Genome 559
Intro to Statistical and Computational Genomics
2009

Lecture 18a:

LD & Association

Larry Ruzzo

(Thanks again to Mary Kuhner for slides)

Mapping in population data

- Basic idea: look for differences between cases and controls
- Problems:
  - Cases and controls must come from the same population
  - Can't use multiple cases from the same family without a correction
  - Requires linkage disequilibrium (LD) between trait and marker, not just linkage

Why would we need LD?

- If we test the actual disease causing mutation, we don't need any LD
- We don't get lucky like this very often
- Usually this only happens when there is some reason to expect that a disease is caused by a specific candidate gene
- Otherwise, we must rely on markers, which means we need LD

Linkage disequilibrium

- Consider a marker locus with alleles A and B and a disease locus with alleles D and H
- If there is no linkage, $p(AD) = p(A)p(D)$
- Even if there is linkage, this may still be true in a population
- In each family the two loci are linked
- But in some families A goes with D and in others A goes with H
- This is linkage equilibrium

Linkage disequilibrium (LD)

- To map in a population we need non-random association between a marker allele and a disease locus allele: linkage disequilibrium

- How could this come about? Useful way:
  - There is linkage between disease locus and marker
  - The disease mutation is relatively recent
  - The disease allele is therefore mainly still on its original haplotype
  - There is positive LD between the new allele and the original haplotype

- Not so useful way:
  - The disease allele and the marker allele are both common in the same subpopulation
  - If your population is heterogeneous enough, you can even see LD between unlinked loci!

Example of unhelpful LD

- Eastern Europeans of Jewish descent have different allele frequencies than other Eastern Europeans

- There are several diseases, such as Tay-Sachs, which are more common in the EE Jewish population than elsewhere

- Population mapping on random Eastern European samples:
  - Every disease common in EE Jews is in LD with loci common in EE Jews
  - No useful map produced

- Successful mapping possible if the populations are carefully sorted

Example of unhelpful LD

- Many individuals may have to be disregarded because their ethnicity is mixed or unclear
- Family studies may be better in highly mixed populations as they don’t rely on LD, only linkage

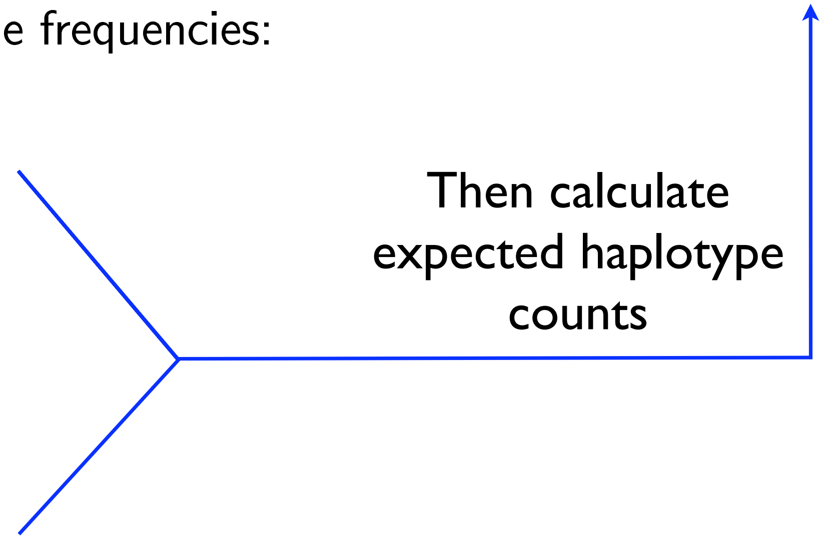
Statistical test for LD

We sampled 200 haplotypes:

Haplotype	Observed	Frequency	Expected
AB	76	0.38	56
Ab	64	0.32	84
aB	4	0.02	24
ab	56	0.28	36

First, calculate allele frequencies:

- $P(A) = 0.7$
- $P(a) = 0.3$
- $P(B) = 0.4$
- $P(b) = 0.6$

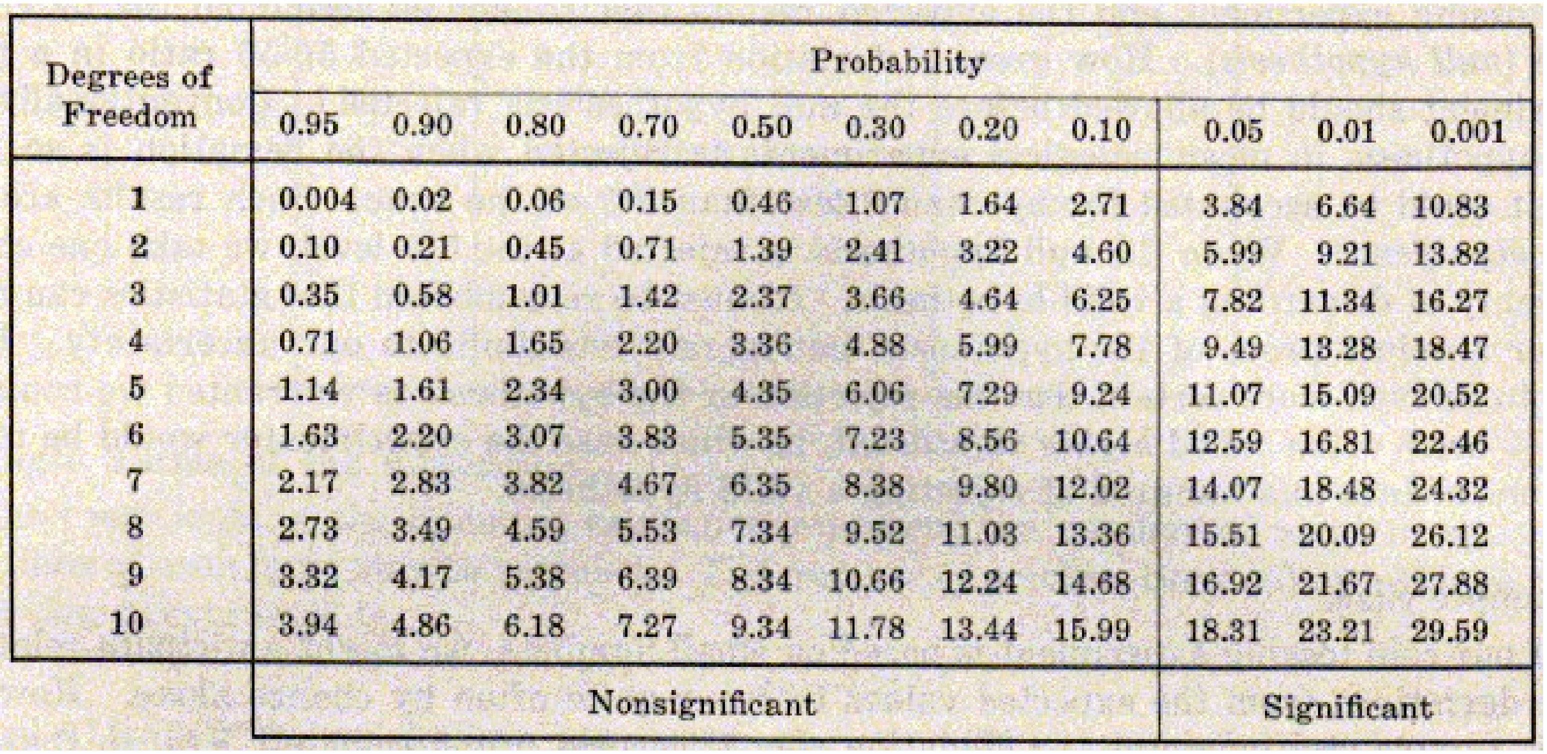


Then calculate expected haplotype counts

Statistical test for LD $\chi^2 = \sum (O\text{-}E)^2/E$

Haplotype	Observed	Expected	$(O - E)^2/E$
AB	76	56	7.14
Ab	64	84	4.76
aB	4	24	16.67
ab	56	36	11.11
Sum	200	200	$\boxed{39.68}$

χ^2 table



Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Statistical test for LD

- How many degrees of freedom?
- We begin with 3 (number of rows - 1)
- 2 are lost due to need to estimate allele frequencies for 2 loci
- This leaves 1 df
- Look up the value 39.68 in the table
- It's significant at more than $p < 0.001$
- We are very confident that this is not linkage equilibrium

Caveats on the χ^2 test

- Not appropriate if there were less than 5 observations expected in the smallest category
- For loci with many alleles, this often requires lumping the rare alleles together
- Test MUST be done on counts of haplotypes, not on frequencies!
- (A frequency difference of 10% is a lot more impressive in a sample of 10,000 than in a sample of 10)

χ^2 test of association

- The previous test asks "Are these two loci in significant LD?"
- A similar test can ask "Is this marker locus correlated with disease?"
- This test is used in a case/control study

Correlation and causation

- A significant test means that marker genotype and disease status are correlated

- This might mean:
  - The marker locus contributes to the disease
  - A different locus linked to the marker locus contributes to the disease
  - Both marker locus and disease locus are tracking a third factor, such as ethnicity
  - The marker locus, or a linked locus, protects from the disease
  - The marker locus affects a person's chance of being diagnosed
  - The marker locus affects a person's chance of being recruited into our study

- We would like to find the causal locus/loci, but may initially only have a correlated locus

Multiple comparisons

- The above test is considered correct for a single marker locus
- If you use a $p < 0.05$ significance cutoff you will have a 5% chance of a false positive
- But what if you go fishing and try 100 well separated marker loci?
- You expect to have 5 false positive results by chance

Bonferroni correction

- If you are making 100 tests, you need a more rigorous significance level to keep the overall chance of a mistake at 5%
- Bonferroni correction divides the target significance level by the number of tests
- Thus, if you do 100 tests you must require $p < 0.0005$ to claim significance at the 5% level

Bonferroni correction

- This seems to make whole-genome scans unfeasible!
- The more loci you scan, the more patients you need to obtain a significant results

- Why so cautious?
- The literature contains large number of unrepeatable association results
- Many researchers want to see two independent reports of association at the same location before they regard linkage as likely

Summary

- Association between a disease phenotype and a marker locus can help locate disease loci
- A χ^2 test is used to detect association
- If multiple independent markers used, a Bonferroni correction is appropriate (though frustrating)