

Population genetic inference

Date	Topic	
22 nd Jan	Good questions in population genetics	GM
29 th Jan	Principles of population genetic inference	GM
5 th Feb	Recombination in the coalescent	JH
12 th Feb	Natural selection	GM
19 th Feb	Demographic models	GM
26 th Feb	Combinatorics of the coalescent	JH
5 th March	Population genetics of disease mutations	GM
12 th March	Linkage disequilibrium and association mapping	GM

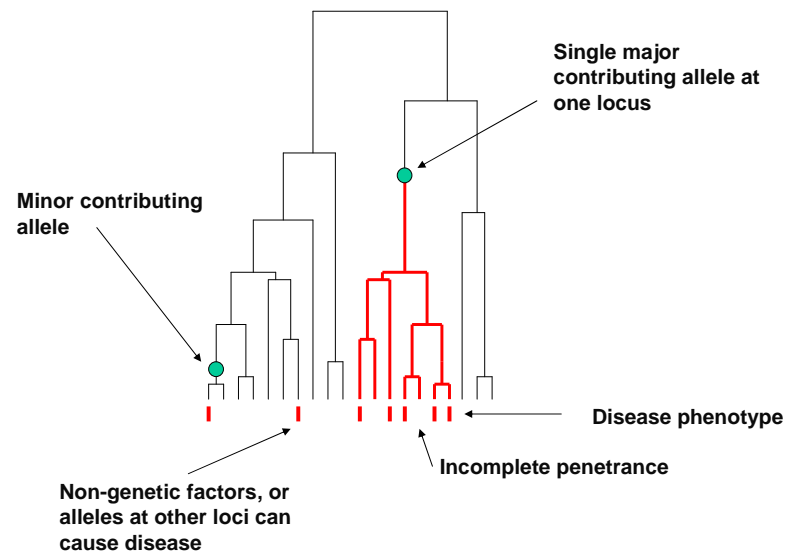
Reading

- Cardon LR & Bell J (2001) Association study designs for complex diseases. *Nature Rev. Genet.* **2**: 91-99
- Pritchard J and Przeworski M (2001) Linkage disequilibrium in humans: models and data. *Am J. Hum. Genet.* **69**: 1-14
- Weiss KM & Clark AG (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**: 19-24.

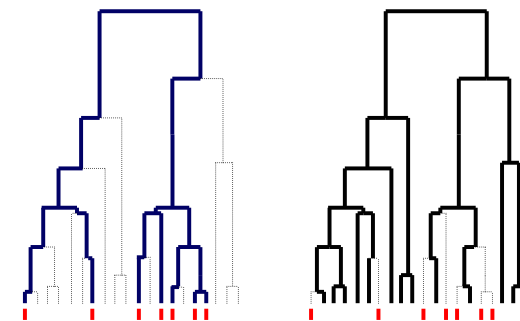
Association mapping of complex disease

- Pedigree studies have limitations for mapping complex disease
 - Low penetrance of complex disease genes means vast amounts of data are needed to identify candidate regions
 - Low resolution of mapping
 - Strong pedigree signal may reflect rare Mendelian forms of complex disease (e.g. BRCA1 & BRCA2 mutations in breast and ovarian cancer)
- Family-based population methods (e.g. TDT) have restrictive design
 - Requires 3x genotyping for a single data point
 - Requires parents (difficult for late onset)
 - Parents must be heterozygous to be informative
- Detect genes/genomic regions associated with a disease through allelic associations in case-control 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

A genealogical view of complex disease

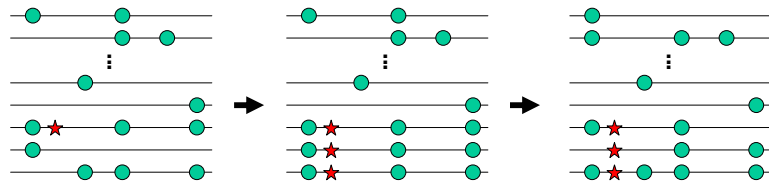


Case-control genealogies



- Rapid coalescence of a subset of case chromosomes generates
 - Strong identity by descent within disease chromosomes
 - Strong associations between disease mutations and linked markers
- BUT allelic heterogeneity, multiple causative factors and incomplete penetrance reduce both effects

The population genetics of associations

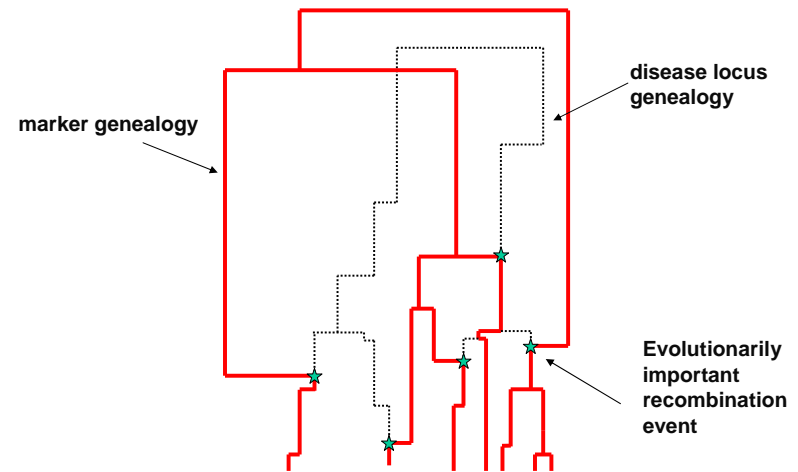


The disease mutation arises on a particular genetic background

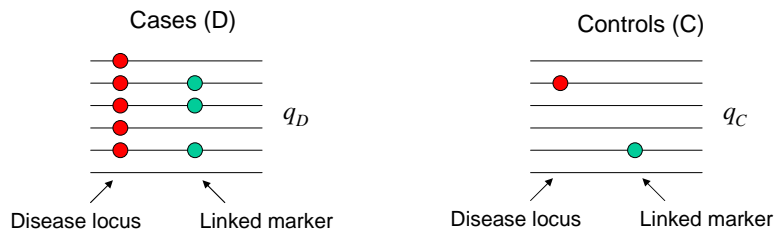
If the disease mutation increases in frequency by drift (or selection) the associated haplotype will also increase in frequency

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

Gene genealogies at linked markers



Measuring associations



$$X_{(D)}^2 = n \frac{(q_D - q_C)^2 x_D (1 - x_D)}{\bar{q} (1 - \bar{q})}$$

Proportion of disease in sample x_D
Average frequency of marker \bar{q}

Q: How do we maximise power?

Deterministic solution: $E[X_{(M)}^2] = (k - 1)e^{-2rt}$

k = Number of alleles at marker locus

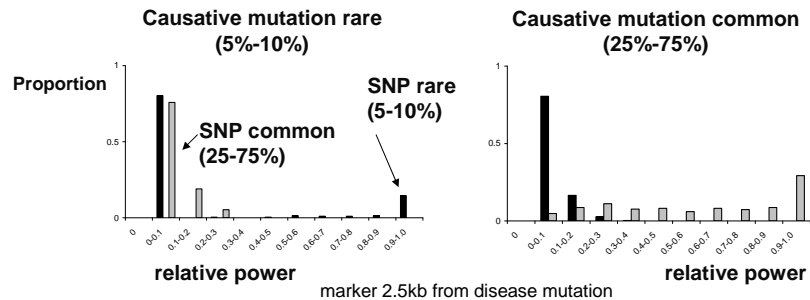
Factors affecting test power

- Recombination
 - Tighter linkage means associations persist longer
- Age of disease mutation
 - Old mutations have had more chances to recombine
- Marker locus allele frequencies
 - Rare alleles highly informative if captured by ancestral disease mutation
 - BUT less likely to be captured
- Marker locus allele number
 - Deterministic solution implies allele number is the most important factor
 - BUT genetic drift in allele frequencies will be important for ancient mutations
 - AND variance of associations can vary considerably with marker allele frequencies and age of disease mutation

Need to model stochastic nature of allelic associations

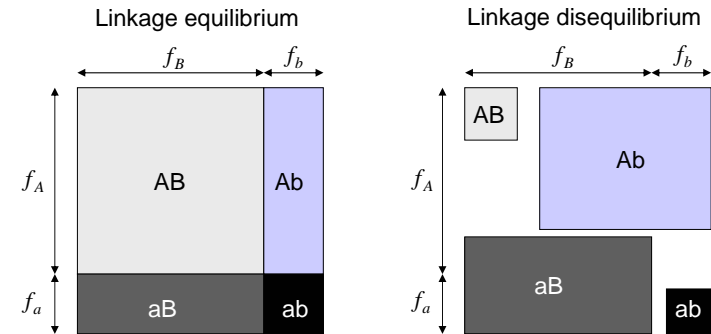
Common SNPs for common disease?

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



Answer depends on frequency of causative SNP

Allelic associations and linkage disequilibrium



$$D_{AB} = f_{AB} - f_A f_B$$

$$= -D_{Ab} = -D_{aB} = D_{ab}$$

Measuring LD

- The expectation of $D = 0$
- Correlation coefficient measure [0,1]
 - Hill & Robertson (1968)
$$r_{AB}^2 = \frac{D_{AB}^2}{f_A f_a f_B f_b} = \rho_{AB}^2$$
- Range constrained by allele frequencies [0,1]
 - Lewontin (1964)
$$|D'_{AB}| = \begin{cases} \frac{D_{AB}}{\min(f_A f_B f_a f_b)} & \text{if } (D > 0) \\ \frac{-D_{AB}}{\min(f_A f_B f_a f_b)} & \text{else} \end{cases}$$
- Odds-ratio formulation
 - Devlin & Risch (1995)
$$\delta_{AB} = \frac{D_{AB}}{f_B f_{ab}}, D_{AB} > 0$$
- Statistics confound lack of information about associations with either evidence for either absence (r^2) or presence ($|D'|$) of association

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

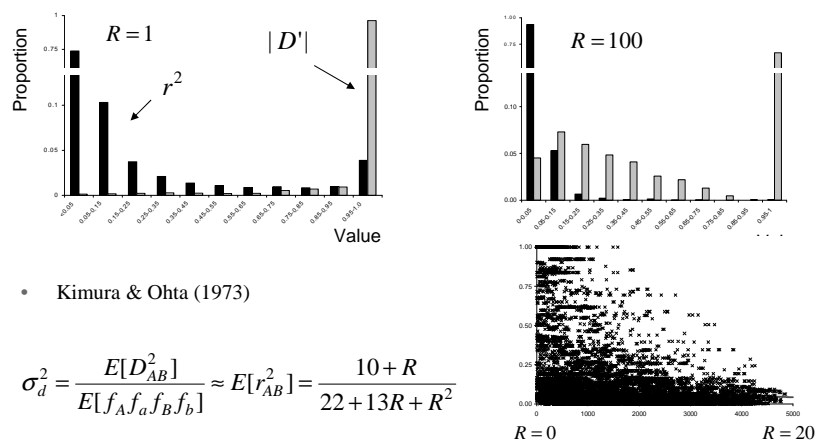
x_D	= Population frequency of disease
p_D	= frequency causative SNP in disease
p_C	= frequency causative SNP in control
q	= Marker frequency in population
D_{pq}	= LD between causative SNP and marker in whole population
D_{pq}^D	= LD between causative SNP and marker in disease population
D_{pq}^C	= LD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

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

If $p_D = 1, p_C = 0$ $X_{(M)}^2 = n \times r_{pq}^2$ ← Square of correlation coefficient measure of LD

The distribution of LD in WF populations

- Coalescent simulations can be used to estimate the distribution of LD under various demographic models
- In constant size populations, key quantity is the compound parameter $R = 4N_e r$



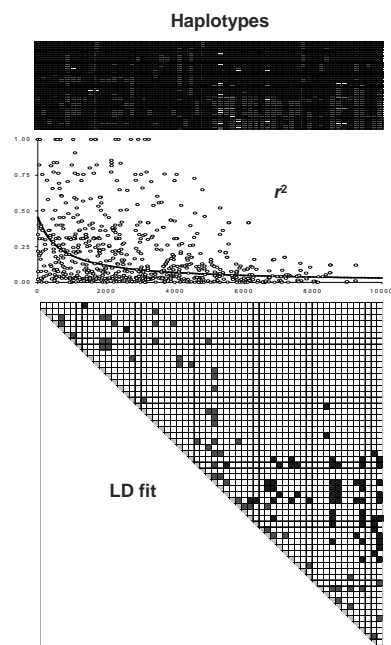
Estimating $4N_e r$

- Fit curve to plot of LD
 - Chakravati *et al.* (1984)
- Variance in pairwise differences
 - Hudson (1987)
- Full likelihood inference
 - Griffiths & Marjoram (1996), Fearnhead and Donnelly (2001)
- Summary-statistic inference
 - HRM: Wall (2000)
- Approximate likelihood inference
 - Composite likelihoods: Hudson (2001), McVean *et al.* (2002)

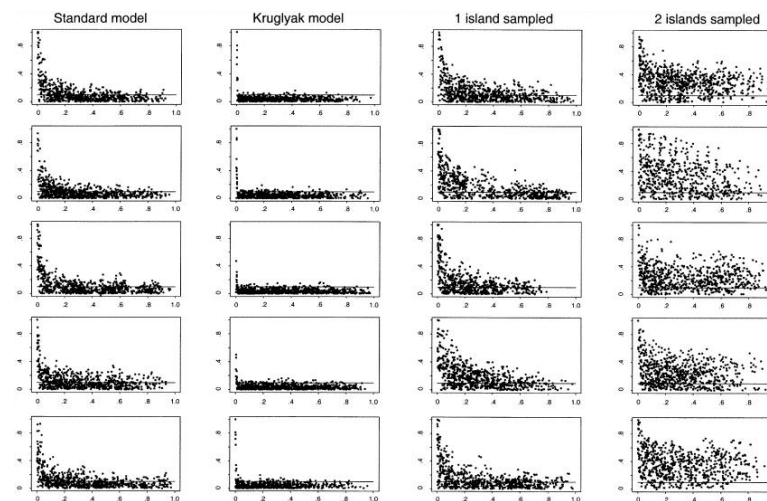
Example: LPL

- 9.7 kb region sequenced in 48 African American chromosomes (Nickerson *et al.* 1998)
- Using $N_e = 10,000$ $r = 1\text{cM/Mb}$: $R = 4$
- Using composite likelihood approach: $\hat{R} = 29$
- BUT
 - Evidence that constant population size, constant recombination rate model does not fit data

- Too little LD
- Too much LD



LD under different demographic models



LD and genealogical histories

- Under infinite-sites assumption, LD measures correlations in genealogical history
 - McVean (2002)

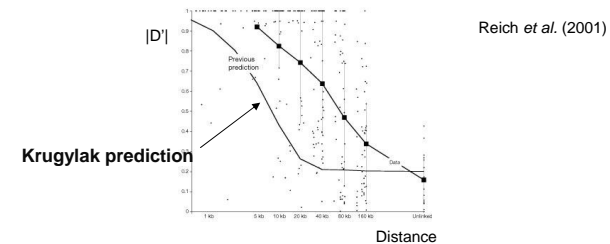
$$\sigma_d^2 = \frac{\rho_C(\tau_x, \tau_y) - 2\rho_T(\tau_x, \tau_y) + \rho_D(\tau_x, \tau_y)}{E[\tau]^2 / \text{Var}(\tau) + \rho_D(\tau_x, \tau_y)}$$

$\rho(\tau_x, \tau_y)$ = Covariance in coalescence time at loci x and y
 $E[\tau]$ = Expected coalescence time for a pair of sequences
 $\text{Var}(\tau)$ = Variance in coalescence time for a pair of sequences

- Correlations mainly depend on recombination
- Mean and variance of τ strongly affected by demographic processes

Empirical patterns of LD

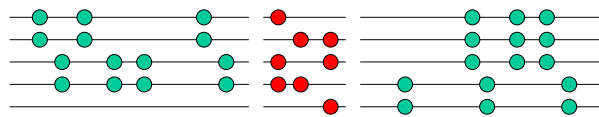
- Large-scale surveys of LD in humans
 - e.g. Huttley *et al.* (1999), Abecasis *et al.* (2001), Reich *et al.* (2001)
 - LD extends over considerable distance (>>10kb) in most populations



- African populations show less LD than European populations (e.g. Frisse *et al.* 2001)
- Small, isolated populations (e.g. Saami, Evenki) show increased LD
- Founder populations (e.g. Finland, Sardinia) do not always show increased LD (e.g. Eaves *et al.* 2000)
- Evidence for heterogeneity in LD along chromosomes
 - Taillon-Miller *et al.* (2000), Jeffreys *et al.* (2001), Daly *et al.* (2001), Patil *et al.* (2001)
 - Block haplotype structure

The implications of block haplotype structure

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



- * Strong haplotype structure
- * Each haplotype defined by multiple SNPs
- * Small number of haplotypes
- * Recombination events rare
- * Shattered haplotype structure
- * Many haplotypes
- * Many recombination events

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

Which population?

- Small populations
 - Founder events create genetic homogeneity and strong linkage disequilibrium (e.g. Iceland)
 - Small populations have more extensive LD (e.g. Saami)
- Large populations
 - Low LD due to historical recombination
 - Genetic heterogeneity high
- BUT
 - Founder events may have captured alleles of small effect in large population (assumed to be the one of interest)
 - G x E 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 Differences in allele frequency contribution of each source population recombination fraction Time since admixture

The diagram shows the equation $D_{pq(t)} = (p_1 - p_2)(q_1 - q_2)x_1x_2(1-r)^t$. Below the equation, four labels are placed: 'LD' under the first term, 'Differences in allele frequency' under the second term, 'contribution of each source population' under the third term, and 'recombination fraction' under the fourth term. To the right, 'Time since admixture' is written with an arrow pointing to the exponent t .

- 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 G x E interactions
 - Epistatic interactions may be difficult to detect

Big questions for association mapping

- Are common variants responsible for common disease?
 - Or are multifactorial diseases influenced by many rare mutations at many loci?
- Are single mutations at disease loci responsible for most variation?
 - Or is allelic heterogeneity a problem?
- Is a marker spacing of X kb (3-50kb) sufficient to capture associations?
 - Or is a much finer map needed to allow for the stochastic nature of associations?
- Does demographic LD help association mapping?
 - Or are the complexities introduced by admixture and structure more hindrance than help?
- Can global haplotype diversity be captured by few well-chosen markers?
 - Or are population differences in block haplotype structure overwhelming?