

Computational Human Genetics

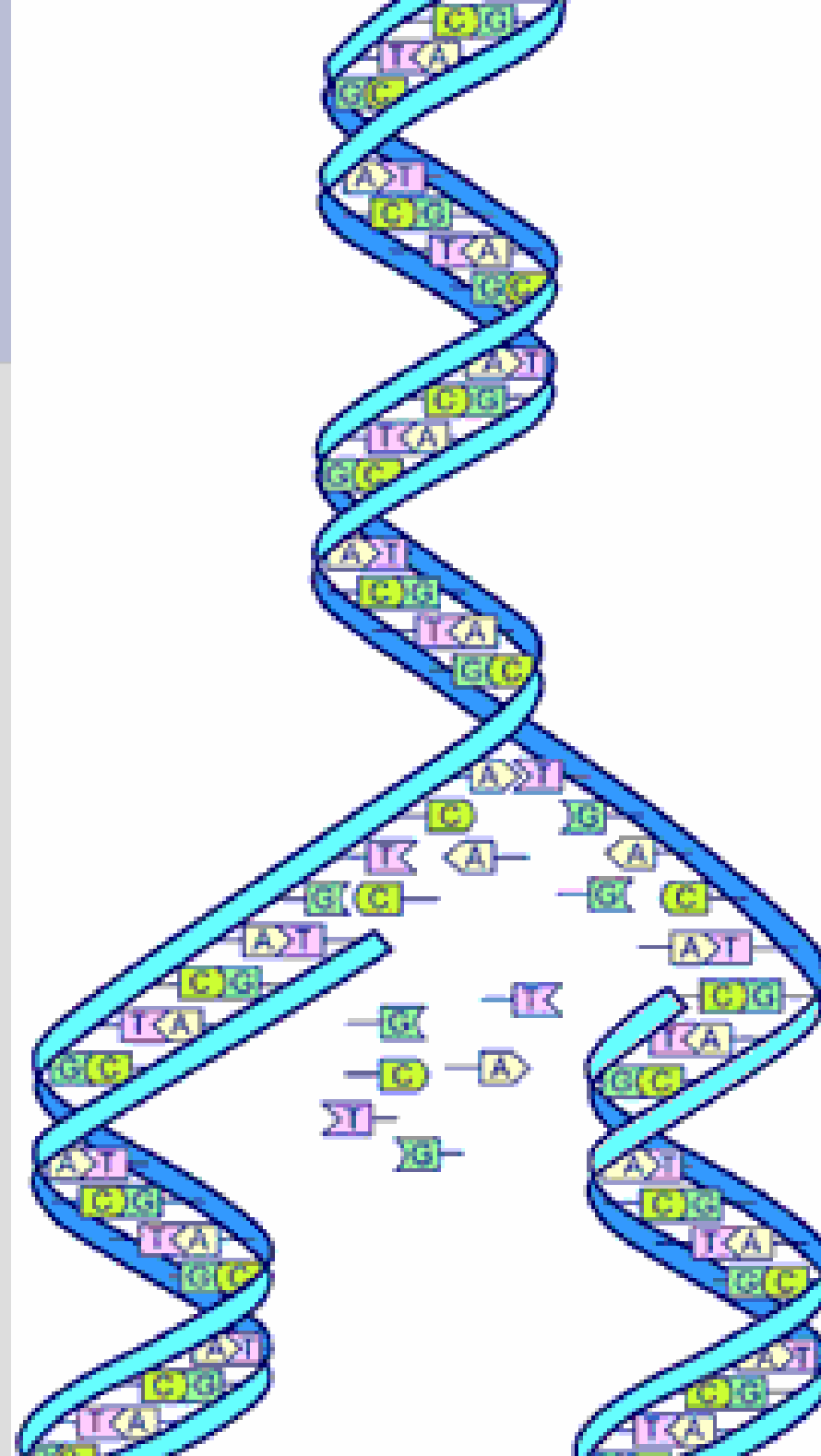
Itsik Pe'er

Department of Computer Science
Columbia University

Fall 2006

Reminder

- Cells
- Genes & DNA
- Genetics



Administration

- Moved to 337MUDD
- Send me email with:
 - Background in biology
 - Background in CS
 - Background in statistics
- Extra credit exercises
- TA wanted!!

Meeting #2

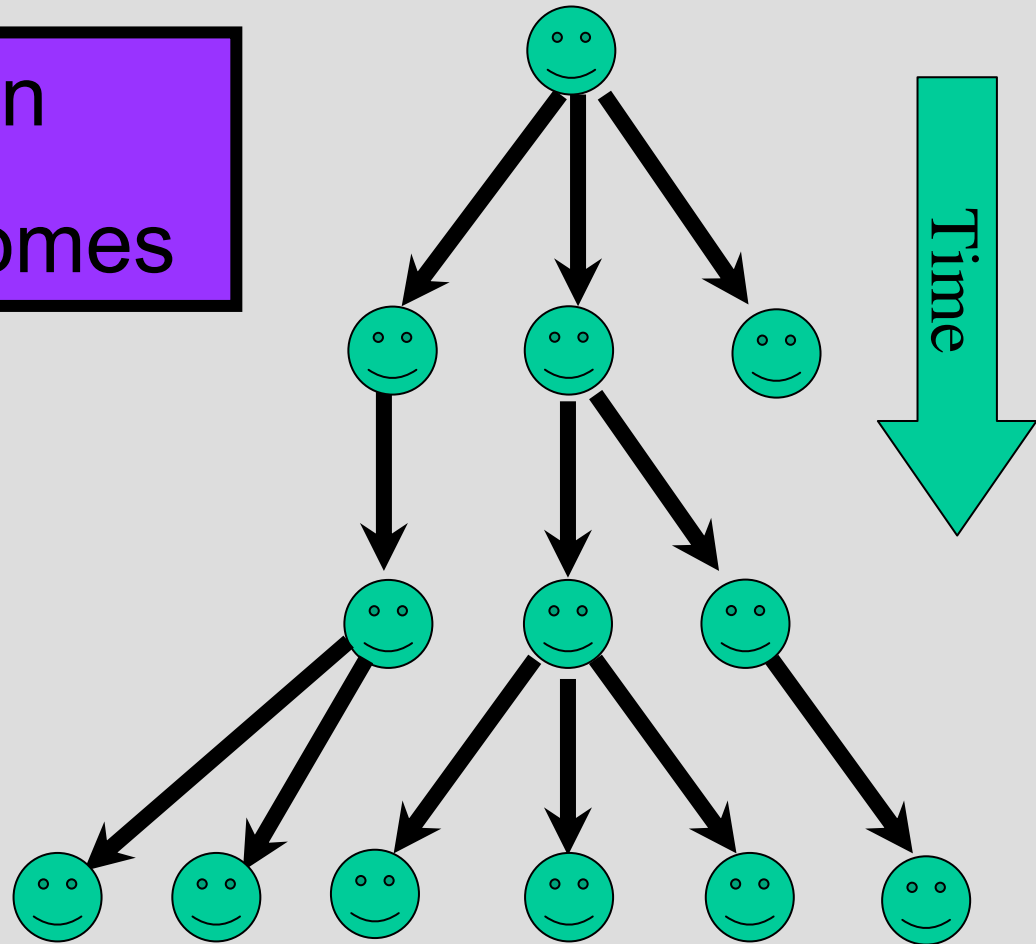
Genetics of a single site

Genetics of a Single Site

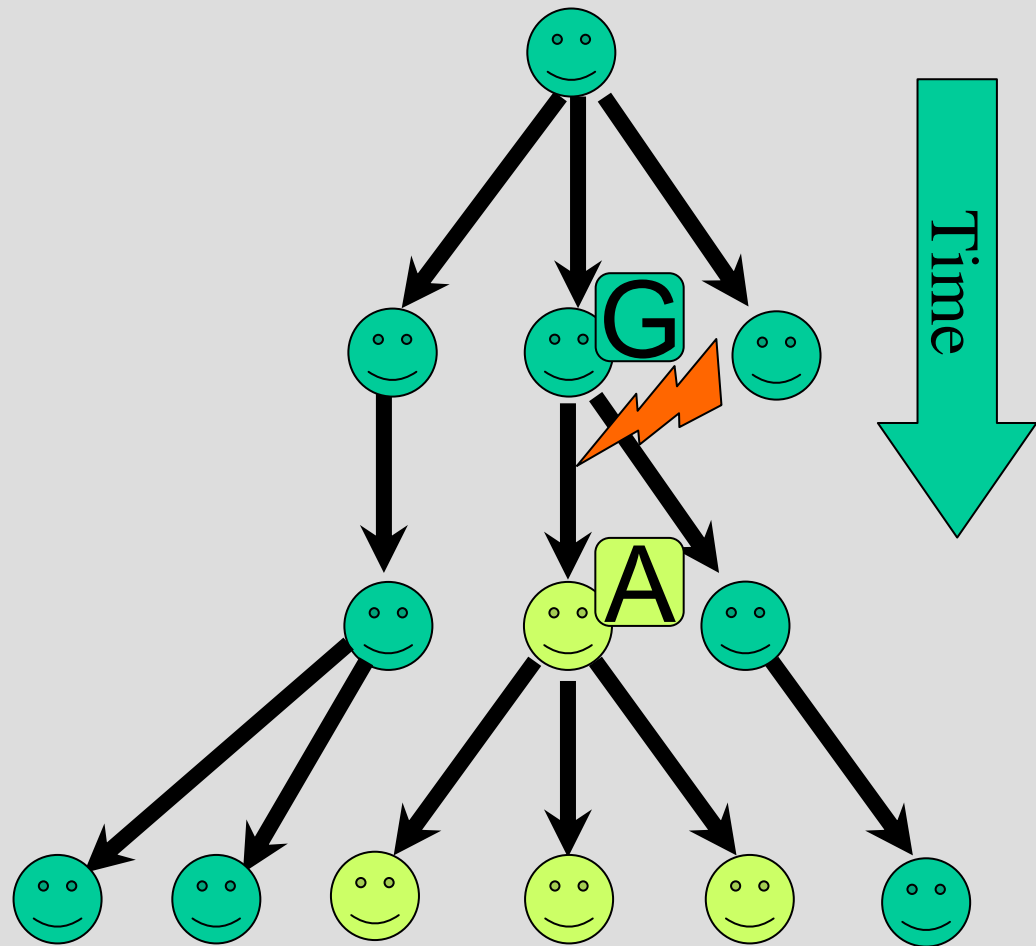
- Nordburg, M., Coalescent theory. Chapter 7 *in* D. J. Balding, M. J. Bishop, and C. Cannings, eds. **Handbook of statistical genetics.**
- Wakeley, J., Coalescent theory. Chapters 2-4
- Gusfield, D., Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology Chapter 17.5

The Divergence History of a Site

- No recombination
- Single chromosomes



Divergence History & Mutation

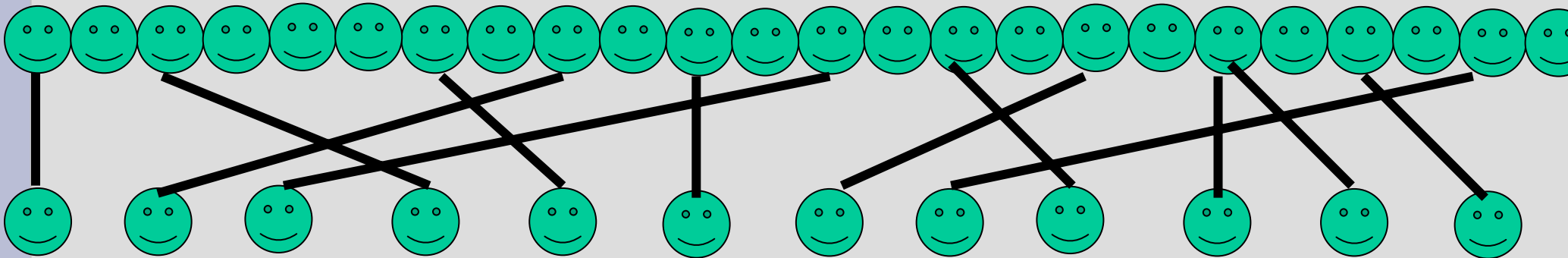


Genetics of a Single Site

- Coalescent models of a single site
- Coalescence and mutation
- Trees for several sites

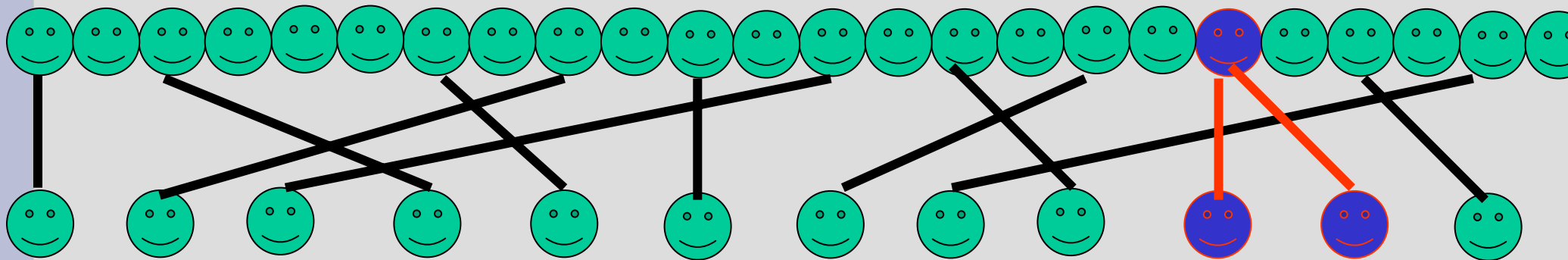
Back in Time

- Each offspring randomly chooses a parent
- Occasional *coalescent* events:
Two offsprings choose the same parent

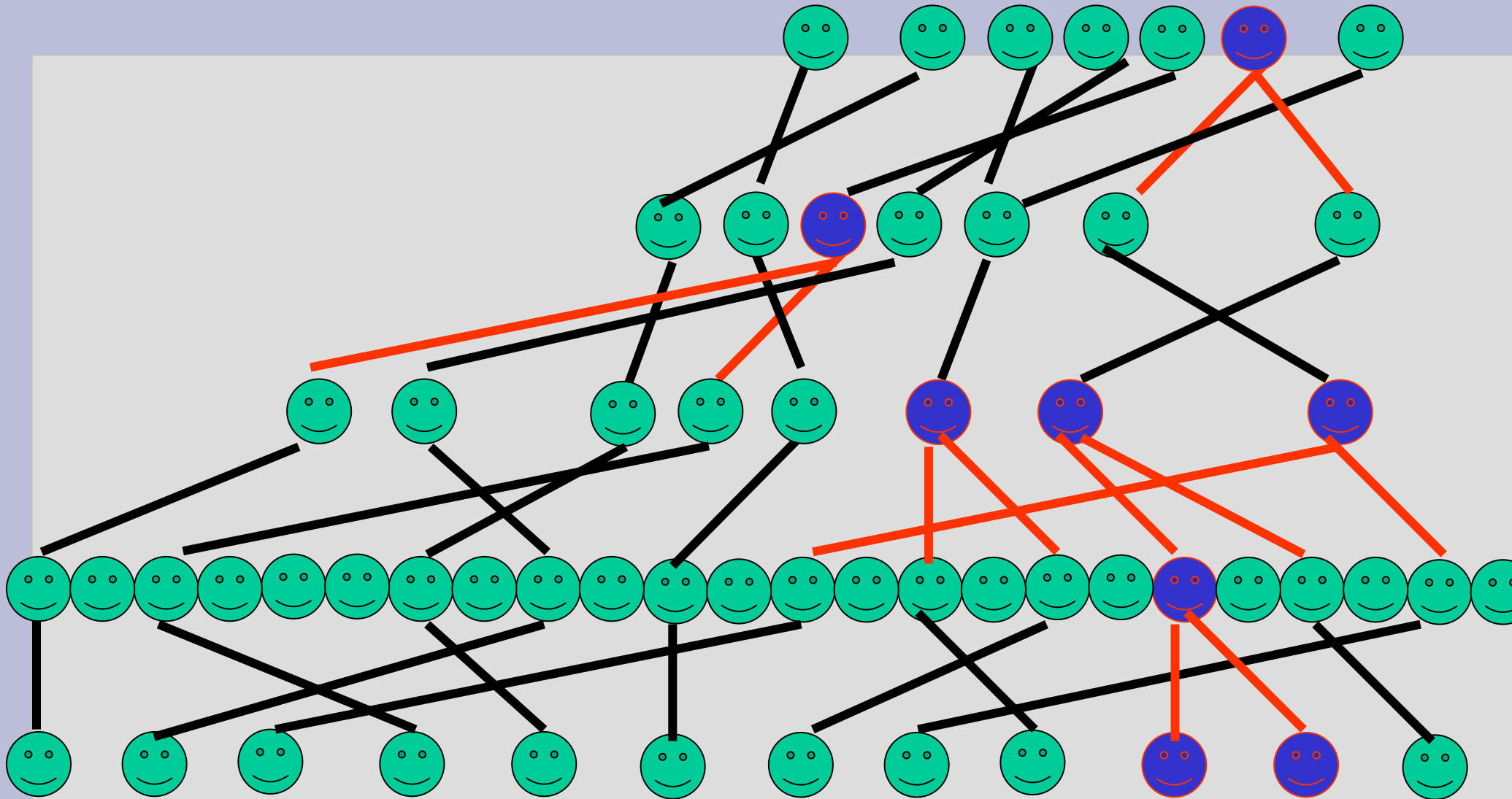


Probability of Coalescence

- Notation:
 - k : the number of individuals we are tracing
 - N_e : the *effective population size* .
- Two specific individuals coalesce with probability $1/N_e$.
- Expected number of events: $\binom{k}{2}/N_e$



Recursive Coalescence



Time to Coalescence

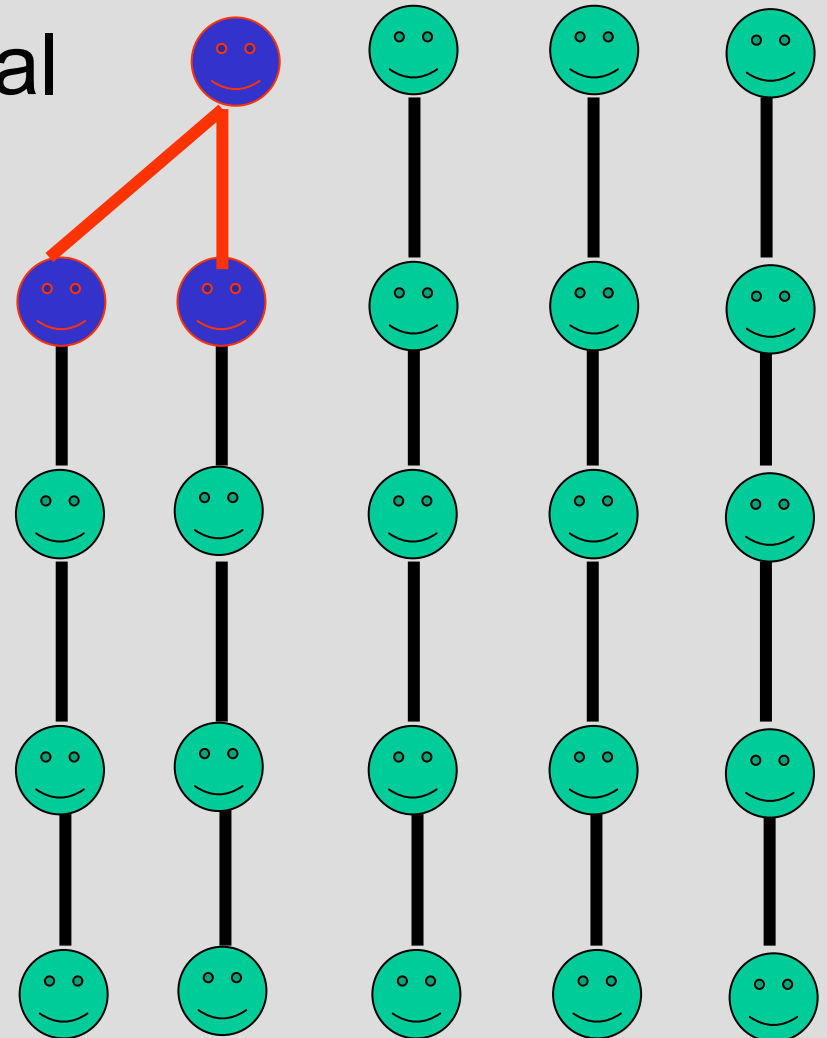
- When $k^2 \ll N_e$:
“No coalescence” is typical

- T_k : time to coalescence of k into $k-1$ individuals

- $T_k \sim \text{Geometric}(p = \binom{k}{2} / N_e)$

- $\text{Exp}(T_k) = 2N_e / [k(k-1)]$

- $\text{Var}(T_k) = (1-p) / p^2$

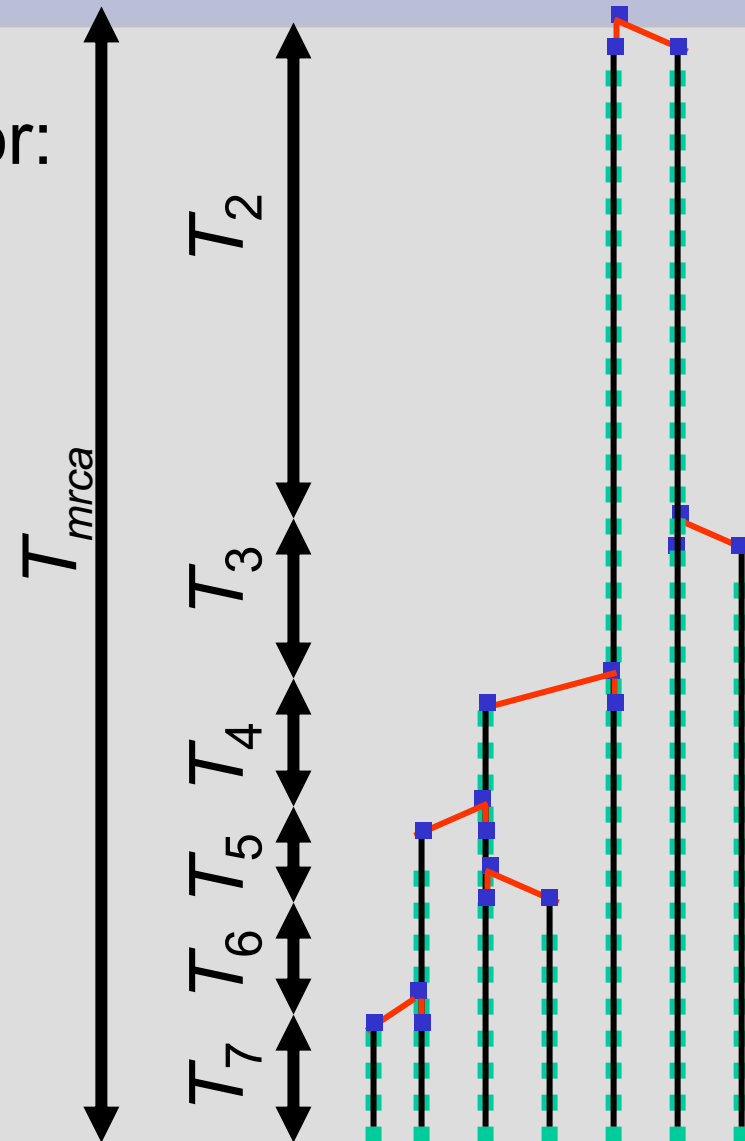


Height of a Coalescence Tree

Time to **m**ost **r**ecent **c**ommon **a**ncestor:

$$Exp(T_{mrca}) = \sum_{i=2}^k Exp(T_i) =$$

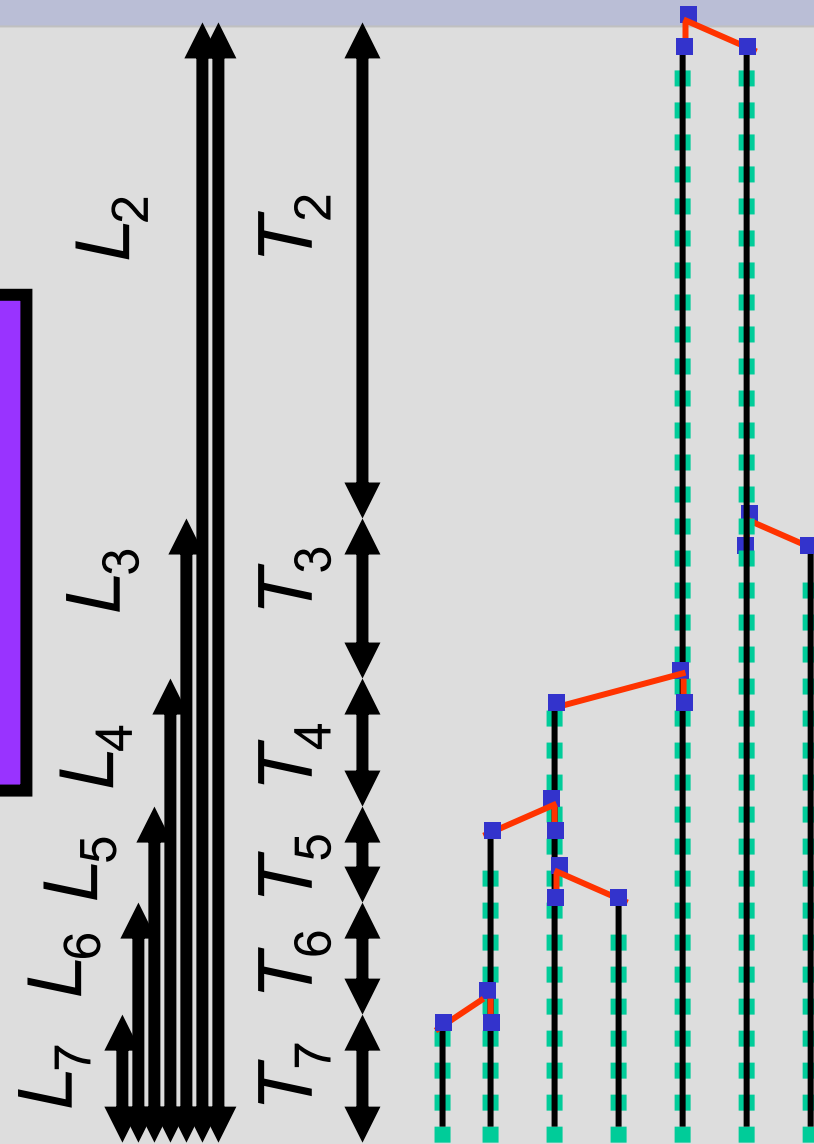
- $T_{mrca} \rightarrow 2N_e$ for large k
- Most of T_{mrca} : at the top



Length of a Coalescence Tree

$$\text{Exp}(L_{total}) = \sum_{i=2}^k i \text{Exp}(T_i) =$$

- $L_{total} \rightarrow \infty$ with k
- Most of L_{total} : at the bottom



Continuous Version

- Unit conversion:
1 *coalescent time* = N_e generations
- $T_k \sim \text{Exponential}(2/[k(k-1)])$
- Allows derivation of distributions for
 T_{mrca} , L_{total}

Model Assumptions

- No recombination:
True for single bases, approx. for short regions

Model Assumptions

- No recombination.
- Constant population size:
False, but:
 - may be fine for most human history
 - Can generalize for variable size.

Unit conversion is

$$t[\textit{generations}] \rightarrow \int_{\tau=0}^t \frac{1}{N(\tau)} d\tau[\textit{coalescent time}]$$

Model Assumptions

- No recombination.
- Constant population size.
- Single chromosomes:
True only for asexual reproduction.
Otherwise: another factor of 2.

$$\text{Exp}(L_{total}) = 4N_e(\ln k + O(1))$$

$$\text{Exp}(T_{mrca}) = 4N_e(1 - 1/k)$$

Model Assumptions

- No recombination.
- Constant population size.
- Single chromosomes.
- Independent, uniform parent selection:
 - False, due to gender
 - False, due to socio-demographic factorsHandled by using $2N_e$ rather than $2N$

Model Assumptions

- No recombination.
- Constant population size.
- Single chromosomes.
- Independent, uniform parent selection.
- **No selective variation**

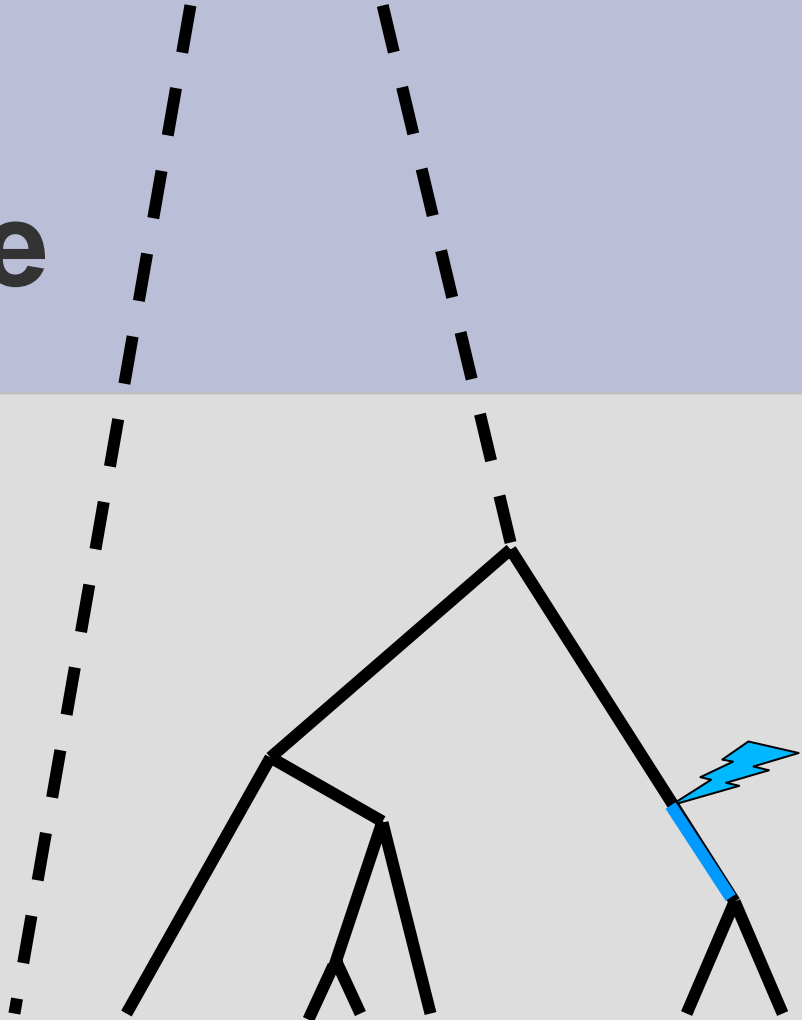
Wright-Fisher model

Genetics of a Single Site

- Coalescent models of a single site
- **Coalescence and mutation**
- Trees for several sites

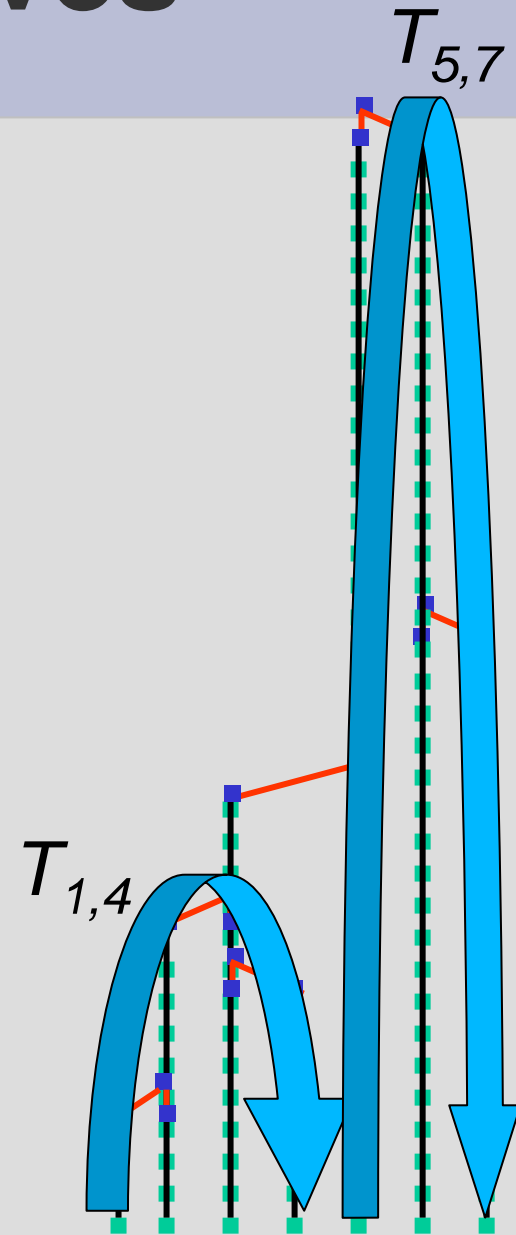
A Mutation on a Tree

- Derived allele is present a continuous subtree
- Ancestral allele can be identified by an outgroup
- Time along the branch doesn't matter
- Assumption:
No recurrent/reverse mutation
(*infinite site model*)



Distance Between Leaves

$$\text{Exp} (T_{a,b}) = 2T_2 = 4N_e$$



#Mutations on a Tree

- Depends on mutation rate, branch length
- Notation:
 - μ - mutations per generation per site
 - θ - *heterozygosity* :
#changes between two chromosomes.
 - $\hat{\theta}$ average heterozygosity across all pairs
- $\hat{\theta} \sim \text{Poisson}(4N_e\mu)$ [*distribution over loci*]
- #Polymorphic sites $\sim \text{Poisson}(L_{total}\mu)$

Some More Properties

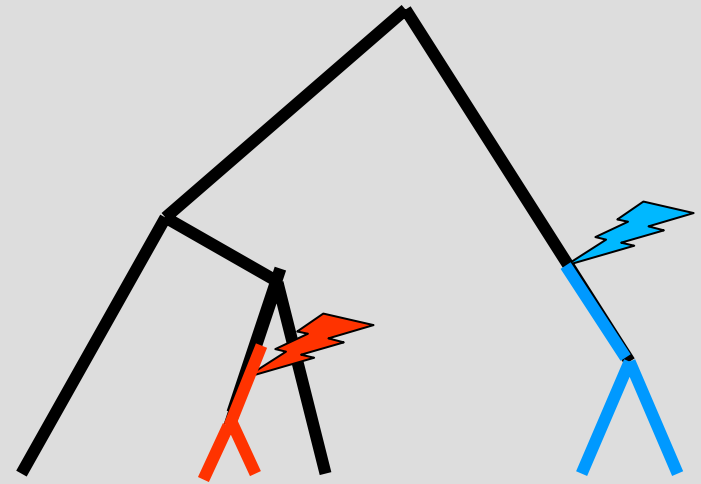
- Total length of branches with j descendants is $\tau_j = 4N_e/j$
- Fraction of polymorphic sites with j mutants is τ_j / L_{total}
- If a site is a difference between two samples, its frequency in additional $2k+1$ samples is uniformly distributed across frequencies

Genetics of a Single Site

- Coalescent models of a single site
- Coalescence and mutation
- **Trees for several sites**

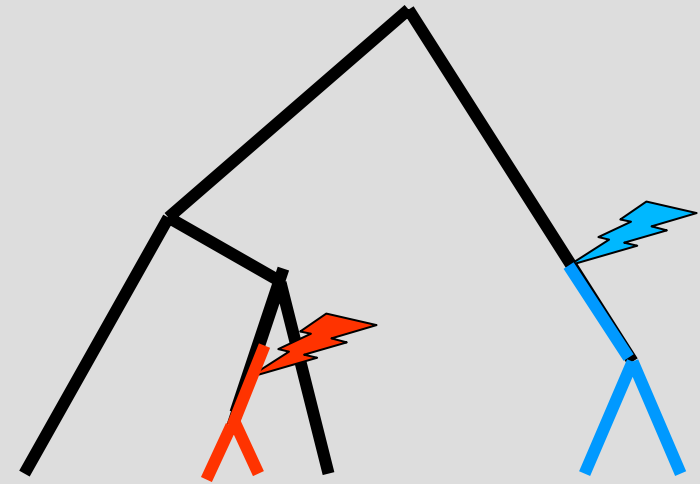
Two Mutations on a Tree

- Subtrees are either disjoint

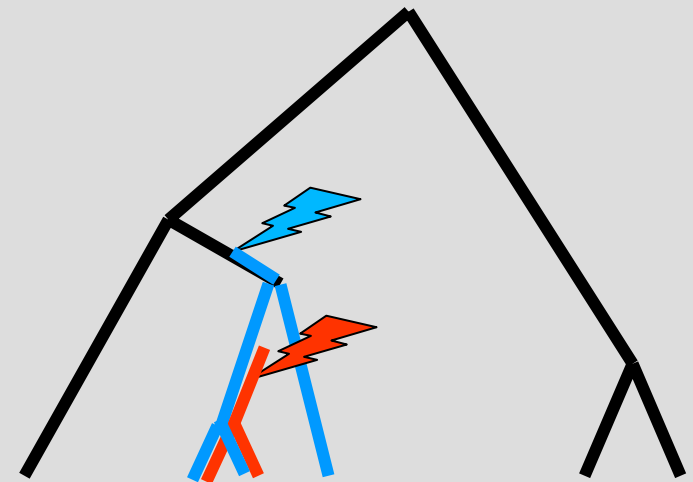


Two Mutations on a Tree

- Subtrees are either disjoint



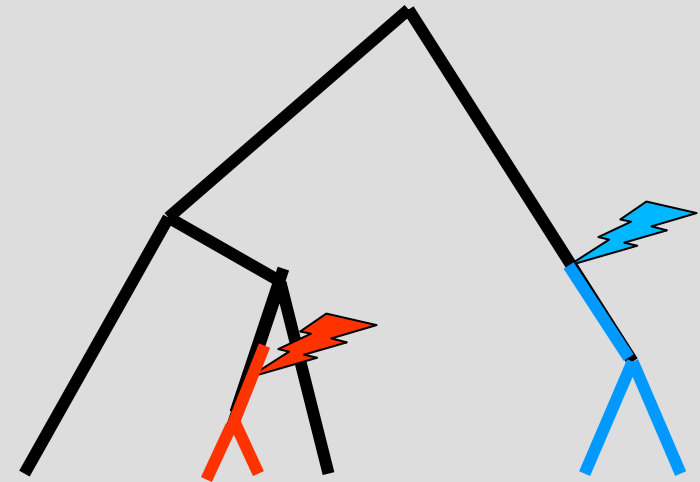
or contained in one another



Two Mutations on a Tree

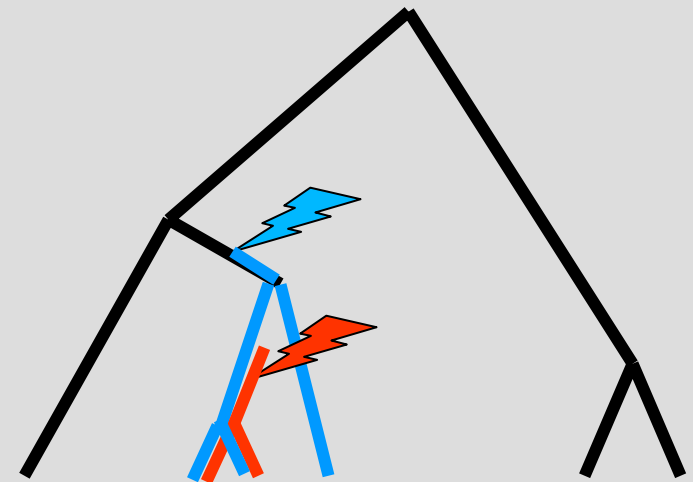
- Subtrees are either disjoint

Haplotypes: 00
01
10



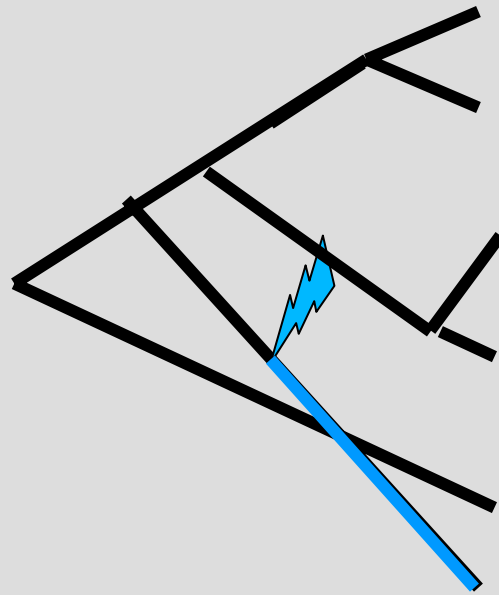
or contained in one another

Haplotypes: 00
01
11



An Unknown Tree

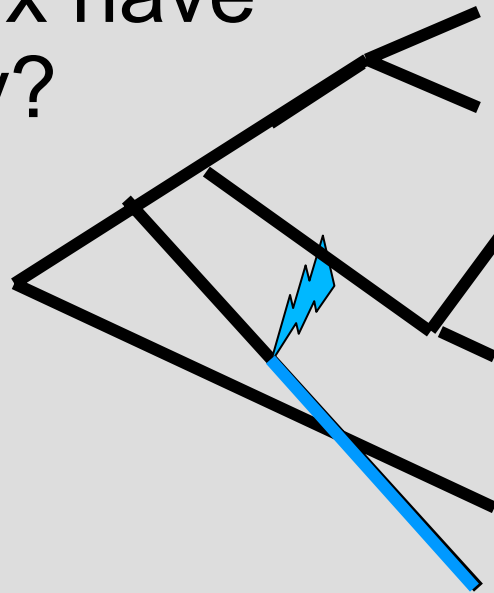
- Typical data: matrix M of *haplotypes*, w/o tree
- A tree + mapping of sites → branches, individuals → leaves, is a *phylogeny*



	sites					
Individuals	1	0	0	0	1	1
	1	0	0	0	1	0
	1	1	1	0	0	0
	1	0	1	0	0	0
	0	0	0	0	0	0
	1	0	0	1	0	0

Perfect Phylogeny

- (*directed*) perfect phylogeny:
each site changes once (only 0→1)
- **Problem:**
Does an input matrix have
a perfect phylogeny?



	sites					
Individuals	1	0	0	0	1	1
	1	0	0	0	1	0
	1	1	1	0	0	0
	1	0	1	0	0	0
	0	0	0	0	0	0
	1	0	0	1	0	0
	1	0	0	0	0	0

Forbidden Submatrices

- **Thm:**

A binary matrix has a directed perfect phylogeny
iff it has no minor

01

10

11

Forbidden Submatrices

- **Thm:**

A binary matrix has a perfect phylogeny
iff it has no minor

00

01

10

11

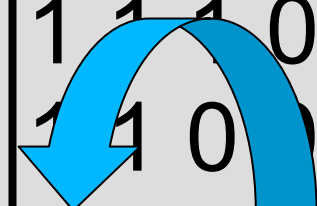
- “4-gamete rule” of non-recombinant
haplotypes

Perfect Phylogeny - Algorithm

- Sort columns, delete duplicates
- For each (i, j) s.t. $M_{i,j} = 1$:
 - $L(i, j) \leftarrow$ closest 1 on the left: k s.t. $k < j$, $M_{i,k} = 1$
- For each column
 - $L(j) \leftarrow \max(L(i, j))$
- Perfect Phylogeny iff $\forall i, j: L(i, j) = L(j)$

1	0	0	0	1	1
1	0	0	0	1	0
1	1	1	0	0	0
1	0	1	0	0	0
0	0	0	0	0	0
1	0	0	1	0	0

1	1	1	0	0	0
1	1	0	0	0	0
1	0	0	1	1	0
1	0	0	1	0	0
0	0	0	0	0	0
1	0	0	0	0	1



Diploids

- Alleles for diploids may be 00/01/10/11
- Technology:
Reads signals on “0” and “1” channels.
- *Homozygous* 0 or 1 states are unambiguous
- Cannot distinguish “01” from “10”:
Ambiguity for *heterozygotes*

Diploid Perfect Phylogeny

- Real input:

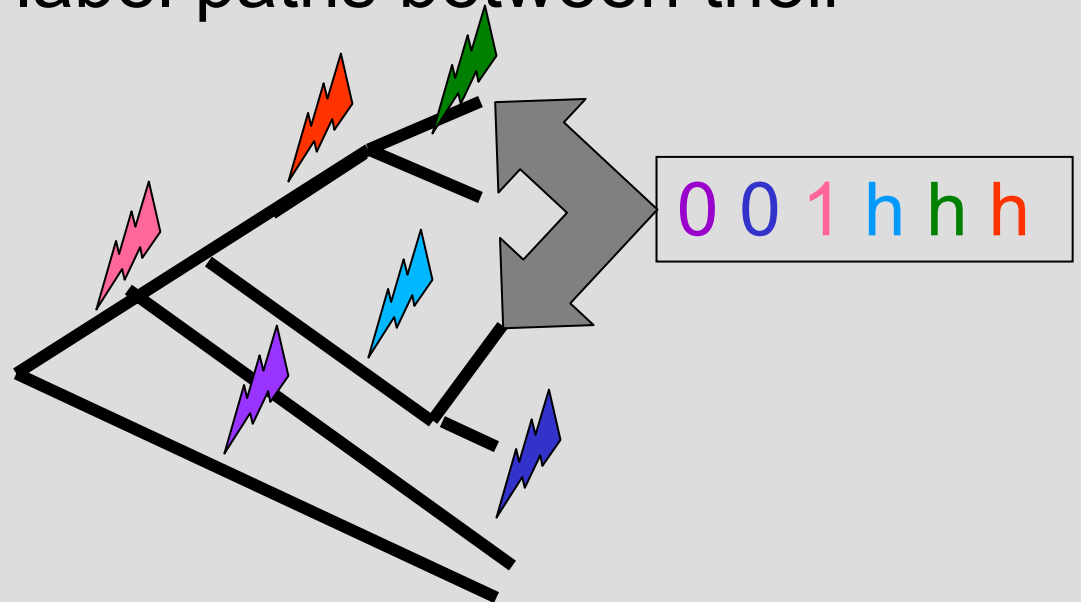
0	0	0	0	0	0
h	0	0	0	h	0
1	0	0	h	h	h
1	0	0	h	h	0
1	h	h	0	0	0
h	h	h	0	0	0

- Forbidden minor

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

Perfect Phylogeny Haplotyping

- Also linear time, but more involved
- Important idea:
 - Heterozygous sites label paths between their leaves



Model Assumptions

- Infinite site model:
 - No recurrent mutation
 - No reverse mutation
 - True when mutation is rare
- No recombination
 - True in short segments – more next week
- No errors in the data
 - Never true

Summary

- Coalescent models of a single site
 - Coalescent process implies height and length of tree
- Coalescence and mutation
 - Inferences regarding frequencies of polymorphisms in a tree
- Trees for several sites
 - Binary perfect phylogenies

Extra Credit if $\exists TA$

1. When you observe sequence from k chromosomes, what is the contribution of derived alleles present in j chromosomes to overall average heterozygosity?
2. In Figure 3 of http://www.hapmap.org/downloads/presentations/Nature_HapMap_phase1.pdf authors report allele frequencies in 90 individuals after discovery in 10x sequencing with/without 16 additional sequenced individuals. Is that what you expect? Explain.

Project Suggestion

- Create and analyze a perfect phylogeny map of the human genome
 - Use advanced algorithms (see work of Eskin & Halperin, or Gusfield and colleagues)
 - Allow errors
 - Run on the entire genome as in http://hapmap.org/downloads/phasing/2006-07_phaseII/phased/
 - Find regions of perfect phylogeny
 - Report connections between perfect phylogeny to genomic features (genes, gene families, repeats, chromosomes)