Haplotype Reconstruction

Lectures 6 – Oct 12, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

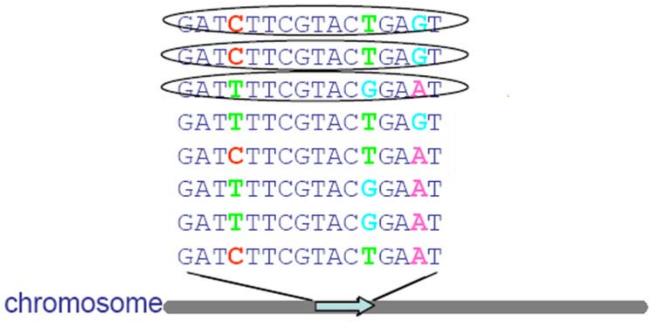
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Course Announcements

- Project proposal
  - Due this Friday
  - 1 paragraph describing what you'd like to work on for the class project.
- Special office hours
  - Today 3-5pm: discussing project topics

Haplotype

- A combination of alleles present in a chromosome
- Each haplotype has a *frequency*, which is the proportion of chromosomes of that type in the population



- Consider N binary SNPs in a genomic region
- There are 2^N possible haplotypes
  - But in fact, far fewer are seen in human population

More on haplotype

- What determines haplotype frequencies?
  - Recombination rate (r) between neighboring alleles
  - Depends on the population
  - r is different for different regions in genome
- Linkage disequilibrium (LD)
  - Non-random association of alleles at two or more loci, not necessarily on the same chromosome.
- Why do we care about haplotypes or LD?

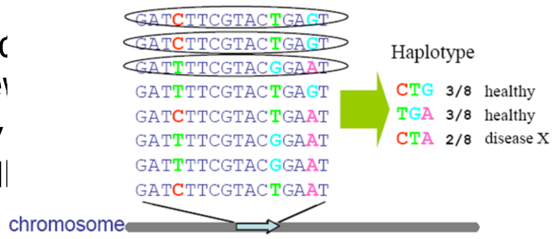
Useful roles for haplotypes

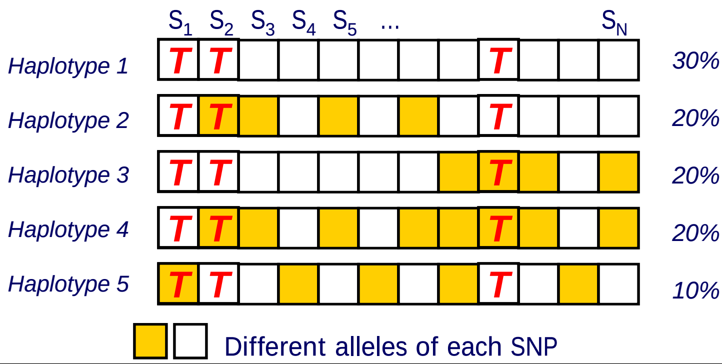
- Linkage disequilibrium studies
  - Summarize genetic variation
  - Learn about population history
- Selecting markers to genotype
  - Identify haplotype tag SNPs

5

Exploiting LD – tag SNPs

- In a typic here are only a fe
- Carefully tus of other SNI





6

Association studies and LD

- Why is LD important for gene mapping (eg QTL mapping)?
- If all polymorphisms were independent at the population level, association studies would have to examine every one of them...
- Linkage disequilibrium makes tightly linked variants strongly correlated producing cost savings for association studies

Useful roles for haplotypes

- Linkage disequilibrium studies
  - Summarize genetic variation
  - Learn about population history
- Selecting markers to genotype
  - Identify haplotype tag SNPs
- Candidate gene association studies
  - Help interpret single marker associations
  - Map capture effect of ungenotyped alleles

The problems...

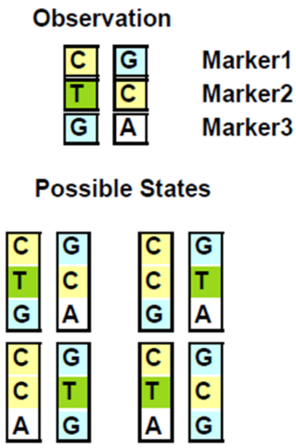
- Haplotypes are hard to measure directly
  - X-chromosome in males
  - Sperm typing
  - Hybrid cell lines
  - Other molecular techniques

- Often, statistical reconstruction required

Typical genotype data

- Two alleles for each individual
  - Chromosome origin for each allele is unknown

- Multiple haplotype pairs can fit observed genotype



Use information on relatives?

- Family information can help determine phase at many markers
- Still, many ambiguities might not be resolved
  - Problem more serious with larger numbers of markers
- Can you propose examples?

11

Example – inferring haplotypes

- Genotype: AT//AA//CG
  - Maternal genotype: TA//AA//CC
  - Paternal genotype: TT//AA//CG
  - Then the haplotype is AAC/TAG
- Genotype: AT//AA//CG
  - Maternal genotype: AT//AA//CG
  - Paternal genotype: AT//AA//CG
  - Cannot determine unique haplotype
- Problem
  - Determine Haplotypes without parental genotypes

12

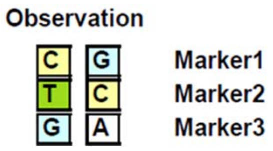
What if there are no relatives?

- Rely on linkage disequilibrium
- Assume that population consists of small number of distinct haplotypes

13

Haplotype reconstruction

- Also called, *phasing, haplotype inference* or *haplotyping*



- Data
- Genotypes on N markers from M individuals
- Goals
- Frequency estimation of all possible haplotypes
- Haplotype reconstruction for individuals
- How many out of all possible haplotypes are plausible in a population?

14

Clark's Haplotyping Algorithm

- Clark (1990) *Mol Biol Evol* **7**:111-122

- One of the first haplotyping algorithms
  - Computationally efficient
  - Very fast and widely used in 1990's
  - More accurate methods are now available

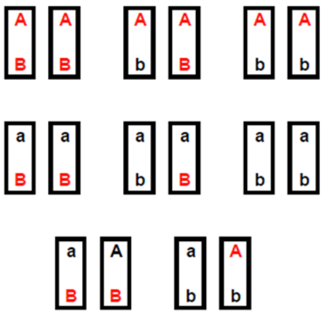
Clark's Haplotyping Algorithm

- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a list of known haplotypes

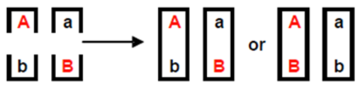
  - Unambiguous individuals
    - Homozygous at every locus (e.g. TT//AA//CC)
      Haplotypes: TAC
    - Heterozygous at just one locus (e.g. TT//AA//CG)
      Haplotypes: TAC or TAG

Unambiguous vs ambiguous

- Haplotypes for 2 SNPs (alleles: A/a, B/b)



Unambiguous Genotypes
Underlying Haplotype is Known



Ambigous Genotype
Multiple Underlying Genotypes Possible

Clark's Haplotyping Algorithm

- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a *list of known haplotypes*
- Resolve ambiguous individuals
  - If possible, use two haplotypes from list
  - Otherwise, use one known haplotype and augment list
- If unphased individuals remain
  - Assign phase randomly to one individual
  - Augment haplotype list and continue from previous step

18

Parsimonious Phasing - Example

- Notation (more compact representation)
  - 0/1: homozygous at each locus (00,11)
  - h: heterozygous at each locus (01)

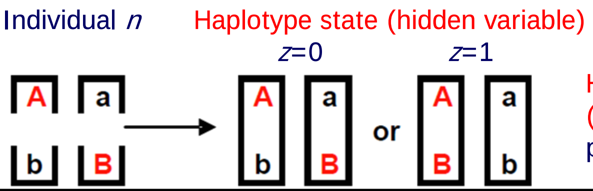
1 0 1 0 0 h	1 0 1 0 0 0 1 0 1 0 0 1
h 0 1 h 0 0	1 0 1 0 0 0 0 0 1 1 0 0
0 h h 1 h 0	0 0 1 1 0 0 0 1 0 1 1 0

Notes ...

- Clark's Algorithm is extremely fast
- Problems
  - No homozygotes or single SNP heterozygotes in the sample
  - Many unresolved haplotypes at the end
  - Error in haplotype inference if a crossover of two actual haplotypes is identical to another true haplotype
  - Frequency of these problems depend on average heterozygosity of the SNPs, no of loci, recombination rate, sample size

The EM Haplotyping Algorithm

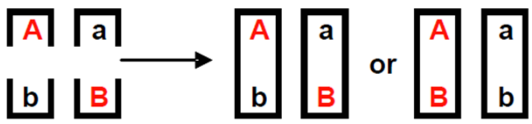
- Excoffier and Slatkin (1995) *Mol Biol Evol* **12:**921-927
- Why EM for haplotyping?
  - EM is a method for MLE with hidden variables.
- What are the hidden variables, parameters?
  - **Hidden variables:** haplotype state of each individual
  - **Parameters:** haplotype frequencies



Haplotype frequencies (parameters)

p_{Ab}, p_{aB}, p_{AB}, p_{ab}

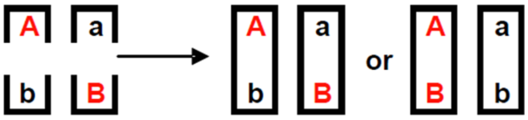
Assume that we know haplotype frequencies



For example, if
P_{AB} = 0.3
P_{ab} = 0.3
P_{Ab} = 0.3
P_{aB} = 0.1

- Probability of first outcome:
  - $2P_{Ab}P_{aB}$ =
- Probability of second outcome:
  - $2P_{AB}P_{ab}$ =

Conditional probabilities are ...



For example, if
P_{AB} = 0.3
P_{ab} = 0.3
P_{Ab} = 0.3
P_{aB} = 0.1

- Conditional probability of first outcome:
  - $2P_{Ab}P_{aB} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) =$
- Conditional probability of second outcome:
  - $2P_{AB}P_{ab} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) =$

Assume that we know the haplotype state of each individual

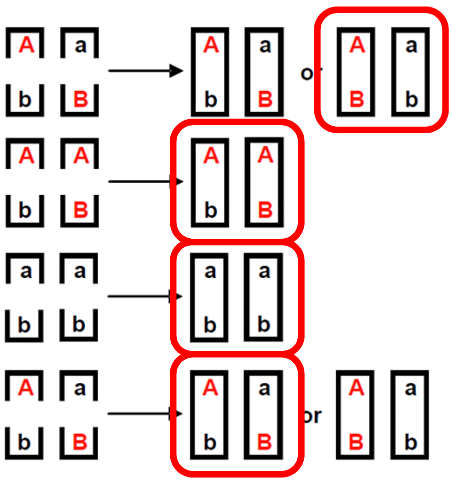
- Computing haplotype frequencies is straightforward

Individual 1

Individual 2

Individual 3

Individual 4

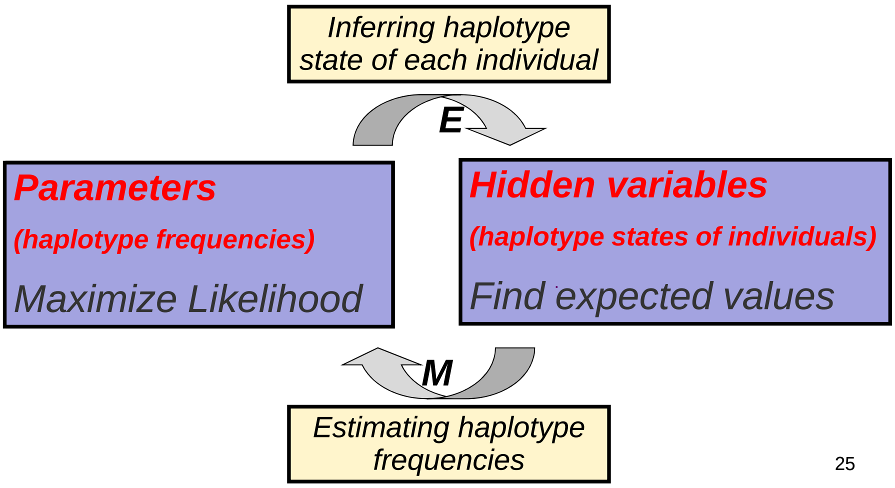


p_{AB} =?
P_{ab} =?
p_{Ab} =?
p_{aB} =?

24

Phasing by EM

- EM: Method for maximum-likelihood parameter inference with hidden variables



EM Algorithm For Haplotyping

- 1. "Guesstimate" haplotype frequencies
- 2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
- 3. Estimate frequency of each haplotype by counting
- 4. Repeat steps 2 and 3 until frequencies are stable

Phasing by EM

Data:

Genotype	Haplotype pair	Weight
1 0 h h 1	*1 0 0 0 1*	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	*0 0 0 1 0*	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	*1 0 0 1 1*	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

27

Phasing by EM

Data:

Genotype	Haplotype pair	Weight
1 0 h h 1	*1 0 0 0 1*	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	*0 0 0 1 0*	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	*1 0 0 1 1*	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

Frequencies	
0 0 0 1 0	*1/12*
0 0 0 1 1	*1/12*
1 0 0 0 1	*1/12*
1 0 0 1 0	*1/12*
1 0 0 1 1	*3/12*
1 0 1 0 1	*1/12*
1 0 1 1 1	*2/12*
1 1 0 1 1	*1/12*
1 1 1 1 1	*1/12*

28

Phasing by EM

Data:

Data	Haplotype pairs	Weight	Haplotypes
1 0 h h 1	1 0 0 0 1	¼	0.4
	1 0 1 1 1	¼	
	1 0 0 1 1	¼	0.6
	1 0 1 0 1	¼	
h 0 0 1 h	0 0 0 1 0	¼	0.75
	1 0 0 1 1	¼	
	0 0 0 1 1	¼	0.25
	1 0 0 1 0	¼	
1 h h 1 1	1 0 0 1 1	¼	0.6
	1 1 1 1 1	¼	
	1 0 1 1 1	¼	0.4
	1 1 0 1 1	¼	

Frequencies

0 0 0 1 0	1/12
0 0 0 1 1	1/12
1 0 0 0 1	1/12
1 0 0 1 0	1/12
1 0 0 1 1	3/12
1 0 1 0 1	1/12
1 0 1 1 1	2/12
1 1 0 1 1	1/12
1 1 1 1 1	1/12

Expectation

29

Phasing by EM

Maximization

Data:

Data	Haplotype pairs	Weight	Haplotypes
1 0 h h 1	1 0 0 0 1	¼	0.4
	1 0 1 1 1	¼	
	1 0 0 1 1	¼	0.6
	1 0 1 0 1	¼	
h 0 0 1 h	0 0 0 1 0	¼	0.75
	1 0 0 1 1	¼	
	0 0 0 1 1	¼	0.25
	1 0 0 1 0	¼	
1 h h 1 1	1 0 0 1 1	¼	0.6
	1 1 1 1 1	¼	
	1 0 1 1 1	¼	0.4
	1 1 0 1 1	¼	

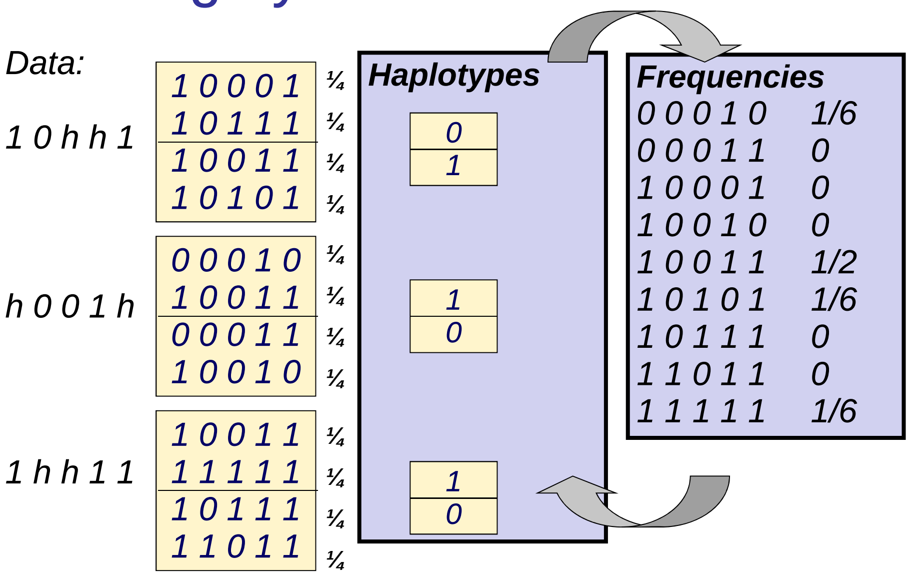
Frequencies

0 0 0 1 0	.125
0 0 0 1 1	.042
1 0 0 0 1	.067
1 0 0 1 0	.042
1 0 0 1 1	.325
1 0 1 0 1	.1
1 0 1 1 1	.067
1 1 0 1 1	.067
1 1 1 1 1	.1

Expectation

30

Phasing by EM



31

Computational Cost (for SNPs)

- Consider sets of *m* unphased genotypes
  - Markers 1..*m*

For example, if m=10

- If markers are bi-allelic
  - 2^m possible haplotypes = 1024
  - $2^{m-1} (2^m + 1)$ possible haplotype pairs = 524,800
  - 3^m distinct observed genotypes = 59,049
  - 2^{n-1} reconstructions for *n* heterozygous loci = 512
- For example, if m = 10

32

EM Algorithm For Haplotyping

- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
  - Fewer microsatellites
- More accurate than Clark's method
- Fully or partially phased individuals contribute most of the information

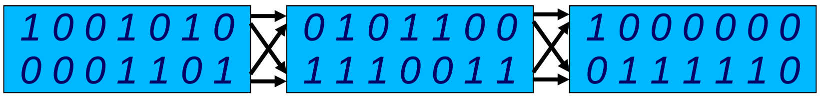
Enhancements to EM

- List only haplotypes present in sample
- Gradually expand subset of markers under consideration, eliminating haplotypes with low estimated frequency from consideration at each stage
  - SNPHAP, Clayton (2001)
  - HAPLOTYPER, Qin et al (2002)

Divide-And-Conquer Approximation

- Number of potential haplotypes increases exponentially
  - Number of observed haplotypes does not

- Approximation
  - Successively divide marker set
  - Locally phase each segment through EM
  - Prune haplotype list as segments are ligated
  - Merge by phasing vectors of haplotype pairs



- Computation order: ~ m log m
  - Exact EM is order ~ 2^m