Algorithms in Bioinformatics: A Practical Introduction

Population genetics

Human population

- Our genomes are not exactly the same.
- Human DNA sequences are 99.9% identical between individuals

- Those genetic variation (polymorphism) give different skin color, different outlook, and also different genetic diseases.

- This lecture would like to have a look of strategy to study human population.

Locus and Alleles

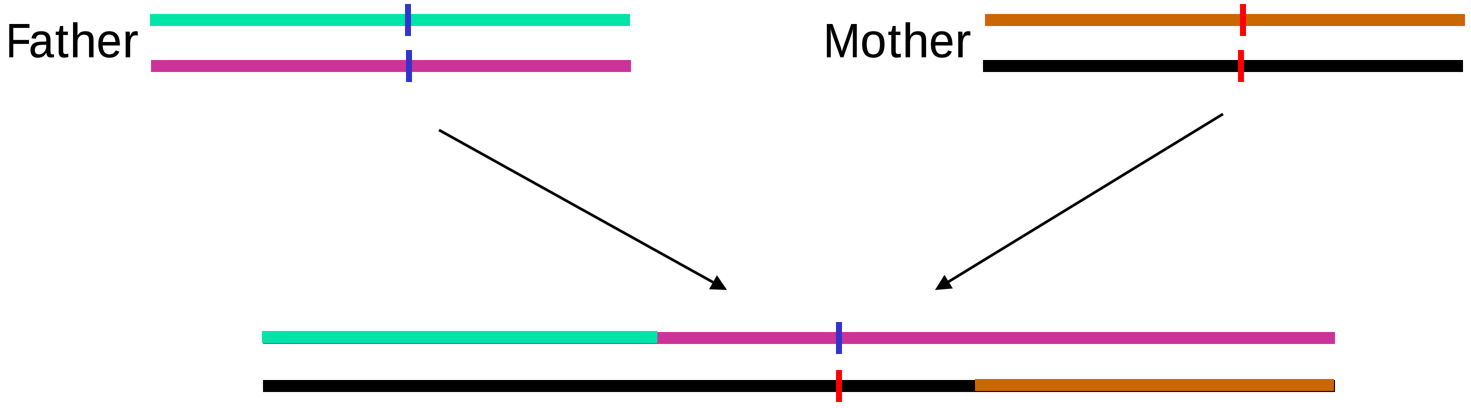
- Locus
  - A particular location in a chromosome

- An allele is a possible nucleotide that occupies a given locus.
- In the human population, a locus may have 4 possible alleles.

- Since mutation is rare, most of the loci are diallelic.

Human are diploid

- We have two copies of each chromosome
- One inherit from father while another one inherit from mother.



Locus and Alleles

- Example: Consider the following chromosome pair.

```
  i j
…ACGTCATG…
…ACGCCATG…
```

- For locus i, the allele is C.
- For locus j, the alleles are T and C.

Genotype: Homozygote vs Heterozygote

- Let *A* and *a* represent a pair of alleles of a given locus
- Then AA, aa, and Aa are the genotypes of the locus.
- *AA* and *aa* are called homozygotes.
- *Aa* is called heterozygote.

Homozygote vs Heterozygote: Example

Individual 1:	…ACGTCATG…
	…ACGCCATG…
Individual 2:	…ACGCCATG…
	…ACGCCATG…
Individual 3:	…ACGTCATG…
	…ACGTCATG…
Individual 4:	…ACGCCATG…
	…ACGTCATG…

- For the loci in red color,
  - Homozygote: Individuals 2, 3
  - Heterozygote: Individuals 1, 4

Dominance vs Recessiveness

- Let *A* and *a* represent a pair of alleles of a given locus
- A is called a **dominant** allele if
  - the appearance or phenotype of the Aa individuals resembles that of the AA type
- a is called a **recessive** allele.

Single-Nucleotide Polymorphisms (SNPs)

- SNP is the loci where there is a single nucleotide variation among different individuals. It is the most common type of polymorphism.
- Below example contains 4 pair of chromosomes.

Individual 1:	…ACGTCATG...
	…ACGCCATG…
Individual 2:	…ACGCCATG...
	…ACGCCATG…
Individual 3:	…ACGTCATG...
	…ACGTCATG…
Individual 4:	…ACGTCATG...
	…ACGTCATG…

- For the loci in red color, there is a SNP with two alleles T and C.
- The allele frequency of T is 5/8 while the allele frequency of C is 3/8.
- In this case, the minor allele frequency is 3/8.

More on SNPs

- SNPs make up 90% of all human genetic variations.
  - SNPs with a minor allele frequency of ≥ 1% occur every 100 to 300 bases along the human genome, on average.
  - Two third of the SNPs substitute cytosine (C) with thymine (T).

HapMap project

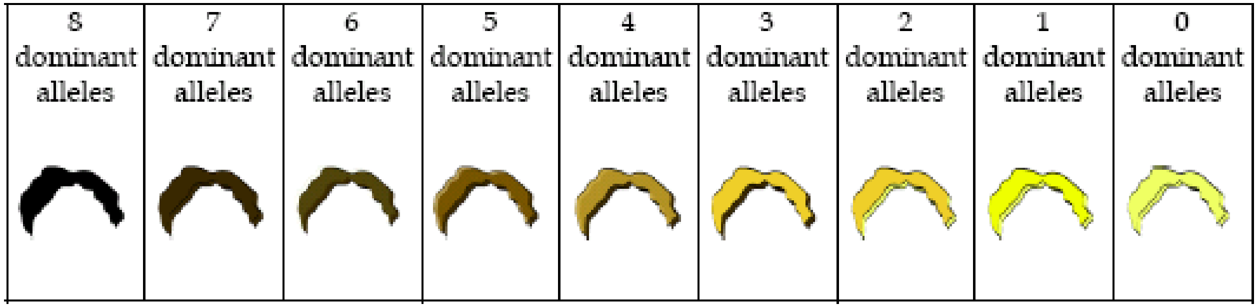
- Through the collaborative effort of many countries,
  - We already have identified the set of common SNPs in human population
  - See http://www.hapmap.org/

SNP and phenotype

- Phenotype
  - The observable structure, function or behavior of a living organism.
  - E.g. The color of the hair
- The variation of SNPs may or may not affect the phenotype.
- The SNPs which do not affect the phenotype are called natural SNPs; Otherwise, they are called causal SNPs.

Example: Hair color

- Hair color varies from black to white.
- The color of hair is control by 4 genes in on chromosome 3, 6, 10 and 18.
- The greater the number of dominant alleles, the darker the hair.



8 dominant alleles	7 dominant alleles	6 dominant alleles	5 dominant alleles	4 dominant alleles	3 dominant alleles	2 dominant alleles	1 dominant alleles	0 dominant alleles

Example: Eyebrow

- Eyebrow thickness is determined by a gene in chromosome 9.
- Thick eyebrow = ZZ or Zz while thin eyebrow = zz.

Bushy (ZZ, Zz)

Fine (zz)

- Eyebrow placement is determined by another gene in chromosome 10.
- Connected = aa while Disconnected = AA or Aa.

Not connected (AA, Aa)

Connected (aa)

Genotype frequency

- Genotype frequency is the relative frequency of a genotype on a genetic locus in a population.
- Example:
  - Let *A* and *a* represent a pair of alleles of a given locus
  - Let the population be AA, Aa, aa, AA, AA, Aa, aa, Aa, AA, Aa
  - f(AA) = 4/10
  - f(aa) = 2/10
  - f(Aa) = 4/10

Allele frequency

- Allele frequency is the relative frequency of an allele on a genetic locus in a population.
- Example:
  - Let *A* and *a* represent a pair of alleles of a given locus
  - Let the population be AA, Aa, aa, AA, AA, Aa, aa, Aa, AA, Aa
  - $p_A = (2+1+0+2+2+1+0+1+2+1)/20 = 0.6$
  - $p_a = (0+1+2+0+0+1+2+1+0+1)/20 = 0.4$

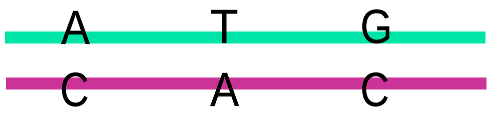
Genotype frequency → Allele frequency

- p_A = f(AA) + 0.5 f(Aa)
- p_a = f(aa) + 0.5 f(Aa)

- Example:
  - Let *A* and *a* represent a pair of alleles of a given locus
  - Let the population be AA, Aa, aa, AA, AA, Aa, aa, Aa, AA, Aa
  - p_A = 0.6, p_a = 0.4
  - f(AA) = 4/10, f(aa) = 2/10, f(Aa) = 4/10

Haplotype

- Haplotype is a combination of alleles at different loci on the same chromosome.
- For example:
  - The following three loci have genotypes AC, AT, CG.
  - There are two haplotypes: ATG and CAC.



Genotype vs haplotype

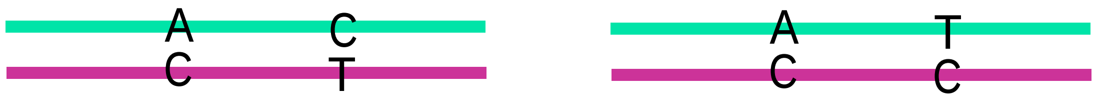
- Example: consider the following two copies of the chromsome.

```
                       i        j
Copy1 of the chr -----A--------B-------
Copy2 of the chr -----a--------b-------
```

- The genotype for loci i and j are Aa and Bb.
- Consider copy1 of the chromosome, the haplotype for loci i and j are A and B.
- Consider copy2 of the chromosome, the haplotype for loci i and j are A and B.

Technologies for studying human population

- There are 100 different genotyping technologies.
- Nowaday, we can perform whole genome genotyping for all the common SNPs found in HapMap!
  - (US$0.1-US$0.01 per genotype)
- Note that genotyping does not tell us the hapotypes appear in the chromosomes.
- E.g. The genotype of two loci are AC and CT. Then, there are two possible cases:



Bioinformatics problems

- Data quality checking
  - Check if the genotyping found by biological experiments are good or not.
- Genotype phasing
  - Identify the hapotypes from the genotypes.
- Tag SNP selection
  - Genotyping all SNPs are expensive and sometimes impossible. Hence, we want to select a subset of SNPs, called tag SNPs, for genotyping.
- Association study
  - Find the relationship between disease and genetic variation

Data quality checking

Hardy Weinberg equilibrium (HWE)

- Let p_A and p_a be the major and minor allele frequencies.
- Under the assumption:
  - Random mating
  - No natural selection

- Then, the expected frequencies are:
  - $e(AA) = p_A * p_A$
  - $e(aa) = p_a * p_a$
  - $e(Aa) = 2\ p_A * p_a$

- We expect the genotype frequencies should be similar to the expected frequencies.

Hardy Weinberg equilibrium (HWE)

- Example:
  - Let *A* and *a* represent a pair of alleles of a given locus
  - Let the population be AA, Aa, aa, AA, AA, Aa, aa, Aa, AA, Aa
  - $p_A = 0.6$, $p_a = 0.4$
  - f(AA) = 4/10, f(aa) = 2/10, f(Aa) = 4/10
- By HWE,
  - e(AA) = 0.6*0.6 = 0.36; $e_{AA} = 3.6$
  - e(aa) = 0.4*0.4 = 0.16; $e_{aa} = 1.6$
  - e(Aa) = 2*0.6*0.4 = 0.48; $e_{Aa} = 4.8$

χ^2 test for HWE

- We can use χ^2 test to determine if the genotype frequencies satisfy HWE.

- χ^2 test with degree of freedom = 1

$$\chi^2 = \frac{(n_{AA} - e_{AA})^2}{e_{AA}} + \frac{(n_{Aa} - e_{Aa})^2}{e_{Aa}} + \frac{(n_{aa} - e_{aa})^2}{e_{aa}}$$

χ^2 test for HWE: Example

- χ^2 test with degree of freedom = 1

$$\chi^2 = \frac{(4-3.6)^2}{3.6} + \frac{(4-4.8)^2}{4.8} + \frac{(2-1.6)^2}{1.6} = 0.278$$

- Pr(χ^2 > 0.278) = 0.5980
  - Which is much bigger than 0.05.
  - So we accept that the SNP satisfies HWE.

Genotype	AA	Aa	aa
Actual	4	4	2
Expected	3.6	4.8	1.6

Fisher's exact test for HWE

- n is the size of the population.
- n_{Aa} = number of Aa
- n_A = number of A.
- Number of combinations where there are n_A's A is $\binom{2n}{n_A}$
- Number of combinations where there are n_{Aa} heterozygotes is $\binom{n}{n_{AA}, n_{Aa}, n_{aa}} 2^{n_{Aa}}$
- $\Pr(n_{Aa} \mid n_A) = \dfrac{\binom{n}{n_{AA}, n_{Aa}, n_{aa}} 2^{n_{Aa}}}{\binom{2n}{n_A}}$

Fisher's exact test for HWE: Example

Genotype	AA	Aa	aa
Actual	4	4	2

- $n = 10, n_A = 12, n_{Aa} = 4.$
- $\Pr(n_{Aa} \mid n_A) = \dfrac{\dbinom{10}{4,4,2} 2^4}{\dbinom{20}{12}}$

  $= 3150 * 2^4 / 125970 = 0.40095 > 0.05$
- So, we accept that the SNP satisfies HWE.

Clean-up the dataset by HWE

- If a SNP derviates from HWE, it may be due to miscall during the genotyping process.
- Usually, we discard SNPs which derivate from HWE at significance level 10^{-3} or 10^{-4}.
- However, this approach may miss some causal SNPs.
  - In real life, there exists different forces to change the frequencies
  - The forces include selection, drift, mutation, and migration.
  - Those forces make the causal SNP derviates from HWE.

Other factors regarding clean-up

- Resolving missing genotypes

Genotype phasing

Genotype phasing

- Genotyping technology allows us to generate genotype of individual easily.
- However, it is difficult to recover the haplotype.
- The process of recovering haplotype from genotype is called genotype phasing.

Example

- Given the genotype of an individual:
  - Aa,BB,cc,DD
- We need to recover the two hapotypes of the individual, which are
  - ABcD; and
  - aBcD

Notation

- For haplotype, we use
  - 0 to represent major allele and
  - 1 to represent minor allele

- For genotype, we use
  - 0 to represent both alleles are major,
  - 1 to represent both alleles are minor, and
  - 2 to represent one is major and one is minor.

- For the previous example,
  - AaBBccDD is represented as 2010
  - ABcD is represented as 0010
  - aBcD is represented as 1010

Experimental method for genotype phasing

- Asymmetric PCR amplification (Newton et al. 1989; Wu et al. 1989)
- Isolation of single chromsome by limit dilution followed by PCR amplification (Ruano et al. 1990)
- Inferring haplotype information by using genealogical information in families (Perlin et al. 1994)

- The above methods are low-throughput, costly, and complicated.

Computational methods

- We study computational methods for genotype phasing.
- We discuss the following:
  - Clark's algorithm
  - Perfect Phylogeny Haplotyping
  - Maximum likelihood
  - Phase (just mention)

Difficulty of genotype phasing

- Consider the following example.

Genotype: 01211201

- Which one is correct? (I) or (II)?

(I) Haplotype: 01011101
01111001

OR

(II) Haplotype: 01111101
01011001

Genotype phasing Problem

- Input:
  - A set of genotypes $G=(G_1, G_2, \ldots, G_n)$.
- Output:
  - A set of haplotypes which can best explain G according to certain criteria.

- Example Criteria:
  - Minimize the number of haplotypes
  - Maximize the likelihood
  - ...

Clark’s algorithm (1990)

- Parsimony approach: Find the simplest solution
  - Minimize the total number of haplotypes.

- He gave a heuristics algorithm.

1. From all homozygotes and single-site heterozygotes genotypes,
   - Unambiguously, we generate a set of haplotypes.
2. For each know haplotype H, we look for unresolved genotype G’,
   - Check if we can resolve G’ by H and some new haplotype H’.
   - If yes, include H’ and resolve G’.
3. Repeat the procedure until all genotypes are resolved.

- Note that Clark’s algorithm may fail to return answer.

Example for Clark's algorithm Step 1

- Example genotype input:
  - G_1 = 10121101
  - G_2 = 10201121
  - G_3 = 20001211
- From G_1, we have
  - H_1 = 10101101
  - H_2 = 10111101

Example for Clark's algorithm Step 2

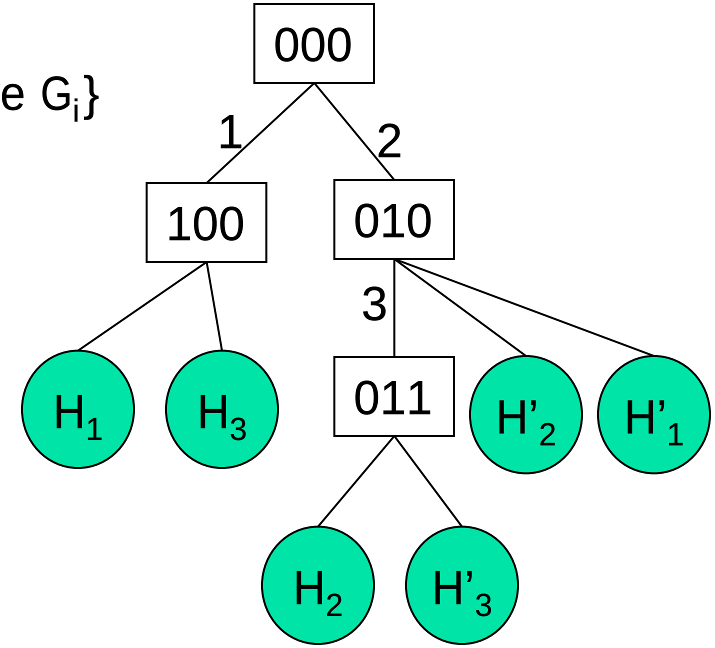
- Example genotype input:
  - G_1 = 10121101
  - G_2 = 10201121
  - G_3 = 20001211
- We have the following haplotypes:
  - H_1 = 10101101
  - H_2 = 10111101

- From H_1 and G_2, we have
  - H_3 = 10001111
- From H_3 and G_3, we have
  - H_4 = 00001011

- Hence, the set of predicted haplotypes is
  - H_1 = 10101101
  - H_2 = 10111101
  - H_3 = 10001111
  - H_4 = 00001011

Perfect Phylogeny Haplotyping

- This problem is first introduced by Gusfield 2002.
- Input:
  - A set of genotypes $G=\{G_1, \ldots, G_n\}$, each G_i is a length-m genotype.
- Output:
  - A set of haplotypes $H=\{H_i,H'_i \mid H_i,H'_i \text{ resolve } G_i\}$ such that $H_1,H'_1 \ldots, H_n,H'_n$ form a perfect phylogeny
- For example,
  - $G=\{G_1=220, G_2=012, G_3=222\}$
  - The solution is $H=\{100, 010, 011\}$



Previous work

- Gusfield (2002) introduced the problem and gives an $O(nm\ \alpha(nm))$ time algorithm by reduction to the graph realization problem
- Eskin et al (2002) gives a simple $O(nm^2)$ time algorithm.
- Bafna et al (2002) gives a simple $O(nm^2)$ time algorithm.
- Gusfield et al (RECOMB 2005) gives an $O(nm)$ time algorithm.

Represent G as a matrix

- To simplify the discussion, we represent $\{G_1,\ldots,G_n\}$ as a nxm matrix G where the entry G(i,j) is the j genotype of G_i.

	1	2	3	4	5	6
G_1	1	1	2	0	2	0
G_2	1	2	2	0	0	2
G_3	1	1	2	2	0	0
G_4	2	2	2	0	0	2
G_5	1	1	2	2	2	0

Our aim

- Given n x m matrix G
  - Each entry is either 0, 1, or 2

	1	2	3
G_1	2	2	0
G_2	0	1	2
G_3	2	2	2

- Construct 2n x m matrix H
  - Each entry is either 0 or 1
  - If $G(r,c) \neq 2$, $H(2r,c)=H(2r-1,c)=G(r,c)$
  - Otherwise, $\{H(2r,c),H(2r-1,c)\}=\{0,1\}$
  - H satisfies a perfect phylogeny

	1	2	3
H_1	1	0	0
H'_1	0	1	0
H_2	0	1	1
H'_2	0	1	0
H_3	1	0	0
H'_3	0	1	1

4-gamete test

- A set of haplotypes admits a perfect phylogeny (whose root is an all-0 haplotypes) if and only if there are no two columns i and j containing all four pairs 00, 01, 10, and 11.

- Proof:
  - Recall that M admits a perfect phylogeny if and only if for every characters i and j, they are pairwise compatible.

In-phase and out-of-phase

- If some columns c and c’ in G contain (1) either 11 or 12 or 21 and (2) either 00 or 02 or 20,
  - columns c and c’ in H must contain both 11 and 00.
  - In such case, c and c’ are called in-phase.

- If some columns c and c’ in G contain (1) either 10 or 20 and (2) either 01 or 02,
  - Columns c and c’ in H must contain both 10 and 01.
  - In such case, c and c’ are called out-of-phase.

- E.g.
  - Columns 2 and 5 are in-phase
  - Columns 4 and 5 are out-of-phase
  - Columns 3 and 4 are neither in-phase or out-of-phase

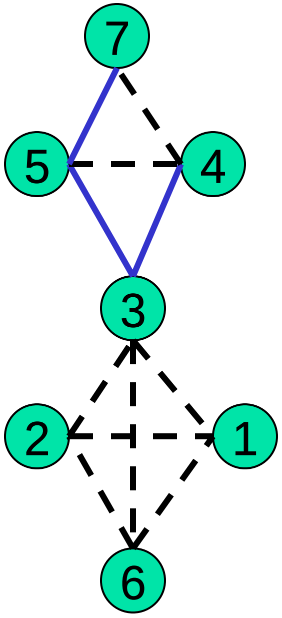
	1	2	3	4	5	6
G_1	1	1	2	0	2	0
G_2	1	2	2	0	0	2
G_3	1	1	2	2	0	0
G_4	2	2	2	0	0	2
G_5	1	1	2	2	2	0

- If columns c and c’ in G are both in-phase and out-of-phase, G has no solution to the PPH problem.
    - Proof: By 4-gamete test

G_M

- In G_M, a pair of columns forms an edge if it contains 22.
- Red: in-phase (color 0)
- Blue: out-of-phase (color 1)

	1	2	3	4	5	6	7
G_1	1	1	0	2	2	0	2
G_2	1	2	2	0	0	2	0
G_3	1	1	2	2	0	0	0
G_4	2	2	2	0	0	2	0
G_5	1	1	2	2	2	0	0
G_6	1	1	0	2	0	0	2



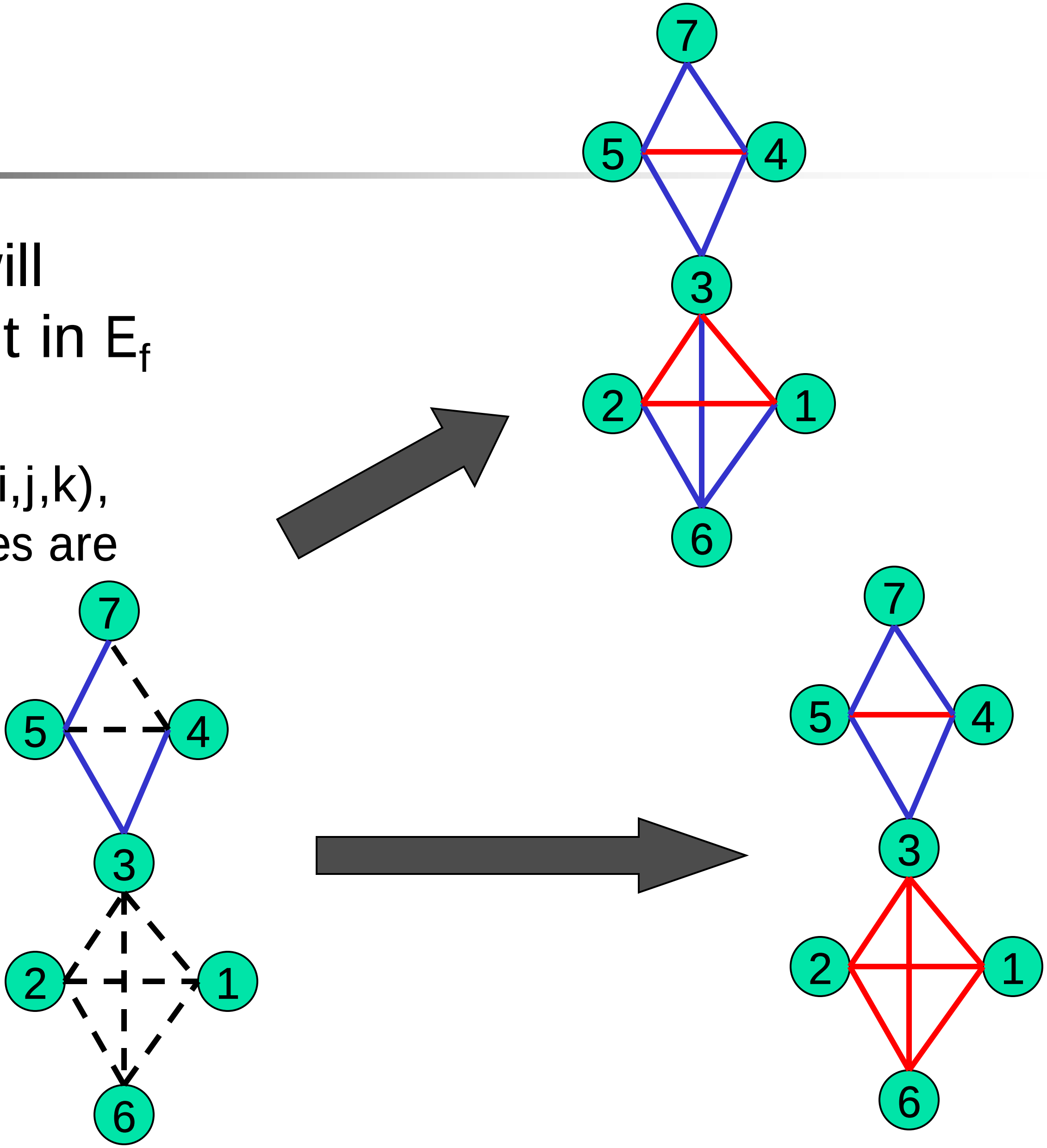
Theorem

- Consider a matrix M such that every pair of columns is not both in-phase and out-of-phase.
- There exists a PPH solution for M if and only if we can infer the colors of all edges in G_M such that
  - All edges which are in-phase and out-of-phase are colored red and blue, respectively. (Denote E_f be the set of these edges);
  - For any triangle (i,j,k) where there exists r s.t. M[r,i]=M[r,j]=M[r,k]=2, either 0 or 2 edges are colored blue.

- If such coloring exists, such coloring is called a valid coloring of G_M.

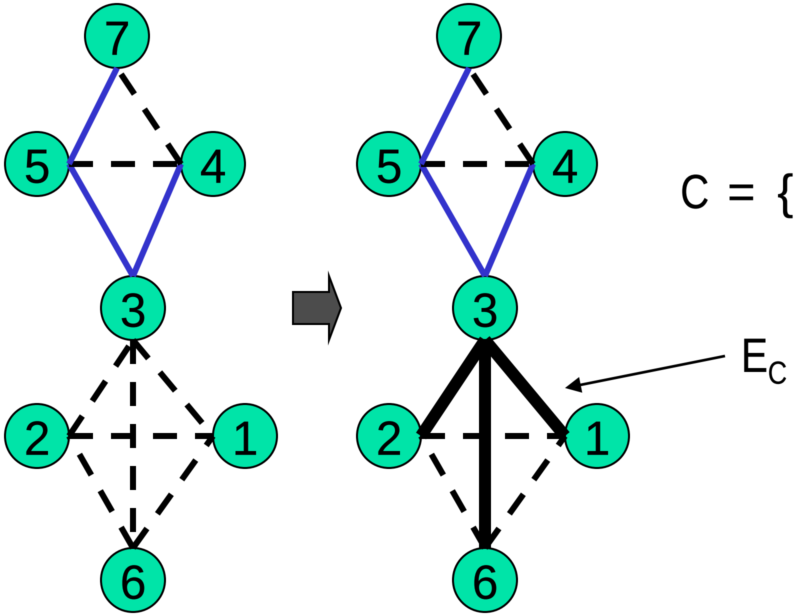
Infer colors for the uncolored edges

- A valid coloring will color all edges not in E_f so that
  - For any triangle (i,j,k), either 0 or 2 edges are colored blue.



How to infer the colors? (I)

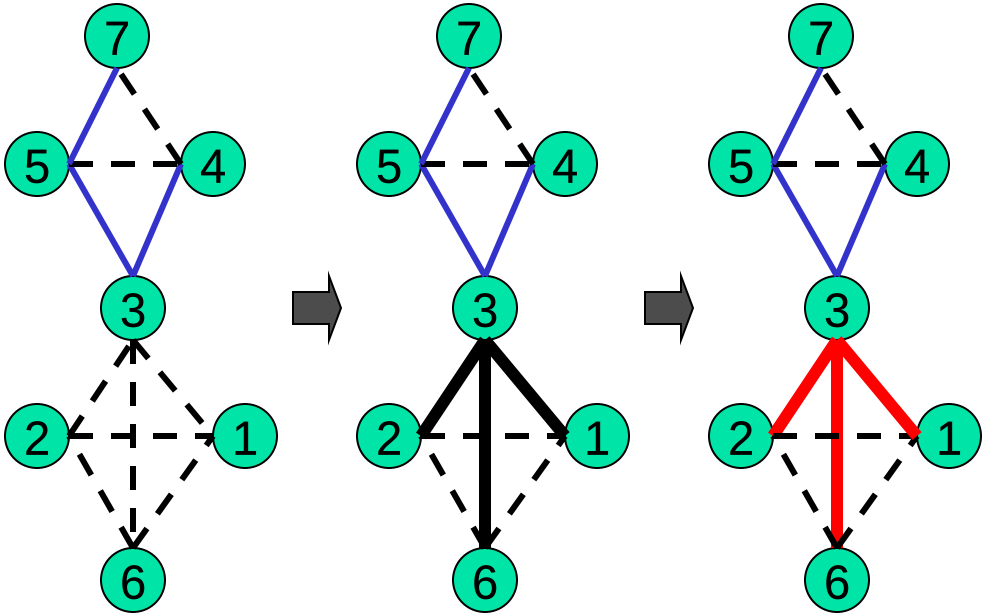
- The colored edges in G_M form a set C of connected components.
- Let E_C be a minimum set of edges, which connect all these connected components.



C = { {3,4,5,7}, {2}, {1}, {6} }

How to infer color? (II)

- Bafna et al. showed the following theorem:
  - Either (1) G_M has no valid solution or (2) any arbitrary coloring of the edges in E_C define a unique valid coloring for G_M. (Thus, there are exactly 2^r valid coloring, where $r=|E_C|$.)

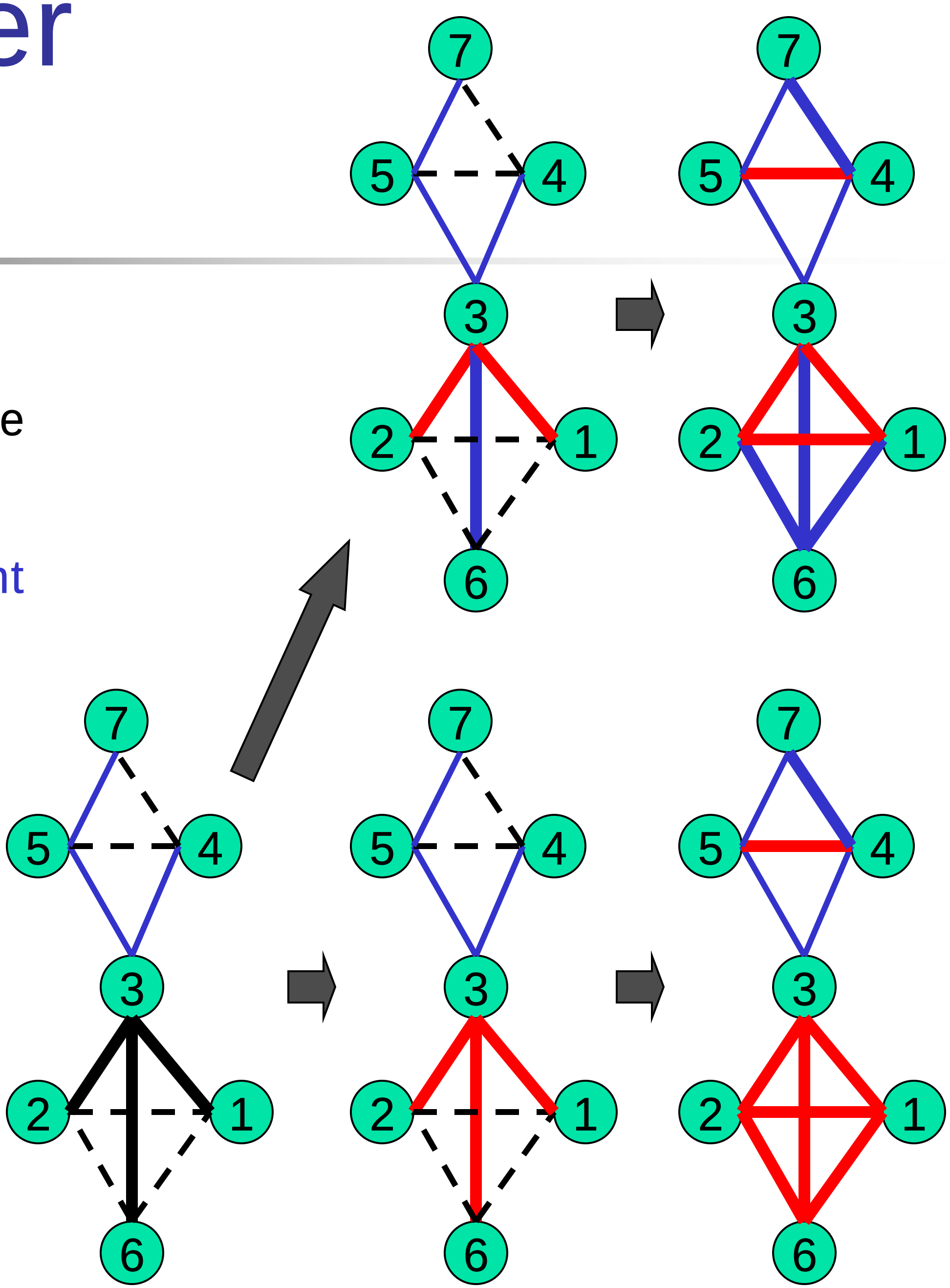


How to infer color? (III)

- Given the coloring of E_C, the colors of the dotted edges can be inferred as follows.

- While a dotted edge e is adjacent to two colored edges,
  - Color e so that the triangle has either 0 or 2 blue edges.

- Bafna et al. showed the above algorithm can infer the color of all dotted edges correctly.



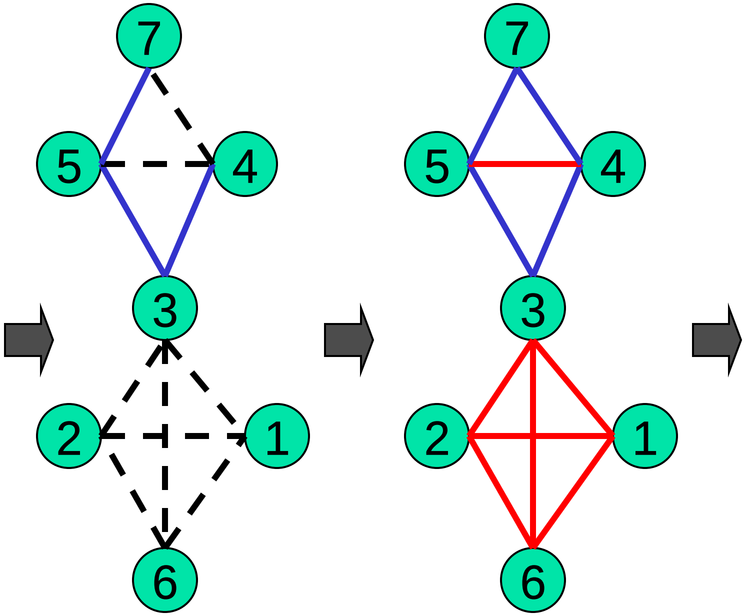
How to infer the haplotypes?

- Given the coloring of all edges of G_M, we can infer the haplotypes as follows.

- For j = 1 to m,
  - For i = 1 to n,
    - if $M[i,j] \in \{0,1\}$, set $H[2i,j]=H[2i-1,j]=M[i,j]$
    - Otherwise, let $k<j$ be a column such that $M[i,k]=2$.
    - If k exists,
      - if (j,k) is colored red, set $H[2i,j]=H[2i,k]$, $H[2i-1,j]=1-H[2i,j]$
      - If (j,k) is colored blue, set $H[2i,j]=1-H[2i,k]$, $H[2i-1,j]=1-H[2i,j]$
    - Else
      - set $H[2i,j]=0$, $H[2i-1,j]=1$

Example

	1	2	3	4	5	6	7
G_1	1	1	0	2	2	0	2
G_2	1	2	2	0	0	2	0
G_3	1	1	2	2	0	0	0
G_4	2	2	2	0	0	2	0
G_5	1	1	2	2	2	0	0
G_6	1	1	0	2	0	0	2



	1	2	3	4	5	6	7
H_1	1	1	0	1	1	0	0
H'_1	1	1	0	0	0	0	1
H_2	1	1	1	0	0	1	0
H'_2	1	0	0	0	0	0	0
H_3	1	1	1	0	0	0	0
H'_3	1	1	0	1	0	0	0
H_4	1	1	1	0	0	1	0
H'_4	0	0	0	0	0	0	0
H_5	1	1	0	1	1	0	0
H'_5	1	1	1	0	0	0	0
H_6	1	1	0	1	0	0	0
H'_6	1	1	0	0	0	0	1

Time analysis

- Checking in-phase and out-of-phase for all pairs of columns takes O(nm^2) time.
- Infering colors for the uncolored edges takes O(m^2) time.
- Compute the matrix H takes O(nm) time.

- In total, the algorithm runs in O(nm^2) time.

More on PPH problem

- Theorem: If every column in M contains at least one 0 and one 1 entry,
  - Then there is either no PPH solution for M or has a unique PPH solution for M.
  - Also, such solution can be found in O(nm) time.

Maximum likelihood approach

- This approach is used by Excoffier and Slatkin (1995).
- Try to infer the haplotype with the most realistic haplotype frequencies
  - under the assumption of Hardy-Weinberg equilibrium

Motivation (I)

- Example: Consider two genotypes
  - G_1 = 0111
  - G_2 = 0221
- Two possible solutions:

G_1:	0111	G_1:	0111
	0111		0111
G_2:	0111	G_2:	0101
	0001		0011

- Which solution is better?

Motivation (II)

- For solution 1:

  G_1: 0111
  0111
  G_2: 0111
  0001

  - There are two haplotypes 0111 and 0001.
  - Their frequencies are ¾ and ¼.
  - The chance of getting G_2=0221 is ¾*¼.

- For solution 2:

  G_1: 0111
  0111
  G_2: 0101
  0011

  - There are three haplotypes 0111, 0101, and 0011.
  - Their frequencies are ½, ¼ and ¼.
  - The chance of getting G_2=0221 is ¼*¼.

- Solution 1 seems better!

Preliminary

- Given a genotype G_i, we can generate the set S_i, which is the set of all haplotype pairs that are phased genotypes of G_i.

- Example: Consider the genotype 0221.
  - Since there are two heterozygous loci,
    - we have $2^2 = 4$ possible haplotypes.
    - h_1=0001, h_2=0011, h_3=0101, h_4=0111
  - The set of all phased genotypes of 0221 is
    - $\{h_1h_4, h_2h_3\}$.

Maximum Likelihood (I)

- Let G = {G_1, G_2, ..., G_n} be the set of n genotypes.
- Let h_1, h_2, ..., h_m be the set of all possible haplotypes that can resolve G.
- Let F={F_1, F_2,..., F_m} be the population frequency of {h_1, h_2, ..., h_m}.
  - Note: $F_1+F_2+\ldots+F_m=1$
- For x = 1, 2, ..., n,

$$\Pr(G_x \mid F) = \sum_{\substack{h_i h_j \text{ is a} \\ \text{phased genotype} \\ \text{of } G_x}} (F_i \cdot F_j)$$

Maximum Likelihood (II)

- We would like to maximize the overall probability product of all P(G_i), that is, the following function L.

$$L(F) = \Pr(G \mid F) = \alpha \prod_{i=1..n} \Pr(G_i \mid F)$$

- In principle, we can solve this equation. But there is no close form.
- Instead, we use EM algorithm.

Formal definition of Maximum likelihood

- Given
  - a set of observations $X=\{x_1,x_2,\ldots,x_n\}$
  - A set of parameters Θ.
- The likelihood function:
  - $L(\Theta)=\Pi_{i=1..n}Pr(x_i|\Theta)=Pr(X|\Theta)$
- Aim:
  - Find $\Theta' = \text{argmax}_\Theta\ Pr(X|\Theta)$
    $= \text{argmax}_\Theta\ \Pi_{i=1..n}\ Pr(x_i|\Theta)$

Hidden data

- x_i is called observed data
  - Each x_i is associated with some hidden data y_i.
- Finding $\Theta' = \text{argmax}_{\Theta} \Pr(X|\Theta)$ may be difficult.
- Moreover, finding $\text{argmax}_{\Theta} \Pr(X,Y|\Theta)$ may be easier.

What is EM algorithm?

- EM algorithm is a popular method for solving the maximum likelihood problem.

- The idea is to alternate between
  - Filling in Y based on the best guess Θ; and
  - Maximizing Θ with Y fixed.

EM Algorithm

- Initialization: A guess at Θ
- Repeat until satisfy
  - **E-step:** Given a current fixed Θ', compute $\Pr(y|x,\Theta')$
  - **M-step:** Given $\Pr(y|x,\Theta')$, find Θ which maximizes $\Sigma_x \Sigma_y \Pr(y|x,\Theta') \log \Pr(x,y|\Theta)$

Explanation of EM-algorithm (I)

- Let Θ’ be the old guess.
- Maximizing L(Θ) is the same as maximizing R(Θ,Θ’) = L(Θ)/L(Θ’)
  - since Θ’ is fixed.

$$
\begin{aligned}
R(\Theta,\Theta') &= \frac{\prod_x \sum_y \Pr(x,y\,|\,\Theta)}{\prod_x \Pr(x\,|\,\Theta')} \\
&= \prod_x \frac{\sum_y \Pr(x,y\,|\,\Theta)}{\Pr(x\,|\,\Theta')} \\
&= \prod_x \sum_y \frac{\Pr(x,y\,|\,\Theta)}{\Pr(x\,|\,\Theta')} \\
&= \prod_x \sum_y \frac{\Pr(x,y\,|\,\Theta')}{\Pr(x\,|\,\Theta')} \frac{\Pr(x,y\,|\,\Theta)}{\Pr(x,y\,|\,\Theta')} \\
&= \prod_x \sum_y \Pr(y\,|\,x,\Theta') \frac{\Pr(x,y\,|\,\Theta)}{\Pr(x,y\,|\,\Theta')}
\end{aligned}
$$

Explanation of EM-algorithm (II)

- By AM≥GM, we have

$$R(\Theta,\Theta') = \prod_x \sum_y \Pr(y \mid x,\Theta') \frac{\Pr(x,y \mid \Theta)}{\Pr(x,y \mid \Theta')}$$

$$\geq \prod_x \prod_y \left[\frac{\Pr(x,y \mid \Theta)}{\Pr(x,y \mid \Theta')} \right]^{\Pr(y \mid x,\Theta')}$$

- By taking log and Θ' is a constant, maximizing R(Θ,Θ') is the same as maximizing Q(Θ,Θ') where

$$Q(\Theta,\Theta') = \sum_x \sum_y \Pr(y \mid x,\Theta') \log \Pr(x,y \mid \Theta)$$

Example: Genotype phasing

- $G = \{G_1, G_2, \ldots, G_n\}$ which are the set of observed genotypes.
- Let $\{h_1, h_2, \ldots, h_m\}$ be the set of all possible haplotypes that can resolve G.
- Θ is set of haplotype frequencies $\{F_1, F_2, \ldots, F_m\}$ where F_x is the frequency of h_x.
- Aim:
  - Find $\Theta' = \mathrm{argmax}_{\Theta} \Pr(G|\Theta)$

Example: Genotype phasing

- For each genotype G_i,
  - The hidden data is its phase $h_x h_y$.

- $\Pr(h_x h_y, G_i | \Theta) = F_x\ F_y$.

Example: Genotype phasing EM algorithm

- Initialization: $F^{(0)} = \{F_1^{(0)}, F_2^{(0)}, \ldots, F_m^{(0)}\}$.
- Repeat the following two steps:
- E-step:
  - For every G_x, estimate the phased genotype frequencies $P(h_i h_j | G_x, F^{(g)})$ for all $h_i h_j$ that is consistent with G_x.
- M-step:
  - Based on the phased genotype frequencies, we estimate a new set $F^{(g+1)}$ of haplotype frequencies.

Example: Genotype phasing E-step

- Suppose $h_x h_y$ is a phased genotype of G_i.

$$P(h_x h_y \mid G_i, F^{(g)}) = \frac{F_x^{(g)} F_y^{(g)}}{\sum \{F_{x'}^{(g)} F_{y'}^{(g)} \mid h_{x'} h_{y'} \text{ is a phased genotype of } G_i\}}$$

Example: Genotype phasing M-step

- M-step: Maximizes Q(Θ,Θ’)

$$Q(\Theta,\Theta') = \sum_{i=1..n} \sum_{\substack{h_x h_y \text{is a phased} \\ \text{genotype of } G_i}} \Pr(h_x h_y \mid G_i, \Theta') \log \Pr(h_x h_y, G_i \mid \Theta)$$

$$= \sum_{i=1..n} \sum_{\substack{h_x h_y \text{is a phased} \\ \text{genotype of } G_i}} \Pr(h_x h_y \mid G_i, \Theta') \log (F_x F_y)$$

$$= \sum_{x} \left(\sum_{i=1..n} \sum_{\substack{h_x h_y \text{is a phased} \\ \text{genotype of } G_i}} \Pr(h_x h_y \mid G_i, \Theta') \right) \log F_x$$

Example: Genotype phasing M-step

- To maximize $\Sigma_x(a_x \log F_x)$ such that $\Sigma_x F_x = 1$
    - The solution is $F_x = a_x / (\Sigma_x a_x)$ for all x.

- Hence, M-step is:

$$F_x^{(g+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{\substack{h_x h_y \text{ is a} \\ \text{phased genotype} \\ \text{of } G_i}} \delta(h_x, h_x h_y) P(h_x h_y \mid G_i, F^{(g)})$$

where $\delta(h,H)$ is the number of occurrences of h in the phased genotype H

Example

- $G=\{G_1=11, G_2=12, G_3=22\}$.
- Possible haplotypes of G: $h_1=11$, $h_2=00$, $h_3=10$, $h_4=01$
- Let F_1, F_2, F_3, and F_4 be the corresponding haplotype frequencies. (Suppose $F_i=0.25$ for all i.)

- h_1h_1 is the only possible phased genotype of G_1.
  - $P(h_1h_1| G_1, F) = 1$
- h_1h_3 is the only possible phased genotype of G_2.
  - $P(h_1h_3| G_2, F) = 1$
- h_1h_2 and h_3h_4 are the possible phased genotype of G_3.
  - $P(h_1h_2|G_3, F) = (F_1 F_2)/(F_1 F_2 + F_3 F_4)=1/2$
  - $P(h_3h_4|G_3, F) = (F_3 F_4)/(F_1 F_2 + F_3 F_4)=1/2$

Example

- $G=\{G_1=11, G_2=12, G_3=22\}$. $(n=3)$
- Possible haplotypes of G: $h_1=11$, $h_2=00$, $h_3=10$, $h_4=01$

- $P(h_1h_1| G_1,F) = 1$
- $P(h_1h_3| G_2,F) = 1$
- $P(h_1h_2| G_3,F) = 1/2$
- $P(h_3h_4| G_3,F) = 1/2$

- $F'_1 = [2P(h_1h_1| G_1,F)+ P(h_1h_3| G_2,F)+ P(h_1h_2|G_3,F)]/2/n = 7/12$
- $F'_2 = P(h_1h_2|G_3,F)/2/n = 1/12$
- $F'_3 = [P(h_1h_3| G_2,F) + P(h_3h_4|G_3,F)]/2/n = 3/12$
- $F'_4 = P(h_3h_4|G_3,F)/2/n = 1/12$

Phase

- When there are many heterozygous loci, EM algorithm becomes slow since there are exponential number of haplotypes.

- Phase resolves this problem. More importantly, it improves the accuracy.

- Phase is a Bayesian-based method which uses Gibbs sampling.

Motivation (I)

- Given a set of known haplotypes
  - 4’s 10001
  - 5’s 11110
  - 3’s 00101

- For the ambiguous genotype 20112, two possible solutions:

(A) 10110
00111 (B) 10111
00110

- Which solution is better?

Motivation (II)

- Given a set of known haplotypes
  - 4’s 10001
  - 5’s 11110
  - 3’s 00101

(A) 10110
00111

(B) 10111
00110

- Solution (A) is better since the two haplotypes look similar to some known high frequency haplotypes.

Mutation model

- Given a set H of haplotypes, for any haplotype h, it is shown that Pr(h|H) is

$$\sum_{\alpha \in H} \sum_{s=0}^{\infty} \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta} \right)^s \frac{n}{n+\theta} \left(P^s \right)_{\alpha h}$$

- where
  - n=|H|, θ is the scaled mutation rate,
  - n_α is the number of occurrences of haplotype α in H, and
  - P is mutation matrix

- Phase try to use Gibbs sampling to predict the haplotype phase of G.

- For any haplotype $H_i=(h_{i1},h_{i2})$
  - $Pr(H_i|G,H_{-i}) \propto Pr(H_i|H_{-i}) \propto Pr(h_{i1}|H_{-i})Pr(h_{i2}|H_{-i})$

Phase algorithm

- Initialization: Let $H^{(0)} = \{H_1^{(0)},\ldots, H_n^{(0)}\}$ be the initial guess of the phase haplotypes of G.

1. Uniformly randomly choose an ambiguous individual G_i (i.e., individuals with more than one possible haplotype reconstruction).
2. Sample $H_i^{(t+1)}$ from $\Pr(H_i \mid G,H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i.
3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j = 1,\ldots,n,\ j \neq i$.

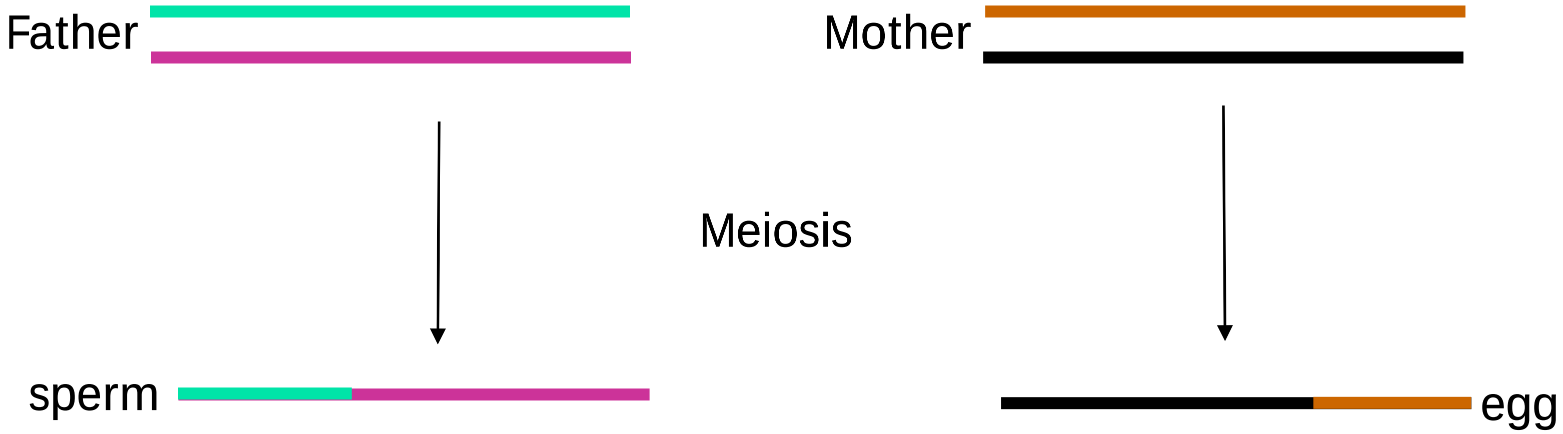
References

- Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122
- Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927. [EM algorithm]
- Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989. [Phase]
- Paul Scheet and Matthew Stephens (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-644. [FastPhase]

Linkage disequilibrium

Is recombination randomly distributed on the genome?

- Recombination occurs in the evolution process.
- Is the recombination cut the genome at random position?



Recombination hotspot evident (I)

- Daly et al (2001) study 500kb region on chromosome 5q31
  - Broken into a series of discrete haplotype blocks that range in size from 3-92kb.
  - Each haplotype block corresponded to a region in which there were just a few common haplotypes (2-4 per block)
- Jeffreys et al (2001) study the class II major histocompatability complex (MHC) region from single-sperm typing.
  - Most of the recombinations are restricted to narrow recombination hotspots.

Recombination hotspot evident (II)

- Many other studies also found that recombination tends to cluster in hotspots that are roughly 102kb in length.
- For haplotype block, it can be very long (says 804kb for a haplotype block on chromosome 22). Most of the haplotype blocks are of length about 5-20kb.
- Hence, it is conjecture that
  - The genome might be divided into regions of high LD that are separated by recombination hotspots.

Correlation between recombination hotspots and genomic features

- By Li et al (AJGH2006), a recombination hotspot is correlated with
  - High G+C content
  - Less repeat. In detail:
    - Less L1
    - More MIR, L2, and low_complexity
  - Less gene region
  - High DNaseI hypersensitivity

Linkage disequilibrium (LD)

- LD refers to the non-random association between alleles at two different loci.
  - that is, two particular alleles can co-occur more often than expected by chance.

- There are two important LD measurements:
  - D;
  - D’; and
  - r^2

D

- Loci 1: either A or a ($p_a + p_A = 1$)
- Loci 2: either B or b ($p_b + p_B = 1$)
- If loci 1 and 2 are independent,
  - $p_{AB}=p_A\ p_B$
  - $p_{Ab}=p_A\ p_b$
  - $p_{aB}=p_a\ p_B$
  - $p_{ab}=p_a\ p_b$
- If LD presents (says, A associate with B), then
  - $p_{AB}=p_A\ p_B + D_1$
  - $p_{Ab}=p_A\ p_b - D_2$
  - $p_{aB}=p_a\ p_B - D_3$
  - $p_{ab}=p_a\ p_b + D_4$
  - We can show that $D_1=D_2=D_3=D_4=D$.
  - D is known as the linkage disequilibrium coefficient
  - D is in the range -0.25 to 0.25. D = 0 under linkage equilibrium

D’

- D is highly dependent on the allele frequency and is not good for measuring the strength of LD.
- Define D’ = D / D_{max}
  - where D_{max} is the maximum possible value for D given p_A and p_B.
  - Note: $D_{max} = \min\{p_A,p_B\}-p_Ap_B$.
- When |D’|=1, we say it is a complete LD.

Example

- AB, Ab, aB, Ab, ab, ab, ab.
- p_{AB}=1/7, p_A=3/7, and p_B=2/7.
- Hence, D = 1/7 – 3/7*2/7 = 1/49.

- Given p_A=3/7, p_B=2/7, the max value for p_{AB} = min{p_A, p_B} = 2/7. Hence, D_{max}=2/7 – 3/7*2/7 = 8/49.

- Hence, D’ = D / D_{max} = 1/8.

r^2

- r^2 measures the correlation of two loci.
- Define $r^2 = D^2 / (p_A\, p_a\, p_B\, p_b)$.
- When $r^2 = 1$,
  - If we know the allele on loci 1, we can deduce the allele on loci 2, and vice versa.
  - Called perfect LD.

Example

- AB, Ab, aB, Ab, ab, ab, ab.
- p_{AB}=1/7, p_A=3/7, and p_B=2/7.
- Hence, D = 1/7 – 3/7*2/7 = 1/49.

- $r^2 = (1/49)^2/(3/7*4/7*2/7*5/7) = 1/120$.

Tag SNP selection

- There are about 10 million common SNPs (SNPs with allele frequency > 1%).
- It accounts for ~90% of the human genetic variation.
- Hence, we can study the genetic variation of an individual by getting its profile for the common SNPs.

- Even though the cost of genotyping is rapidly decreasing, it is still impractical to genotype every SNP or even a large proportion of them.

- Fortunately, nearby SNPs using show strong correlation to each other (i.e. strong LD).
- It is possible to define a subset of SNPs (called tag SNPs) to represent the rest of the SNPs.

Idea of Zhang et al PNAS 2002

- Assume the genome can be blocked so that the SNPs in each block has high LD.

- Partition the genome into blocks.
- Within each block, we select a minimum set of tag SNPs which can distinguish the haplotypes in the block.

- Aim: minimizing the total number of tag SNPs.

- **Input**: a set of K haplotypes, each is described by n SNPs.
- Denote $r_i(k)$ be the allele of the i-th SNP in the k-th haplotype.
  - where $r_i(k)$ = 0, 1, 2 where 0 means missing data.

- **Output**: A set of blocks, each block is r_i ... r_j.
  - For each block, a set of tag SNPs which can distinguish at least α% of the unambiguous haplotypes (defined in the next slide).
  - The total number of tag SNPs is minimized.

Example

- (1,2,1, 2,1,0,1, 1,1,2,1)
- (1,0,1, 1,0,1,2, 1,1,0,1)
- (0,2,1, 0,1,2,1, 1,0,2,2)
- (2,1,2, 2,1,2,1, 2,2,1,2)
- (2,0,2, 1,2,1,0, 2,0,1,2)
- (2,1,0, 1,2,0,2, 1,2,2,2)

- For the above example, we may want to partition them into 3 blocks: $r_1..r_3$, $r_4..r_7$, $r_8..r_{11}$.
- For block $r_1..r_3$, we select r_1 as the tag SNP.
- For block $r_4..r_7$, we select r_4 as the tag SNP.
- For block $r_8..r_{11}$, we select r_8 and r_{11} as the tag SNPs.

Ambiguous

- Two haplotypes in a block are compatible if the alleles are the same for all loci with no missing values.
- Example:
  - h_1=(1, 2, 0, 0), h_2=(0, 2, 1, 2), h_3=(1, 2, 1, 1).
  - h_1 is compatible with h_2 and h_3. However, h_2 is not compatible with h_3.

- A haplotype h in a block is ambiguous if h is compatible with h' and h'' but h' is not compatible with h''.
- For the above example, h_1 is ambiguous in the block.

block(r_i, …, r_j)

- Within a block, we can cluster the haplotypes into different groups,
  - Each group contains unambiguous haplotypes which are compatible.
  - A haplotype in a group is called common if its group is of size at least two.
- We want most of the haplotypes in a block are unambiguous.
- Formally, we define block(r_i, …, r_j) = 1 if there are >β% common unambiguous haplotypes.

$f(r_i \ldots r_j)$

- We denote $f(r_i \ldots r_j)$ = the minimum number of tag SNPs that can uniquely distinguish at least α% of the common unambiguous haplotypes in the block $r_i \ldots r_j$.

- Example: In the block $r_3 \ldots r_5$, we have the following haplotypes.
    - (1,1,2), (1,0,2), (1,1,0), (2,1,1), (2,1,0), (2,0,1)
    - All haplotypes are unambiguous and form two groups:
        - {(1,1,2), (1,0,2), (1,1,0)} and {(2,1,1), (2,1,0), (2,0,1)}
    - To distinguish 100% of these haplotypes, we need 1 tag SNP, that is, r_3.

Dynamic programming (I)

- Let S(i) = minimum number of tag SNPs to uniquely distinguish at least α% of the unambiguous haplotypes in $r_1 \ldots r_i$.
- Base case:
  - S(0) = 0
- Recursive case:
  - $S(i)=\min\{S(j-1)+f(r_j \ldots r_i) \mid 1 \leq j \leq i, block(r_j \ldots r_i)=1\}$

Dynamic programming (II)

- In practice, there may exist several block partitions that give the minimum number of tag SNPs.
- We want to minimize the number of blocks.
- Let C(i) = minimum number of blocks so that the number of tag SNPs is S(i).
- We have
  - $C(0) = 0;$
  - $C(i) = \min\{\ C(j-1) + 1 \mid 1 \leq j \leq i, \mathrm{block}(r_j \ldots r_i) = 1, S(i) = S(j-1) + f(r_j \ldots r_i)\}$

IdSelect (Carlson et al. Am. J. Hum. Genet. 2004)

- Aim: Among all SNPs exceeding a specified minor allele frequency (MAF) threshold, select a set of tag SNPs S such that
    - For every SNP i, there exists a SNP j in S so that their r^2 > a certain threshold th.

Algorithm IdSelect

- IdSelect is a greedy algorithm.

Algorithm **IdSelect**

1. Let S be the set of SNPs that are above the MAF threshold.
2. Let $T = \phi$
3. While S is not empty,
    - Select $s \in S$ which maximizes the size of the set $\{s' \in S \mid r^2(s,s') > th\}$.
    - $T = T \cup \{s\}$;
    - $S = S - \{s\} - \{s' \in S \mid r^2(s,s') > th\}$.

Disadvantage of IdSelect

- Since rare SNPs are harder to link with other SNPs, IdSelect tends to include many rare SNPs as the tag SNPs,
  - which is not nature.

Reference

- Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74: 106–120.
- Zhang, K., Deng, M., Chen, T., Waterman, M.S., and Sun, F. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci.* **99:** 7335–7339.

Association study

What is association study?

Case

(Disease sample)

```
ACGTACCGGTCACTCGCCCACTTCAGGCATA
ACGTGCCGGTCACTCACTCACTTCAGGCCTA
ACGTACAGGTCACTCGCTCACTTCAGGCATA
ACGTACCGGTCACACGCTCACTTTAGGAATA
AGGTACCGGTCACTCGCTCACTTCAGGCATA
ACCTACAGGTGACTCGCTCACTTCTGGCATG
ACGTACCGGTCACTCACTCTCTTCAGGCATG
ACGTACCGGTCAATCGCTCACTTCAGGCATA
```

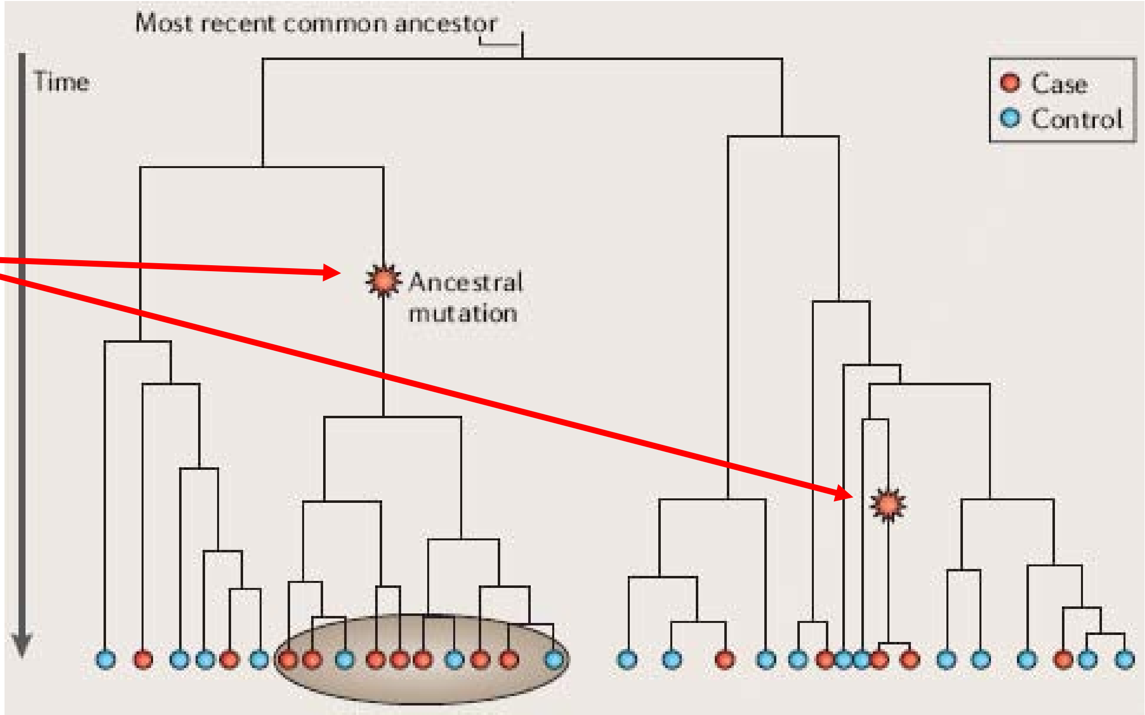
Control

(Normal sample)

```
ACCTACCGGTCACTCACTCACTTCAGGCCTA
ACGTACCGGACACTCACTCACTTTAGGCATA
GCGTACCGGTCACACACTCACTTCAGTCATA
ACGTACCGGTCACTCACTCACTTCAGGCCTA
ACCTGCCGGTGACTCACTCACTTTAGGCATG
ACGTACCGGTCACTCGCTCTCTTCAGGCATA
ACGTACAGGTCACTCACTCACTTCAGGCATA
ACGTACCGGTCACTCACTCACTTCAGGCATA
```

Rationale for association studies

- Case: individuals with disease
- Control: normal individuals



Risk enhancing mutation

Why association studies?

- Identify genetic variation which are correlated to disease
  - Such information help to identify
    - Drug target
    - Disease marker
- Understand how genetic variation affects the respond to pathogens or drugs.
- Understand the different among different races.
  - E.g. Why Asian has higher chance of getting Hapatitis B infection?

Single SNP association study

- Relative risk and odds ratio
- Logistic regression

Relative risk and odds ratio

- Let x and y be the two possible alleles in a loci.
- To check if Case is associate with allele x.
- Relative risk (RR) is [a/(a+b)] / [c/(c+d)].
- Odds ratio (OR) is ad/bc.

- The bigger the value of RR and OR, the SNP is more related to the disease.
- We use the Odds ratio to rank the SNPs.

Actual	Allele x	Allele y
Case	a	c
Control	b	d

Relative risk and odds ratio (II)

Actual	Allele G	Allele A
Case	6	2
Control	1	7

- RR = (6/7)/(2/9) =3.86
- OR = (6*7)/(2*1) = 21

- Since the values are big, this SNP is highly related to the disease.

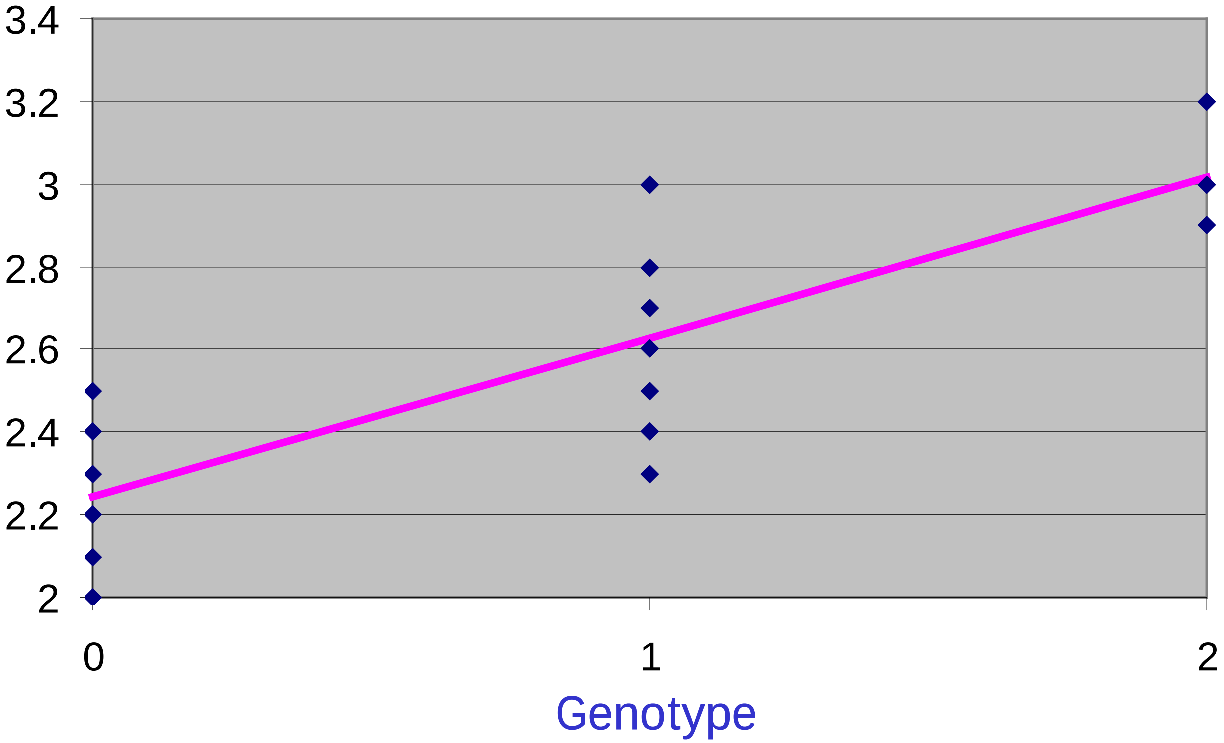
```
ACGTACCGGTCACTCGCCCACTTCAGGCATA
ACGTGCCGGTCACTCACTCACTTCAGGCCTA
ACGTACAGGTCACTCGCTCACTTCAGGCATA
ACGTACCGGTCACACGCTCACTTTAGGAATA
AGGTACCGGTCACTCGCTCACTTCAGGCATA
ACCTACAGGTGACTCGCTCACTTCTGGCATG
ACGTACCGGTCACTCACTCTCTTCAGGCATG
ACGTACCGGTCAATCGCTCACTTCAGGCATA
ACCTACCGGTCACTCACTCACTTCAGGCCTA
ACGTACCGGACACTCACTCACTTTAGGCATA
GCGTACCGGTCACACACTCACTTCAGTCATA
ACGTACCGGTCACTCACTCACTTCAGGCCTA
ACCTGCCGGTGACTCACTCACTTTAGGCATG
ACGTACCGGTCACTCGCTCTCTTCAGGCATA
ACGTACAGGTCACTCACTCACTTCAGGCATA
ACGTACCGGTCACTCACTCACTTCAGGCATA
```

Linear regression

Find the straight line which best fit the data!

Genotype	phenotypic score
0	2
0	2.1
0	2.4
0	2.3
0	2.2
0	2.5
1	2.4
1	2.5
1	2.6
1	3
1	2.7
1	2.8
1	2.3
2	2.9
2	3.2
2	3

$$y = 2.2415 + 0.3874x + \varepsilon$$



Formal definition

- Given (x_i, y_i), i=1, 2, …,n
  - where x_i is the genotype of the SNP and y_i is the phenotypic score.
- We would like to compute β_0 and β_1 such that
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; and
  - $\Sigma_{i=1..n}\, \varepsilon_i^2 = \Sigma_{i=1..n}(y_i-\beta_0-\beta_1 x_i)^2$ is minimized.

- $\Sigma\, \varepsilon_i^2$ is called the sum of squares error (SSE).
- Denote $\hat{y}_i = \beta_0 + \beta_1 x_i$

β_0 and β_1

- By partial differentiation with respect to β_0 and β_1, we can show that
  - $\beta_1 = \dfrac{\Sigma_{i=1..n} (x_i - \mu_x)(y_i - \mu_y)}{\Sigma_{i=1..n} (x_i - \mu_x)^2}$
  - $\beta_0 = \mu_y - \beta_1 \mu_x$.
- μ_x and μ_y are the means of x and y respectively.

Significant test for linear regression

- Mean sum of squares error (MSE) is $\Sigma_{i=1..n}(y_i- \hat{y}_i)^2$ / (n-2).
- Regression sum of squares (MSR) is $\Sigma_{i=1..n}(\hat{y}_i-\mu_y)^2$.
- MSR/MSE follows the F distribution.
- H_0: $\beta_1=0$, H_1: $\beta_1\neq0$
- We reject H_0 if MSR/MSE > $F_{1,n-2,0.95}$

Example

- $n=16$
- $\mu_y = 2.55625$
- $MSE = \Sigma_{i=1..n}(y_i - \hat{y}_i)^2 / (n-2) = 0.040931$
- $MSR = \Sigma_{i=1..n}(\hat{y}_i - \mu_y)^2 = 1.266338$
- $MSR/MSE = 30.03819 > F_{1,14,0.95} = 4.6$
- We reject H_0: $\beta_1=0$.

Genotype	phenotypic score
0	2
0	2.1
0	2.4
0	2.3
0	2.2
0	2.5
1	2.4
1	2.5
1	2.6
1	3
1	2.7
1	2.8
1	2.3
2	2.9
2	3.2
2	3

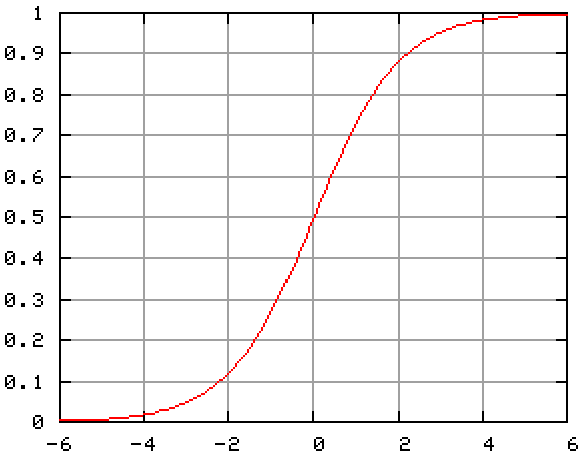
Regression when Y is binary

- For case and control study,
  - Y usually has only 2 values: 0 and 1.

- In this case, we would like to fit
  - $\Pr(D) = \alpha + \beta X + \varepsilon$.
- However, such function is difficult to fit since Pr(D) is in a narrow range [0,1].

Sigmoid function (standard logistic function)

- $F(t) = 1 / (1 + e^{-t})$
  - $t = 0$ → $F(t) = 0.5$
  - $t = +\infty$ → $F(t) = 1$
  - $t = -\infty$ → $F(t) = 0$



- We try to fit
  - $Pr(D) = 1 / (1 + e^{-(\alpha+\beta X)})$
  - Hence, $Pr(D)/(1-Pr(D)) = e^{-(\alpha+\beta X)}$

Logistic regression

$$\log(\frac{\Pr(D)}{1-\Pr(D)}) = \alpha + \beta X$$

- D is the disease status
- X has 3 values:
  - 2 if the genotype is xx;
  - 1 if the genotype is xy; and
  - 0 if the genotype is yy.
- Test if β=0