CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Apr 20

- Finish population genetics review (what does Tajima's d mean?)
- Lab 8 analysis recap
- HMM review + Viterbi worksheet

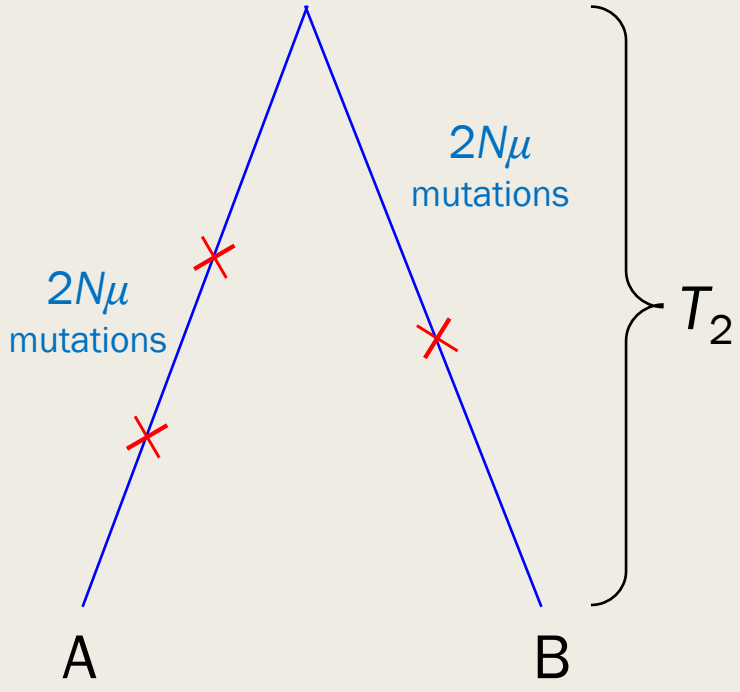
Deviations from neutrality: Tajima's D

Expected value of π (average pairwise heterozygosity)

- Let μ be the per site, per generation mutation rate
- Considering two samples, the expected time to coalescence is 1 coalescent unit or $2N$ generations
- Therefore the expected number of mutations separating the two samples is

$$E[\pi] = 4N\mu$$

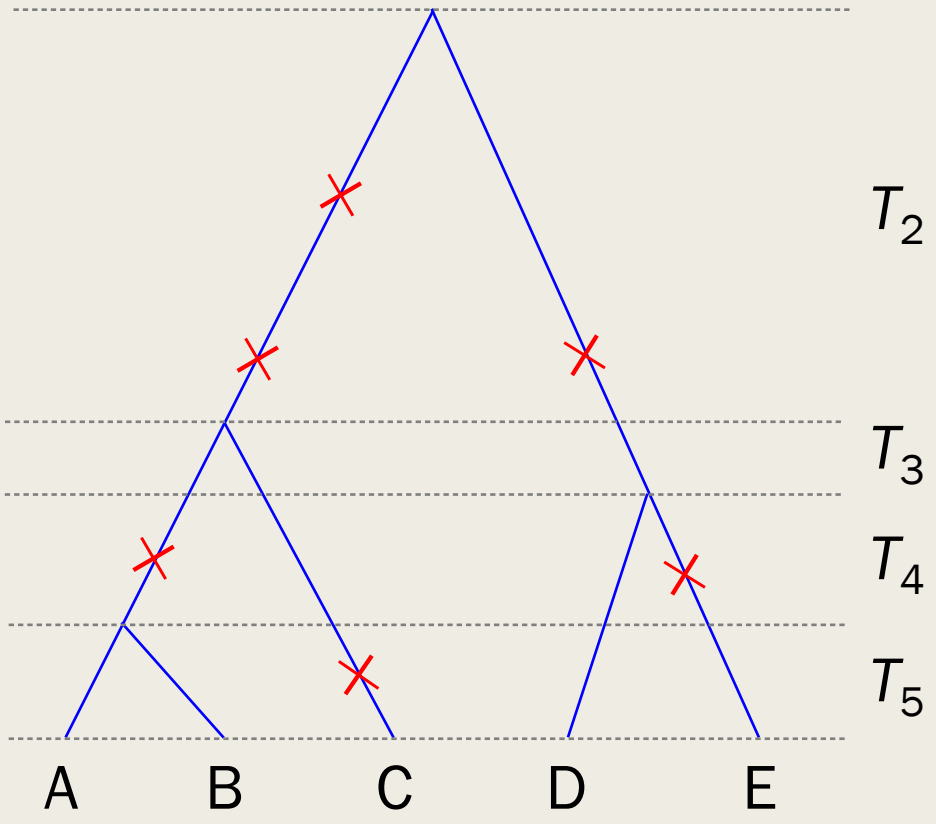
- Need to multiply by L if considering more than one site



Expected value of S (number of segregating sites)

- For E[S], we need to compute the total branch length

$$T_{\text{total}} = \text{total length of all branches in the tree}$$

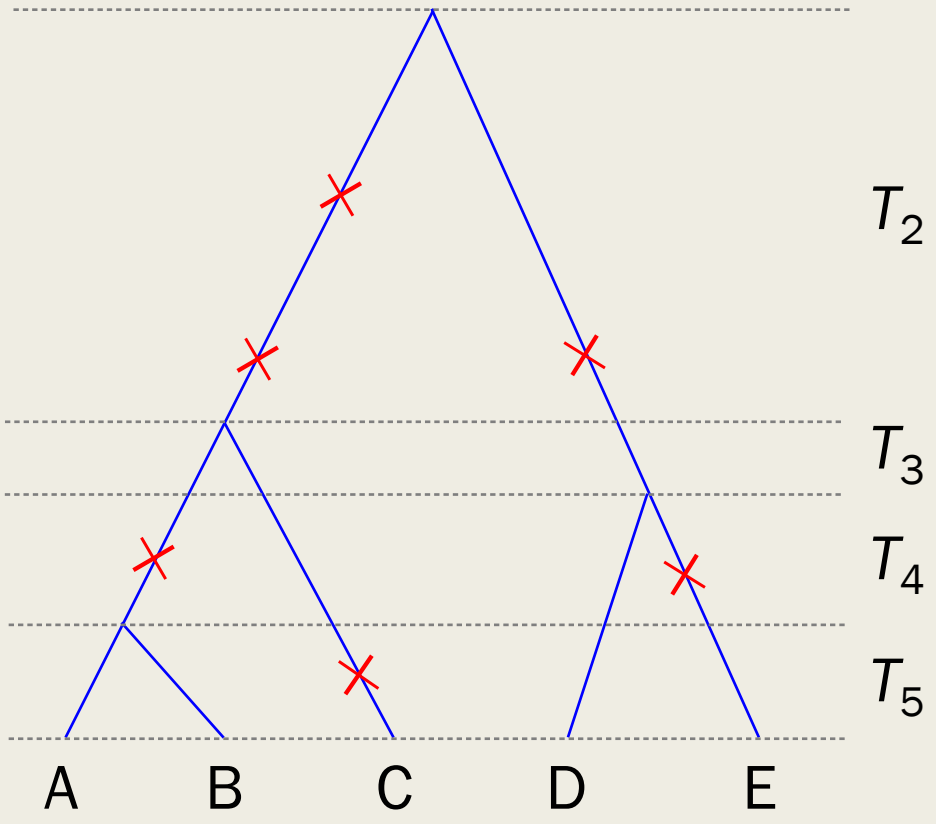


Expected value of S (number of segregating sites)

- For E[S], we need to compute the total branch length

$$T_{\text{total}} = \text{total length of all branches in the tree}$$

$$E[T_{\text{total}}] = \sum_{i=n}^{2} E[T_i] \cdot i$$



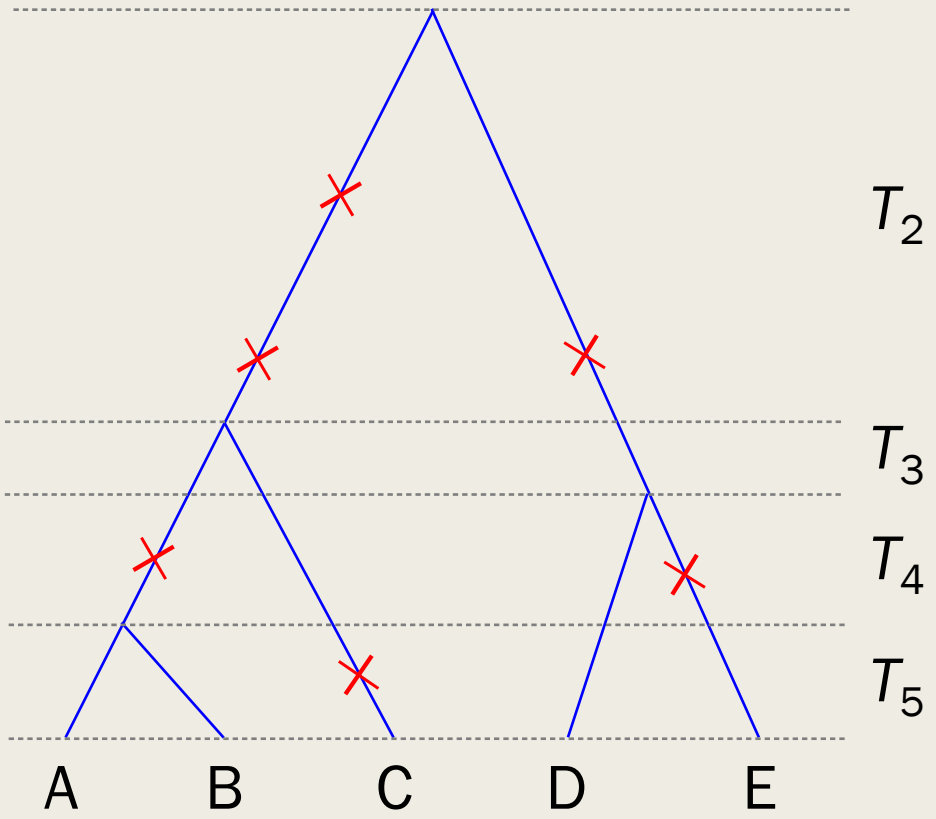
Expected value of S (number of segregating sites)

- For E[S], we need to compute the total branch length

$$T_{\text{total}} = \text{total length of all branches in the tree}$$

$$E[T_{\text{total}}] = \sum_{i=n}^{2} E[T_i] \cdot i$$

$$= \sum_{i=n}^{2} \frac{2}{\cancel{i}(i-1)} \cdot \cancel{i}$$



Expected value of S (number of segregating sites)

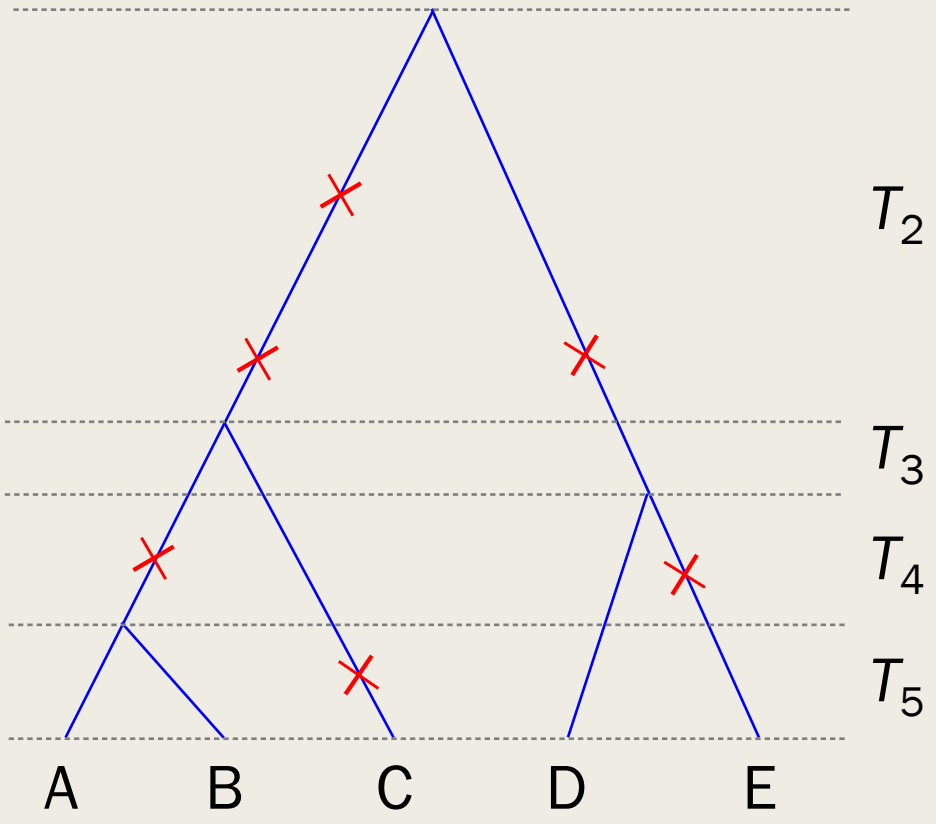
- For E[S], we need to compute the total branch length

$$T_{\text{total}} = \text{total length of all branches in the tree}$$

$$E[T_{\text{total}}] = \sum_{i=n}^{2} E[T_i] \cdot i$$

$$= \sum_{i=n}^{2} \frac{2}{\cancel{i}(i-1)} \cdot \cancel{i}$$

$$= 2 \sum_{i=1}^{n-1} \frac{1}{i}$$



Expected value of S (number of segregating sites)

- For E[S], we need to compute the total branch length

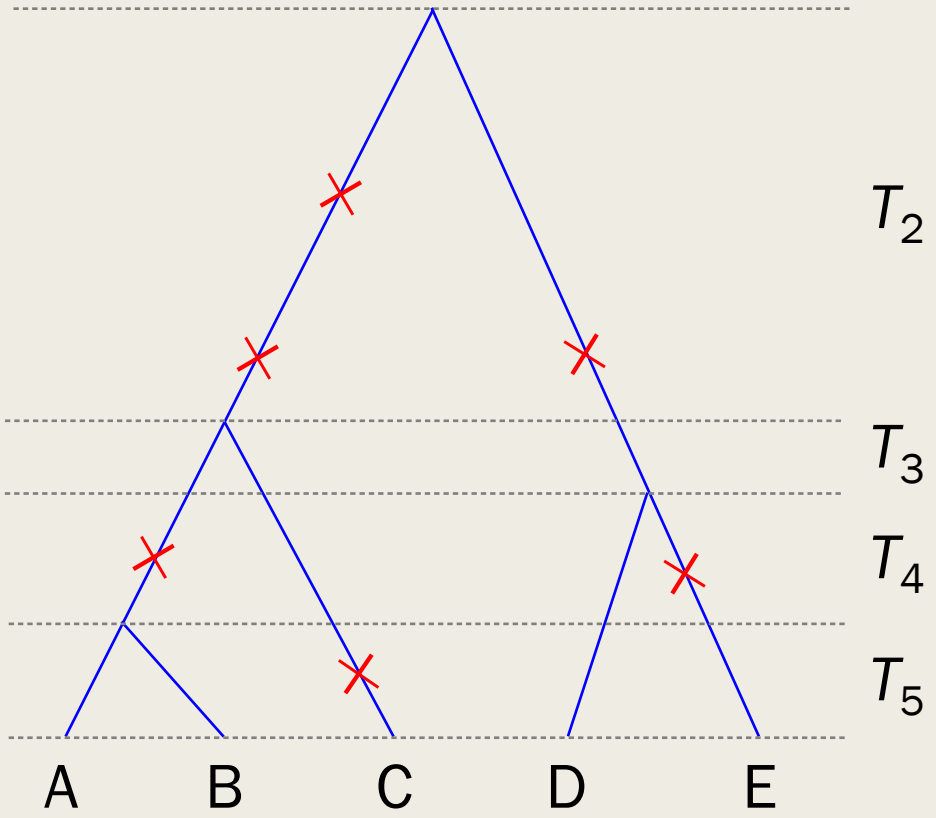
$$T_{\text{total}} = \text{total length of all branches in the tree}$$

$$E[T_{\text{total}}] = \sum_{i=n}^{2} E[T_i] \cdot i$$

$$= \sum_{i=n}^{2} \frac{2}{\cancel{i}(i-1)} \cdot \cancel{i}$$

$$= 2 \sum_{i=1}^{n-1} \frac{1}{i}$$

$$= 2a_1$$



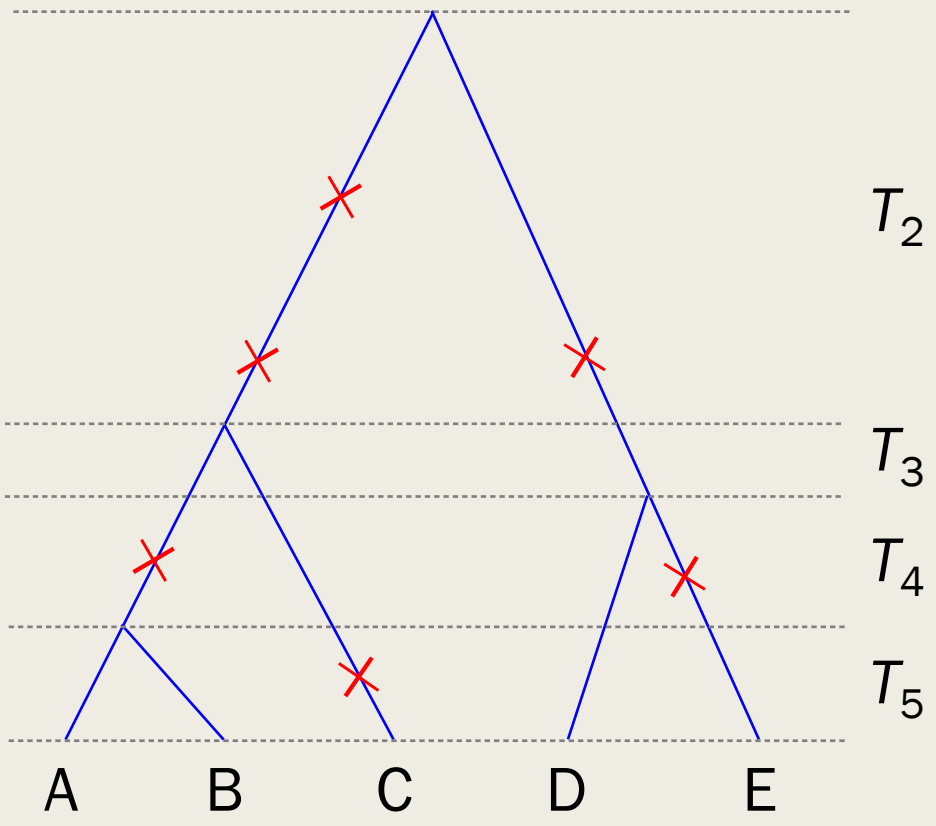
Expected value of S (number of segregating sites)

- After we have the total branch length, we can multiple by $2N\mu$, the rate of mutations per unit of coalescent time

$$E[S] = E[T_{\text{total}}] \cdot (2N\mu)$$

- We can simplify this to get an expression similar to the expected value for π

$$E[S] = 4N\mu \cdot a_1$$



Putting this together, we get Tajima's d

- We will consider lowercase *d*, whose expectation is E[*d*] = 0

$$d = \pi - S/a_1$$

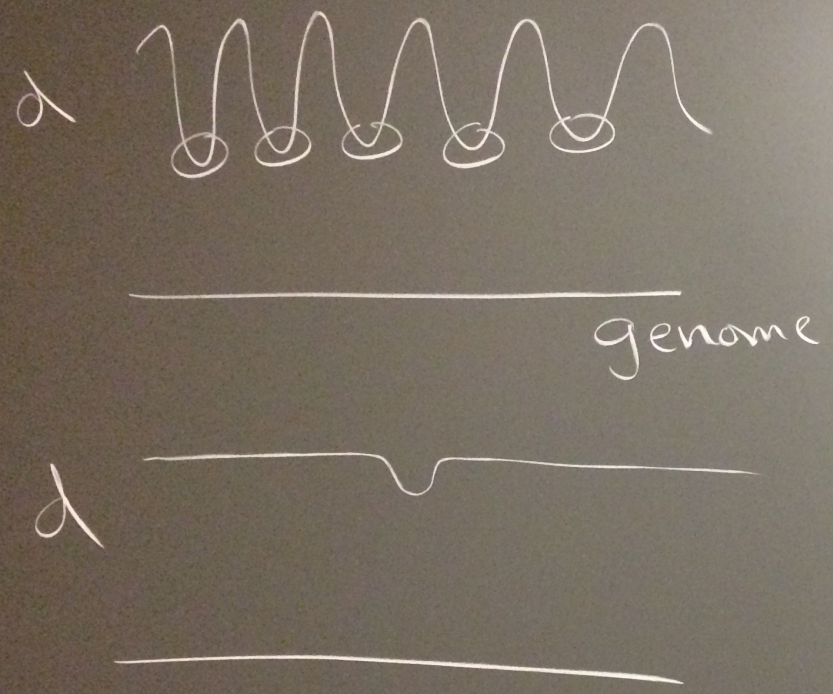
- Tajima's (capital) *D* is defined as:

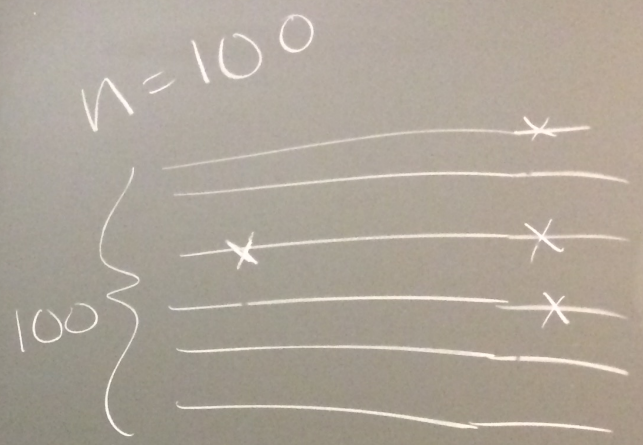
$$D = \frac{d}{\sqrt{\mathrm{Var}(d)}}$$

- We will mainly focus on the sign of *d* so we'll ignore the denominator

Takeaways from yesterday

- All variation contributes to *S*
- Rare variation contributes less to π
- Middle frequency variation (i.e. “common”) contributes more to π





rare :

common

less

a

$\underbrace{99 \cdot 1}_{\text{Contribution to pi}}$ $\underbrace{50 \cdot 50}_{\text{Contribution to pi}}$

99 $\boxed{2500}$

What do deviations from *d*=0 mean?

- If d is close to 0, neutral expectations (probably) hold (i.e. constant population size, random mating, no natural selection)
- If d > 0, the pairwise heterozygosity is higher than we expect relative to the number of segregating sites => excess of **middle** frequency SNPs
- If d < 0, the number of segregating sites is higher than we expect relative to the pairwise heterozygosity => excess of **rare** variation

What do deviations from *d*=0 mean?

- If d is close to 0, neutral expectations (probably) hold (i.e. constant population size, random mating, no natural selection)

- If d > 0, the pairwise heterozygosity is higher than we expect relative to the number of segregating sites => excess of **middle** frequency SNPs

  - Bottleneck or population decline
  - Population structure or isolation with migration

- If d < 0, the number of segregating sites is higher than we expect relative to the pairwise heterozygosity => excess of **rare** variation

What do deviations from *d*=0 mean?

- If d is close to 0, neutral expectations (probably) hold (i.e. constant population size, random mating, no natural selection)

- If d > 0, the pairwise heterozygosity is higher than we expect relative to the number of segregating sites => excess of **middle** frequency SNPs

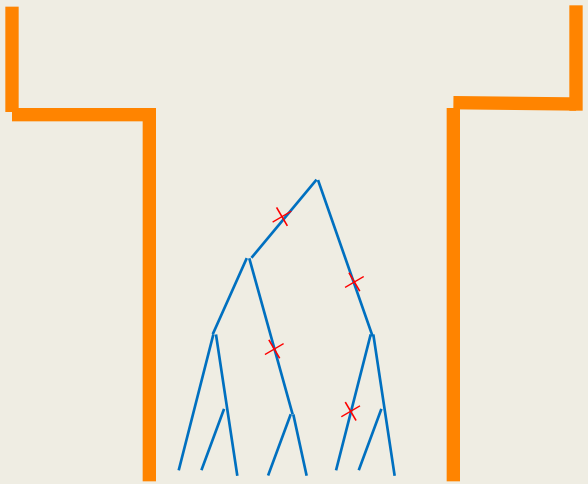
  - Bottleneck or population decline
  - Population structure or isolation with migration

- If d < 0, the number of segregating sites is higher than we expect relative to the pairwise heterozygosity => excess of **rare** variation

  - Population growth
  - Natural selection

Tajima's d > 0

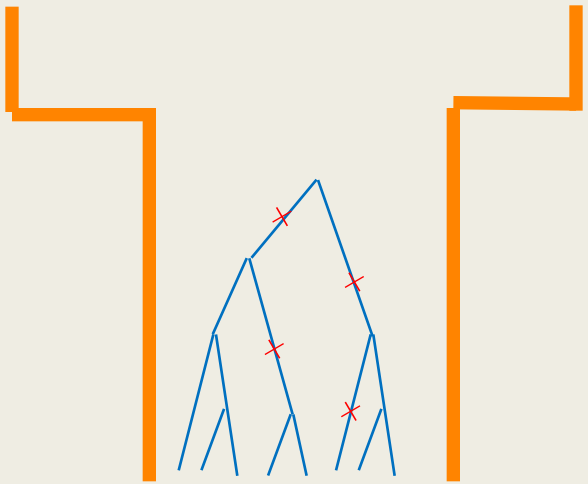
Population size & structure



- Population decay or recent bottleneck
- In the recent past, find common ancestors quickly
- Small T_{total} (tree size) => small S

Tajima's d > 0

Population size & structure

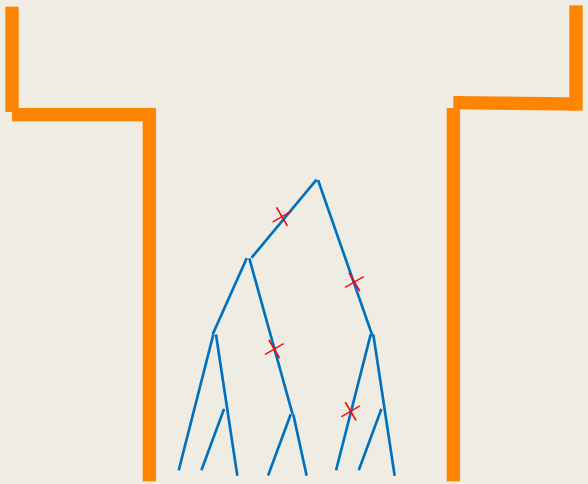


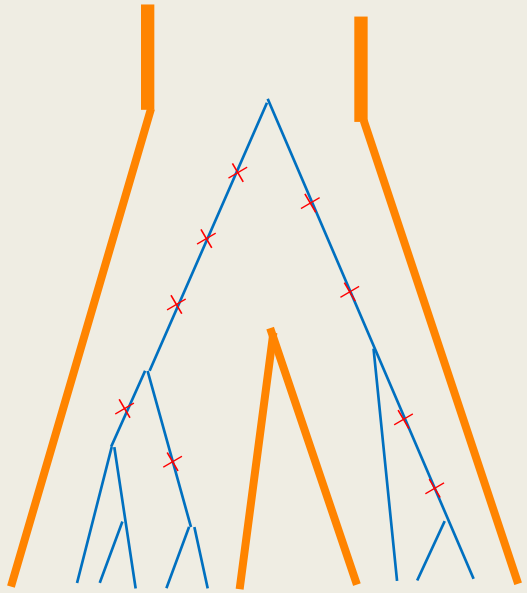
- Population decay or recent bottleneck
- In the recent past, find common ancestors quickly
- Small T_{total} (tree size) => small S
- d is positive because S is small

$$d = \pi - S/a_1$$

Tajima's d > 0

Population size & structure





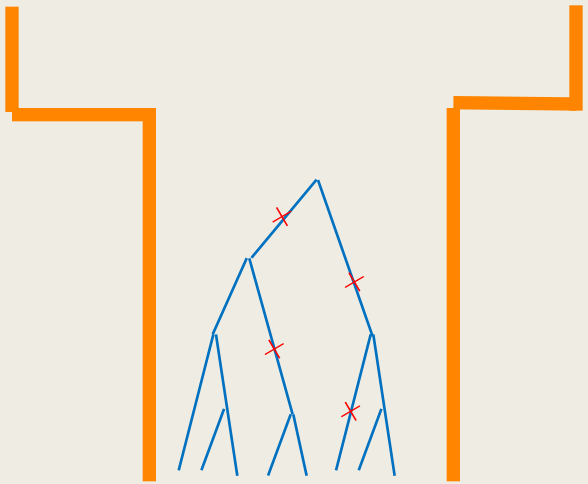
- Population decay or recent bottleneck
- In the recent past, find common ancestors quickly
- Small T_{total} (tree size) => small S
- d is positive because S is small

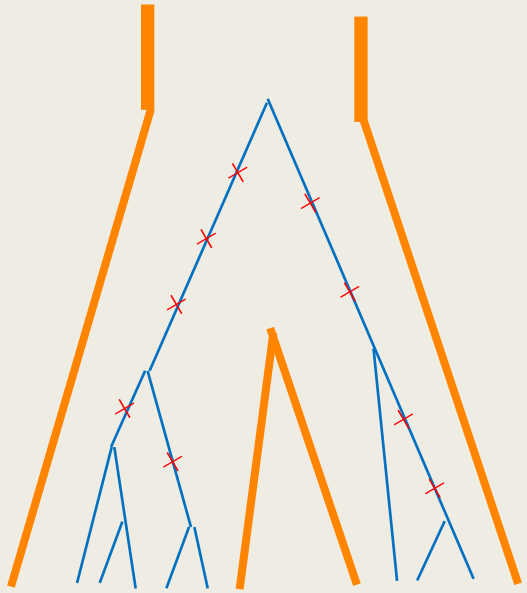
$$d = \pi - S/a_1$$

- Population structure (i.e. splits, non-random mating)
- Within population: find common ancestors quickly
- Long time to find common ancestors across pops
- Excess of common variation in upper branches

Tajima's d > 0

Population size & structure





- Population decay or recent bottleneck
- In the recent past, find common ancestors quickly
- Small T_{total} (tree size) => small S
- d is positive because S is small

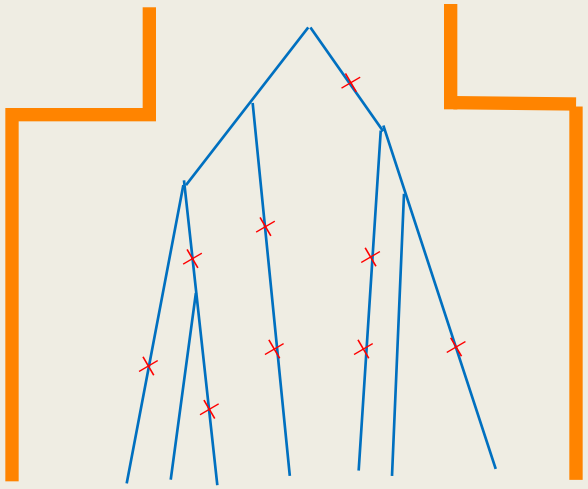
$$d = \pi - S/a_1$$

- Population structure (i.e. splits, non-random mating)
- Within population: find common ancestors quickly
- Long time to find common ancestors across pops
- Excess of common variation in upper branches
- d is positive because π is large

$$d = \pi - S/a_1$$

Tajima's d < 0

Population size & structure

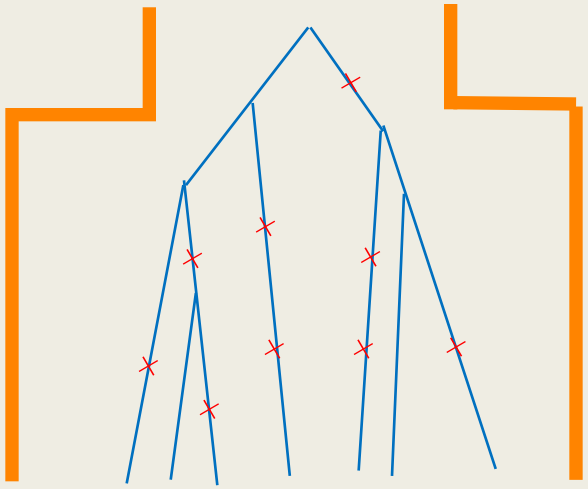


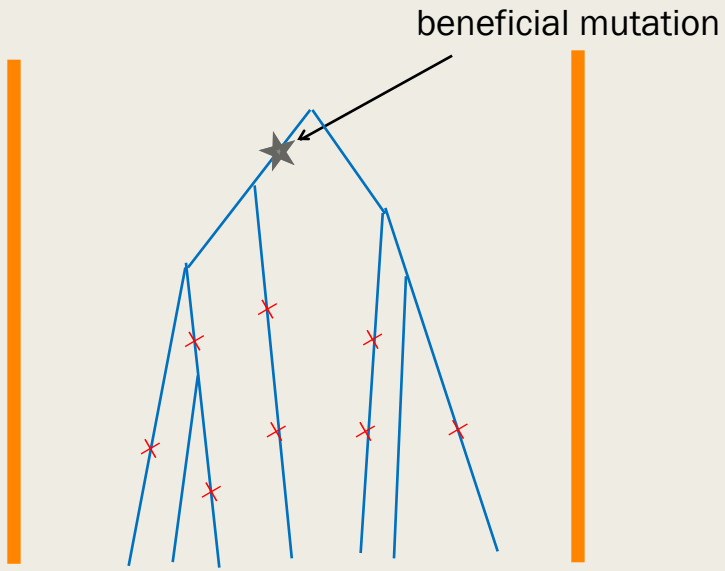
- Population growth
- Difficult to find common ancestors in recent past
- Excess of rare variation on terminal branches
- π could be large, but S is larger (relatively)

$$d = \pi - S/a_1$$

Tajima's d < 0

Population size & structure





- Population growth
- Difficult to find common ancestors in recent past
- Excess of rare variation on terminal branches
- π could be large, but S is larger (relatively)

$$d = \pi - S/a_1$$

- Natural selection
- Similar pattern to growth, but only in a single region!
- *d* more negative relative to background rate

$$d = \pi - S/a_1$$

Tajima’s D in practice

Example of Tajima's D from the literature

Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA

- Why is Tajima's D greater than 0?
- Hypothesis: bottleneck in European and Asian populations is still affecting patterns of variation
- Population structure is playing a role in African populations

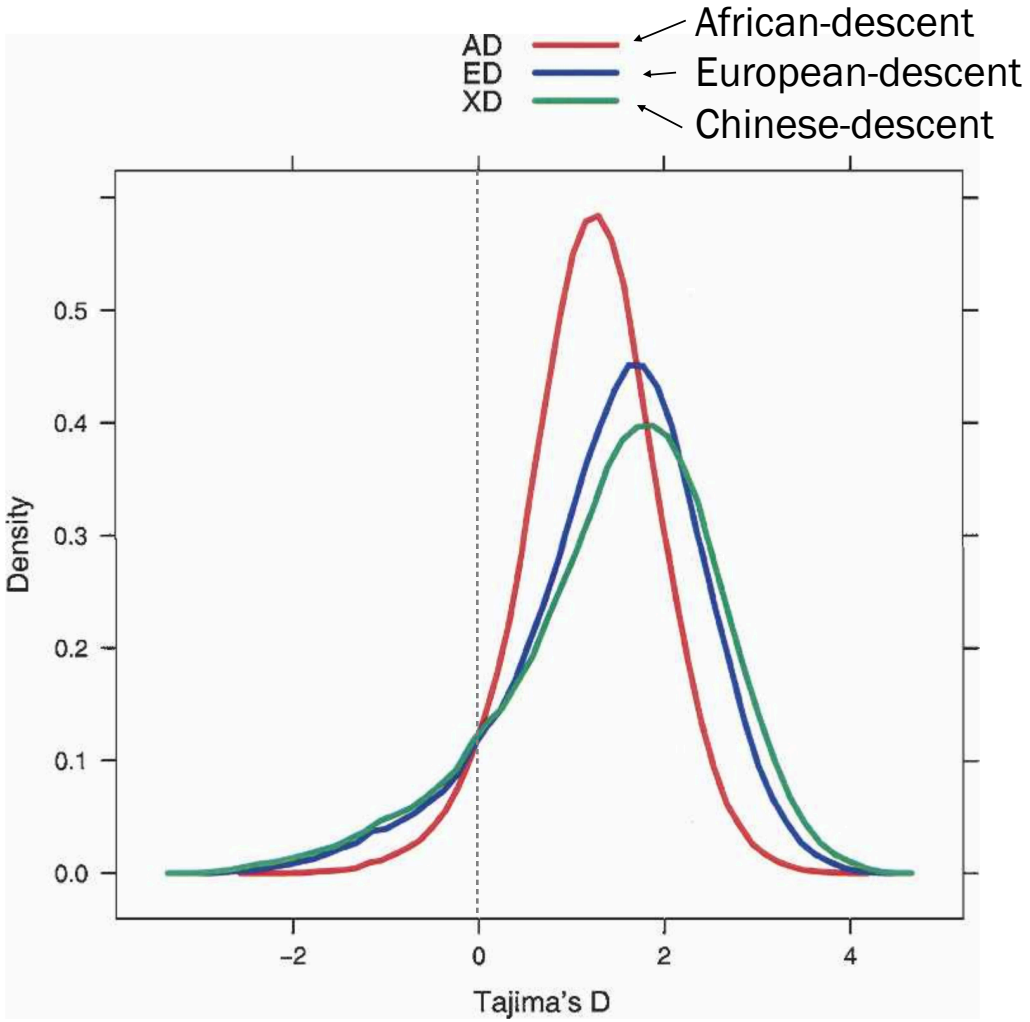
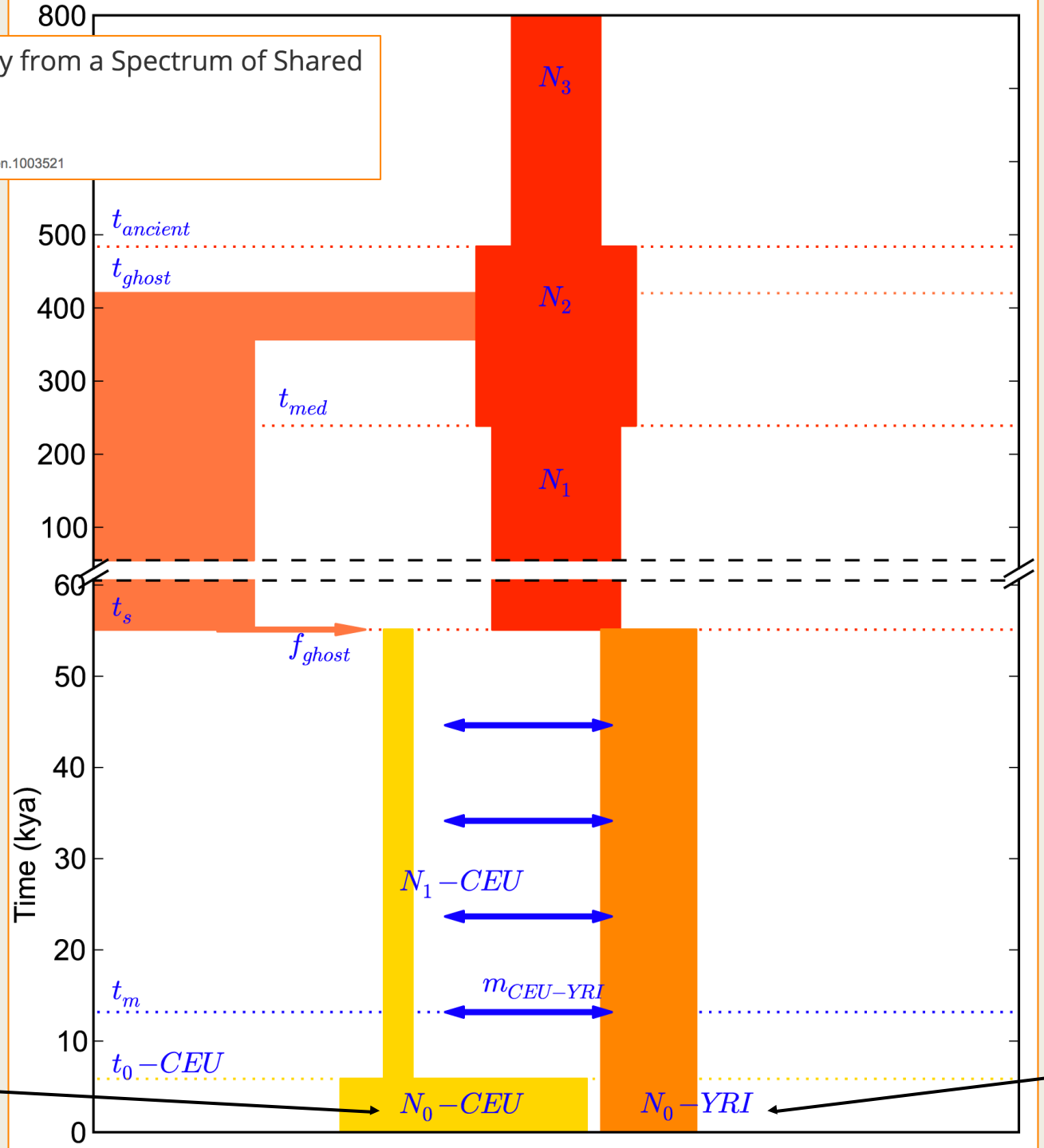


Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

Inferring Demographic History from a Spectrum of Shared Haplotype Lengths

Kelley Harris, Rasmus Nielsen

Published: June 6, 2013 • https://doi.org/10.1371/journal.pgen.1003521



Method that builds on the idea of deviation from neutrality

CEU: European population

YRI: African population

Lab 8 Analysis

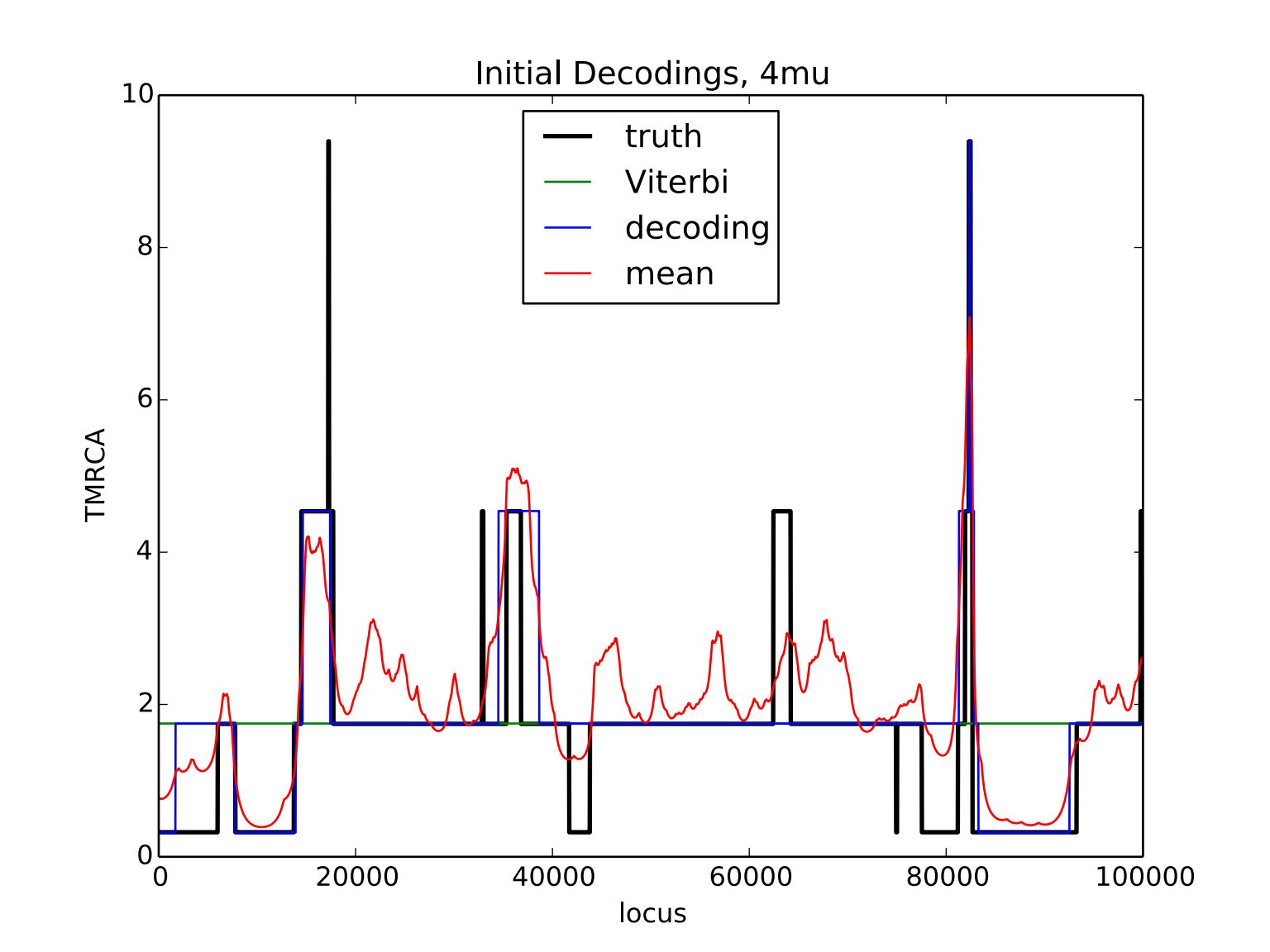
Analysis Questions

1.What general observations do you make about the differences/similarities between our 3 ways of estimating the hidden state sequence (Viterbi, decoding, mean)? Which would you choose if you could only pick one?

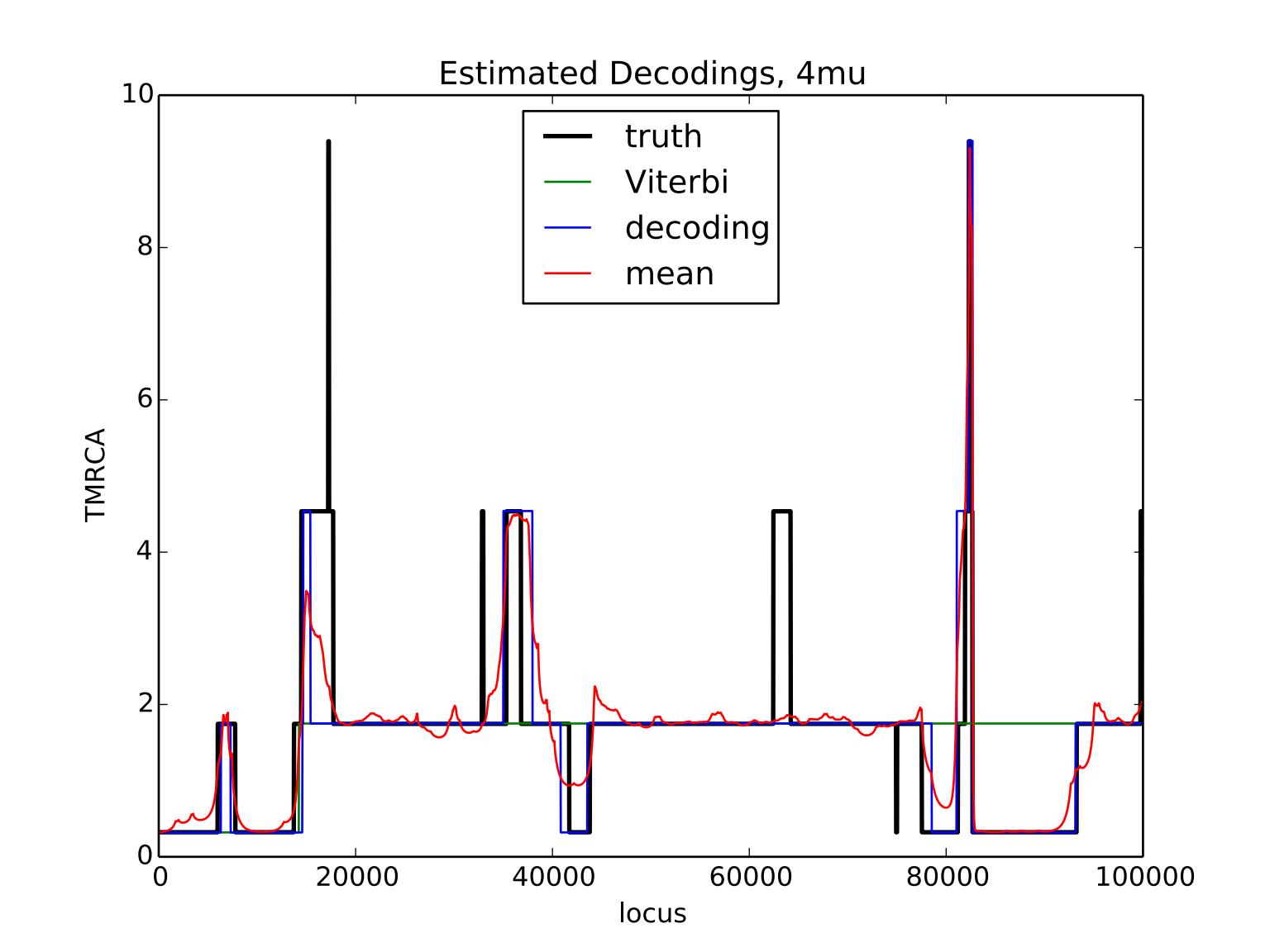
2.For the two true Tmrca sequences (true_tmrca_test.txt and true_tmrca.txt), how did the estimated state sequences change between using the "initial" parameters and using the "estimated" parameters? Why might we observe this difference? How did the log-likelihoods change for these two datasets?

3.For the true_tmrca.txt dataset, what was the difference in your output plots between mu, 2mu, and 5mu? Why might we observe this difference as the mutation rate increases?

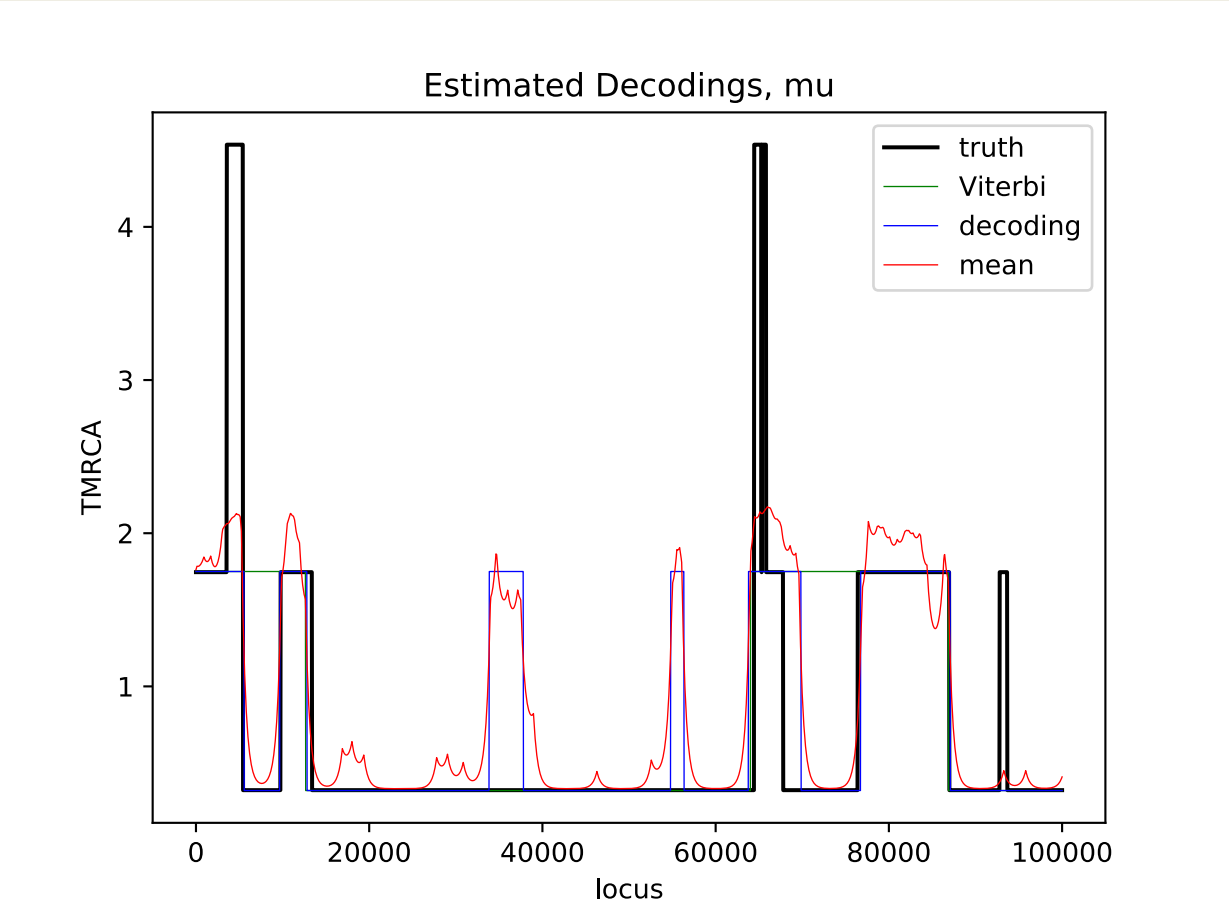
4mu: initial decodings



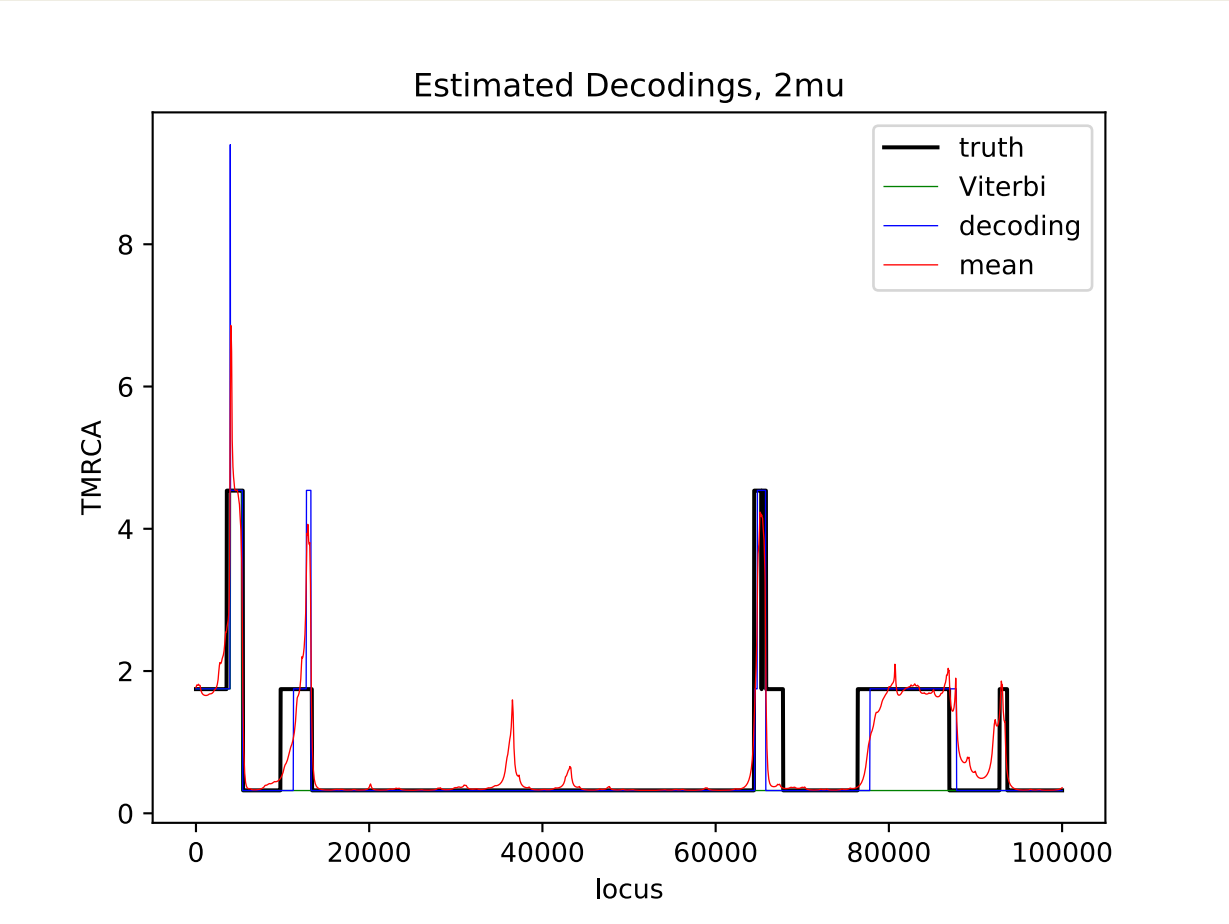
4mu: estimated decodings



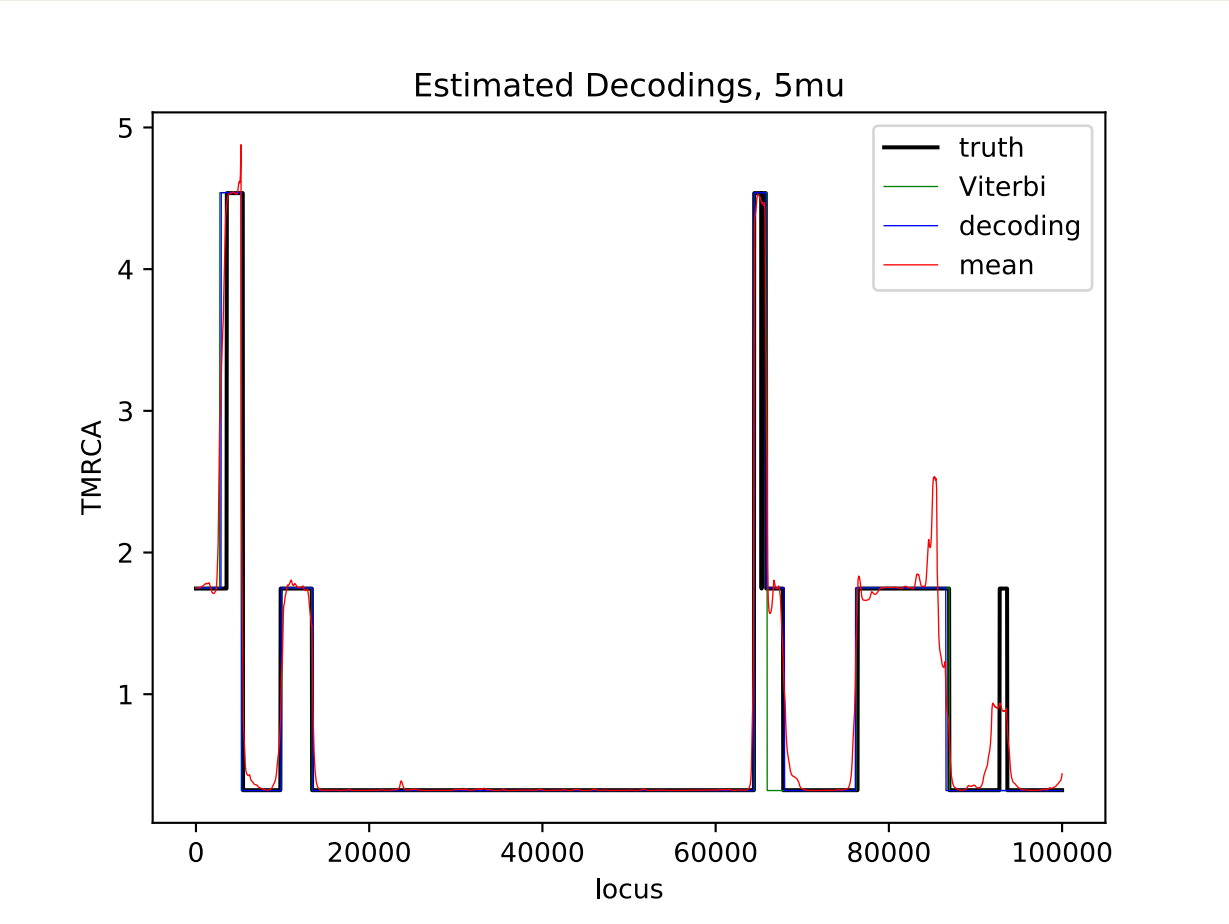
mu estimated



2mu estimated



5mu estimated



rare : MAF < 0.05

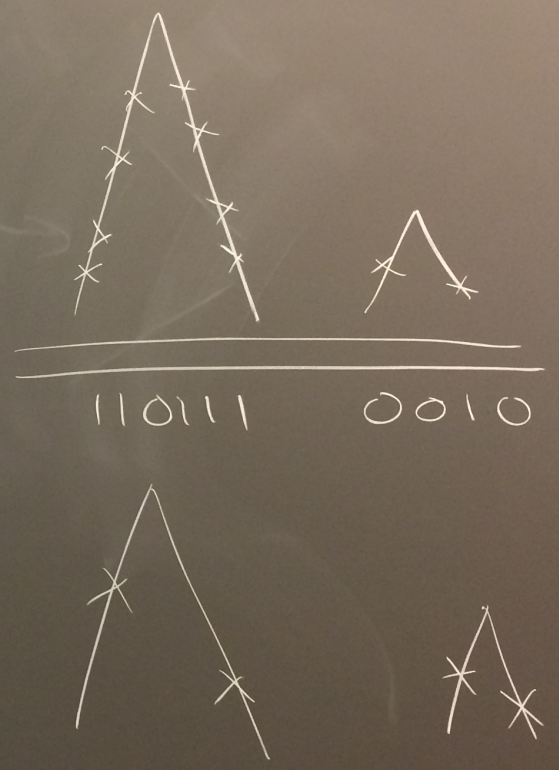
common: MAF > 0.05

less common (minor) allele frequency

50

contribution to pi

2500



Handout 26

$a^{(t)}$	1	2
1	1/2	1/2
2	1/5	4/5

e	H	T
1	2/3	1/3
2	1/4	3/4

π	
1	1/2
2	1/2

① $V_1(1) = \pi_1 e_1(H) =$

$V_2(1) = \pi_2 e_2(H) =$

② $V_1(2) = e_1(T) \cdot m$

$= \frac{1}{3} ma$

$= \frac{1}{18}$

① $\vec{x}$

	H	T	H
1	1/3	1/18	1/54
2	1/8	1/8	1/40

V?

② z: 1 2 2

② z: 2 1 2 1 1 1 2 2 2 2 2 2 2 2 1 2 2

x: T H H H T H T T H T T T T H H T T

L=17

a e ?

① $V_1(1) = \pi_1 e_1(H) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$

$V_2(1) = \pi_2 e_2(H) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$

② $V_1(2) = e_1(T) \cdot \max\{V_1(1) a_{11}, V_2(1) a_{21}\}$

$= \frac{1}{3} \max\{\frac{1}{3} \cdot \frac{1}{2}, \frac{1}{8} \cdot \frac{1}{5}\}$

$= \frac{1}{18}$

2 2 2 → 1 → 2 2

T T H H T T

A	1	2
1	2	3
2	3	8

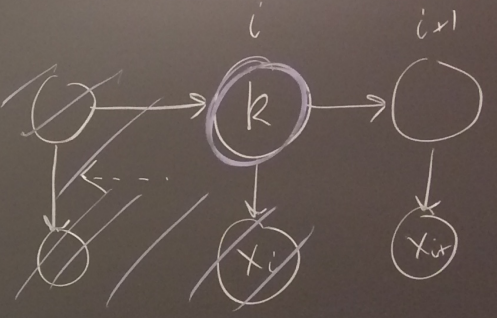
E-step

$a^{(t+1)}$	1	2
1	2/5	3/5
2	3/11	8/11

M-step

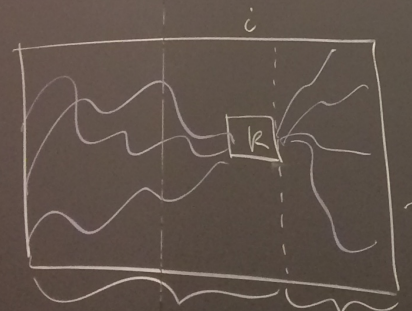
Case 2: hidden sequence unknown

$$\boxed{P(z_i = k \mid \vec{x})} = \frac{P(\vec{x}, z_i = k)}{P(\vec{x})} \qquad \left(P(B \mid A) = \frac{P(A,B)}{P(A)} \right)$$



$$= P(x_1 \cdots x_i, z_i = k, x_{i+1} \cdots x_L) / P(\vec{x})$$

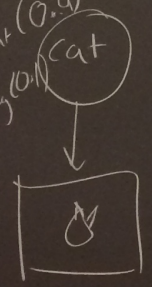
$$= \underbrace{P(x_1 \cdots x_i, z_i = k)}_{f_k(i)} \cdot \underbrace{P(x_{i+1} \cdots x_L \mid \cancel{x_1 \cdots x_i}, z_i = k)}_{b_k(i)} / P(\vec{x})$$

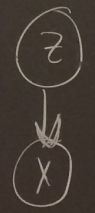


$$\boxed{= \frac{f_k(i)\, b_k(i)}{P(\vec{x})}}$$

posterior prob

$P_{cat}(0.9)$
$P_{dog}(0.1)$





$P_{\pi, t, e}(\vec{x})$