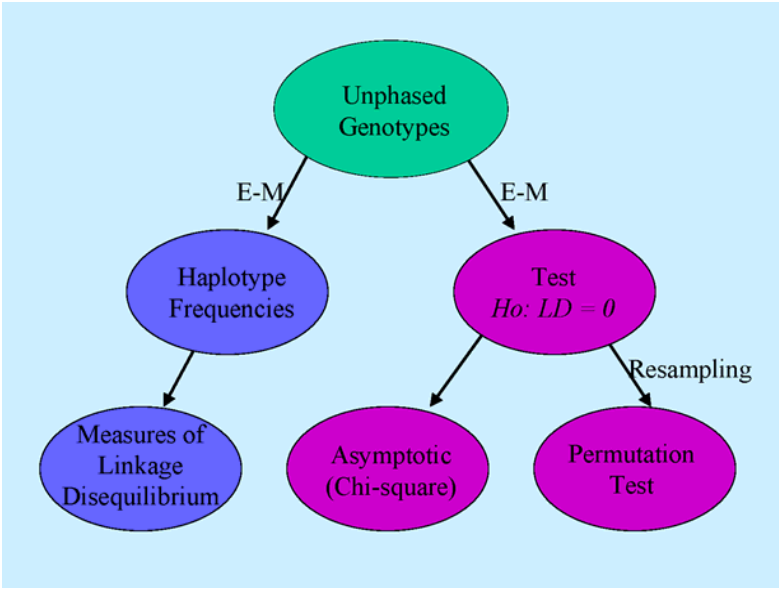
Haplotype Frequency Estimation



2 Locus 2 Allele Haplotypes

9 Phase-Unknown Categories ("Y")		10 Phase-Known Categories ("X")
AABB	⇒	AB/AB
AABb	⇒	AB/Ab
AAbb	⇒	Ab/Ab
AaBB	⇒	AB/aB
Aabb	⇒	AB/ab
aaBB	⇒	aB/aB
aaBb	⇒	aB/ab
aabb	⇒	ab/ab
AaBb	⇒	AB/ab
		Ab/aB

$$n_{AaBb} = n_{AB/ab} + n_{Ab/aB}$$

Haplotype Frequency Estimation

~~phenotype~~ an unphased genotype e.g., $AaBB$

haplotype sequence of alleles on 1 chromosomal strand
e.g., $AaBB \Rightarrow AB/aB$

P_A frequency of allele A

P_{Aa} frequency of genotype Aa

P_{aB} frequency of haplotype aB

Hardy-Weinberg Disequilibrium

$$D_A = P_{Aa} - P_A P_a$$

Linkage (gametic) Disequilibrium

$$D_{aB} = P_{aB} - P_a P_B$$

The Expectation Maximization Algorithm

1. Start with random haplotype frequencies (HFs).
2. Compute expected genotype frequencies given these HFs.
3. Use relative genotype frequencies to estimate new HFs by gene counting.
4. Check for improvement using the log-likelihood function.
5. Return to #2 with the new HFs.

The EM Algorithm for 2 Locus 2 Allele Haplotypes

E-step:

$$n_{AB/ab}^{(k)} = E[n_{AB/ab}|Y,p] = n_{AaBb}\frac{2p_{AB}^{(k)}p_{ab}^{(k)}}{2p_{AB}^{(k)}p_{ab}^{(k)}+2p_{Ab}^{(k)}p_{aB}^{(k)}}$$

$$n_{Ab/aB}^{(k)} = E[n_{Ab/aB}|Y,p] = n_{AaBb}\frac{2p_{Ab}^{(k)}p_{aB}^{(k)}}{2p_{AB}^{(k)}p_{ab}^{(k)}+2p_{Ab}^{(k)}p_{aB}^{(k)}}$$

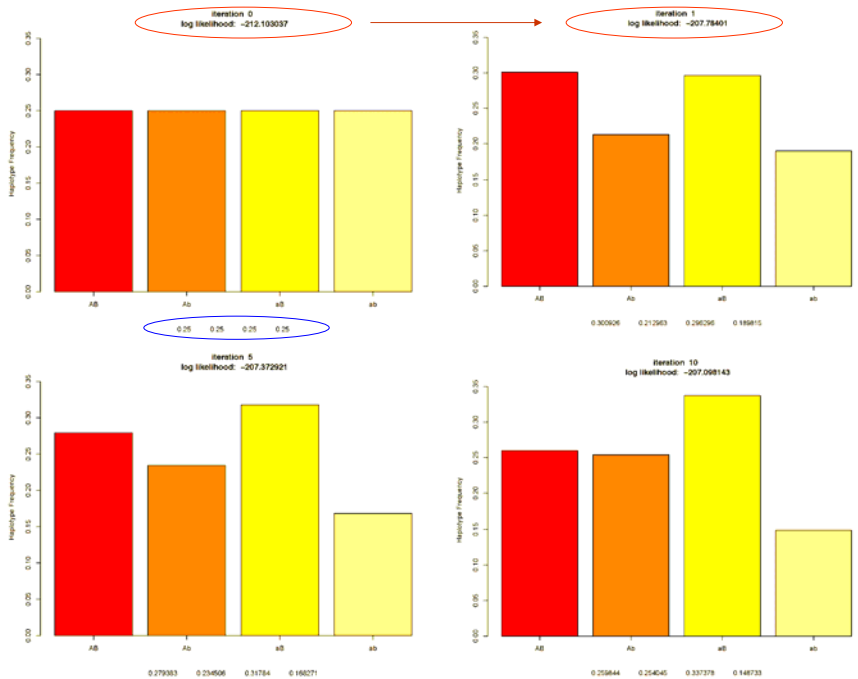
M-step:

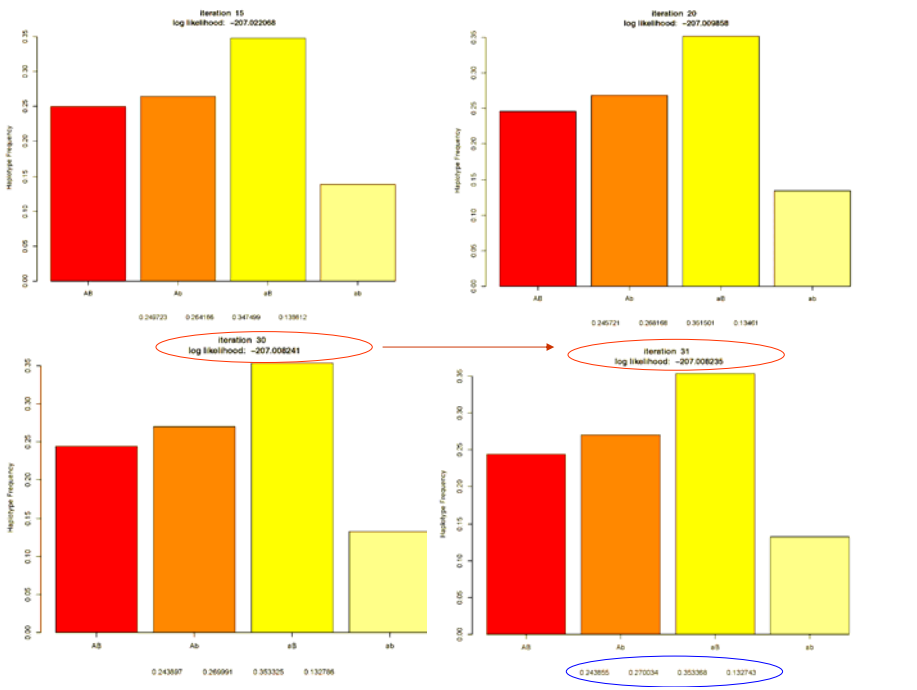
$$p_{AB}^{(k+1)} = \frac{2n_{AABB}+n_{AABb}+n_{AaBB}+n_{AB/ab}^{(k)}}{2n}$$
$$p_{Ab}^{(k+1)} = \frac{2n_{AAbb}+n_{AABb}+n_{Aabb}+n_{Ab/aB}^{(k)}}{2n}$$
$$p_{aB}^{(k+1)} = \frac{2n_{aaBB}+n_{aaBb}+n_{AaBb}+n_{Ab/aB}^{(k)}}{2n}$$
$$p_{ab}^{(k+1)} = \frac{2n_{aabb}+n_{aaBb}+n_{Aabb}+n_{AB/ab}^{(k)}}{2n}$$

An Example:

genotype data for 108 individuals

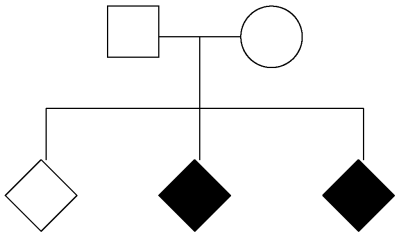
	BB	*Bb*	*bb*
AA	**10**	**10**	**4**
Aa	**10**	**50**	**3**
aa	**13**	**3**	**5**



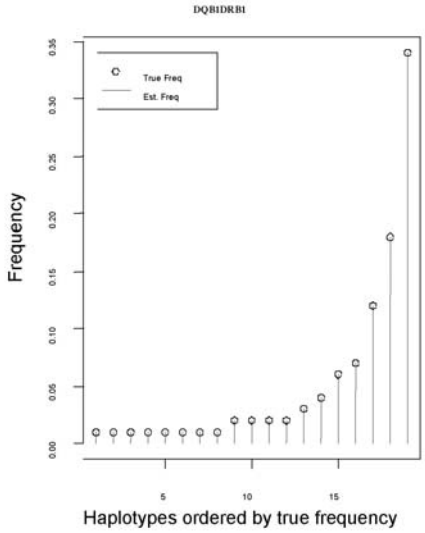


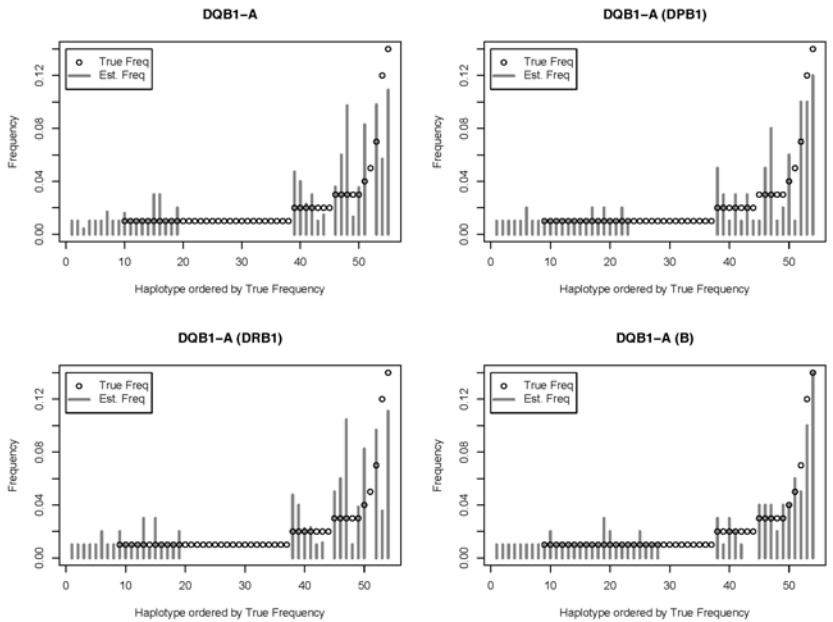
Human Biological Data Interchange (HBDI)

- A repository of cell lines for the study of type I diabetes
- Nuclear families with unaffected parents
- At least 2 siblings affected with type I diabetes
- Loci typed: DPB1, DQB1, DQA1, DRB1, B, A



DRB1-DQB1 haplotype estimation





Single et al., 2002, Genetic Epidemiology

EM Caveats & Recommendations

- EM iterations may converge on a local maximum
  - * Multiple starting conditions should be used
- ← Not just equal HFs as in example
- Low frequency HF estimates may not be reliable.
- EM iterations assume Hardy-Weinberg Proportions
  - * Deviations from HWP may decrease estimation accuracy
- Collapsing 3-locus HF estimates over a locus can improve estimation of 2-locus HFs.
  - * The "best" locus to collapse over depends on the degree of LD and HWD.