CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Apr 18

- Perfect phylogeny review
- Population genetics review
- Friday: HMM and PCA review
  - *(try to do back of Handout 26 before Friday)*

Midterm tentatively moved to next week (April 26 in lab)

- In lab tomorrow: project meetings for all groups (if you can come to a different section to keep your group together, that's great but no problem if not)
- In lab tomorrow: you can also work on Lab 9 (PCA), can be a nice supplementary analysis for your project (email me if you would like a repo)
- Project proposals are still do Monday
- Take advantage of the extra time to study:
  - Work through Handouts 10-26
  - Read the book (good for phylogenetics and HMMs)
  - Come to office hours or arrange a meeting

Perfect Phylogeny

Lab 6 Analysis questions

1. For the human dataset, run your perfect phylogeny algorithm on the first x sites and experiment with the value of x. What is the largest value of x where there is still a perfect phylogeny?

x = 21, still a perfect phylogeny, x = 22, not a perfect phylogeny

2. Although there are some sites with repeated mutations in the human dataset, that is actually not the main reason why there is not a perfect phylogeny when analyzing a large number of sites. Based on our discussion of population genetics in class so far, what might be the main reason? Explain your answer.

3.Does your runtime plot agree with your theoretical runtime analysis from Part 2? Explain why or why not.

Lab 6 Analysis questions

1. For the human dataset, run your perfect phylogeny algorithm on the first x sites and experiment with the value of x. What is the largest value of x where there is still a perfect phylogeny?

x = 21, still a perfect phylogeny, x = 22, not a perfect phylogeny

2. Although there are some sites with repeated mutations in the human dataset, that is actually not the main reason why there is not a perfect phylogeny when analyzing a large number of sites. Based on our discussion of population genetics in class so far, what might be the main reason? Explain your answer.

Our starting assumption that there should only be one tree is wrong. Due to recombination (which occurs between individuals from the same species but not across species), the tree changes along the genome. In each recombination "block", it is quite likely there will be a perfect phylogeny.

3.Does your runtime plot agree with your theoretical runtime analysis from Part 2? Explain why or why not.

Lab 6 Analysis questions

1. For the human dataset, run your perfect phylogeny algorithm on the first x sites and experiment with the value of x. What is the largest value of x where there is still a perfect phylogeny?

x = 21, still a perfect phylogeny, x = 22, not a perfect phylogeny

2. Although there are some sites with repeated mutations in the human dataset, that is actually not the main reason why there is not a perfect phylogeny when analyzing a large number of sites. Based on our discussion of population genetics in class so far, what might be the main reason? Explain your answer.

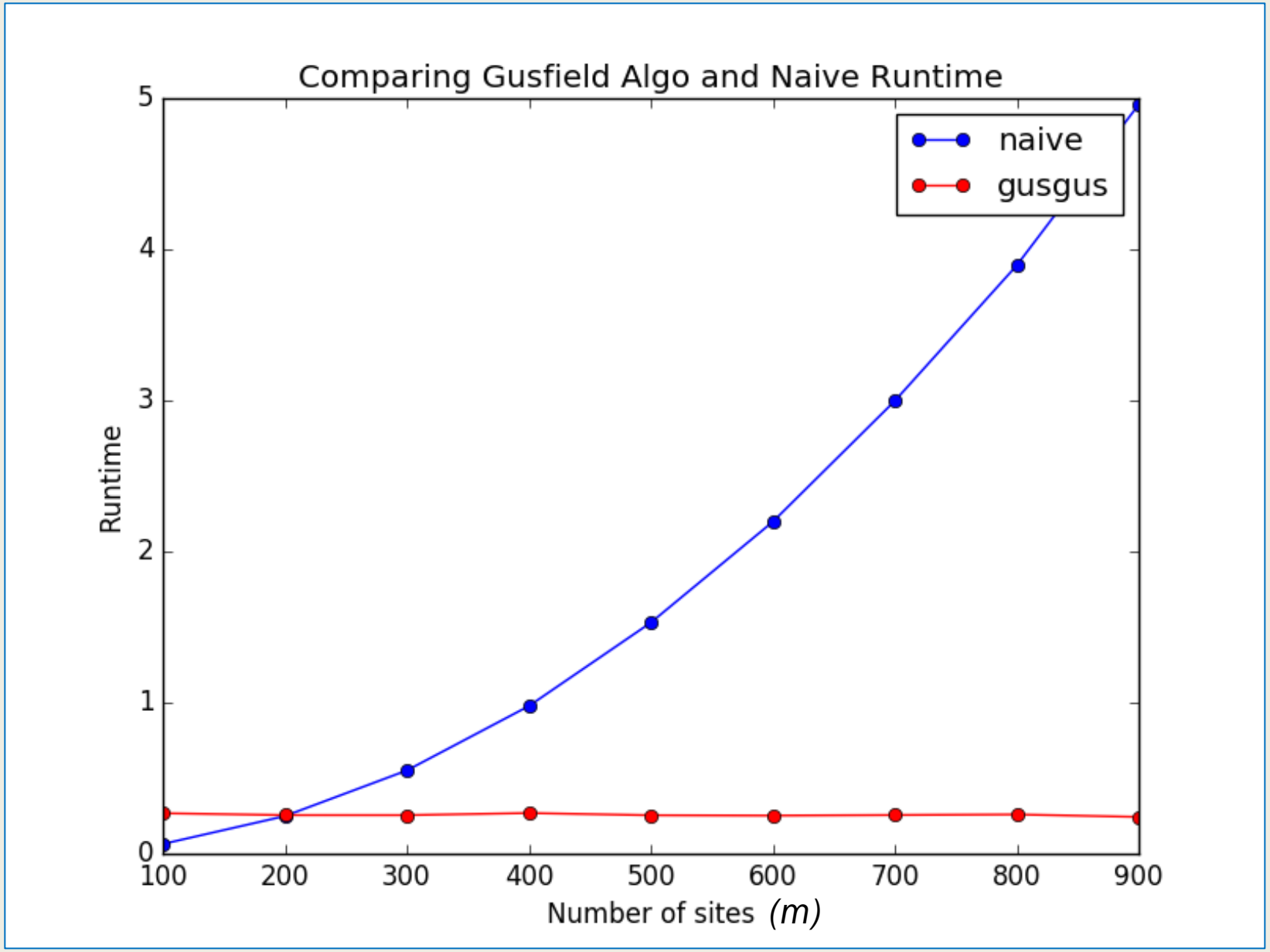
Our starting assumption that there should only be one tree is wrong. Due to recombination (which occurs between individuals from the same species but not across species), the tree changes along the genome. In each recombination "block", it is quite likely there will be a perfect phylogeny.

3.Does your runtime plot agree with your theoretical runtime analysis from Part 2? Explain why or why not.

n=number of samples, m=number of sites
Naïve algorithm: $O(nm^2)$, Gusfield: $O(nm)$

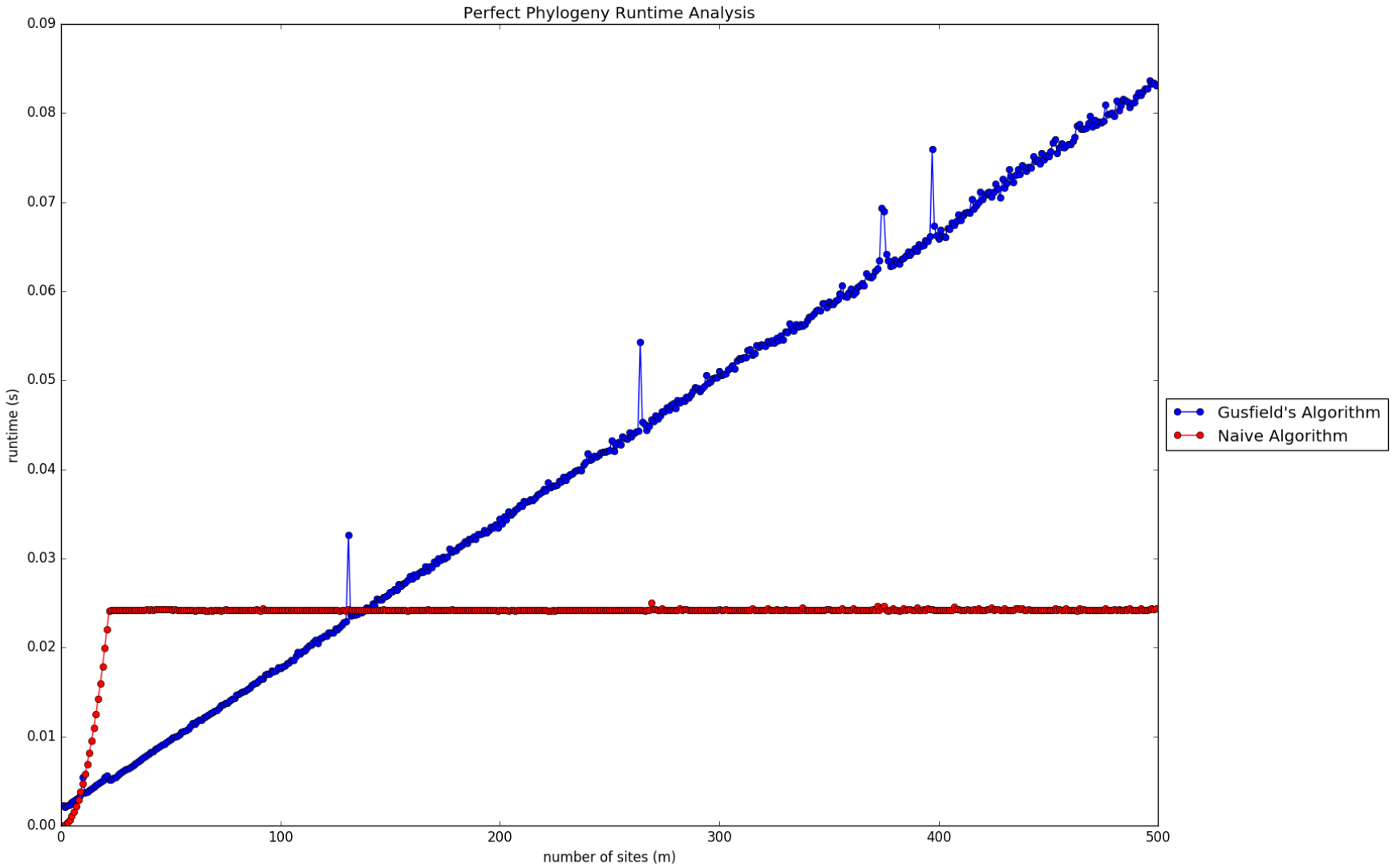
What is the runtime of Gusfield's algorithm?

Linear in both *n* & *m*: O(*nm*)



Theint & Ben

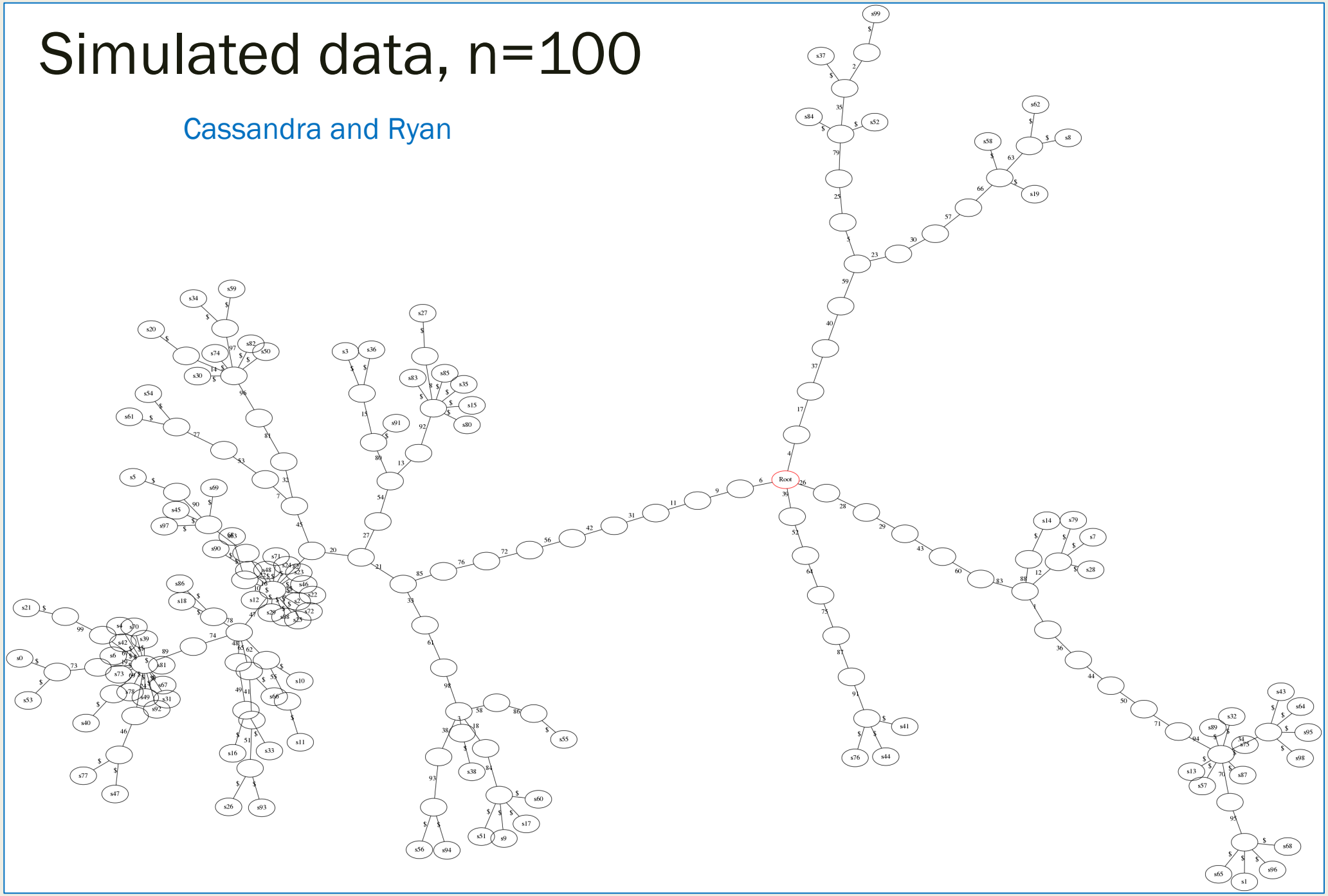
On human data, naïve does much better



Genji and Gabriel

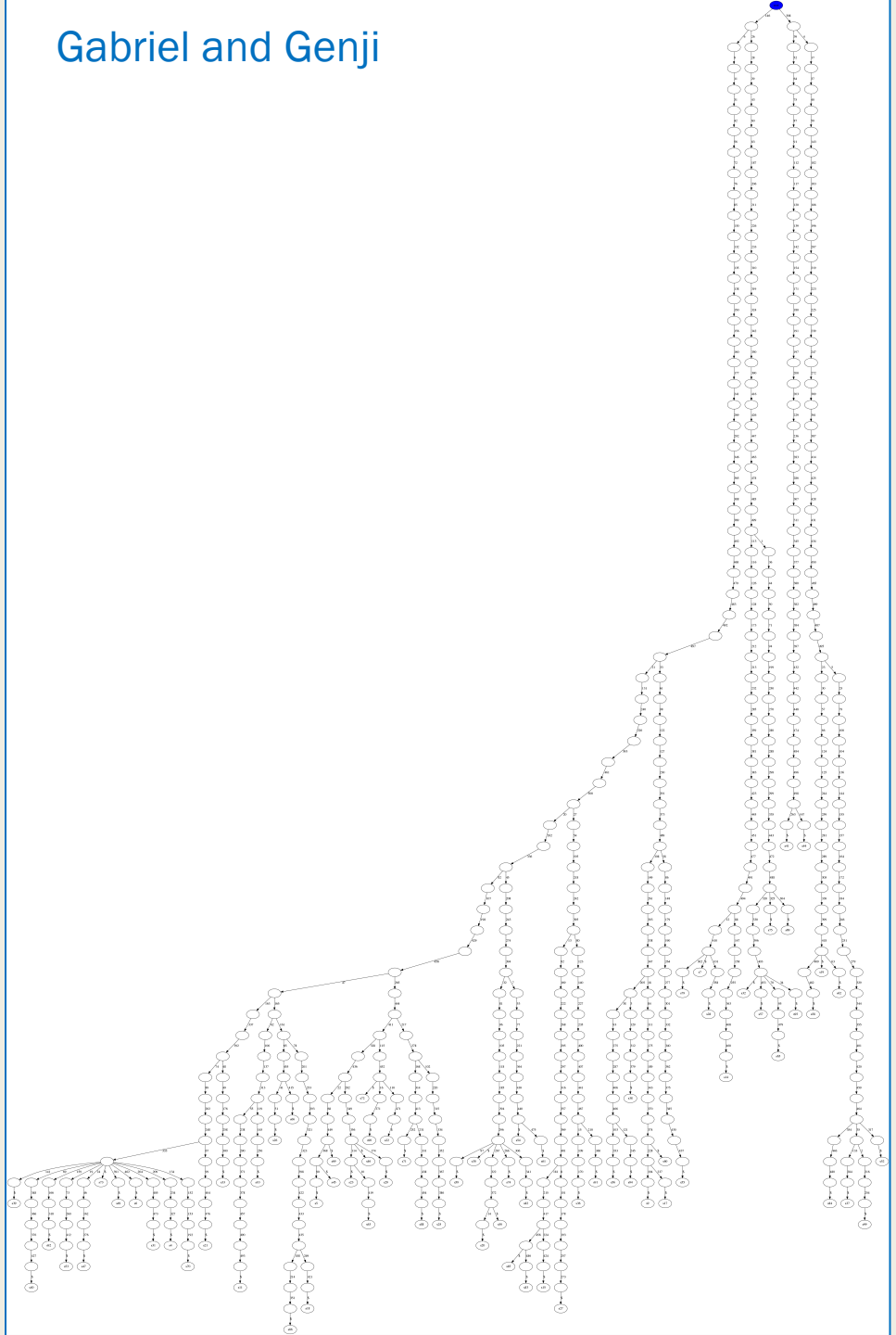
Simulated data, n=100

Cassandra and Ryan



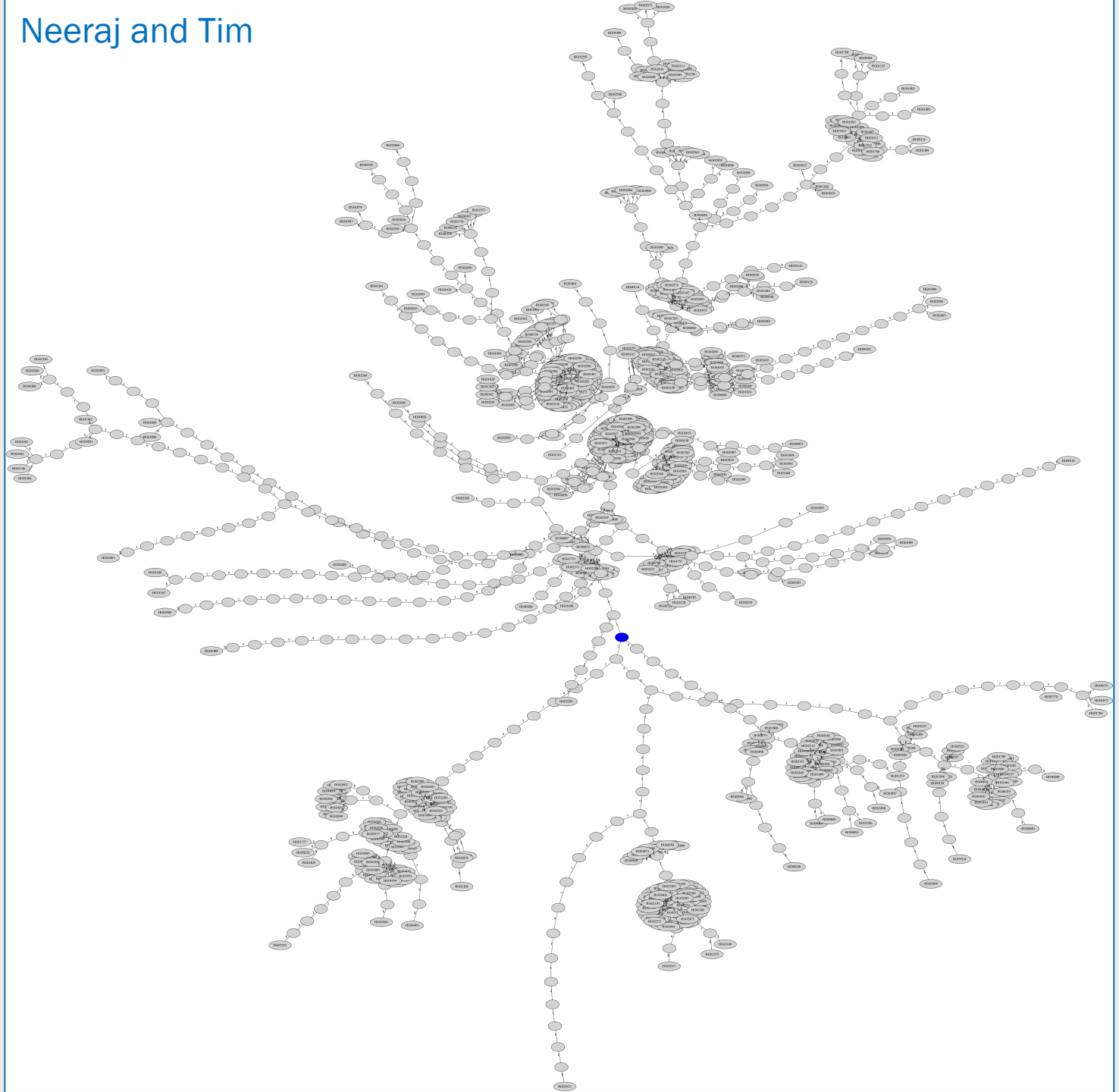
Simulated data, n=100

Gabriel and Genji



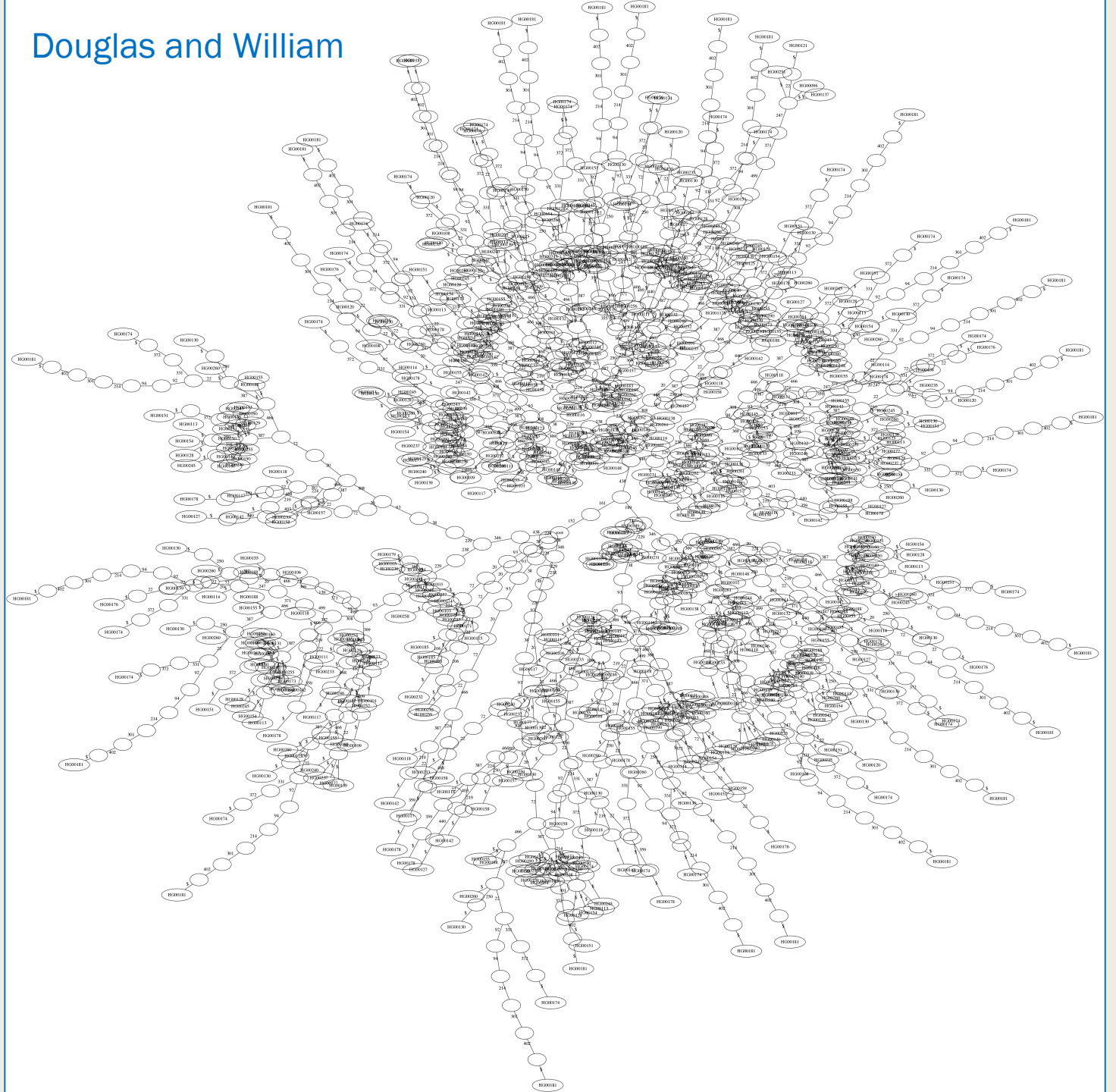
Human data, m=250

Neeraj and Tim

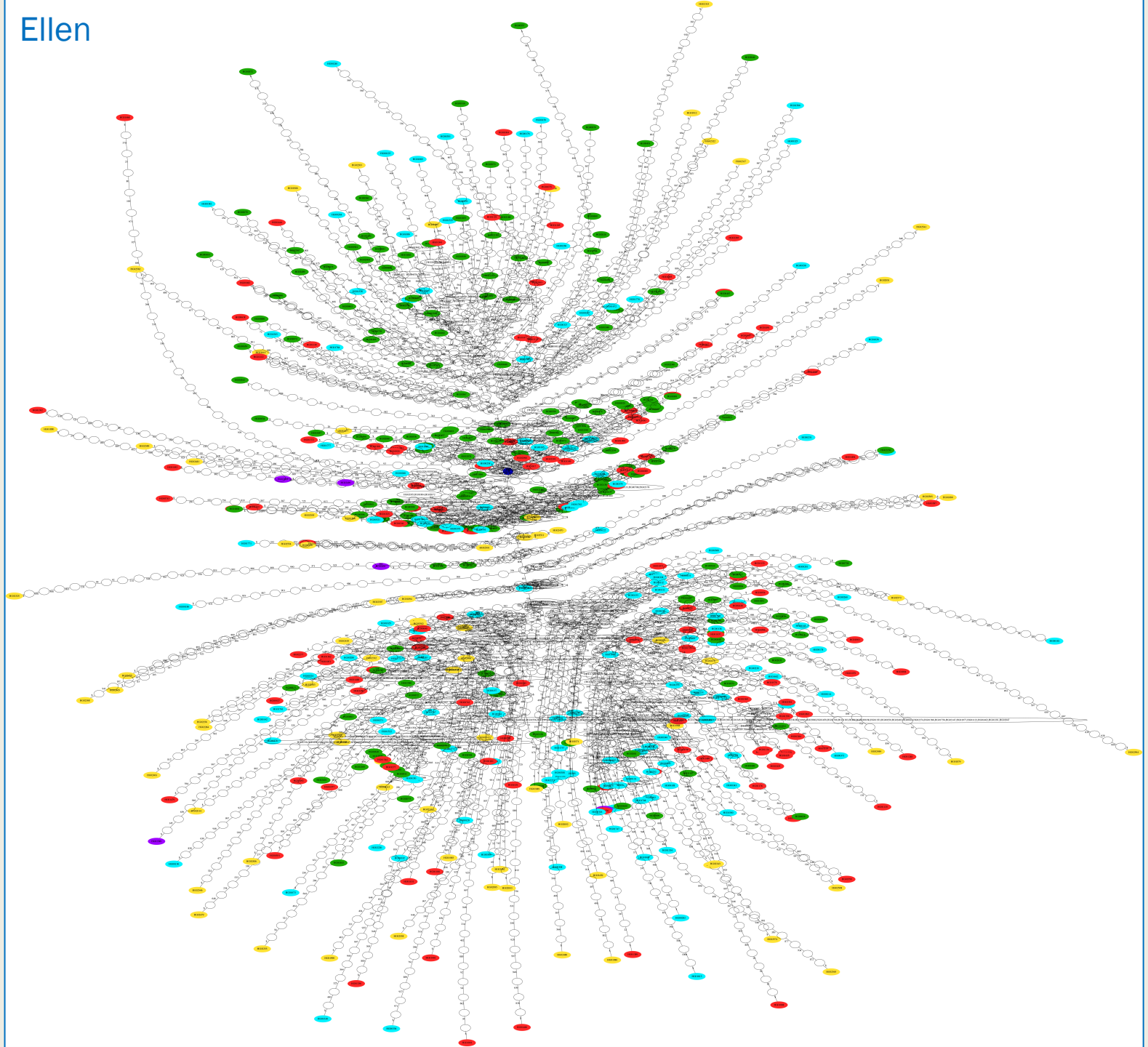


Human data

Douglas and William



Human data extension, m=1000



Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

2	1	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	0	0	0

Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

2	1	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	0	0	0

3	4	2	1	5
0	0	1	1	0
1	0	0	0	0
0	0	1	1	1
1	1	0	0	0
0	0	1	0	0

Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

2	1	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	0	0	0

3	4	2	1	5
0	0	1	1	0
1	0	0	0	0
0	0	1	1	1
1	1	0	0	0
0	0	1	0	0

2	1	5	3	4
1	1	0	0	0
0	0	0	1	0
1	1	1	0	0
0	0	0	1	1
1	0	0	0	0

Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

2	1	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	0	0	0

3	4	2	1	5
0	0	1	1	0
1	0	0	0	0
0	0	1	1	1
1	1	0	0	0
0	0	1	0	0

2	1	5	3	4
1	1	0	0	0
0	0	0	1	0
1	1	1	0	0
0	0	0	1	1
1	0	0	0	0

3	2	1	5	4
0	1	1	0	0
1	0	0	0	0
0	1	1	1	0
1	0	0	0	1
0	1	0	0	0

Radix sort columns high to low

Handout 18:
Example 1

1	2	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
0	1	0	0	0

2	1	3	4	5
1	1	0	0	0
0	0	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	0	0	0

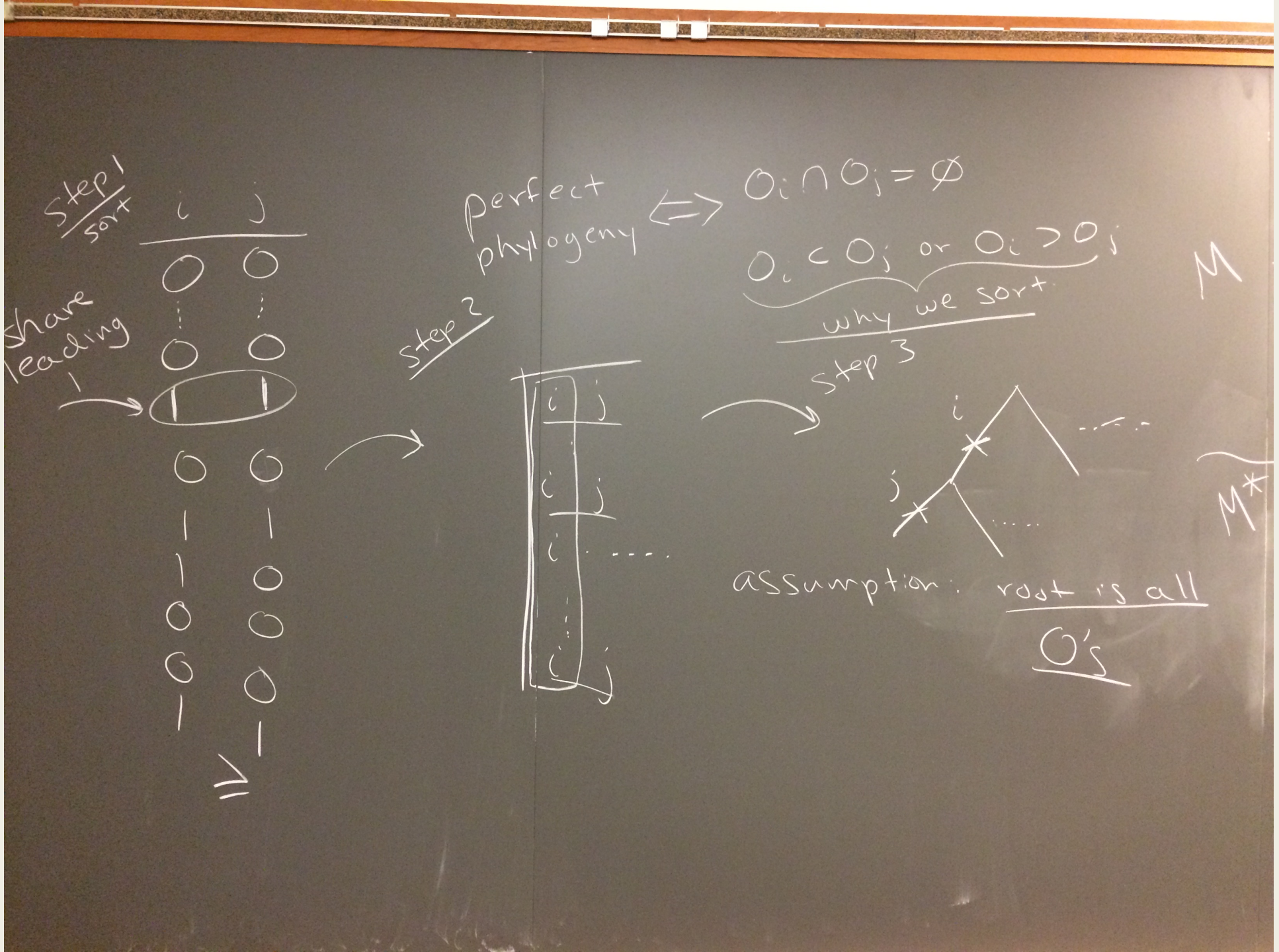
3	4	2	1	5
0	0	1	1	0
1	0	0	0	0
0	0	1	1	1
1	1	0	0	0
0	0	1	0	0

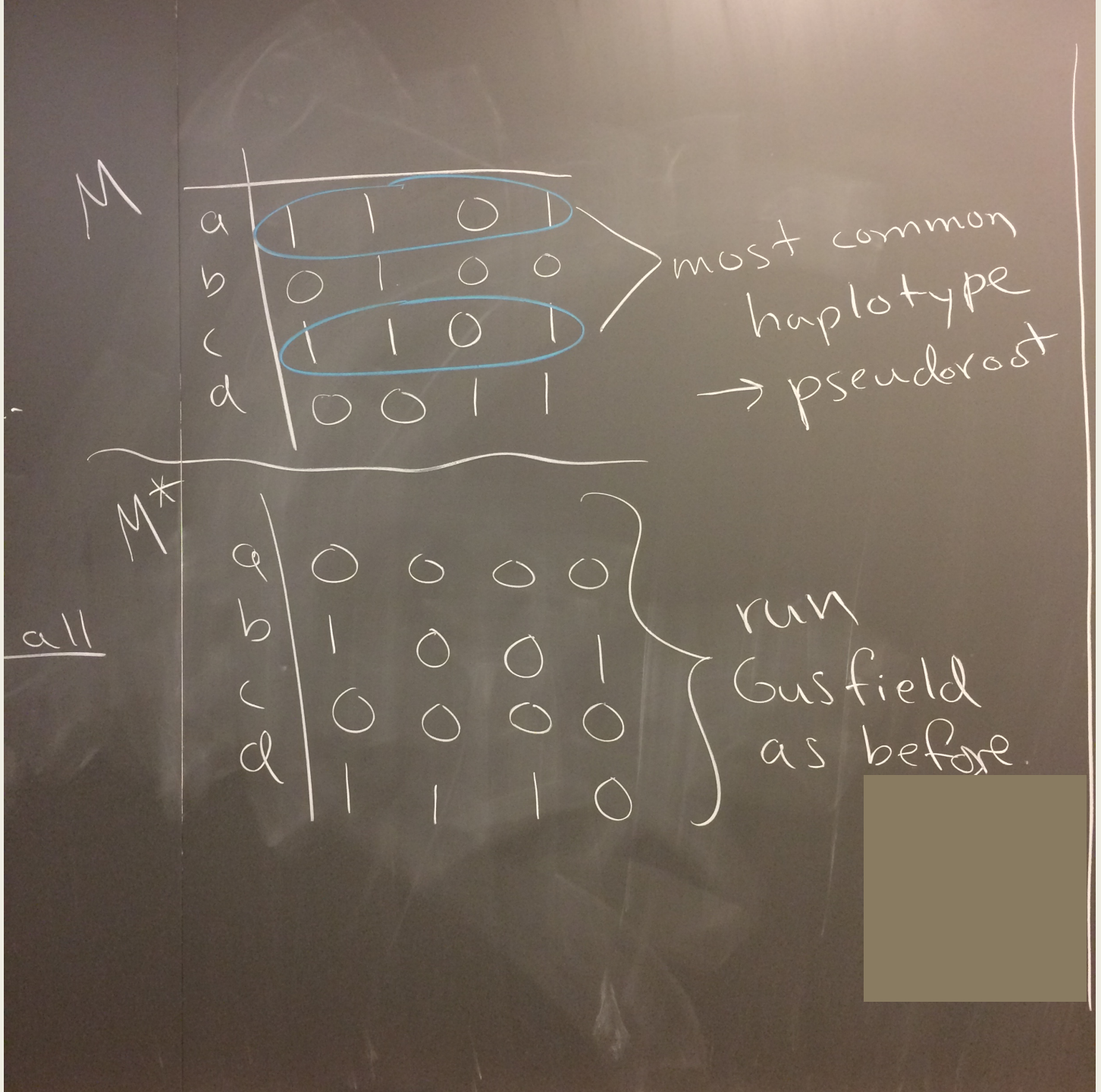
2	1	5	3	4
1	1	0	0	0
0	0	0	1	0
1	1	1	0	0
0	0	0	1	1
1	0	0	0	0

3	2	1	5	4
0	1	1	0	0
1	0	0	0	0
0	1	1	1	0
1	0	0	0	1
0	1	0	0	0

2	1	3	5	4
1	1	0	0	0
0	0	1	0	0
1	1	0	1	0
0	0	1	0	1
1	0	0	0	0

Why do we radix sort?


step1
sort
i
j
share
leading
1
step 2
perfect
phylogeny
$O_i \cap O_j = \emptyset$
$O_i \subset O_j$ or $O_i \supset O_j$
why we sort
step 3
M
M*
i
j
assumption: root is all
0's


M
a 1 1 0 1
b 0 1 0 0
c 1 1 0 1
d 0 0 1 1
most common haplotype → pseudoroot
M*
all
a 0 0 0 0
b 1 0 0 1
c 0 0 0 0
d 1 1 1 0
run Gusfield as before.

Population Genetics

Lab 7, Part 2, Question 1

1. Integrate this probability distribution over $t \in [0, \infty)$ to demonstrate that the result is 1 (i.e. that it is a proper probability distribution).

Proof. Observe that by the definition of the binomial coefficient,

$$\binom{i}{2} = \frac{i!}{2!(i-2)!} = \frac{i(i-1)}{2}.$$

Before integrating, we demonstrate the this improper integral exists in the extended sense. Let A be the open set in $\mathbb{R}$ defined by $A = \{x | x > 0\}$. Observe that by the continuity of the exponential function, P_{T_i} is clearly continuous. Now we let $U_N = (0, N)$ be a sequence of open sets; clearly the union of all U_N is A. Then since $|P_{T_i}|$ exists and is bounded by $i(i-1)/2$ on each U_N, it follows by theorem that $\int_{U_N} |f|$ exists and is bounded. Moreover, we find that

$$\int_A P_{T_i} = \lim_{N \to \infty} \int_{U_N} P_{T_i}.$$

Computing $\int_{U_N} P_{T_i}$, it follows that

$$\int_{U_N} P_{T_i} = \frac{i(i-1)}{2} \int_0^N e^{-\frac{i(i-1)}{2} t} = \frac{i(i-1)}{2} \left[\frac{2}{i(i-1)} e^{-\frac{i(i-1)}{2} t} \right]_0^N = e^{-\frac{i(i-1)}{2} N} + 1.$$

Therefore, we can conclude that

$$\int_A P_{T_i} = \lim_{N \to \infty} e^{-\frac{i(i-1)}{2} N} + 1 = 1,$$

as desired.

Alex and Dayo

Lab 7, Part 2, Question 3

Angelina and Rye

Integration by parts:

$$u = t \quad \frac{dv}{dt} = xe^{-\binom{i}{2}t}$$

$$du = dt \quad v = -e^{-\binom{i}{2}t}$$

$$-te^{-\binom{i}{2}} - \int_0^\infty -e^{-\binom{i}{2}t}dt$$

$$[-te^{-\binom{i}{2}t} + \frac{-1}{\binom{i}{2}}e^{-\binom{i}{2}t}]_0^\infty$$

$$(-\infty e^{-\infty} + \frac{-1}{\binom{i}{2}}e^{-\infty}) - (0e^0 + \frac{-1}{\binom{i}{2}}e^0)$$

$$\frac{1}{\binom{i}{2}} = \frac{2}{i(i-1)}$$

Charlotte and Emily

Note: $\infty e^{-\infty}$ evaluates to $\infty * 0$ which is an indeterminate form. We take f(t) $= \infty$ and g(t) $= 0$. The $\lim_{x\to c} f(x) = \infty$ and $\lim_{x\to c} g(x) = 0$ for some constant c. By applying L'Hopital's rule, we can look at $\lim_{x\to c} \frac{1}{f(x)} = \lim_{x\to c} \frac{1}{\infty} = 0$ Thus, we have $\lim_{x\to c} f(x)$ and $\lim_{x\to c} \frac{1}{g(x)}$ both go to $0 \implies \lim_{x\to c} f(x)g(x) = 0, \implies \infty e^{-\infty}$ goes to 0.

Lab 7, Part 2, Question 3

Julia and Gabriel

$$
\begin{aligned}
E[T_{MRCA}] &= \sum_{i=n}^{2} E[T_i] \\
&= \sum_{i=n}^{2} \frac{2}{i(i-1)} \\
&= \sum_{i=2}^{n} \frac{2}{i(i-1)} \\
&= 2\sum_{i=1}^{n-1} \frac{1}{(i+1)i} \\
&= 2\sum_{i=1}^{n-1} \left(\frac{1}{i} - \frac{1}{i+1}\right) \\
&= 2\left[\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \ldots + \left(\frac{1}{n-1} - \frac{1}{n}\right)\right] \\
&= 2\left[1 + \left(-\frac{1}{2} + \frac{1}{2}\right) + \left(-\frac{1}{3} + \frac{1}{3}\right) + \ldots + \left(-\frac{1}{n-1} + \frac{1}{n-1}\right) - \frac{1}{n}\right] \\
&= 2\left[1 - \frac{1}{n}\right] \\
&= 2 - \frac{2}{n}
\end{aligned}
$$

As n grows, T_{mrca} approaches a limit of 2.

We found that $\frac{1}{(i+1)i} = \left(\frac{1}{i} - \frac{1}{i+1}\right)$ through:

$$
\begin{aligned}
\frac{1}{i(i+1)} &= \frac{A}{i} + \frac{B}{i+1} \\
1 &= A(i+1) + Bi \\
1 &= Ai + A + Bi \\
1 &= A \qquad\qquad 0i = (A+B)i \\
B &= -A = 1 \\
\frac{1}{i(i+1)} &= \frac{1}{i} + \frac{-1}{i+1}
\end{aligned}
$$

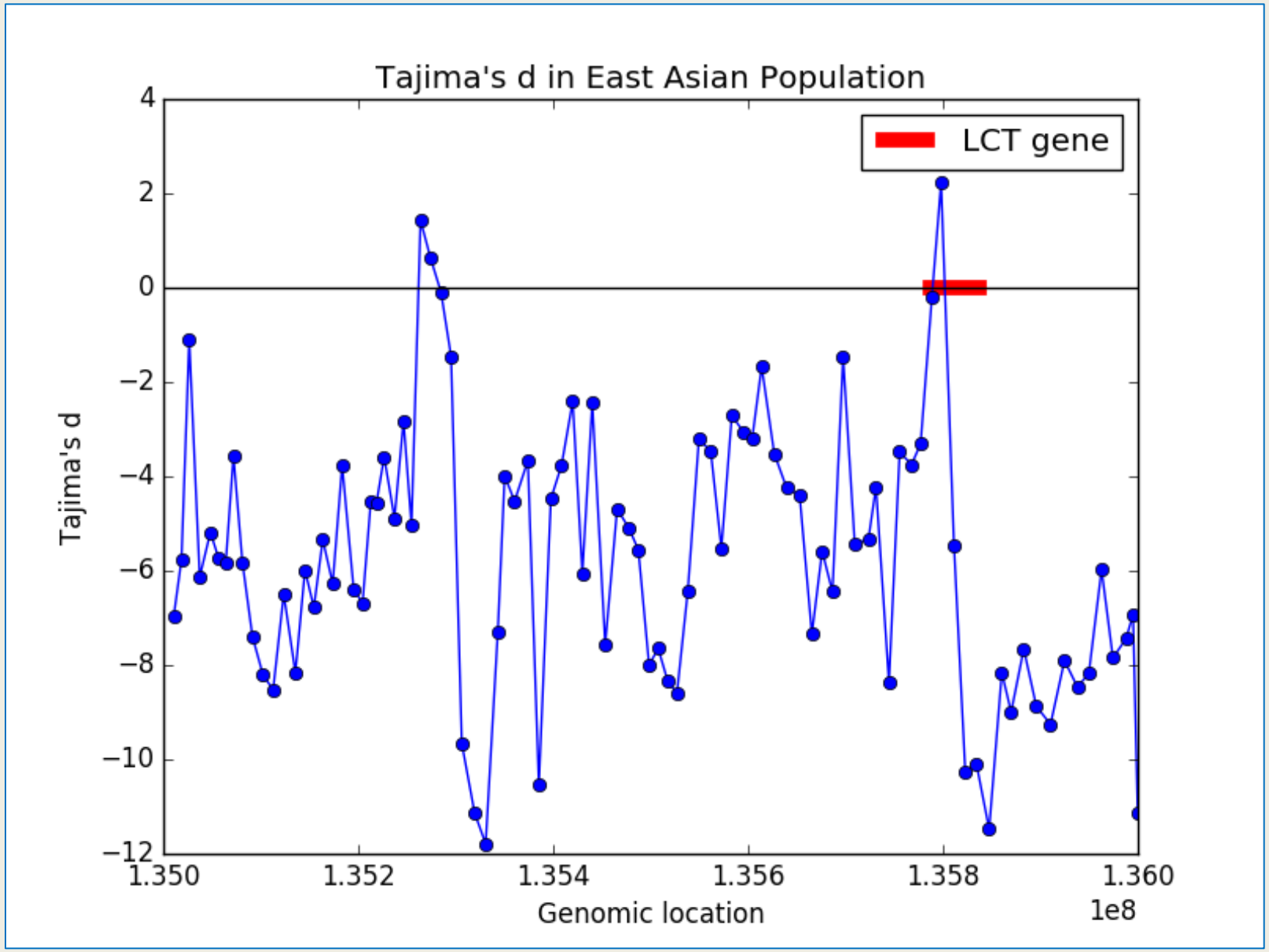
Lab 7 Analysis questions

1. What does the LCT gene do? Why might it have been under natural selection in the past?

   The lactase (LCT) gene regulates lactose metabolism. A variation of this gene allows individuals to digest milk into adulthood. This variation has been under natural selection in the past in some human populations. This selection has often coincided with cattle domestication (individuals who could take advantage of more nutritional resources were more likely to survive and reproduce).

2. What conclusions can you draw from your plot of Tajima's d and the LCT gene?

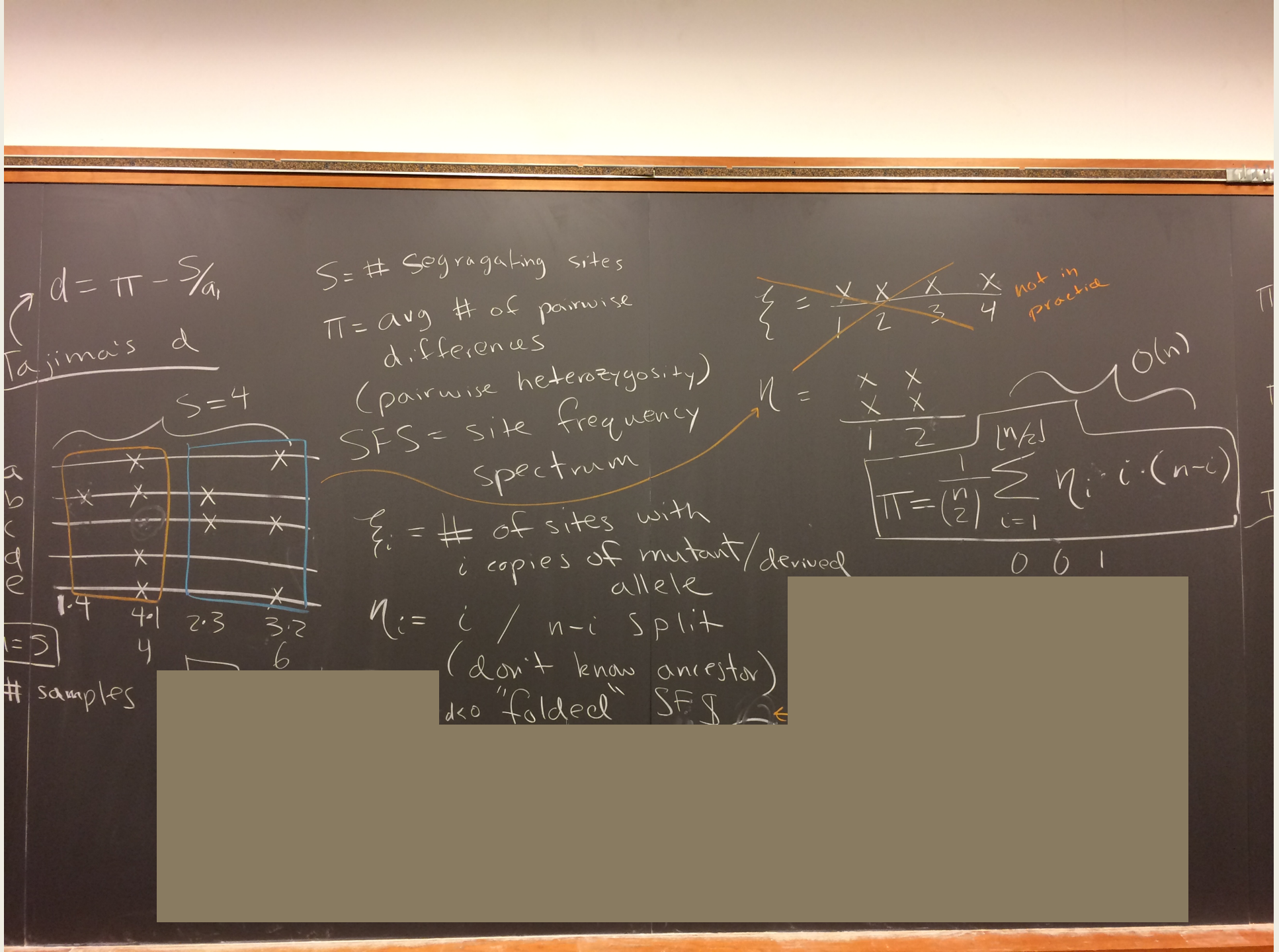
Tajima’s d on human data



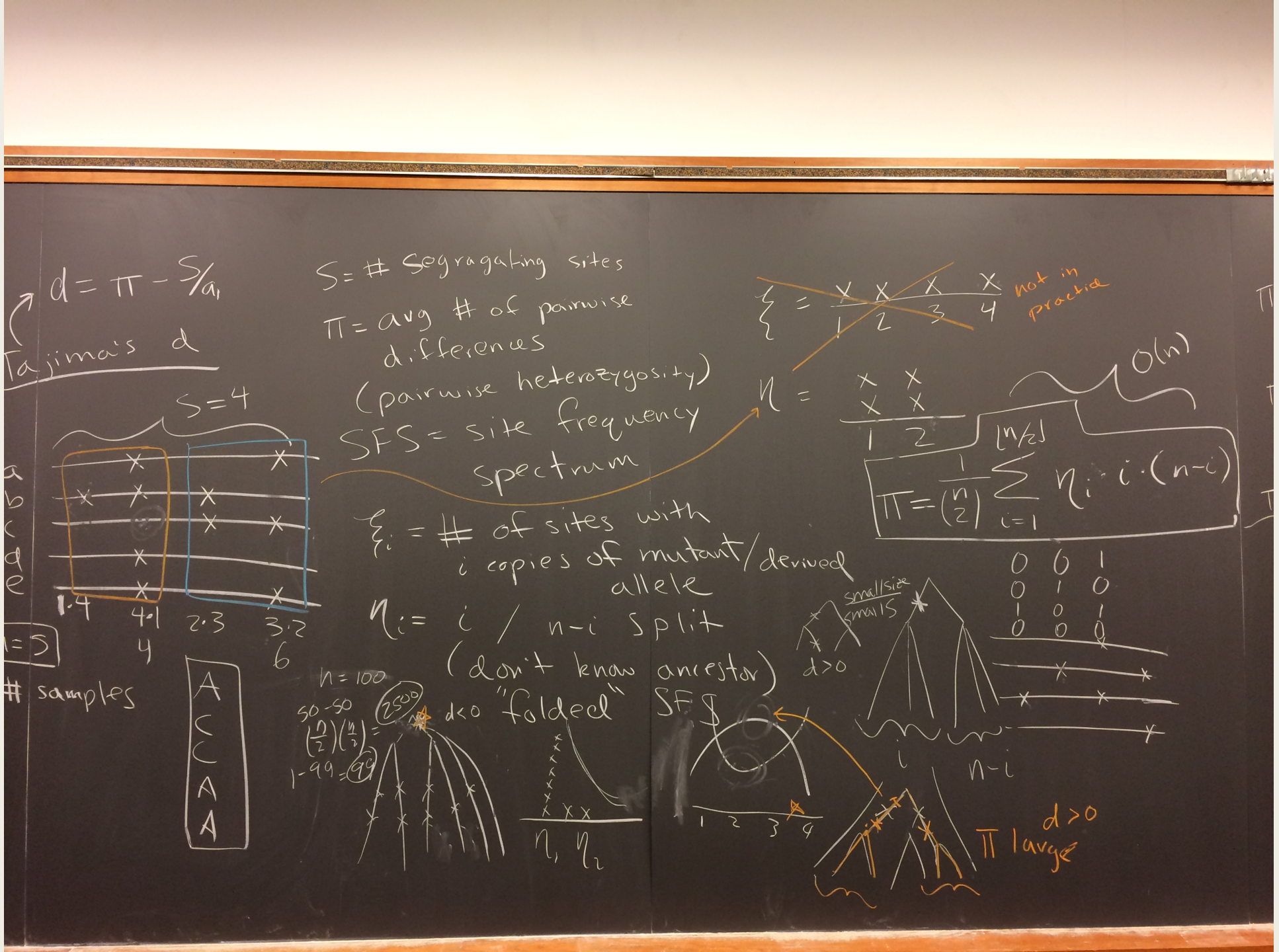
Linda & Lesia

Goals of population genetics review:

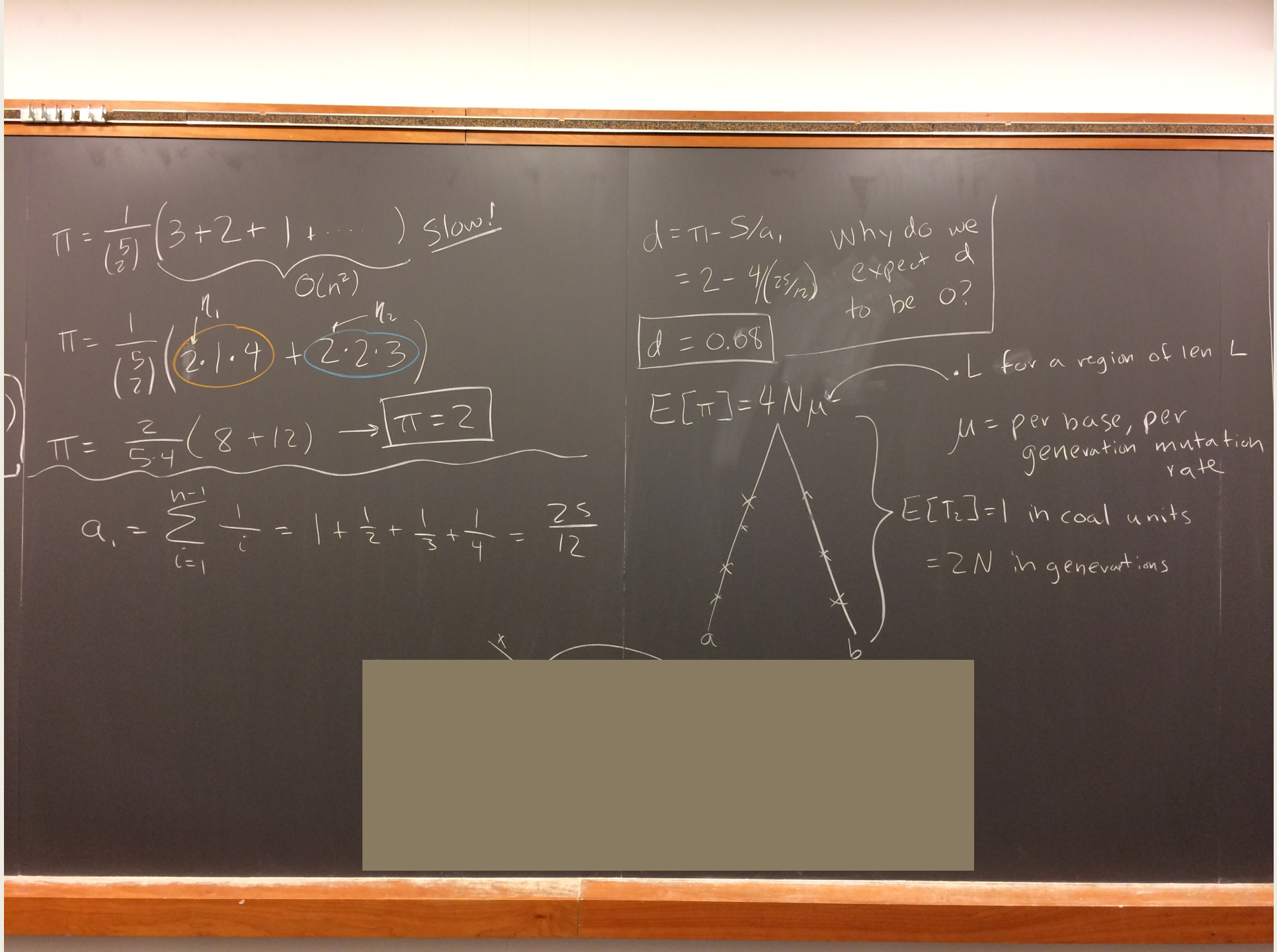
- Compute Tajima's *d* in practice (i.e. for an actual dataset)
- Compute the theoretical (expected) value of Tajima's *d* under neutrality
- See why $d > 0$ or $d < 0$ mean deviations from neutrality



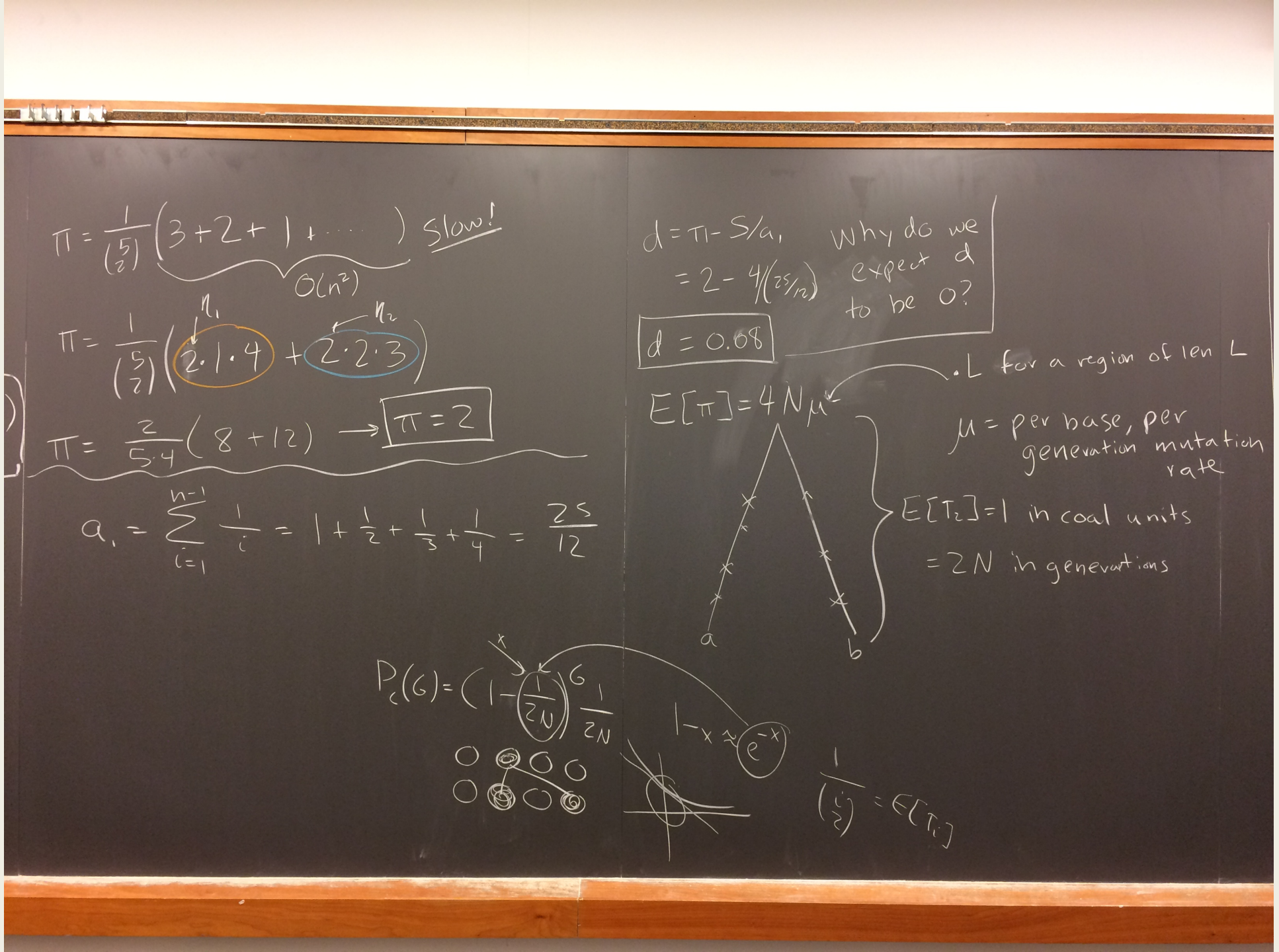
Note: some parts of the board were from questions after class, so I included both versions.



Note: some parts of the board were from questions after class, so I included both versions.



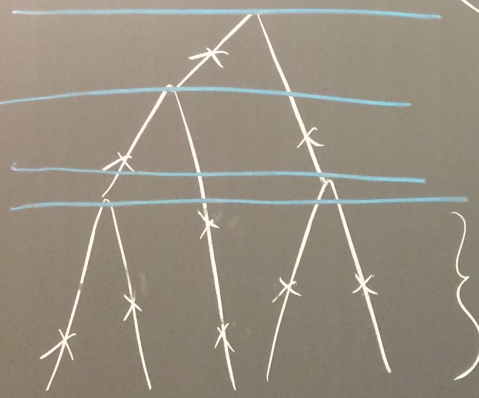
Note: some parts of the board were from questions after class, so I included both versions.



Note: some parts of the board were from questions after class, so I included both versions.

$E[S] = E[T_{total}] \cdot \mu$ ← $\cdot L$ for region len

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



5 lineages
T_5

T_{total} = total tree length
(sum of all branches)