

## Population genetic inference

Date	Topic
22 <sup>nd</sup> Jan	Good questions in population genetics
29 <sup>th</sup> Jan	Principles of population genetic inference
5 <sup>th</sup> Feb	Recombination in the coalescent
12 <sup>th</sup> Feb	Natural selection
19 <sup>th</sup> Feb	Demographic models
26 <sup>th</sup> Feb	Combinatorics of the coalescent
5 <sup>th</sup> March	Population genetics of disease mutations
12 <sup>th</sup> March	Model organisms

## Reading

Cardon LR & Bell J (2001) Association study designs for complex diseases. *Nature Rev. Genet.* **2**:91-99

Slatkin M & Rannala B (2001) Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* **1**:225-249

Weiss KM & Clark AG (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**:19-24.

## Questions about genetic disease

- What is the genetic contribution to disease?
- How many genes contribute to the disease phenotype?
- What is the genomic location of disease-associated genes?
- What is the disease risk associated with a particular genotype?
- Are interactions (G x E, epistasis) important?

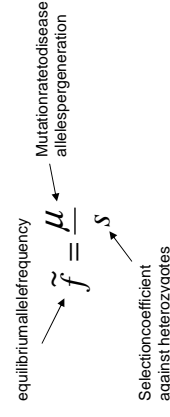
## Simple Mendelian disease

- Single gene
- Rare
- Dominant or recessive
  - Huntington's disease, Cystic fibrosis
- High penetrance
- Often strong selection against disease alleles
- Point mutations, deletions, insertions, translocations, triplet expansions - repeat

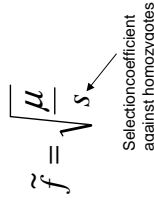
## Mutation selection balance

- frequency of disease alleles is a balance between recurrent mutation and selection against the disease

### Dominant



### Recessive



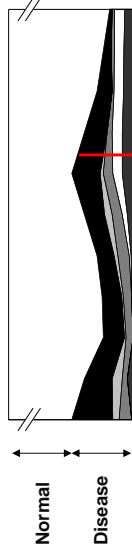
e.g. Huntington's:  $s \approx 2\%$ ,  $\mu \approx 10^{-6}$ ,  $f \approx 1/20,000$

e.g. Cystic fibrosis:  $s \approx 100\%$ ,  $\mu \approx 10^{-5}$ ,  $f \approx 1/140$

From increase in disease incidence in breeding (4% outbred - 16% in 1<sup>st</sup> cousin marriages), each person carries several deleterious recessive mutations

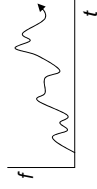
## Stochastic modelling of deleterious mutations

- frequency of disease alleles varies due to genetic drift
  - What is the frequency spectrum of disease mutations?
  - How old are disease mutations?
- Assume disease mutations neutral with respect to each other and death process
  - Population dynamics among disease chromosomes described by neutral model (Slatkin & Rannala 1997)
  - Sampling distribution under IAM = Ewens sampling formula



$$E[k] = \sum_{i=1}^n \frac{4N_e \mu_{del} i}{4N_e \mu_{del} + i - 1} \quad P(n_1, n_2, \dots, n_k | n, k) = \frac{n!}{k! i_1! i_2! \dots i_k!}$$

- BUT: Recurrent mutation (e.g. CpGs) and unequal mutation rates to different alleles



described by birth -

death process

neutral model

(Slatkin & Rannala 1997)

Sampling distribution under IAM = Ewens sampling formula

## Estimating allele age

- Treat the age of the mutation as a parameter
  - Thomson (1976): Branching process
  - Slatkin & Rannala (1997): Birth-death process
  - Distribution of number of descendants of a mutation conditional on geometric distribution

$t$  generations after origin =

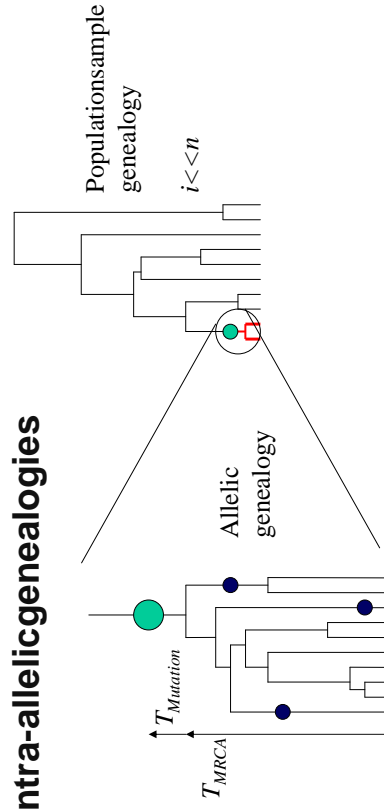
$$u = \frac{1 - e^{-st}}{1 - e^{-st}(1 - 2s)}$$

$$\hat{t} = \frac{1}{s} \ln \left[ \frac{4N_e s(i-1) + 1}{n} \right]$$

MLE of age (generations) given frequency alone =

- Example
    - $\Delta F508$  in CF: 70% of disease alleles (population frequency CF = 3%)
    - Estimated age = 2339 generations
- no selection in heterozygotes, population growth rate of 0.5%,  $N = 3 \times 10^8$

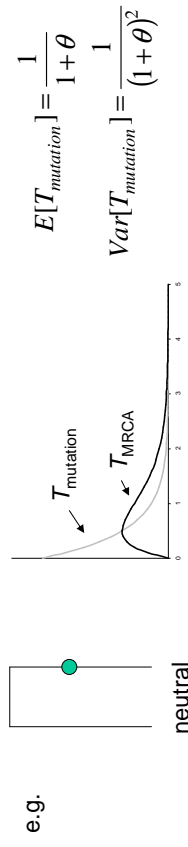
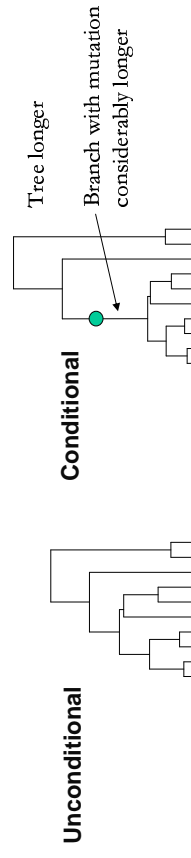
## Intra-allelic genealogies



- Treat intra-allelic genealogy as reconstructed (Nee *et al.* 1994) birth-death process determined by  $i, t$  and the proportion of populations sampled (Slatkin & Rannala 1997)
- Allelic variability determined by length in intra-allelic genealogy and mutation rate
- CF  $\Delta F508$ 
  - 46 mutations in  $\approx 1705$ , with a sample of 10<sup>-3</sup>% of the population
  - Estimated age of mutation = 146 generations (very different!)

## Estimating allele age

- Treat the age of a mutation as a random variable
  - Kimura & Ohta (1973), Griffiths and Tavaré (1998), Wiuf (2001)
- Conditional on observing a mutation, the genealogical process is distorted

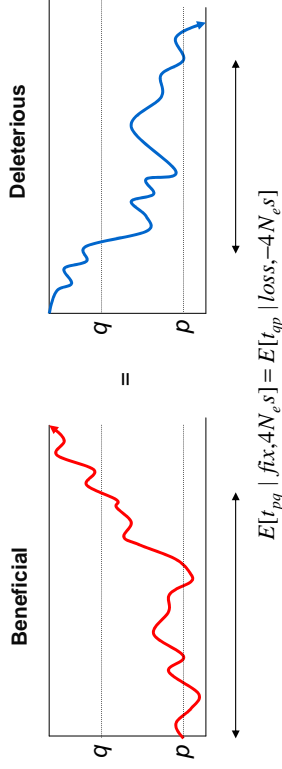


e.g.

neutral

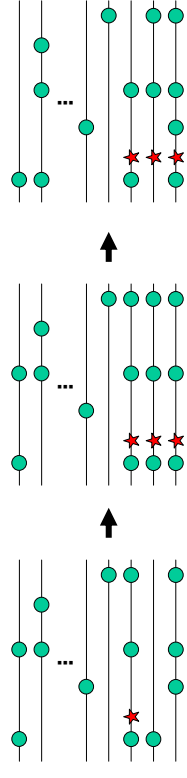
## Simulating selected genealogies

- Can use reversible nature of diffusion process to simulate time
  - Maruyama and Kimura (1971), Kimura (1973), Slatkin (2001)
- The processes of fixation of an advantageous mutation, and extinction of a deleterious mutation are reciprocal
  - e.g. the expected age of the mutation of a beneficial mutation is estimated from the expected time to loss of a deleterious mutation



- Genealogy of linked neutral alleles can be treated as a coalescent subdivision with variable population size

## The population genetics of association



The disease mutation arises on a particular genetic background

If the disease mutation increases in frequency by drift (or selection) the associated haplotype will also increase in frequency

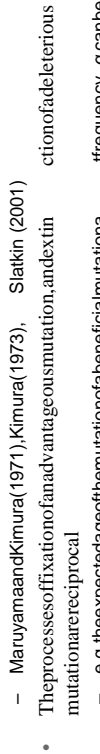
Over time the association between the disease mutation and linked mutations will decay by recombination

The probability that there has been no recombination between the ancestral marker (separated by a recombination fraction of  $r$ ) in the  $t$  generations since the mutation arose =

$$(1 - r)^t$$

## Fine-scale mapping of disease genes

- Pedigrees used to detect linkage

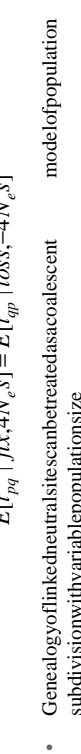


Power =  $f$  (#meioses, marker frequencies, penetrance, dominance)

Without huge pedigrees, scale of resolution =  $o(cM)$  = 100s genes

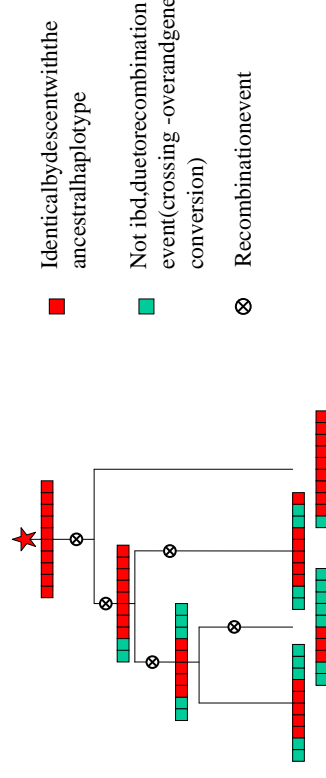


- Pedigrees in populations...
  - Can increase resolution (number of meioses) by considering population
  - Will only work if most cases share a particular disease mutation by descent (i.e. mutation arose only once)



## Identity by descent

- Identity by descent decays over time



- BUT even if sites are not identical by descent, they can be identical in state if they recombined by an ancestral chromosome carrying the same mutations
- ALSO sites with a common descent can be different in state through mutation

## A deterministic view of identity in state

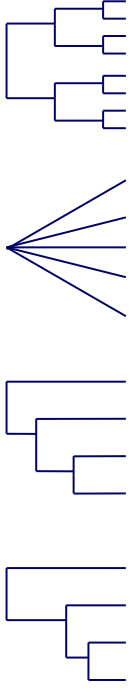
D	C	Time	Allele frequency among disease haplotypes ( $q_D$ )	Strength of association ( $q_D/q_C$ )
		0	1	$1/q_C$
		1	$(1-r) + r q_C$	$(1-r)/q_C + r$
		.	.	.
		t	$(1-q_C)e^{-rt} + q_C$	$(1/q_C - 1)e^{-rt} + 1$

$q_D$  = frequency of allele among disease haplotypes  
 $q_C$  = Population frequency of allele on ancestral disease haplotype

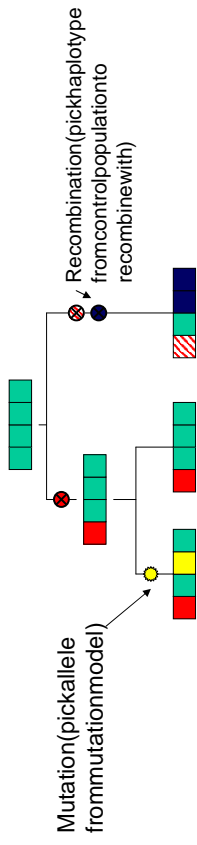
- Key points
  - Allelic association decays in an exponential fashion over time
  - Rare alleles captured by ancestral disease haplotypes will be more informative (though also less likely to be captured)
  - Generally, more alleles per marker means more power

## Modelling the stochastic nature of associations

- Haves to model
  - Stochastic generation of haplotype genealogy
  - Coalescent, branching process, star-genealogies, Luria-Delbruck



- Stochastic effects of recombination (and mutation?)

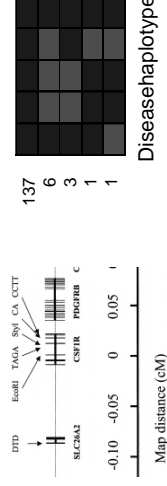


## Methods for fine-scale mapping

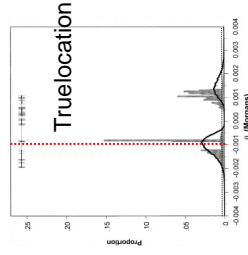
- Single-marker methods
  - Kaplan *et al.* (1995), Rannala & Slatkin (1998), ...
  - PROBLEM: non-independence between linked markers
- Haplotype methods with star-shaped genealogies (assume all disease chromosomes independent)
  - Terwilliger (1995), Xiong & Guo (1997), Graham & Thomson (1998), McPeck and Strahs (1999), Morris *et al.* (2000)
  - PROBLEM: non-independence due to shared genealogical history
- Haplotype methods with genealogical process
  - Rannala and Reeve (2001): Birth-death process
  - Morris *et al.* (2002): Shattered coalescent
- Main unknown is allelic heterogeneity at disease locus

## Example: DTD in Finland

- Diastrophic dysplasia (DTD) in Finland (Hästbacka *et al.* 1992)
  - Autosomal recessive disease (1-2% carriers)
  - High frequency, presumably due to founder event
  - Major mutation 90% Finnish disease chromosomes = GT → GC in 5' UTR

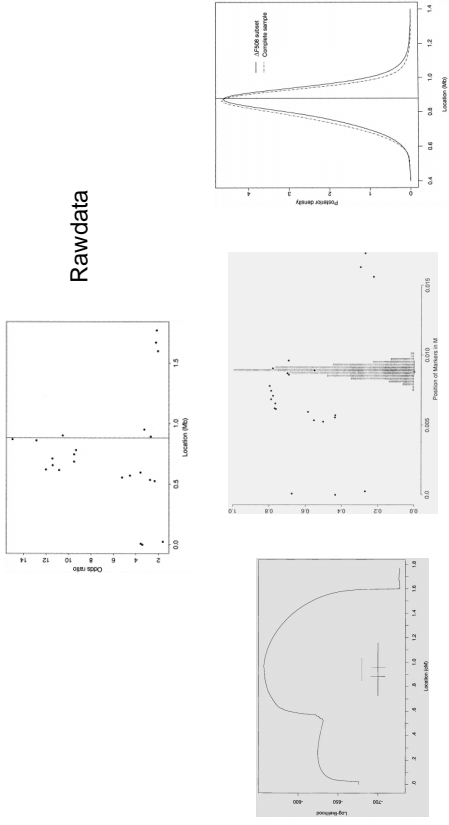


Marker	EcoRI	TAGA	SpyI	CA	CcTt
frequency in disease chromosomes	0.007	0.061	0.061	0.054	0
frequency in population	0.088	0.362	0.256	0.161	0.049
Deterministic point estimate of $rt$	0.08	0.19	0.27	0.41	0



## Example $\Delta F508$ in CTFR

- Most new methods for fine-scale mapping 'tested' on same data



Raw data

McPeck and Strahs (1999)

Liu *et al.* (2001)

Morris *et al.* (2001)

## Estimating genetic risk

- Sibling relative risk ratio
    - Ratio of increase in risk of disease given sibling has disease
- $$\lambda_s = \frac{K_s}{K} \leftarrow \begin{matrix} \text{Probability that a sibling of an affected individual is also affected} \\ \text{Population frequency of disease} \end{matrix}$$

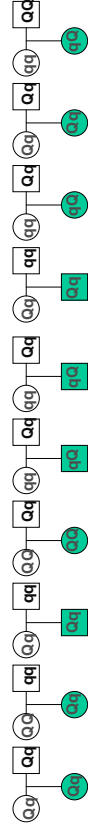
Relationship	Relatedness	$\lambda_s$ , Schizophrenia
MZ-twin	1.0	52.1
Offspring	0.5	10.0
DZ-twin	0.5	14.2
Full-sib	0.5	8.6
Half-sib	0.25	3.5
Niece/nephew	0.25	3.1
First Cousin	0.125	1.8
$K$		0.85%

## Complex (multifactorial) diseases

- Multiple interacting genes
- Low penetrance
- Non-genetic factors important
- Diagnosis can be problematic
- Examples
  - Alzheimer's disease
  - Schizophrenia
  - Hereditary heart disease
  - Asthma

## Family-based population methods

- Transmission-disequilibrium test (TDT)
  - Spielman *et al.* (1993)
  - Family-based controls
  - Classify markers in parents of probands as transmitted / untransmitted
  - Untransmitted alleles are internal controls



Transmitted allele	Untransmitted allele	
	Q	q
Q	a	b
q	c	d

Test statistic

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Test requires both linkage disequilibrium AND linkage

$$\frac{E[b-c]}{2N} = \frac{(1-2r)D_{PQ}}{x_D}$$

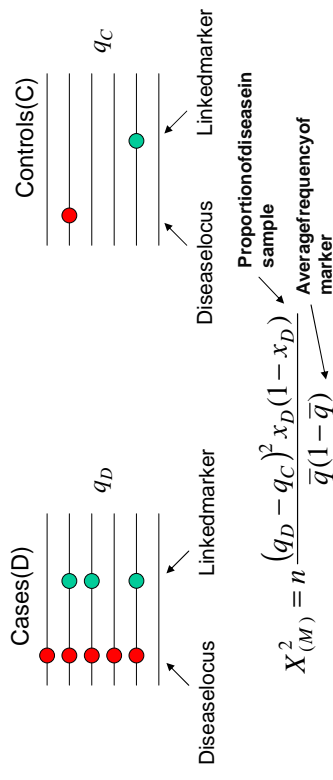
- But parents must be heterozygous and acquiring parents (even sib disease) difficult for late onset

## Association mapping

- Pedigree studies have limitations for mapping complex disease
  - Poor resolution
  - Low penetrance means low power
  - Strong pedigree signal may reflect rare Mendelian forms of complex disease (e.g. BRCA1 & BRCA2 mutations in breast and ovarian cancer)
- Detect genes/genomic regions associated with disease through association studies
  - Causal variants will be associated with the disease phenotype
  - Linked neutral variants will be associated with the disease phenotype through linkage disequilibrium with the causal variant
- The CDCV model for complex disease
  - Most genetic variants in complex disease are caused by a few rare mutations, common to most populations, but each of low penetrance

## Whole genome LD scanning

- Markers with intermediate frequency alleles identified from small panels
  - Common alleles likely to be ancient, hence shared by most populations, though frequencies may vary
  - Density such that associations between adjacent markers are weak
- Allelic associations assessed by comparing frequencies in case and control populations



## The relationship between LD and the power of association mapping

- The power to detect allelic association depends on the strength of association between disease mutations and linked neutral variants

$x_D$  = Population frequency of disease  
 $p_D$  = frequency of causative SNP in disease  
 $p_C$  = frequency of causative SNP in control  
 $q$  = Marker frequency in population  
 $D_{pq}$  = LD between causative SNP and marker in whole population  
 $D_{pq}^D$  = LD between causative SNP and marker in disease population  
 $D_{pq}^C$  = LD between causative SNP and marker in control population  
 $n$  = Sample size

$$X^2_{(M)} = n \frac{[D_{pq} - x_D D_{pq}^D - (1 - x_D) D_{pq}^C]^2}{q(1 - q)x_D(1 - x_D)} \times \frac{1}{(p_D - p_C)^2}$$

If  $p_D = 1, p_C = 0$

$$X^2_{(M)} = n \times r_{pq}^2$$

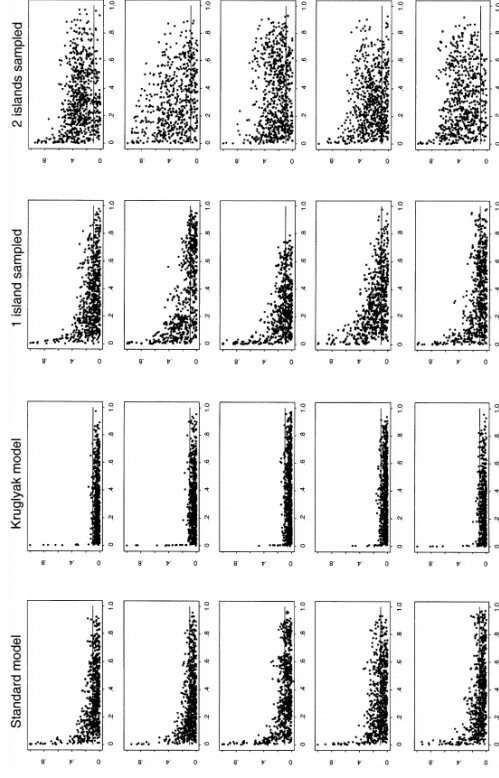
← **Square of correlation coefficient measure of LD**

$$r_{pq} = \frac{D_{pq}}{\sqrt{p(1 - p)q(1 - q)}}$$

## Factors affecting the choice of marker density

- Recombination
  - Instable populations LD determined by  $4 N_e r$  (balance between recombination and genetic drift)
  - Heterogeneity in recombination rate, inversions, deletions
- Allele frequency
  - Recent alleles have less opportunity for evolutionary recombination
  - Rare alleles can have little power to detect associations
- Demographic history
  - Population growth leads to a decrease in LD
  - Population structure, admixture and bottlenecks increase LD
  - Demographic LD may make LD due to linkage hard to detect
- Stochasticity
  - For any given set of parameters, allelic associations are very variable

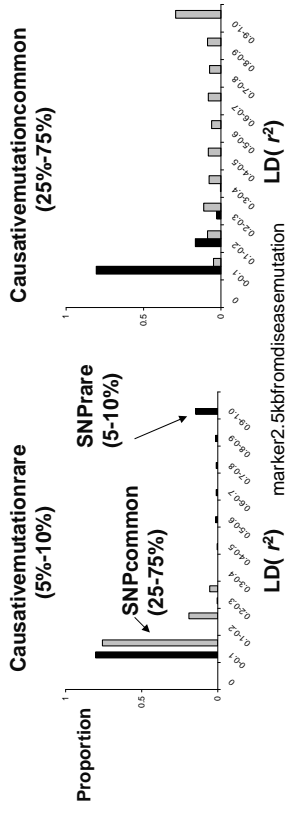
## Modelling LD with coalescent simulations



Pritchard and Przeworski (2001)

## Common SNPs for common disease?

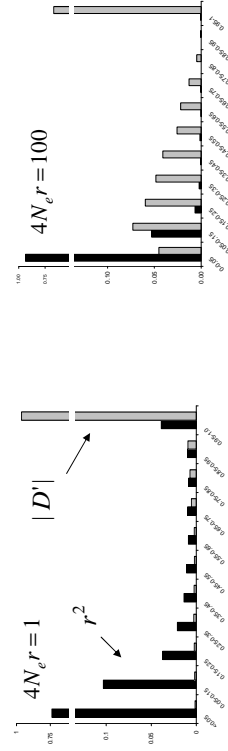
- Current methodology focuses on SNPs at intermediate frequency common to most populations
- BUT Ancient SNPs will tend to have experienced much recombination
  - Are rare
  - Address by simulating LD between common and rare SNPs



Answer depends on frequency of causative SNP

## What does linkage disequilibrium measure?

- Multiple statistics of LD
  - $D$ : standard coefficient
  - $r^2$ : ( $\Delta^2$ ,  $\rho^2$ ) correlation coefficient (Hill & Robertson 1968)
  - $|D'|$ : Ranganormalised by allele frequencies (Lewontin 1964)
  - Others, e.g.  $\delta$ : Odds ratio formulation (Devlin & Risch 1995)



- Statistics found lack of information about recombination with either evidence for recombination ( $r^2$ ) or evidence for strong LD ( $|D'|$ )
- Model-based methods that use mutation to infer the missing data of (e.g. coalescent likelihoods) are preferable

## Which population?

- Small populations
  - Founders events create genetic homogeneity and strong linkage disequilibrium (e.g. Iceland)
  - Small populations have more extensive LD (e.g. Sardinia)
- Large populations
  - Low LD due to historical recombination
  - Genetic heterogeneity high
- BUT
  - Founders events may have captured alleles of small effect in large population (assumed to be of interest)
  - GxEx interactions may play an important role

## Linkage disequilibrium and admixture

- Admixture generates linkage disequilibrium between unlinked markers if there are differences in allele frequency between the source populations

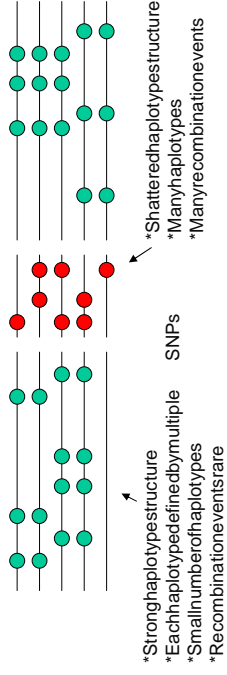
$$D_{pq(t)} = (p_1 - p_2)(q_1 - q_2)x_1x_2(1 - r)^t$$

LD Difference in allele frequency contribution of each source population Times since admixture recombination fraction

- Examples
  - Jamaica, South Africa, UK, Iceland...
- Admixture (or stratification) can lead to false positives
  - Use LD between unlinked markers to test for population admixture (e.g. Pritchard and Rosenberg 1999)
- Use the linkage disequilibrium generated by admixture to *aid* association mapping
  - Chakraborty & Weiss (1988), McKeigue (1997)
  - Though will need to correct for high background LD
  - Stratification may have hidden GxE interactions
  - Epistatic interactions may be difficult to detect

## The implications of block haplotype structure

- Some evidence that recombination, or at least LD, is not uniform along the genome (e.g. Yu *et al.* 2001, Tallon-Miller *et al.* 2000, Daly *et al.* 2001)



- SNP density within blocks can be much lower than within regions of high recombination – as diversity captured by few, well chosen markers
- BUT
  - Recombination haplotypes may provide considerable information
  - Distribution of blocks may vary between populations if causes is demographic
  - Block distribution has to be assessed empirically from dense SNP maps