 1
- Weedon et al. (2008): Association tested in GWAS on ~34,000 individuals
- Influenced by 20 genes
  - Each variant has ‘tall’ and ‘small’ allele
  - Body length ~ number of tall alleles
  - 6 cm difference between 15 and 30 tall alleles

Body length

15 tall alleles

30 tall alleles



More recent results on body length

- October 2018
- 700,000 individuals
- 3290 variants at genome-wide significance ($P < 1 \times 10^{-8}$)
  - together explain 24,6% of the heritability for adult height.
- All common variants together capture about 60% of the heritability
- Enriched for genes, pathways and tissue types known to be involved in growth
- Several genes and pathways not previously connected with human skeletal growth

Yengo et al, Hum Mol Genet 27:3641, 2018

Pitfalls of genetic association studies

- Multiple testing
- Missing heritability

Multiple testing

- When is an association “proven”?
- Classical threshold of $p < 0.05$?
- 5% of the test are expected to be significant ($p<0.05$) just by chance
  - Testing 100 SNPs: expect 5 p-values < 0.05 by chance
  - Testing 500K SNPs: expect 25,000 p-values < 0.05 by chance
- Multiple testing leads to increased type I error (α-error, false positive)

Solutions for multiple testing problem

- Adjusting significance level
  - Declare significant if $p < 0.05/\#$ tests (Bonferroni correction)
  - Too strict for GWAS due to dependence of tests (LD)
  - Consensus on GWAS significance threshold of 5×10^{-8}

    (Similar to LOD score genome wide threshold of 3.3)
- Replicate significant findings in independent population

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.

Missing heritability

Disease	Number of loci	Heritability explained
Age-related macular degeneration	5	50%
Crohn's disease	32	20%
Systemic lupus erythematosus	6	15%
Type 2 diabetes	18	6%
HDL cholesterol	7	5.2%
Height	40	5%
Early onset myocardial infarction	9	2.8%
Fasting glucose	4	1.5%

Manolio et al, Nature **461**, 747-753, 2009

Missing heritability

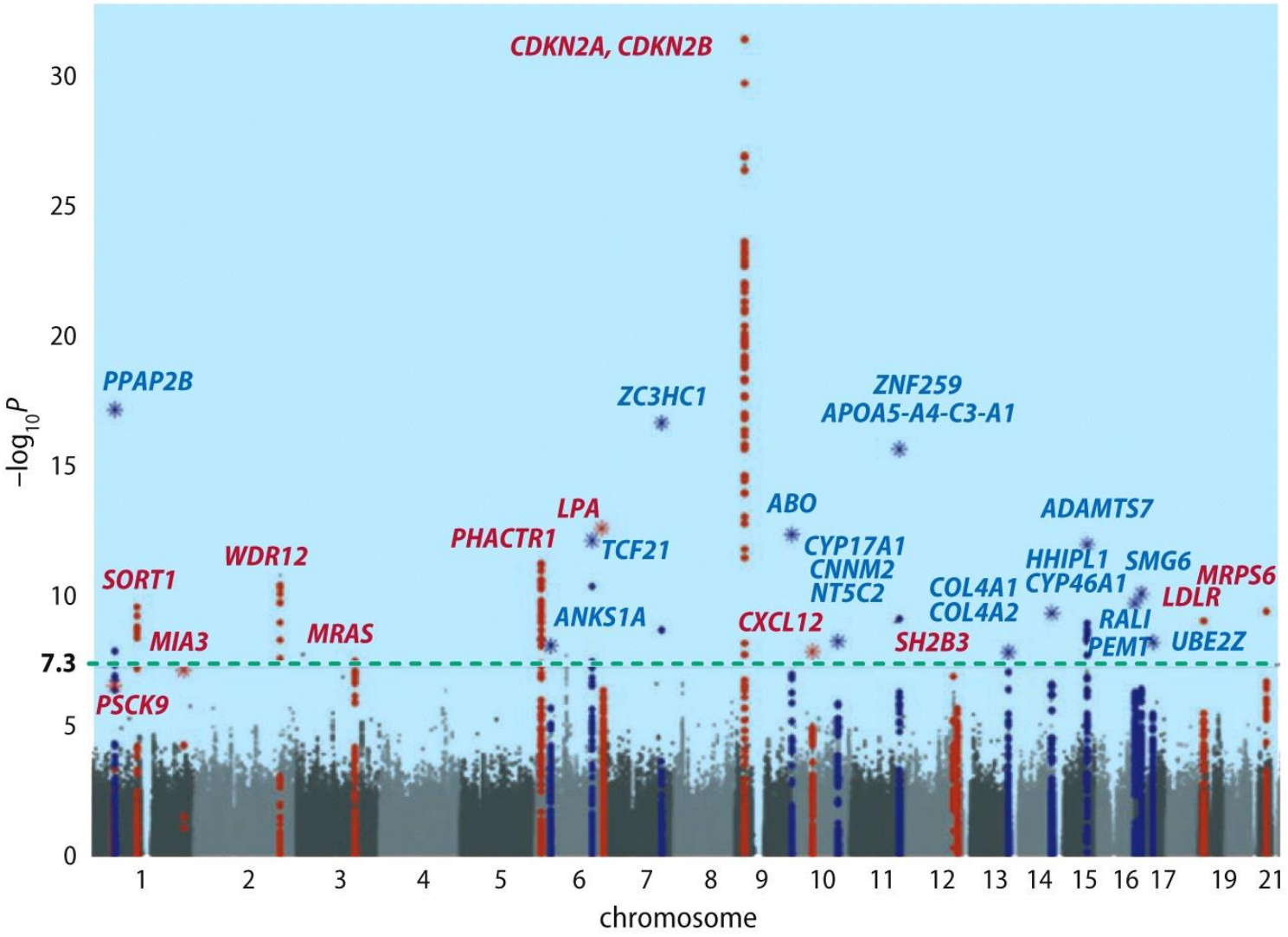
Possible origin:

- Variants (of smaller effect) not reaching significance
- Rare variants
- Gene–gene interactions
- Inadequate accounting for shared environment among relatives

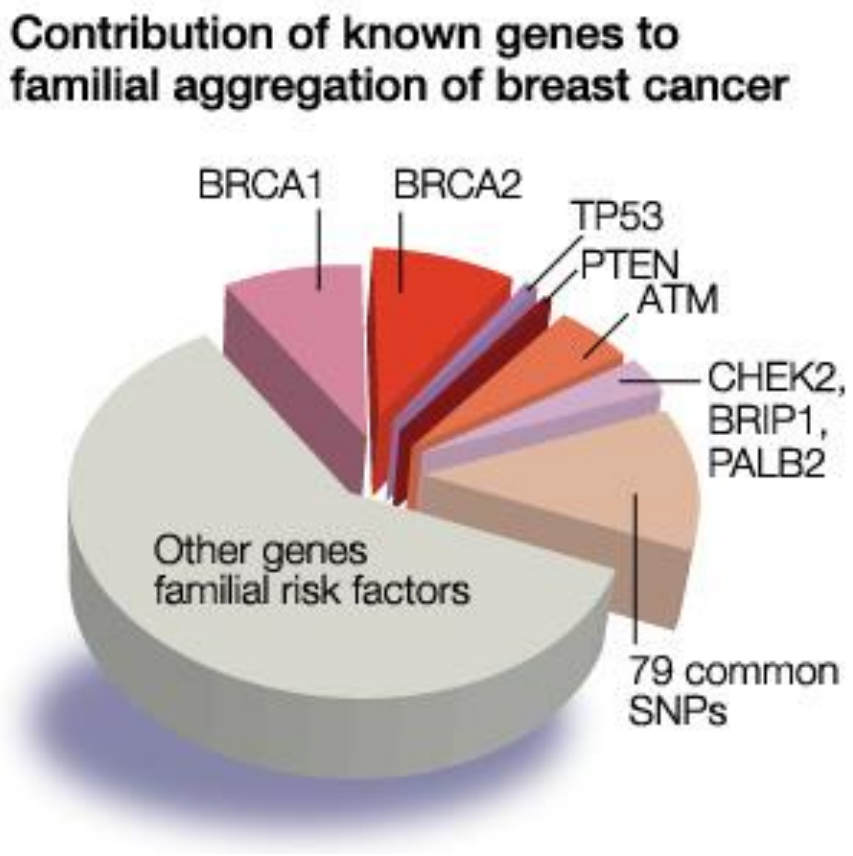
  (Inflated heritability, ghost heritability)
- Structural variants poorly captured by existing microarrays

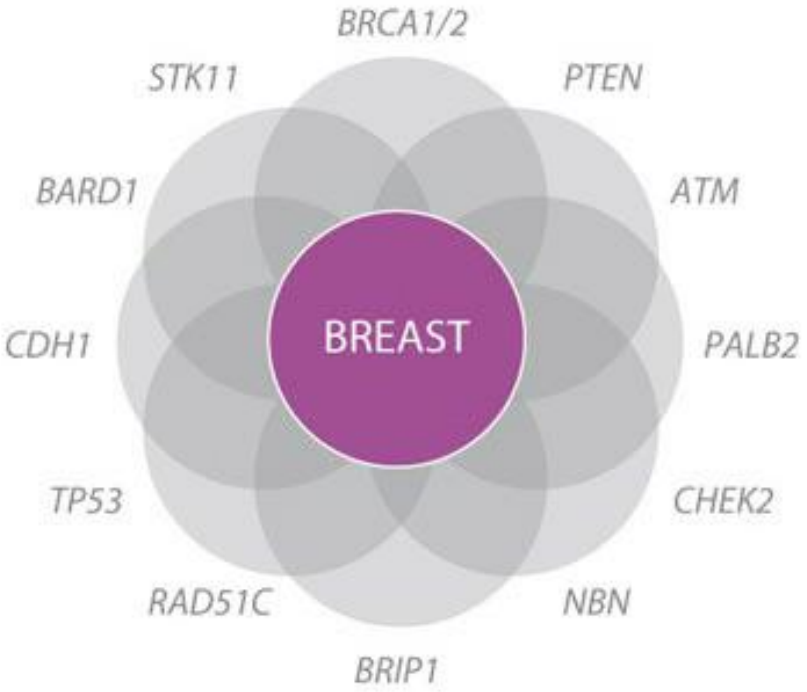
Manhatten plot

Manhatten plot coronary artery disease

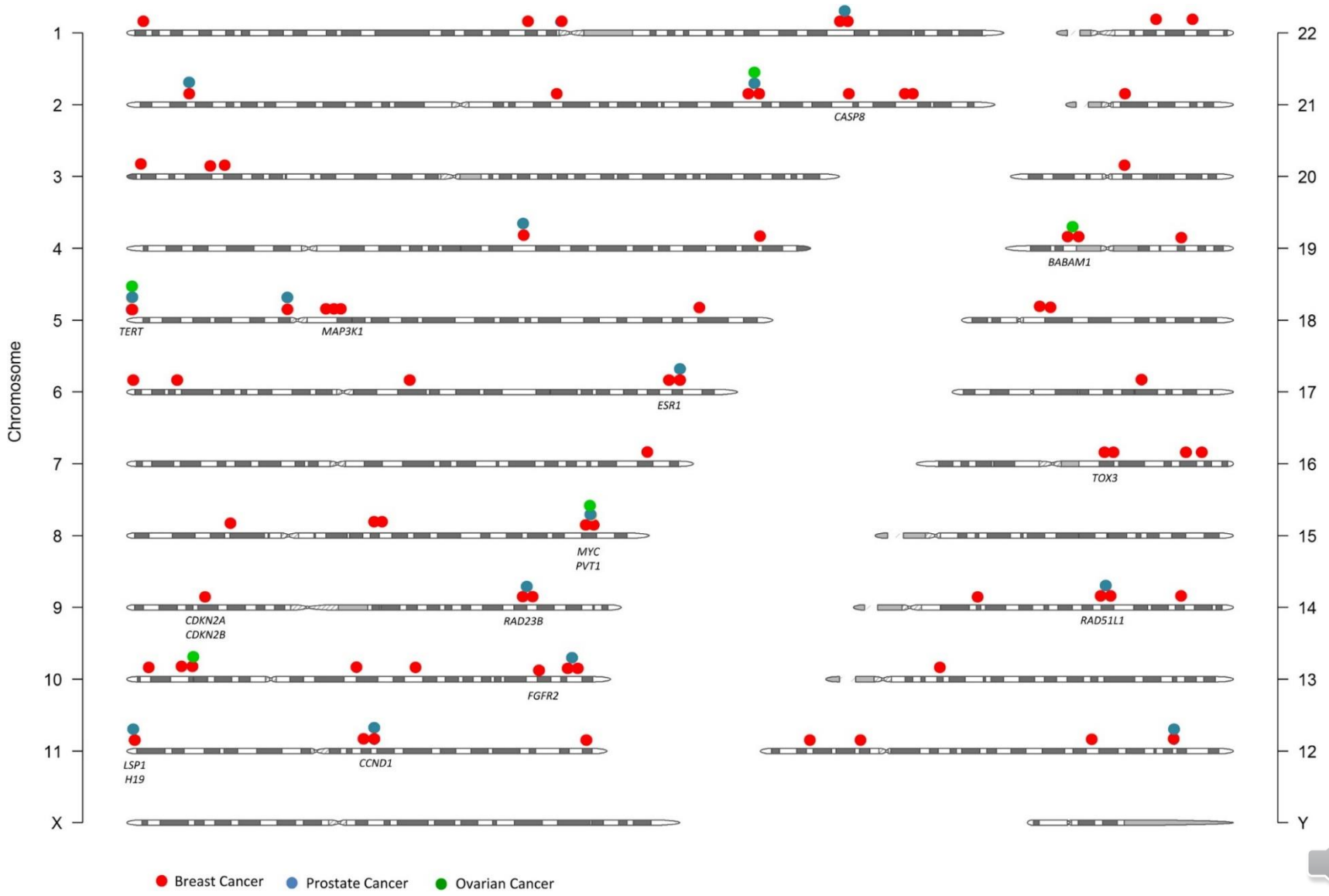


Breast cancer susceptibility



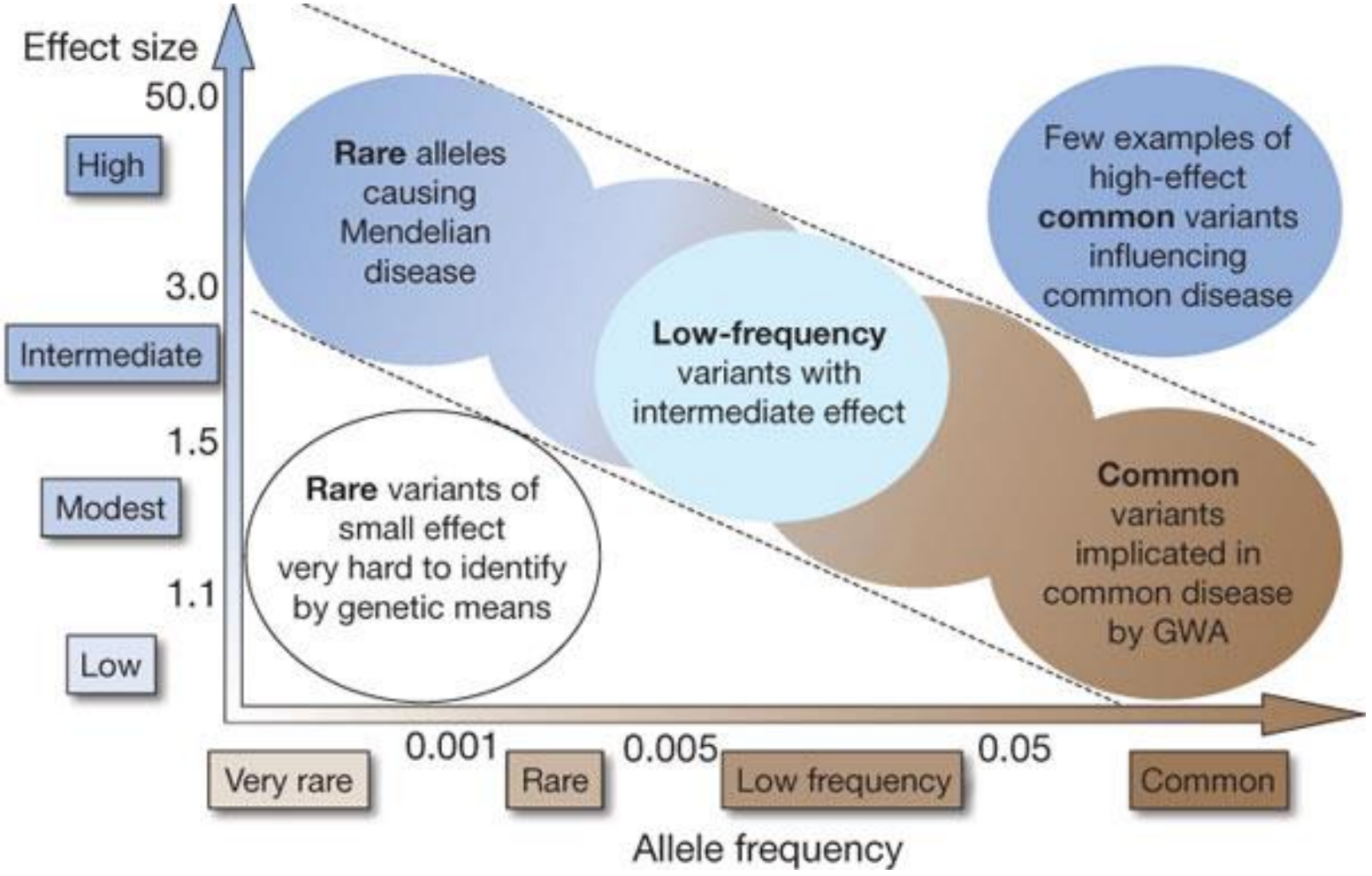


Common breast cancer susceptibility loci



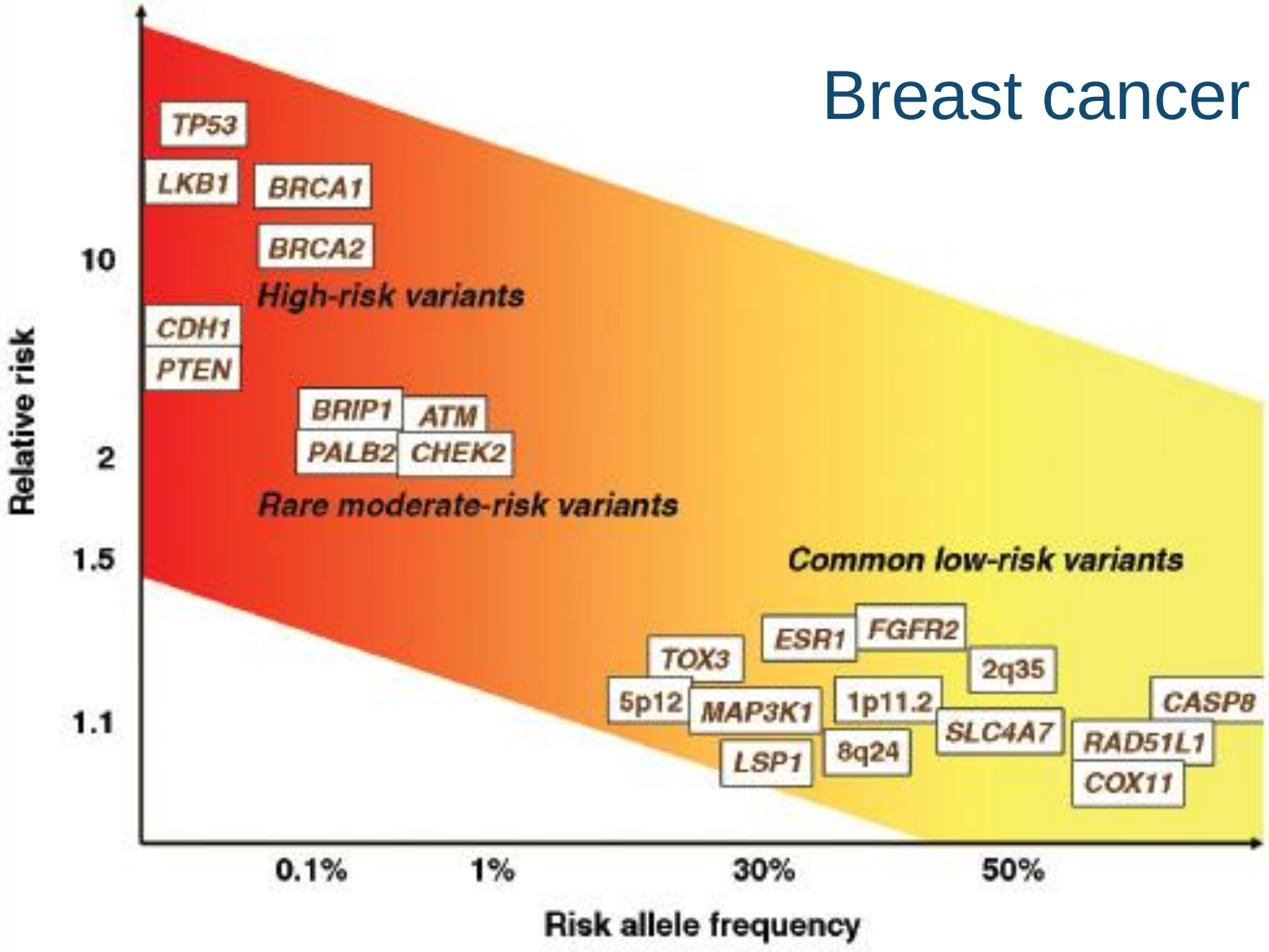
The American Journal of Pathology 2013 183, 1038-1051

odds ratio

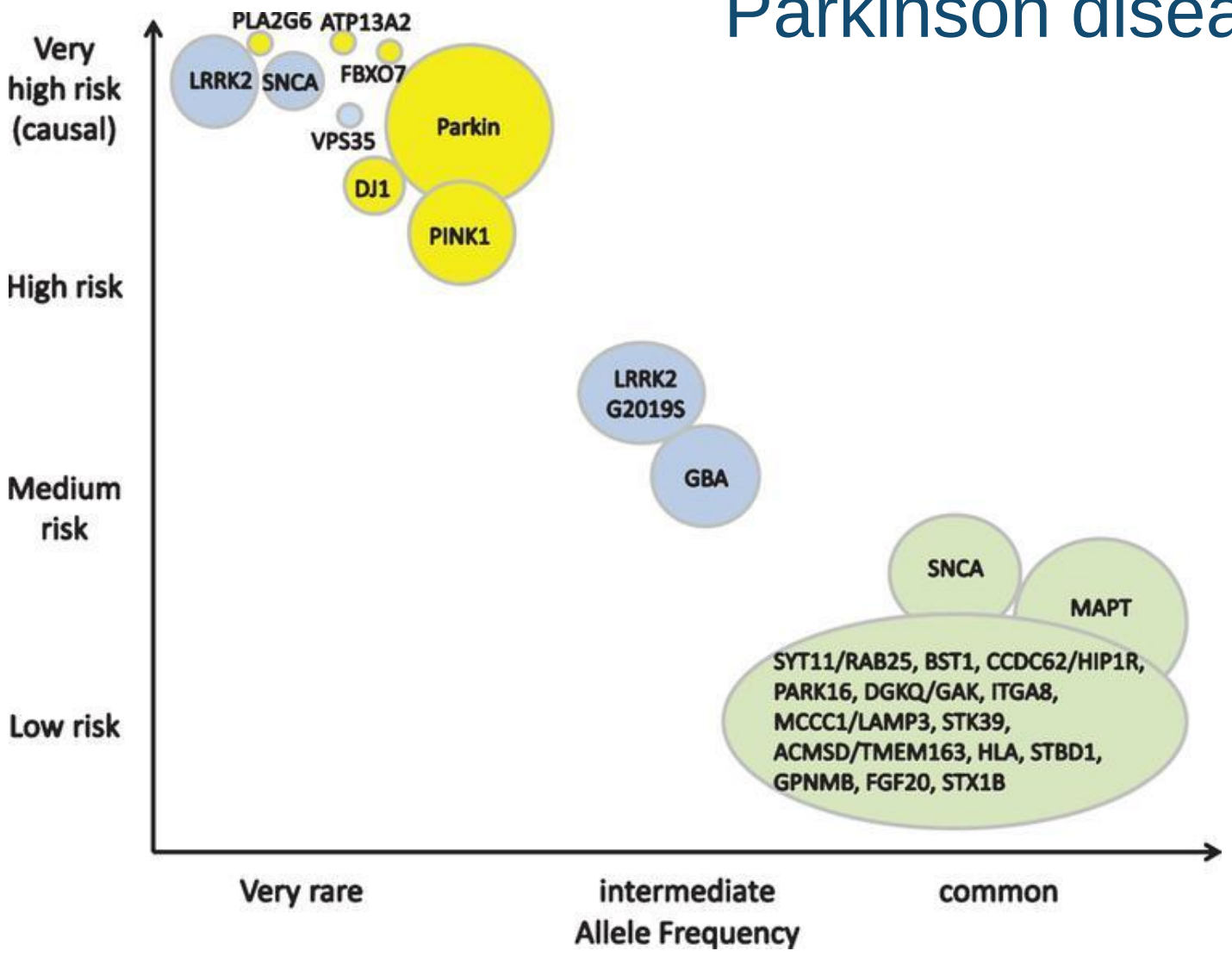


Manolio *et al. Nature* **461**, 747-753 (2009)

Breast cancer

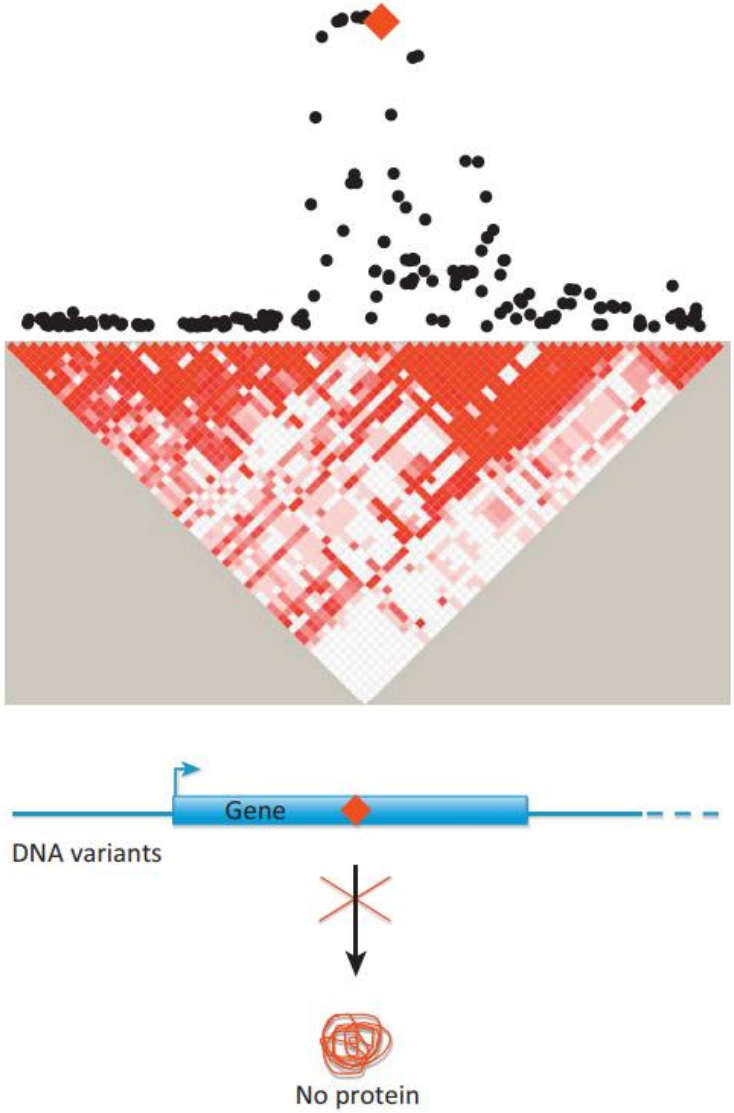


Parkinson disease



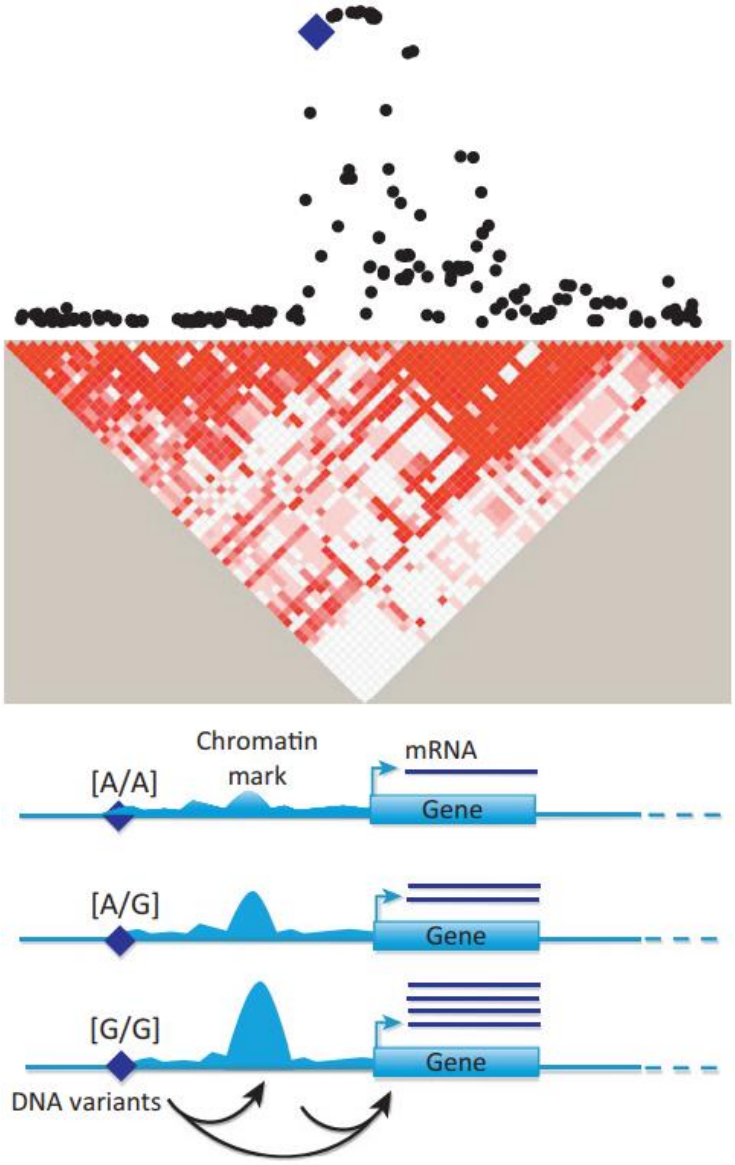
How do genetic variants exert an effect?

- Effect on the protein
- But … many associations are found outside coding regions



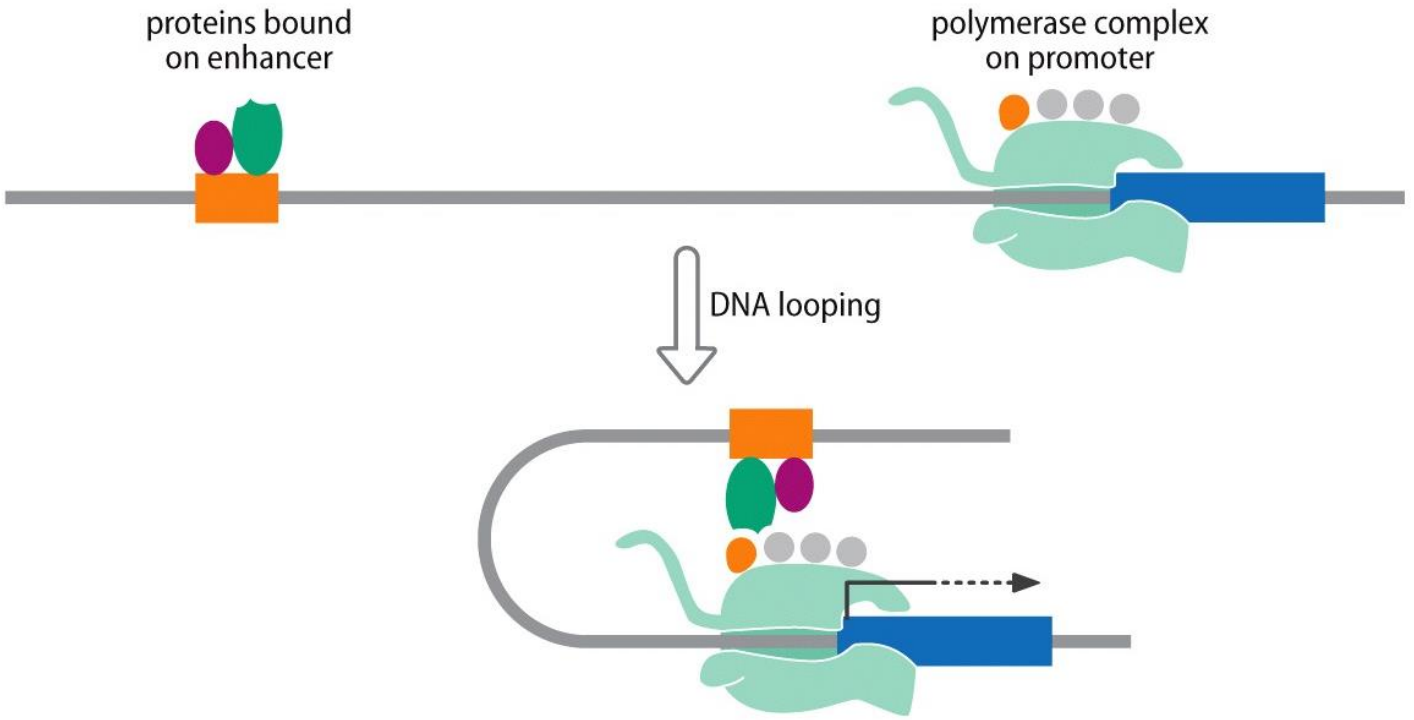
How do genetic variants exert an effect?

- Gene regulation
- TAD domain interactions

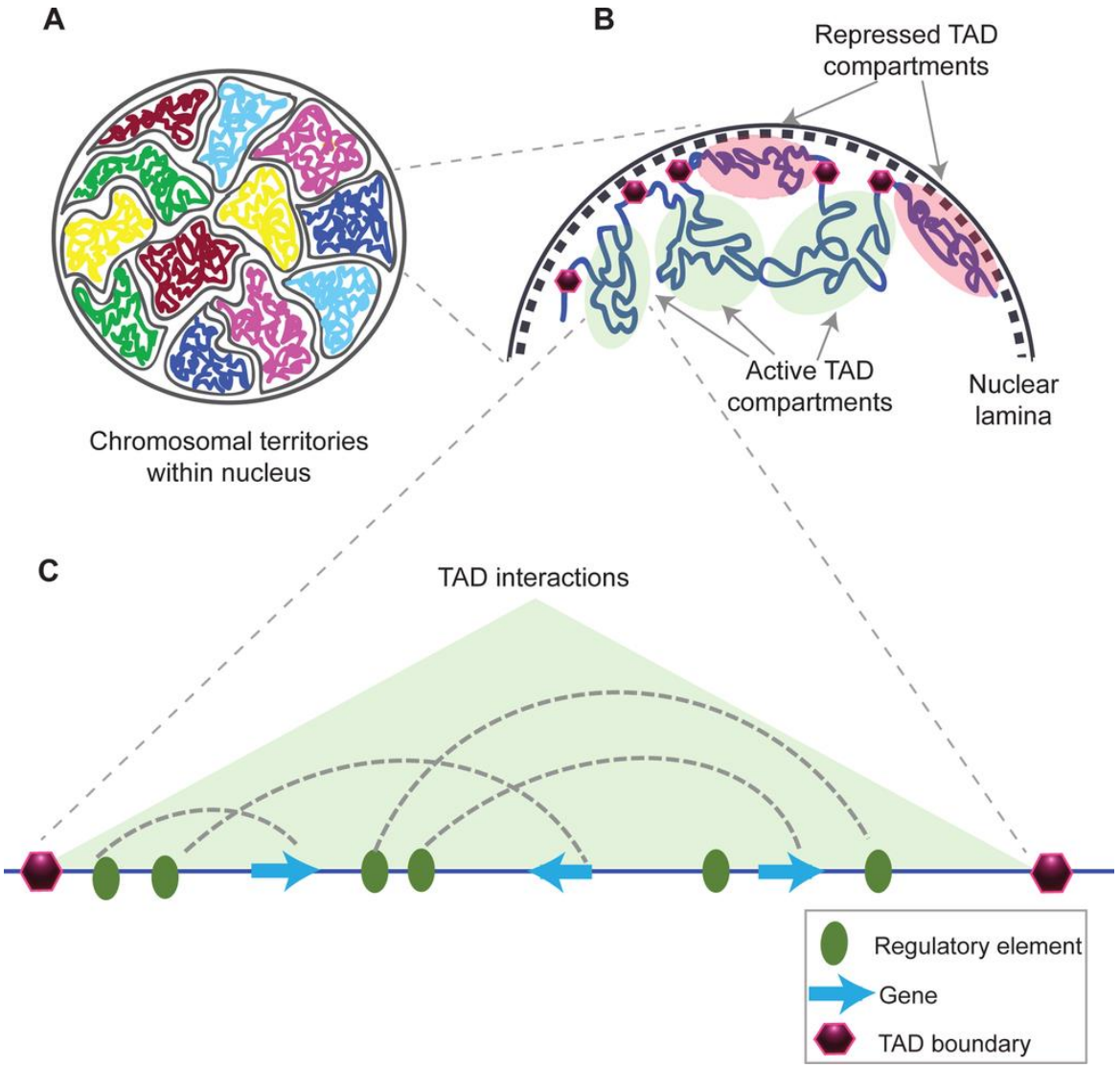


Gene regulation

- DNA binding proteins and lncRNA (trans acting factors)
- Bind DNA sequences (cis acting factors) (enhancers, silencers, promoters, …)

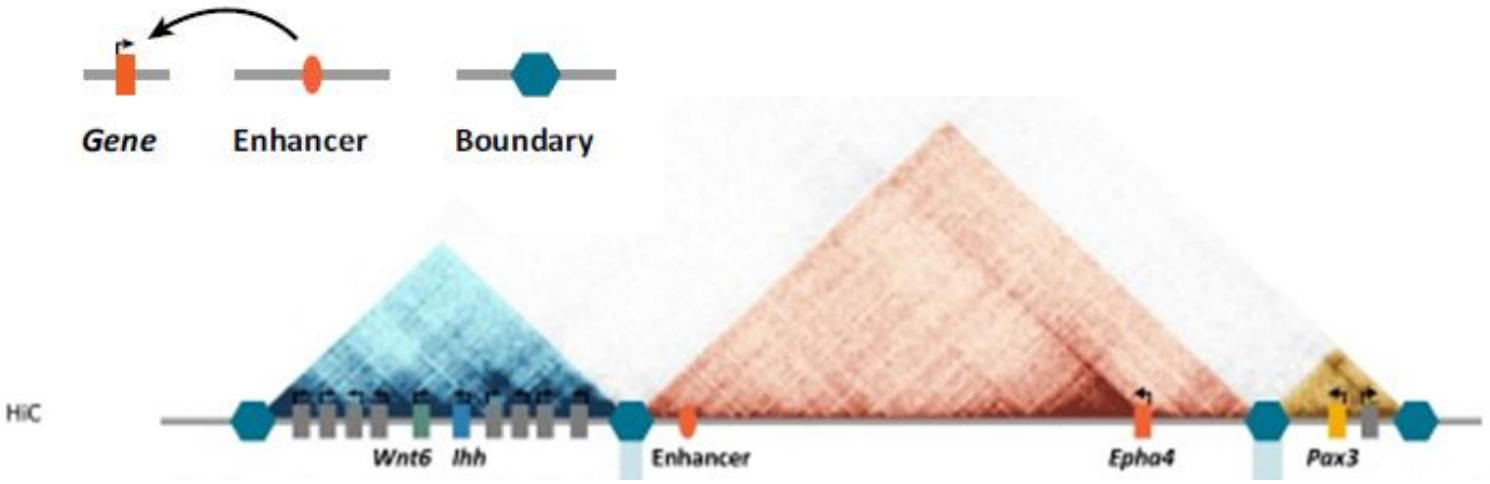


Topologically associating domains (TADs)



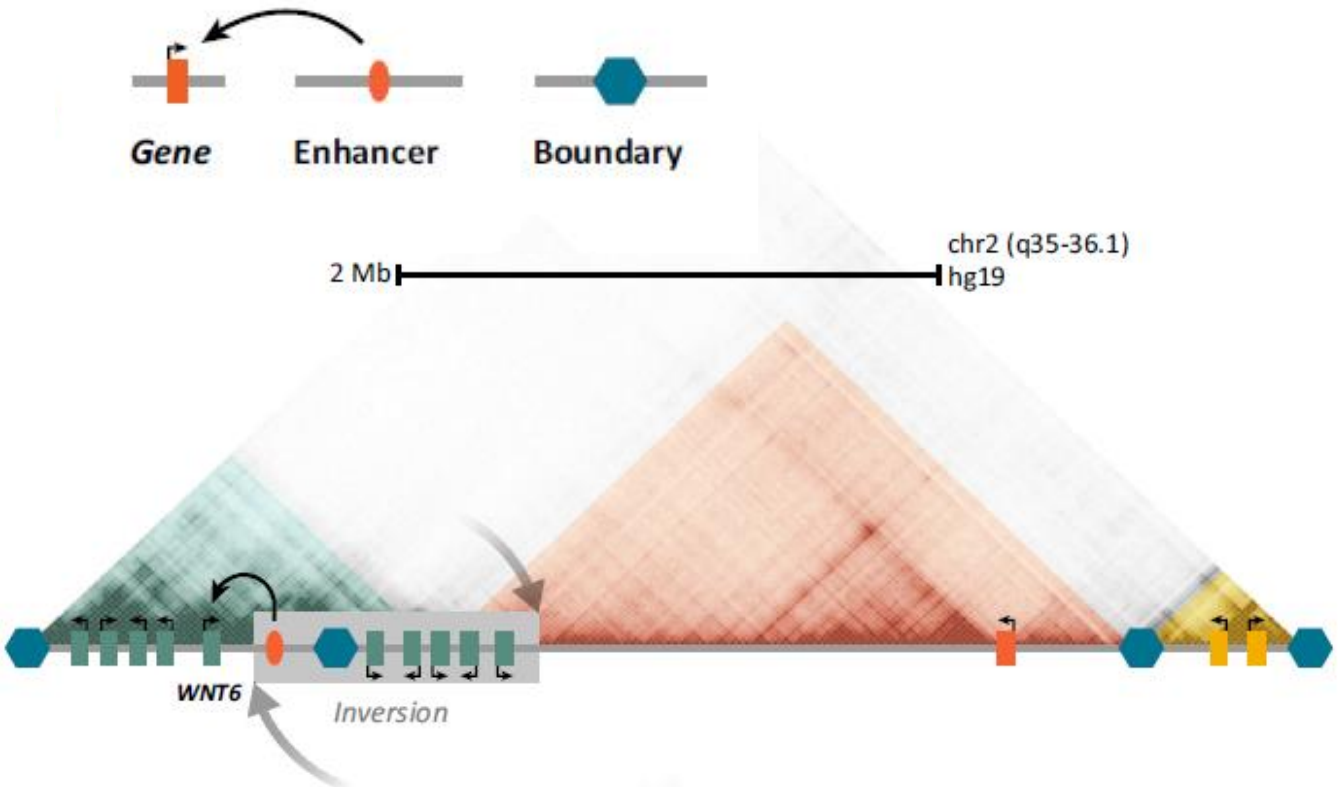
Topologically associating domains (TADs)

- Genomic region that limits promotor enhancer interactions
- Delimited by boundaries
- Evolutionary conserved



Topologically associating domains (TADs)

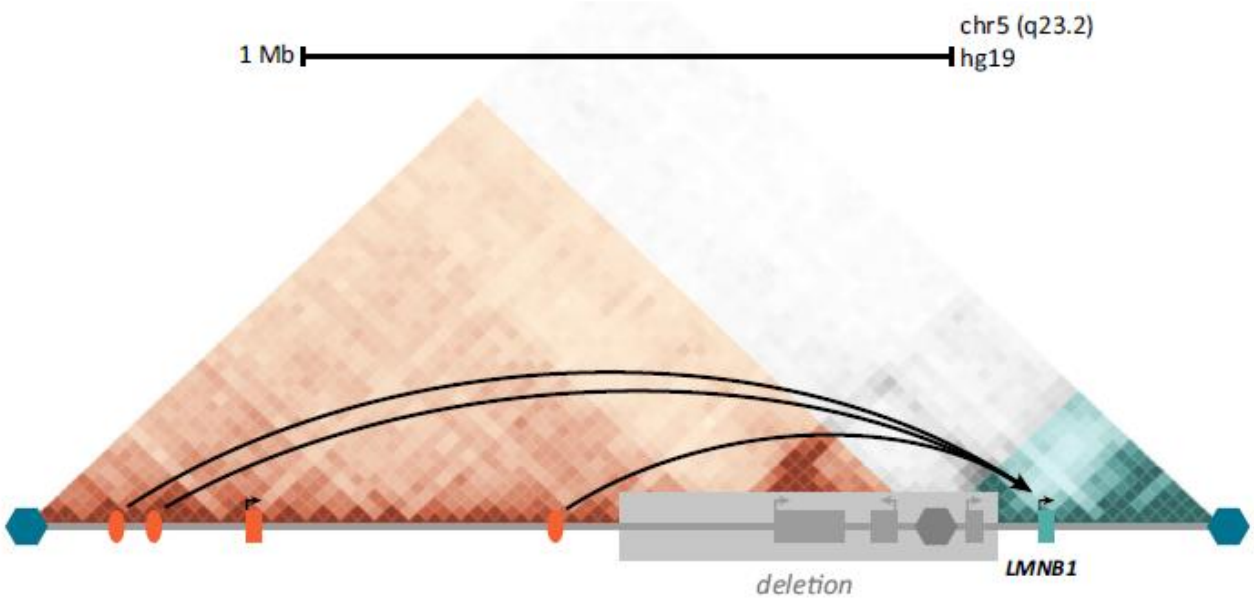
Wrong expression of WNT6 by mislocalisation of enhancers of a neighbouring gene leads to syndactyly

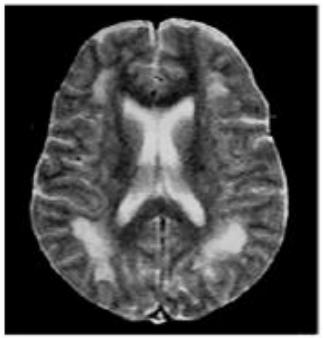


F-syndrome

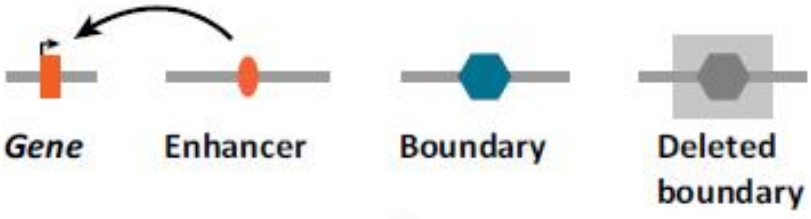
Topologically associating domains (TADs)

Overexpression of LMNB1 because of a deletion of a boundary leads to ADLD





Adult-onset demyelinating leukodystrophy



Gene-environment interactions

Predicted values of high-density lipoprotein cholesterol (HDL-C) for different hepatic lipase (LIPC) genotypes at different total levels of dietary fat intake

Manolio et al, Nat Rev Genet 7:812-820