Analysis of genome-wide genotype data

Acknowledgement: Several slides based on a lecture course given by
Jonathan Marchini & Chris Spencer, Cape Town 2007

Introduction & definitions

- Allele: A version of a genetic locus
- Single Nucleotide Polymorphism (SNP): Common variants (>1% frequency of minor allele?)
- An allele can be the
    - Minor allele (the less frequent)
    - Ancestral
    - Derived (opposite to ancestral)
- At a polymorphic locus an individual can be
    - Homozygous (two alleles identical)
    - Heterozygous

Human variation

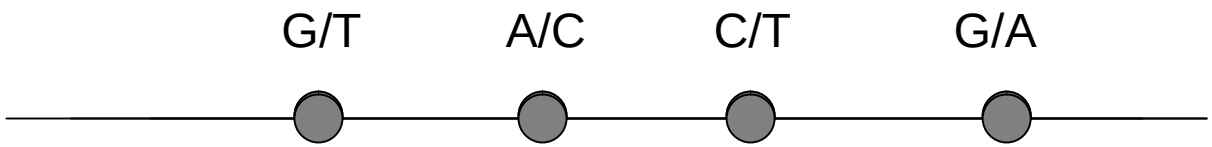
A pair of randomly sampled human chromosomes differ at about 1 site per 1,200

A total of ~ 10 million common SNPs in human (account for 90% of human variation)

Many variants account for the rest

Linkage, haplotypes and HapMap

Haplotype: A set of alleles co-occurring on a chromosome (contraction of haploid genotype)



G A T G

T C T G

Linkage, haplotypes and HapMap

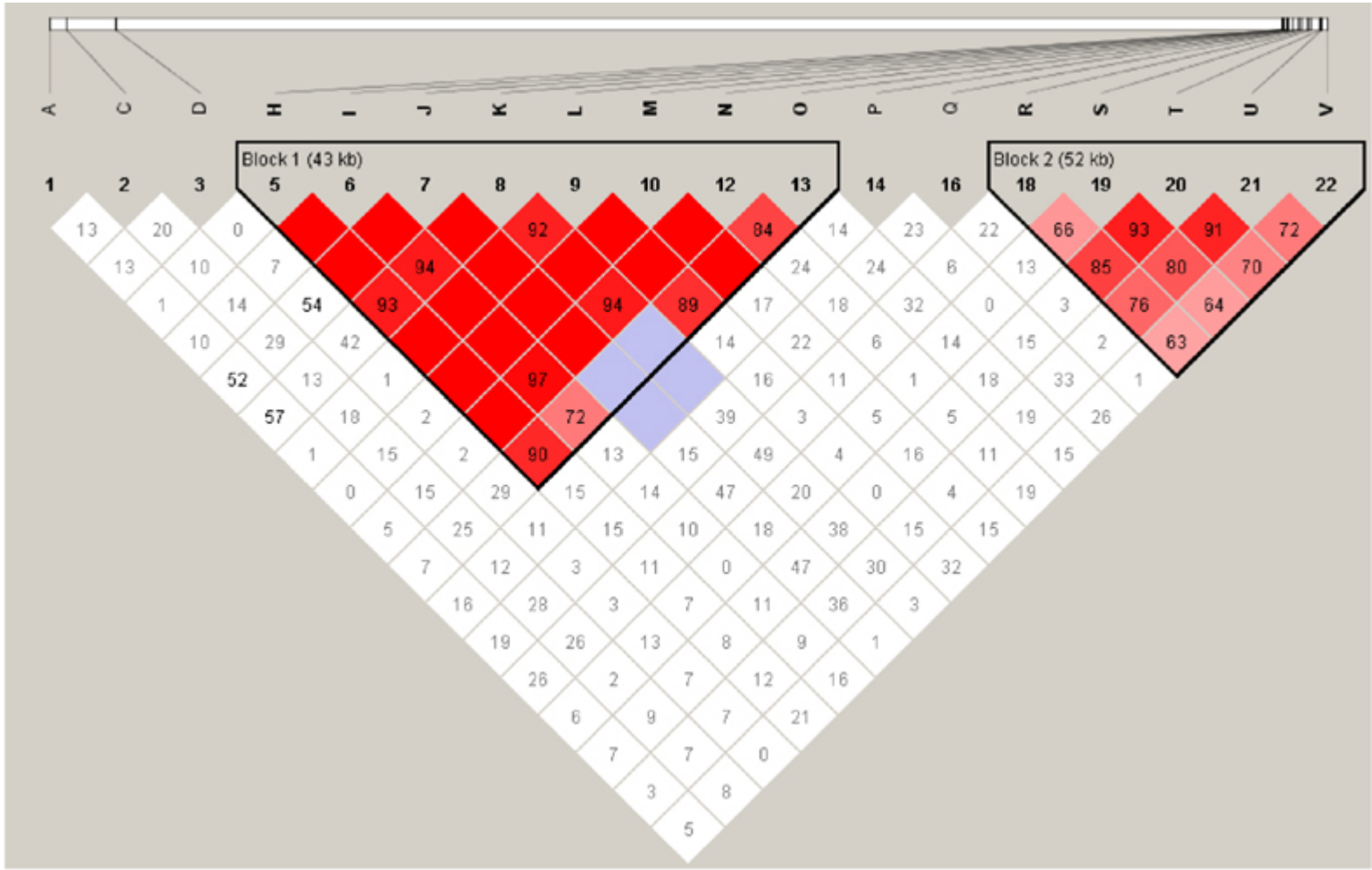
Haplotype: A set of alleles co-occurring on a chromosome (contraction of haploid genotype)

Alleles located close to one another are inherited together (said to be linked or in linkage disequilibrium)

This pattern is disrupted by recombination (shuffles chromosomes)

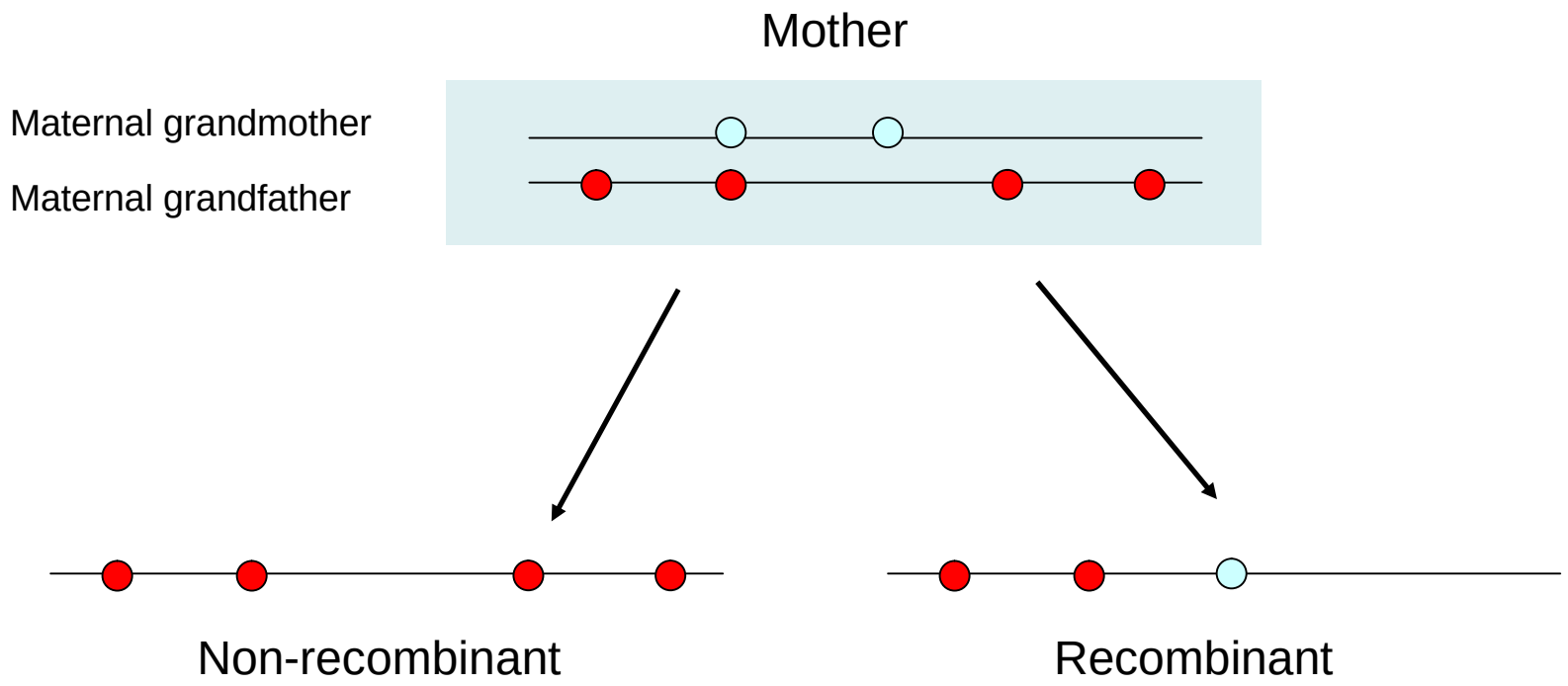
But recombination finite and time since origin of humans (~150,000 years) insufficient to break down all linkage => haplotype blocks

Plots of linkage disequilibrium



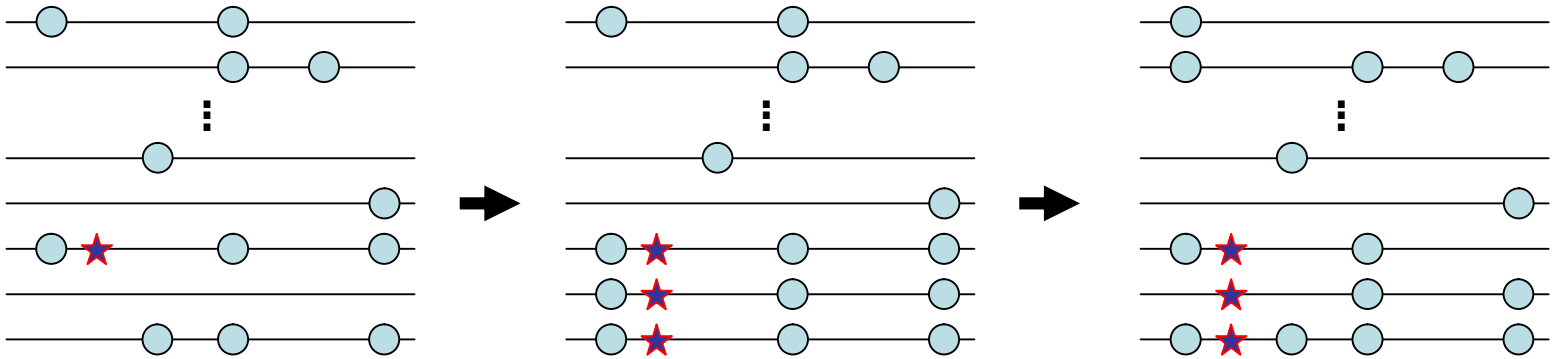
Jeronimo et al 2007

Recombination breaks down LD



- Recombination breaks down LD by making mosaic of existing haplotypes.

LD and association mapping -> Genome-wide association studies (GWAS)



Disease mutation arises on single haplotype

Increase in frequency by drift.

Shuffling occurs due to recombination

Measuring linkage disequilibrium

Linkage disequilibrium measure, D: $f_{AB} - f_A f_B$

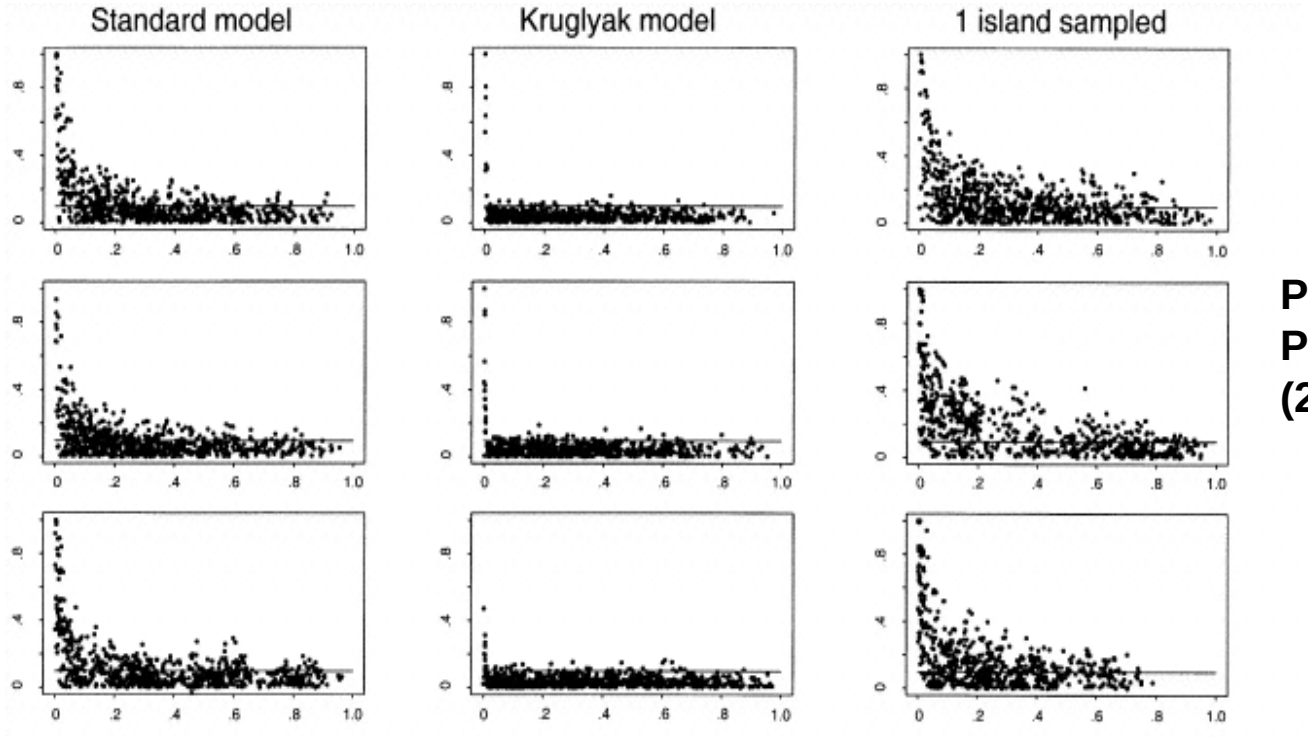
But D tends to depend on the frequency of the alleles (highest for equal frequency alleles). A normalized equivalent of D exists which does not have this dependency - D'

$$D' = D/D_{max}$$

R^2 = Pearson correlation coefficient (the proportion of the variation at one locus which is 'explained' by the variation at the other)

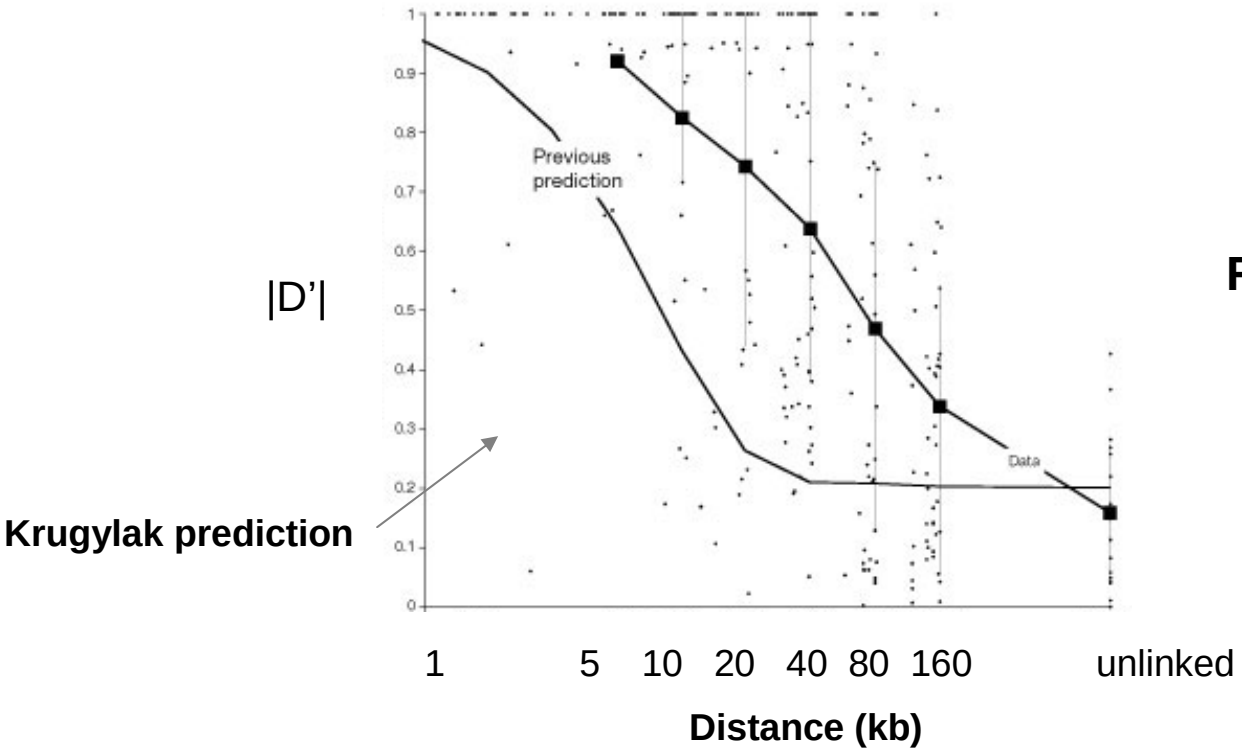
Linkage disequilibrium and association mapping

R^2 > about 0.5 is considered useful for association studies (below this you start to need much larger sample sizes). This level of linkage disequilibrium breaks down on a distance scale of about 5-10 kb in human populations (depending on the population and the presence or absence of recombination hotspots in the region of interest).



Pritchard and Przeworski (2001)

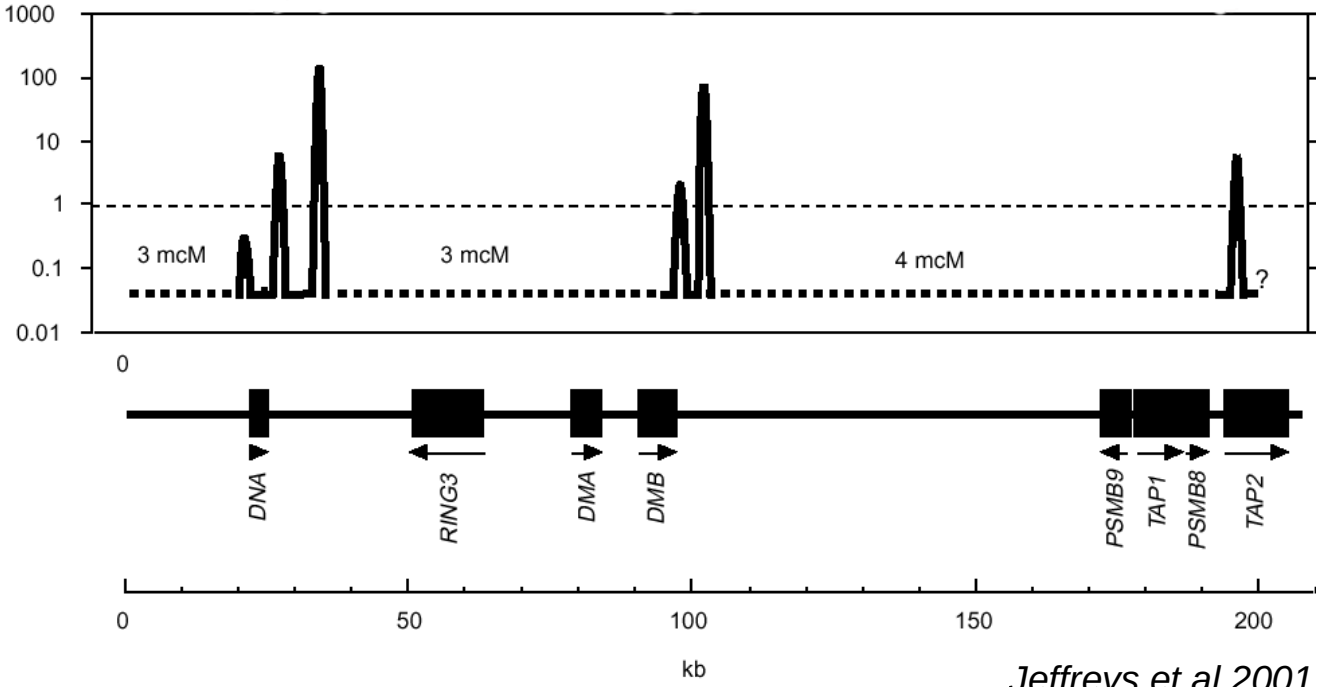
Decay of LD under demographic models



Reich *et al.* (2001)

- More appropriate to observe LD in natural populations
- LD extends over considerable distance **(>>10kb)** in most populations

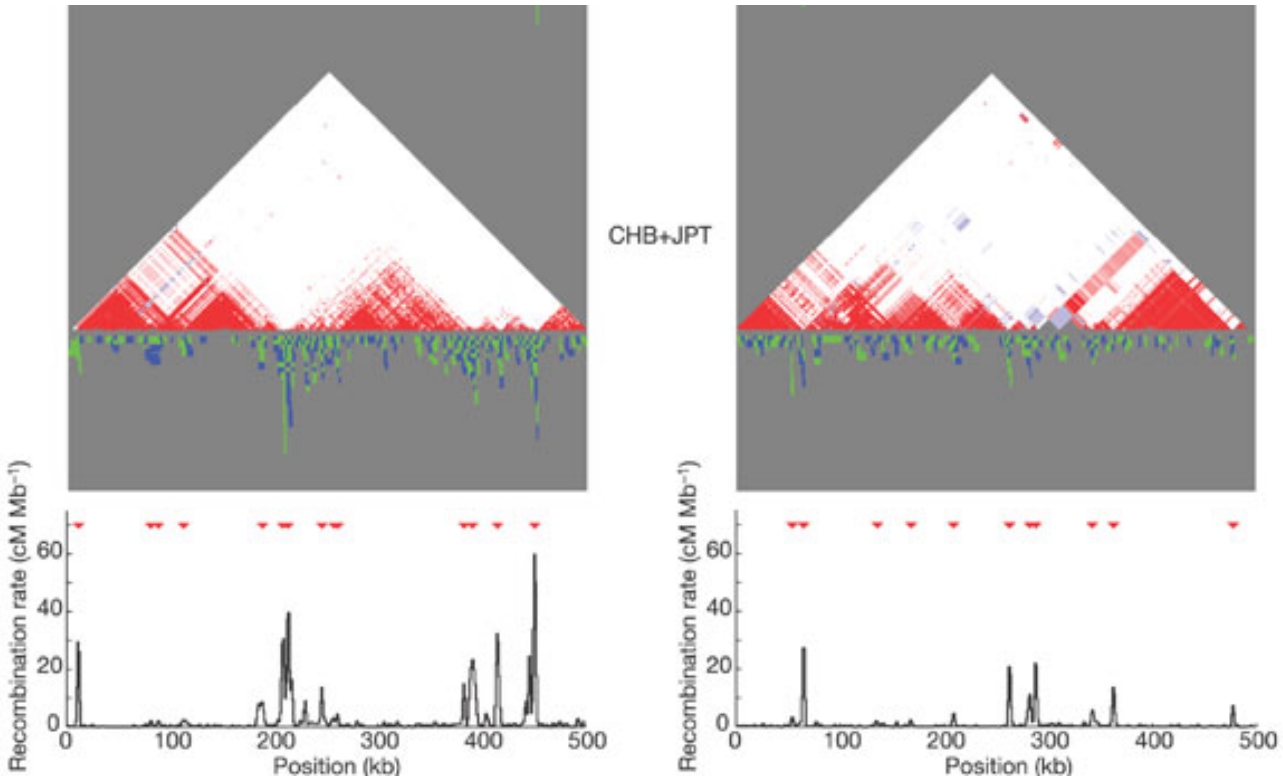
Why? Hotspots of recombination



Jeffreys et.al 2001

- Experimental evidence shows that recombination is clustered in hotspots
- Hotspots are interspersed with long cold spots

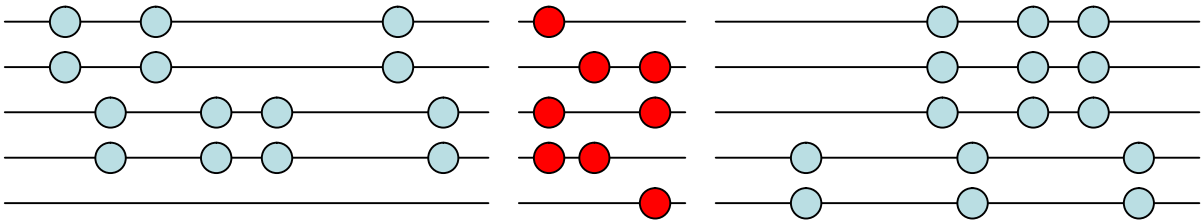
Recombination and LD



HapMap Consortium 2005

- Recombination hotspots demark break down in LD
- Inter-hotspot expanses have multiple markers in strong LD

Implication of recombination hotspots

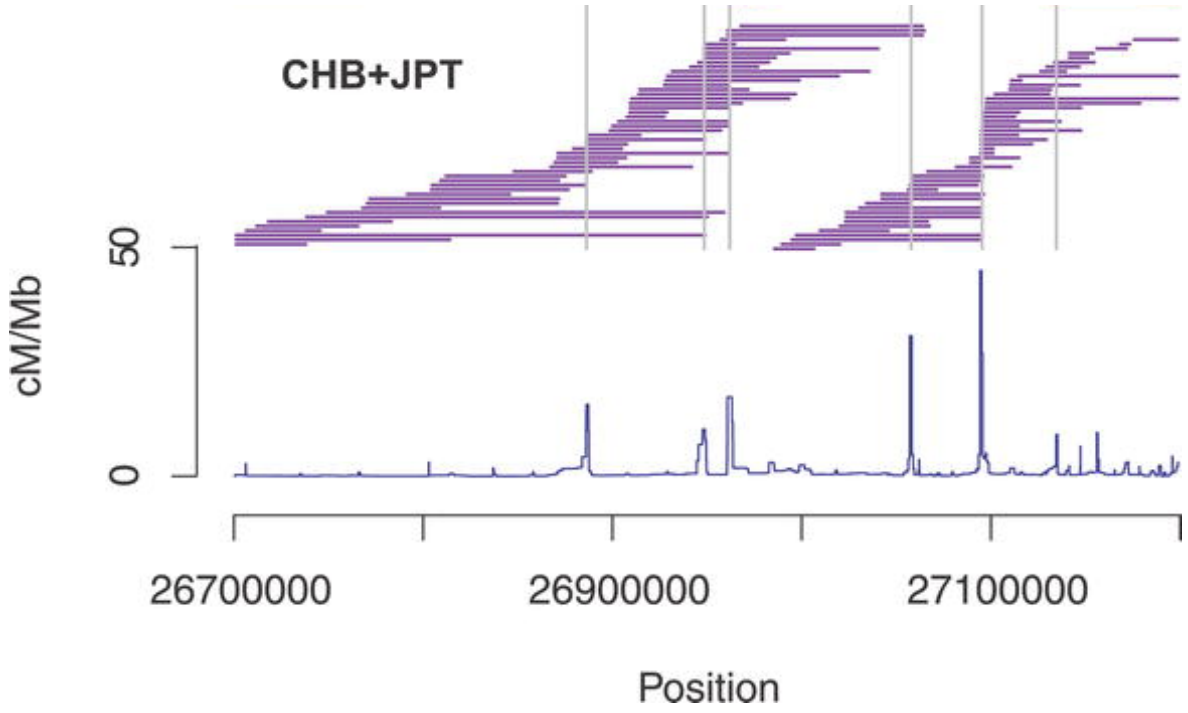


Haplotype “block”

Haplotype “block”

- Block-like structure to LD across the human genome
- Few markers may tag a large number of SNPs
- More markers needed in recombination hotspots

Relationship between haplotype blocks and recombination hotspots



McVean, Spencer Chaix, 2005

Imputation and Tag SNPs

Because of linkage the allele at a given SNP can predict the allele at a linked SNP

Therefore sufficient to genotype a subset of SNPs (Tag SNPs) and predict the rest

The HapMap project (begun in 2002)

Aims

Map variation across human populations

Determine tag SNPs (~600,000)

- one of the major reasons for the HapMap project

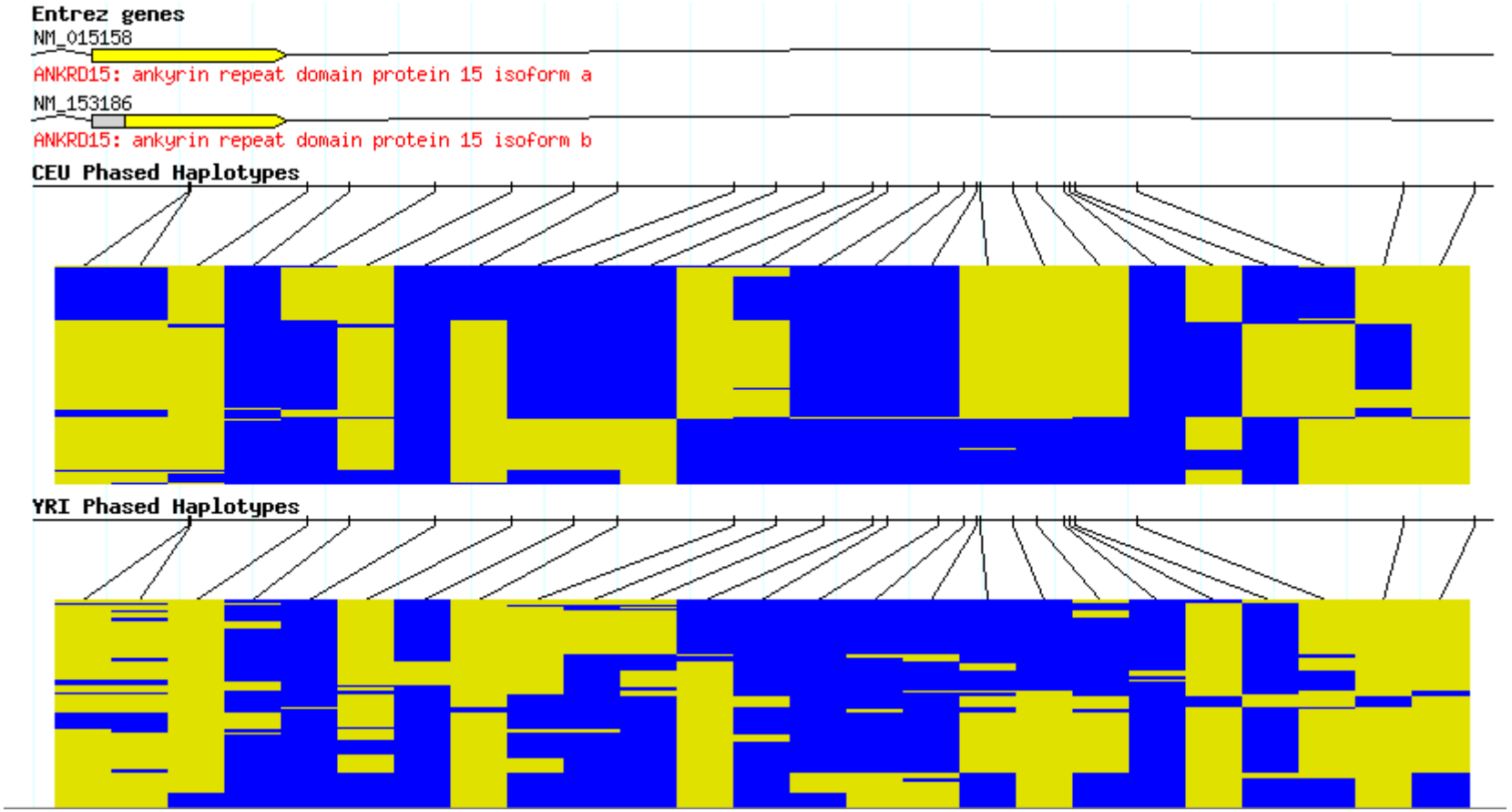
- several methods have been proposed to choose the optimal set of tags

Which leads to reduced cost of genotyping individuals

Facilitates research into genetic disorders

Make all the data publicly available

Haplotype blocks



Haplotype data from HapMap

rs7038561	rs10815523	rs9650708	rs12684491	rs2885162	rs2885163	rs13283481	rs7047065	rs7022766	rs12345967	rs7866110	r
711523	711717	711867	712423	712756	712946	712983	713429	713773	714043	714139	
C	G	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	T	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	
T	G	T	G	C	A	C	T	G	G	G	

Generation of ‘genome-wide’ genotype data

Generally carried out using ‘microarray’ technologies

- small glass slides typically containing millions of ‘spots’ consisting of immobilized nucleic acid sequences (or probes)

- genomic DNA is first fragmented (using enzymes) and labelled with a fluorescent dye

- genomic DNA washed over slide, hybridizes to complementary probes

- unhibridized DNA washed off

- fluorescent signal detected by scanning with a laser

Examples of some commercially available chips

- Affymetrix 100K
- Affymetrix 500K
- Affymetrix 1M

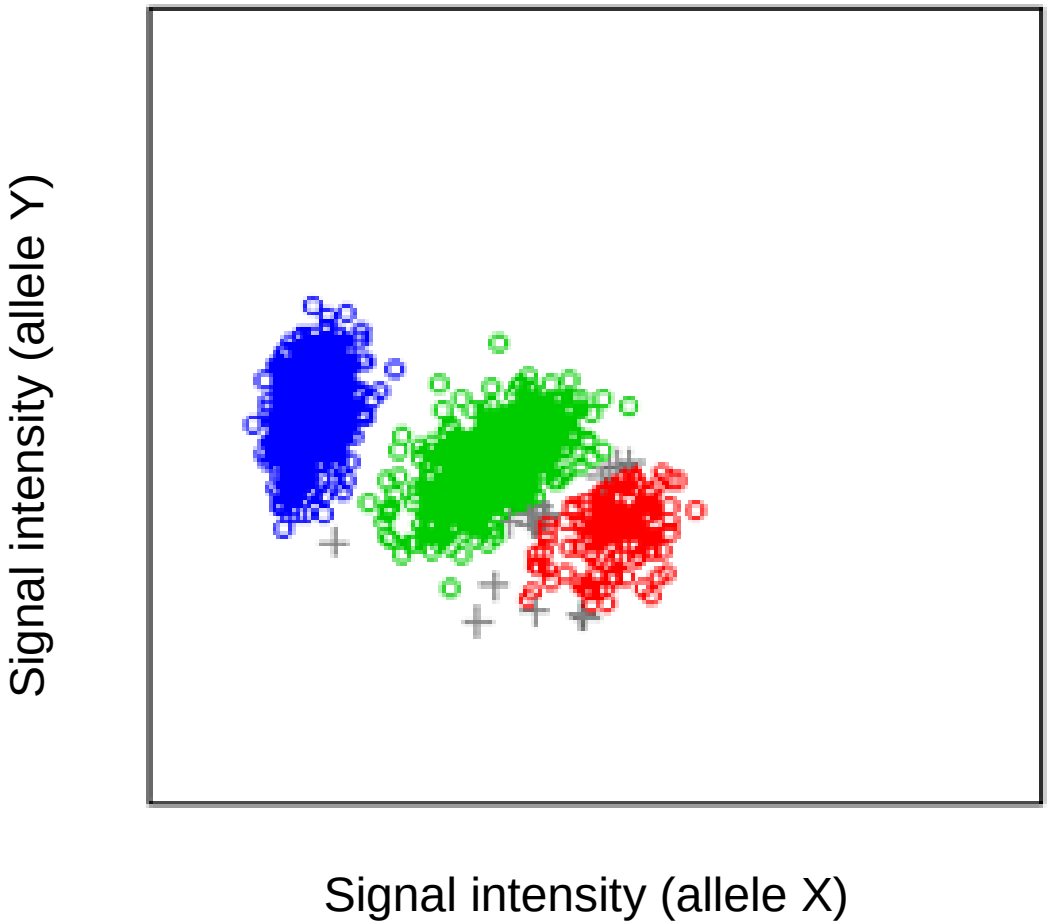
} Essentially random sets of SNPs

- Illumina 317K
- Illumina 550K
- Illumina 1M

} Designed using the HapMap

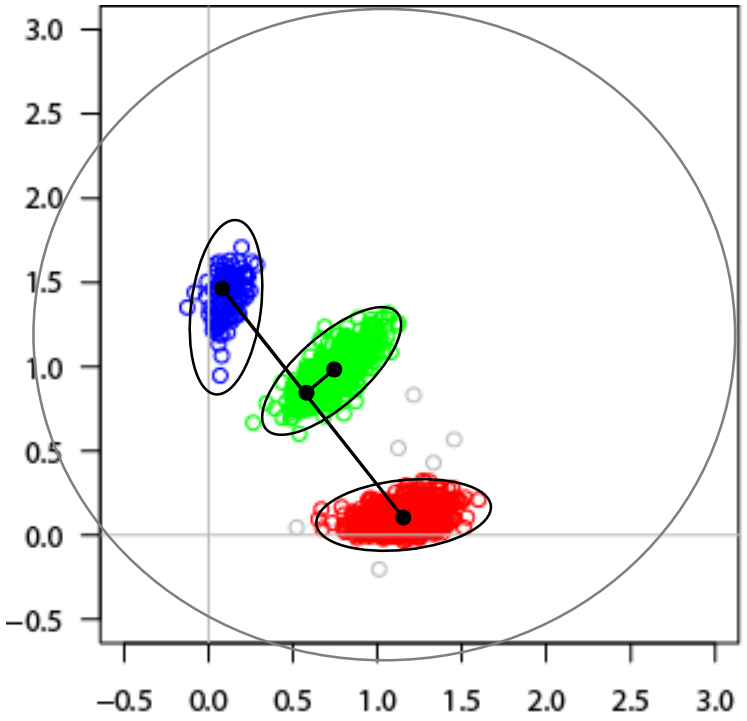
- Illumina 650Y (550K + 100K YRI fill-in)

Genotype Calling



E.g. Chiamo (Marchini et al.)

Model



Example: Wellcome Trust Case Control Consortium, *Nature 2007*

Investigated genetics of seven common complex disorders

~ 17,000 samples, ~500,000 SNPs

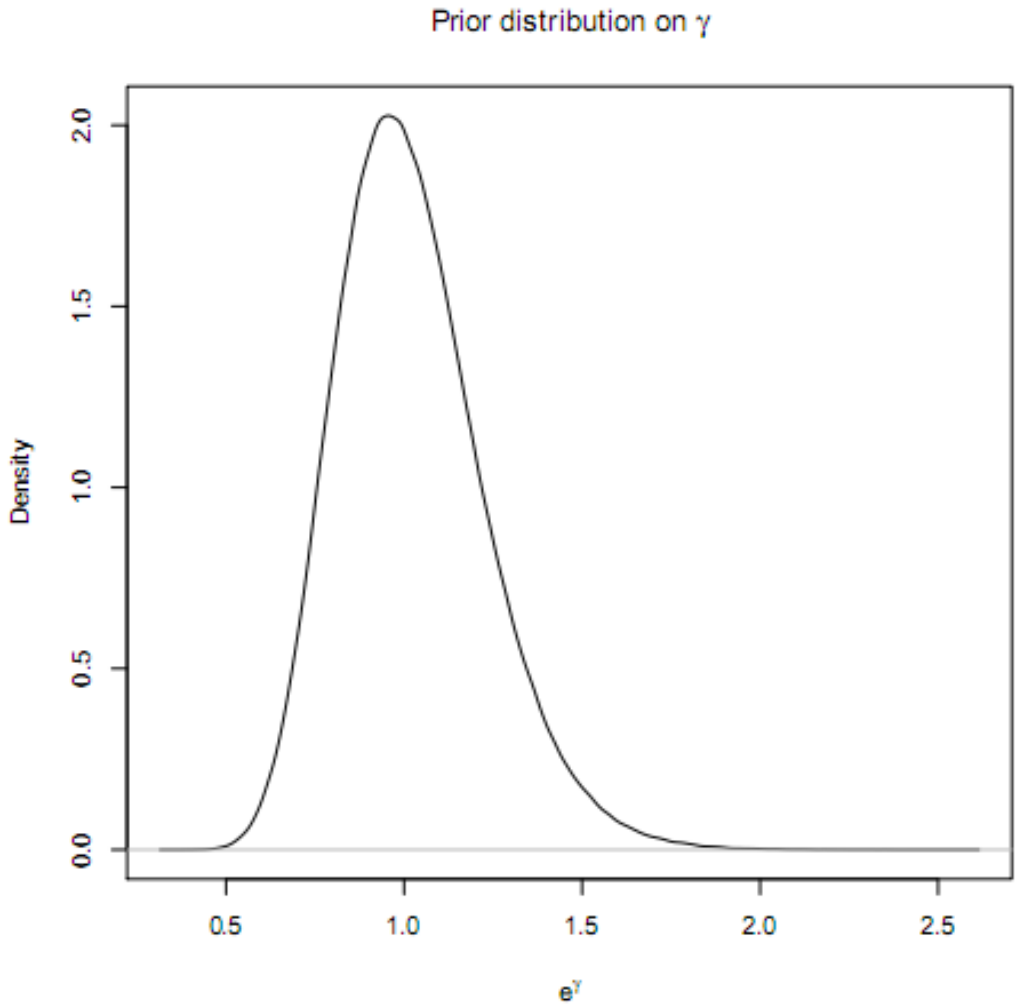
Used a logistic regression model;

$$M_0 : \log(\tfrac{p}{1-p}) = \mu$$
$$M_1 : \log(\tfrac{p}{1-p}) = \mu + \gamma Z_i$$
$$M_2 : \log(\tfrac{p}{1-p}) = \mu + \gamma I(Z_i = 1) + \phi(2\gamma I(Z_i = 2))$$

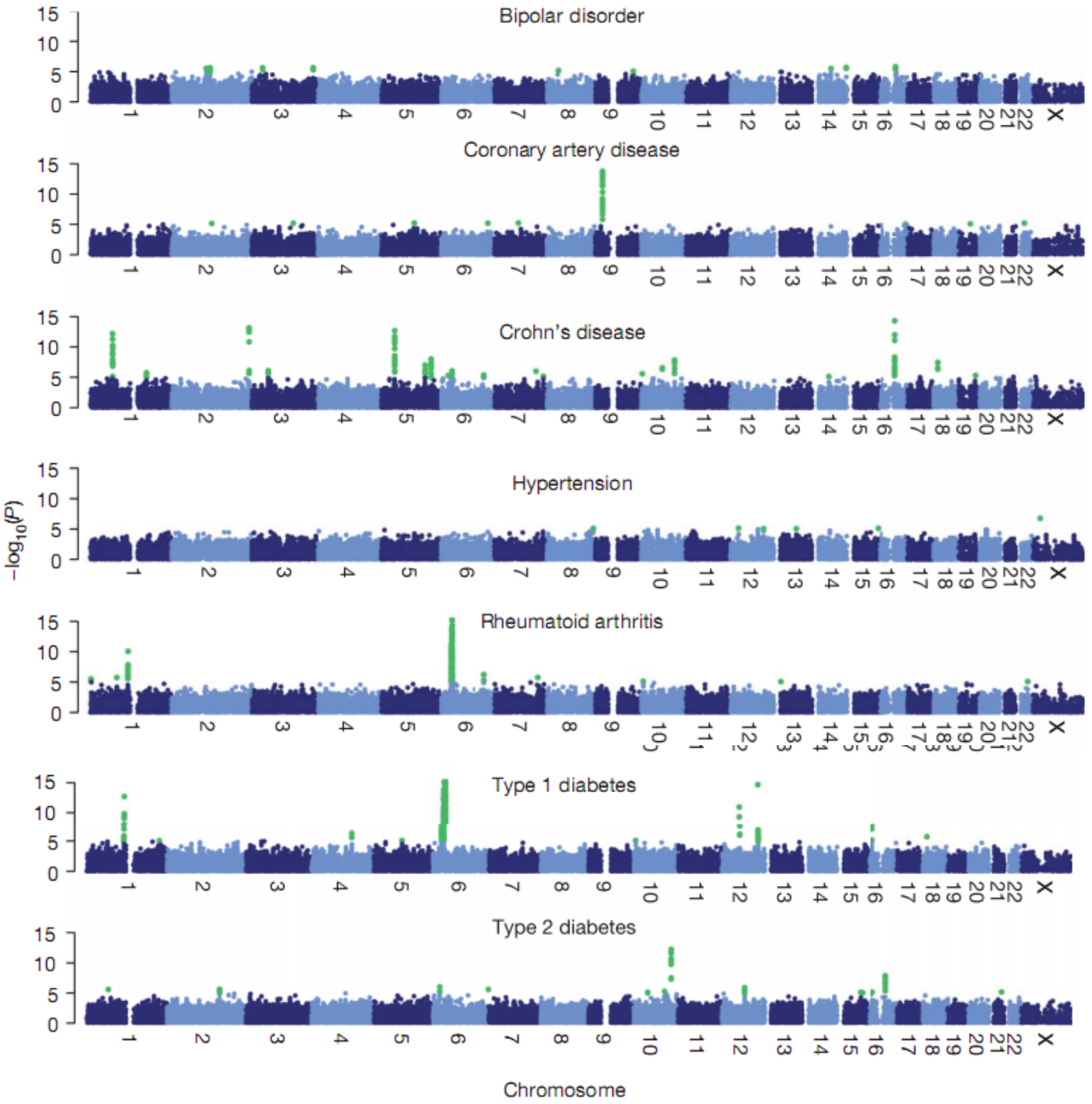
We can calculate Bayes factors, as before, as e.g. $\frac{P(D|M_1)}{P(D|M_0)}$

$$= \frac{\int P(D|\theta_1, M_1)P(\theta_1|M1)d\theta_1}{\int P(D|\theta_0, M_0)P(\theta_0|M0)d\theta_0}$$

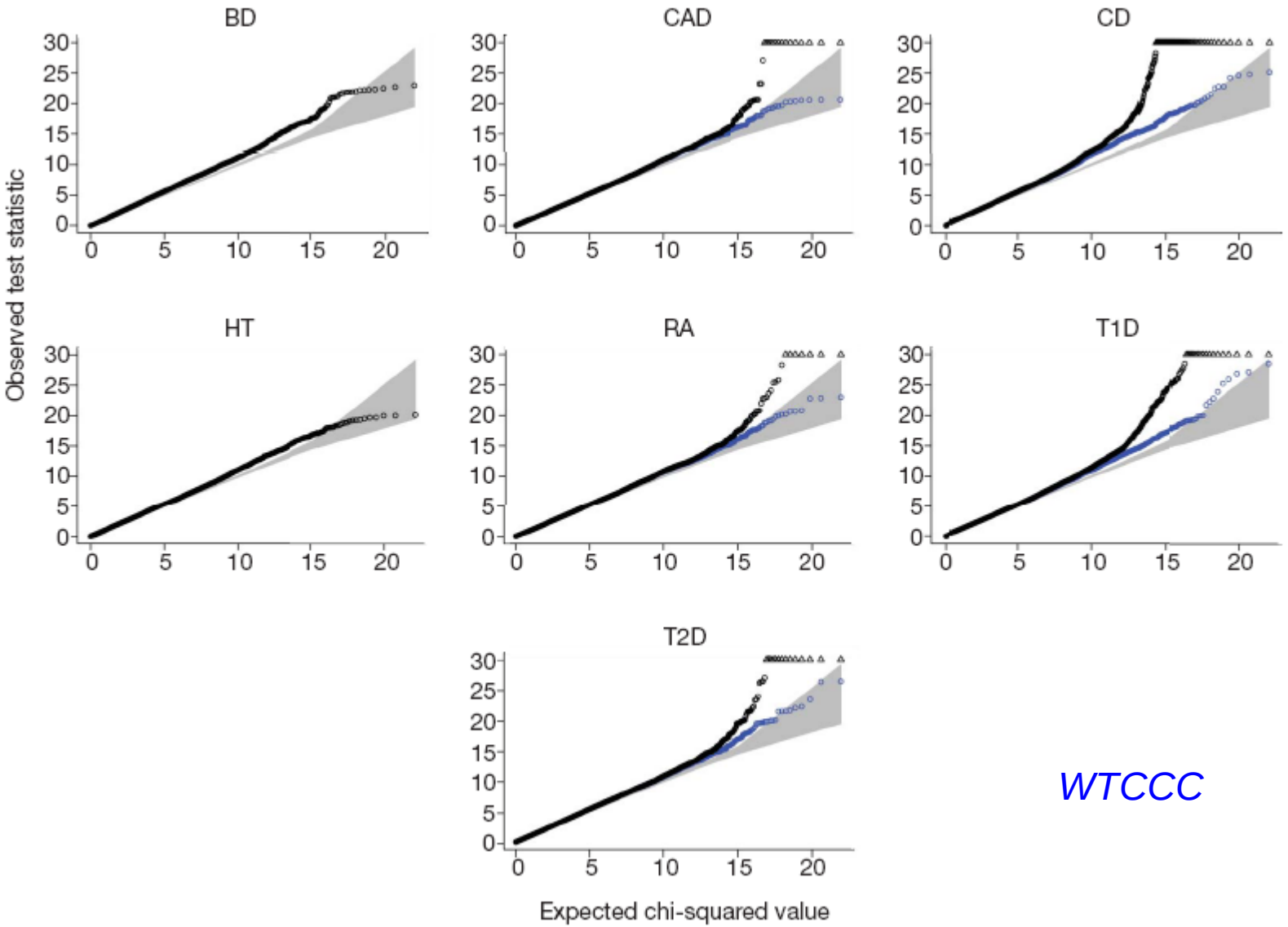
Deciding a prior for model parameters: “it is widely believed that the genetic variants underlying common disease will have risk allele odds-ratios in the range 1-2 with substantially more weight on the values between 1-1.5”



e.g. Results of Wellcome Trust Case Control Consortium (Nature, 2007)

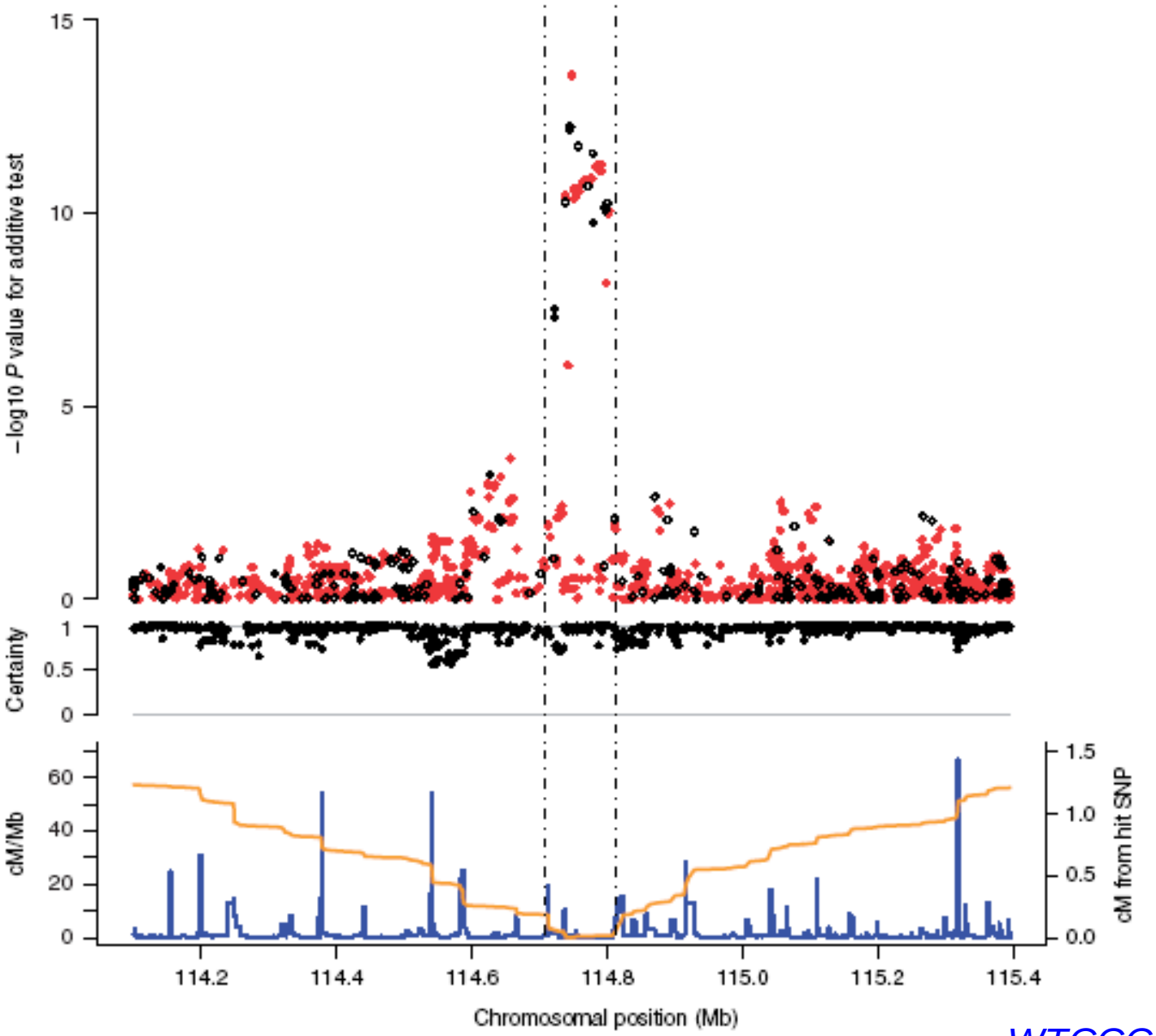


Accounting for multiple testing a key issue in GWAS



WTCCC

T2D hit region, chromosome 10



WTCCC