Computational Haplotype Phasing of Unrelated Individuals

October 3, 2013

Review

- Algorithms we've covered
  - Clark Phasing
  - Expectation Maximization (EM) Phasing
- Topics we've covered
  - Graph theory
  - Population genetics models (e.g. LD)
  - Maximum Likelihood (ML) Estimation
  - EM for ML

What’s next

- Coalescent methods
  - Modeling populations as descendants from a most recent common ancestor
- Hidden Markov Models
- Identity by descent
  - Everyone is related to some degree
  - Relationships that are distant but still have significant haplotype overlap are called ‘cryptically related’
- Algorithms for tagging SNPs and associations

Known haplotype frequencies

- If we know haplotype frequencies in the popluation we can phase using statistical methods
- E.g. genotype *g* has 3 explanations *h1h2, h3h4, h5h6*
  - We know haplotype frequencies, we multiple the frequencies for each explanation and take the maximum value

Coalescent model

- In general, the frequencies are unknown
- Approximate computation of haplotype frequencies in the population
- Approximation coalescent models
  - Intuition: New haplotypes are created from haplotype existing in the population from mutation and recombination

Coalescent phasing

- Phase using haplotypes that were used before in other phasings
- Algorithms of this type are based on the approximate coalescent
  - Modeled well by hidden Markov Models
  - We will estimate the parameters with EM

Algorithms

- PHASE
  - Can handle about 100 SNPs by 100 people
  - The gold standard for phasing small regions
  - Employs markov chain monte carlo
- fastPHASE
  - Can handle larger data than PHASE, uses HMM
- BEAGLE
  - Faster than fastPHASE and more accurate on large (>1000 people) data but less accurate when you have <100 individuals
- SHAPE-IT
  - More efficient

Algorithm continued

- MACH
- IMPUTE2
  - Based on seminal work on recombination due to Li-Stephens model
  - Use HMMs

Identity by descent

- Two regions are identical by descent
- Suppose we know
  - Haplotype frequencies
  - Where IBD regions are

Example

SNP	Unphased Genotypes	Shared Haplotypes	IBD-phased genotypes	Phasing Individual 1	Phasing Individual 2
1	{A/C} {A/C}	?	? ? ? ?	A C C A	A C C A
2	{C/T} {C/C}	C	C T C C	C T C T	C C C C
3	{T/T} {T/G}	T	T T T G	T T T T	T G T G
4	{G/G} {A/G}	G	G G G A	G G G G	G A G A
5	{C/C} {C/C}	C	C C C C	C C C C	C C C C

Individual 1 in black, 2 in ref

If haplotype frequencies for all haplotypes in the phasings of individuals 1 and 2 are known, multiplying the haplotypes of the explanations will give you the probability of the phasing.