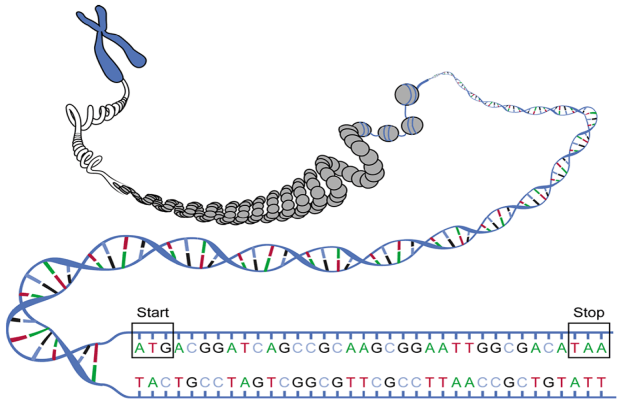
Use of full Genome Sequence information in fine mapping of QTL

Goutam Sahana
Centre for Quantitative Genetics and Genomics (QGG)
Aarhus University



Outline

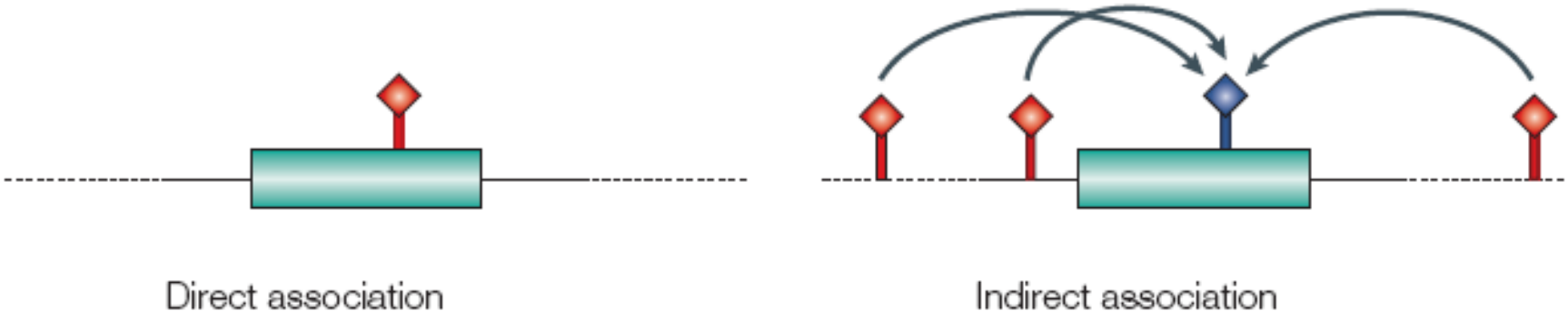
- Gene mapping with SNP array
- What we have learnt?
- Gene mapping using sequence data
- Discuss if it is a game changer

Quantitative traits

- Phenotypic variation for quantitative traits results from the segregation of alleles at multiple loci
- QTL – Quantitative trait loci
- QTL effect depends on
  - Genotype
  - Environment
  - Their interactions
- Major challenge is to map the molecular polymorphisms responsible for the variation in quantitative traits

Genome-wide association studies (GWAS)

- An approach that scans markers across genome to find genetic variations associated with a particular quantitative trait or complex disease

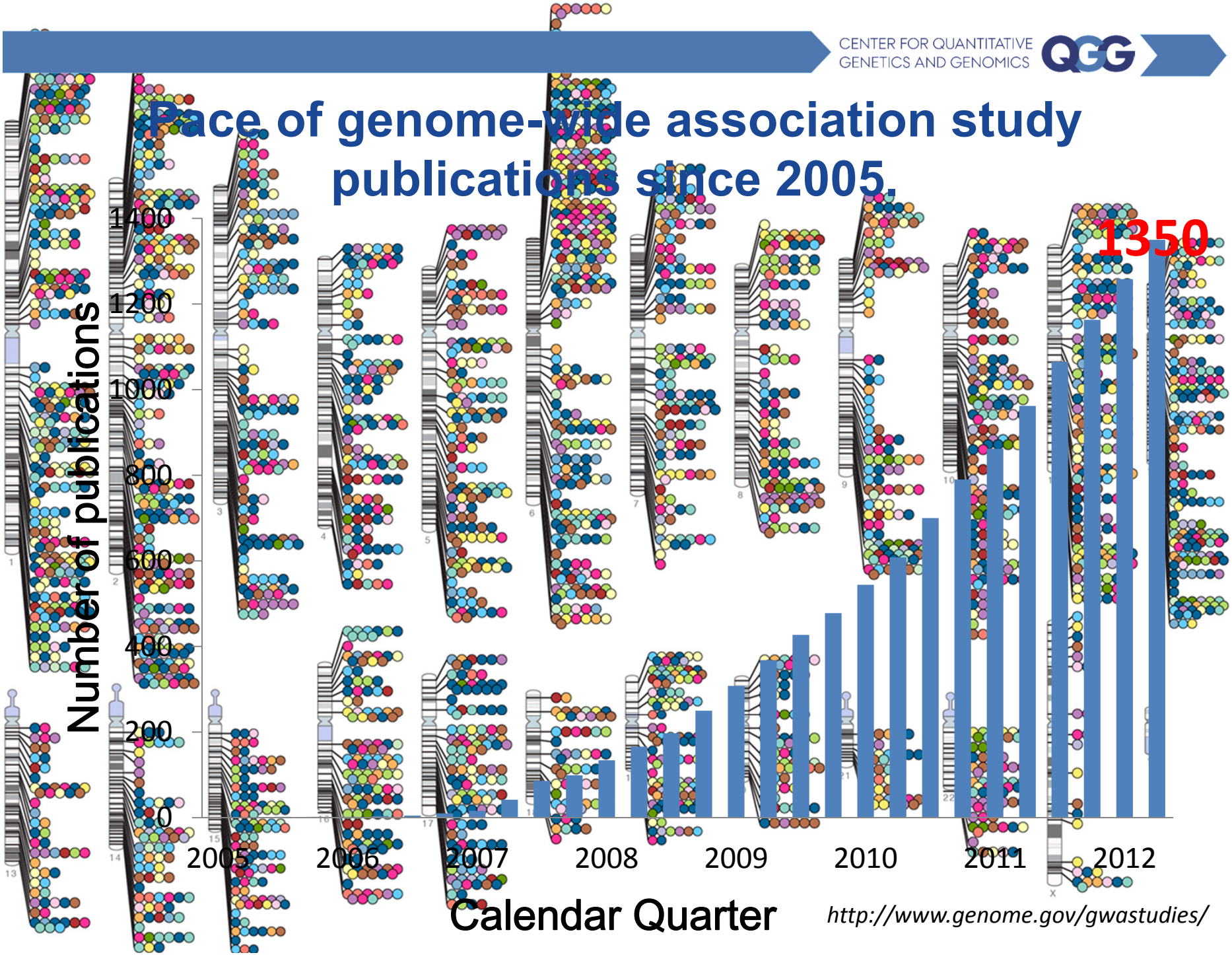


Causal locus directly typed

Marker correlated with causal locus

Hirschhorn & Daly, Nat Rev Genet 2005

Pace of genome-wide association study publications since 2005.



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

What have we learned from GWAS?

- 1000s trait-associated genetic variants have been identified by GWAS
- Majority with small effect on the trait
- Very little of apparent heritability is explained
- Biological effect behind majority of the associated variants remains unclear

Heritability explained by the identified loci

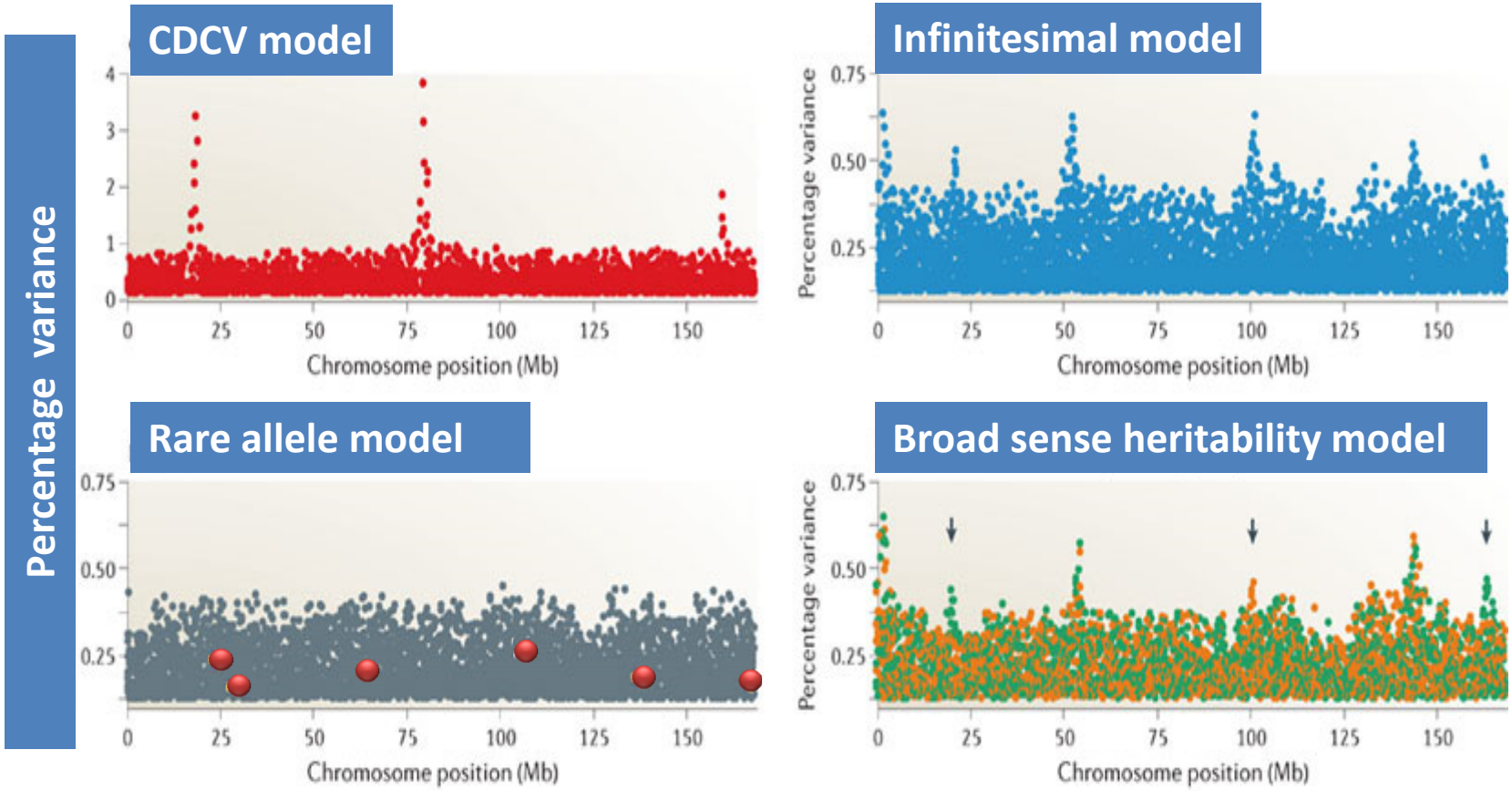
Disease	Number of loci	Proportion of heritability explained (%)
Age-related macular degeneration	5	50.0
Crohn's disease	32	20.0
Systemic lupus erythematosus	6	15.0
Type 2 diabetes	18	6.0
HDL cholesterol	7	5.2
Height	40	5.0
Early onset myocardial infarction	9	2.8
Fasting glucose	4	1.5

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

Missing heritability

- Very little of apparent heritability is explained
- Heritability estimates can be overstated
  - Shared environment
  - Non-additive gene effects
  - Gene x environment
  - Epigenetics including parent-of-origin effects
- Poor tagging
  - Common variants
  - Structural variants
  - Rare mutations of large effect

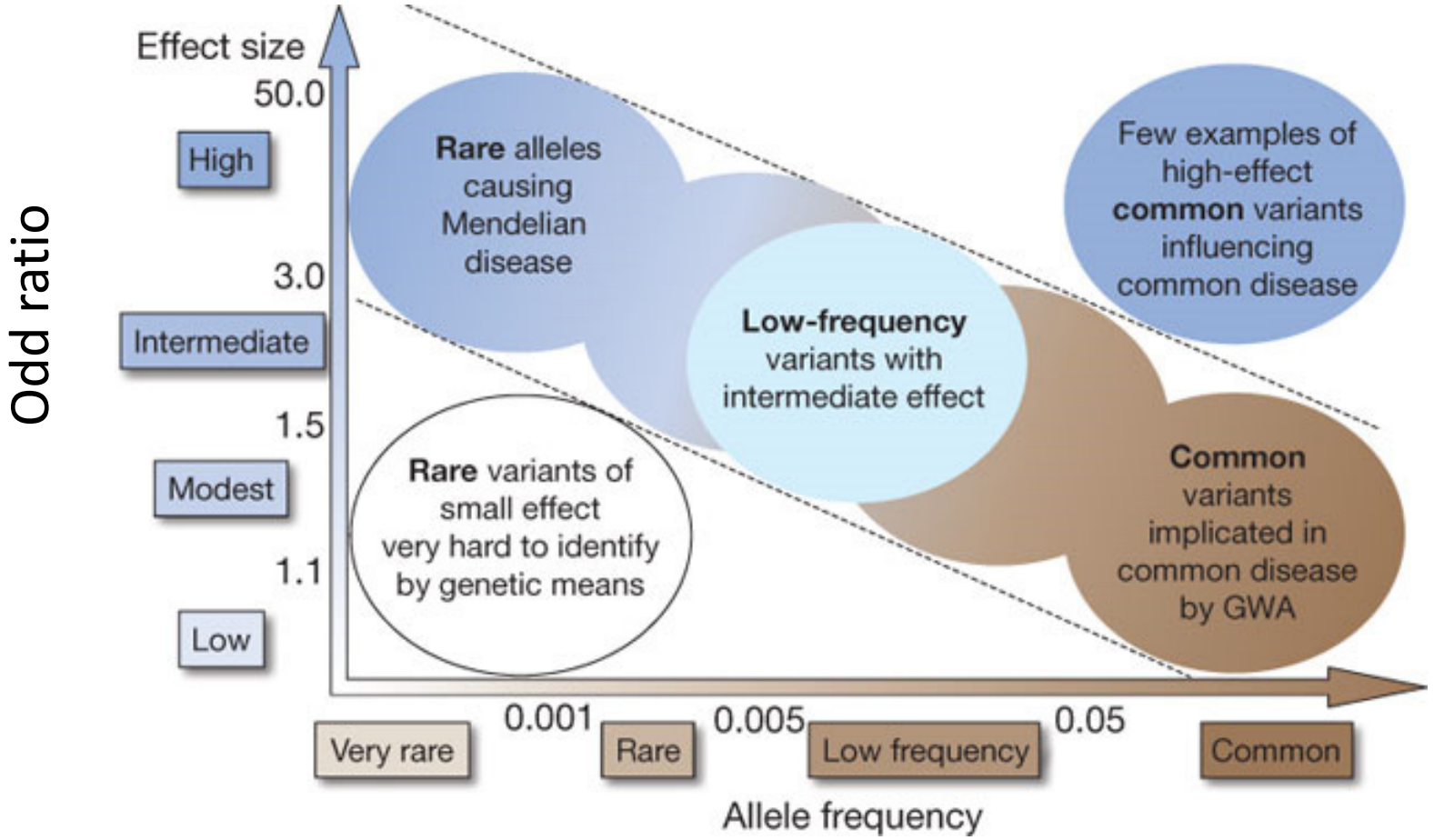
Different expected signatures from genome-wide association studies for four models



Nature Reviews | Genetics

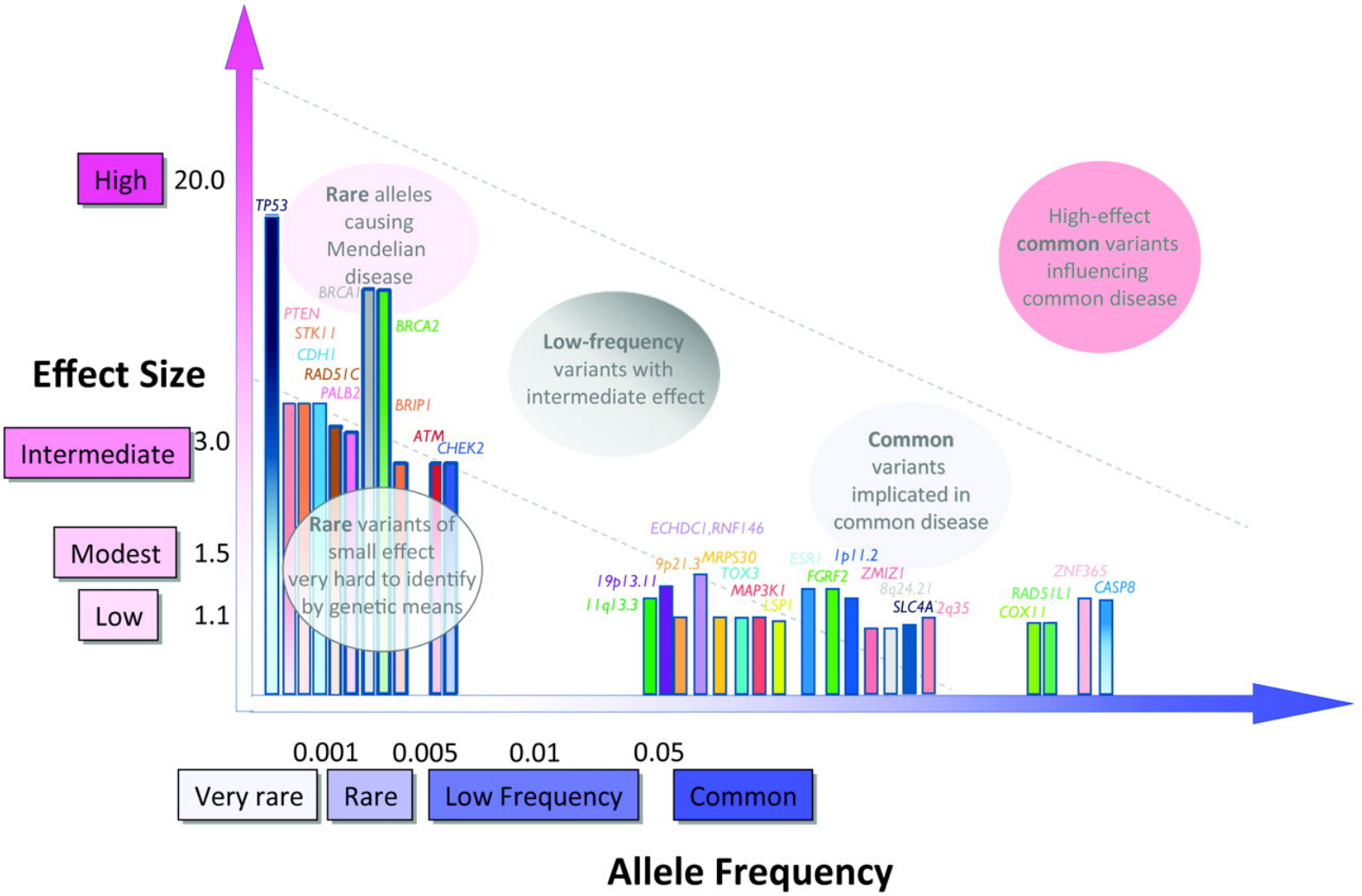
Gibson G. Nat Rev Genet 2012, 13:135

Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect



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

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

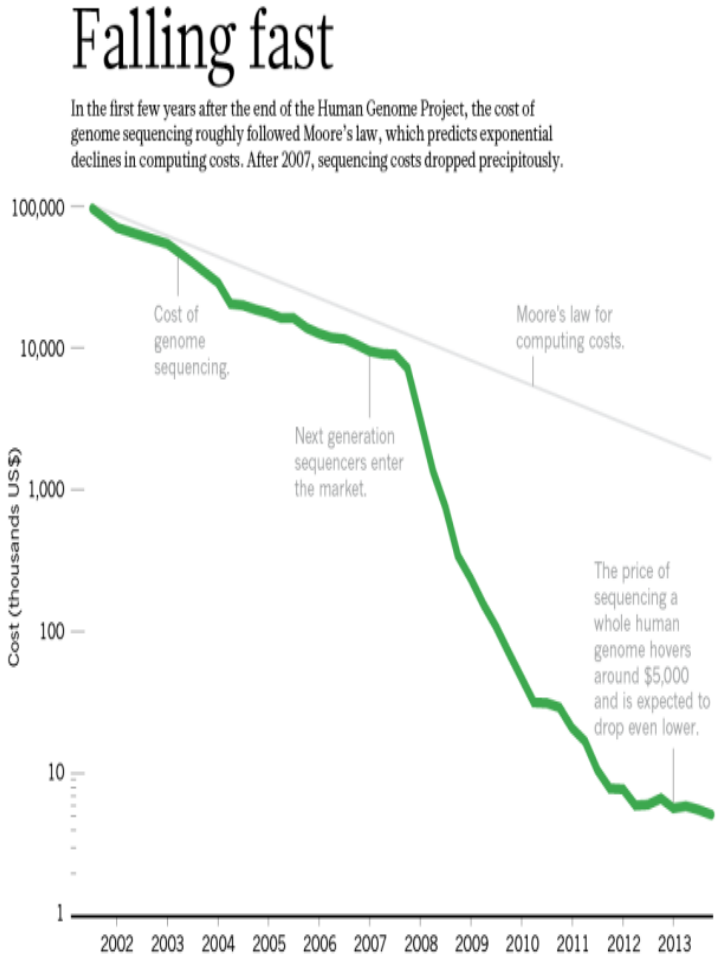


Hindorff L A et al. Carcinogenesis 2011;32:945-954

Sequence based design for genome-wide association

Advantage of sequence data

- The causal mutations are in the data
- No longer depend on linkage disequilibrium
- Discover mechanisms of disease/ complex traits
- Information is usable across population



Hayden EC, Nature 2014, 507:294

Genome / exome sequencing

- **Genome:** All the genetic component within in organism (~3×10^9 bp)
- **Exome:** the portions of a gene or genome that code information for protein synthesis
- Approximately 180,000 exons in the human genome, arranged into approximately 22,000 genes
- Exome represent about 2-3% of the genome

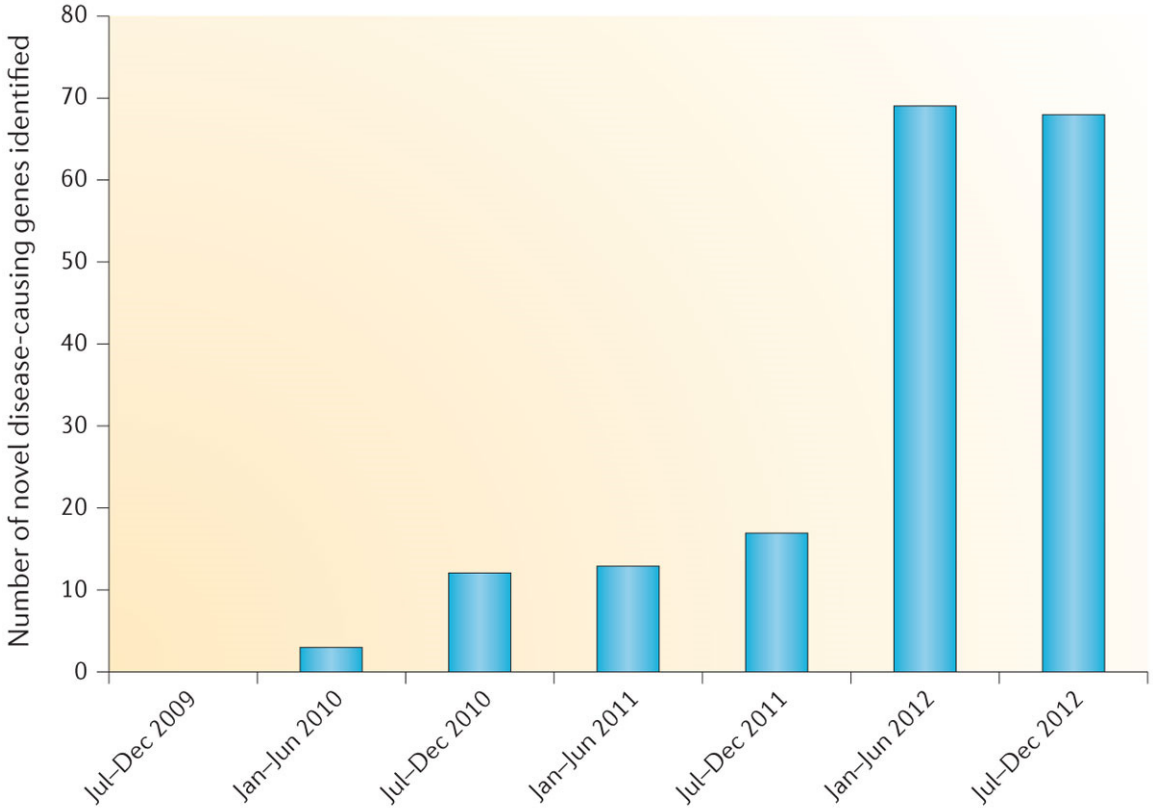
Exome Sequencing

- Estimated that majority of disease causing mutations will be identified within exome
- Lower cost, better depth of coverage
- Less data analysis /storage
- Fewer variants
- Biological function is known

Sequence based design for GWA

- Greater potential to identify causal variants
  - Direct application
- Rare variants
- Spontaneously arising (De novo) discovery
- Expensive
  - International collaborations (1000 genomes project, UK10K)
  - Imputation works (Hybrid design)
- Handling vast amount of information (IT support)
- Interpretation problems

Pace of discovery of novel rare-disease-causing genes



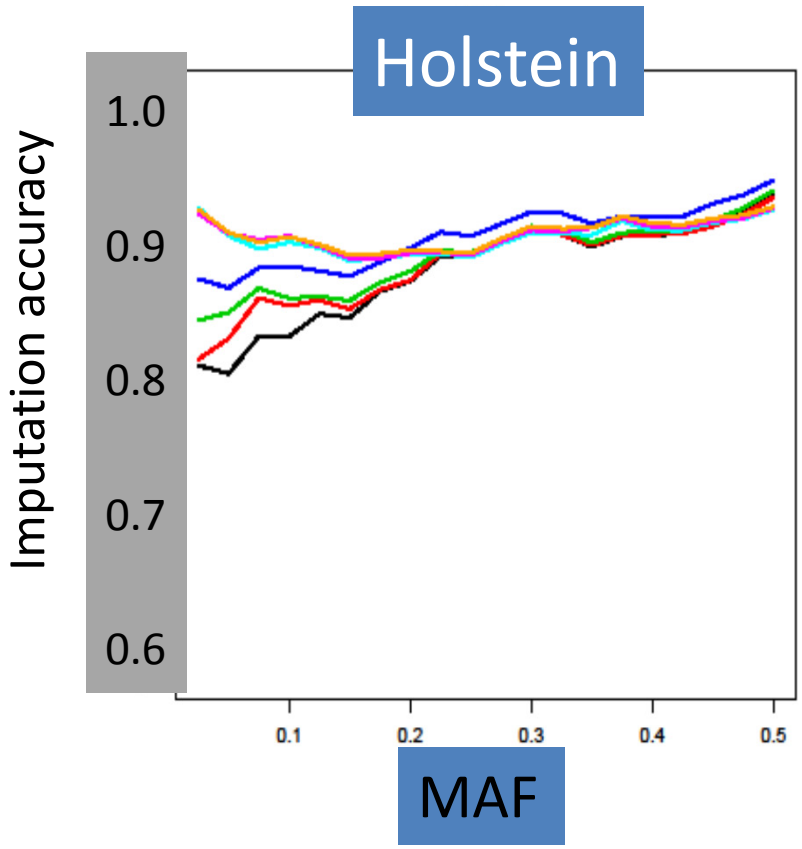
Nature Reviews | Genetics

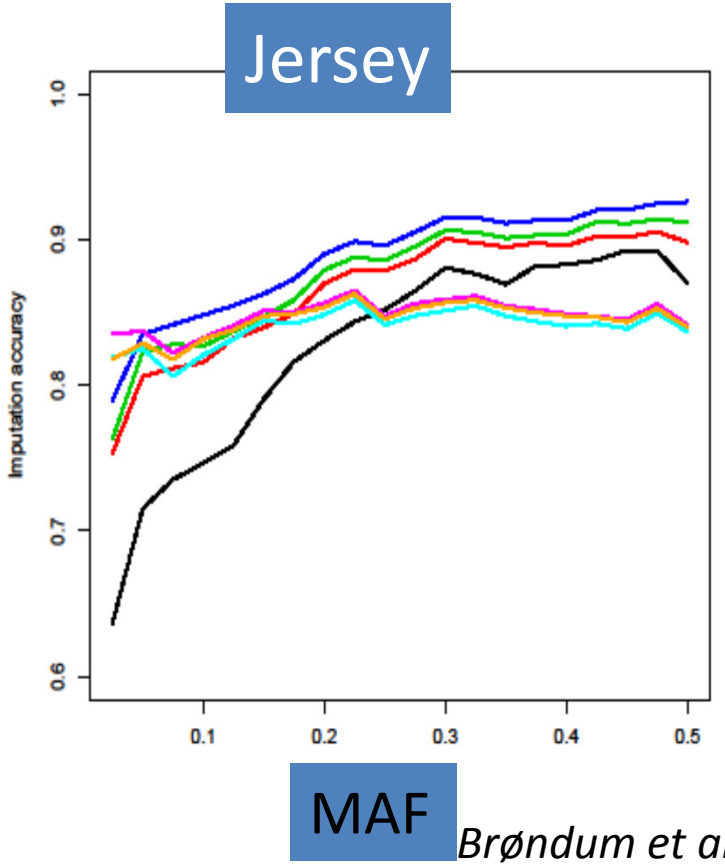
Boycott et al. 2013

Whole genome sequence and association with quantitative traits

Imputation to full sequence

- 242 whole genome sequenced bulls
- ~9 million variants





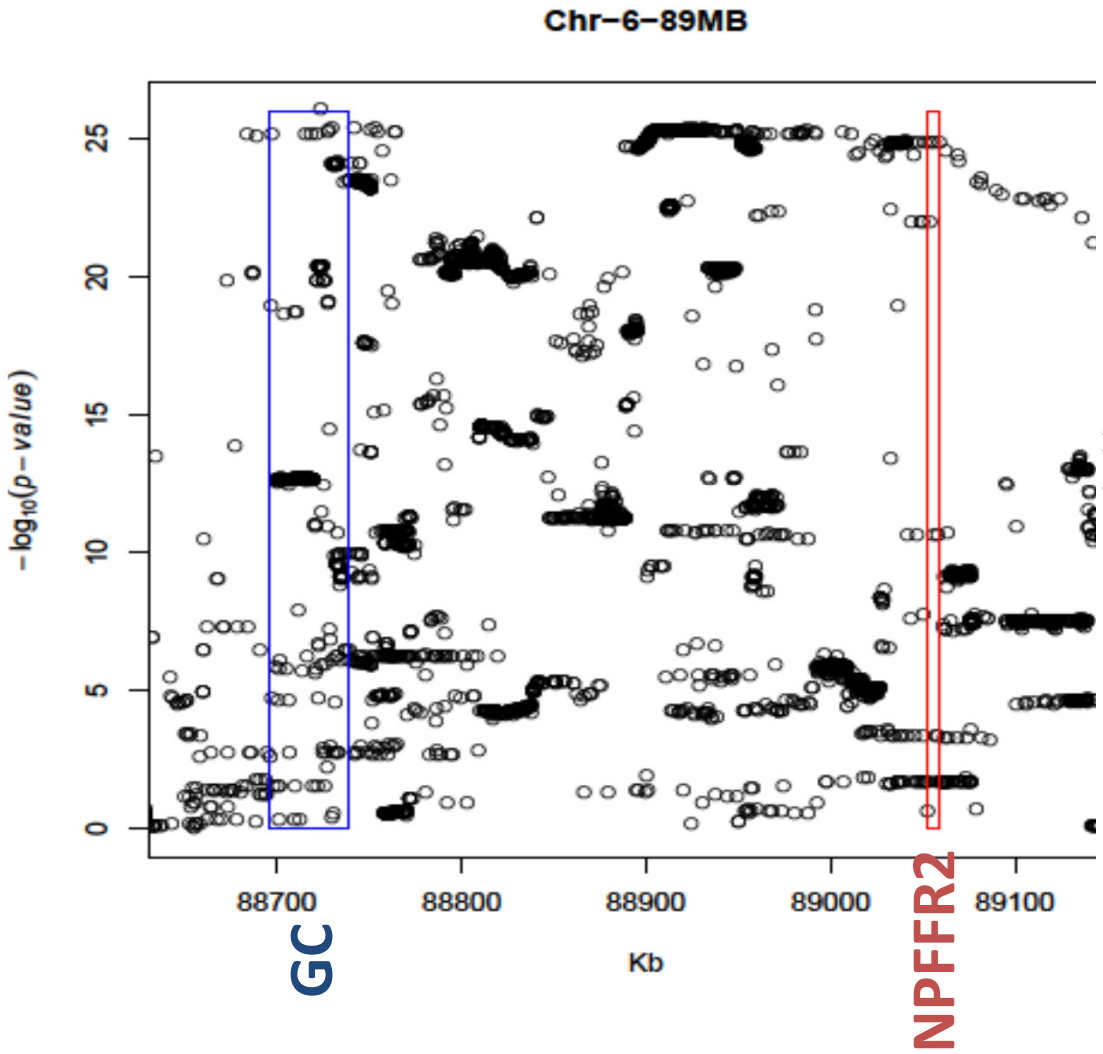
Brøndum et al. 2014

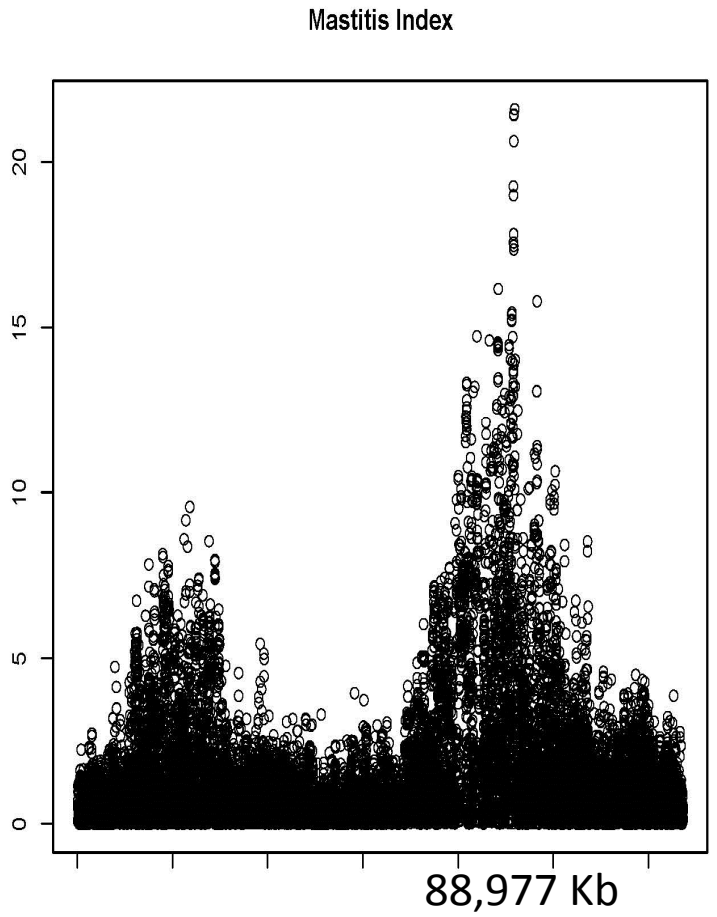
GWAS with NGS data in dairy cattle

Additive genetic variance (%) explained by lead QTLs

QTL	Milk	Fat	Protein	Mastitis	Fertility
1	10.22	12.21	2.63	4.18	1.66
2	2.29	2.39	0.96	2.09	1.66
3	1.67	1.43	0.96	1.98	1.58
4	1.49	1.36	0.96	1.76	1.50
5	1.30	1.23	0.91	1.54	1.50
Ten lead QTLs	22.61	23.86	10.88	18.58	14.83

LD obscuring the location of specific causative loci

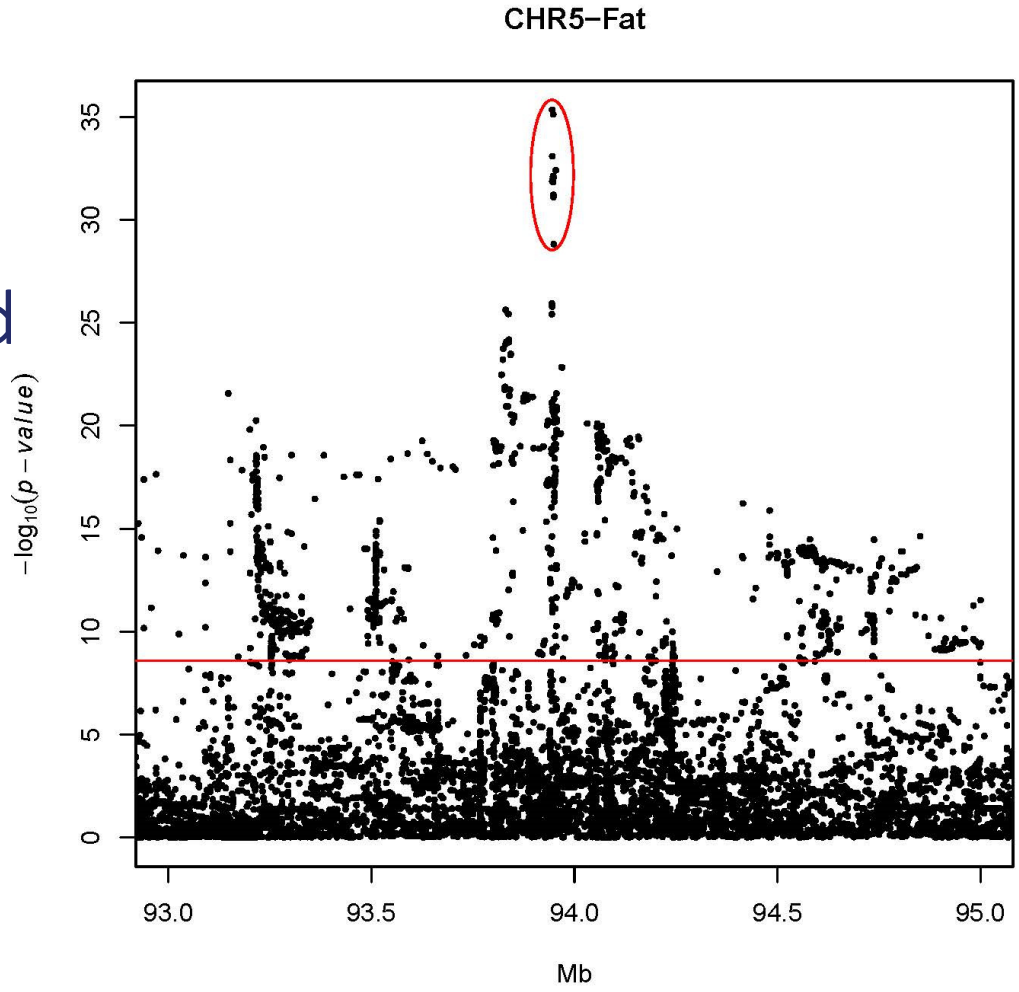




Sahana et al. 2014

Strongest association with intronic variants

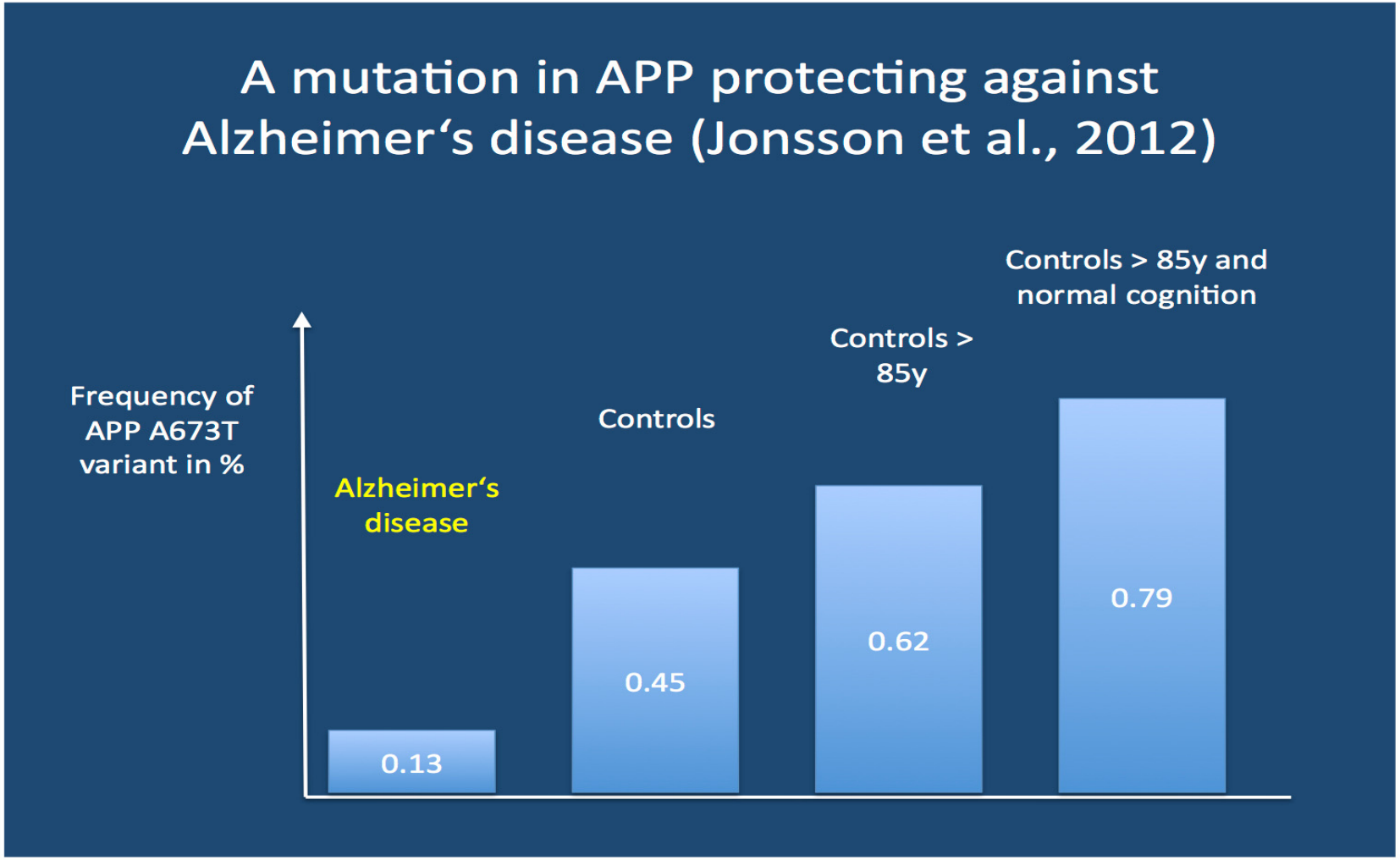
- A major QTL for fat on chromosome 5
- Strongest associated SNPs are from MGST1 gene
- But are intronic variants



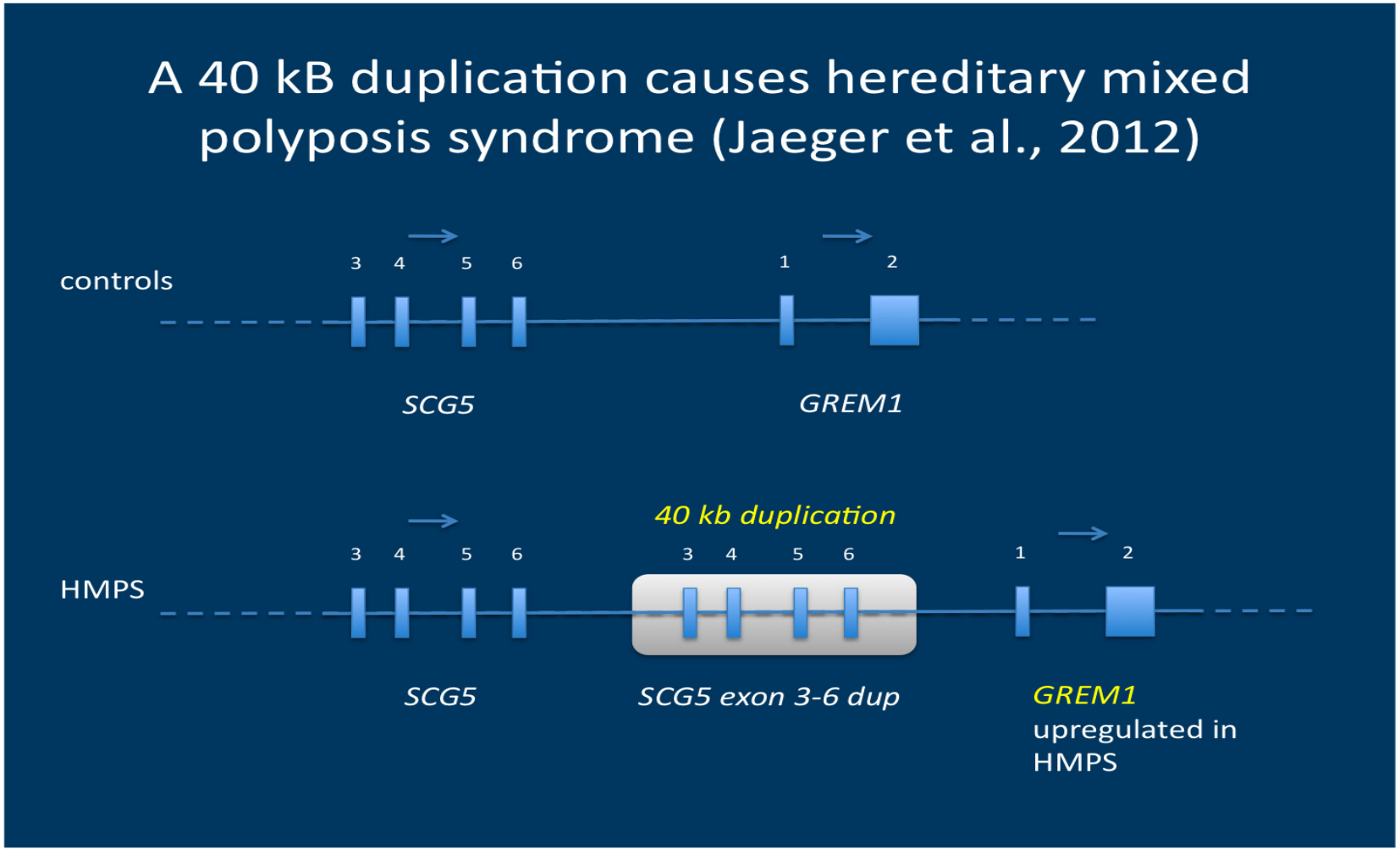
Biological effect behind variants remains unclear

- Associated loci are located outside recognized genes
- No known function of the allelic variant
- LD obscuring the location of specific causative loci
  - Leads to inability to ascribe function
  - Pinpointing the causal variant among the many variants present in the genome remains a major challenge
- Association information alone is often insufficient for quantitative traits

Genetic heterogeneity

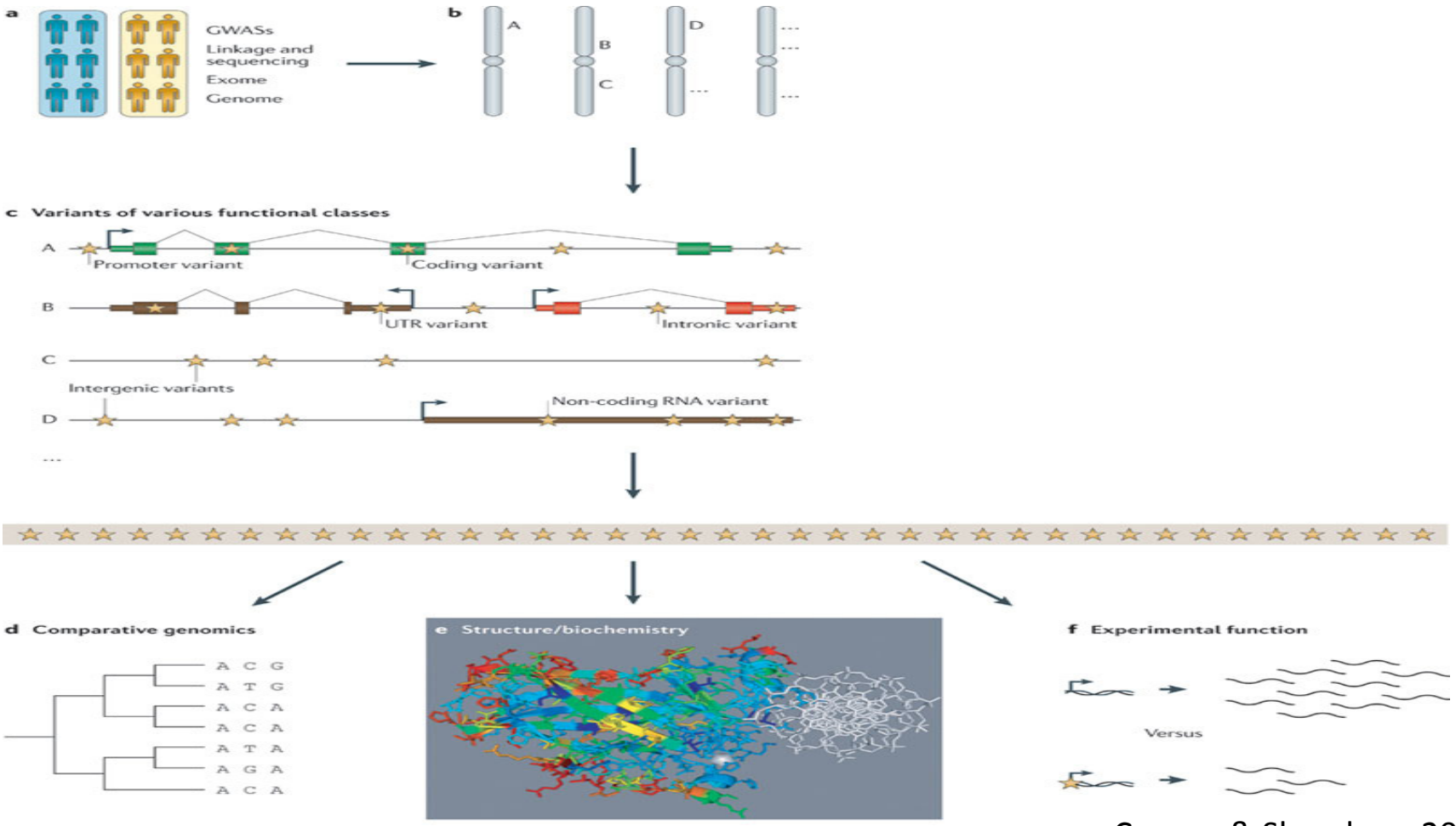


Mutations are outside genes



How do we prioritize candidate variants?

Steps to prioritize candidate variants



Cooper & Shendure, 2011

Nature Reviews | Genetics

Evidence relevant to the implication of sequence variants to phenotype

Evidence level	Evidence class	Example
Gene level	Genetic	Gene burden
	Experimental	Protein interaction, biochemical function, expression, gene disruption, model system
Variant level	Genetic	Association, segregation, population frequency
	Informatics	Conservation, predicted function
	Experimental	Gene disruption, phenotype recapitulation, rescue

MacArthur et al. Nature 2014, 508:469-476

Challenges ahead

- Structural variants
- Merging massive amount of data (WGS, WES, 'omics' data, phenotypes, environment etc.)
- Greatest challenge will be to deciphering functional mechanism and clinical relevance

Concluding remarks

- Sequence data
  - Increased discovery of QTL
  - Precise localization of QTL
  - Very successful to identify causal variants for monogenic trait
  - But not so successful for quantitative trait
    - Limitation to ascribe function
    - Difficulty in separating causal variants from the makers in LD

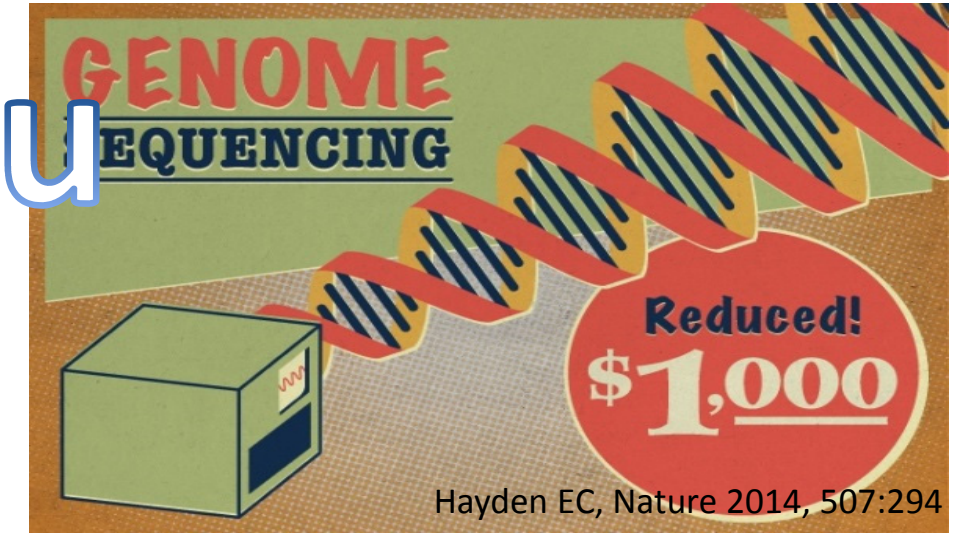
Concluding remarks

- We need very large data sets, larger than any entity can collect on its own
  - Foster cooperation: knowledge and data
- It is now time to utilize the information for improved care for diseases / prediction

Acknowledgements

- Bernt Guldbrandtsen
- Mogens Sandø Lund

Thank you



Hayden EC, Nature 2014, 507:294