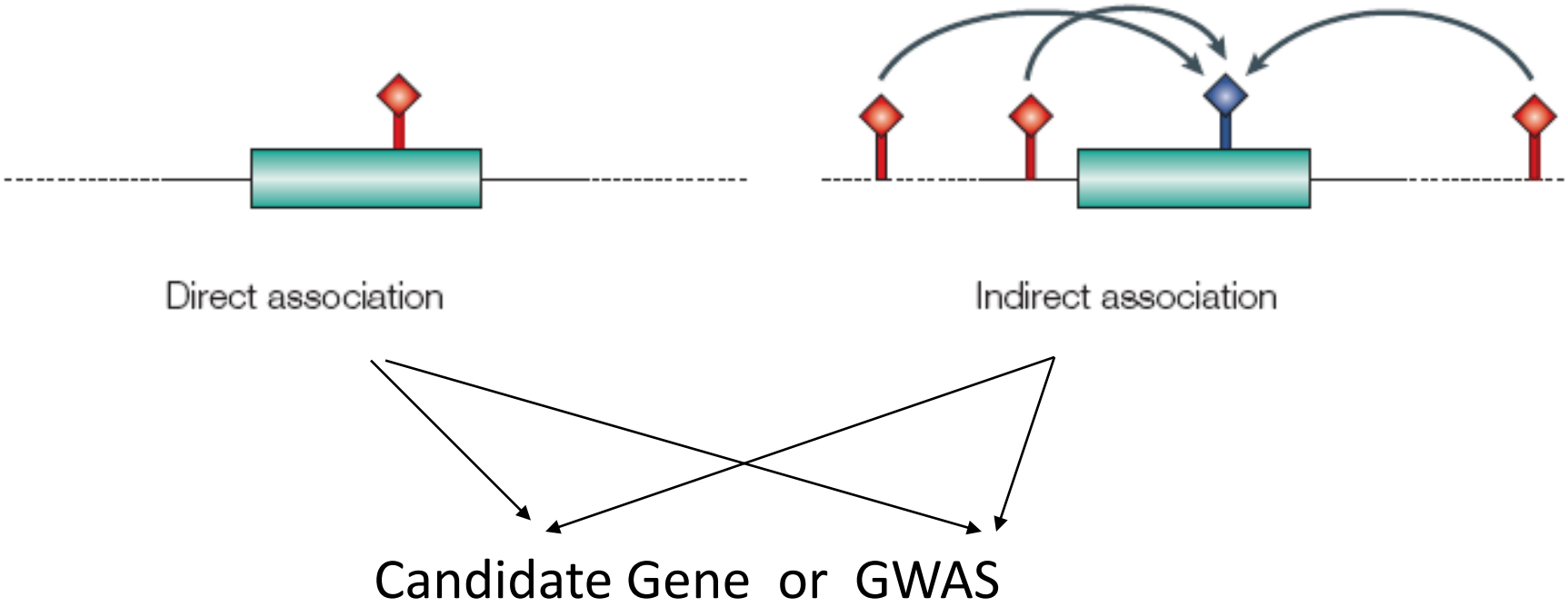
Genome-wide association studies – lessons learned and future directions

Goutam Sahana
Center for Quantitative Genetics and Genomics (QGG)
Dept. Molecular Biology and Genetics
Aarhus University

What is a GWAS?

- Scan markers across genome to find genetic variations associated with a particular phenotype
- A large number of subjects are needed because
  - Effect of the causal variants are expected to be small
  - High level of significance needed to pass multiple testing correction
- Useful for finding genetic variation affecting to quantitative and complex diseases phenotypes

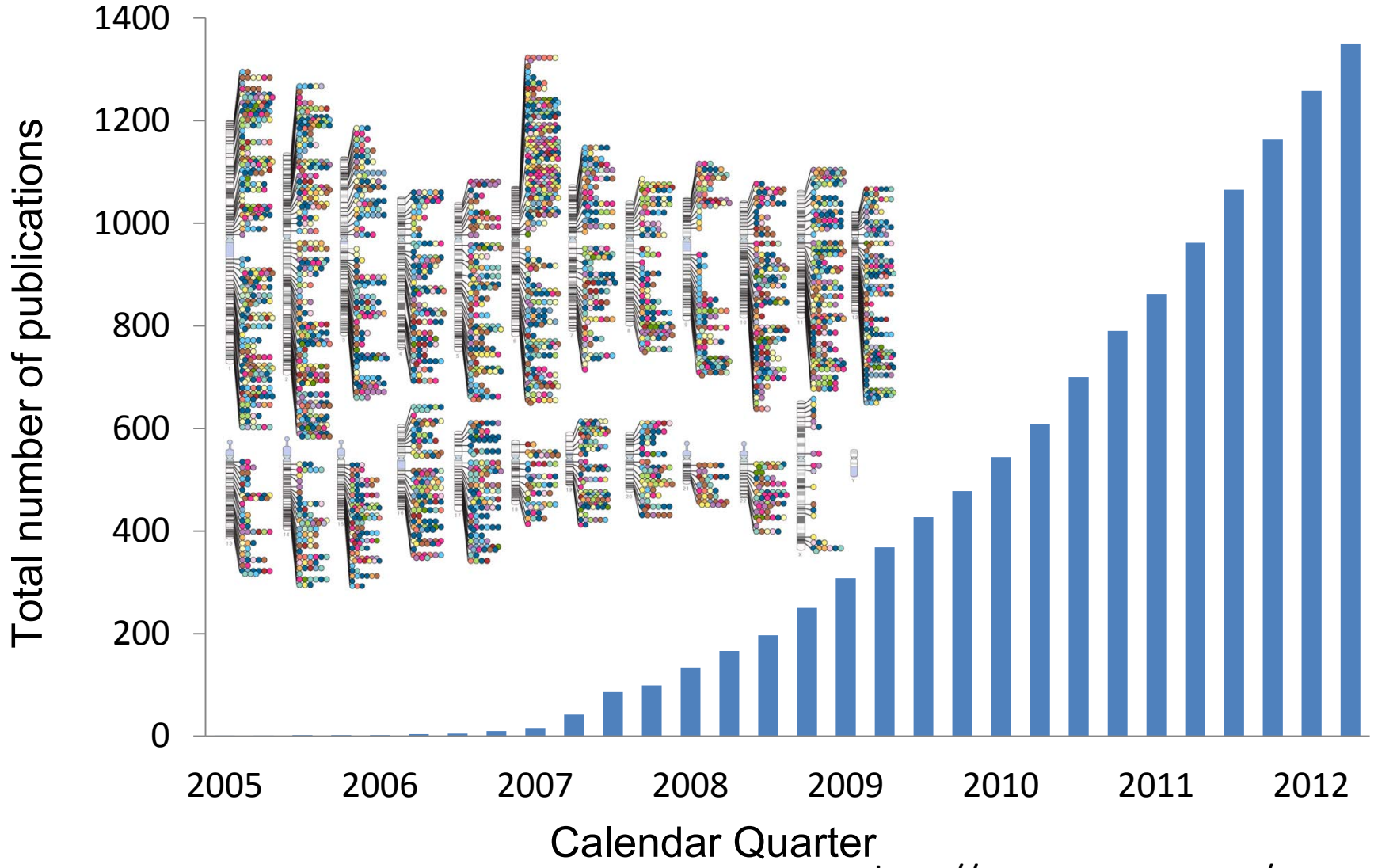
Association Studies



Hirschhorn & Daly, Nat Rev Genet 2005

CENTER FOR QUANTITATIVE GENETICS AND GENOMICS

Published GWA Reports, 2005 – 6/2012



http://www.genome.gov/gwastudies/

QTL / SNP association reported in livestock

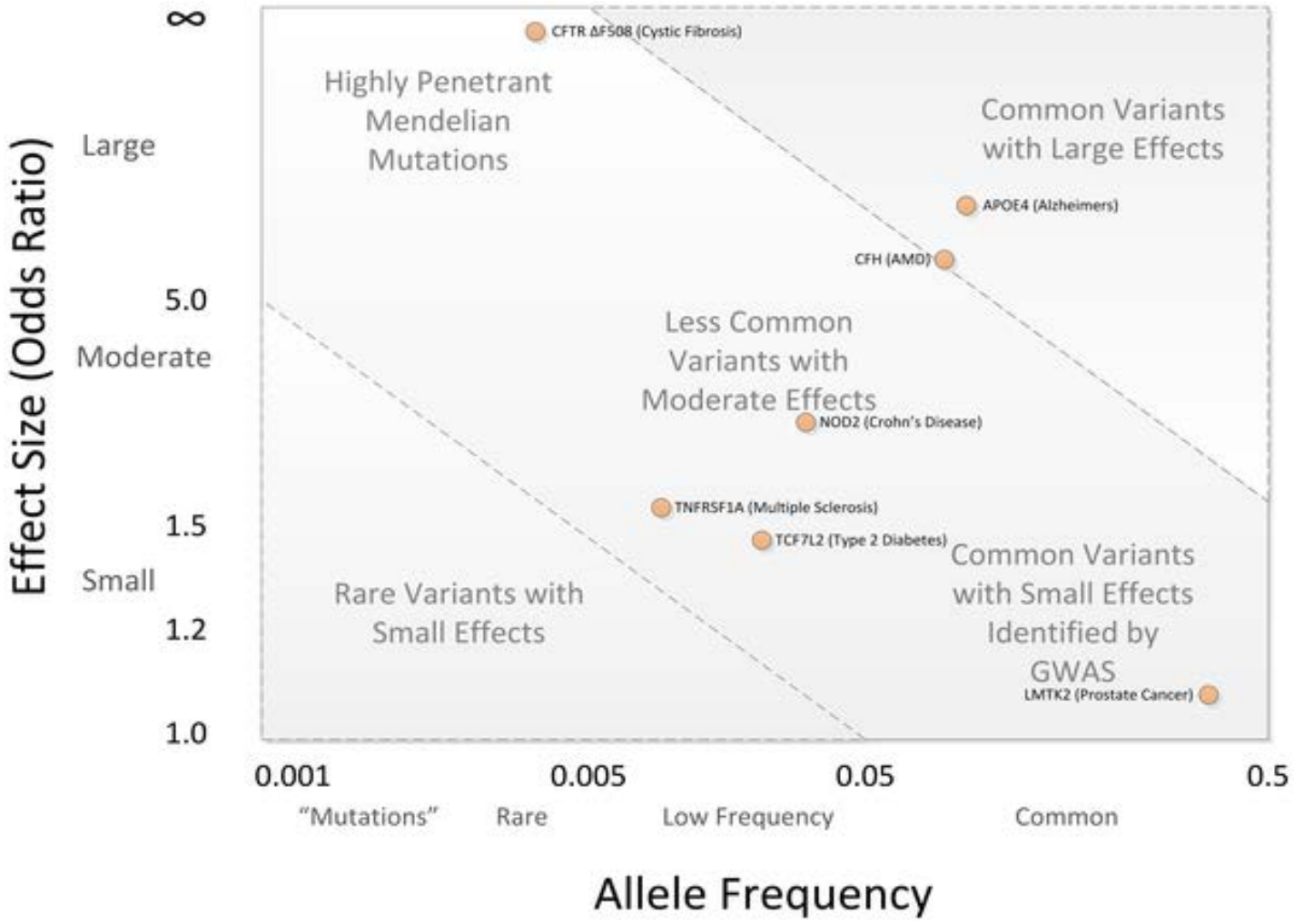
Species	QTLs or SNP associations
Pigs	8402
Cattle	7091
Chicken	3808
Sheep	789
Rainbow trout	127

http://www.animalgenome.org/cgi-bin/QTLdb/

What have we learned from GWAS?

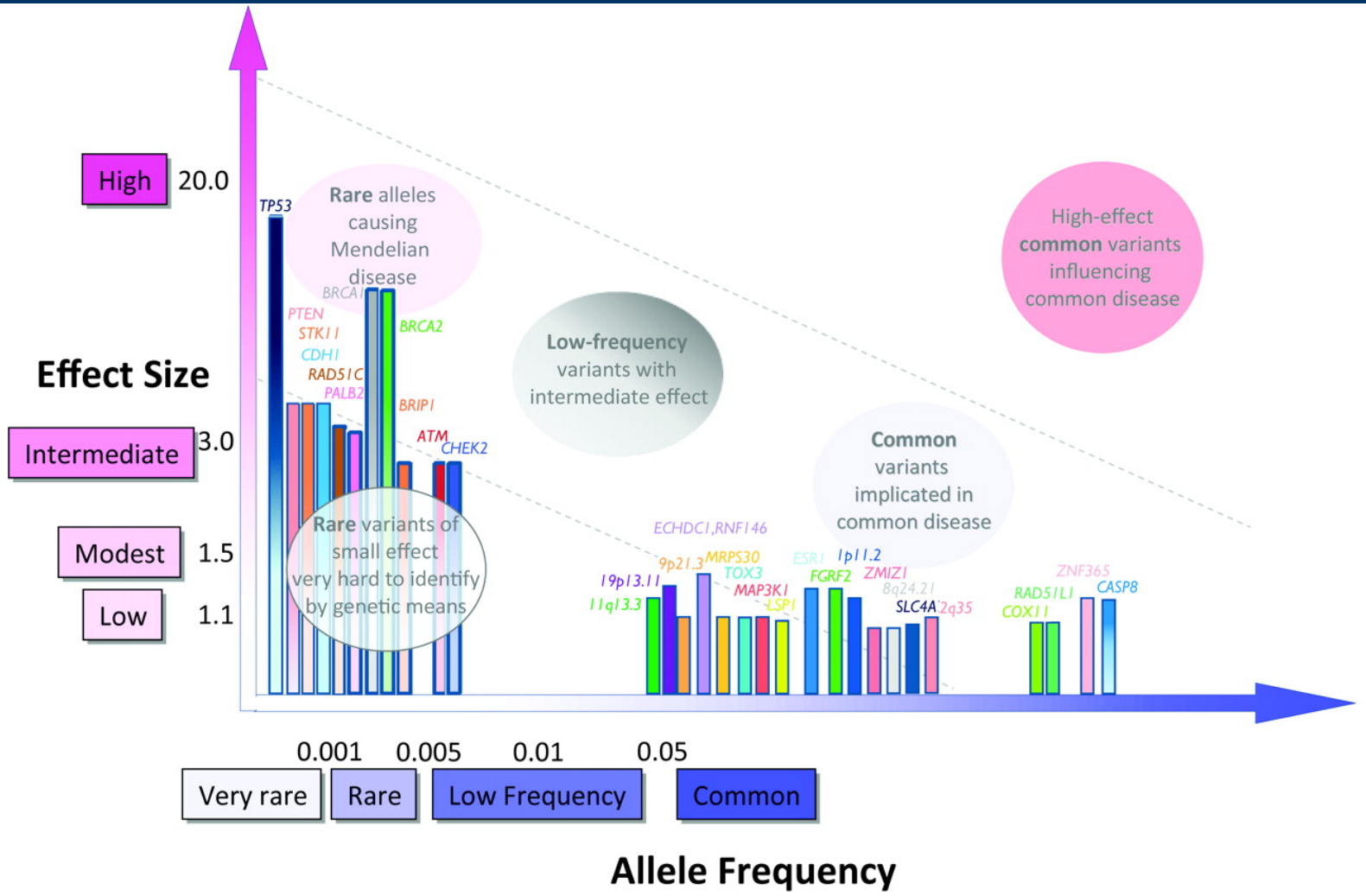
- 100s of trait-associated genetic variants identified by GWAS
- Majority with small effect on the traits
- Tiny part of heritability explained
- Biological mechanism of majority of the associated variants unclear
- New insights into biological pathways controlling complex traits

Spectrum of disease allele effects



Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822.

Allele frequency and effect sizes for genetic variants associated with breast cancer

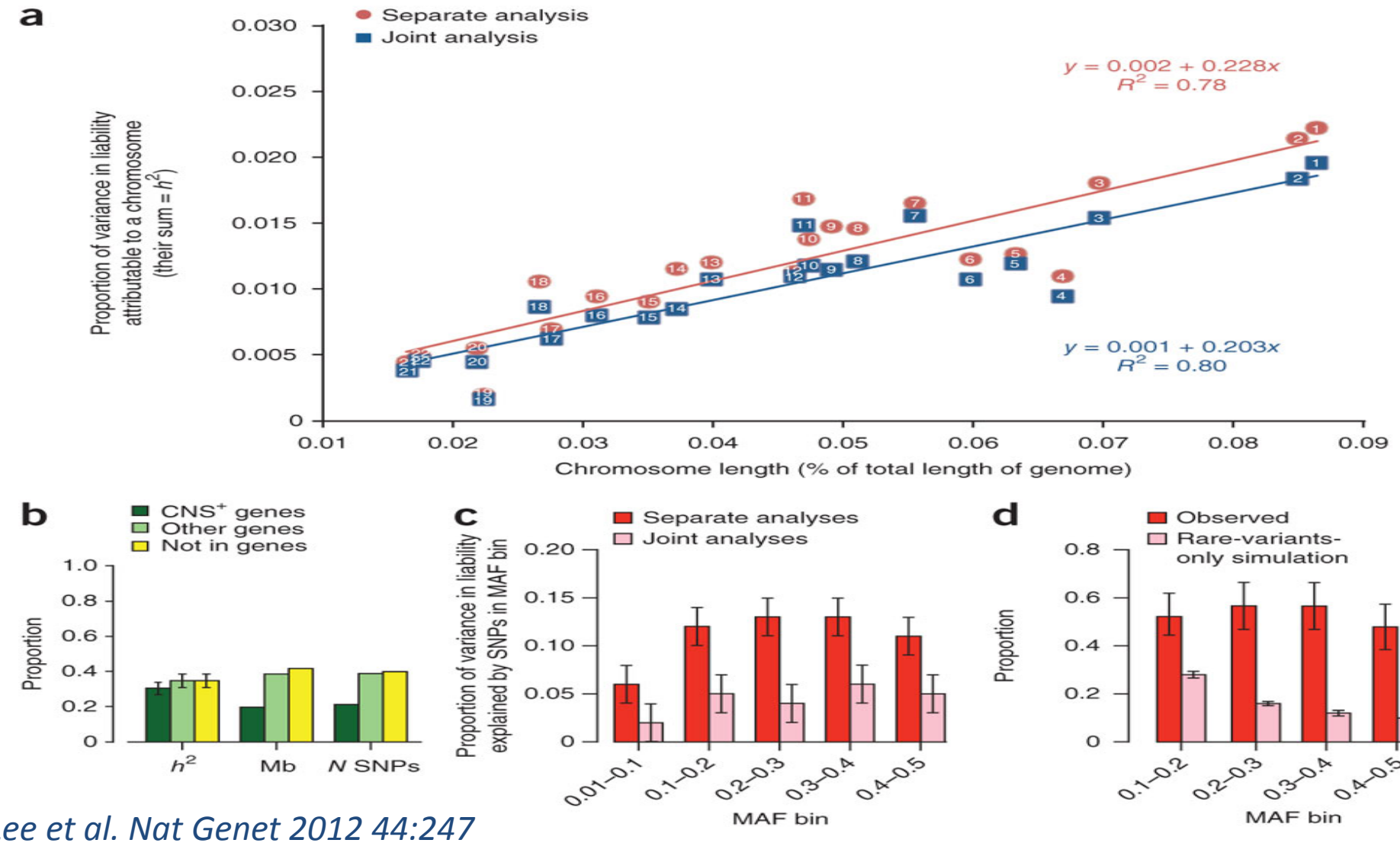


Hindorff L A et al. Carcinogenesis 2011;32:945-954

Missing heritability

- Tiny part of heritability explained
  - 30 loci for type 2 diabetes explain ~10% of heritability
- Is heritability overestimated?
- Disease heterogeneity - lots of different diseases with the same phenotype
- Poor tagging of rare mutations of large effect (including CNVs)
- Statistical modeling

Proportion of variation in susceptibility to schizophrenia captured by common SNPs

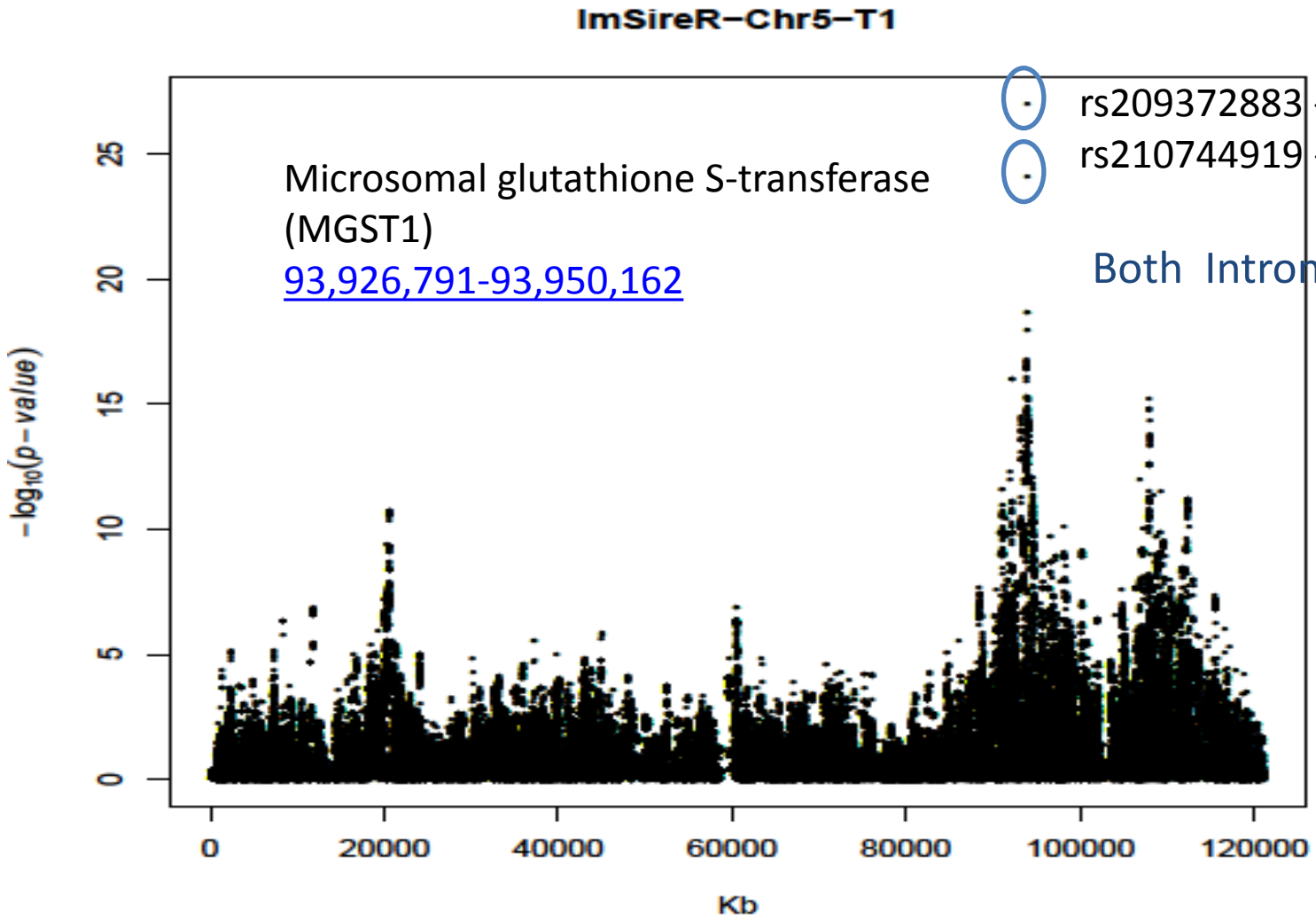


Lee et al. Nat Genet 2012 44:247

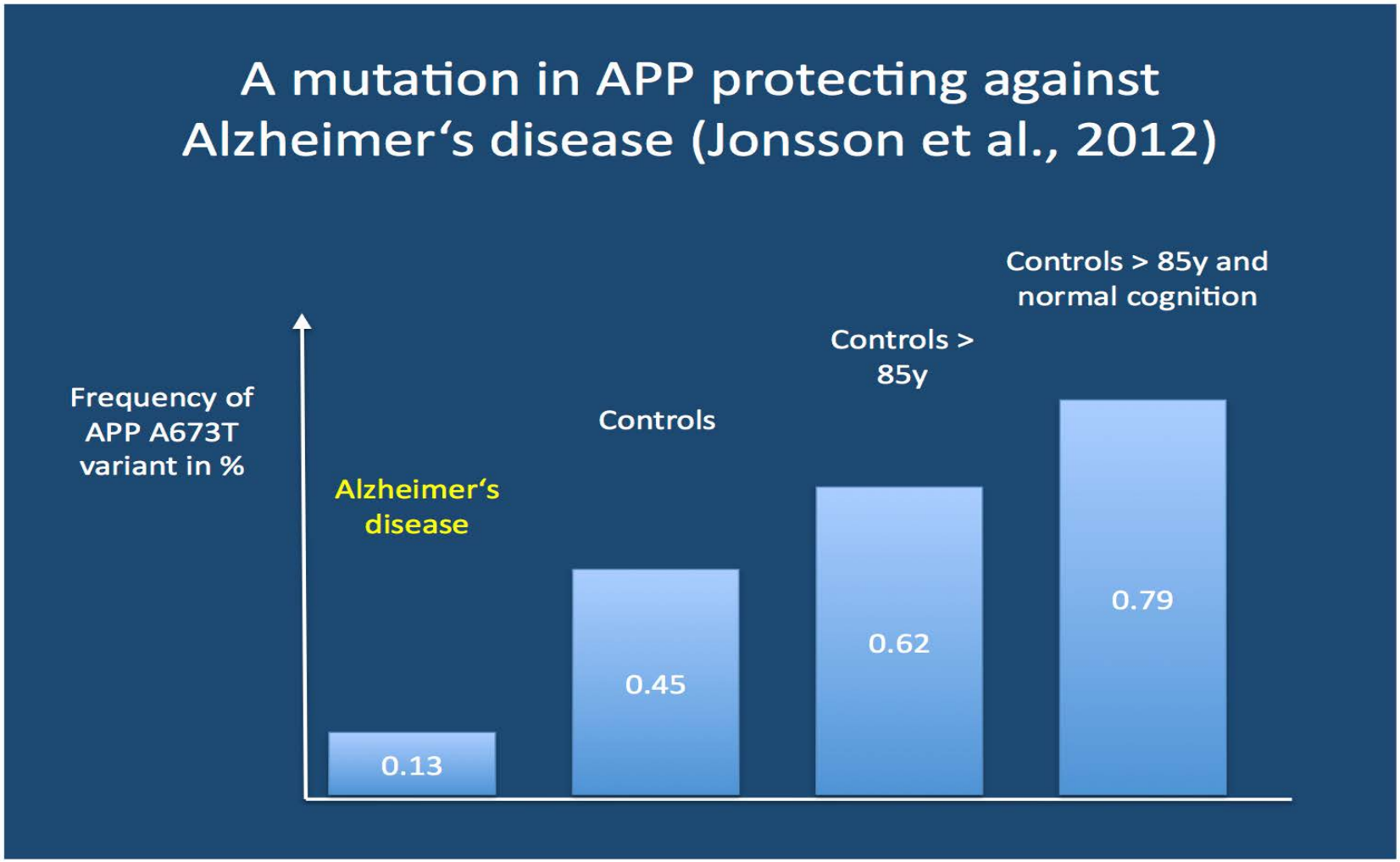
Biological effect behind variants remains unclear

- ~30% of associated variations inter-genic
  - ENCODE: ‘80% of the genome has biochemical function’
- Many within-gene variations have no known function
- LD obscures the location of specific causative loci
  - Reduced ability to identify function
  - Pinpointing the exact causal variant in the genome remains a major challenge

Strongest association with intronic variants

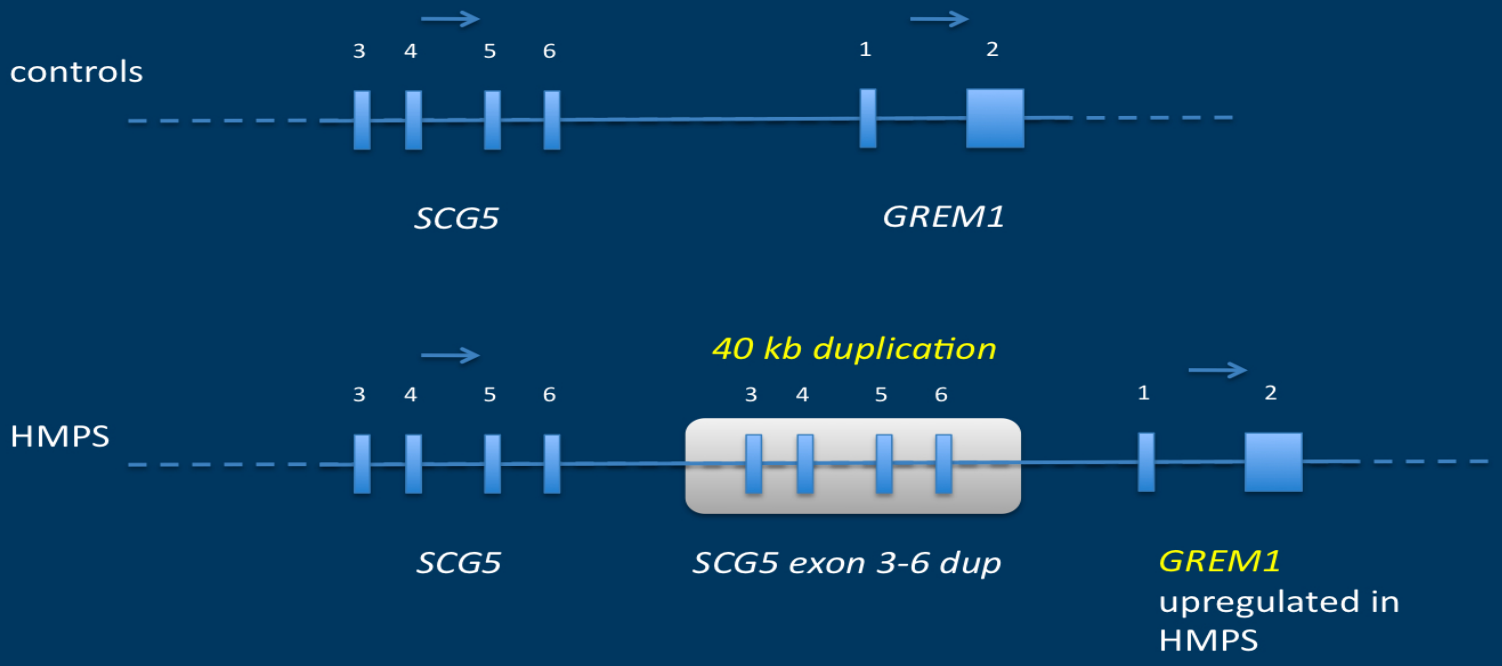


Genetic heterogeneity

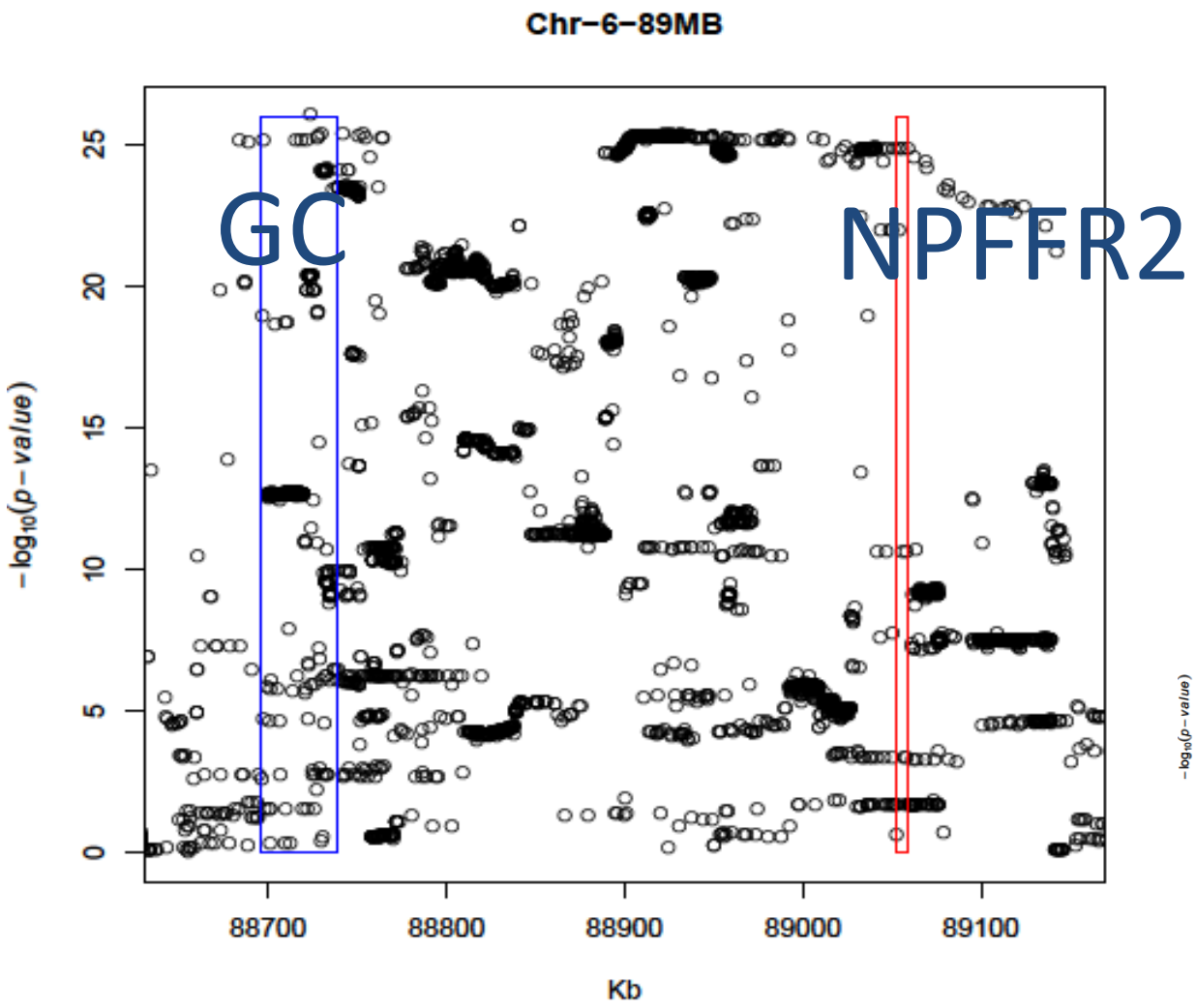


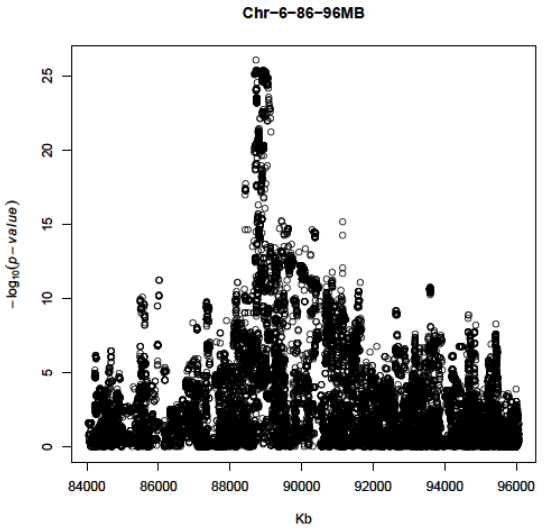
Gene action due to duplication in another gene

A 40 kB duplication causes hereditary mixed polyposis syndrome (Jaeger et al., 2012)

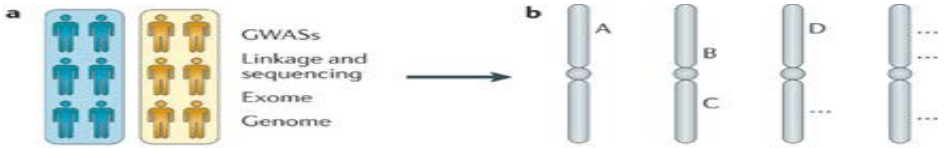


LD obscures the location of causative mutation

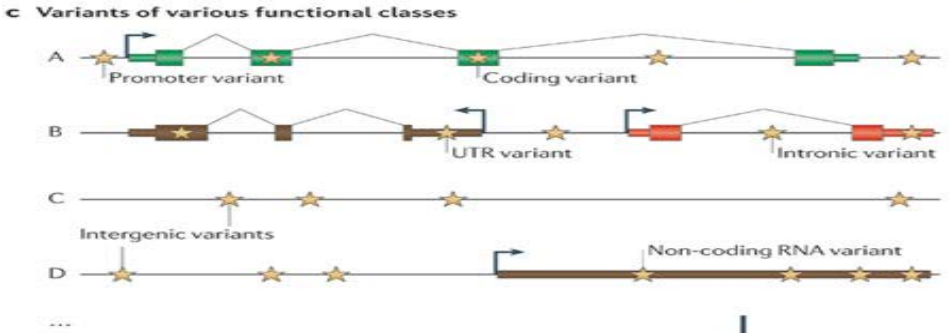




How to prioritize candidate variants?



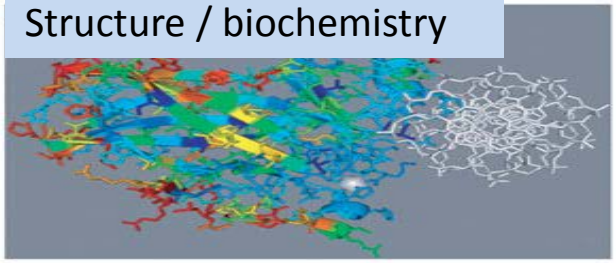
GWAS



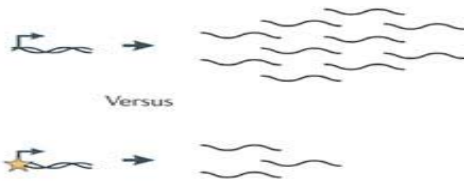
Functional classes

Comparative Genomics

Structure / biochemistry



Functional study



Cooper & Shendure, 2011
Nature Reviews | Genetics

Tsunami of ‘new’ data

- Improvement in genotyping array technology
- Greater access to low-cost sequencing
- ‘New’ technologies
  - Chromatin immuno-precipitation (ChIP) with NGS (ChIP-seq)
  - Gene expression
  - Whole exome sequencing
  - RNA-seq
  - Omics data

Challenges ahead

- Structural variants poorly tagged by current SNP-chips
- Current CNV arrays only detect large variants;
  - No systematic coverage of the vast number of small CNVs (including microsatellites)
- Merging massive amount of data (WGS, omics data, phenotypes, environment etc.)
- Greatest challenge will be to deciphering functional mechanism and clinical relevance

Conclusions

- Important advance towards deciphering genetic basis of complex trait through GWAS despite limitations
- 'Difficult' types of genetic variations become mappable through scientific and technological advances
- Explore biological meaning by combining genome-wide and knowledge-based approaches