Statistical Genetics – Part I

Lecture 3: Haplotype reconstruction

Su-In Lee, CSE & GS, UW
suinlee@uw.edu

1

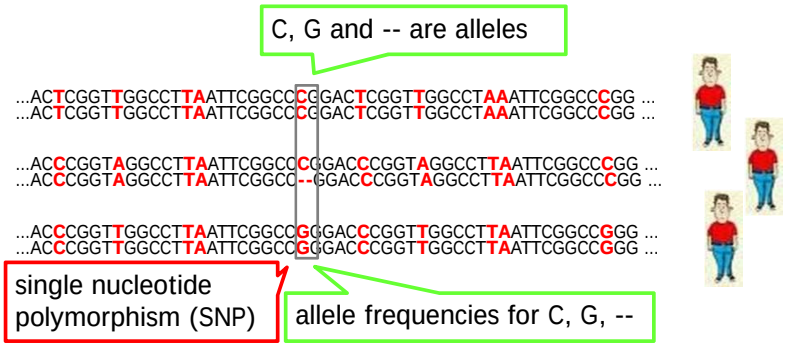
Outline

- Basic concepts
  - Allele, allele frequencies, genotype frequencies
  - Haplotype, haplotype frequency
  - Recombination rate
  - Linkage disequilibrium
- Haplotype reconstruction
  - Parsimony-based approach
  - EM-based approach
- Next topic
  - Disease association studies

2

Alleles

- Alternative forms of a particular sequence
- Each allele has a frequency, which is the proportion of chromosomes of that type in the population



3

Allele Frequency Notations

- For two alleles
  - Usually labeled p and $q = 1 - p$
  - e.g. p = frequency of C, q = frequency of G
- For more than 2 alleles
  - Usually labeled p_A, p_B, p_C ...
  - … subscripts A, B and C indicate allele names

4

Genotype

- The pair of alleles carried by an individual
  - If there are **n** alternative alleles …
  - … there will be **n(n+1)/2** possible genotypes
  - In most cases, there are 3 possible genotypes
- **Homozygotes**
  - The two alleles are in the same state
  - (e.g. CC, GG, AA)
- **Heterozygotes**
  - The two alleles are different
  - (e.g. CG, AC)

5

Genotype Frequencies

- Since alleles occur in pairs, these are a useful descriptor of genetic data.
- However, in any non-trivial study we might have a lot of frequencies to estimate.
- p_{AA}, p_{AB}, p_{AC},… p_{BB}, p_{BC},… p_{CC} …

6

The Simple Part

- Genotype frequencies lead to allele frequencies.
- For example, for two alleles:
  - $p_1 = p_{11} + ½\ p_{12}$
  - $p_2 = p_{22} + ½\ p_{12}$
- However, the reverse is also possible!
  - We just need an additional assumtion

7

Hardy-Weinberg Equilibrium (HWE)

- Relationship described in 1908
  - Hardy, British mathematician
  - Weinberg, German physician
- Shows **n** allele frequencies determine **n(n+1)/2** genotype frequencies
  - Large populations
- Random union of the two gametes produced by two individuals

8

Random Mating: Mating Type Frequencies

- Denoting the genotype frequency of A_iA_j by p_{ij}, and the allele frequency A_i by p_i (i, j ∈ {1,2}),
  - $p_1 = p_{11} + ½\ p_{12}$; $p_2 = p_{22} + ½\ p_{12}$

Mating	Frequency
$A_1A_1*A_1A_1$	p_{11}^2
$A_1A_1*A_1A_2$	$2p_{11}p_{12}$
$A_1A_1*A_2A_2$	$2p_{11}p_{22}$
$A_1A_2*A_1A_2$	p_{12}^2
$A_1A_2*A_2A_2$	$2p_{12}p_{22}$
$A_2A_2*A_2A_2$	p_{22}^2
Total	1.0

9

Mendelian Segregation: Offspring Genotype Frequencies

		Offspring		
Mating	Frequency	A_1A_1	A_1A_2	A_2A_2
$A_1A_1*A_1A_1$	p_{11}^2	1	0	0
$A_1A_1*A_1A_2$	$2p_{11}p_{12}$	0.5	0.5	0
$A_1A_1*A_2A_2$	$2p_{11}p_{22}$	0	1	0
$A_1A_2*A_1A_2$	p_{12}^2	0.25	0.5	0.25
$A_1A_2*A_2A_2$	$2p_{12}p_{22}$	0	0.5	0.5
$A_2A_2*A_2A_2$	p_{22}^2	0	0	1

10

Mendelian Segregation: Offspring Genotype Frequencies

Mating	Frequency	Offspring A_1A_1		A_1A_2	A_2A_2
$A_1A_1 * A_1A_1$	p_{11}^2	1	x p_{11}^2	0	0
$A_1A_1 * A_1A_2$	$2p_{11}p_{12}$	0.5	x $2p_{11}p_{12}$	0.5	0
$A_1A_1 * A_2A_2$	$2p_{11}p_{22}$	0	x $2p_{11}p_{22}$	1	0
$A_1A_2 * A_1A_2$	p_{12}^2	0.25	x p_{12}^2	0.5	0.25
$A_1A_2 * A_2A_2$	$2p_{12}p_{22}$	0	x $2p_{12}p_{22}$	0.5	0.5
$A_2A_2 * A_2A_2$	p_{22}^2	0	x p_{22}^2	0	1

$= p_{11}^2 + 2p_{11}(0.5\ p_{12}) + (0.5\ p_{12})^2$
$= (p_{11} + 0.5\ p_{12})^2$
$= p_1^2$

11

Mendelian Segregation: Offspring Genotype Frequencies

Mating	Frequency	Offspring A_1A_1	A_1A_2		A_2A_2
$A_1A_1 * A_1A_1$	p_{11}^2	1	0	x p_{11}^2	0
$A_1A_1 * A_1A_2$	$2p_{11}p_{12}$	0.5	0.5	x $2p_{11}p_{12}$	0
$A_1A_1 * A_2A_2$	$2p_{11}p_{22}$	0	1	x $2p_{11}p_{22}$	0
$A_1A_2 * A_1A_2$	p_{12}^2	0.25	0.5	x p_{12}^2	0.25
$A_1A_2 * A_2A_2$	$2p_{12}p_{22}$	0	0.5	x $2p_{12}p_{22}$	0.5
$A_2A_2 * A_2A_2$	p_{22}^2	0	0	x p_{22}^2	1

$= p_1^2$ $= 2p_1p_2$

12

Mendelian Segregation: Offspring Genotype Frequencies

Mating	Frequency	Offspring A_1A_1	Offspring A_1A_2	Offspring A_2A_2	
$A_1A_1 * A_1A_1$	p_{11}^2	1	0	0	x p_{11}^2
$A_1A_1 * A_1A_2$	$2p_{11}p_{12}$	0.5	0.5	0	x $2p_{11}p_{12}$
$A_1A_1 * A_2A_2$	$2p_{11}p_{22}$	0	1	0	x $2p_{11}p_{22}$
$A_1A_2 * A_1A_2$	p_{12}^2	0.25	0.5	0.25	x p_{12}^2
$A_1A_2 * A_2A_2$	$2p_{12}p_{22}$	0	0.5	0.5	x $2p_{12}p_{22}$
$A_2A_2 * A_2A_2$	p_{22}^2	0	0	1	x p_{22}^2
		$= p_1^2$	$= 2p_1p_2$	$= p_2^2$	

Frequency of A_1 in offspring $= p_1^2 + \frac{1}{2} 2p_1p_2 = p_1(p_1 + p_2) = p_1$

13

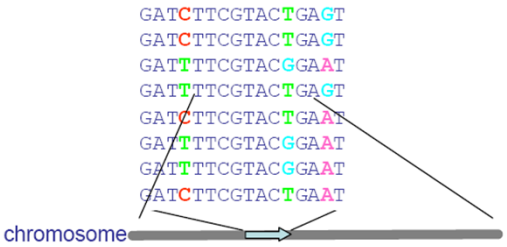
Conclusion: HWE

- Allele frequencies and genotype ratios in a randomly-breeding population *remain constant* from generation to generation.
- Genotype frequencies are function of allele frequencies.
  - Equilibrium reached in one generation
  - Independent of initial genotype frequencies
  - Random mating, etc. required

14

Review: Genetic Variation

- Single nucleotide polymorphism (SNP)
  - Each variant is called an *allele*; each allele has a *frequency*



- Hardy Weinberg equilibrium (HWE)
  - Relationship between allele frequency and genotype frequencies
- How about the relationship between alleles of neighboring SNPs?
  - We need to know about linkage (dis)equilibrium

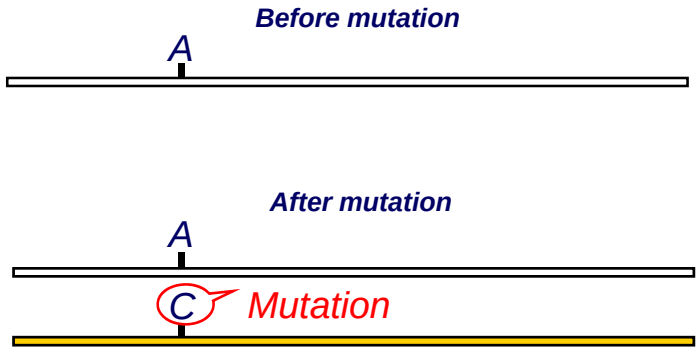
15

Let's consider the history of two neighboring alleles…

16

History of Two Neighboring Alleles

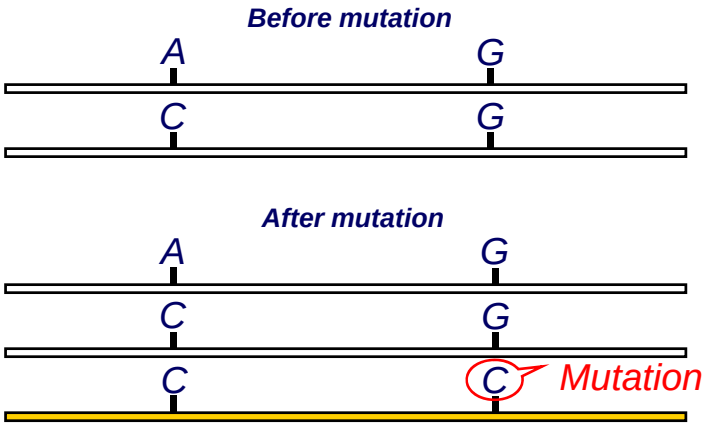
- Alleles that exist today arose through ancient mutation events…



17

History of Two Neighboring Alleles

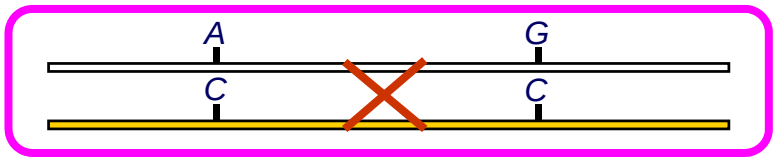
- One allele arose first, and then the other…



Haplotype: combination of alleles present in a chromosome

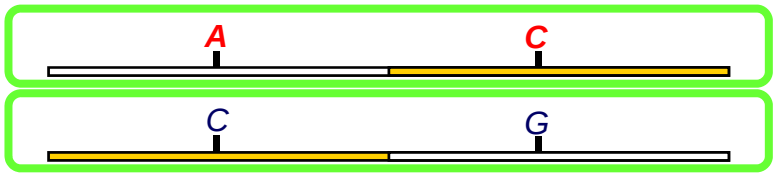
18

Recombination Can Create More Haplotypes

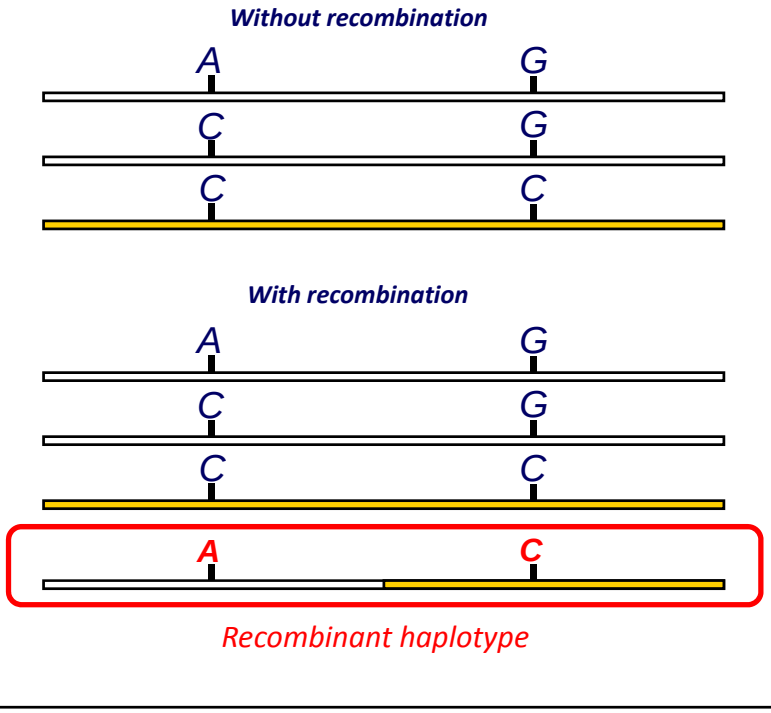


- No recombination (or 2n recombination events)

- Recombination



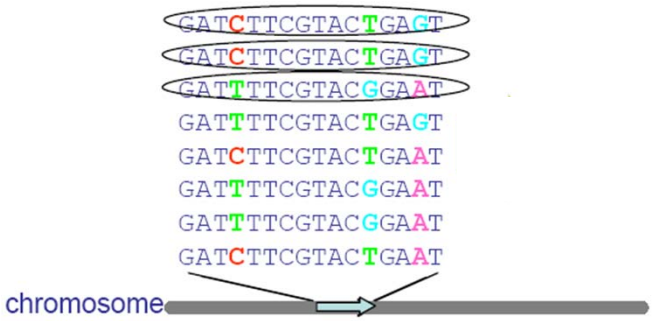
19



20

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

21

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?

22

Useful Roles For Haplotypes

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

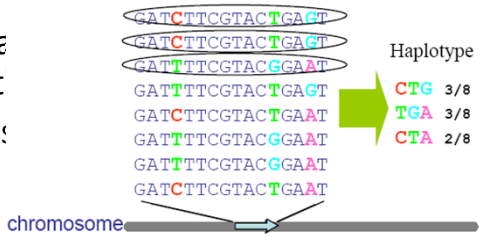
What is genotyping?

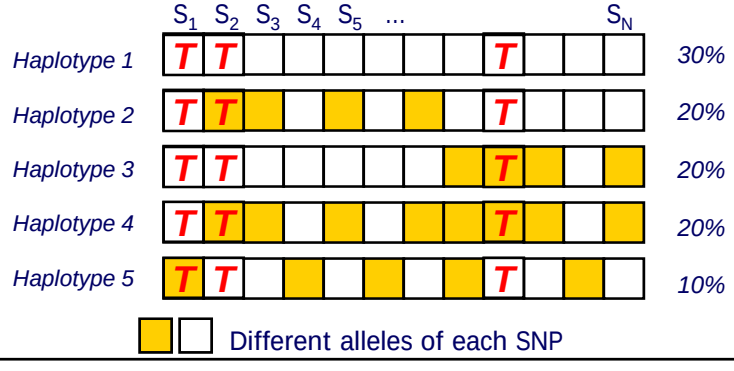
- Genome-wide sequencing is still too expensive
- There are sites that are known to vary across individuals (e.g. SNPs)
- "genotyping" means determining the alleles in each SNP for a certain individual.

23

Exploiting LD – Tag SNPs

- In a typica … are only a few dist
- Carefully … of other SNPs





24

Association Studies and LD

- Why is LD important for disease association studies?
- 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 genotyping in association studies

25

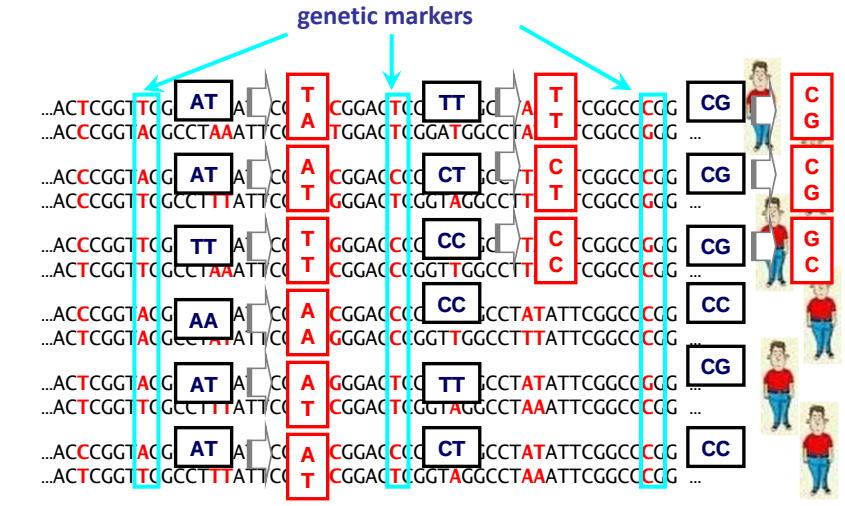
The Problems…

- Haplotypes are hard to measure directly
  - X-chromosome in males
  - Sperm typing
  - Hybrid cell lines
  - Other molecular techniques
- Often, statistical reconstruction required

26

Goal

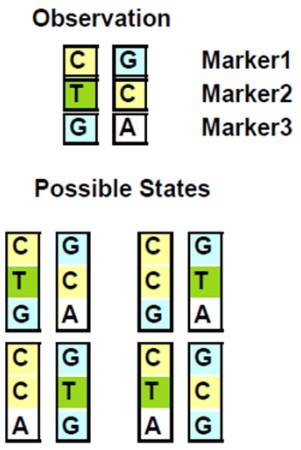
- Haplotype reconstruction



27

Typical Genotype Data

- Two alleles for each individual
  - Chromosome origin for each allele is unknown
- Multiple haplotype pairs can fit observed genotype



28

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?

29

Example – Inferring Haplotypes

- Genotype: AT//AA//CG
  - Maternal genotype: TA//AA//CC → TAC/AAC
  - Paternal genotype: TT//AA//CG → TAC/TAG
  - 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

30

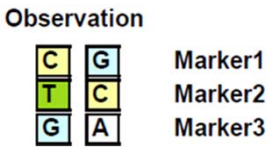
What If There Are No Relatives?

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

31

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?

32

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

33

Clark's Haplotyping Algorithm

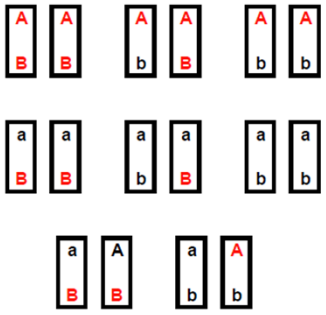
- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a list of known haplotypes

  - 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

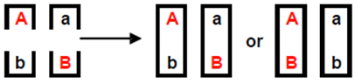
34

Unambiguous vs. Ambiguous

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



Unambigous Genotypes
Underlying Haplotype is Known



Ambigous Genotype
Multiple Underlying Genotypes Possible

35

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

36

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

37

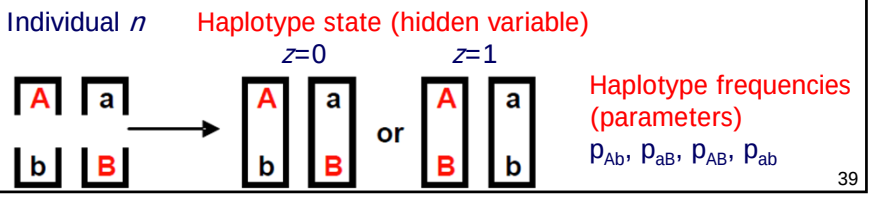
Notes …

- 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

38

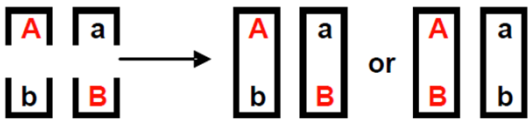
The EM Haplotyping Algorithm

- 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



39

Assume That We Know Haplotype Frequencies

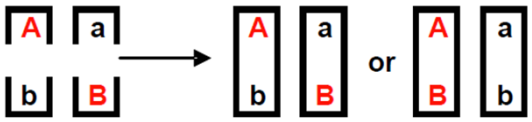


For example, if
$P_{AB} = 0.3$
$P_{ab} = 0.3$
$P_{Ab} = 0.3$
$P_{aB} = 0.1$

- Probability of first outcome:
  - $2P_{Ab}P_{aB} = 0.06$
- Probability of second outcome:
  - $2P_{AB}P_{ab} = 0.18$

40

Conditional Probabilities Are …



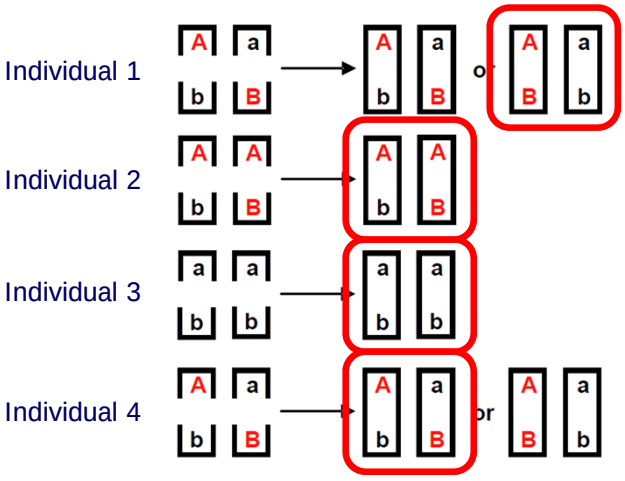
For example, if
$P_{AB} = 0.3$
$P_{ab} = 0.3$
$P_{Ab} = 0.3$
$P_{aB} = 0.1$

- Conditional probability of first outcome:
  - $2P_{Ab}P_{aB} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) = 0.25$
- Conditional probability of second outcome:
  - $2P_{AB}P_{ab} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) = 0.75$

41

Assume That We Know The Haplotype State Of Each Individual

- Computing haplotype frequencies is straightforward

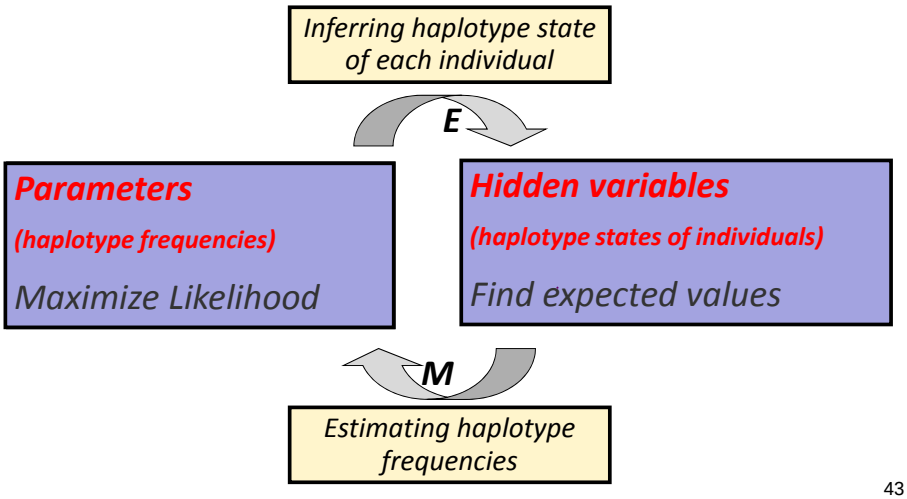


$p_{AB} = ?$
$P_{ab} = ?$
$p_{Ab} = ?$
$p_{aB} = ?$

42

Phasing By EM

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



43

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

44

Phasing by EM

Data:

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

45

Phasing by EM

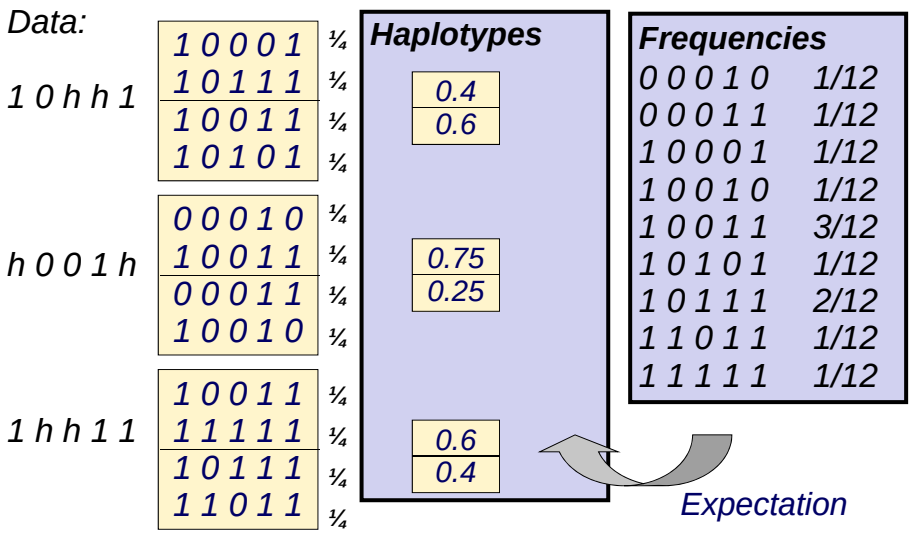
Data:

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*

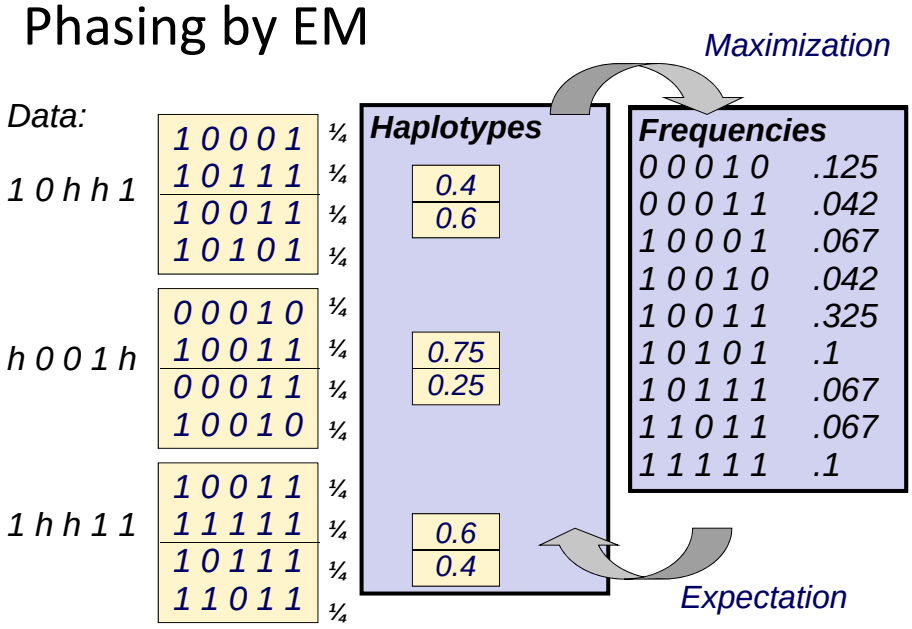
46

Phasing by EM



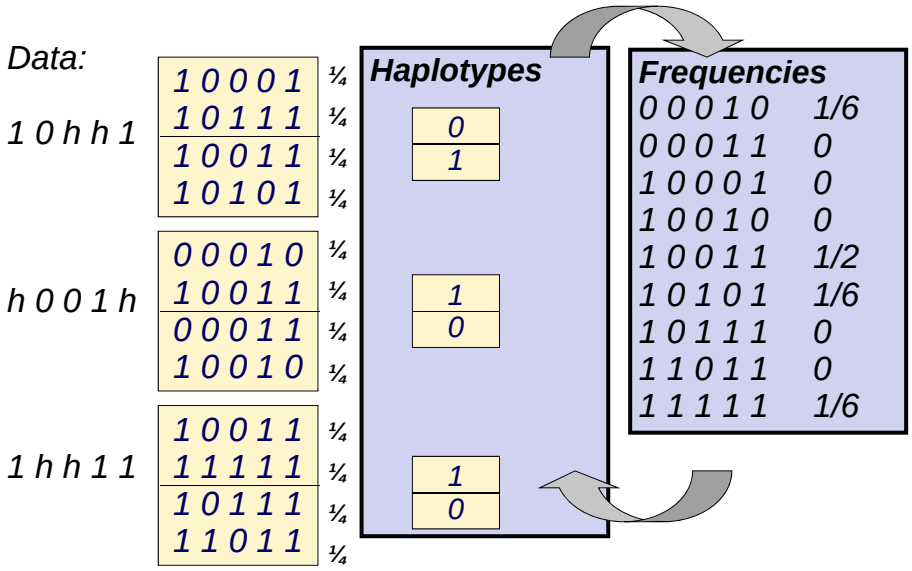
47

Phasing by EM



48

Phasing by EM



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

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

51

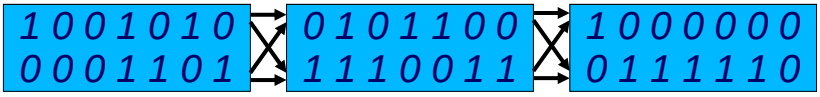
Enhancements to EM

- 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)

52

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

53

Next Topic:

DISEASE ASSOCIATION STUDIES

Why are we so different?

- Human genetic diversity

Any observable characteristic or trait

- Different **"phenotype"**
  - Appearance
  - Disease susceptibility
  - Drug responses
  - :
- Different **"genotype"**
  - Individual-specific DNA
  - 3 billion-long string

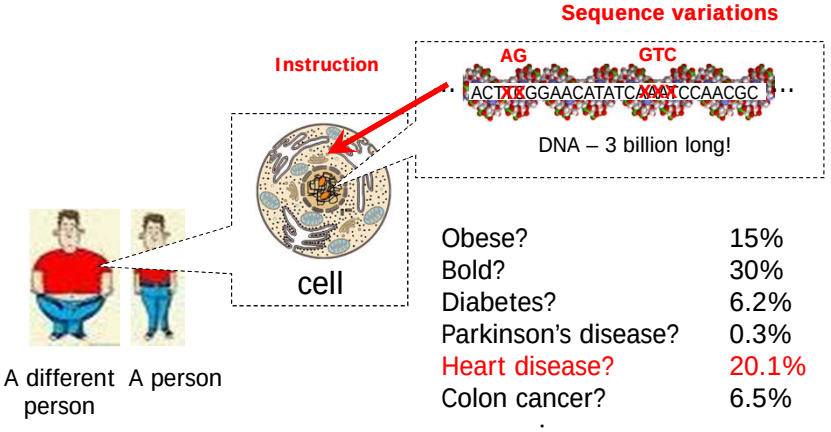
.....ACTGTTAGGCTGAGCTAGCCCAAAATTTATAGC
GTCGACTGCAGGGTCCACCAAAGCTCGACTGCAGTCGACGACCTA
AAATTTAACCGACTACGAGATGGGCACGTCACTTTTACGCAGCTTG
ATGATGCTAGCTGATCGTAGCTAAATGCATCAGCTGATGATCGTAG
CTAAATGCATCAGCTGATGATCGTAGCTAAATGCATCAGCTGATGA
TCGTAGCTAAATGCATCAGCTGATTCACTTTTACGCAGCTTGATGA
CGACTACGAGATGGGCACGTTCACCATCTACTACTACTCATCTACT
CATCAACCAAAAACACTACTCATCATCATCATCTACATCTATCATCA
TCACATCTACTGGGGGTGGGATAGATAGTGTGCTCGATCGATCGAT
CGTCAGCTGATCGACGGCAG.....

55

Motivation

Appearance, Personality, Disease susceptibility, Drug responses, ...

- Which sequence variation affects a trait?
  - Better understanding disease mechanisms
  - Personalized medicine

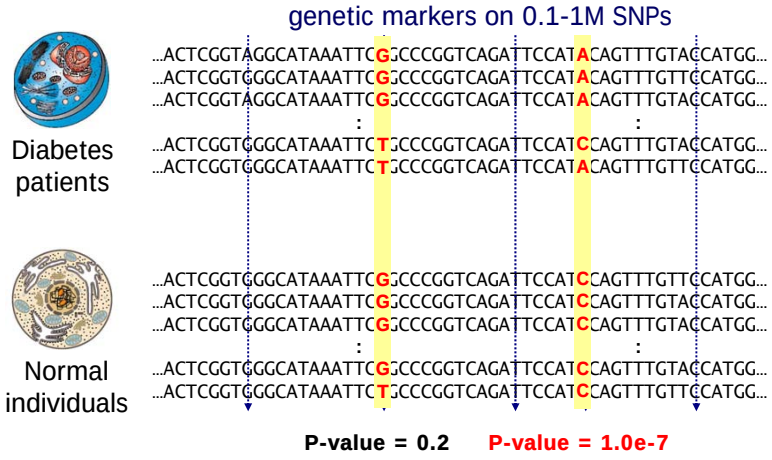


Obese?	15%
Bold?	30%
Diabetes?	6.2%
Parkinson's disease?	0.3%
Heart disease?	20.1%
Colon cancer?	6.5%
:	

56

Detecting Genetic Basis for Disease

- Genome-wide association study (**"GWAS"**)
  - P-value: The probability that we see that much correlation given that the SNP is not relevant to the disease



Outline

- Disease association studies
  - Single marker based association tests
  - Haplotype-based approach
  - Indirect association – predicting unobserved SNPs
  - Selection of tag SNPs

A single marker association test

- Data
  - Genotype data from case/control individuals
    - e.g. case: patients, control: healthy individuals
- Goals
  - Compare frequencies of particular alleles, or genotypes, in set of cases and controls
  - Typically, relies on standard contingency table tests
    - Chi-square goodness-of-fit test
    - Likelihood ratio test
    - Fisher's exact test

59

Construct contingency table

- Organize genotype counts in a simple table
  - Rows: one row for cases, another for controls
  - Columns: one of each genotype (or allele)
  - Individual cells: count of observations

i: case, control j: 0/0, 0/1, 1/1		j=1	j=2	j=3	
		0/0	0/1	1/1	
i=1	Case (affected)	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,.}=o_{1,1}+o_{1,2}+o_{1,3}$
i=2	Control (unaffected)	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,.}=o_{2,1}+o_{2,2}+o_{2,3}$
		$O_{.,1}=O_{1,1}+O_{2,1}$	$O_{.,2}=O_{1,2}+O_{2,2}$	$O_{.,3}=O_{1,3}+O_{2,3}$	

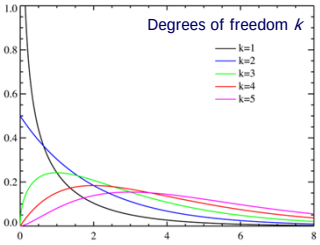
- Notation
  - Let O_{ij} denote the observed counts in each cell
  - Let E_{ij} denote the expected counts in each cell
    - $E_{ij} = O_{i,.} O_{.,j} / O_{.,.}$

60

Goodness of fit tests (1/2)

- Null hypothesis
  - There is no statistical dependency between the genotypes and the phenotype (case/control)
- P-value
  - Probability of obtaining a test statistic at least as extreme as the one that was actually observed
- Chi-square test

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$



  - If counts are large, compare statistic to chi-squared distribution
    - p = 0.05 threshold is 5.99 for 2 df (degrees of freedom, e.g. genotype test)
    - p = 0.05 threshold is 3.84 for 1 df (e.g. allele test)
  - If counts are small, exact or permutation tests are better

61

Goodness of fit tests (2/2)

- Likelihood ratio test
  - The test statistics (usually denoted D) is twice the difference in the log-likelihoods:

$$D = -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right)$$

$$= -2\ln\frac{\prod_{i,j}\left(E_{i,j}/O\right)^{O_{i,j}}}{\prod_{i,j}\left(O_{i,j}/O\right)^{O_{i,j}}} = 2\sum_{i,j} O_{i,j}\ln\frac{O_{i,j}}{E_{i,j}}$$

- How about we do this for haplotypes?
  - When does it out-perform the single marker association test?

62