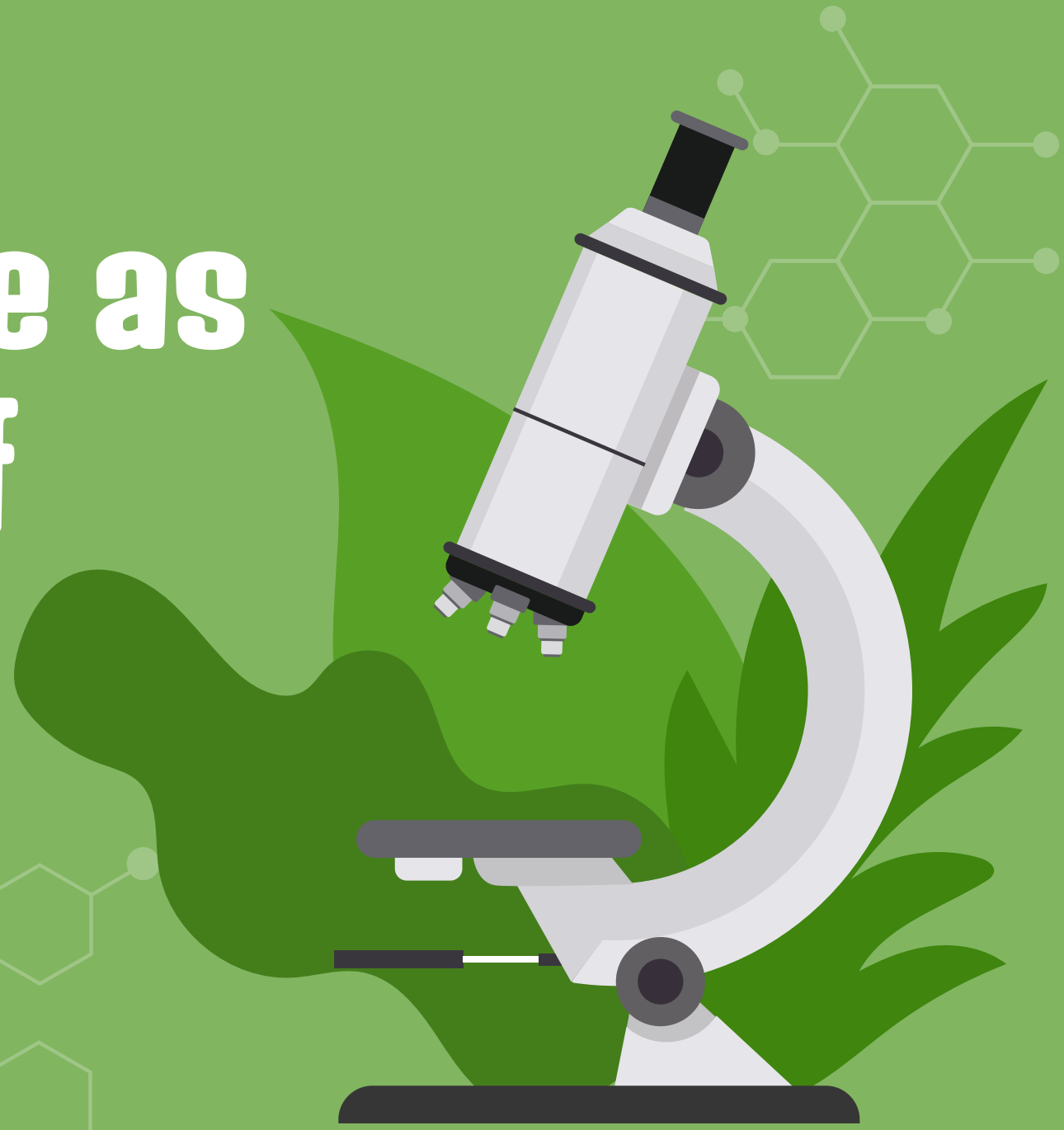
Genealogical Tree Distance as a Measure of Linkage

BIOL 1435 Final Project



Overview

01

Introduction/Motivation: Tree Distance and Linkage Disequilibrium

02

Methods: Ancestry Simulations

03

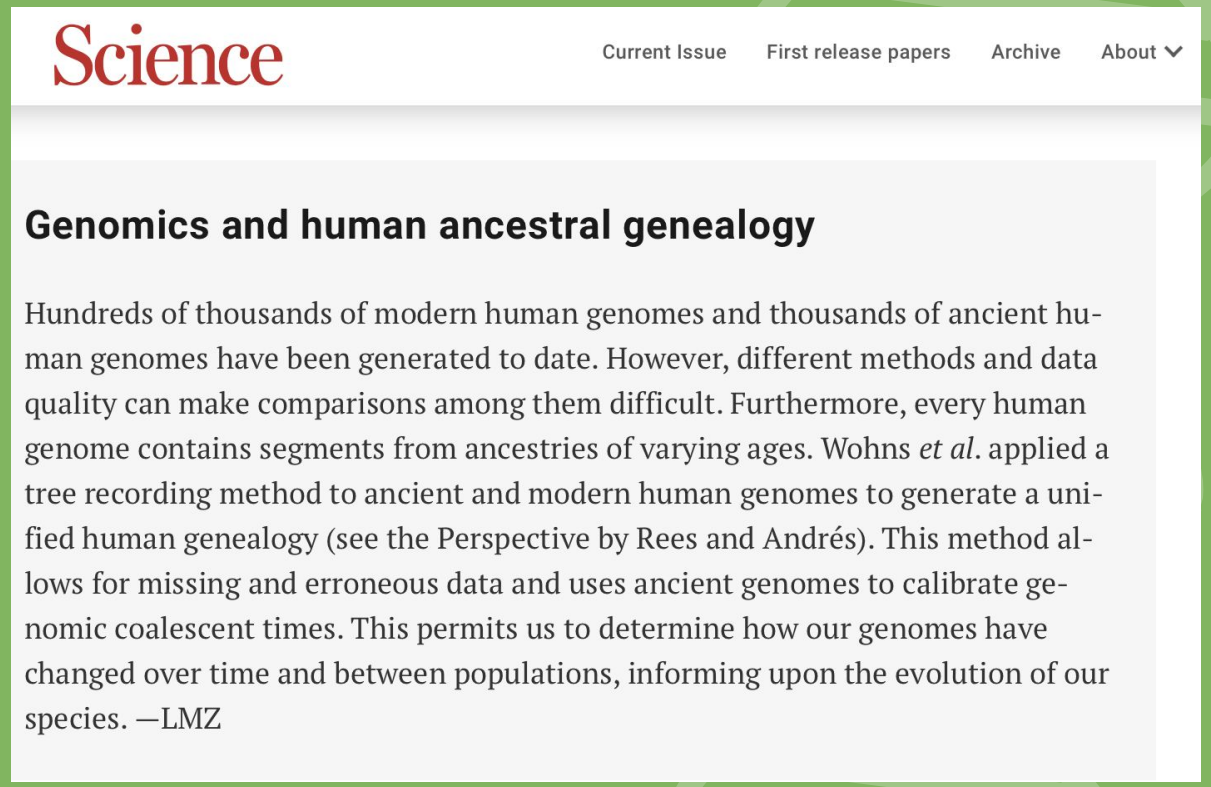
Results: Linkage Patterns

04

Discussion: Significance and Further Steps

Background

- Each region of the genome has a corresponding genealogical tree
- Use of tree sequences to represent whole genome genealogies is becoming increasingly popular relative to genotype data, which is storage intensive
- Search for tree-based metrics comparable to our genotype metrics

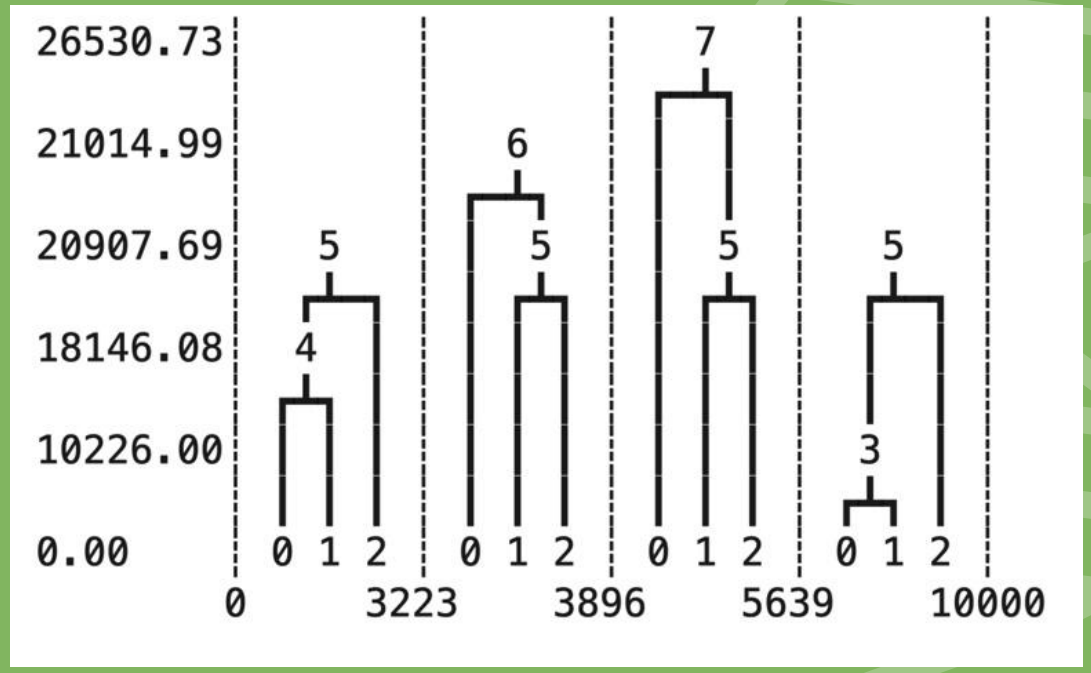


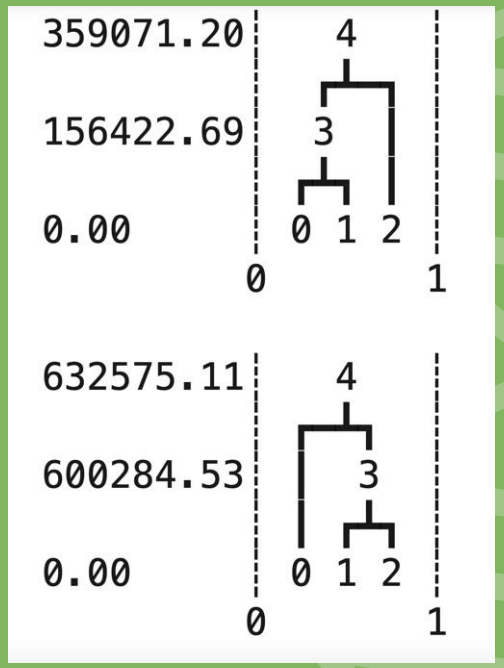
Use of a tree sequence structure to infer “a unified genealogy of modern and ancient samples”

Motivation

Trees for linked loci are expected to be more similar to each other than those for unlinked loci

Does this correlation between genealogical trees follow similar trends as the correlation of variants along the genome (linkage disequilibrium)?

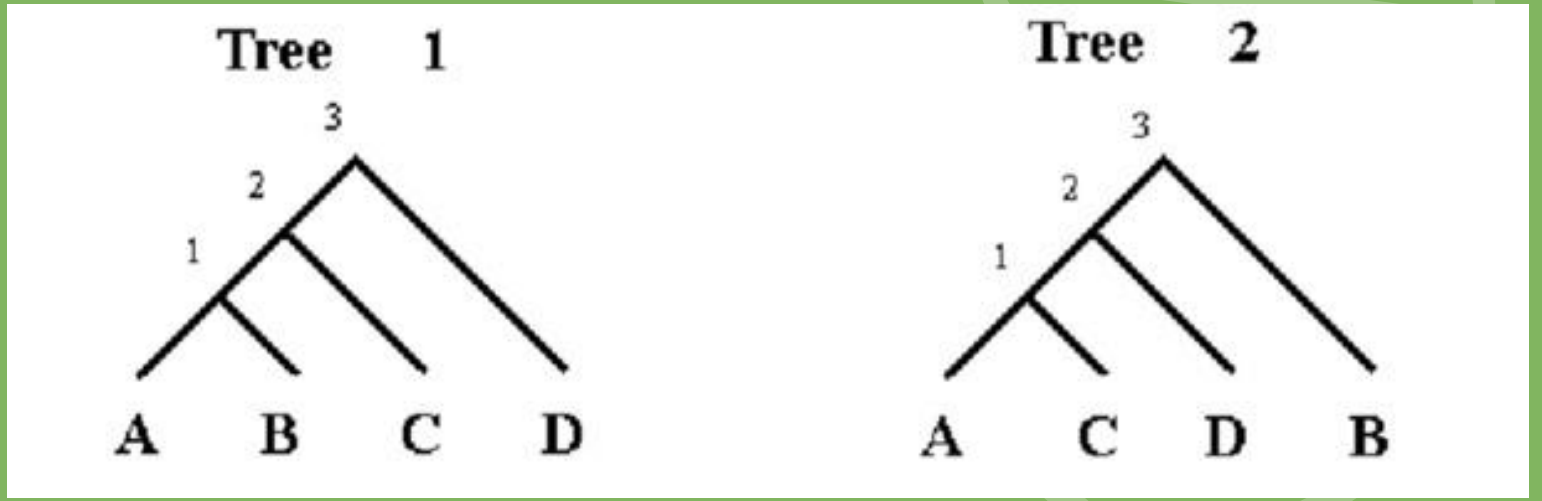




How do we measure “similarity” between genealogical trees?

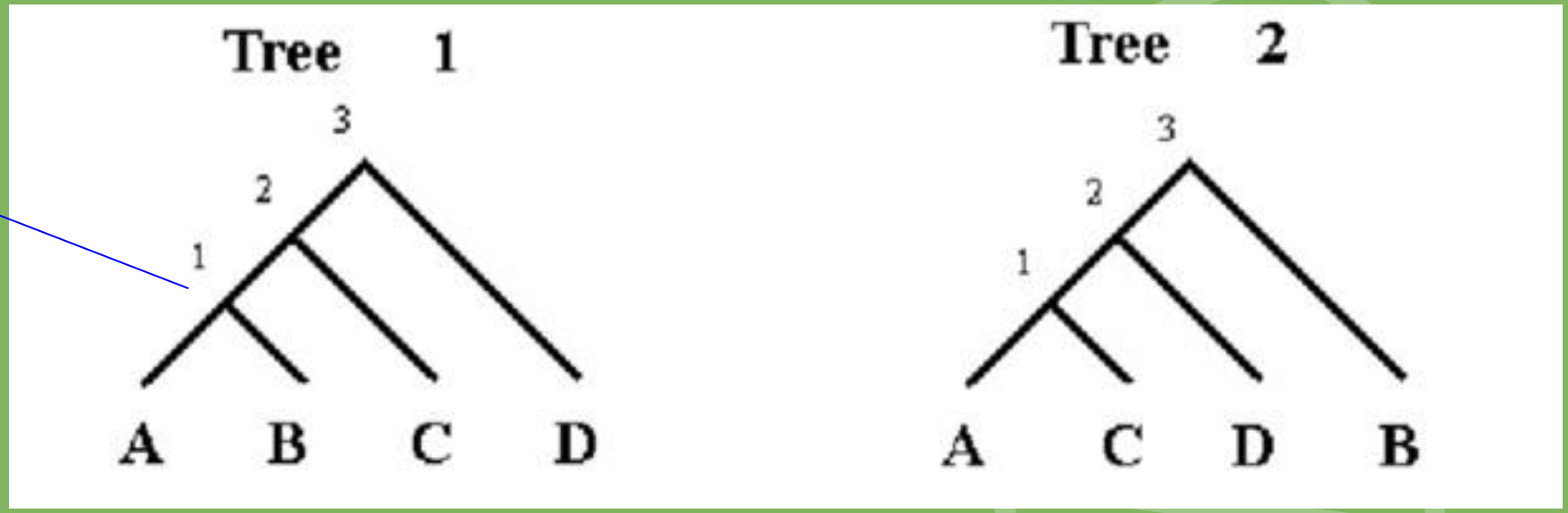
Robinson Foulds (RF) Distance

1) Make a list of the internal nodes of each tree and the clades that they define
2) Count the number of clades that exist in Tree 1 but not in Tree 2 and vice versa
3) Sum these counts to get the RF Distance



RF Distance

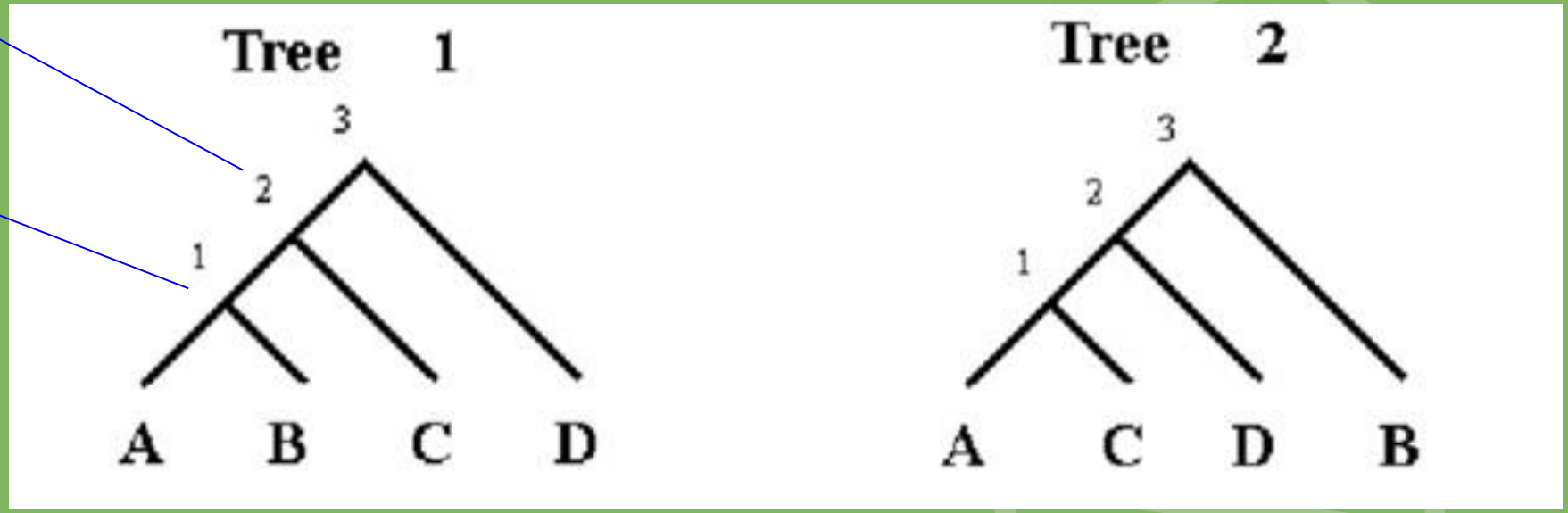
{A, B}



RF Distance

{A, B, C}

{A, B}

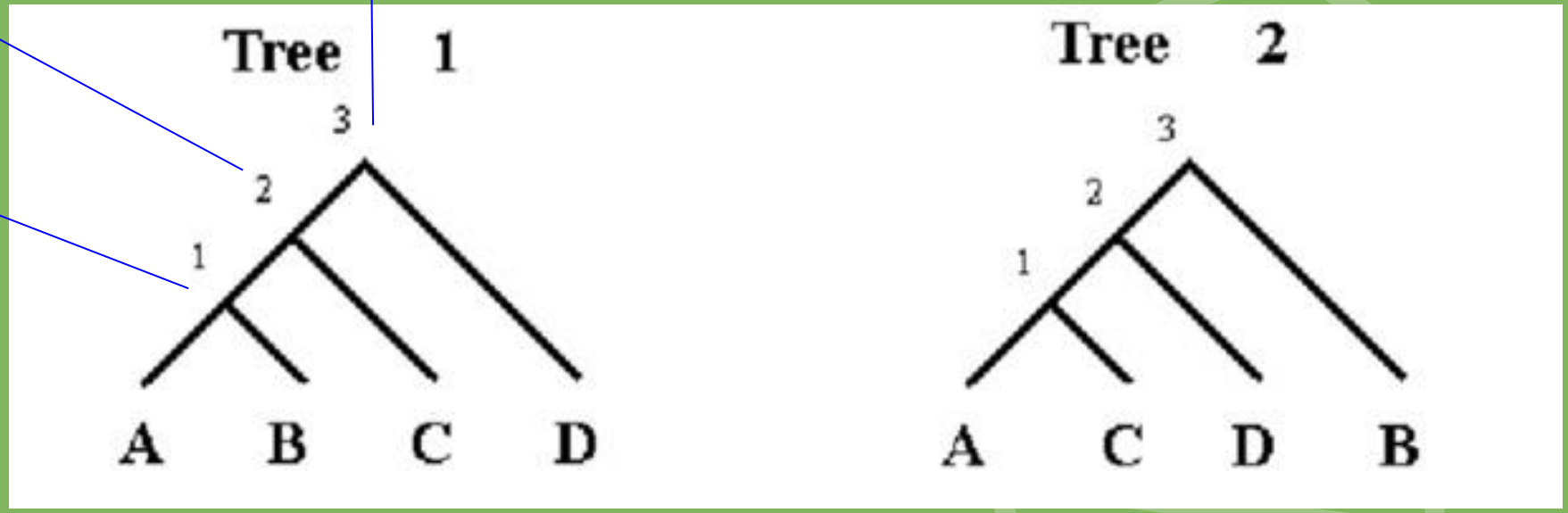


RF Distance

{A, B, C, D}

{A, B, C}

{A, B}



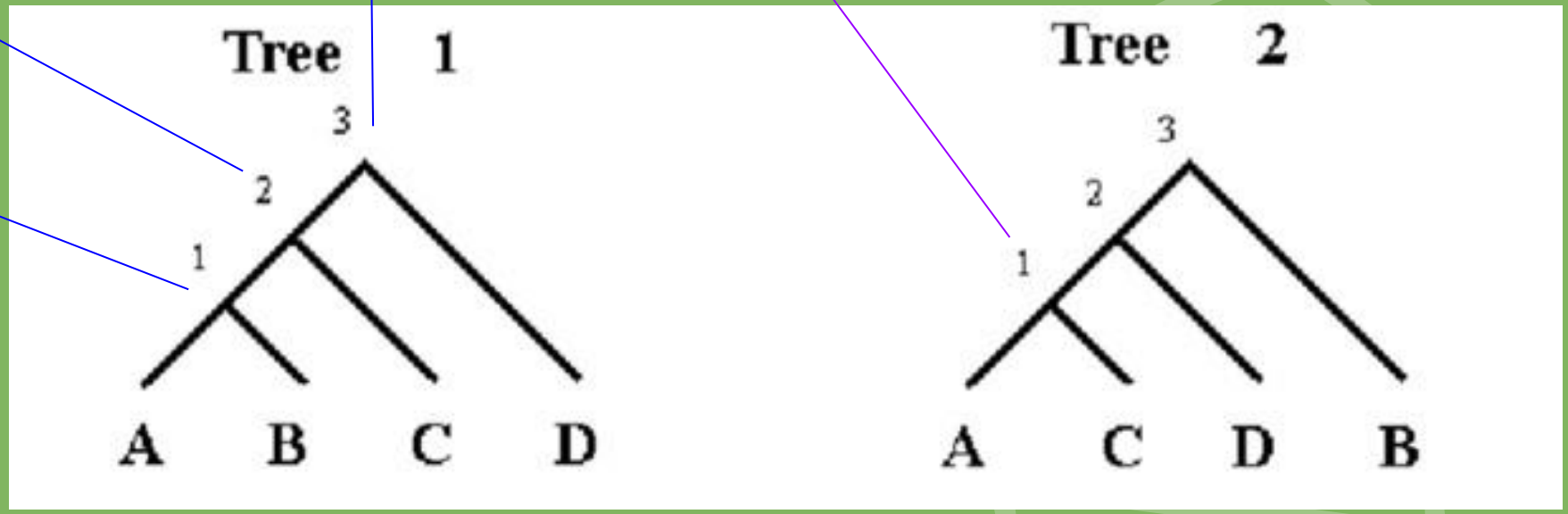
RF Distance

{A, B, C, D}

{A, C}

{A, B, C}

{A, B}

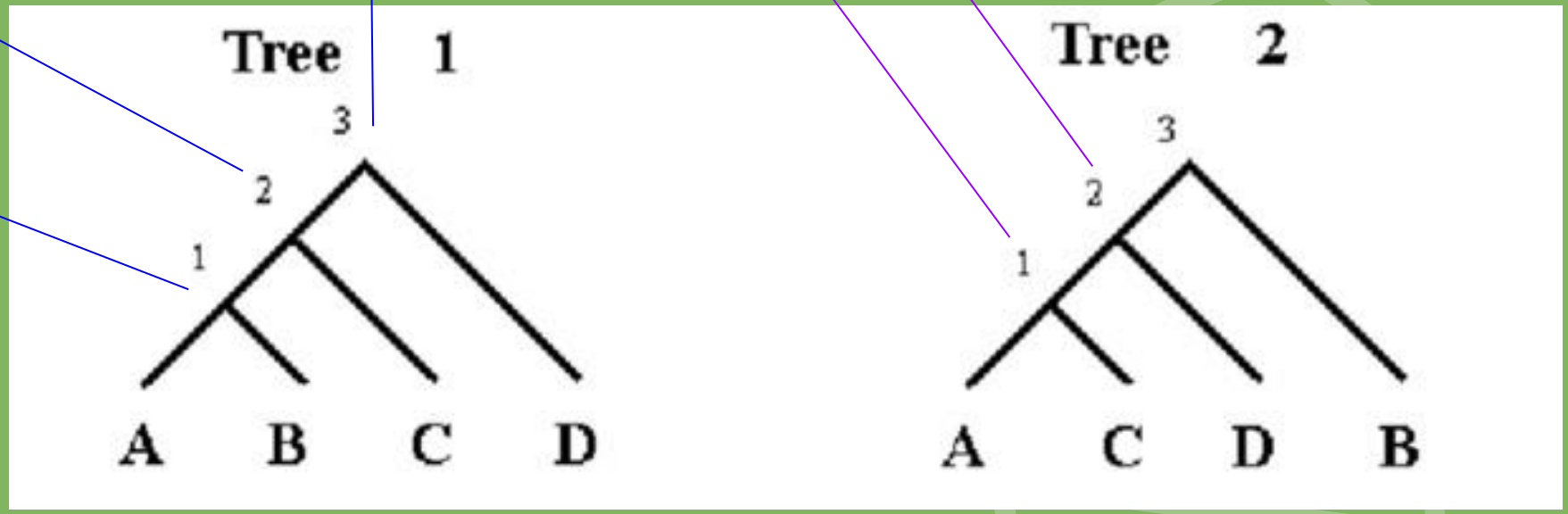


RF Distance

{A, B, C, D}

{A, B, C}

{A, B}



{A, C}

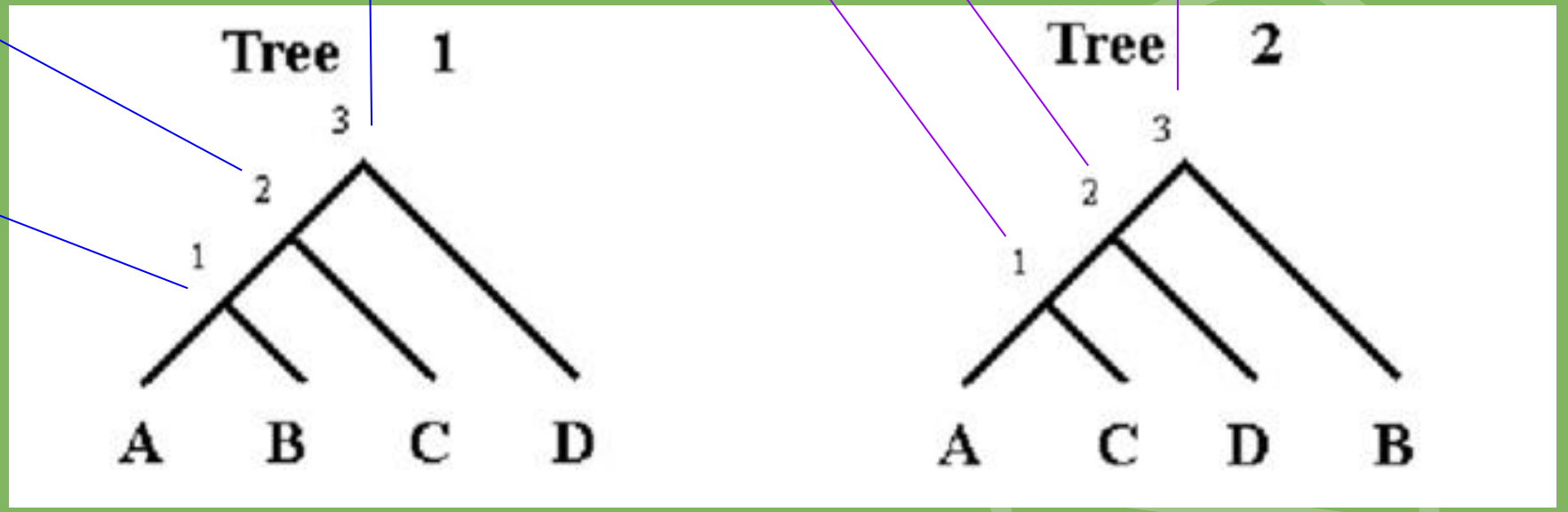
{A, C, D}

RF Distance

{A, B, C, D}

{A, B, C}

{A, B}



{A, C}

{A, C, D}

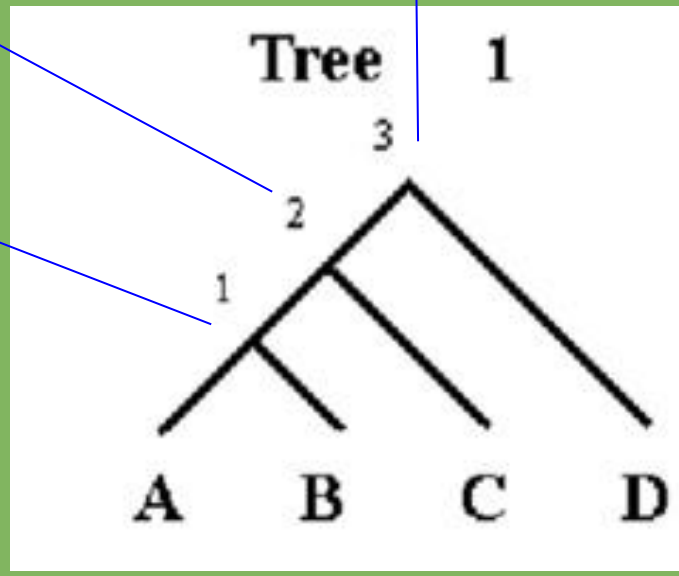
{A, B, C, D}

RF Distance

{A, B, C, D}

{A, B, C}

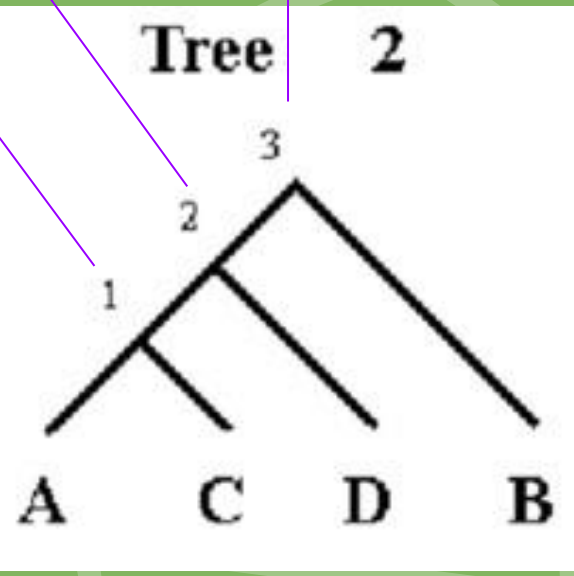
{A, B}



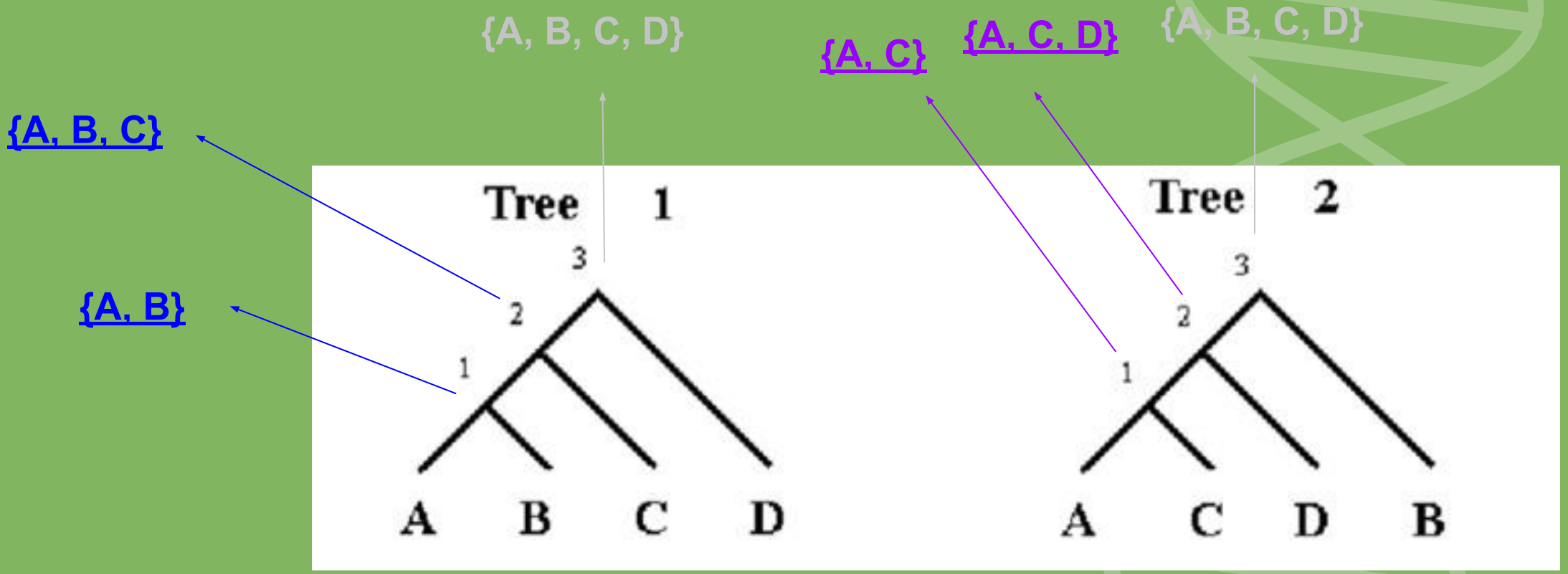
{A, C}

{A, C, D}

{A, B, C, D}



RF Distance = 4

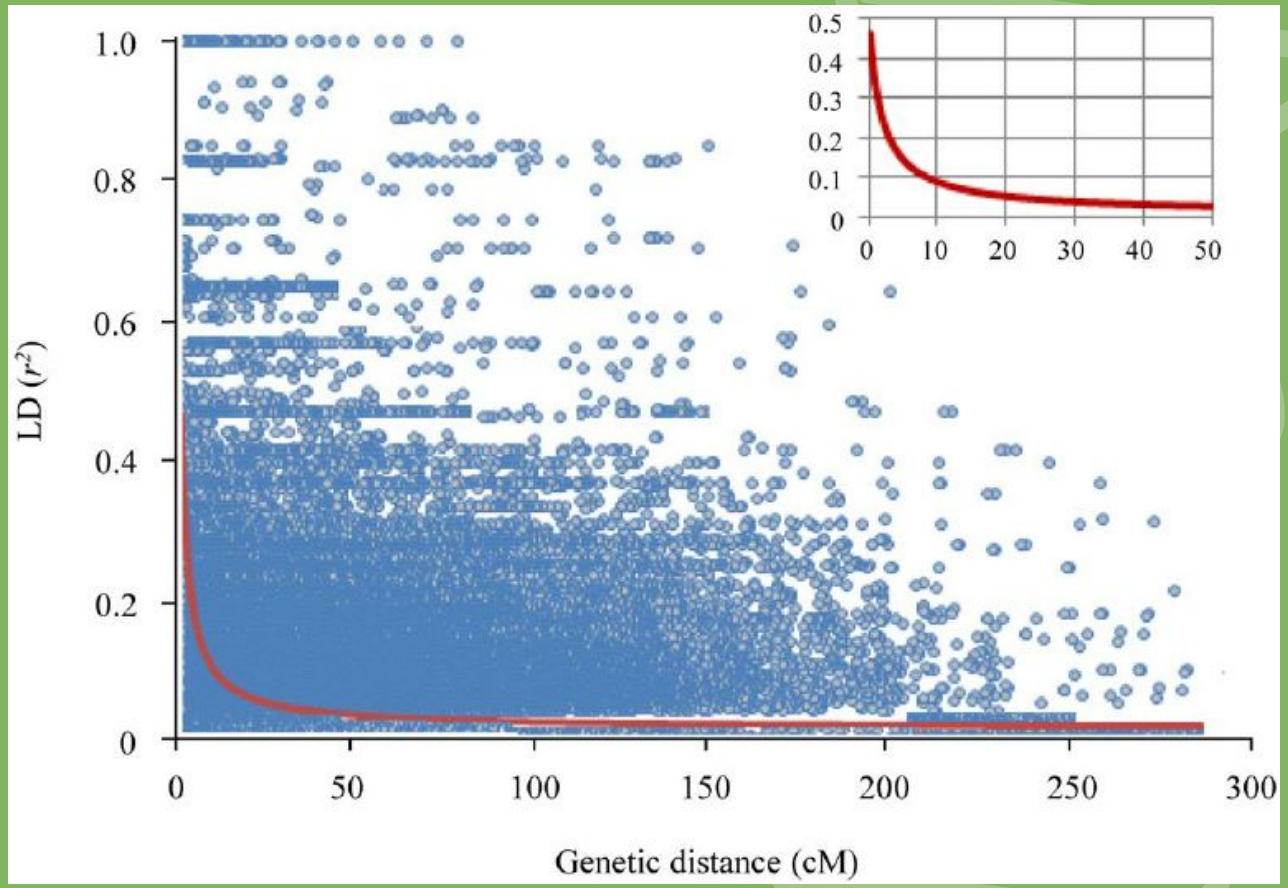


What do we know about linkage disequilibrium?

Correlation between two SNPs based on their allele frequencies

Decays with increasing distances between loci

Due to higher likelihood of recombination between loci that are farther apart



Methods

Ancestry simulations for linked and unlinked loci (100 samples, Recombination Rate = 10^-9 for linked loci)

Calculation of RF Distance between all pairwise combinations of trees

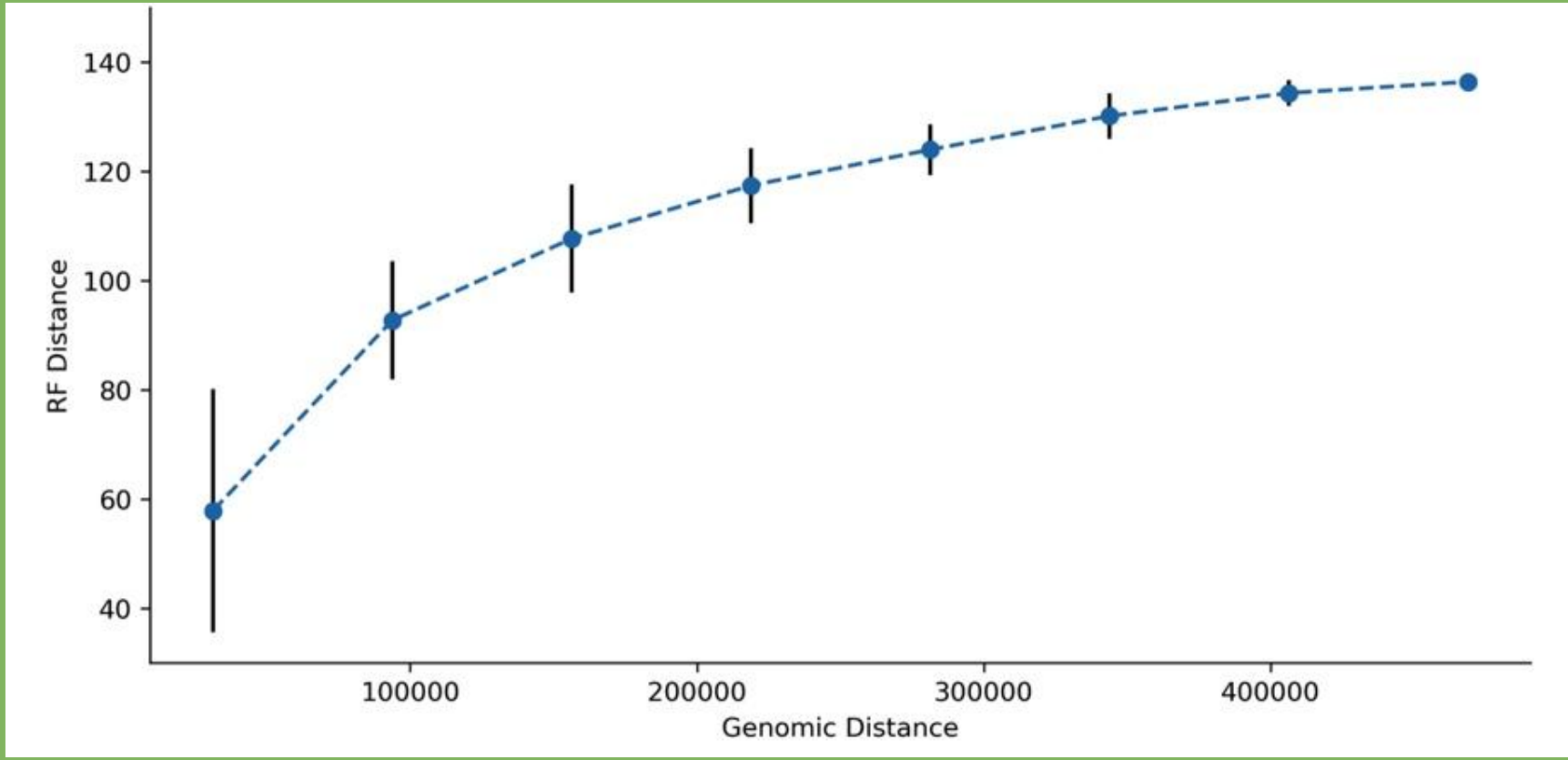
Plot of RF Distance as a function of the distance (in bp) on the chromosome between the two loci, using midpoint of each region

Introduce mutations in order to measure Linkage Disequilibrium in the same ancestry simulation

Results

- RF Distance increases as the distance between loci increases
- Avg RF Distance for unlinked trees: 195.55
- Max RF for trees with 100 samples: 196

RF Distance vs. Genomic Distance



Genealogical Linkage

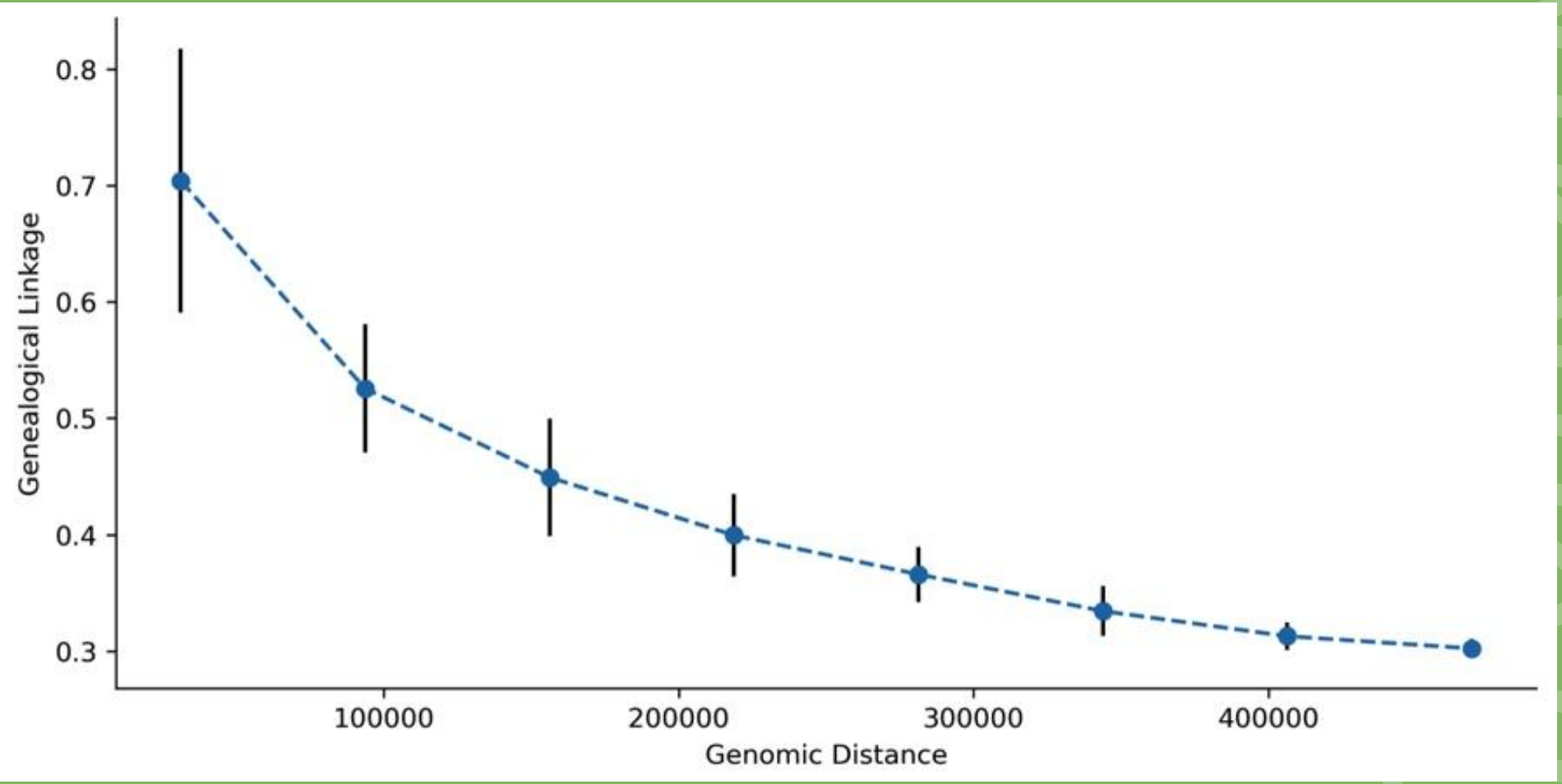
Genealogical Linkage =

1 - (RF Distance Linked / RF Distance Unlinked)

This statistic is normalized by the average difference between unlinked trees and inverted such that higher values represent more similarity and 0 represents no significant similarity between the trees

Results

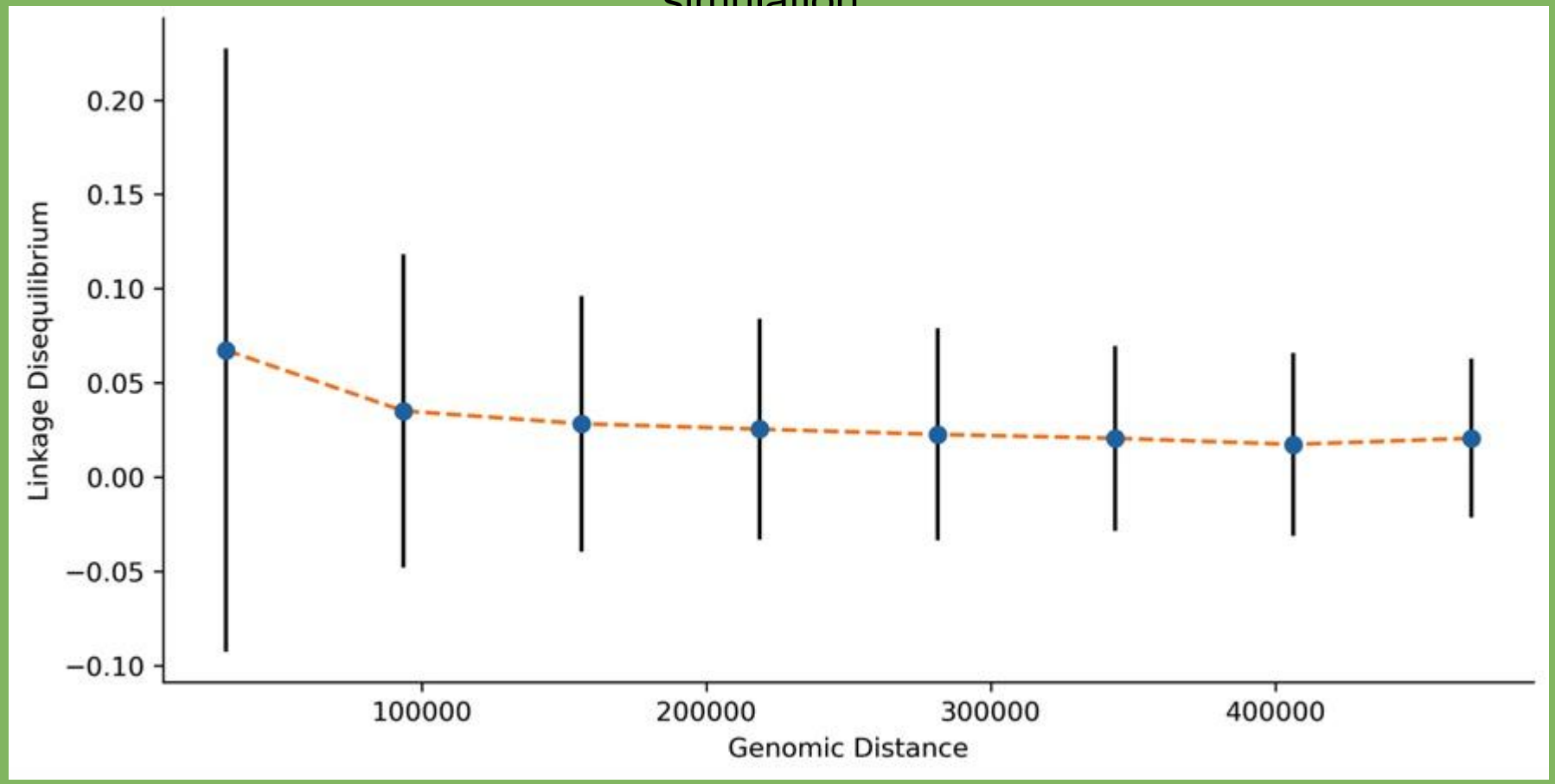
- GL decays with distance
- How does it compare to linkage disequilibrium?



Results

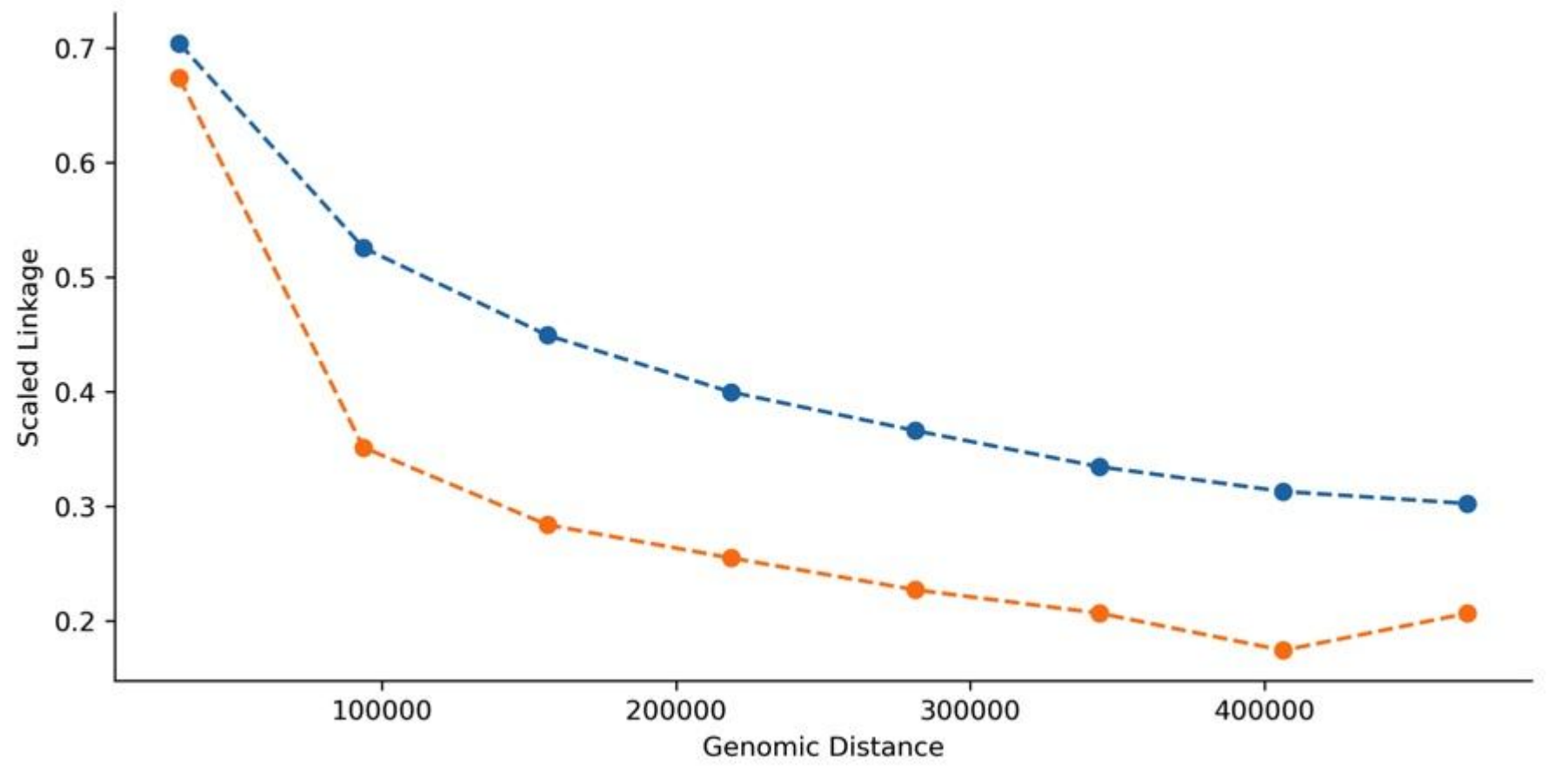
Linkage disequilibrium decays as distance between variants increases, as expected

Variance is high due to low recombination rate in relative to the effective population size in the simulation



Results

When scaled in order to fit on the same axes, the two statistics show similar trends as genome distance increases



Discussion

- Genealogical linkage follows similar trends as linkage disequilibrium
- It also has a lower variance, suggesting that it may be a more precise measure of linkage
- It also continues steadily increasing at genomic distances where linkage disequilibrium levels off
- Genealogical linkage is defining more fine scale trends in linkage

Further Steps

- Test how genealogical linkage changes with the simulation parameters (number of samples, recombination rate, effective population size) in order to more robustly prove its utility as a measure of linkage
- Branch Score !!!

References

McKenzie, P. F., & Eaton, D. A. (2020). The multispecies coalescent in space and Time. https://doi.org/10.1101/2020.08.02.233395