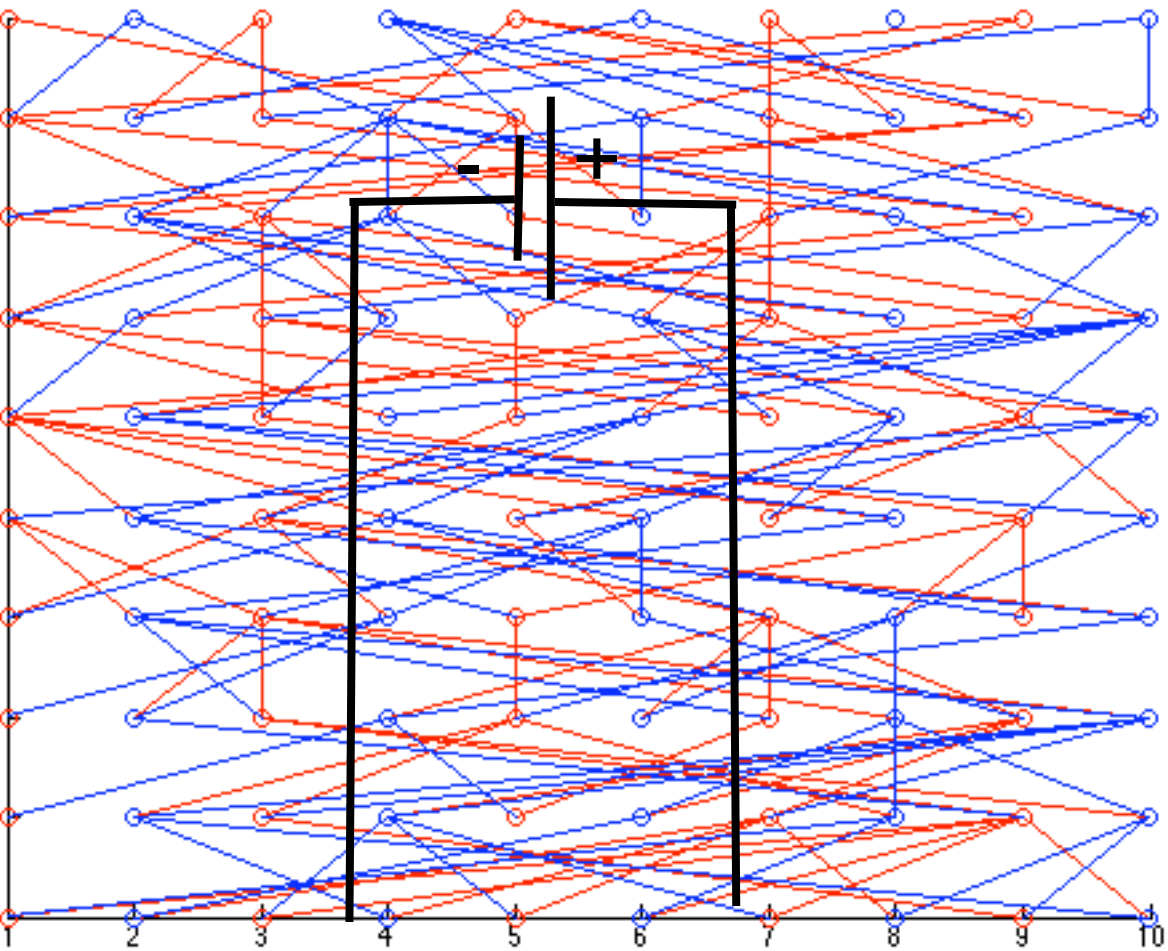
Can you hear the shape of the human genealogy?
Lecture 3, CSSS09



Greg Leibon
Memento, Inc
Dartmouth College

The Plan

Part 1: Towards a *Genealogical Conductance*

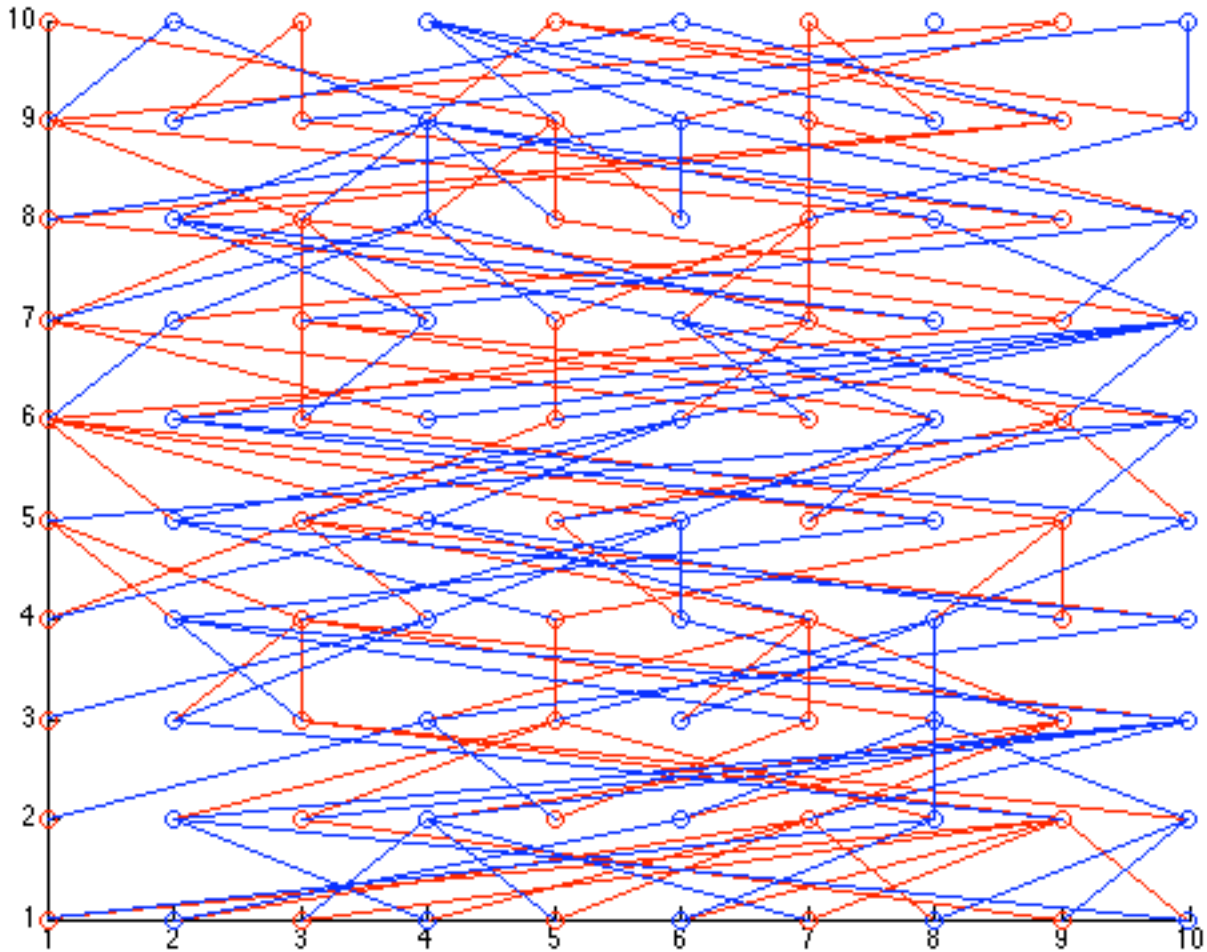
Part 2: The hunt for *IBD Regions*

Appendix: An introduction to the *HMM*

Part 1: Towards a metric on genealogical history

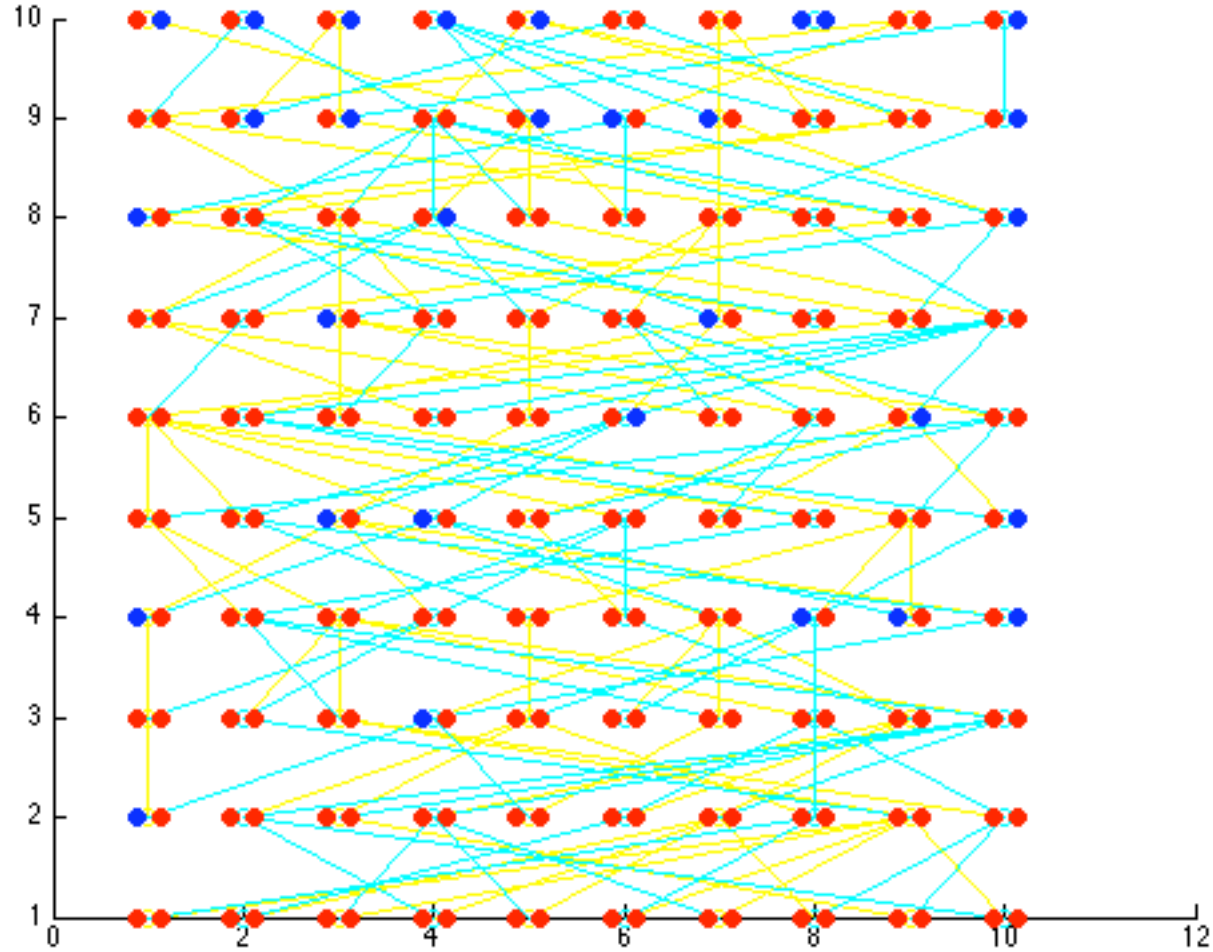
Begins with another flavor of Markov chain....

The Wright-Fisher Model



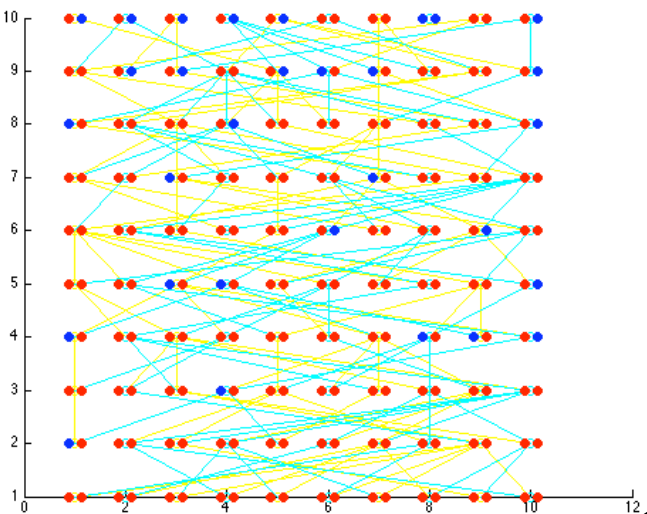

```
N=5;
T=10;
[O E]=WFsim(N,T);
```

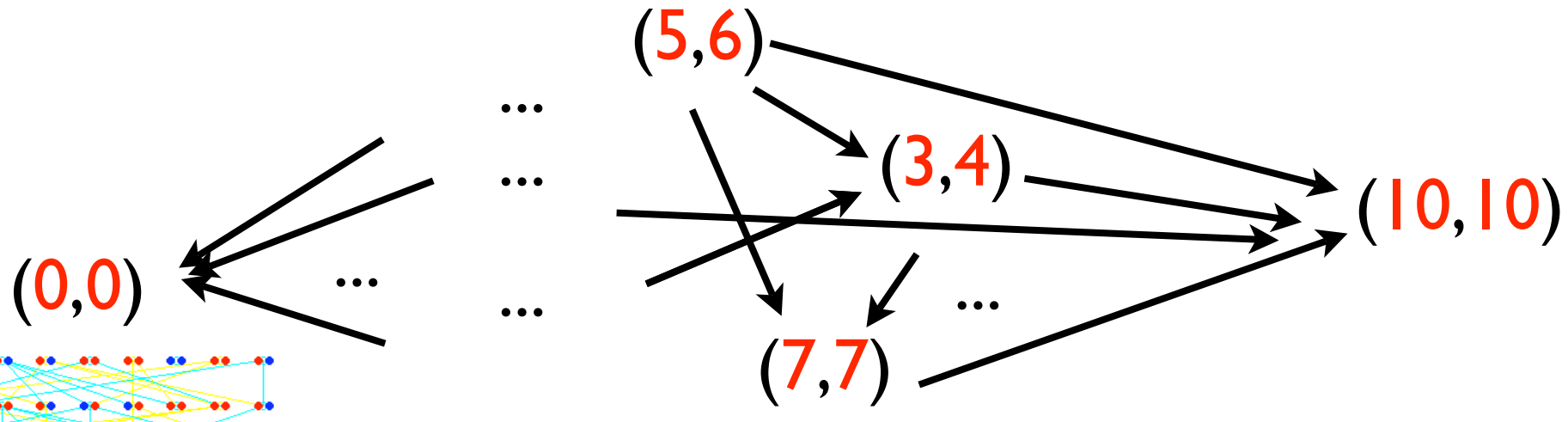
Hardy-Weinberg equilibrium...



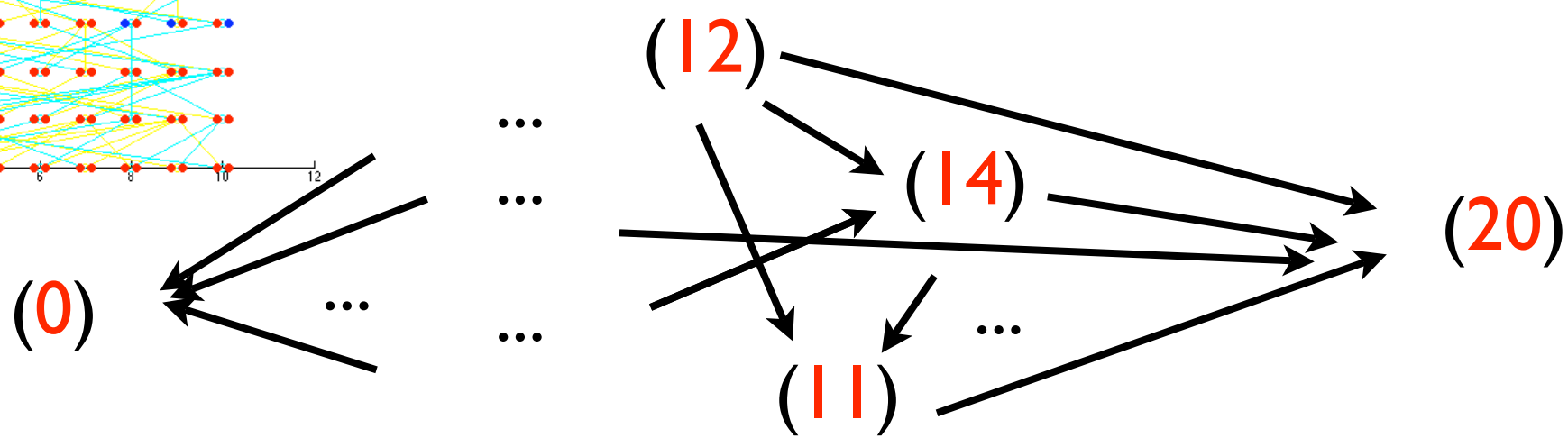

```
close all
GraphBase(O,E,T,N,'y','c');
LabeAlle(O,E,T,N,'r','b',1/8);
```

The Allele Frequency Chain: An absorbing chain



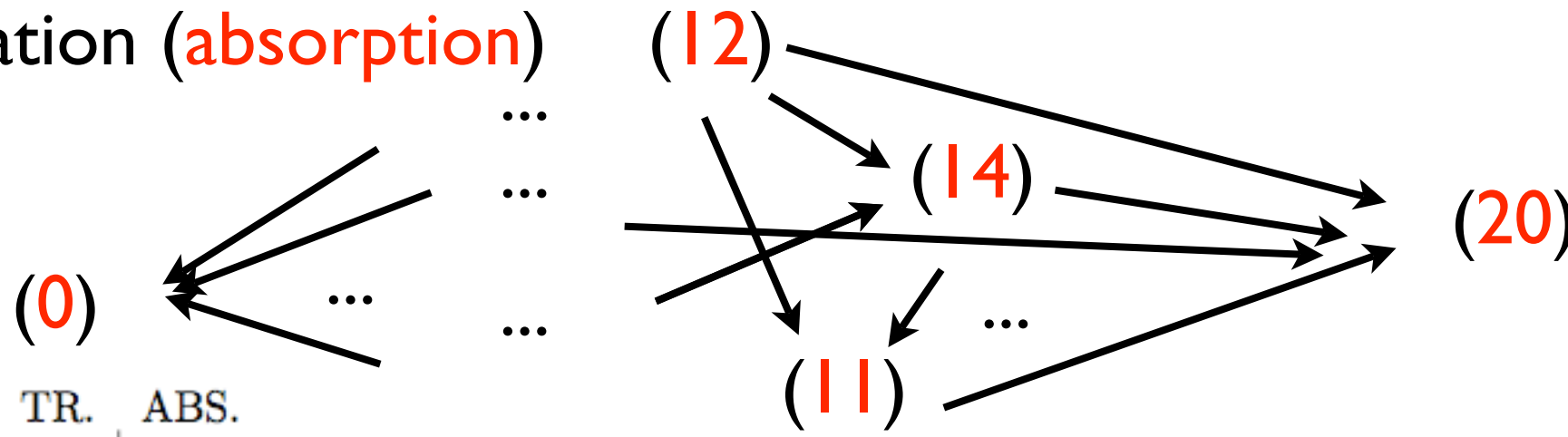


Ignoring the male female thing...



$$P(R(k) = L \mid R(k-1) = K) = \binom{2N}{L} \left(\frac{K}{2N}\right)^{L} \left(1 - \frac{K}{2N}\right)^{2N-L}$$

Time to fixation (absorption)



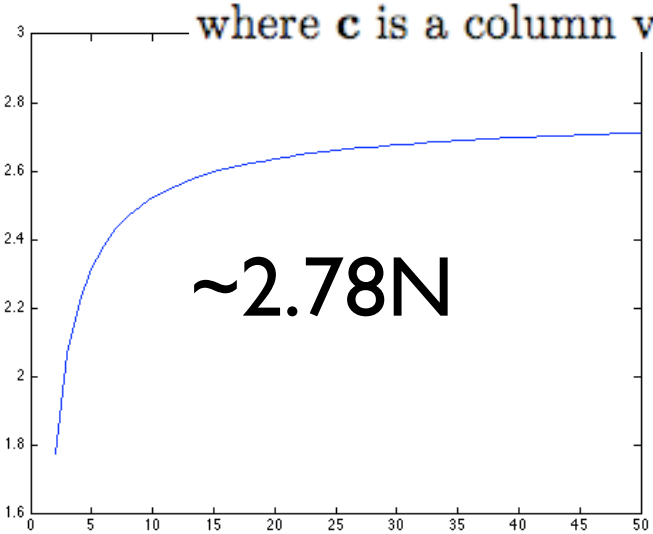
$$\mathbf{P} = \begin{array}{c} \\ \text{TR.} \\ \\ \text{ABS.} \end{array} \begin{array}{c} \begin{array}{cc} \text{TR.} & \text{ABS.} \end{array} \\ \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right) \end{array}$$

$$(\mathbf{I} - \mathbf{Q})^{-1} = \mathbf{N}$$

Theorem 11.5 Let t_i be the expected number of steps before the chain is absorbed, given that the chain starts in state s_i, and let $\mathbf{t}$ be the column vector whose ith entry is t_i. Then

$$\mathbf{t} = \mathbf{Nc} \ ,$$

where $\mathbf{c}$ is a column vector all of whose entries are 1.



~2.78N

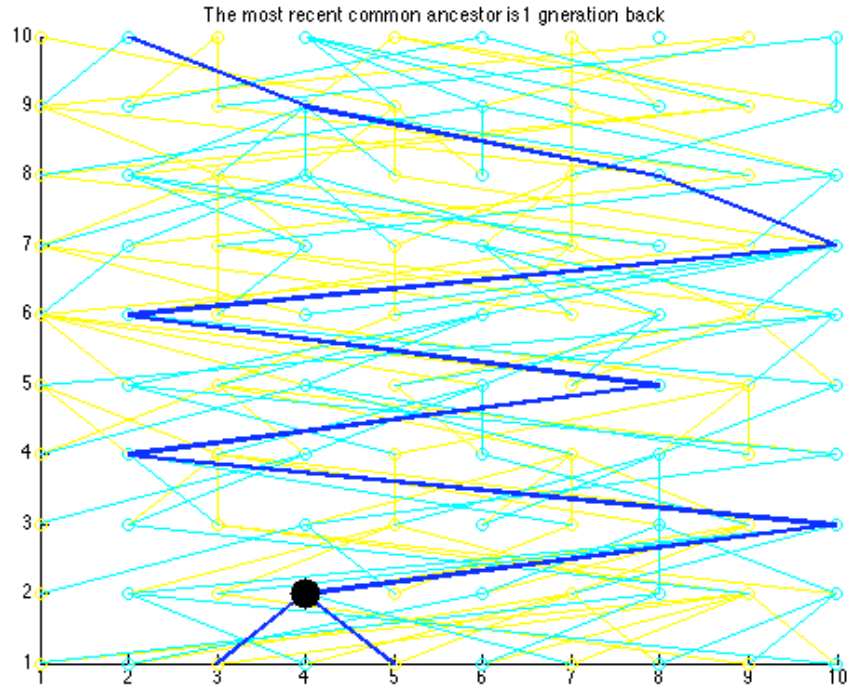
See Grinstead and Snell...the best things in life are free!

http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html

We give this argument because of our needs later, but the diffusion approximation is really the way to go (go Laplacian!). See...

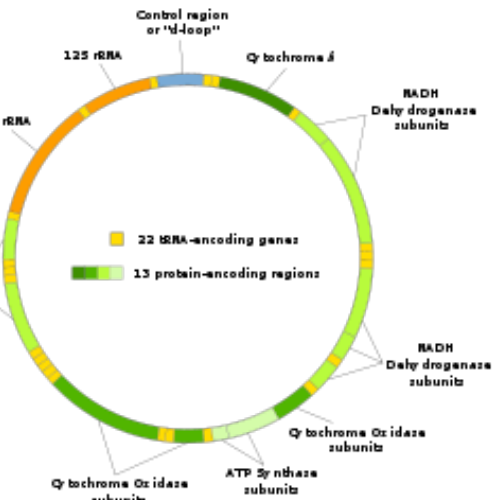
C. Neuhauser 2000. Mathematical Models in Population Genetics. In Handbook of Statistical Genetics. Pp. 153-178. Wiley.

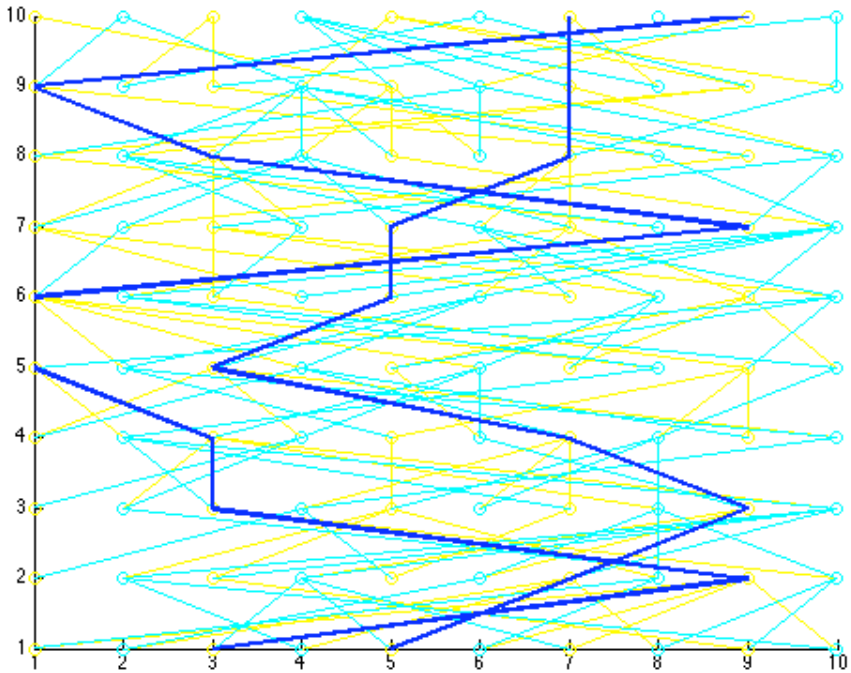
What is a good notion of genealogical distance?




```
p1=3; p2=5; MotherHist(p1,p2,O,E,T,N);
```

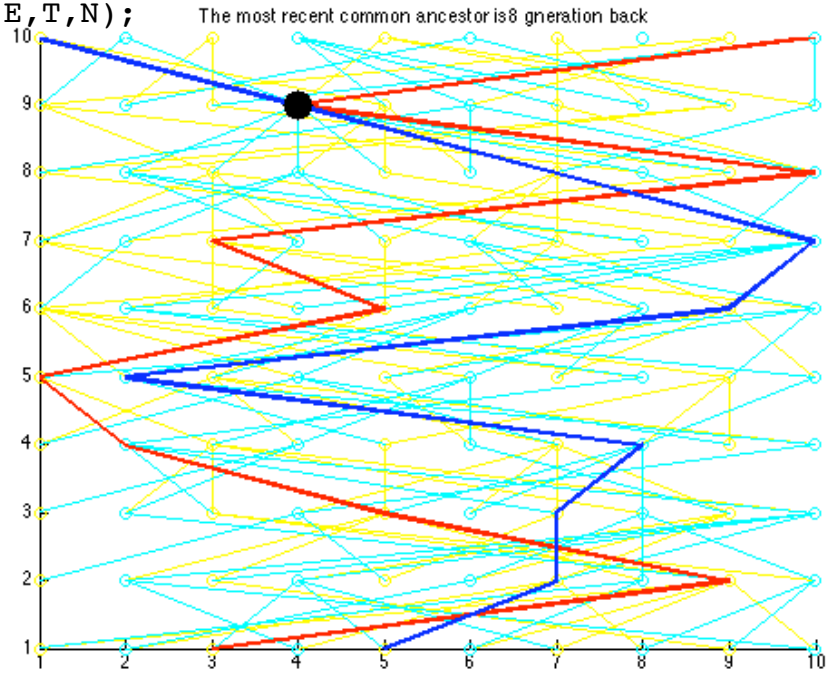
Mothers



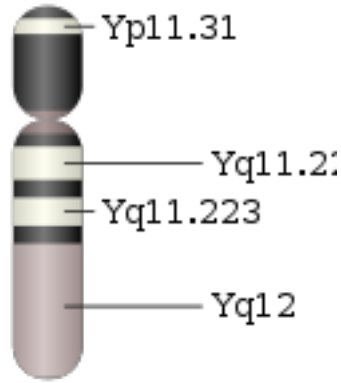


```
FatherHist(p1,p2,O,E,T,N);
```

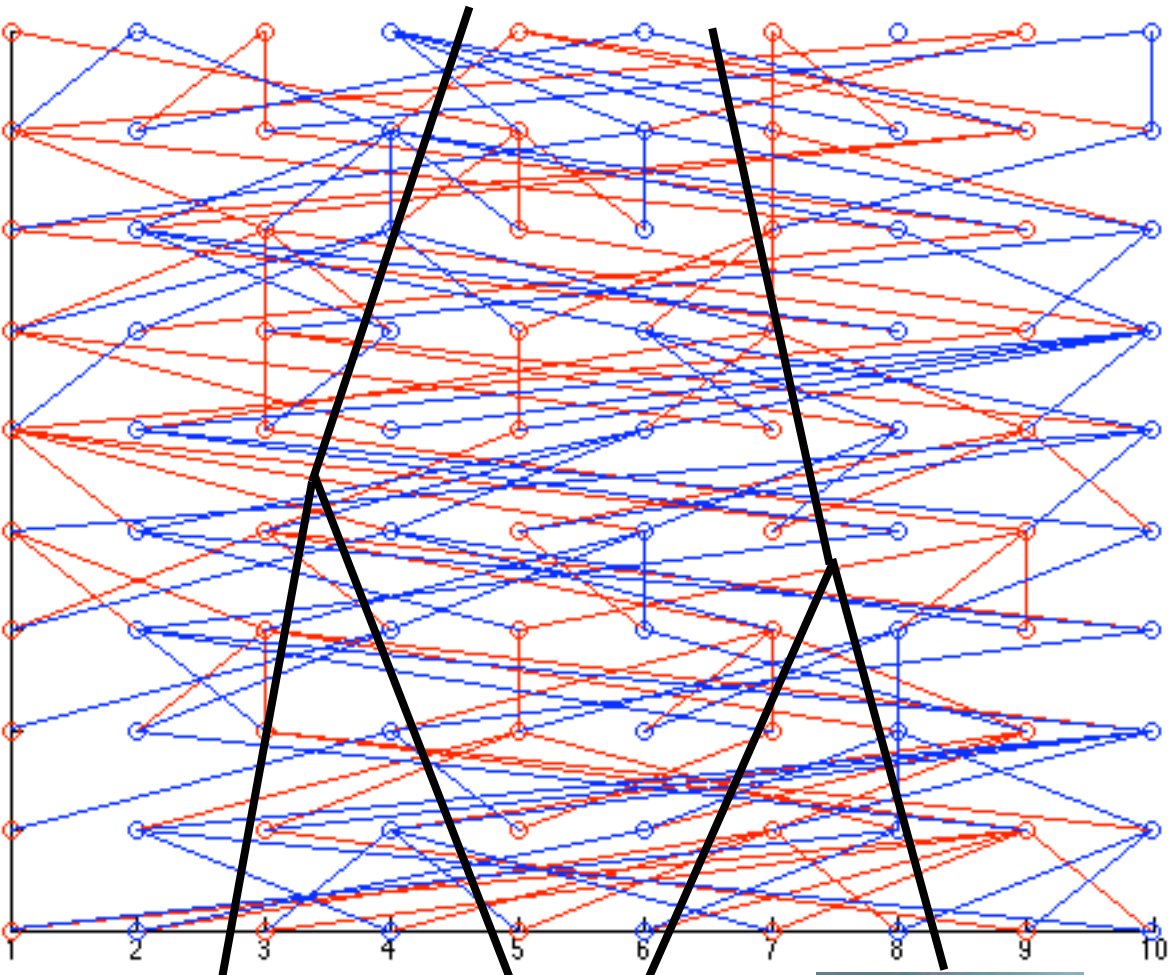
Fathers

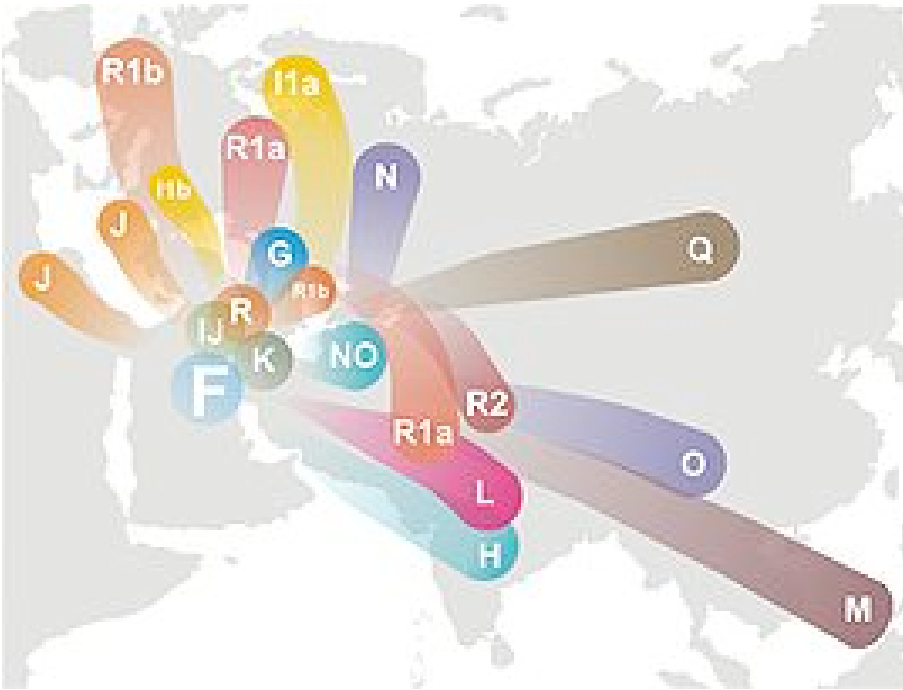


Random



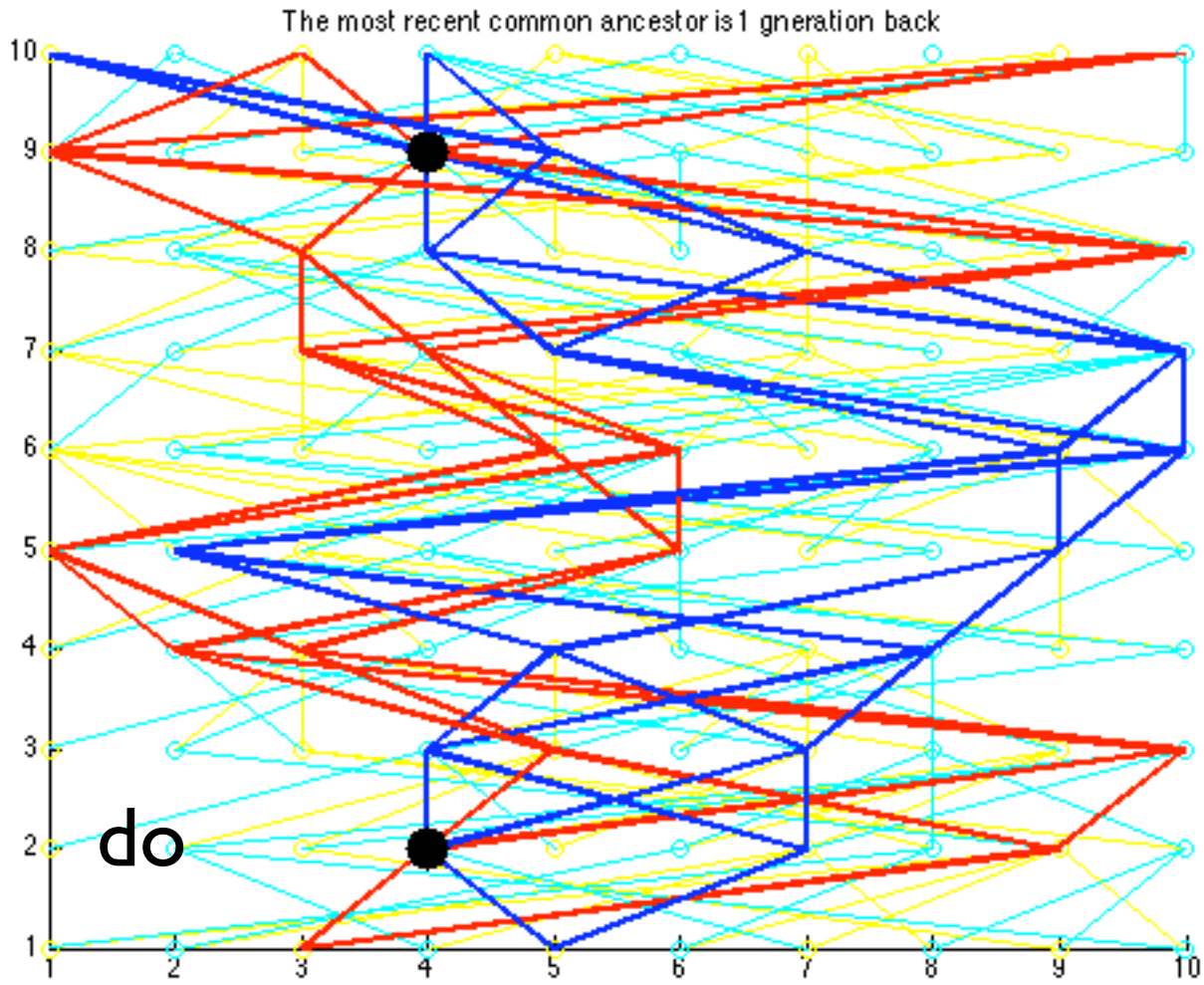

```
PairHist(p1,p2,O,E,T,N,'r','b');
```





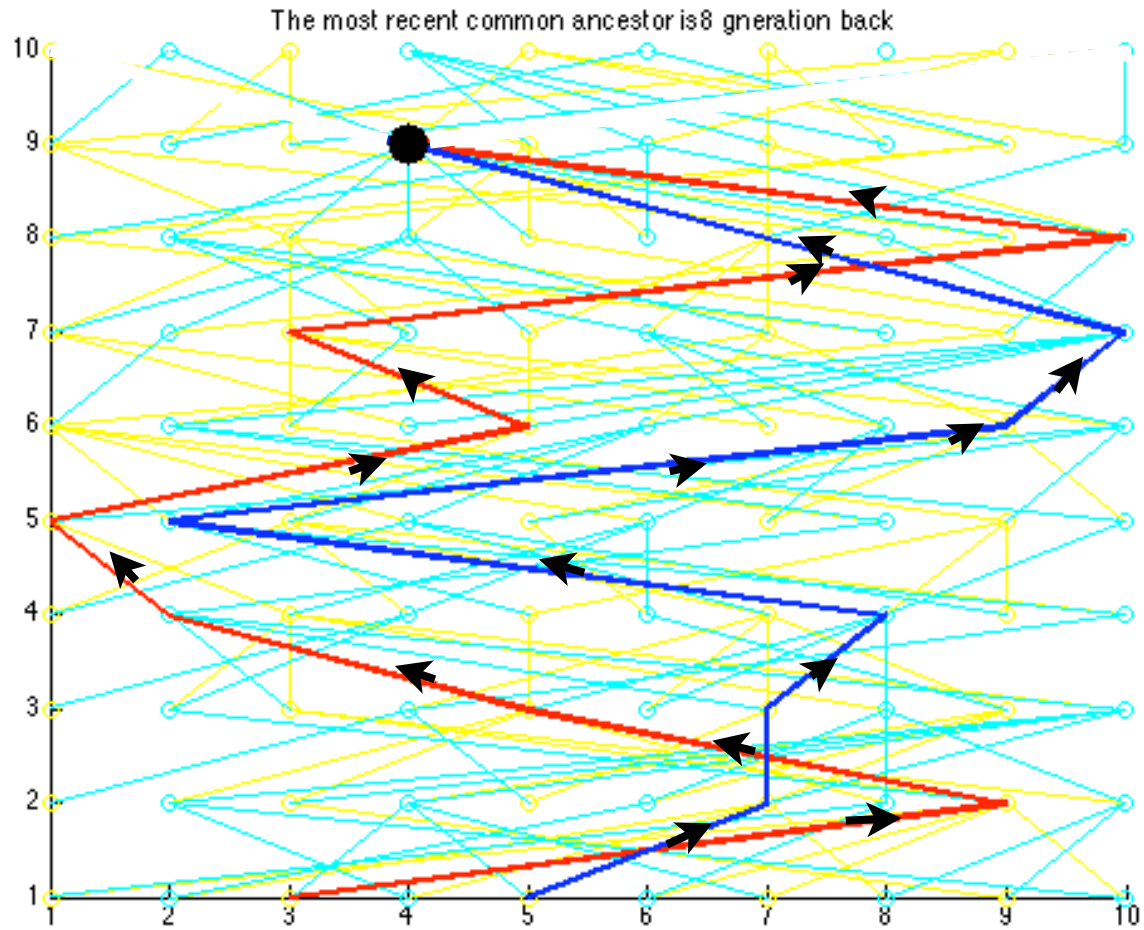
A research study conducted by the [New England Historic Genealogical Society] has turned up a variety of interesting relations to the current Democratic & Republican presidential candidates. According to the study, **Barack Obama** is distantly related to such luminaries as **George W. Bush** and **Brad Pitt**, while **Hillary Clinton** has familial ties to **Angelina Jolie** and **Madonna**. **John McCain** turns out to be a cousin of First Lady **Laura Bush**.

Genealogical distance?



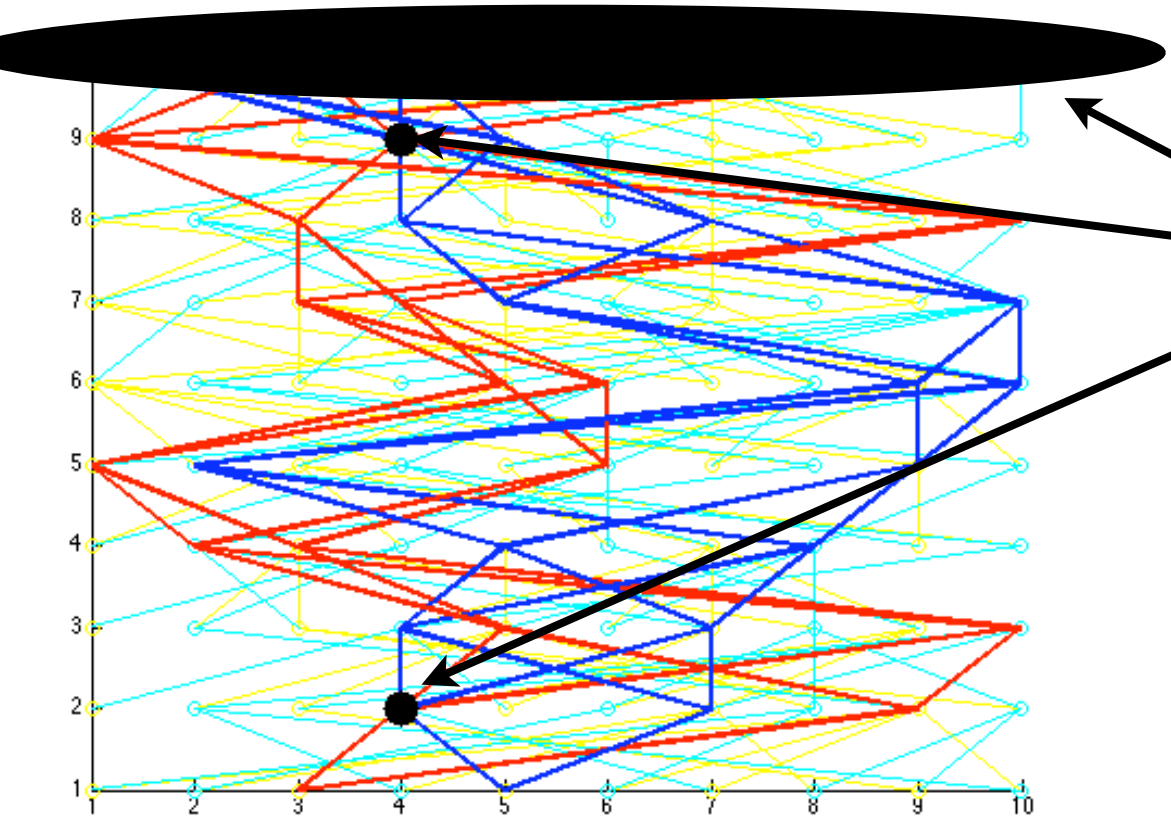
$$E(T_{ab})$$

....or borrow some ideas from Coalescence Theory



Use the historically direct chain... and turn *coalescent events* into absorbing states.

$$E(T_{MRCA})$$



genealogical distance

$$E(T_{MRCA}^{max})$$

genealogical conductance

$$E\left(\frac{1}{T_{MRCA}^{max}}\right)$$

Can we compute them on real data?

Yes...thanks to recombination!

Recombination events

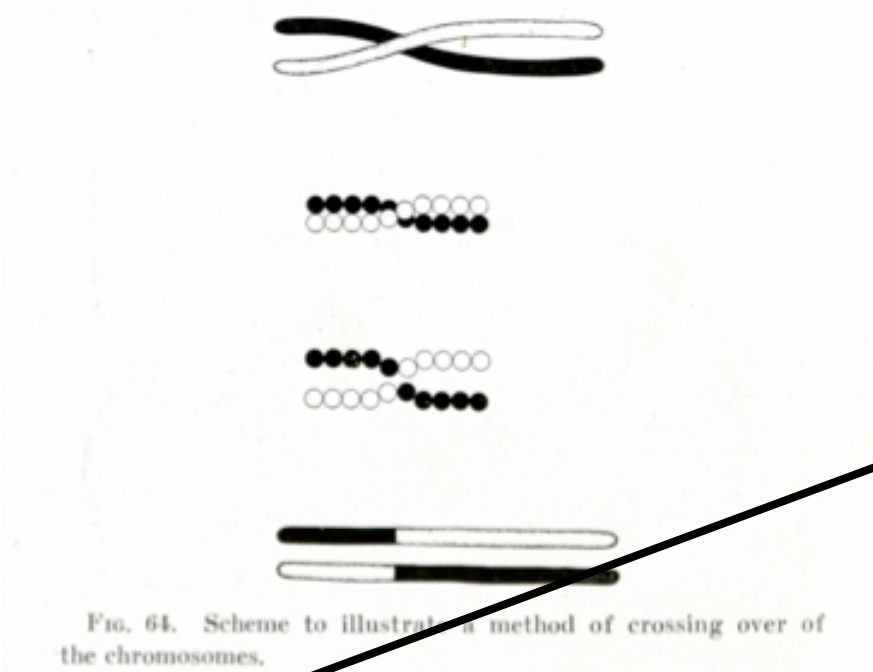
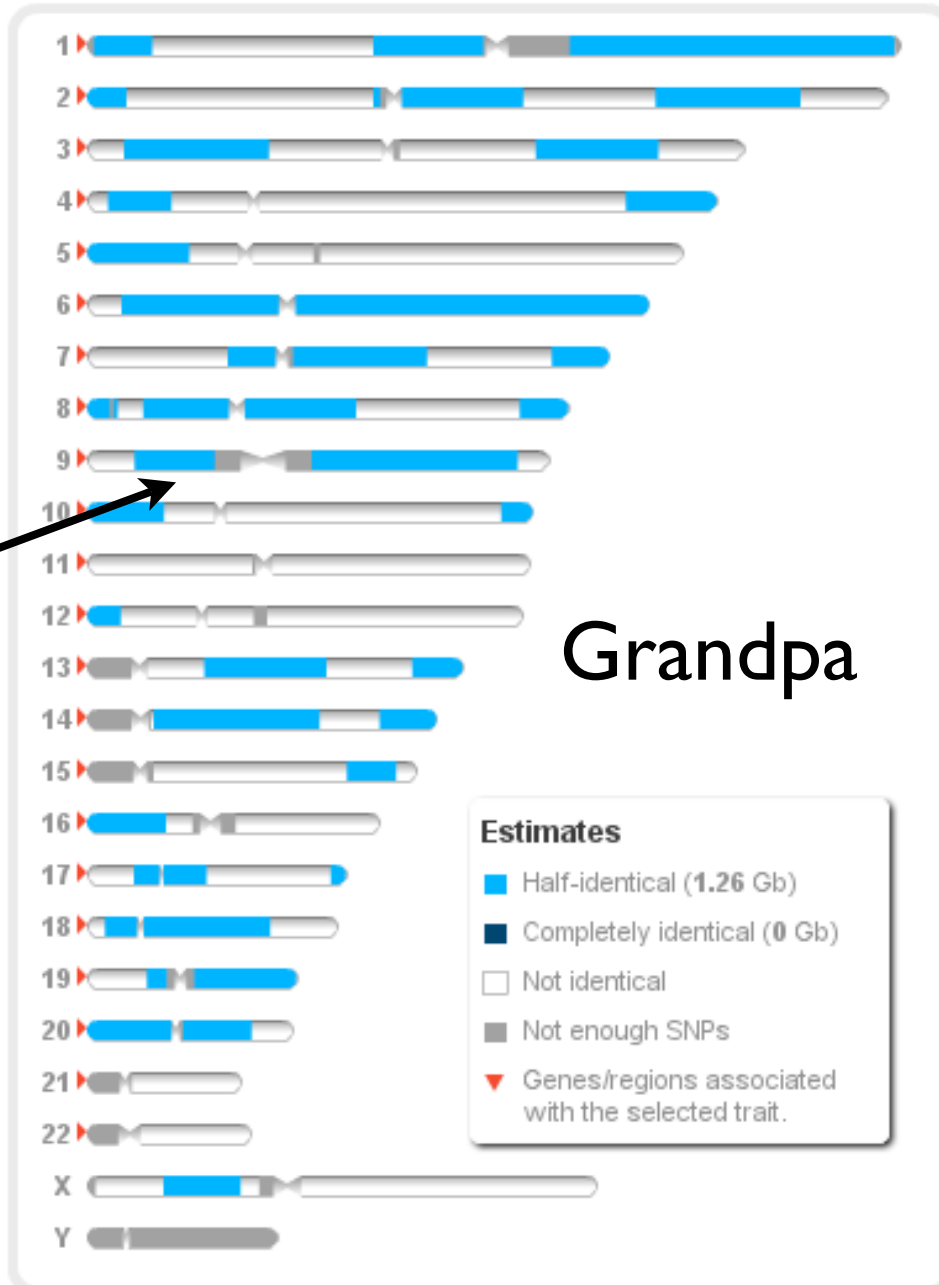


Fig. 64. Scheme to illustrate a method of crossing over of the chromosomes.

IBD (identical by decent)



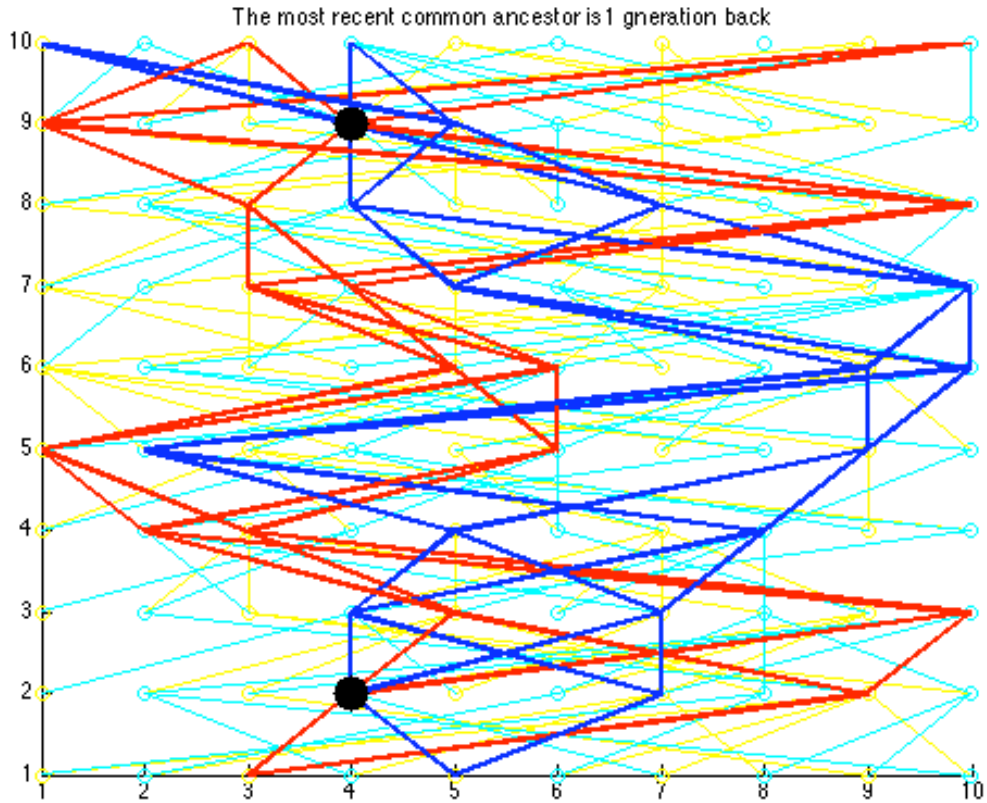
Grandpa

Key fact:

$$E(L(IBD) \mid T_{MRCA}) = \frac{1}{2T_{MRCA}}$$

length in Morgans

Image 23&me



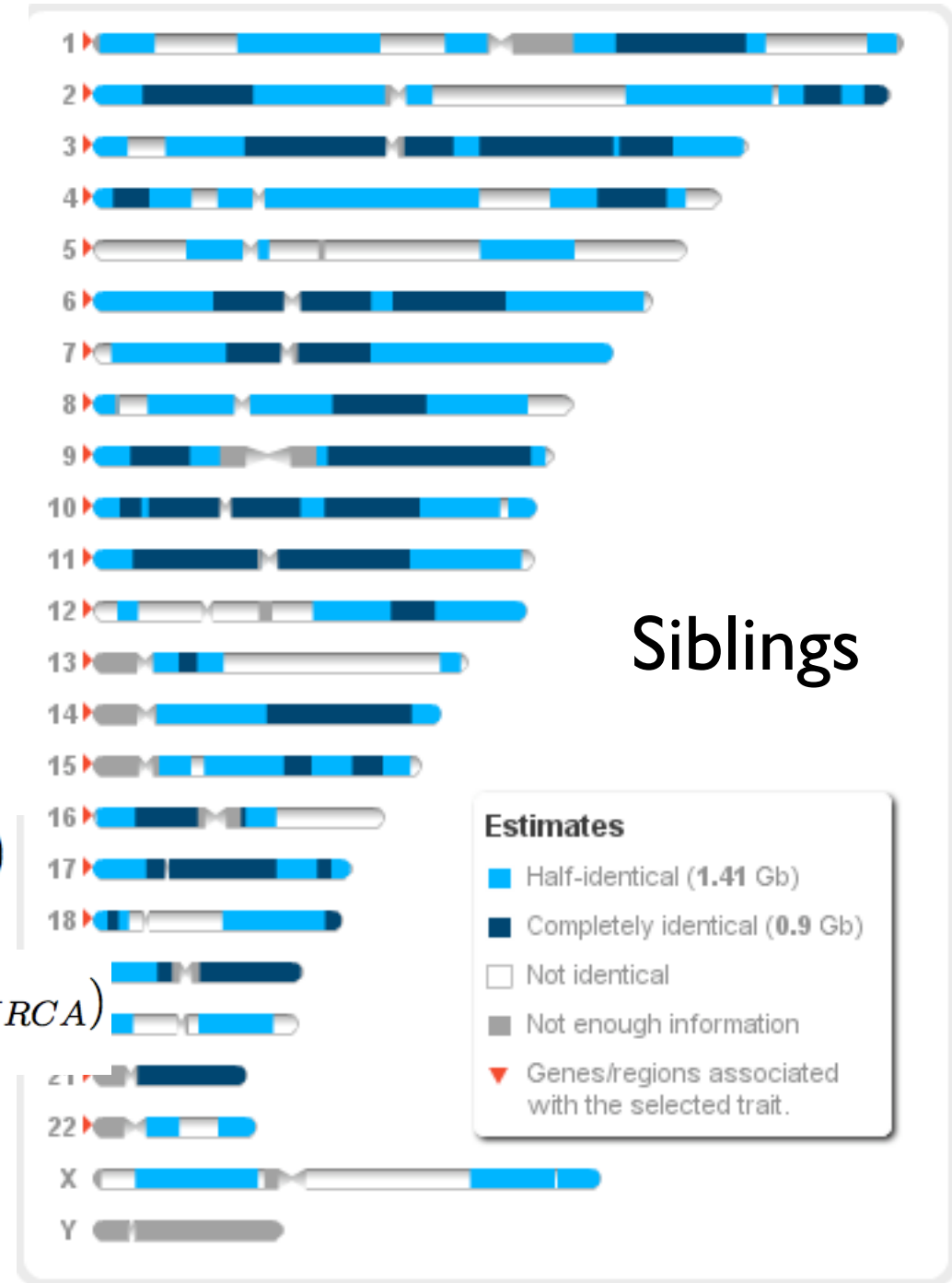
$$E\left(\frac{1}{T_{MRCA}}\right) = \sum_T \frac{1}{T} P(T = T_{MRCA})$$

$$= 2\sum_T E(L(IBD) \mid T = T_{MRCA}) P(T = T_{MRCA})$$

$$= 2E(E(L(IBD) \mid T_{MRCA}))$$

$$= 2E(L(IBD))$$

second fundamental mystery
of probability theory

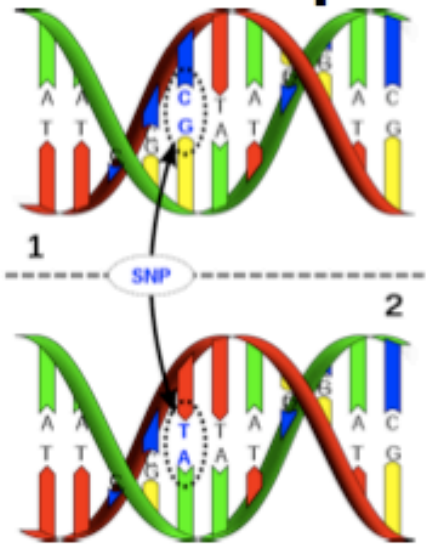


Siblings

Image 23&me

Part 2: The hunt for IBD region

Single nucleotide polymorphism



Roughly 10 million of the 12,000 million bits of the genome in are this form

A single nucleotide polymorphism is a variation of a DNA sequence that affects a single nucleotide in such a way that some individuals carry a variant, for example AAGCCTA, and some individuals carry a different variant, for example AAGCTTA.

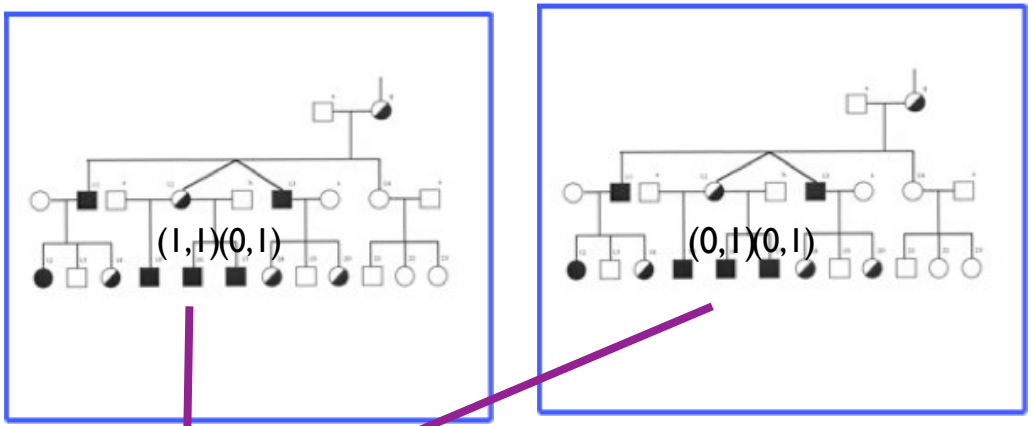
Genotype Data Provided

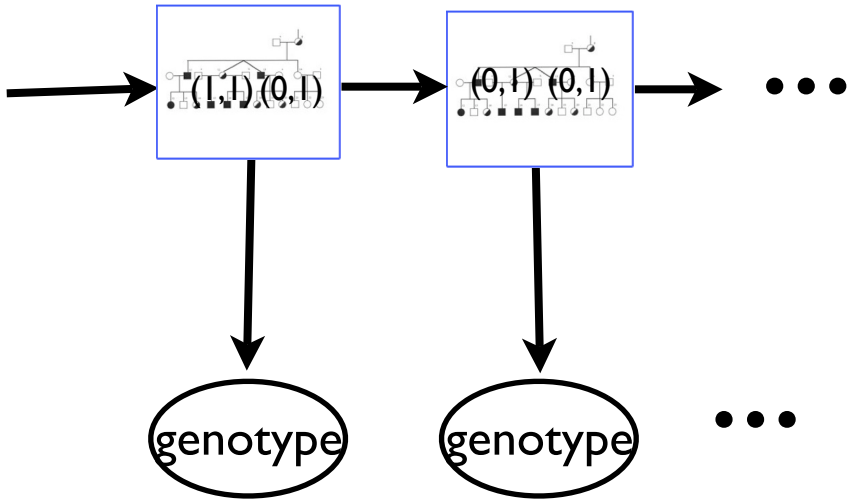
$A_1 B_1$	$\{A_1, A_1\}$
$A_2 A_2$	$A_2 B_2$
$B_3 B_3$	$A_3 A_3$
$B_4 B_4$	$B_4 B_4$
$A_5 A_5$	$A_5 B_5$

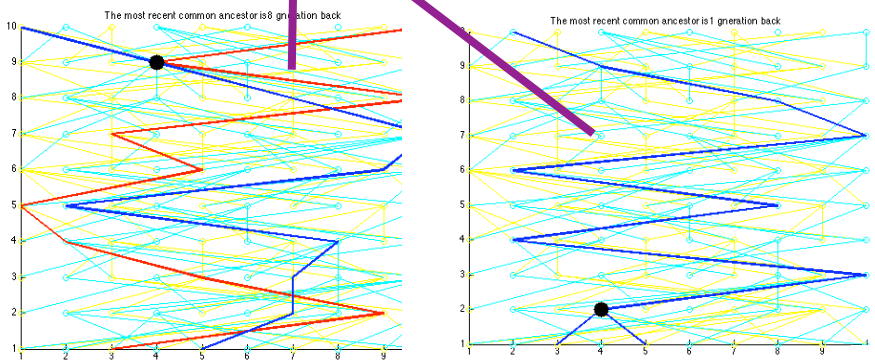
an unordered set

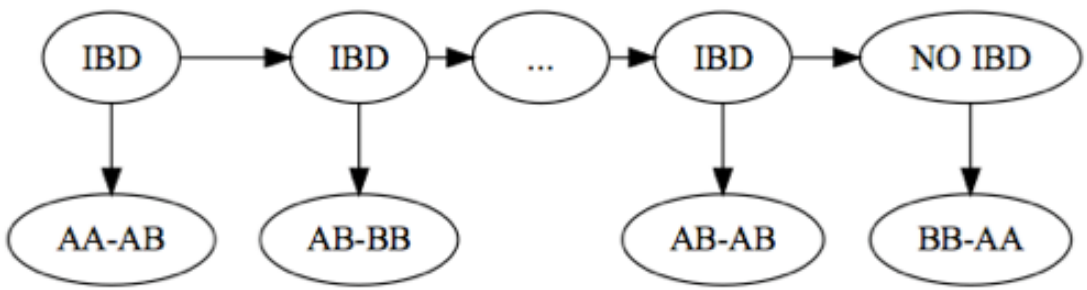
Hidden Markov Models

Lander & Green









Detection of IBD segments using an HMM with the emission probabilities:

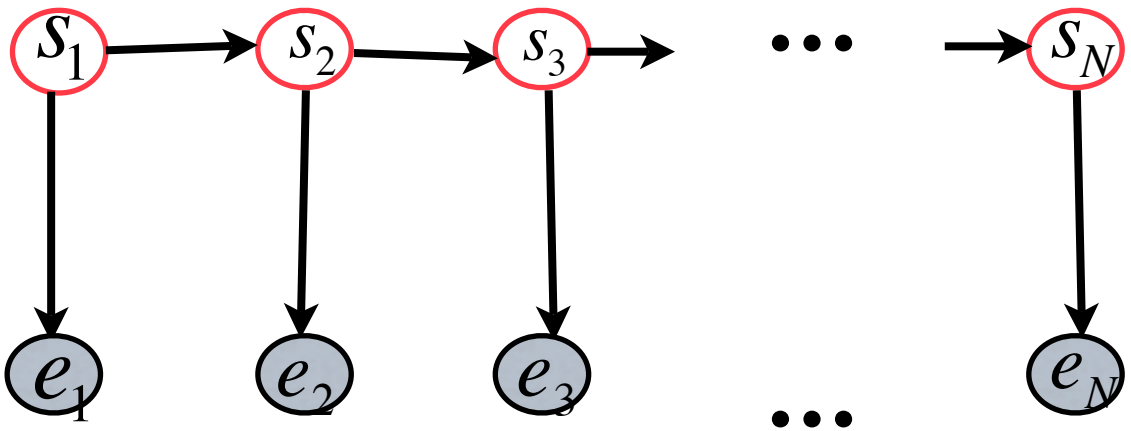
Genotype	{AA,AA}	{AA,AB}	{AA,BB}	{AB,AB}	{AB,BB}	{BB,BB}
NO IBD	p^4	$4p^3q$	$2p^2q^2$	$4p^2q^2$	$4pq^3$	q^4
IBD	p^3	$2p^2q$	0	$p^2q + pq^2$	$2pq^2$	q^3

Hidden Markov Model, HMM

$$\pi_{s_1} = P(s_1)$$

$$P_i^j$$

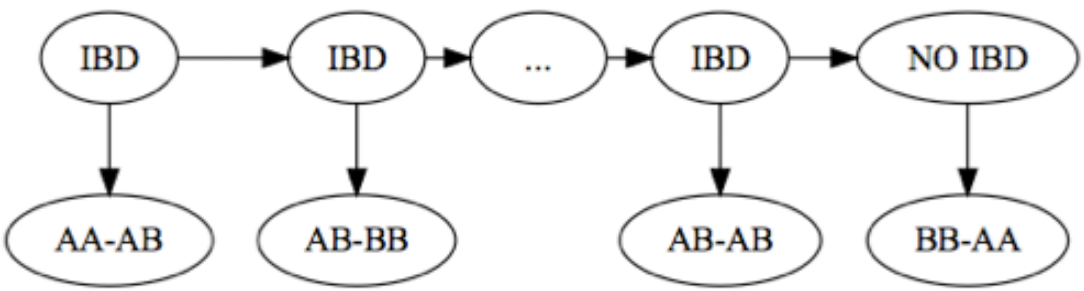
Markov Chain



$$E_{s_t, e_t} = P(e_t \mid s_t)$$

Emissions

Inference: Given an emission sequence find the most likely state sequence.



MatLab Example

First we can fake some data...

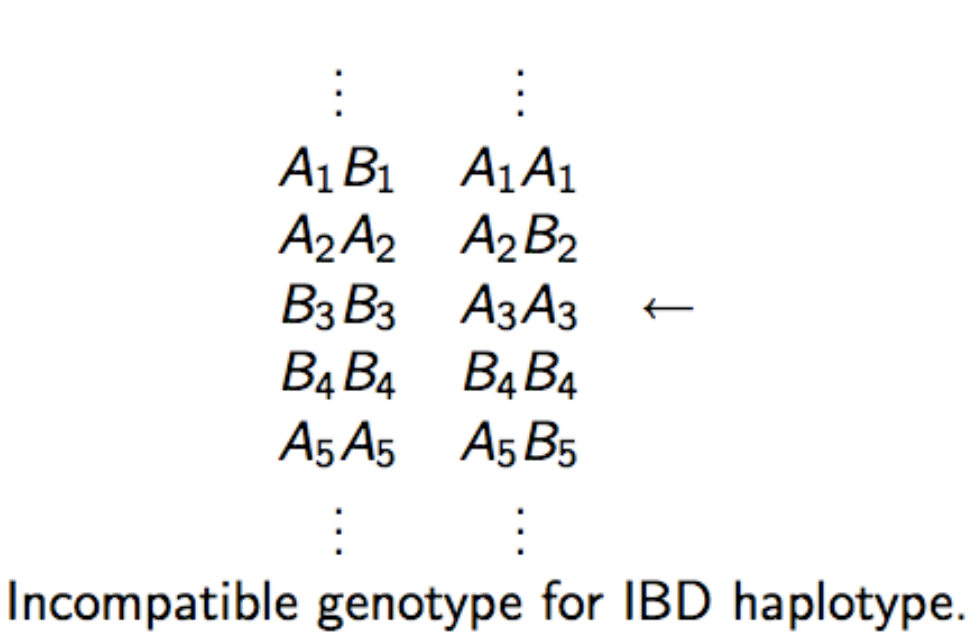
```
trans = [0.95,0.05;  0.10,0.90];
 emis = [   1/6  1/6   1/6   1/6   1/6 1/6;
            1/10 1/10 1/10 1/10 1/10 1/2];
[seq,states] = hmmgenerate(20,trans,emis)
```

Inference: Given an emission sequence find the most likely state sequence.

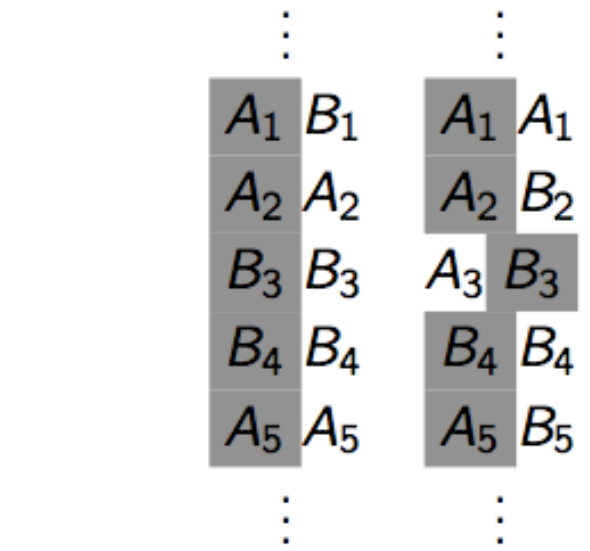
The Viterbi Algorithm

```
STATES = hmmviterbi(seq,trans,emis)
```

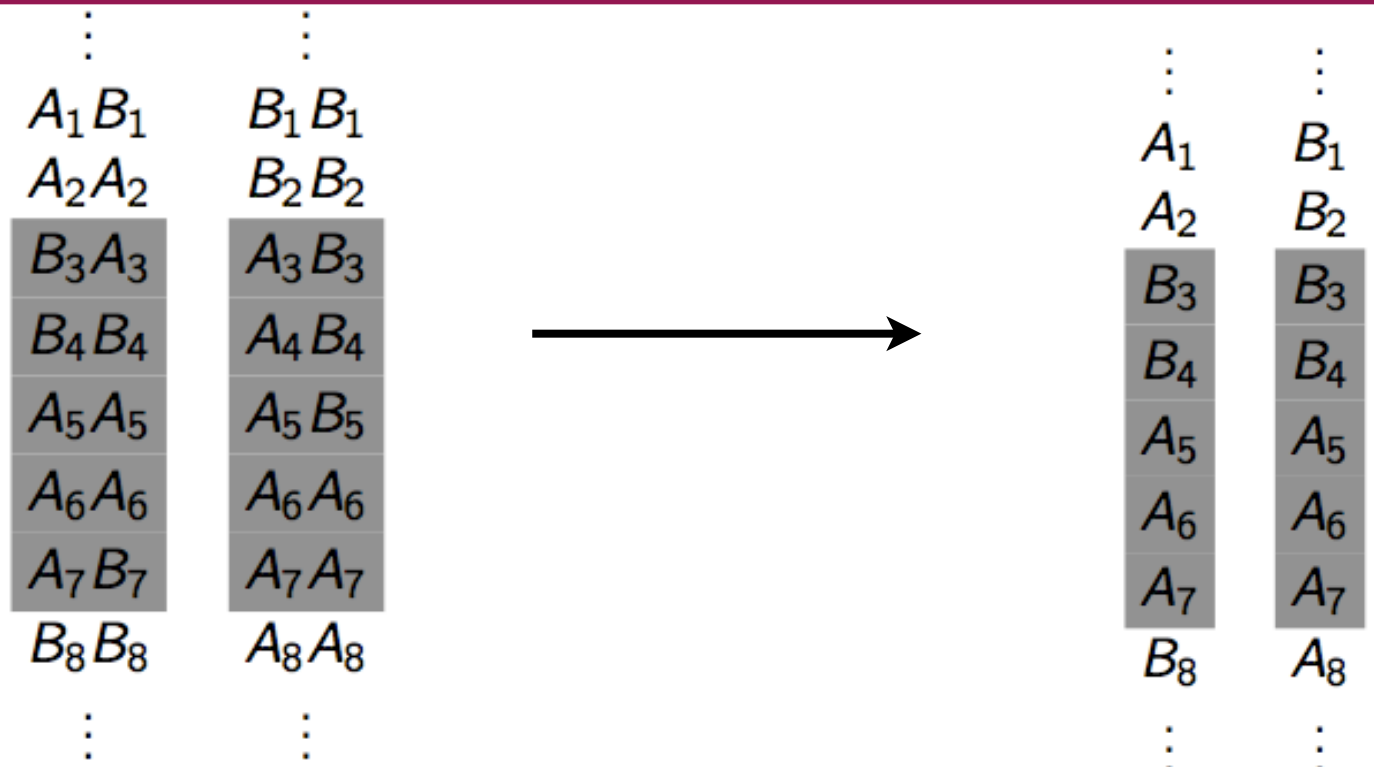
Haplotype vs Genotype



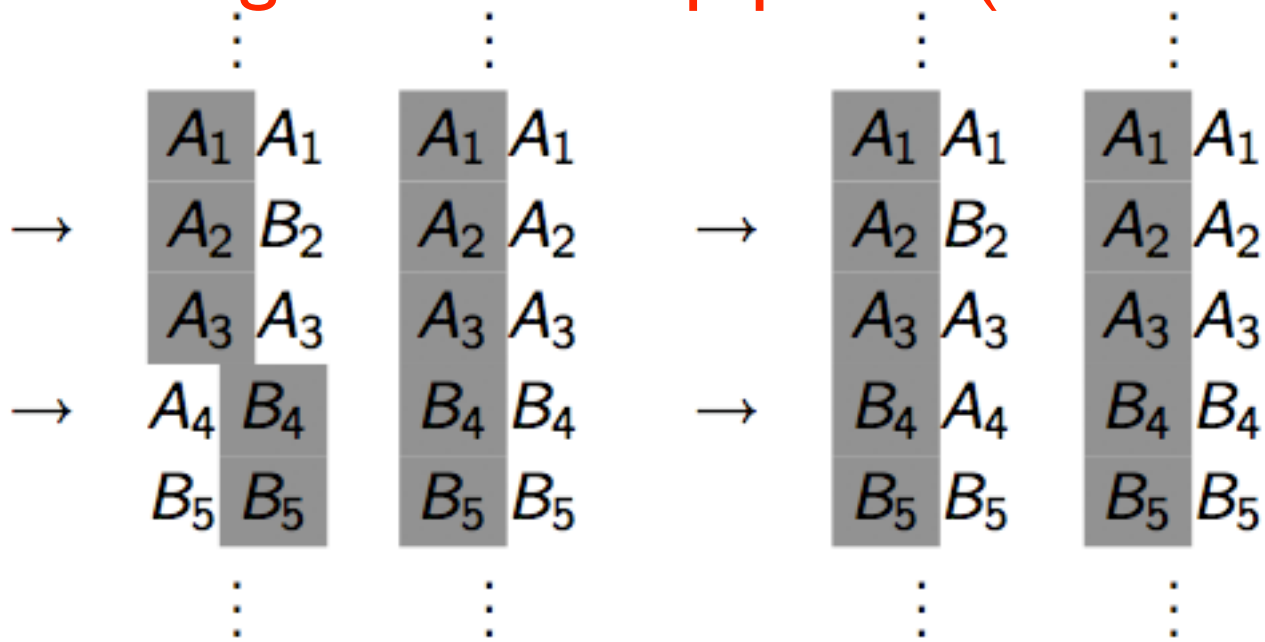
Incompatible genotype for IBD haplotype.



Potential IBD haplotype shared



IBD regions can help phase (MAX GEN2SAT)

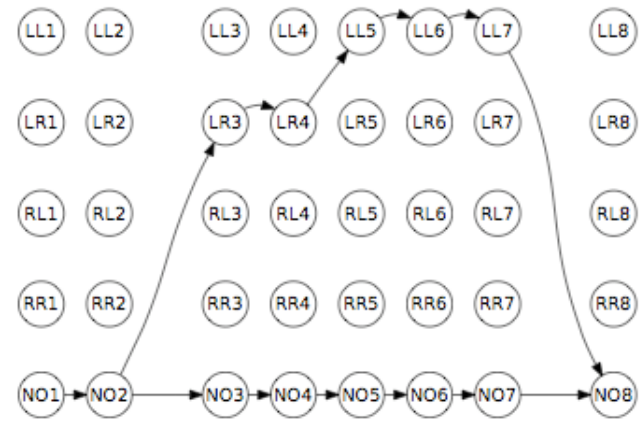


(k) Before phasing. (l) After phasing.

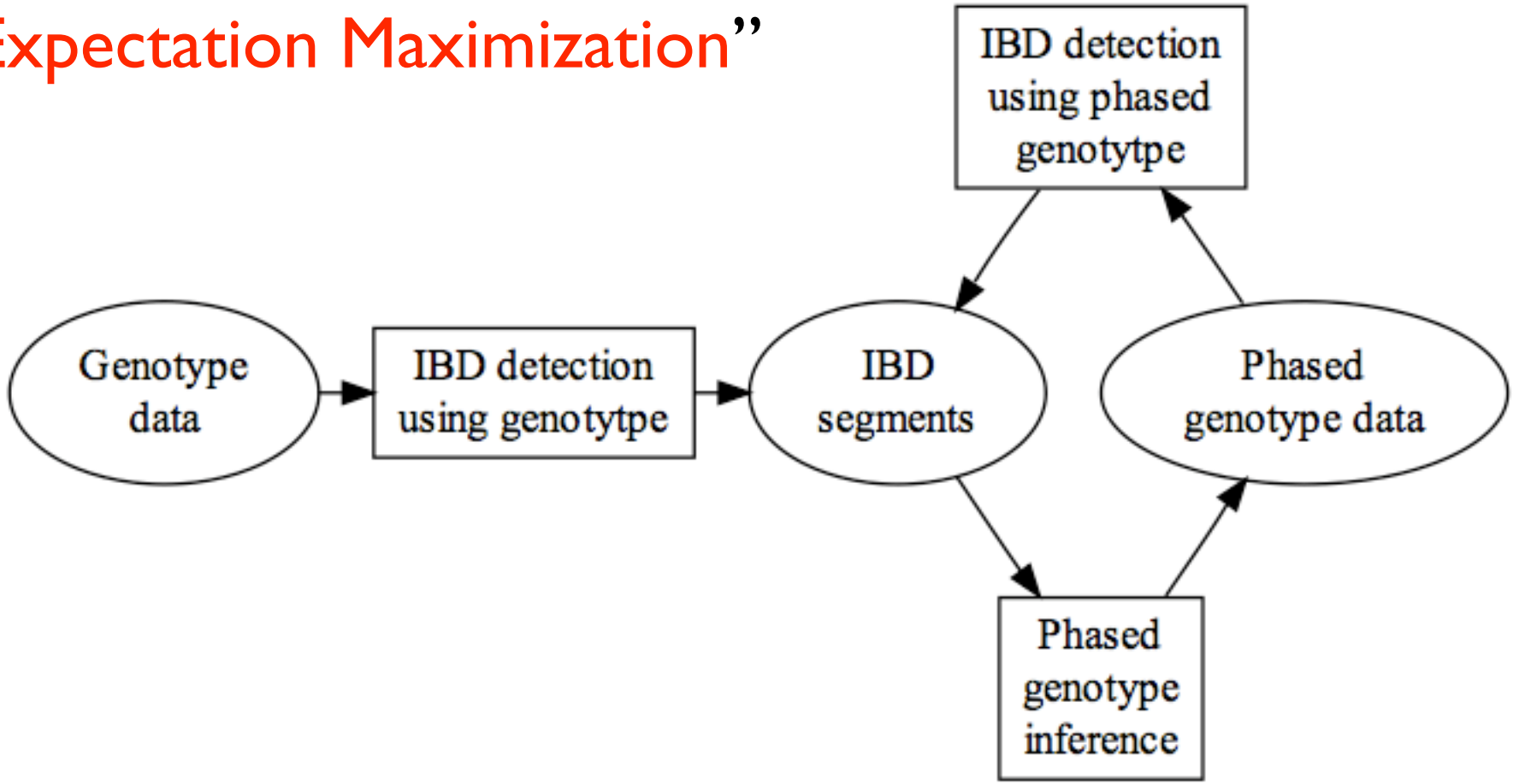
HMM with haplotype information

Haplotype	NO IBD	IBD LL	IBD LR	IBD RL	IBD RR
(AA,AA)	p^4	p^3	p^3	p^3	p^3
(AA,AB)	p^3q	p^2q	0	p^2q	0
(AA,BA)	p^3q	0	p^2q	0	p^2q
(AA,BB)	p^2q^2	0	0	0	0
(AB,AA)	p^3q	p^2q	p^2q	0	0
(AB,AB)	p^2q^2	pq^2	0	0	p^2q
(AB,BA)	p^2q^2	0	pq^2	p^2q	0
(AB,BB)	pq^3	0	0	pq^2	pq^2
(BA,AA)	p^3q	0	0	p^2q	p^2q
(BA,AB)	p^2q^2	0	p^2q	pq^2	0
(BA,BA)	p^2q^2	p^2q	0	0	pq^2
(BA,BB)	pq^3	pq^2	pq^2	0	0
(BB,AA)	p^2q^2	0	0	0	0
(BB,AB)	pq^3	0	pq^2	0	pq^2
(BB,BA)	pq^3	pq^2	0	pq^2	0
(BB,BB)	q^4	q^3	q^3	q^3	q^3

..and phasing allows for better IBD detection



"Expectation Maximization"



Improved IBD detection using incomplete haplotype information

G. Genovese[1], G. Leibon[1], D. Rockmore[1], M.R. Pollak[2]
[1] *Department of Mathematics, Dartmouth College, 6188 Kemeny Hall, Hanover NH 03755, United States,*
[2] *Renal Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston MA 02115, United States*

* a special thanks to Giulio Genovese for letting me pilfer his slides, thanks!