



CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018



Outline: Mar 28

- Finish Tajima's D and population genetics
- Example of how Tajima's D can be used in practice
- Begin: HMMs (Hidden Markov Models)
- Today: Markov chains

Notes:

- Office hours TODAY 1-3pm
- Lab 6 due tonight (late days are an option!)

Technical summer opportunity on campus:

Swarthmore Projects for Educational Exploration and Development (SPEED) <https://goo.gl/EBv42o>

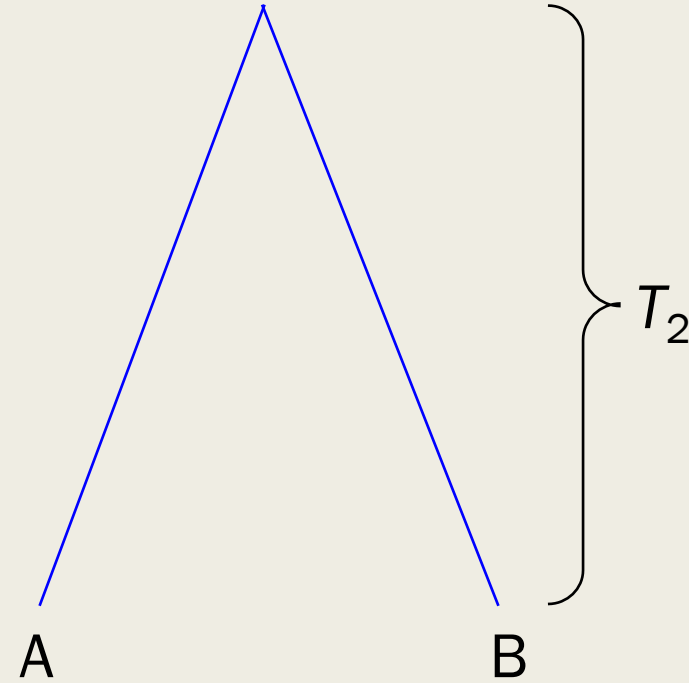
Deviations from neutrality: Tajima's D

Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- For now we will consider a single site
- Let μ be the per site, per generation mutation rate

Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

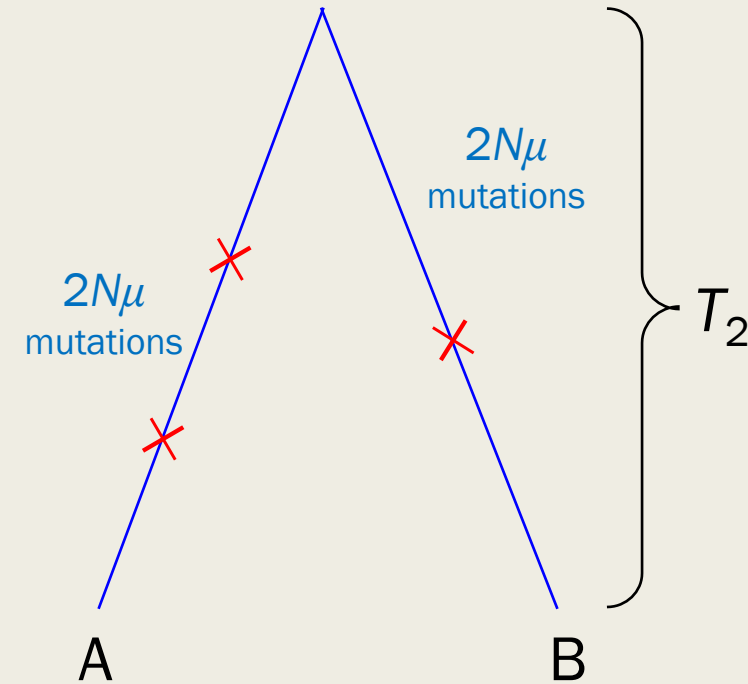
- For now we will consider a single site
- Let μ be the per site, per generation mutation rate
- Considering two samples, the expected time to coalescence is 1 coalescent unit or $2N$ generations



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- For now we will consider a single site
- 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

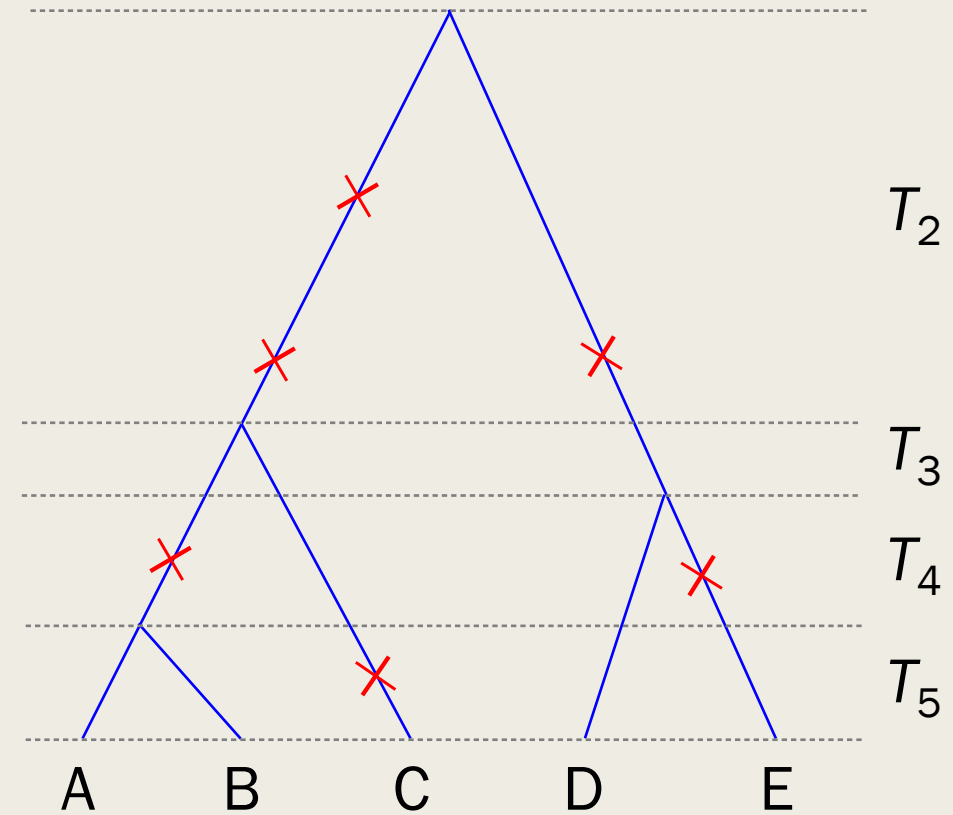
$$E[\pi] = 4N\mu = \theta$$



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

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

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

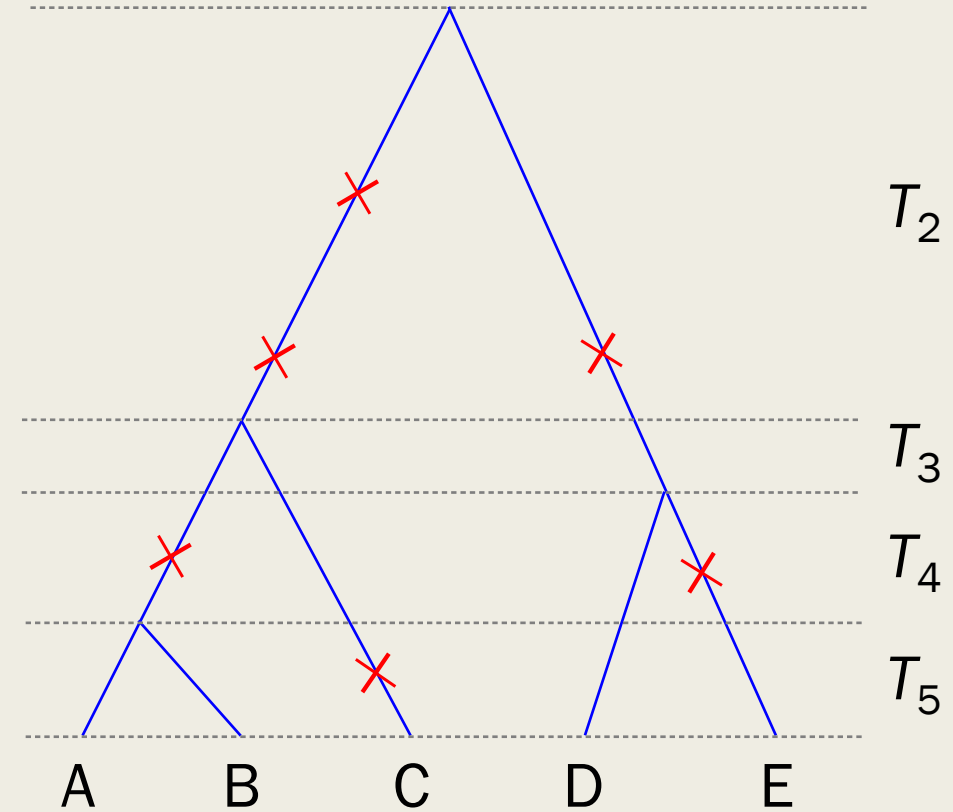


Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

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

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

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



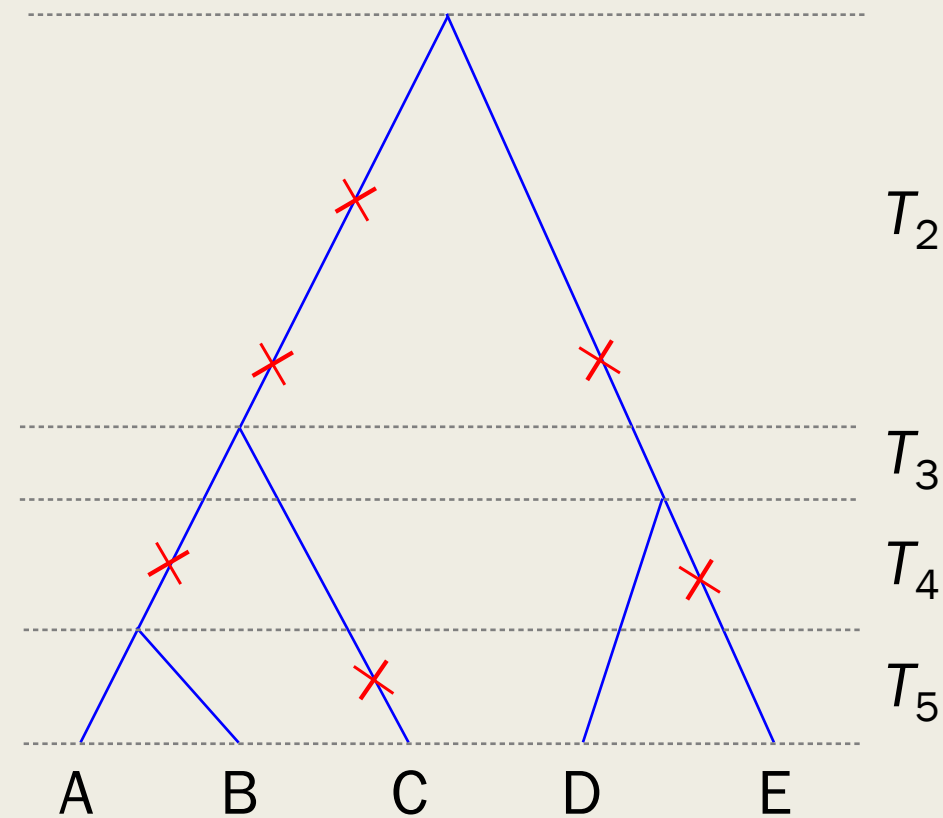
Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

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

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

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

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



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

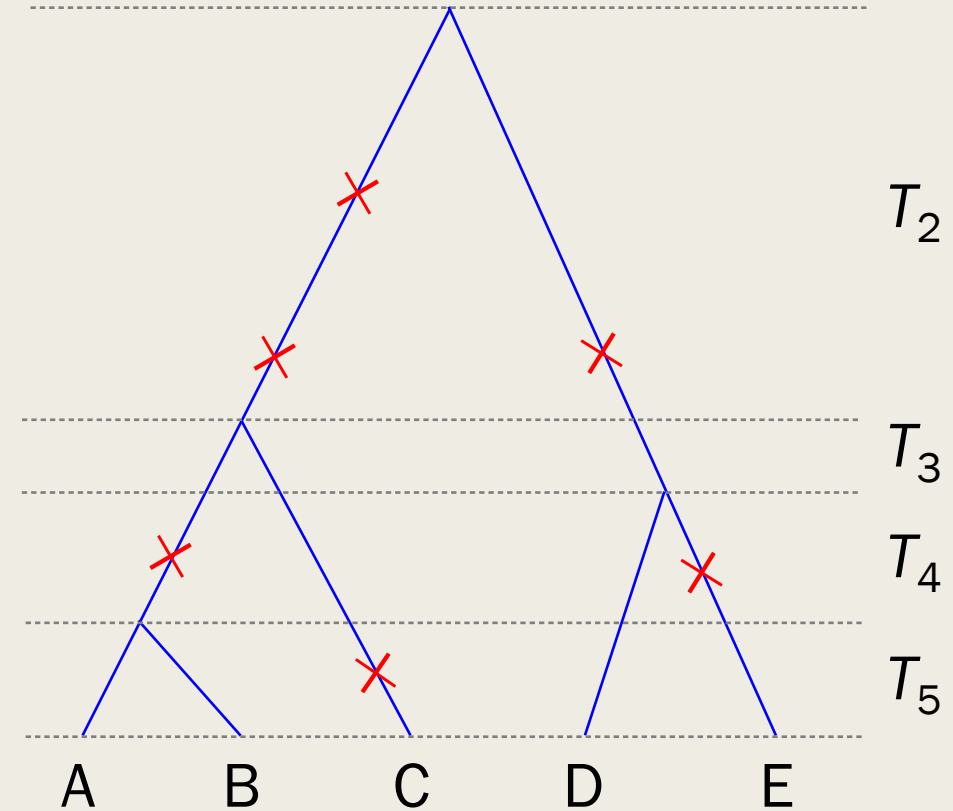
- For $E[S]$, we need to compute the total branch length

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

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

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

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



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

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

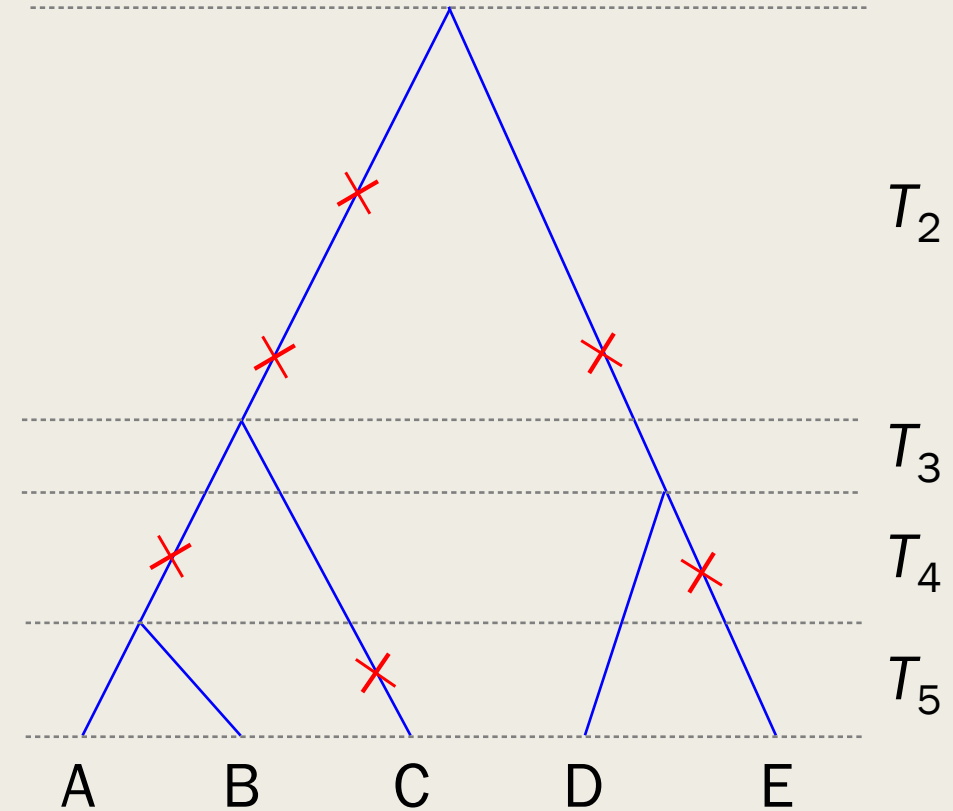
T_{total} = total length of all branches in the tree

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

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

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

$$= 2a_1$$



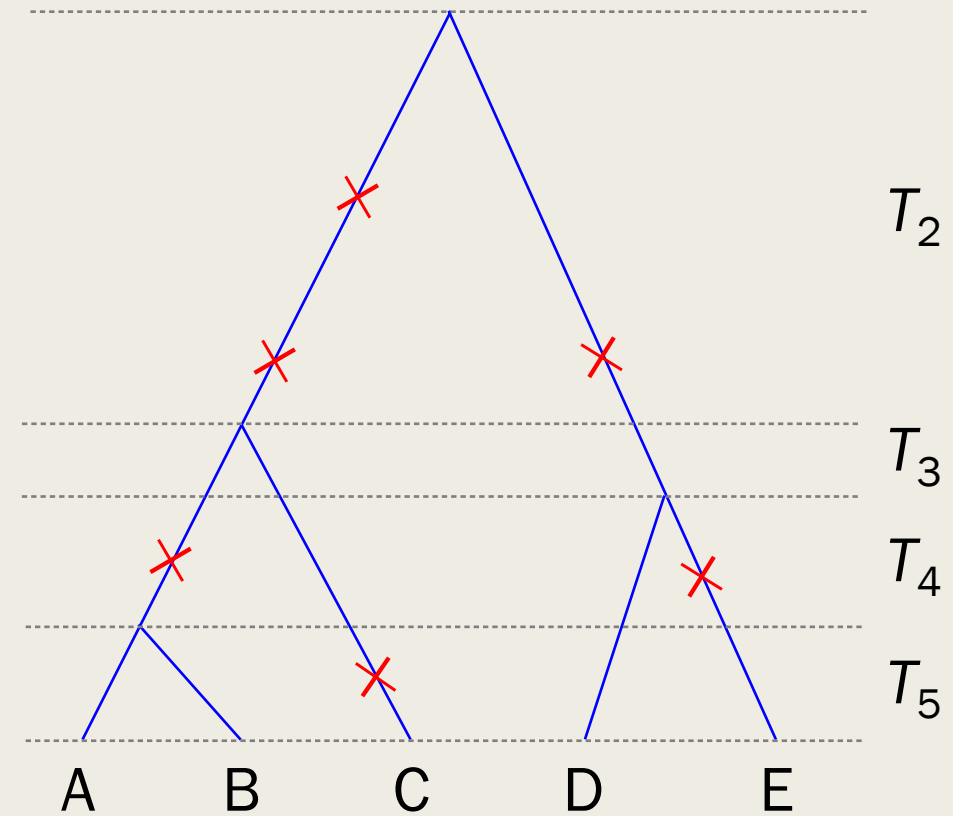
Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- After we have the total branch length, we can multiply 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

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



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:

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

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

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

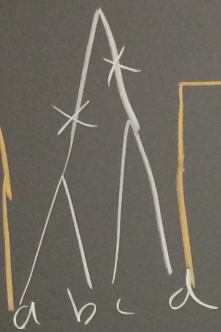
What do deviations from $d=0$ mean?

- 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

$d > 0$

(a)

- bottleneck
- decay

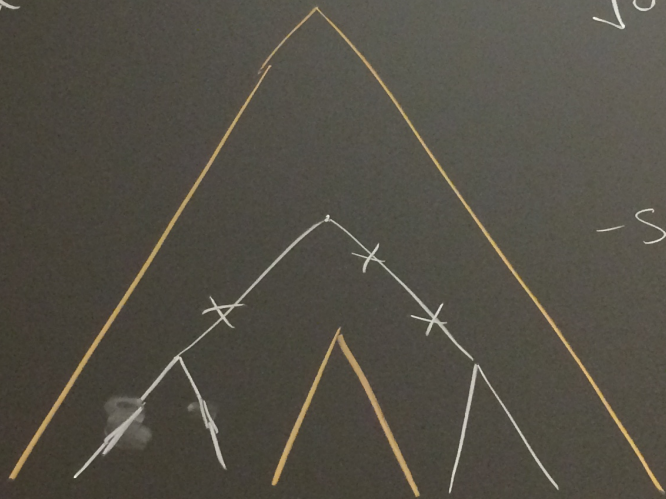


excess
of common
variation

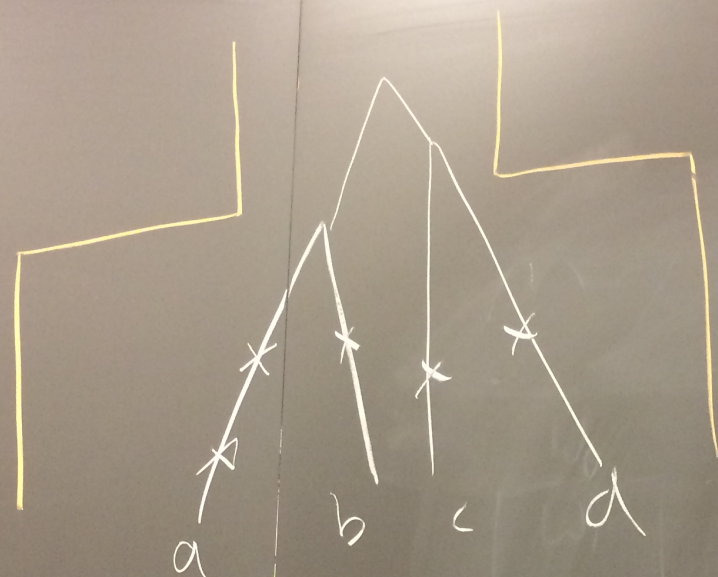
$\mu =$

(b)

- structure

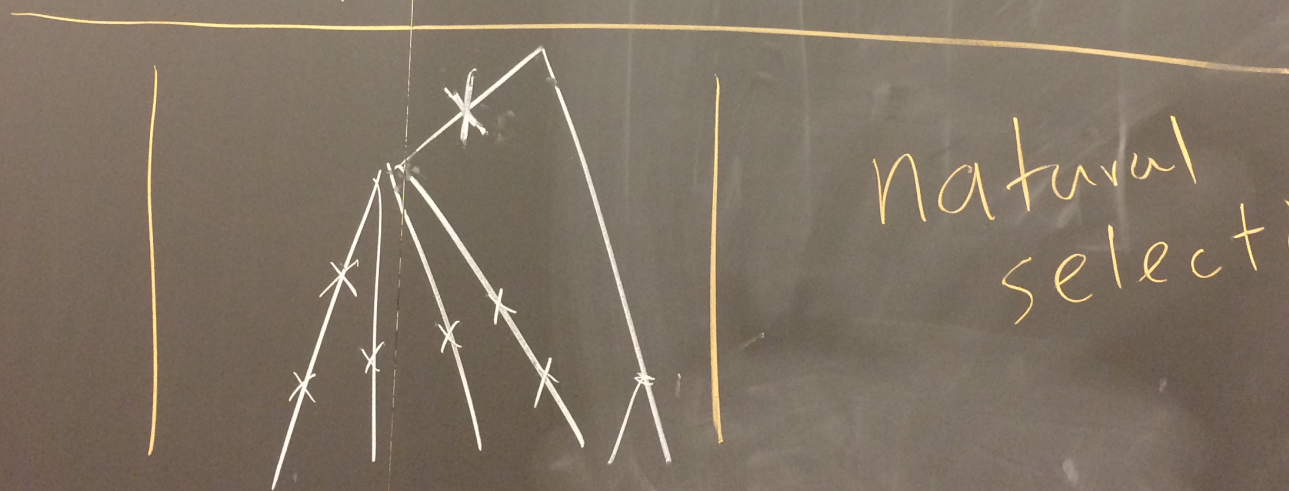


$d < 0$



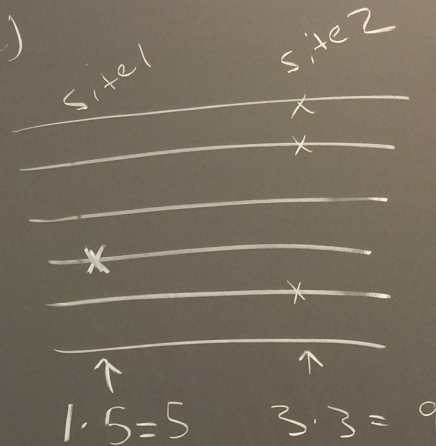
growth

rare
variation



natural
selection

$$\mu = \frac{1.25 \times 10^{-8} \text{ mut}}{\text{base} \cdot \text{gen}} (2N \text{ gen}) (1 \text{ base})$$



rare: < 0.05

common: > 0.05

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. Nickerson¹

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

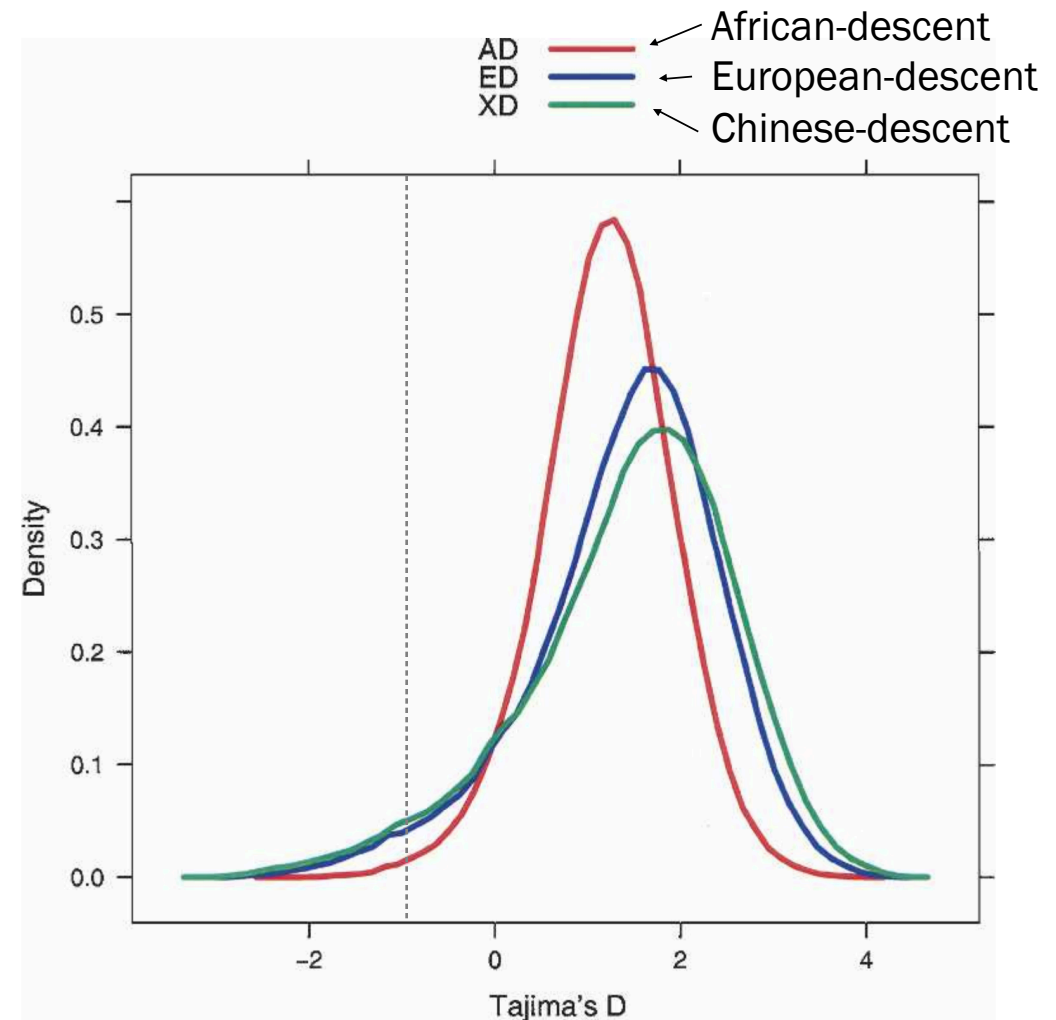


Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

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. Nickerson¹

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

- Why is Tajima's D greater than 0?

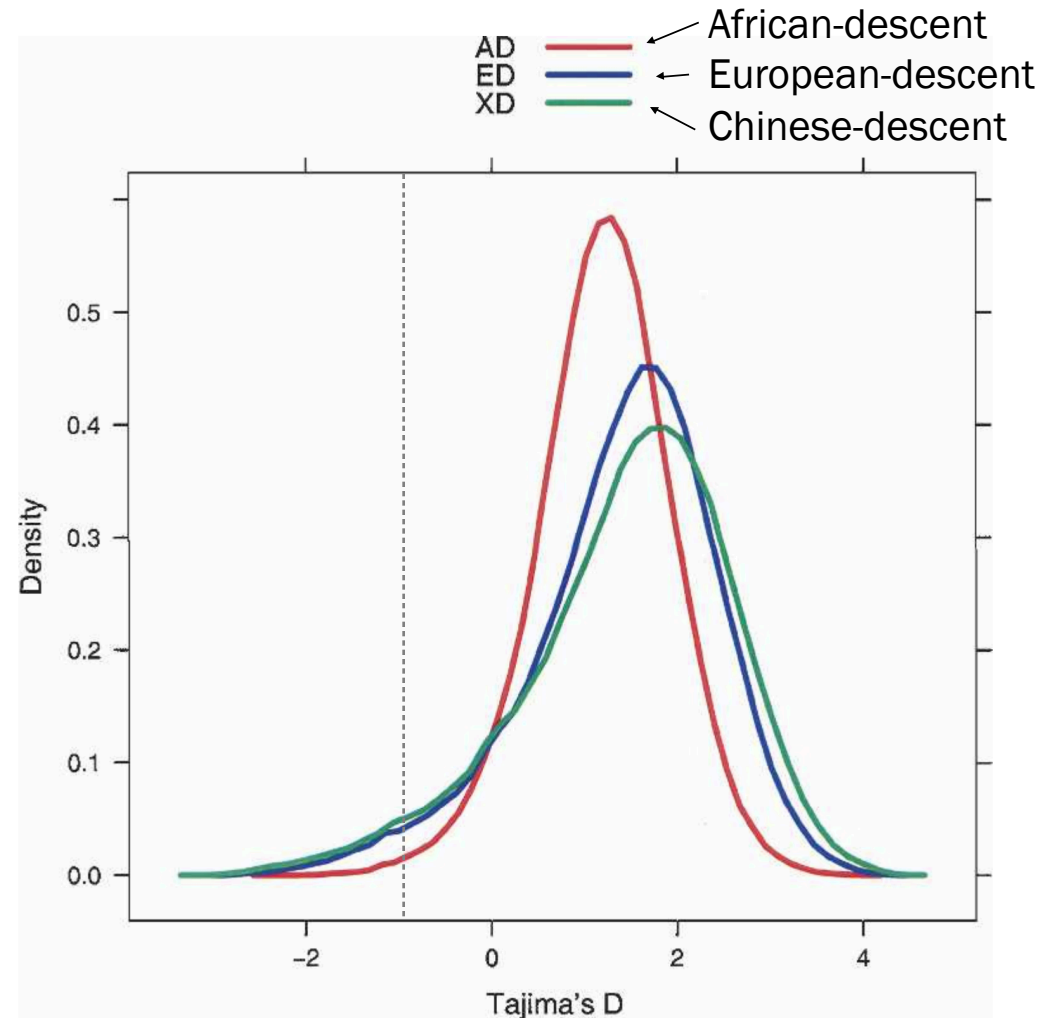


Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

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. Nickerson¹

¹Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; ²Center 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

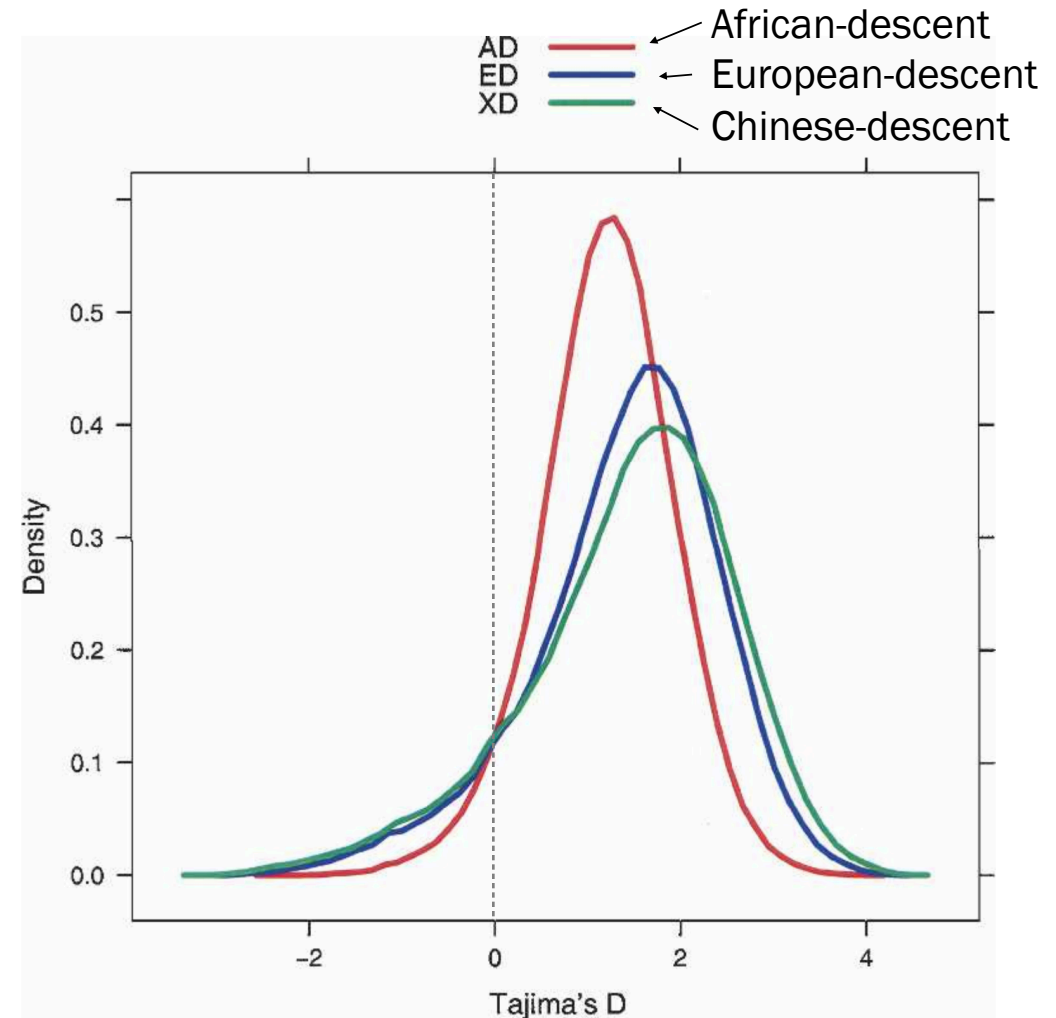


Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

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. Nickerson¹

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

- Regions where Tajima's $D < 0$, probably natural selection (could be random)

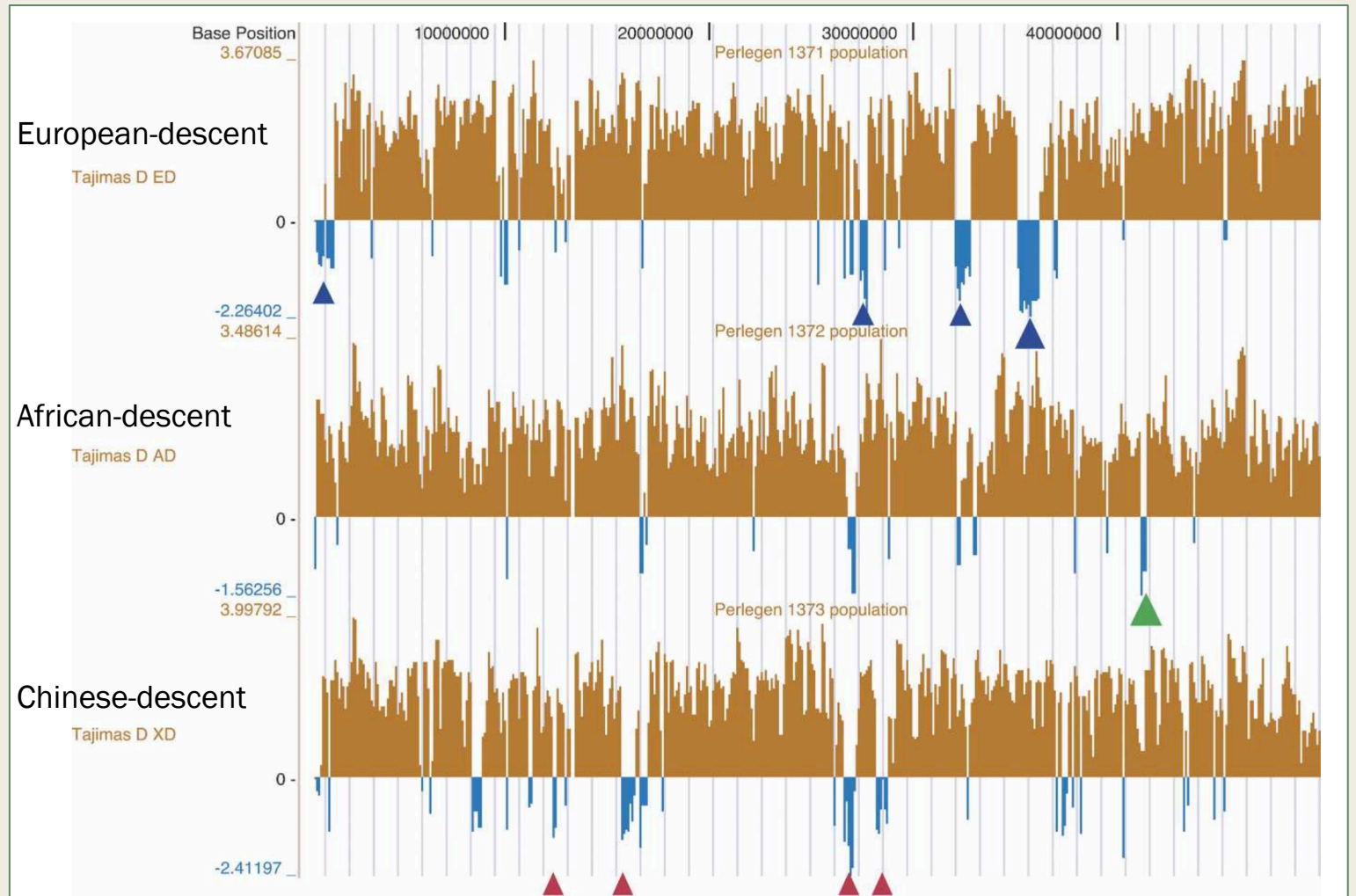


Figure 3. Tajima's D in 100-kbp sliding windows with 10-kbp steps is shown across the first 50 megabases of chromosome 1. Several CRTRs are visible, including a region near 35M in the ED population containing *CLSPN* (large blue arrowhead) and a region near 41M in the AD population spanning *CTPS*, *FLJ23878*, and *SCMH1* (large green arrowhead). CRTRs at the less stringent 5% level are also indicated in the ED population as small blue arrowheads and in the XD population as small red arrowheads.

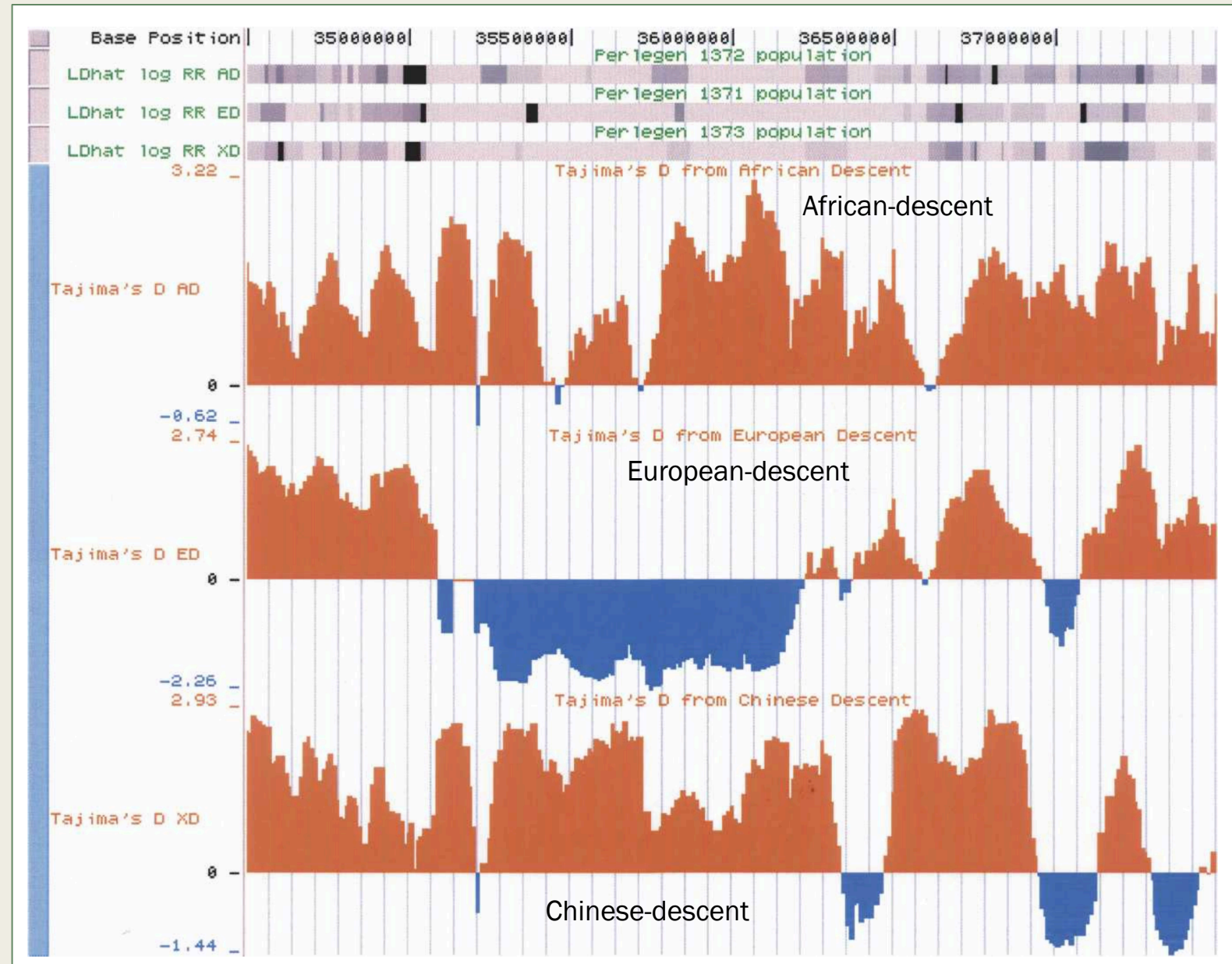
Example of Tajima's D from the literature

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. Nickerson¹

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

- Extended regions of low D , could be strong selection
- This paper found several regions under selection in European and Chinese populations that are linked to drug metabolism



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. Nickerson¹

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

Note: not our formulas!
But exactly the same idea/goal

Nucleotide diversity analysis

There are several statistics that can be used to describe nucleotide diversity, including θ_s (equation 1), π (equation 2), and θ_H (equation 3). These statistics can be calculated for a given resequencing data set by using the following parameters: n is the number of chromosomes resequenced, Sn is the number of polymorphic sites observed, p_i is the derived (nonancestral) allele frequency of the i th SNP, and q_i is the ancestral allele frequency of the i th SNP.

$$\theta_s = \frac{Sn}{n-1} \sum_{i=1}^{Sn} \frac{1}{i} \quad (1)$$

$$\pi = \frac{n}{n-1} \sum_{i=1}^{Sn} 2p_i q_i \quad (2)$$

$$\theta_H = \frac{n}{n-1} \sum_{i=1}^{Sn} 2p_i^2 \quad (3)$$

There are many statistics that can evaluate departures from the expected patterns of neutral variation. One of these is Tajima's D (Tajima 1989), equation 4:

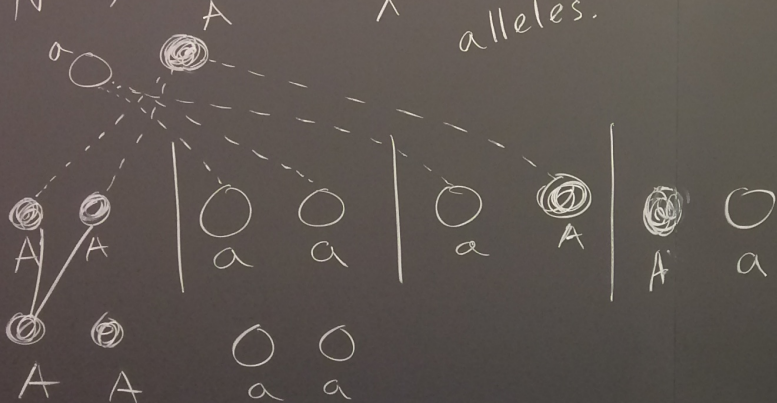
$$D = \frac{\pi - \theta_s}{\sqrt{\text{Var}(\pi - \theta_s)}} \quad (4)$$

Next topic: Hidden Markov Models (HMMs)

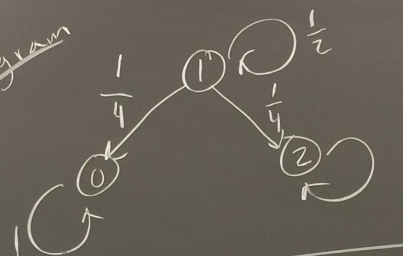
Markov chains

$$N=1, 2N=2$$

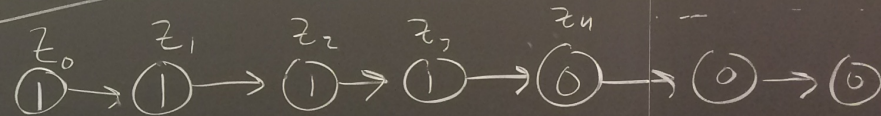
random variable
 $X = \# \text{ of } A \text{ alleles.}$



State diagram



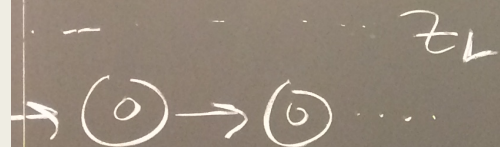
State sequence



$$P(z_0, \dots)$$

$$P(z_0, z_1, z_2, \dots, z_L) = P(z_0) \prod_{i=1}^L \underbrace{P(z_i | z_{i-1})}_{\text{"given"}}$$

Conditional
prob.



Bayes Thm
conditional prob

$$P(a, b) = P(a)P(b|a)$$

↑
"and"

$$= P(b)P(a|b)$$

Bayes Thm

$$P(a)P(b|a) = P(b)P(a|b)$$

U = umbrella

$$P(u) = P(u, r) + P(u, s)$$

rain sun

$$= P(r) \underbrace{P(u|r)}_{0.9} + P(s) \underbrace{P(u|s)}_{0.2}$$