E-M for Haplotyping Revisited

Biostatistics 666

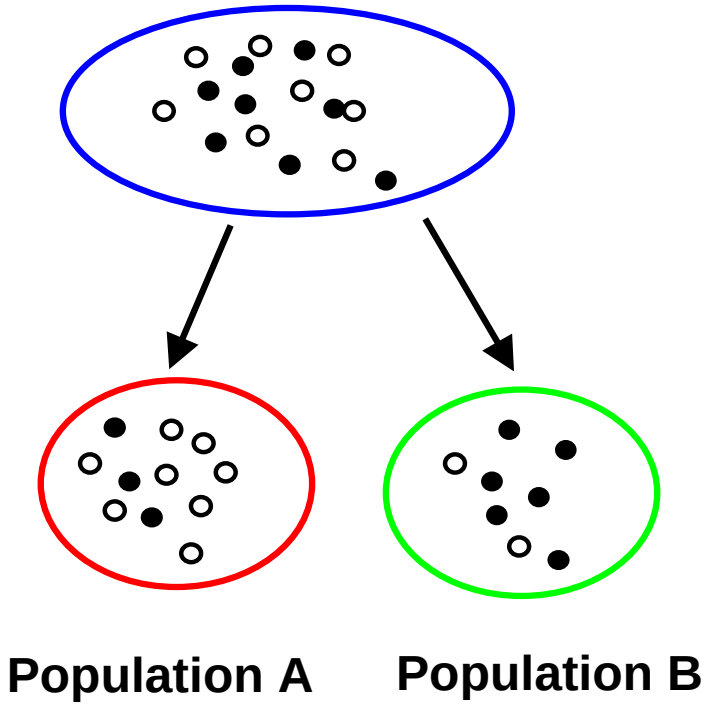
Lecture 11

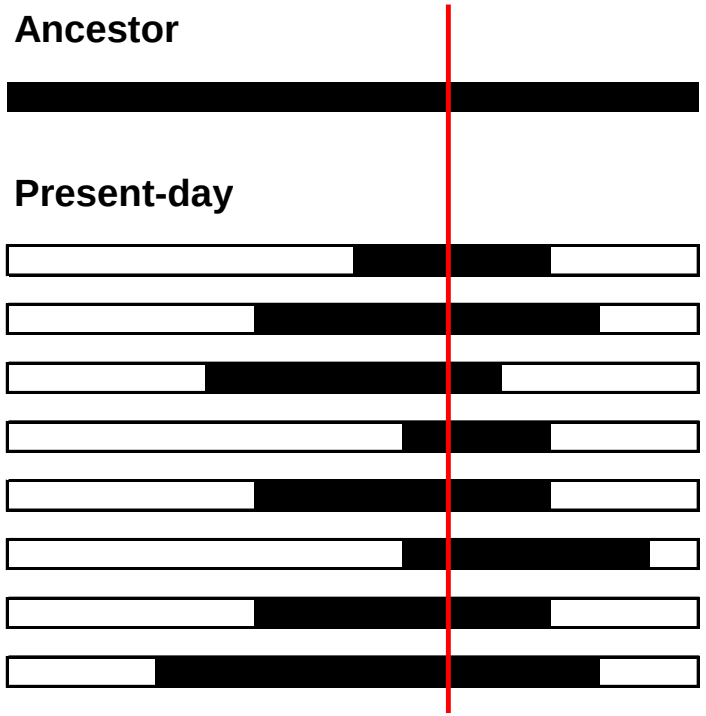
Example

- We'll estimate haplotype frequencies for the example below …

	BB	Bb	bb
AA	25	30	9
Aa	20	12	0
aa	4	0	0

Stratification *vs* Disequilibrium





Stratification

- Due to non-random mating
  - Eg. Mating based on proximity or culture
- Allele frequencies drift apart in each group
  - Eg. Allele frequency differences at many genes between African-Americans and Caucasians
- Disease prevalences may also differ
  - Eg. Glaucoma has prevalence of ~2% in elderly Caucasians, but ~8% in African-Americans

For example...

- Allele frequency
  - $p = 0.2$ $p' = 0.4$
- Background prevalence
  - $k = 0.01$ $k' = 0.02$
- Relative Risk
  - $r = r' = 1$
- Penetrances
  - $f_{11} = f_{12} = f_{22} = k$
  - $f'_{11} = f'_{12} = f'_{22} = k'$

Implies...

Population 1 Sample...

	1/1	1/2	2/2
Affecteds	4	32	64
Unaffecteds	4	32	64

Population 2 Sample...

	1/1	1/2	2/2
Affecteds	16	48	36
Unaffecteds	16	48	36

Sample from Mixed Population (50/50)

	1/1	1/2	2/2
Affecteds	12.0	42.7	45.3
Unaffecteds	10.0	40.0	50.1

The Problem …

Population 1

	Allele1	Allele2	Total
Normal	20	80	100
Affected	20	80	100
Total	40	160	

Chisq = 0

Population 2

	Allele1	Allele2	Total
Normal	120	30	150
Affected	40	10	50
Total	160	40	

ChiSq = 0

	Allele1	Allele2	Total
Normal	140	110	350
Affected	60	90	150
Total	200	200	

ChiSq = 19.5

p-value = ~10^{-5}

The Stratification Problem

- If phenotypes differ between populations
- And allele frequencies have drifted apart
- Unlinked markers exhibit association
- Not very useful for gene mapping!

Possible solutions

- Collect a better matched sample
- Identify population groupings
  - Using self reported ethnicity or genetic markers
  - Carry out association analysis within each group
- Account for inflated false-positive rate
  - Devlin and Roeder (1999)
- Use family based controls

Genomic Control

- Test null markers throughout the genome
  - Markers that are unlikely to be associated
    - Markers that are outside genes
    - Markers in genes that are unlikely to be involved
- ~50 markers is a reasonable number

Define Inflation Factor

- Compute chi-squared for each marker
- Inflation factor λ
  - Average Observed Chi-Squared
  - Median Observed Chi-Squared / 0.456
  - Should be >= 1
- Adjust statistic at candidate markers
  - Replace χ^2_{biased} with $\chi^2_{fair} = \chi^2_{biased}/\lambda$