

Computational Human Genetics

Itsik Pe'er

Department of Computer Science
Columbia University

Fall 2006

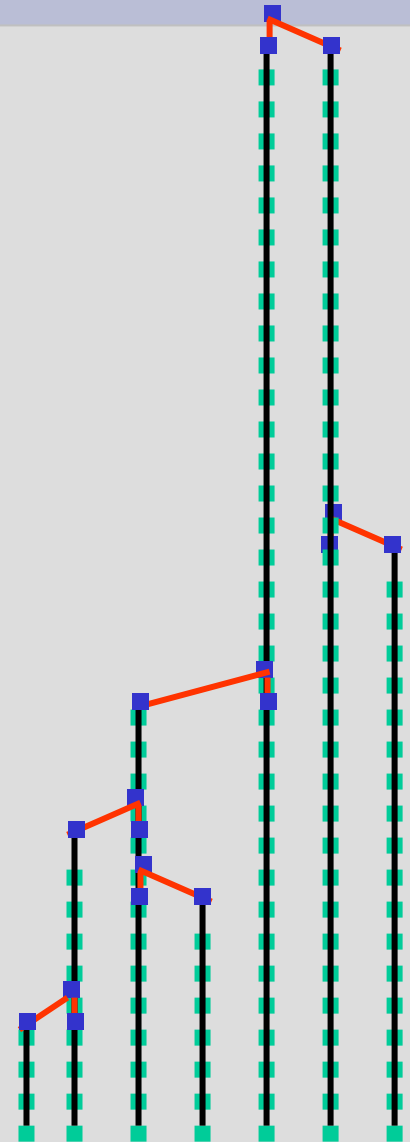
Administration

- Welcome Nalini Kartha, TA
- Classes on Oct 11th, Nov 29th

Reminder

- Coalescent models:
 - of a single site
 - + mutation to create single nucleotide polymorphisms (SNPs)
 - + several sites

What happens on
the next base over?



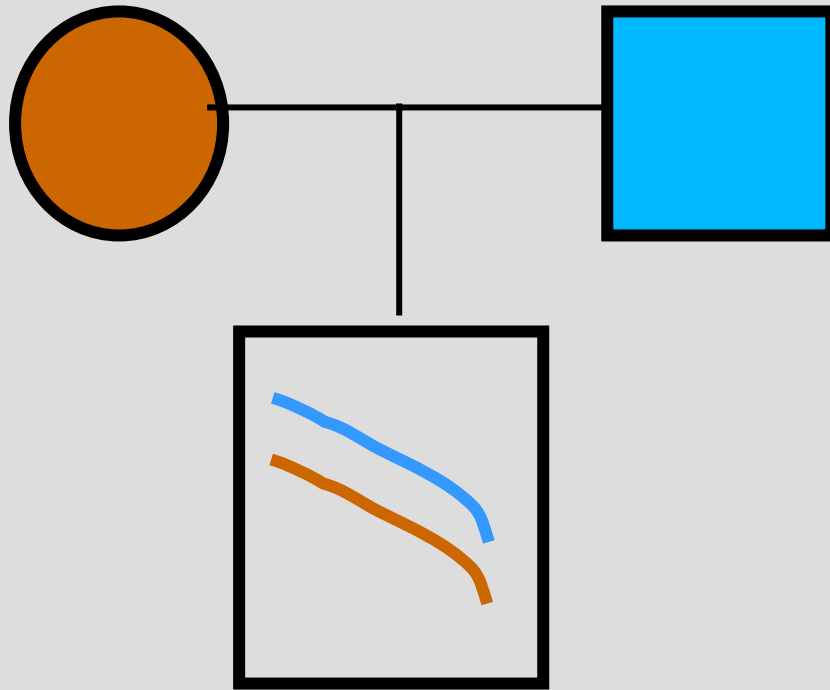
Meeting #3

Coalescence with Recombination

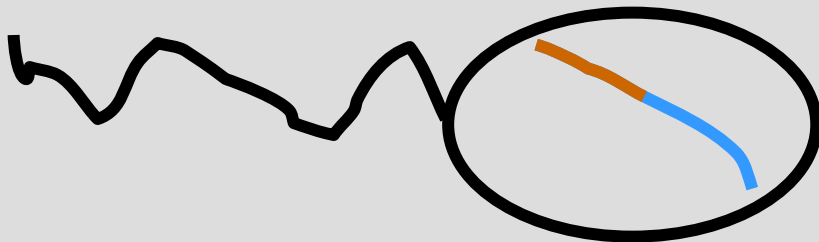
Coalescence with Recombination

- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

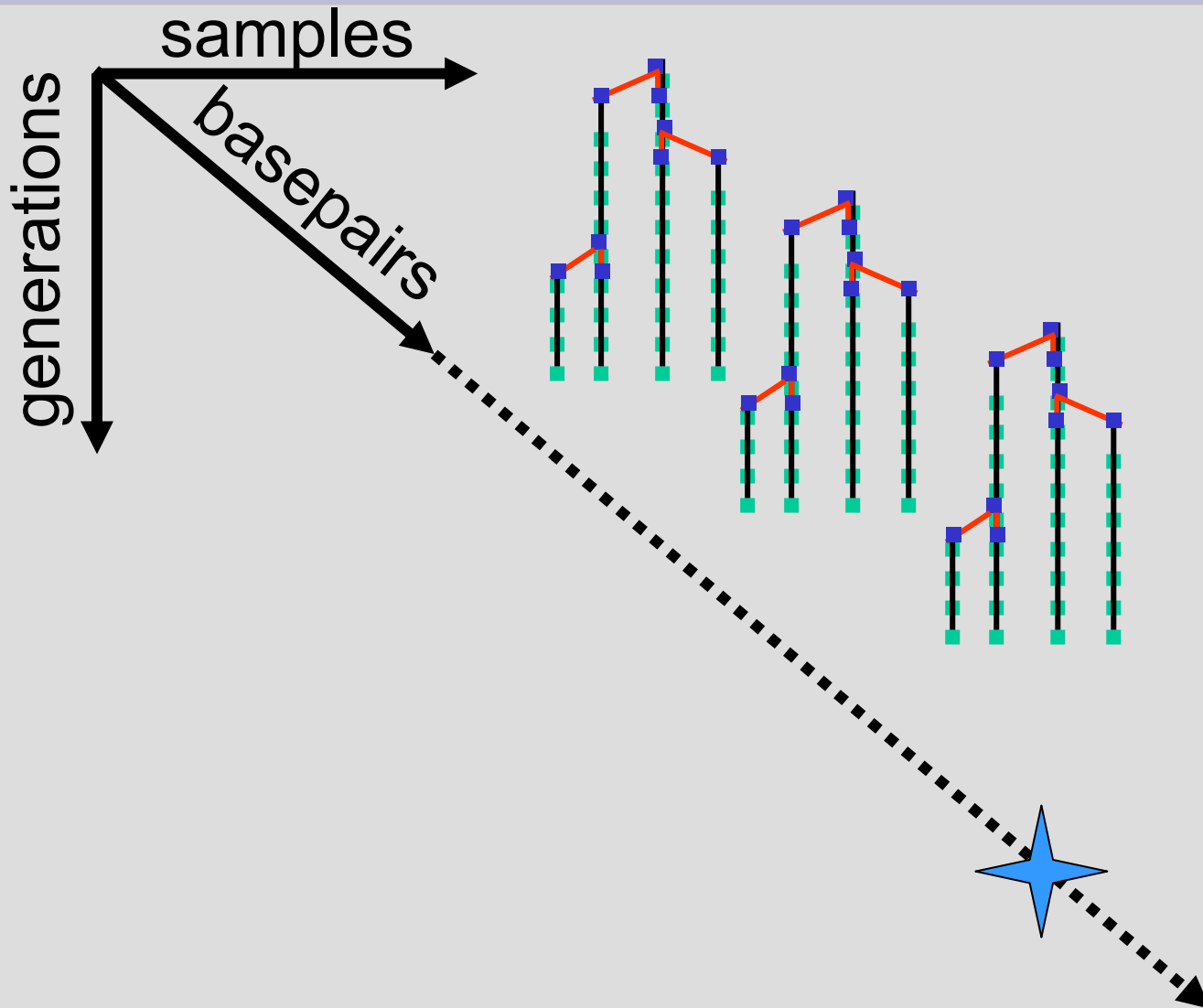
Recombination: Choosing between Mom & Dad



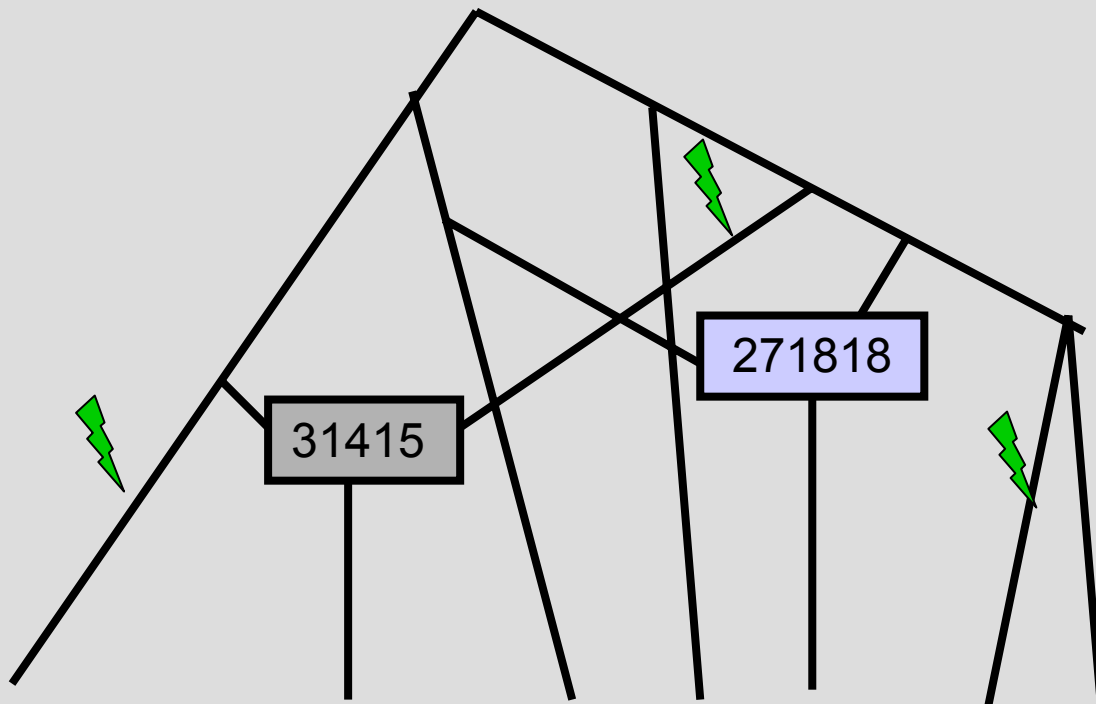
- At each site, one parent's DNA is transmitted on
- This changes at recombination sites



Recombination Rewires the Tree



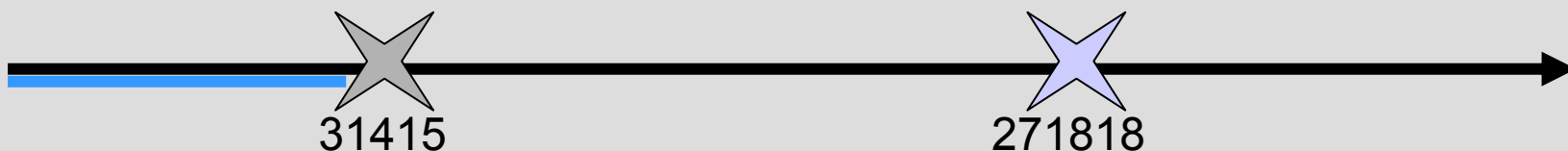
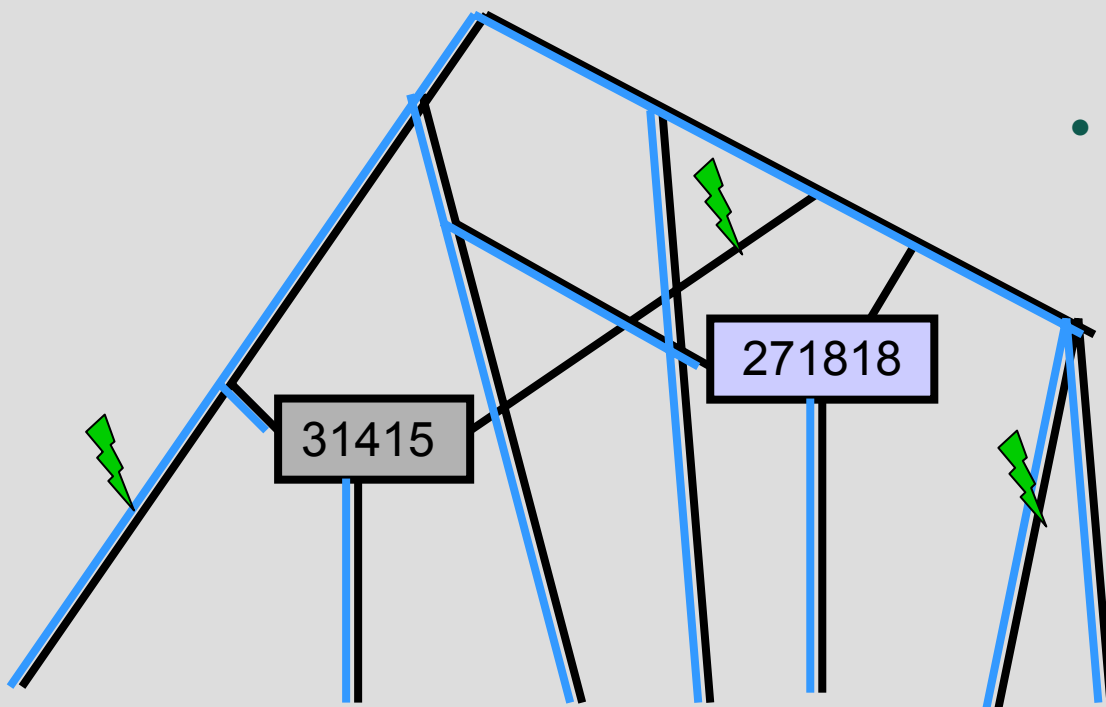
Ancestral Recombination Graph (ARG)



- Directed acyclic graph
- In-degree ≤ 2
- If in-degree = 2:
 - Out-degree = 1
 - Numeric label
- Mutations on branches

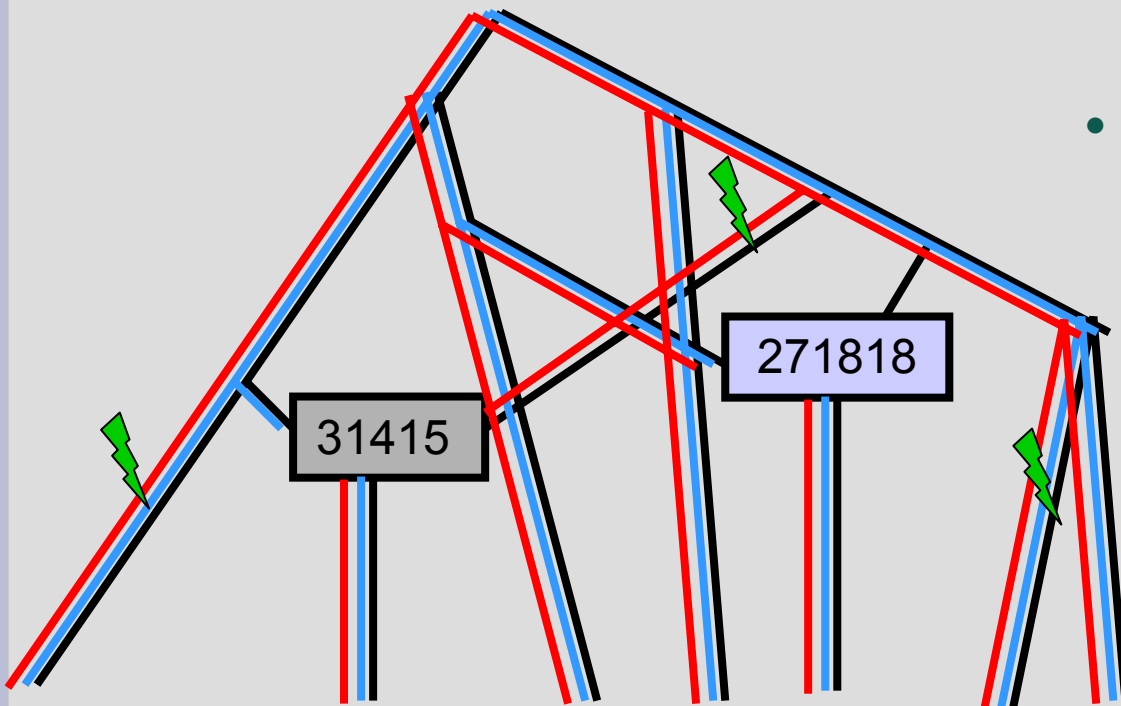
Induced Trees

- Segment the genome by node labels
- Derive a tree/segment:
 - Start from leaves
 - Turn \leftarrow/\rightarrow by label



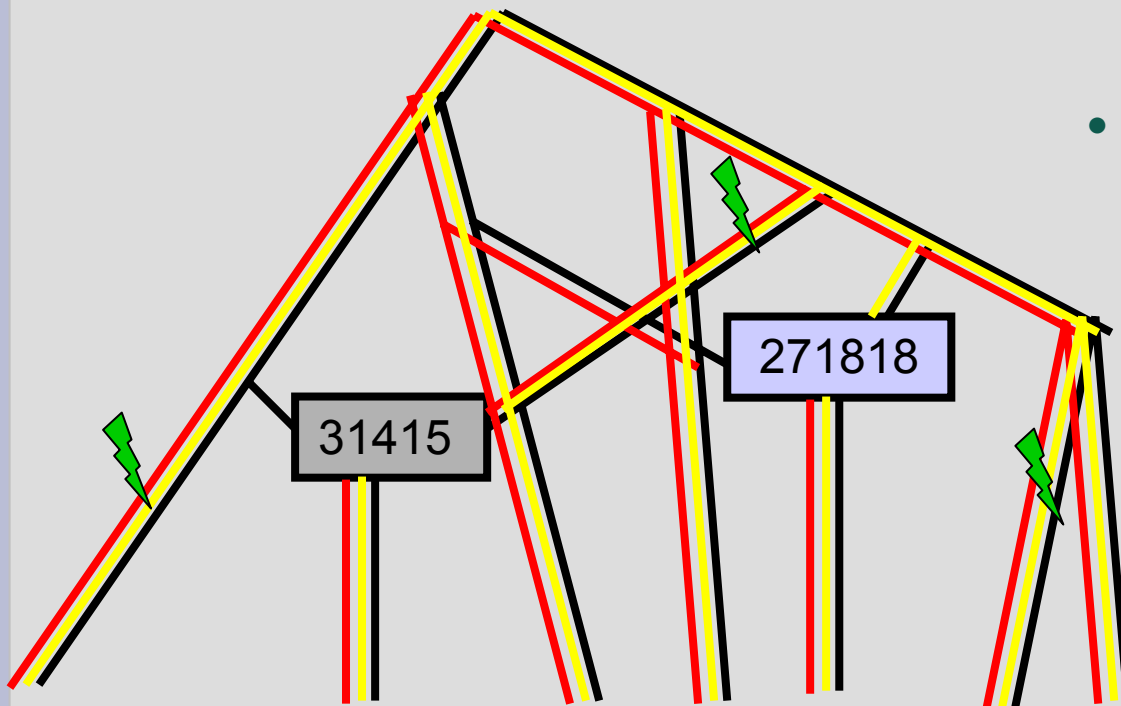
Induced Trees

- Segment the genome by node labels
- Derive a tree/segment:
 - Start from leaves
 - Turn \leftarrow/\rightarrow by label



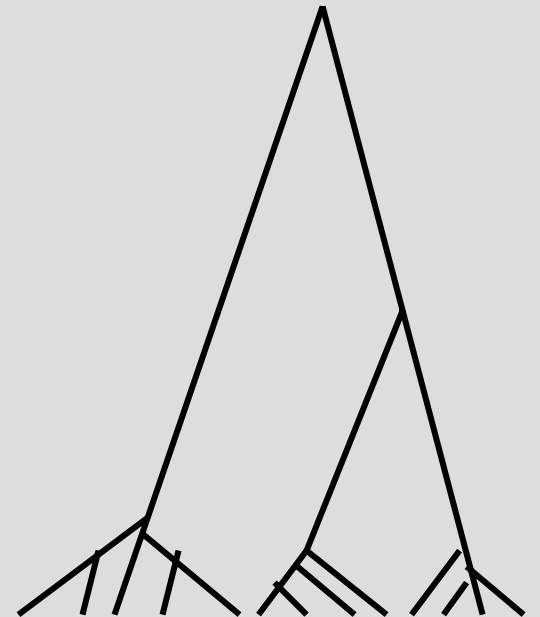
Induced Trees

- Segment the genome by node labels
- Derive a tree/segment:
 - Start from leaves
 - Turn \leftarrow/\rightarrow by label



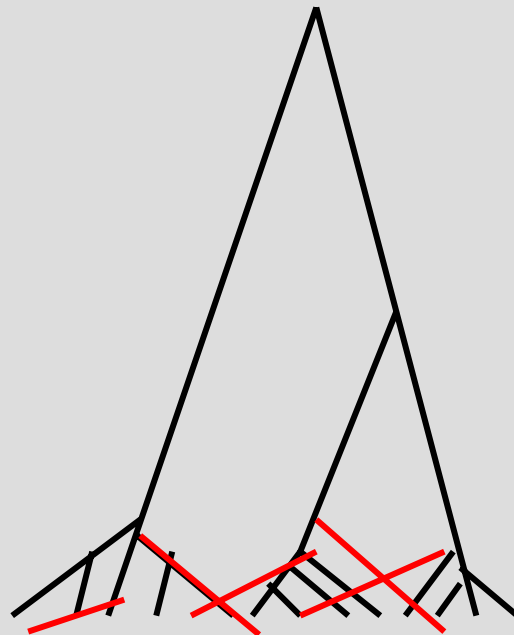
Recombination Rates

- Reminder (mutation rates):
 - $\theta = 4N\mu$ (between two sequences/per generation)
- Same logic for recombination:
 - $\rho = 4Nr$
- Most recombination is recent
- For a typical pair of sequences:
Most recombination is ancient



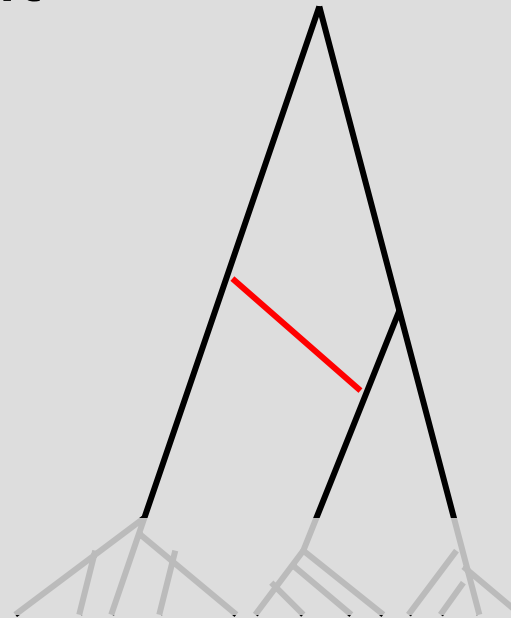
Implications to Haplotypes

- Perfect phylogeny –
violated by many rare recombinants



Implications to Haplotypes

- Perfect phylogeny – violated by many rare recombinants
- Usually only few common haplotypes, potentially recombinant



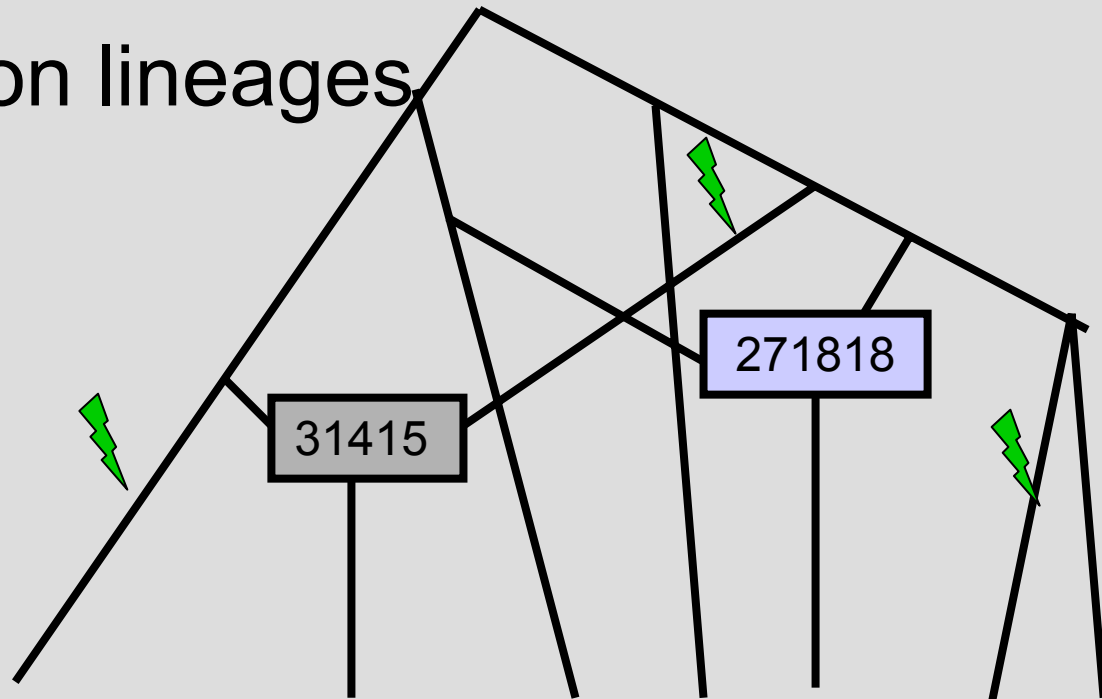
Coalescence with Recombination

- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Simulating Data: Back in Time

1. Generate ARG topology:
 - Start with k contemporary samples
 - Randomize links to previous generation
 - Continue recursively

2. Sprinkle mutations on lineages

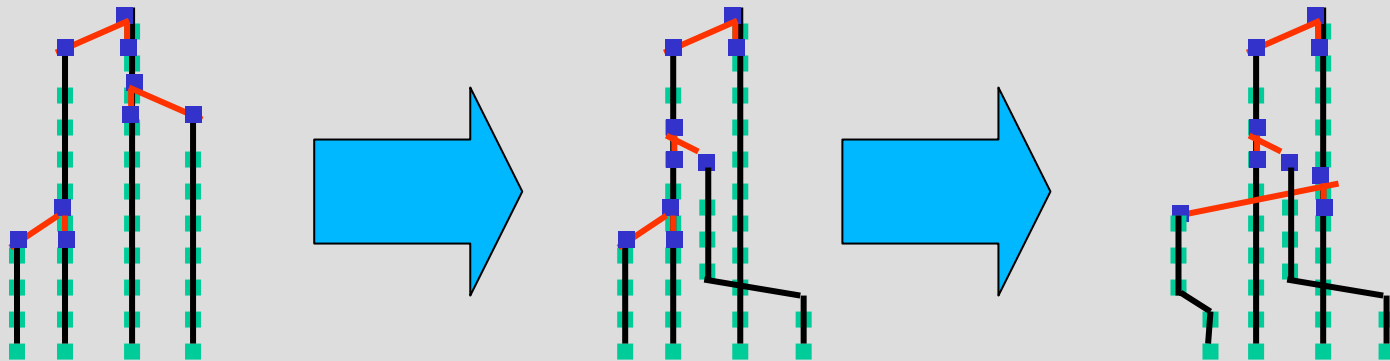


Simulating Data: Back in Time

1. Generate ARG topology:
 - Start with k contemporary samples
 - Randomize links to previous generation
 - Continue recursively
2. Sprinkle mutations on lineages
 - Faster implementation:
 - Randomize time till last event
 - Caveat:
 - Space requirement:
 $O(\text{ARG width}) = O(LrT_{tot})$

Simulating Data: Along the Genome

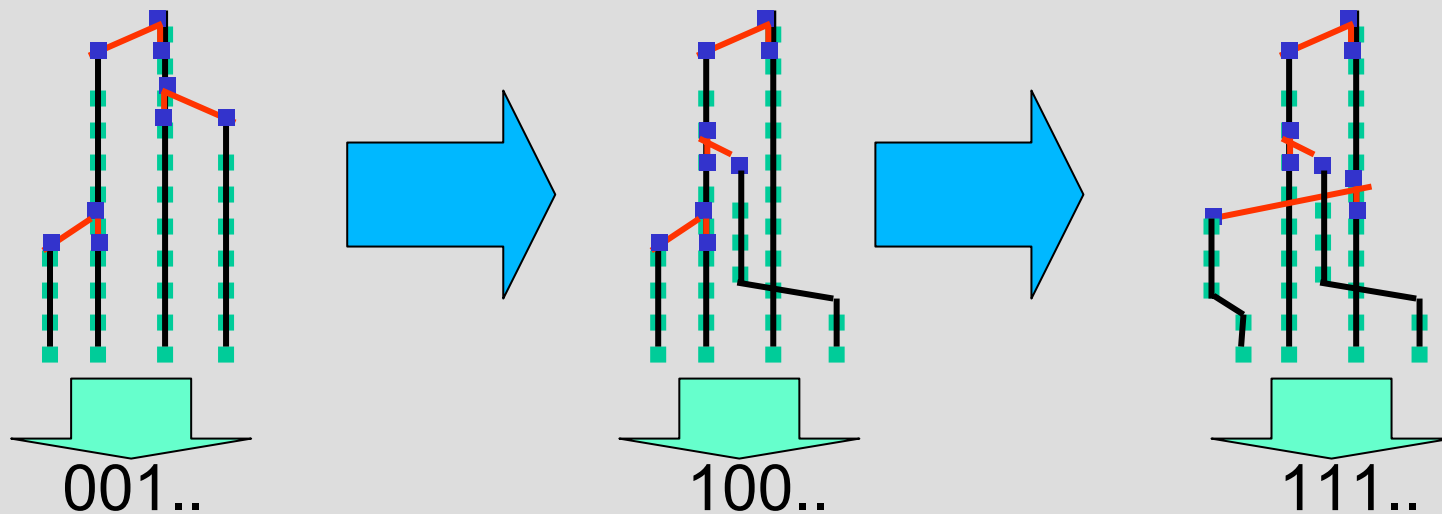
- Alternative approach:
 - Simulate a non-recombinant coalescent tree
 - Compute next event on basepair axis
 - Randomly rewire tree



- A Markov model on trees

Simulating Data: Along the Genome

- Alternative approach:
 - Simulate a non-recombinant coalescent tree
 - Compute next event on basepair axis
 - Randomly rewire tree



- A Markov model on trees
- Hidden Markov Model on the data

Coalescence with Recombination

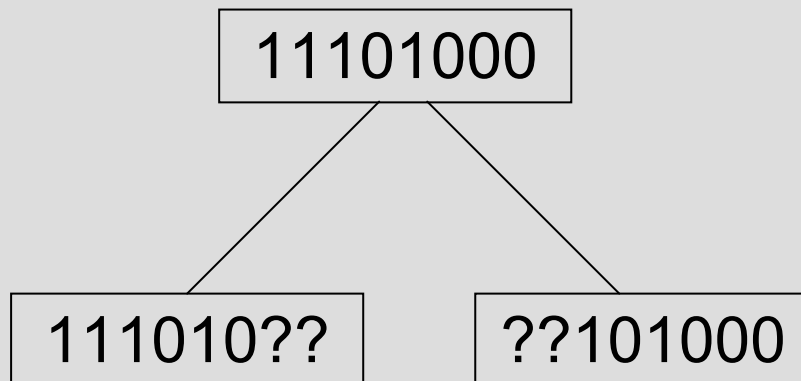
- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
 - Heuristic
 - Probabilistic by coalescence
 - Probabilistic approximation
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Recovering the ARG

- **Input:**
 - Genetic data across a region:
 - Binary haplotype vectors
 - or
 - Trinary {00,Het,11} genotype vectors
- **Output:**
 - ARG giving rise to the data
 - Prioritize solutions by heuristic/formal criteria
- **Goal:**
 - Disease studies
 - Studies of human history

Heuristic ARG Reconstruction

- Alphabet includes “?”
- \underline{V} , \underline{U} *consistent* if equal or “?” \forall coordinate
- Apply rules to backward-evolve vectors:
 - **Split** into consistent vectors



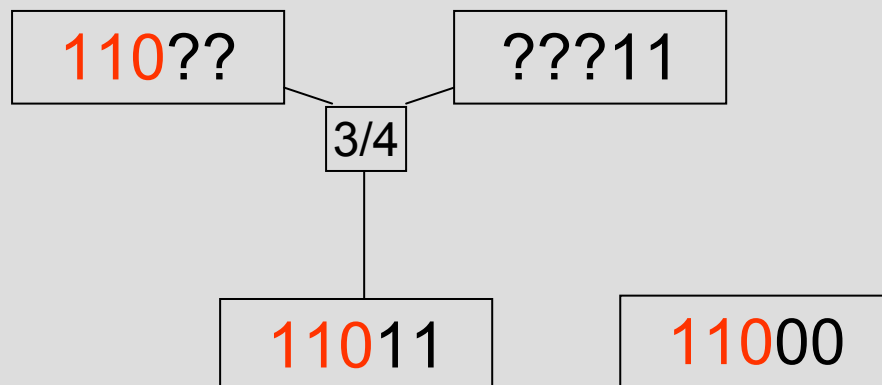
Heuristic ARG Reconstruction

- Alphabet includes “?”
- $\underline{V}, \underline{U}$ consistent if equal or “?” \forall coordinate
- Apply rules to backward-evolve vectors:
 - **Split** into consistent vectors
 - **Mutate** a new allele



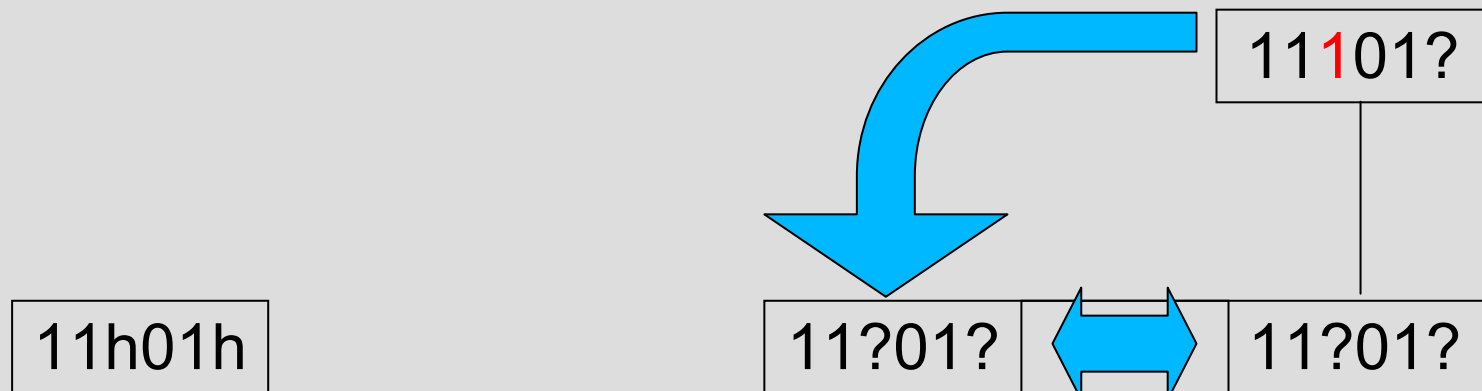
Heuristic ARG Reconstruction

- Alphabet includes “?”
- $\underline{V}, \underline{U}$ consistent if equal or “?” \forall coordinate
- Apply rules to backward-evolve vectors:
 - **Split** into consistent vectors
 - **Mutate** a new allele
 - **Recombine** at the end of a shared tract



Heuristic ARG Reconstruction

- Which rule to apply?
 - Recombination as a last resort
 - Prefer recombination with longer sharing
 - Coalesce recombination parents
- How to treat diploids?
 - Treat heterozygotes as coupled “?”-s, that must be resolved differently



Limitations

- Non-determinism –
 - Generates a distribution over ARGs.
- Implicit goal
 - Minimum recombination assuming no errors, recurrent/reverse mutations

Coalescence with Recombination

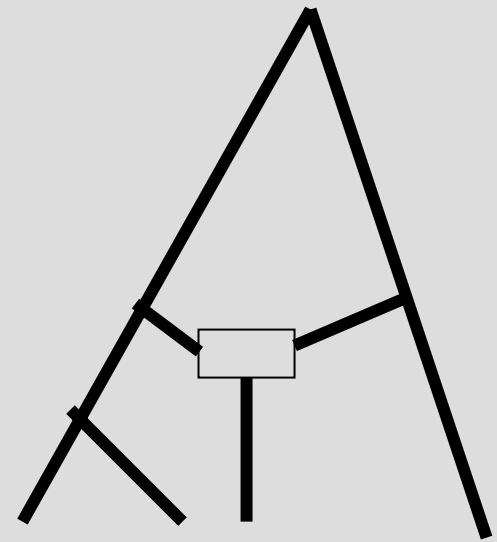
- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
 - Heuristic
 - Probabilistic by coalescence
 - Probabilistic approximation
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Bayesian Methods

- Compute:
 - $Pr(data|ARG)$ by mutation/error probability
 - $Pr(ARG)$ by recombination/coalescence probability
- Ideally:
 - Traverse all ARGs and pick the most probable
- Problem:
 - Impractical for reasonable k, L

Composite Likelihood

- If only $L=2$ polymorphisms:
 - Small L
 - Also small k – only 4 haploid samples
- Strategy:
 - likelihood $\approx \prod$ pairwise likelihood



Parameter Inference

- Primarily θ, ρ
- Summary statistics:
 - #polymorphic sites
 - Heterozygosity
 - #haplotypes
 - Length of non-recombinant segments
- Likelihood methods under the coalescence

Inference under the Coalescence

- **Markov chain Monte Carlo (MCMC):**

Sample a complex distribution space by a Markov chain walk with desired stationary distribution

$$L(\theta, \rho) = \int \Pr(\text{Data} \mid \text{ARG}, \theta, \rho) P(\text{ARG}, \theta, \rho) d\text{ARG}$$

- **Importance sampling:**

Focus sampling on regions that matter:

ARGs with high contribution to likelihood

$$L(\theta, \rho) = \int_{\text{consistent ARG}} P(\text{ARG}, \theta, \rho) d\text{ARG}$$

Importance Sampling

- Sample ARGs by randomized selection of the preceding event
- A probabilistic formulation of ARG reconstruction

Coalescence with Recombination

- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
 - Heuristic
 - Probabilistic by coalescence
 - Probabilistic approximation
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

A Simpler Model

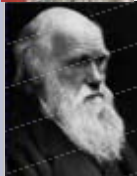
(Stephens & Donnelly)

- Approximate coalescence by resampling
- The next sample has recently diverged from a lineage leading to an existing sample
- Upon recombination closest sample is reselected



Hidden Markov Model

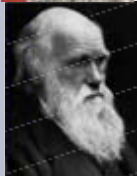
- Generates:
 - The next chromosome in the sample
- States:
 - $S \times M$
 - S : Existing chromosomes in the sample
 - M : markers



genome

Hidden Markov Model

- Transitions:
 - Typically the same sample, rare recombination
- Emission:



genome

Hidden Markov Model

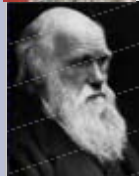
- Transitions:
 - Typically the same sample, rare recombination
- Emission:
 - Typically same as sample genotype, rare mutation or error



0 1 1 1 1 0



1 0 1 0 1



0 1 0 0 1

genome

Approximated Likelihood

- Randomize sample order
- For $i=2 \dots |S|$
 - Compute $P_i = \Pr(S_i | S_1, \dots, S_{i-1})$
- Return $\prod P_i$

Application: Phasing

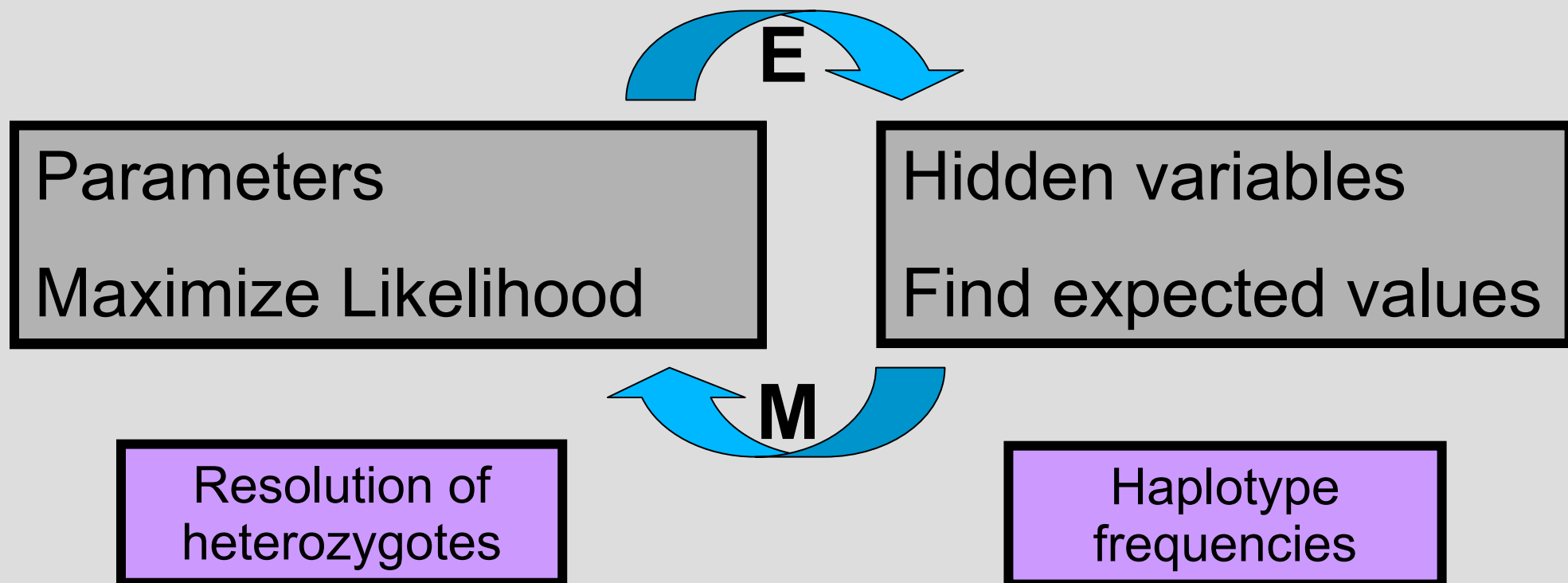
- **Input:** diploid samples
- **Output:** samples as pairs of haploids
(heterozygotes resolved)
- **Method:** Modify HMM to output a diploid;
Follow the best paths to phase

Coalescence with Recombination

- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Phasing by E-M

- E-M:
 - Method for maximum-likelihood parameter inference with hidden variables



Phasing by E-M

Data:

1 0 h h 1

1	0	0	0	1	1/4
1	0	1	1	1	1/4
1	0	0	1	1	1/4
1	0	1	0	1	1/4

h 0 0 1 h

0	0	0	1	0	1/4
1	0	0	1	1	1/4
0	0	0	1	1	1/4
1	0	0	1	0	1/4

1 h h 1 1

1	0	0	1	1	1/4
1	1	1	1	1	1/4
1	0	1	1	1	1/4
1	1	0	1	1	1/4

Maximization

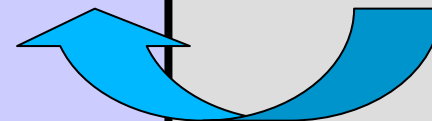
0.4
0.6

0.75
0.25

0.6
0.4

Expectation

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



Phasing by E-M

Data:

1 0 h h 1

1	0	0	0	1	1/4
1	0	1	1	1	1/4
1	0	0	1	1	1/4
1	0	1	0	1	1/4

h 0 0 1 h

0	0	0	1	0	1/4
1	0	0	1	1	1/4
0	0	0	1	1	1/4
1	0	0	1	0	1/4

1 h h 1 1

1	0	0	1	1	1/4
1	1	1	1	1	1/4
1	0	1	1	1	1/4
1	1	0	1	1	1/4

Maximization

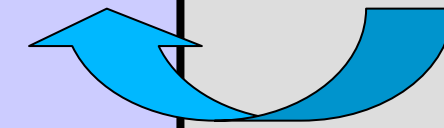
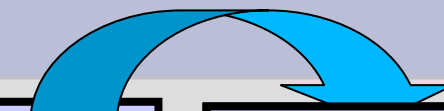
0.4
0.6

0.75
0.25

0.6
0.4

Expectation

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



Phasing by E-M

Data:

1 0 h h 1

1	0	0	0	1	1/4
1	0	1	1	1	1/4
1	0	0	1	1	1/4
1	0	1	0	1	1/4

h 0 0 1 h

0	0	0	1	0	1/4
1	0	0	1	1	1/4
0	0	0	1	1	1/4
1	0	0	1	0	1/4

1 h h 1 1

1	0	0	1	1	1/4
1	1	1	1	1	1/4
1	0	1	1	1	1/4
1	1	0	1	1	1/4

Maximization

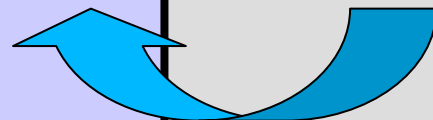
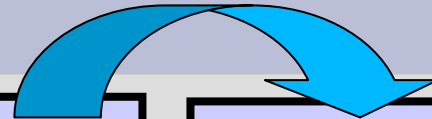
0
1

1
0

1
0

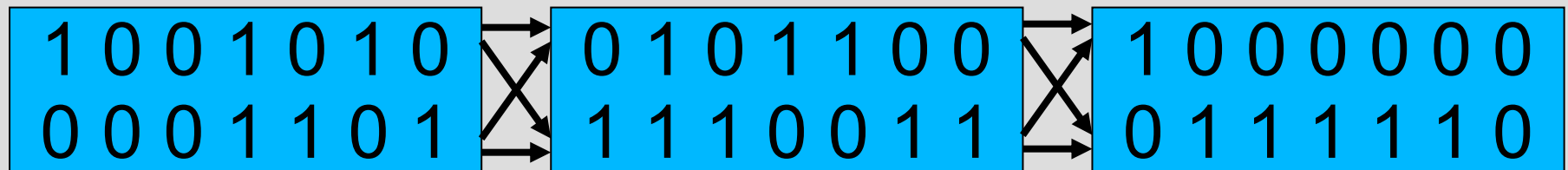
Expectation

0	0	0	1	0	1/6
0	0	0	1	1	0
1	0	0	0	1	0
1	0	0	1	0	0
1	0	0	1	1	1/2
1	0	1	0	1	1/6
1	0	1	1	1	0
1	1	0	1	1	0
1	1	1	1	1	1/6



Partition-Ligation EM

- In practice:
 - #variables is exponential in too many sites
- Solution:
 - Locally phase each region
 - Merge by phasing vectors of haplotype pairs



Coalescence with Recombination

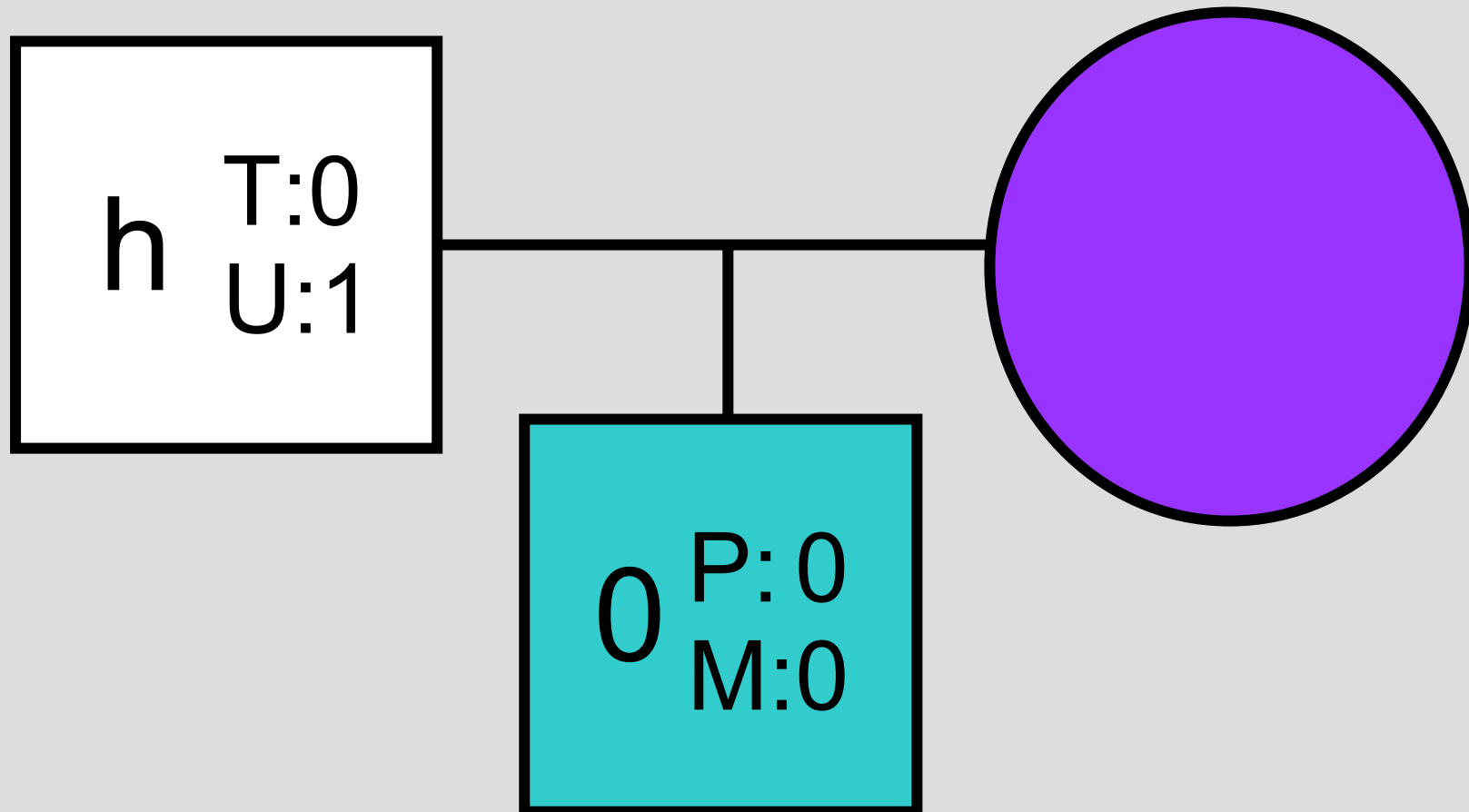
- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Family Trios



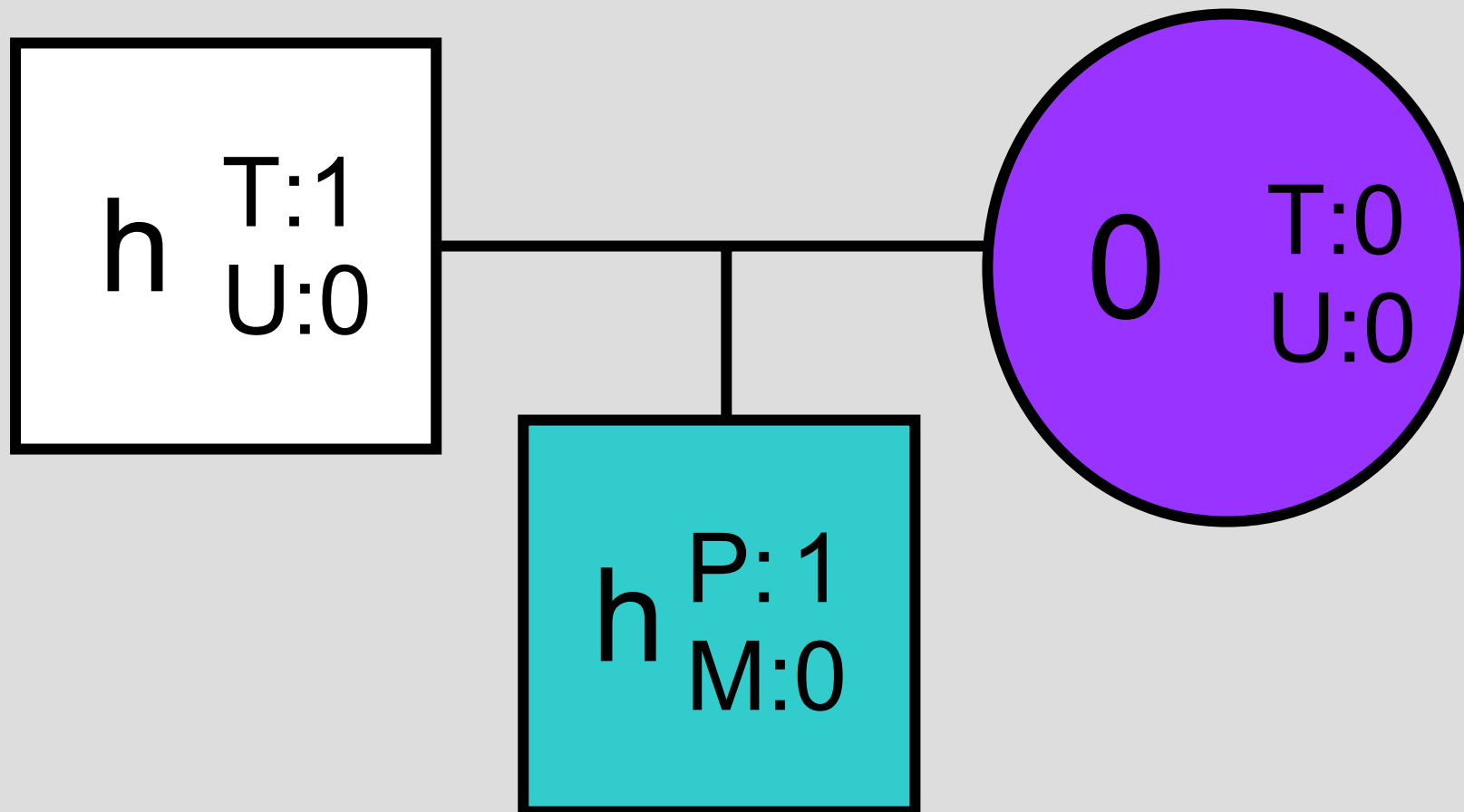
Phasing Family Trios

- Resolve **T**ranmitted/**U**ntranmitted
- Resolve **P**aternal/**M**aternal



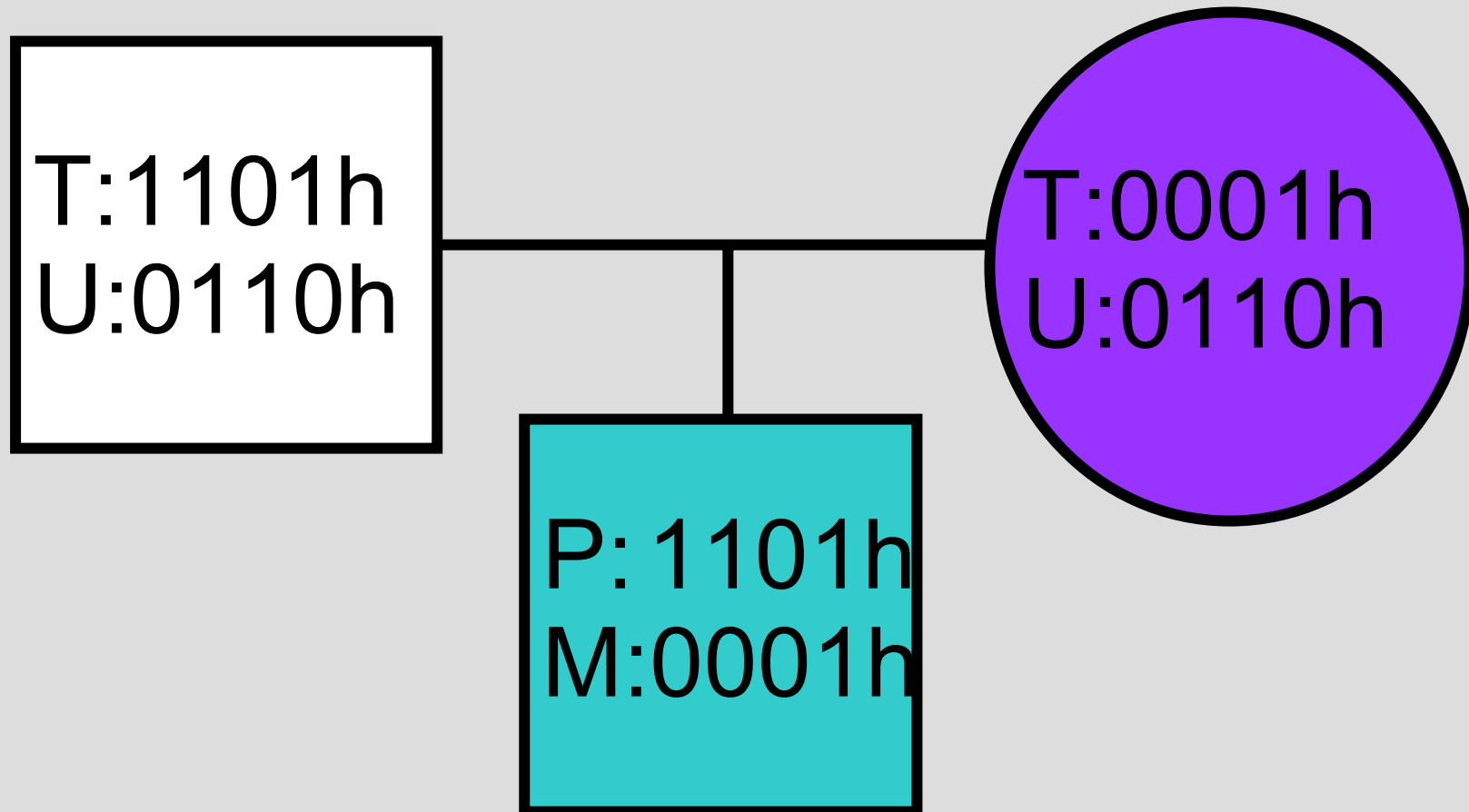
Phasing Family Trios

- Resolve **T**ranmitted/**U**ntranmitted
- Resolve **P**aternal/**M**aternal



Phased Chromosomes in Trios

- Triple h is unresolved
- Rare recombination in parental chromosomes



Phased Chromosomes in Trios

- Triple h is unresolved
- Rare recombination in parental chromosomes
- T/U label is relative to the offspring currently under investigation.



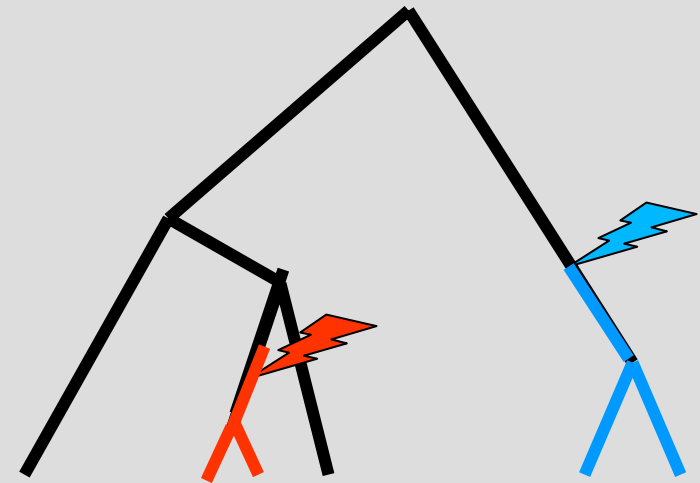
Coalescence with Recombination

- Ancestral recombination graph
 - Model
 - Simulation
 - Inference
- Haplotype inference
 - EM
 - Mendel-based
 - Relations between polymorphisms

Reminder: Mutations on a Tree

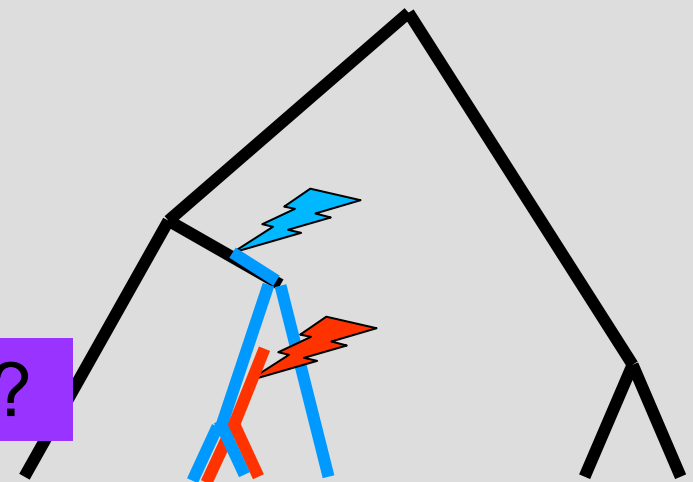
- Subtrees are either disjoint

Haplotypes: 00
01
10



or contained in one another

Haplotypes: 00
01
11



How to deal with some recombination?

Linkage (Dis)Equilibrium

- If independent SNPs of frequencies p_0, p_1 and p'_0, p'_1 :
$$\Pr(ij) = p_i p'_j$$

- Otherwise: Deviation

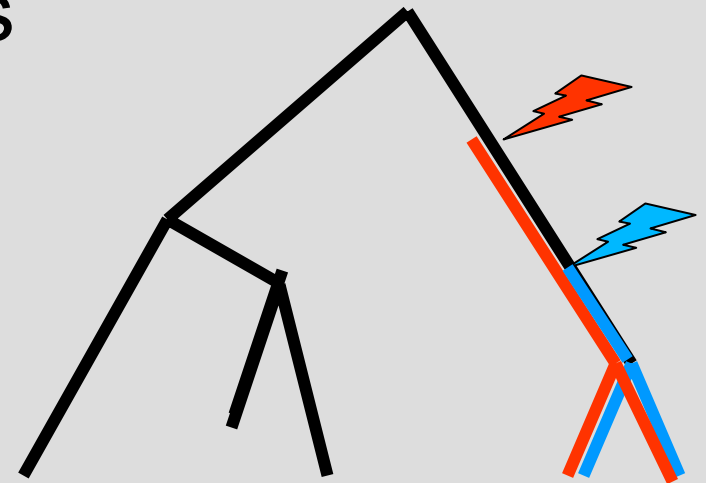
$$D = \Pr(ij) - p_i p'_j$$

- D_{\max} : when a gamete is missing
- $D' = D/D_{\max}$

Mutations on a Lineage

- Genetically equivalent SNPs
- With rare recombination:
 - SNPs with high r^2

$$r^2 = D' / \sqrt{p_0 p'_0 p_1 p'_1}$$



Summary

- Ancestral recombination graph models historical lineages
 - Facilitates simulation and inference
- Haplotypes can be inferred by probabilistic or combinatorial methods

Bibliography

- Wiuf C, Hein J. Related Articles, Recombination as a point process along sequences. *Theor Popul Biol.* 1999 Jun;55(3):248-59.
- Minichiello MJ, Durbin R. Mapping trait loci using inferred ancestral recombination graphs. *Am J Hum Genet.* Epub 2006 Sep
- Stephens M, Smith NJ, Donnelly P. Related Articles, A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001 Apr;68(4):978-89.
- Fearnhead P, Donnelly P. Related Articles, Estimating recombination rates from population genetic data. *Genetics.* 2001 Nov;159(3):1299-318.
- Excoffier L, Slatkin M. Related Articles, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995 Sep;12(5):921-7.
- Niu T, Qin ZS, Xu X, Liu JS. *Am J Hum Genet.* 2002 Jan;70(1):157-69
- R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge University Press, 1998.

Extra Credit

1. Formally write down the transition and emission probabilities for the Stephens-Donnelly HMM
2. Describe how would the following seem to violate Mendel's laws:
 1. A hemizygous (appears in one copy, whereas the other is deleted) region in some of the samples.
 2. A SNP in a repeat region

Project Suggestion

- Prediction of genetic variants
 - Implement the Stephens-Donnelly HMM for diploid samples
 - Assume you have chosen a SNP every 5kb on average in ENCODE regions, and typed them
 - Use HapMap ENCODE data to evaluate your ability to predict other SNPs.
 - Report your ability to predict variation based on properties of the predicted SNP: chromosome, coding status, etc.