

Medical applications of  
population genetics

Jonathan Pritchard  
Department of Statistics  
University of Oxford

`pritch@stats.ox.ac.uk`

`www.stats.ox.ac.uk/~pritch/home.html`

# Outline

This lecture will provide a rapid introduction to the techniques used in mapping human disease genes, with particular emphasis on aspects that rely on population genetics.

Identifying genes that contribute to complex diseases is now one of the major research areas of human genetics.

- Background
  - Mendelian disorders
  - multifactorial disorders
  - non-genetic techniques for studying inheritance
- Introduction to gene mapping techniques
  - linkage methods
  - association/LD mapping
  - fine mapping based on LD
- Genome-wide association mapping
  - rationale
  - current status
- Association mapping and population structure
  - problems and solutions

# Mendelian disorders

E.g, Sickle-cell anaemia, Huntingdon's disease, Myotonic dystrophy, Haemophilia, Cystic Fibrosis, etc.

- single gene
- may be either dominant or recessive
- usually highly penetrant—ie, having the disease genotype virtually ensures having the phenotype
- various mutation mechanisms: eg insertions/deletions, unstable triplet repeats, mutations at key amino acids
- often strong selection against disease alleles

Mutations for Mendelian diseases are usually at low frequency in the population. In most cases they seem to be maintained by mutation/selection balance.

E.g., Haemophilia B (X-linked recessive disease). Strong selection against the disease in males; mutant allele frequency  $\approx 4 \times 10^{-5}$ . Many independent mutations (Green et al 1999).

Possible exception (balancing selection?): Cystic Fibrosis (recessive). Mutant allele frequency of 1.7% in Europeans. A single mutation,  $\Delta F508$ , makes up 70% of mutant alleles in Europe.

# Complex (Multifactorial) diseases

It is now technically straightforward to find the genes responsible for Mendelian disorders, and many of these have now been mapped.

But Mendelian diseases are rare compared to complex diseases: eg, diabetes, hypertension, schizophrenia, asthma, etc.

Our understanding of complex diseases is still very limited:

- multiple interacting genes
- low penetrance
- non-genetic factors are important
- diagnosis can be problematic

So far, there has been limited success in finding “complex disease genes”. One gene that has been found is ApoE, which contributes to Alzheimer’s.

There are also some highly penetrant genes which give rise to Mendelian forms of complex diseases—eg BRCA1 and BRCA2 which lead to breast and ovarian cancer.

# Non-genetic techniques

Prior to performing genetic studies, we can learn a bit about the genetic basis of disease by studying familial inheritance. [Note that this is usually confounded with shared environmental factors.] Define

$K$  as the population frequency of a disease;

$K_s$  as the probability that a sibling of an affected individual (proband) is affected. [Similarly, we can look at other relatives.]

Then set

$$\lambda_s = \frac{K_s}{K}.$$

$\lambda_s$  is known as the *sibling recurrence risk ratio*, and measures the proportionate increase in risk to siblings. If  $\lambda_s = 1$  we might conclude that there is no genetic component.

**Example:** Schizophrenia;  $K = 0.85\%$  [Risch 90a].

Relationship	Estimated value of $\lambda$
MZ-twin	52.1
offspring	10.0
DZ-twin	14.2
sibling	8.6
half-sib	3.5
niece/nephew	3.1
cousin	1.8

# Gene mapping techniques

**Linkage analysis:** goal is to find genetic markers which are passed through a pedigree along with the disease of interest.

For example, suppose that we collect DNA from individuals in the following pedigree. We would genotype a series of microsatellites spread across the genome, and test whether the transmission of any these is correlated with the transmission of the phenotype.

Current experiments typically genotype each individual at 300-400 microsatellite loci, spread across the genome (ie, every 7-10 cM).

# Association/LD mapping

An alternative approach involves direct testing of DNA polymorphisms for *association* with the disease:

**Association mapping:** If a particular mutation increases susceptibility to a disease of interest, we should find that it is *more common among affected individuals than among random controls*.

**Linkage Disequilibrium mapping:** Due to LD, random markers *near* a disease susceptibility mutation may also be associated with the disease.

Association/LD mapping are used in several different contexts:

- Testing for association at a candidate gene
- Examining regions of “suggestive linkage” from pedigree studies
- Genome screens for association [in the future?]
- Narrowing down the location of a mutation

# Fine mapping by LD

- Suppose that a particular disease is caused by a recessive mutation that arose once in the history of the population. We start by using linkage methods to identify a region of perhaps 10 cM that contains the gene.
- We then genotype a series of microsatellites and SNPs in this region to narrow down the location of the gene using LD.
- The population genealogy is like an unobserved pedigree. Within a typical pedigree, we don't see enough recombination events to narrow down the location very well. LD mapping allows much finer resolution.
- Eventually, we hope to narrow the region down enough to be able to sequence the candidate region in multiple individuals, and look for possible functional mutations.



# Fine mapping for complex traits

Fine mapping for complex traits is considerably more difficult than in this example, for various reasons:

- Mutations for complex traits will generally have low penetrance:
  - Not all chromosomes from affected individuals will have a mutation at this locus
  - Some control chromosomes *will* carry the mutation
- There may be multiple independent mutations
- Often we may not know haplotype phase

The problem then is to infer location of a disease mutation, using data from SNPs or microsatellites, taking into account the evolutionary history of the locus, which induces correlations among the data (between chromosomes, and between markers).

# Inferring location of mutations, contd.

Several statistical approaches have been proposed to infer location. These include:

- Terwilliger (1995) proposed an approximation to the likelihood which is based on the disequilibrium coefficient at each marker.
- Lazzeroni (1997) described an approach which fits regression lines through the disequilibrium coefficients to infer the most likely location.
- McPeck and Strahs (1999) and Morris et al (2000) both used a hidden Markov model based on approximations to the coalescent to infer location.

# Inferring location of mutations, contd.

Comments:

- Use of disequilibrium coefficients seems to have limited promise:
  - likely to be inefficient, because it ignores the haplotype structure
  - ignores correlations in the data and hence may actually be misleading
- In principle, we would like to be able to analyse a full population genetics model of the coalescent process with recombination and mutation (and unknown phase?). However, this is computationally daunting.
- In the short term, approximations to the coalescent process (eg Strahs & McPeck; Morris et al) may be sensible, but it is unknown how well these perform in practice.
- Little attention has been given to the difficult problem of allelic heterogeneity.

# Genome-wide association mapping

One advantage of LD mapping is that it allows much finer resolution of location than possible with linkage methods.

Another advantage—greater statistical power—was pointed out in an influential paper by Risch & Merikangas (1996).

*Sample sizes needed to achieve 80% power using linkage and association:*

Genotypic risk factor	Allele frequency	Linkage (# families)	Association (# families)
4.0	0.01	4260	235
	0.10	185	48
	0.50	297	61
2.0	0.01	296710	1970
	0.10	5382	264
	0.50	2498	180

Risch & Merikangas argued that the way to map complex disease genes of small effect is by whole-genome association mapping [but various key assumptions!].

## Practical issues...

- Need a dense map of polymorphic markers across the genome
  - now available: 1.4 million SNPs
- Need cheap and effective SNP genotyping
  - in progress?
- How much LD is there in the human genome?
  - What density of SNPs across the genome do we need in order to have good power to detect association?
- Will allelic heterogeneity be a serious problem?

Kruglyak (1999) used a population genetics model to argue that useful levels of LD might only extend over very short distances ( $\tilde{3}$ KB), which would require an unreasonable number of markers to cover the genome.

Recent empirical work, however, is more encouraging, finding LD over 100s of KB (eg Taillon-Miller et al, 2000; Abecasis et al, 2001).

# LD mapping and population structure

The presence of population structure can lead to linkage disequilibrium between unlinked markers (and hence a large Type I error rate). This is because both disease rates and gene frequencies can vary widely across ethnic groups.

## Example...

Consider a sample which contains a mixture of two ethnic groups. Each individual is genotyped at a marker with two possible types:  $A$ , and  $a$ .

- Suppose that the frequency of the disease is higher in Group 1 than in Group 2. Then Group 1 will be over-represented in the sample of cases.
- Suppose that the frequency of  $A$  is higher in Group 1 than in Group 2.

Then the expected frequency of allele  $A$  will be higher in cases than in controls, leading to an *apparent* association.

# The TDT

One solution to the problem of structure makes use of *family-based* controls:

- collect affected offspring and their parents
- At each locus, classify parental alleles as “transmitted” (to the offspring) or “untransmitted”.
- Now, the untransmitted alleles are effectively matched controls.

This approach was outlined most clearly by Spielman et al (1993), in their TDT (transmission disequilibrium test):

Transmitted allele	Untransmitted Allele	
	$M_1$	$M_2$
$M_1$	$a$	$b$
$M_2$	$c$	$d$

Let  $N$  be the total number of families.

# Transmission probabilities:

Transmitted allele	Untransmitted Allele	
	$M_1$	$M_2$
$M_1$	$m^2 + \frac{m\delta}{p}$	$m(1 - m) + \frac{(1-\theta-m)\delta}{p}$
$M_2$	$m(1 - m) + \frac{(\theta-m)\delta}{p}$	$(1 - m)^2 - \frac{(1-m)\delta}{p}$

$m$ =marker allele frequency

$p$ =disease frequency

$\theta$ =recombination distance (!!)

$\delta$ =LD (D) between disease mutation and marker

Then

$$\frac{E(b - c)}{2N} = \frac{(1 - \theta - m) - (\theta - m)\delta}{p} = \frac{(1 - 2\theta)\delta}{p}.$$

In the absence of linkage,  $\theta = 0.5$ , and hence  $1 - 2\theta = 0$ .

Hence, Spielman et al propose the test statistic

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

which has asymptotically a  $\chi^2$  distribution with 1 d.f. Note that this test rejects the null only in the presence of :

- Linkage ( $\theta \neq 0.5$ ) and
- Allelic association ( $\delta \neq 0$ ).



# Case-control alternatives

There are often good reasons for preferring case-control studies with unrelated controls:

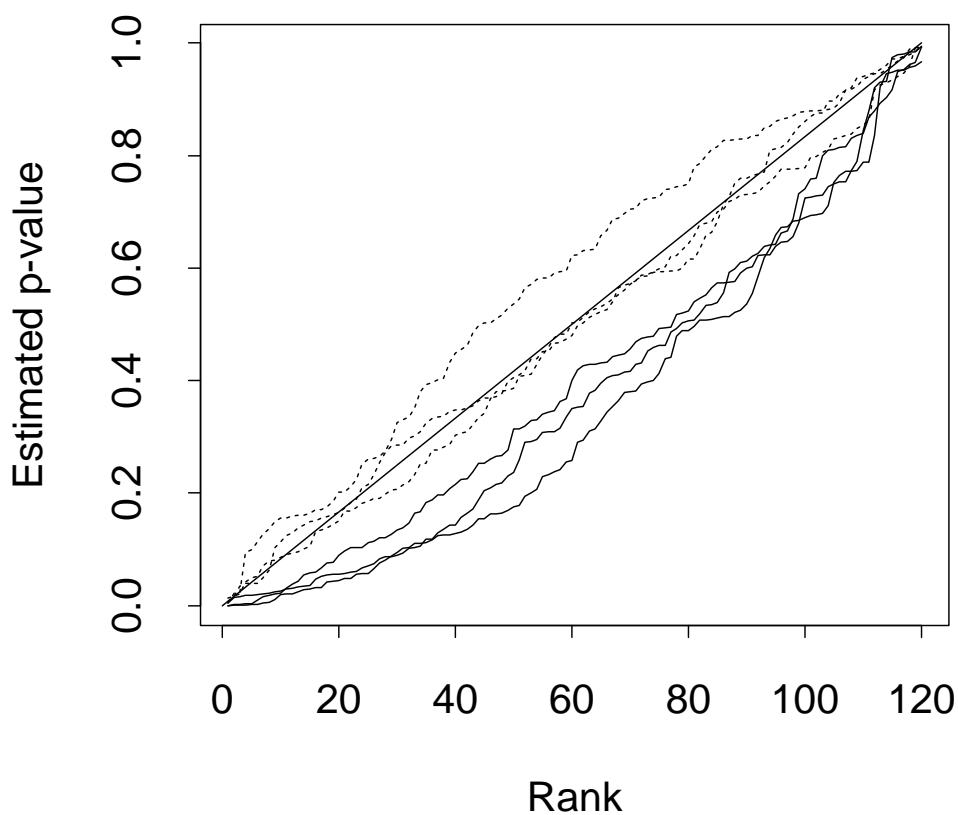
- May be difficult to collect parents or other family members (esp. for late-onset diseases)
- Genotyping controls is more efficient
- Possibility of re-using controls for different studies

Recent work has focused on the following idea:

- If we find an association at a candidate locus, we don't know whether it is due to a linked mutation, or due to population structure.
- But population structure should have a similar effect on loci across the genome.

## Case-control alternatives, contd.

- Pritchard and Rosenberg (1999) propose a test for population structure based on genotyping a series of unlinked marker loci in the same individuals.
- Devlin and Roeder (1999) describe a related test (for SNPs) which computes a correction factor for the effect of population structure. Briefly, they propose a parametric correction for the effect of population structure, based on unlinked marker data.
- Pritchard et al (2000a, 2000b) develop a method for inferring the details of population structure in a sample, and assigning individuals to populations, before testing for association.



# Summary

- Mapping complex disease loci will be a major challenge for human genetics in the coming years.
- There are important population genetic problems in this field. These include the development of:
  - effective methods for inferring the location of disease mutations
  - more powerful tests of association, particularly for the problem of allelic heterogeneity
  - methods that enable valid analysis of case-control tests of association.

# Bibliography

- Abecasis, G., Noguchi, E., Heinzmann, A., Traherne, J., Bhattacharya, S. and *et al*, N. L. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68**, 191–197.
- Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
- Giannelli, F., Anagnostopoulos, T. and Green, P. (1999) Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental mutations inferred from Hemophilia B. *Am. J. Hum. Genet.*, **65**, 1580–1587.
- Green, P., Saad, S., Lewis, C. and Giannelli, F. (1999) Mutation rates in humans. I. Overall and sex-specific rates obtained from a population study of Hemophilia B. *Am. J. Hum. Genet.*, **65**, 1572–1579.
- Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*, **22**, 139–144.
- Lazzeroni, L. C. (1997) Empirical linkage-disequilibrium mapping. *Am. J. Hum. Gen.*, **62**, 159–170.
- McGuffin, P. and Huckle, P. (1990) Simulation of Mendelism revisited: the recessive gene for attending medical school. *Am. J. Hum. Genet.*, **46**, 994–999.
- McPeck, M. and Strahs, A. (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.*, **65**, 858–875.
- Morris, A., Whittaker, J. and Balding, D. (2000) Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am. J. Hum. Genet.*, **67**, 155–169.
- Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Gen.*, **65**, 220–228.

- tion structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.
- Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.*, **46**, 222–228.
- Risch, N. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
- Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
- Risch, N., Spiker, D., Lotspeich, L., Nouri, N. and *al*, D. H. (1999) A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.*, **65**, 493–507.
- Slatkin, M. and Rannala, B. (1997) The sampling distribution of disease-associated alleles. *Genetics*, **147**, 1855–1861.
- Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–513.
- Taillon-Miller, P., Bauer-Sardina, I., Saccone, N., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. and Kwok, P. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics*, **25**, 324–328.
- Terwilliger, J. D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.*, **56**, 777–787.