Data Set assembly 2

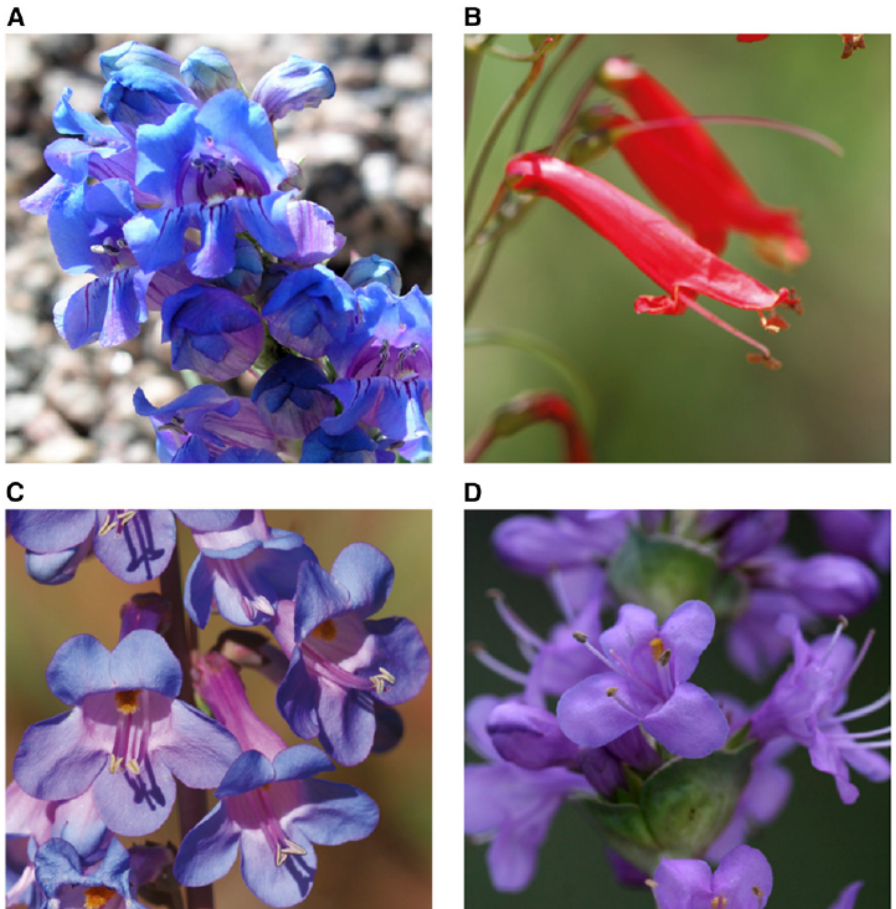
Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

Two case studies:
Phylogenetics of Penstemon
Tracing gonorrhea outbreaks

Phylogenetics of Penstemon using RADseq data

Question: How often have transitions between hummingbird and bee pollination occurred in Penstemon?



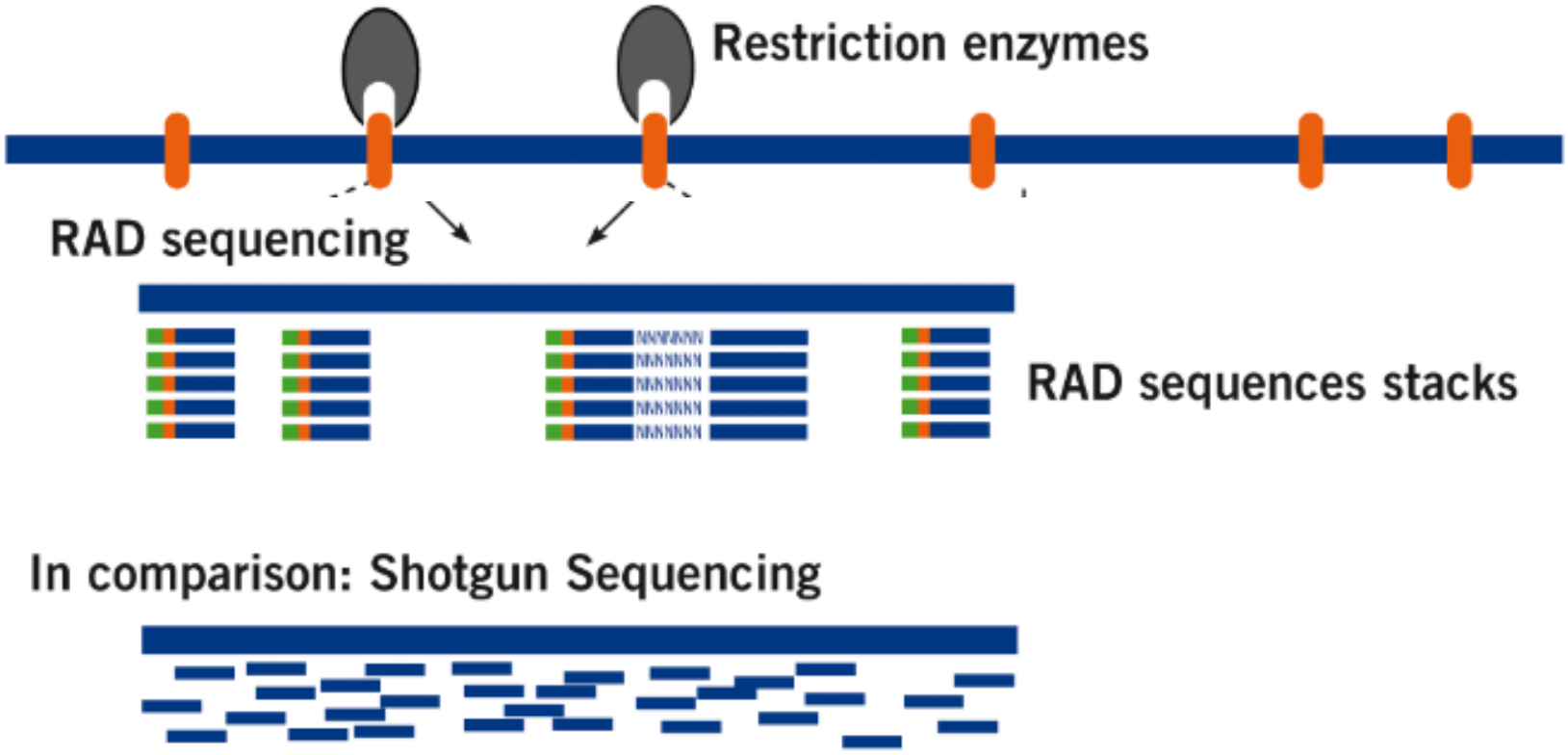
Data:

- Restriction site-associated DNA sequencing (RADSeq)
- 83 species, two samples per species
- No closely related reference genome

RADseq

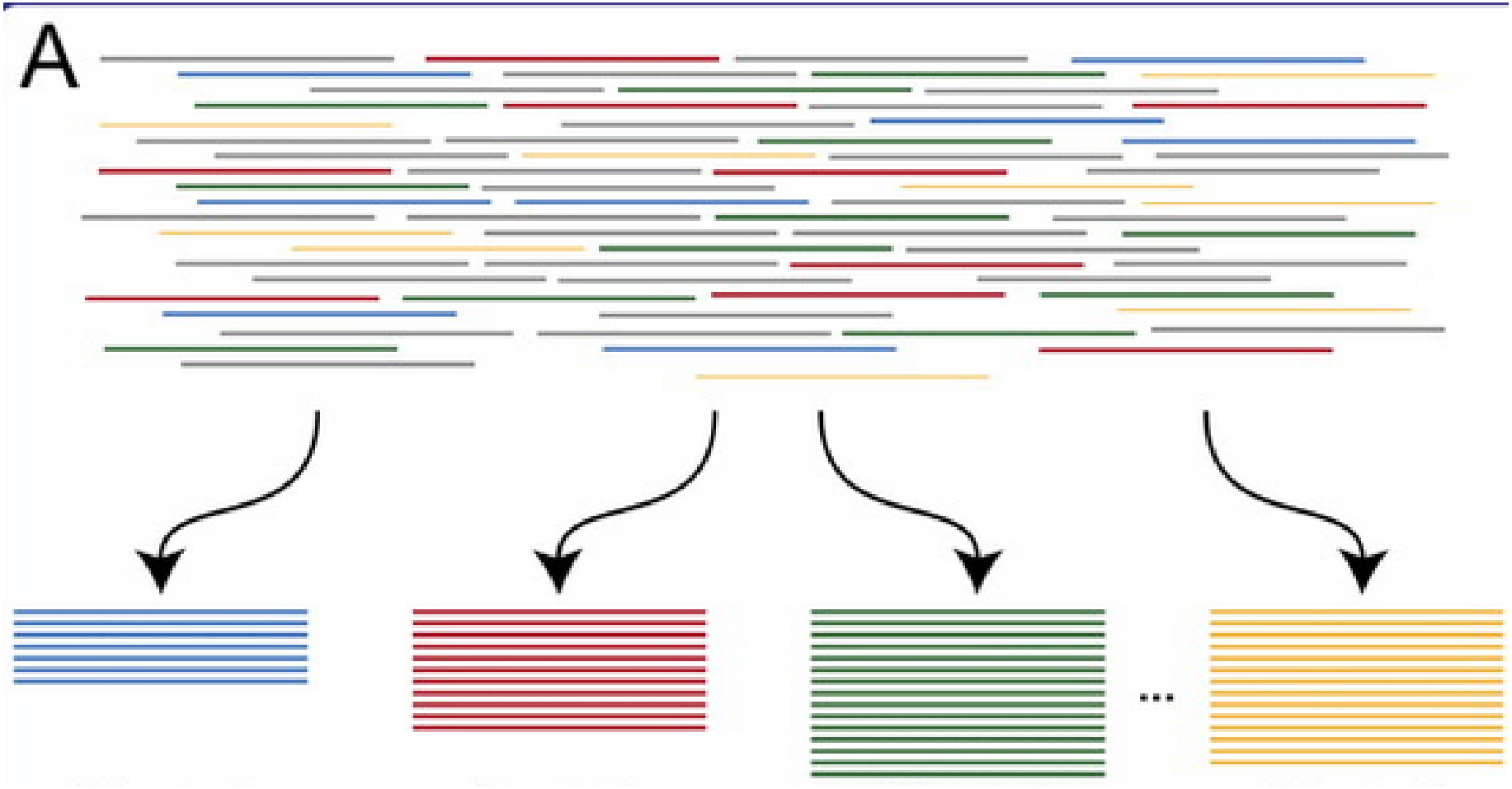
Uses restriction enzymes to fragment DNA

Targets sequencing to the same regions across taxa



(figures from floragenex.com)

In the absence of a reference genome, you need to cluster reads
A 'cluster' is an inference of homology



Clustered using Stacks (?)

Several factors can cause drop-out of alleles in RAD-seq data (i.e. not observing homologous alleles)

- Mutations at restriction digest sites
- Clustering parameters exclude homologous regions
- Low coverage

There have been many conflicting studies on the importance of missing data in phylogenetic analyses,
broadly, as long as missing data is random, it shouldn't be very problematic, but phylogenetically-biased missing data is likely to be. (Roure et al., 2013; Lemmon et al., 2009)

Missing data in RADseq can mislead inference

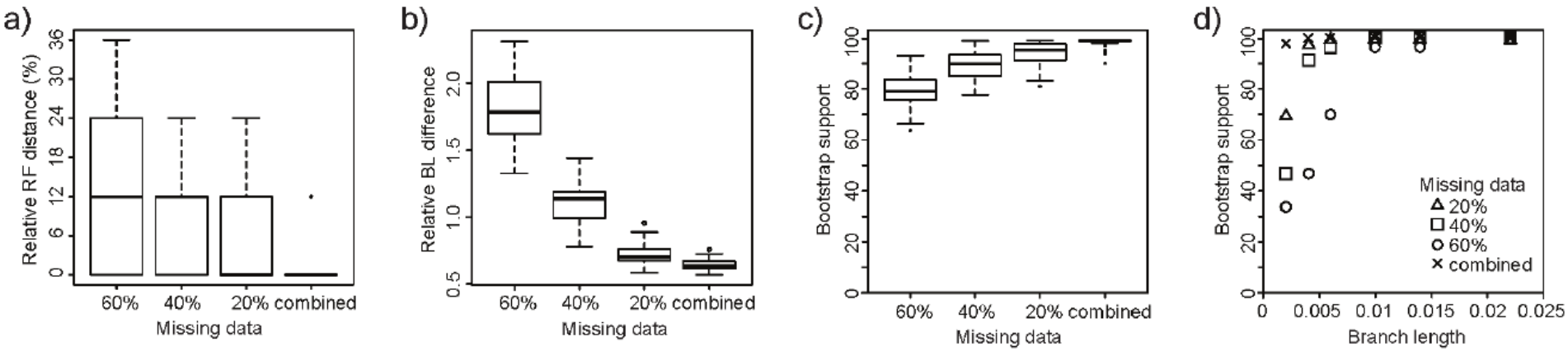
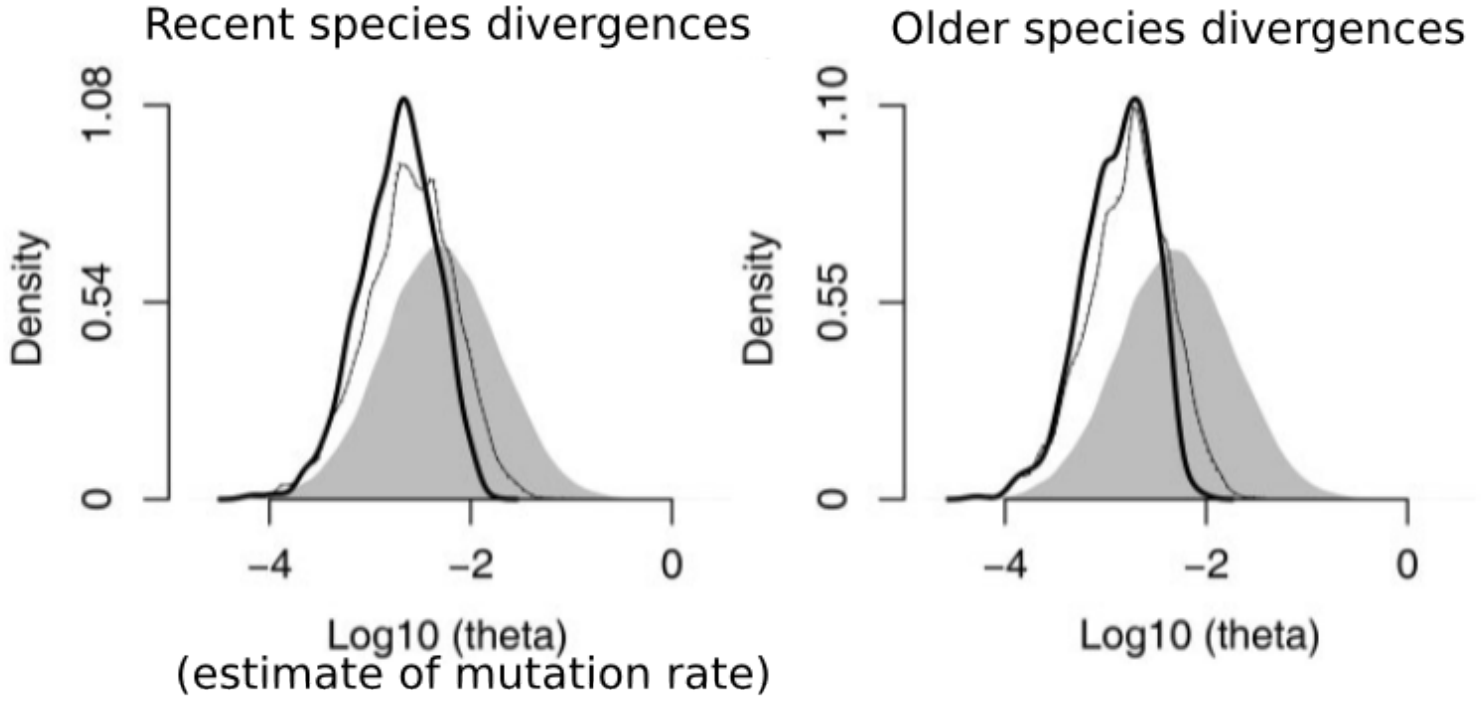


Figure 4: Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support for shorter branches (d).

(Leaché et al., 2015)

But excluding sites with high levels of missing data doesn't solve the problem.

But excluding sites with high levels of missing data doesn't solve the problem.
It biases rate estimation downwards by preferentially removing high rate loci



Gray shading is simulated rates, dashed line is shift due to loss of RAD sites, black line is shift due to loss of cut sites, black line shift due to loss of cut sites + post sequencing processing.

(Huang and Knowles, 2014)

Advice?

Advice?

"Given that the data matrix reflects complex interactions between aspects of library construction and processing with the divergence history itself, our results also suggest that general rules-of-thumb are unlikely."

(Huang and Knowles, 2014)

Advice?

"Given that the data matrix reflects complex interactions between aspects of library construction and processing with the divergence history itself, our results also suggest that general rules-of-thumb are unlikely."

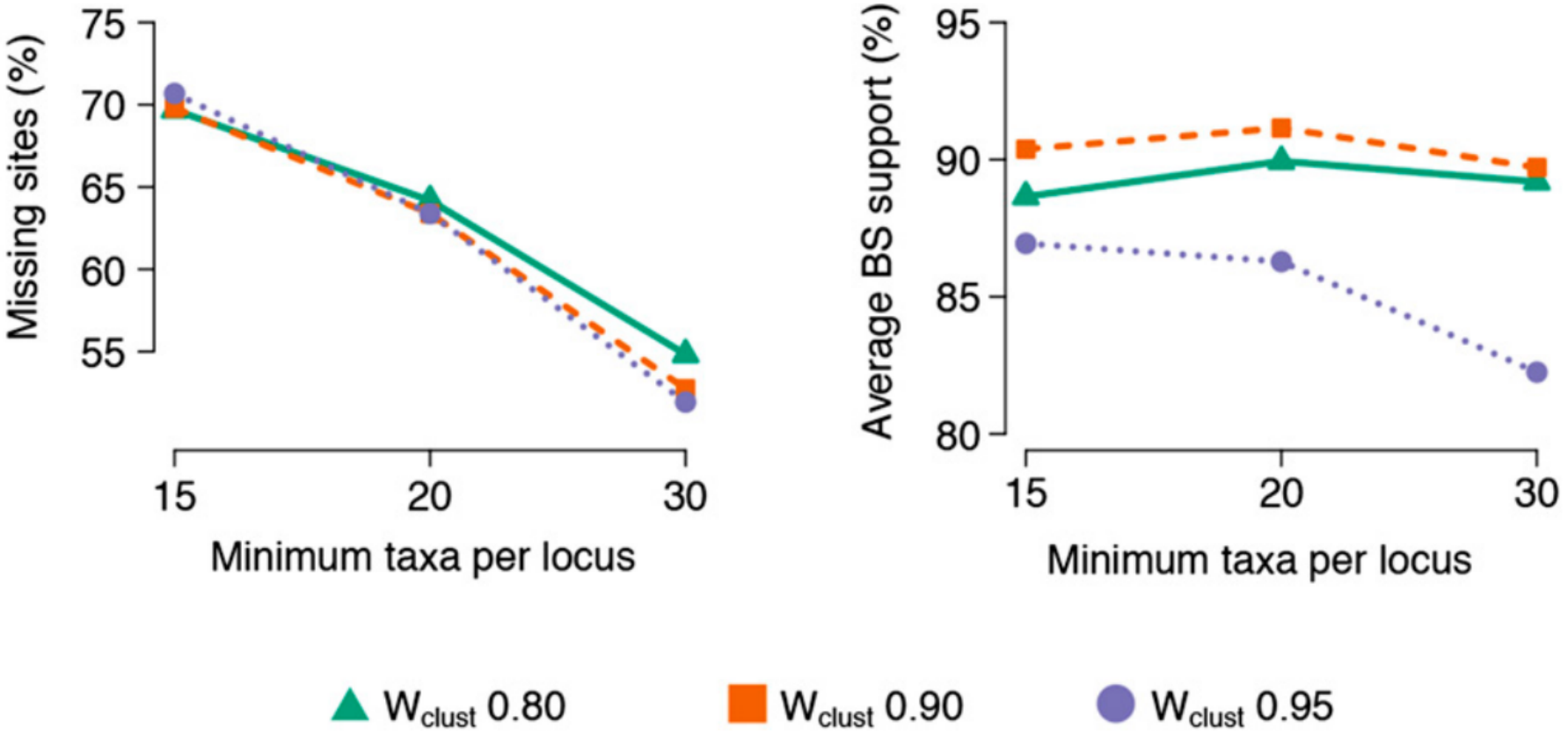
(Huang and Knowles, 2014)

Tradeoffs:

- Decreasing similarity cutoff captures more loci shared across the tree, at risk of incorrect homology
- Decreasing taxon representation threshold allows you to capture more loci, but representing fewer individuals

Approach

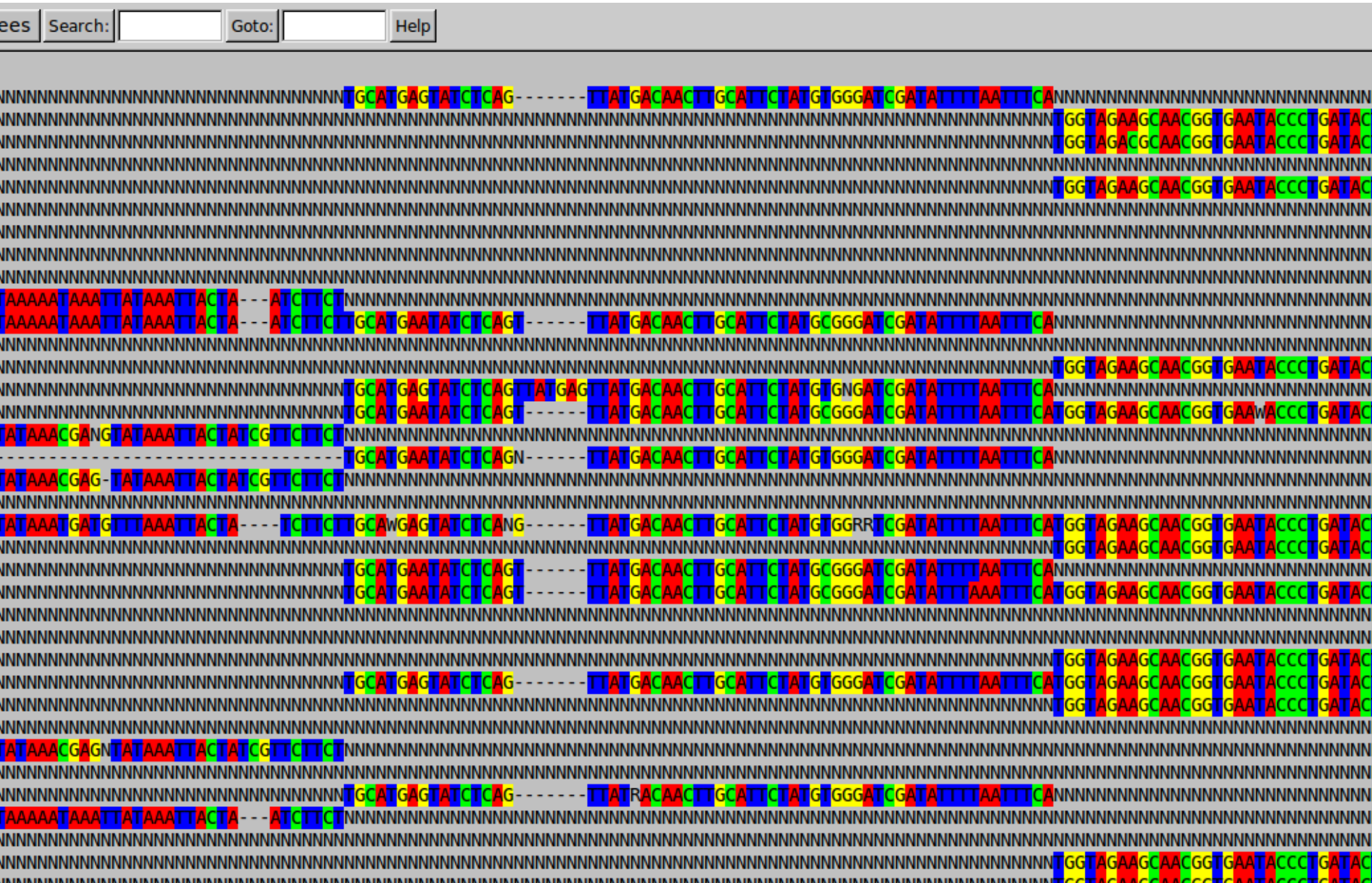
Investigate a range of parameters



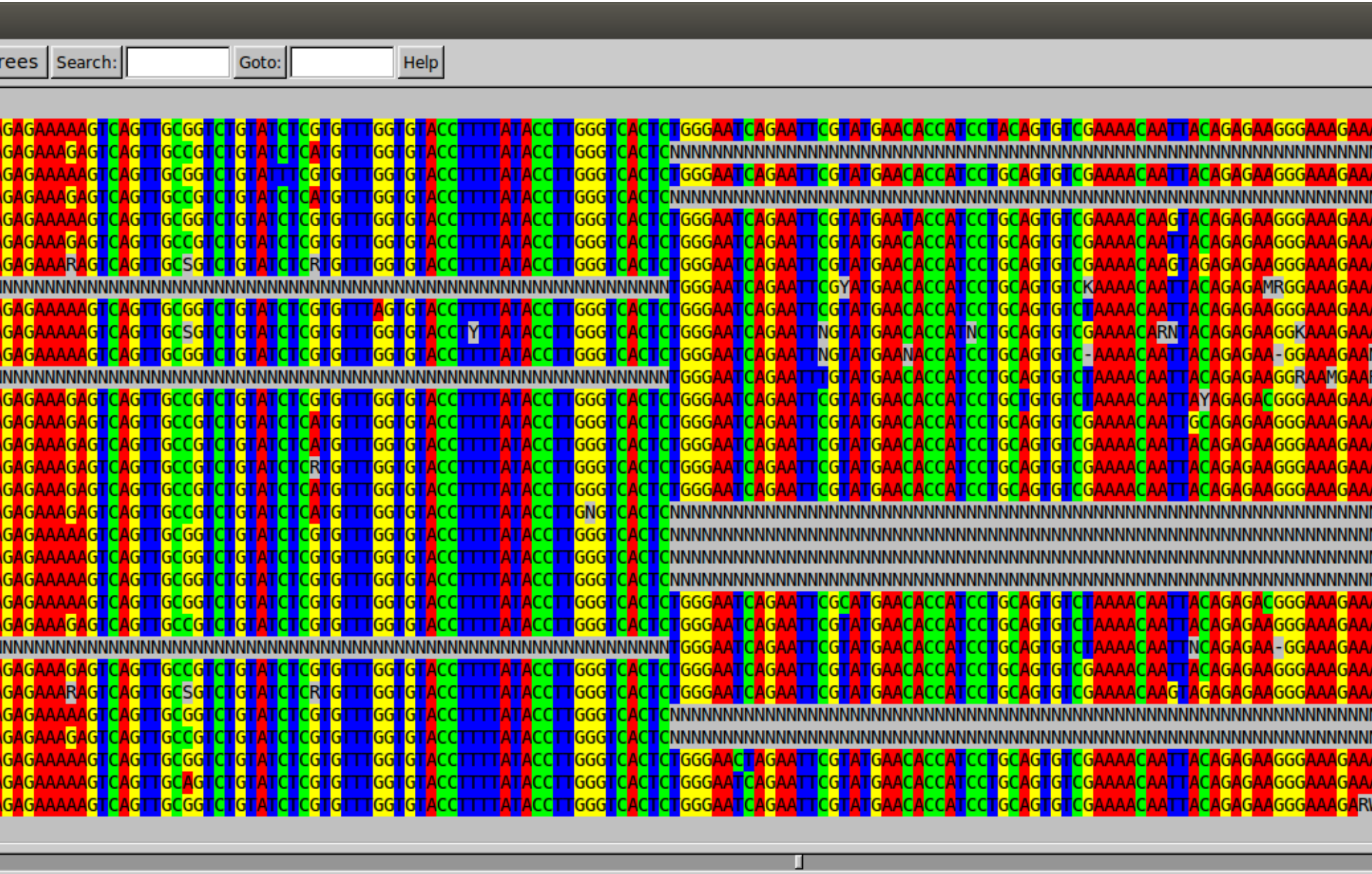
(Wessinger et al., 2016)

Missing data is phylogenetically biased

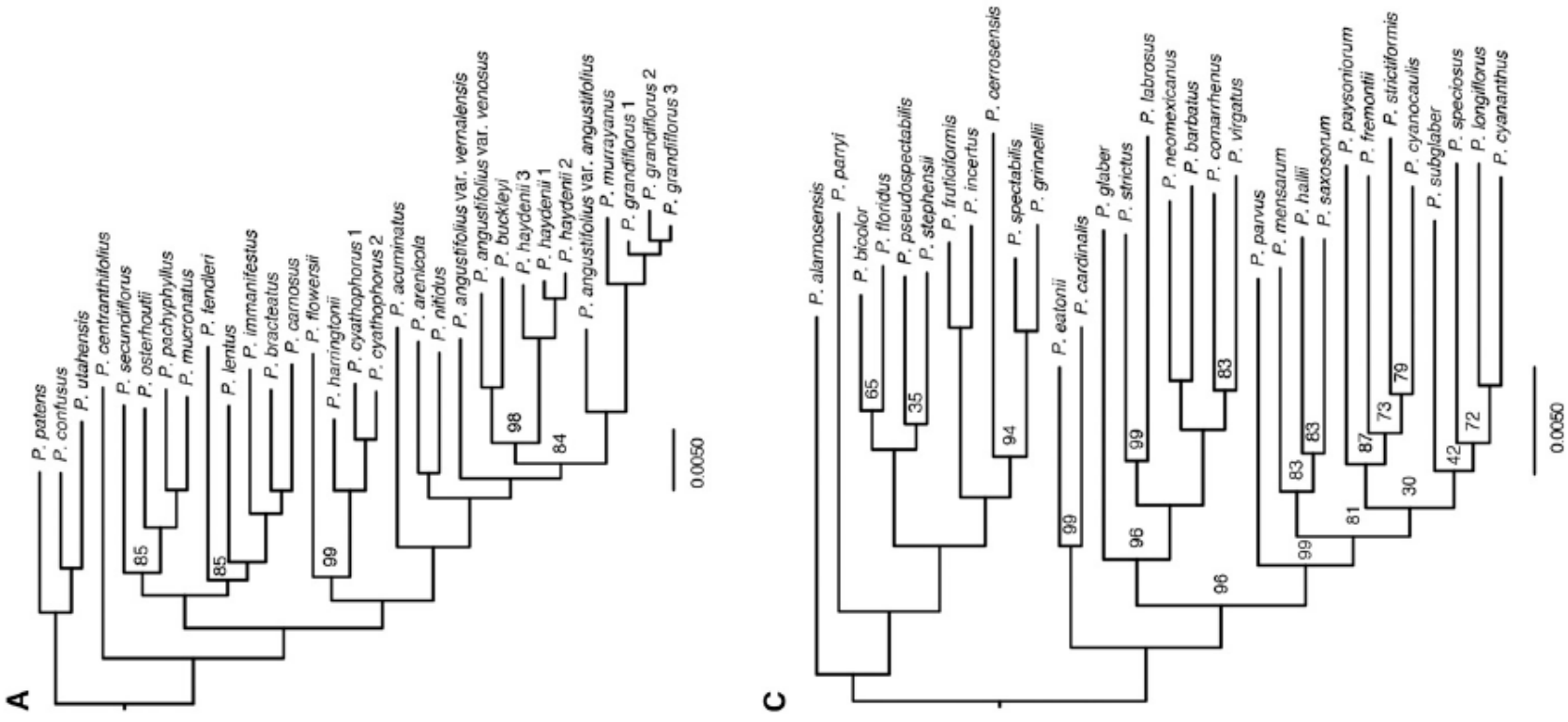
Across full dataset, many loci are only found in one of the major clades



Variation within clades is better captured by dividing the data set and clustering separately



Build (and report!) multiple trees using different filtering parameters



Trees from separate clade analyses (Wessinger et al., 2016)

Summary:

Bias:

Clustering parameters drive non-random missing data

Potential effect on inference:

No topological resolution
Tip branch lengths are shortened
Non-homologous regions align

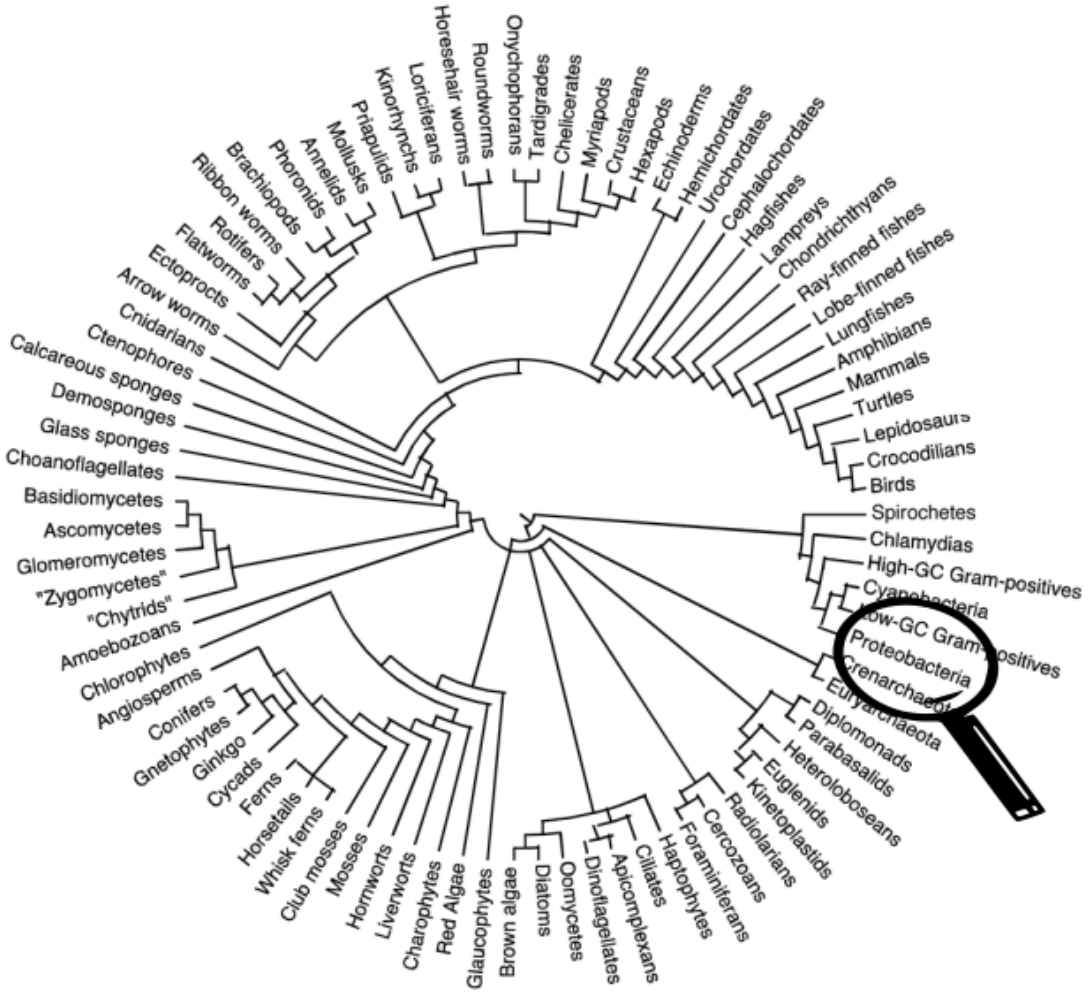
Mitigation:

Estimate relationships under a range of filtering parameters

Conclusions:

Branch lengths and bootstrap support differ across filtering parameters
Different data sets may be appropriate at different phylogenetic scales
Evolutionary inferences about pollinator shifts need to be robust to this uncertainty

Case study - tracing gonorrhea outbreaks

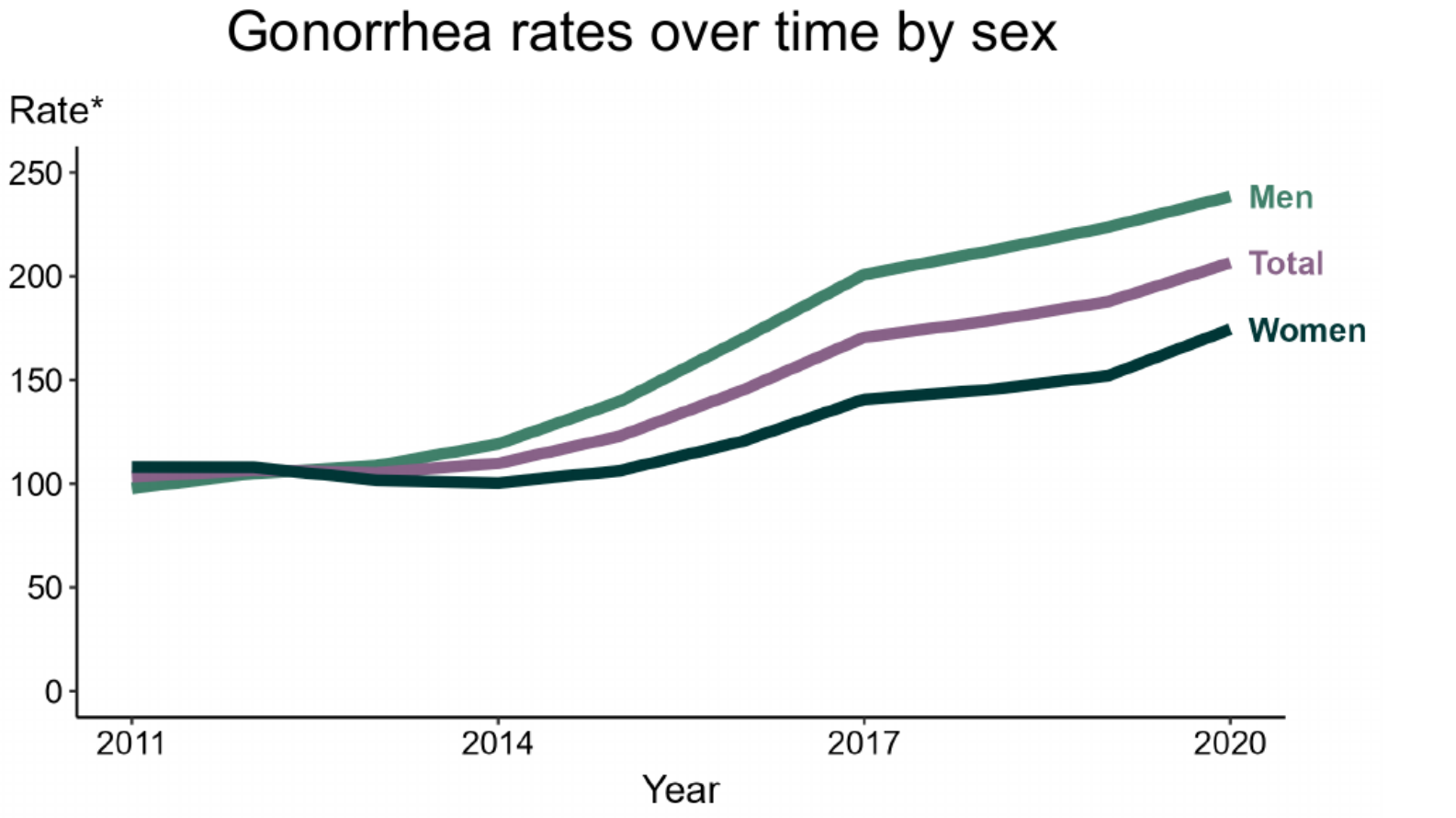


Rapid phylogenetic updating to trace gonorrhea outbreaks

Collaboration with
Jack Cartee CDC, Jeanine Abrams-McLean CDC, and Jasper Toscani Field
(PhD student, UC Merced)

Neisseria gonorrhoeae

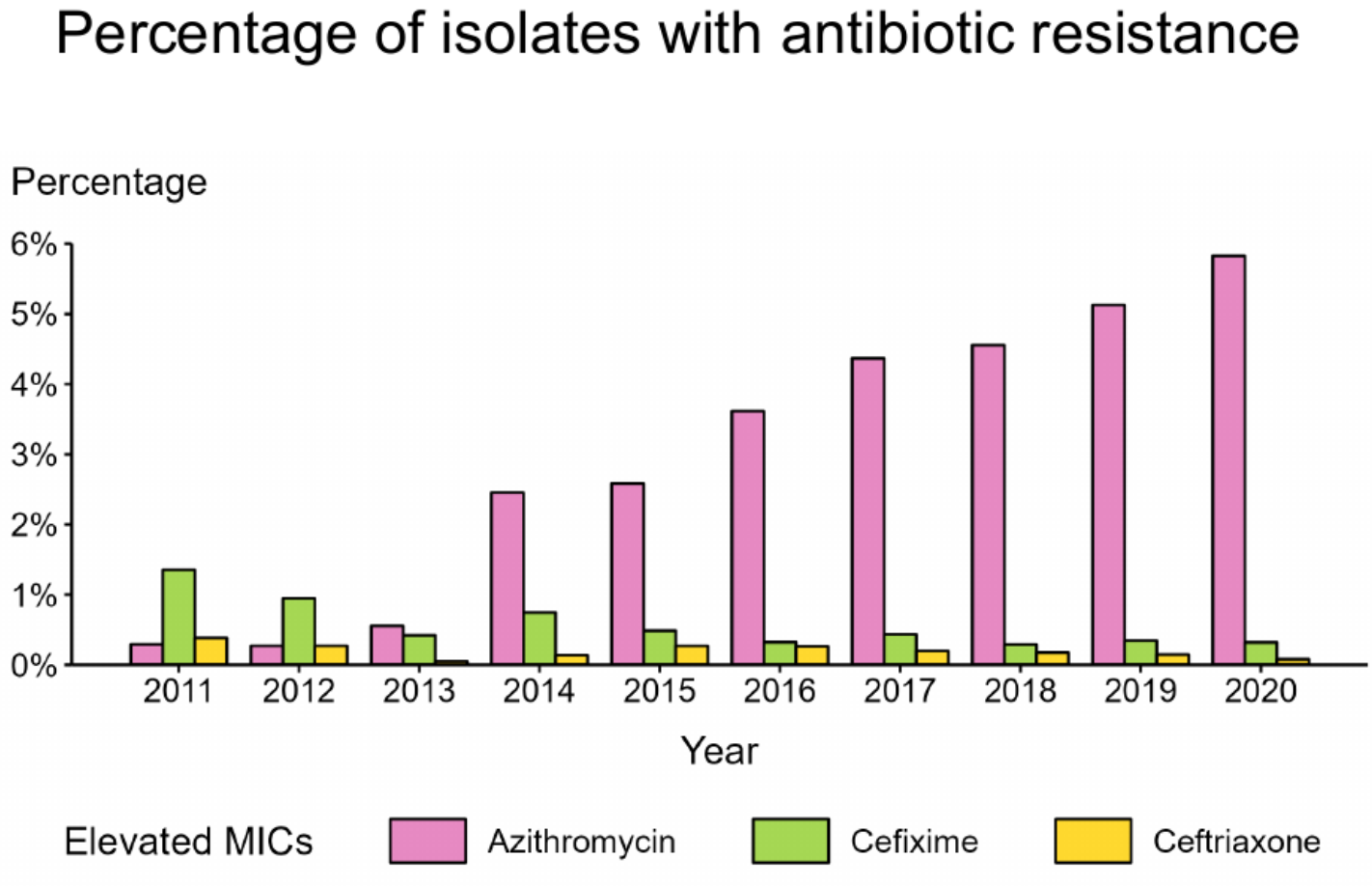
- Gram-negative, diplococci bacteria
- Responsible for the sexually transmitted infection known as gonorrhea
- One of two pathogenic *Neisseria* species known to infect humans
- WHO estimated 82 million new cases among adults worldwide in 2020



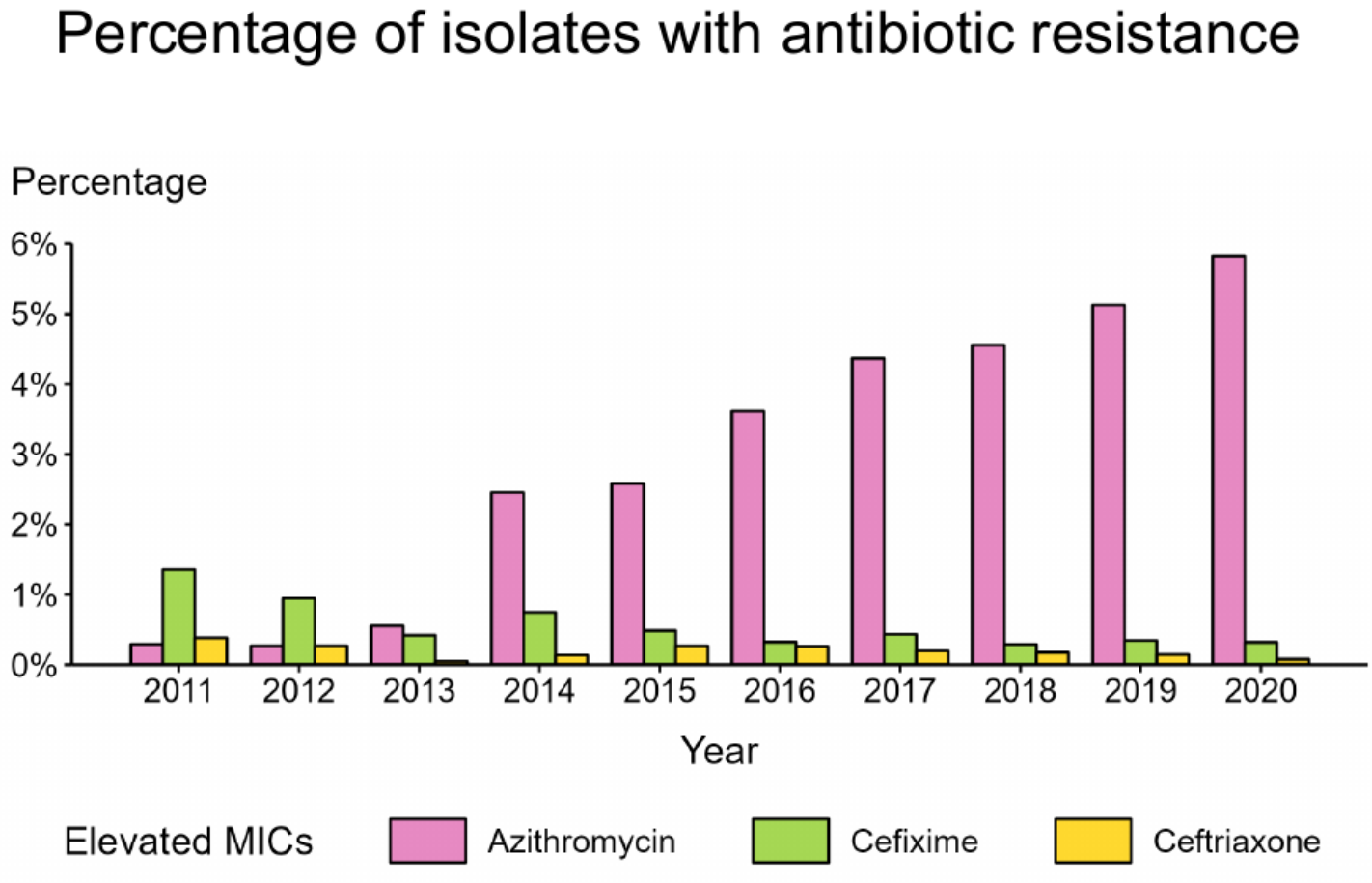
https://www.cdc.gov/std/statistics/2020/figures/GC-2.htm

Recent increase in rates of gonorrhea infections

Neisseria gonorrhoeae has progressively developed resistance to each single dose antibiotic.



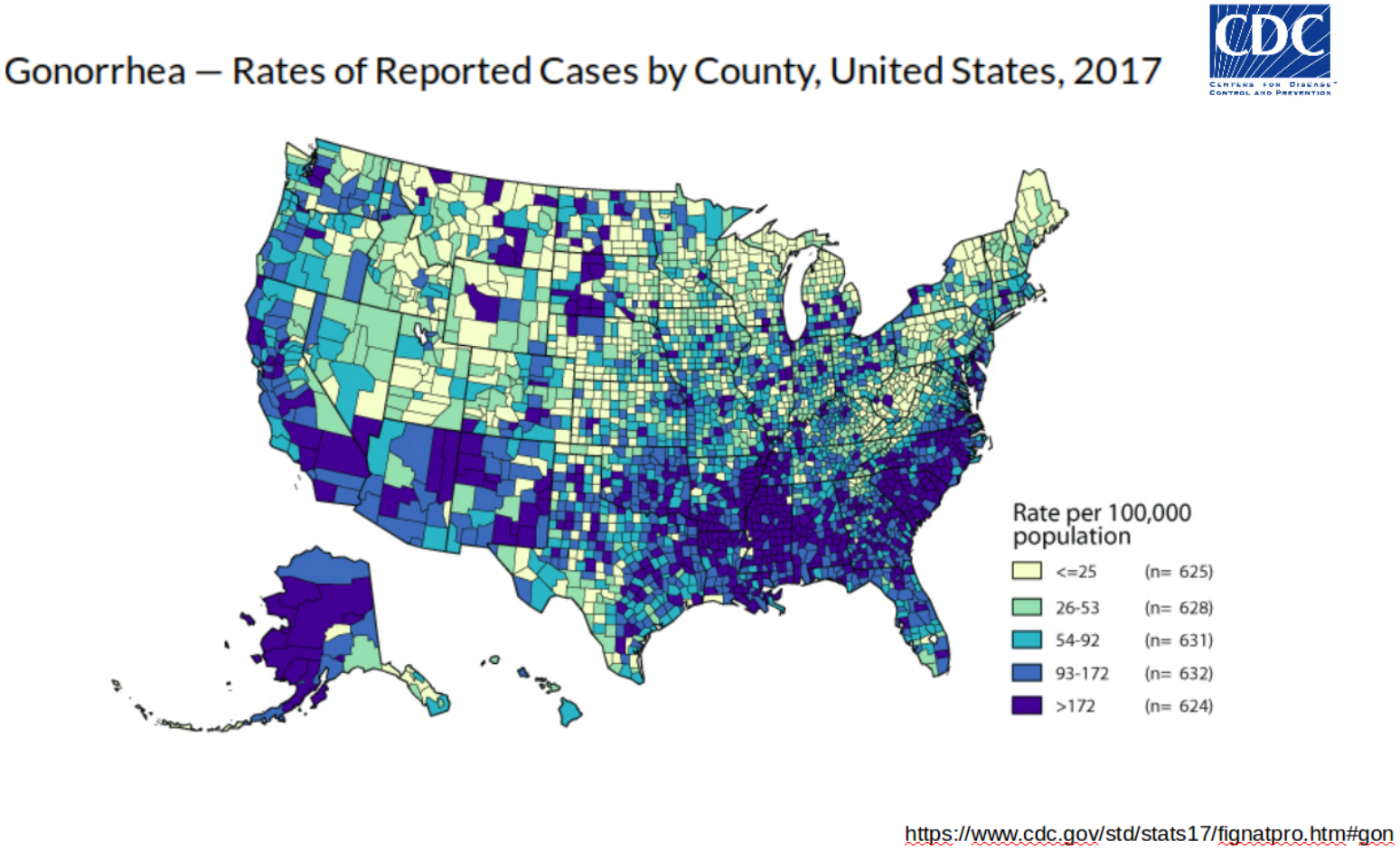
Neisseria gonorrhoeae has progressively developed resistance to each single dose antibiotic.



Only remaining recommended treatment option is dual therapy with a ceftriaxone plus azithromycin

"It is widely recognised that few antimicrobials remain effective in the treatment of *Neisseria gonorrhoeae* infection and that gonorrhoea could become untreatable in the future."
(Chisholm et al. Sex Transm Infect 2015)

To track and control outbreaks, the CDC is tracing evolutionary history of gonorrhea, across the US and globally.

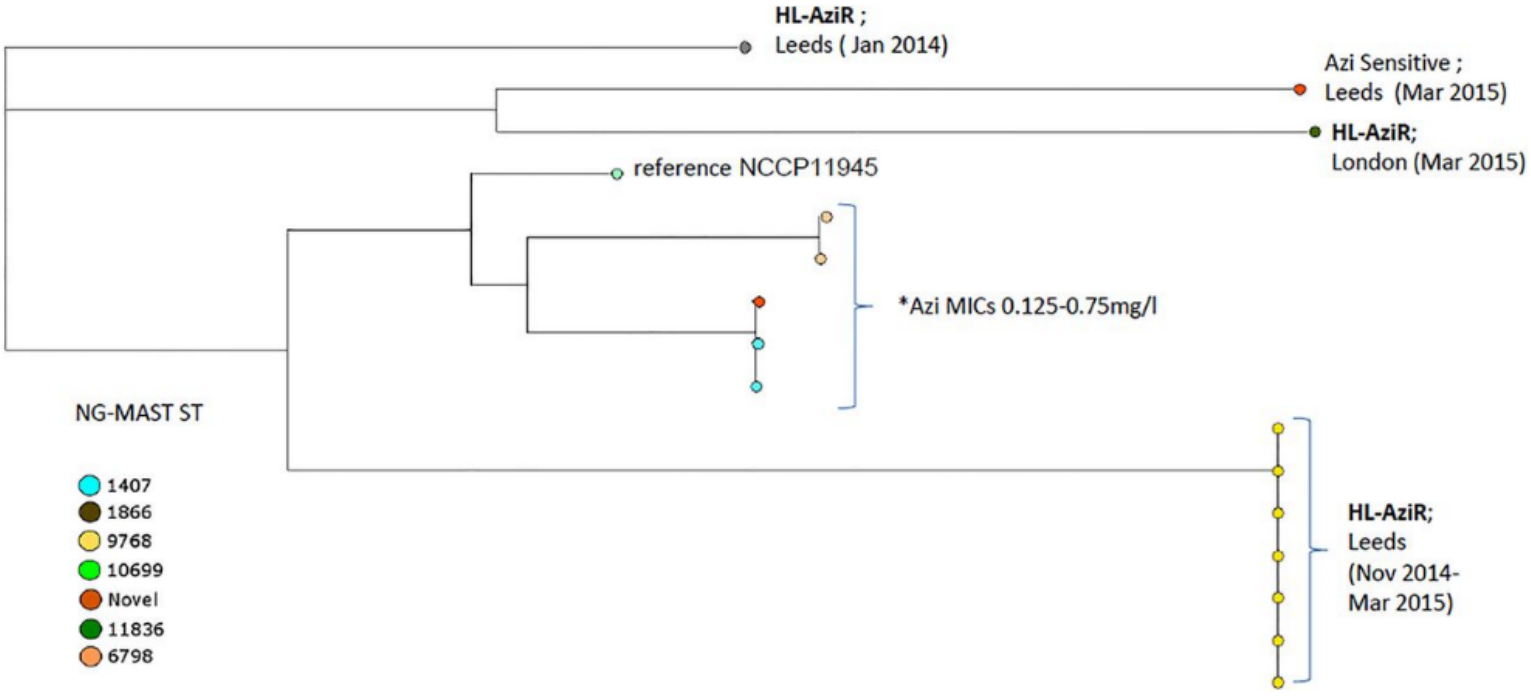


https://www.cdc.gov/std/stats17/figpro.htm#gon

Approach:

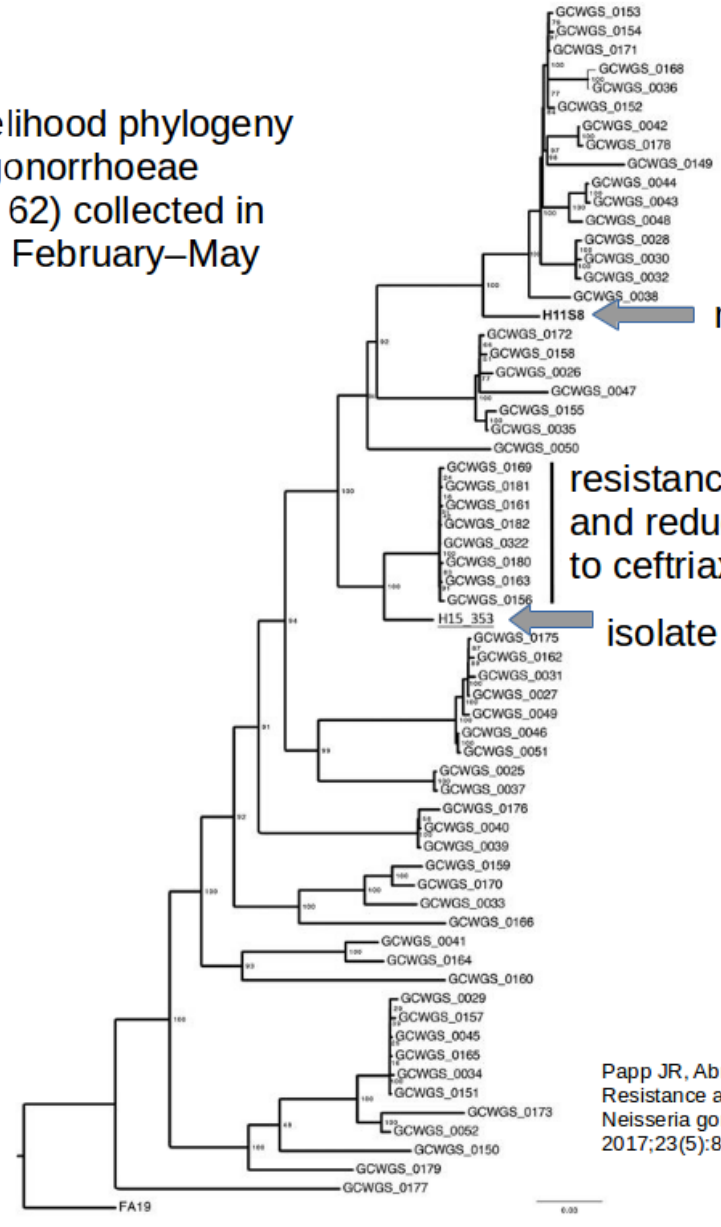
- Whole genomic sequencing of *Neisseria gonorrhea* isolates - up to thousands of lineages
- Phylogenetic inference to track geographic spread and horizontal gene transfer of resistance genes

Combining geographic and evolutionary information can trace transmission, and transfer of resistance alleles across lineages



Stephanie A Chisholm et al. Sex Transm Infect doi:10.1136/sextrans-2015-052312

Maximum-likelihood phylogeny of Neisseria gonorrhoeae samples (N = 62) collected in Hawaii during February–May 2016



resistance to azithromycin

resistance to azithromycin **and reduced** susceptibility **to ceftriaxone**

isolate sampled in the UK

Papp JR, Abrams AJ, Nash E, et al. Azithromycin Resistance and Decreased Ceftriaxone Susceptibility in Neisseria gonorrhoeae, Hawaii, USA. Emerg Infect Dis. 2017;23(5):830-832.

Challenges:

- Thousands of samples; new isolates sequenced every day
- Speed from sampling $\rightarrow$ phylogeny important
- Need to rely on phylogenies for public health action (requires high confidence)
- Often very little nucleotide variability, but horizontal gene transfer is common.

Potential issues:

- Sequencing error
- Effect of choice of reference genome

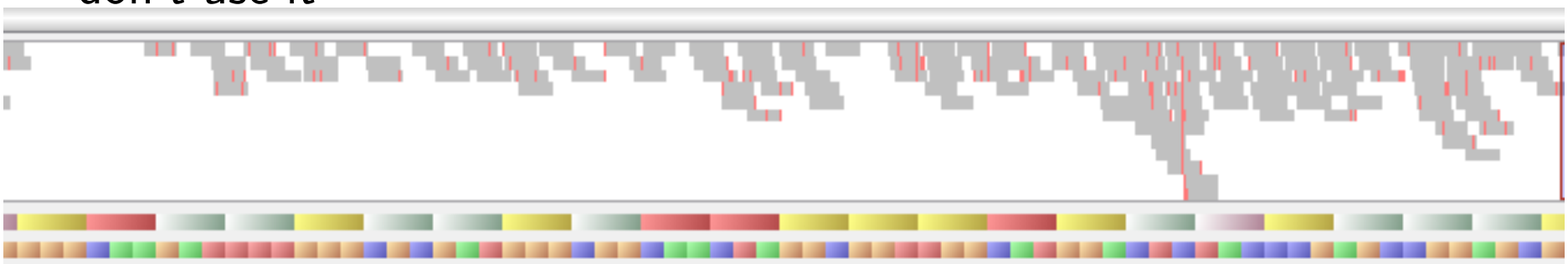
Sequencing error

Potentially problematic when real variable sites are rare

Sequencing errors are likely to be singletons

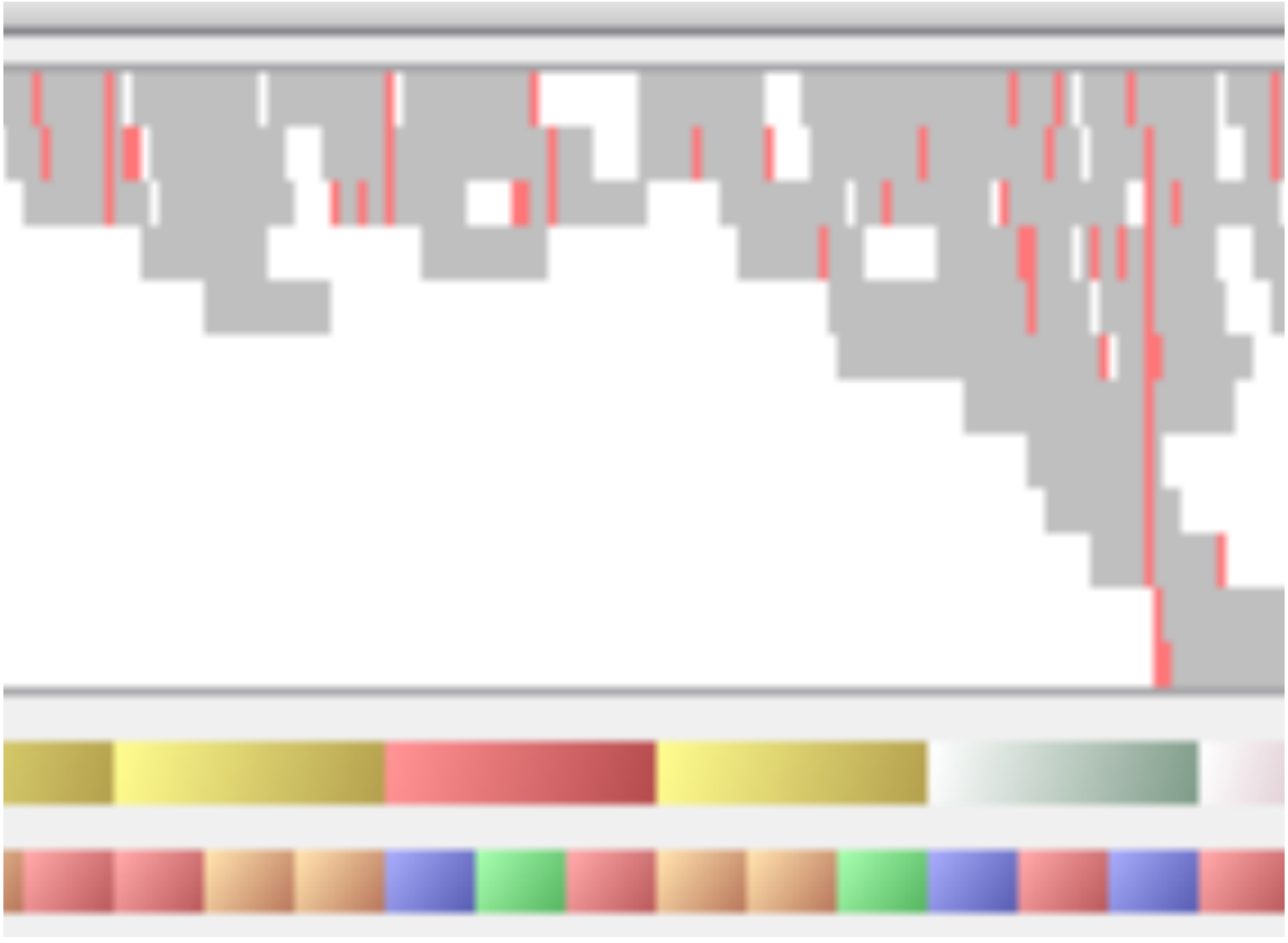
Will overestimate tip branch lengths

Currently, coverage and error information from sequence reads are discarded following  to 
We have information on confidence in individual base calls, but don't use it



Kuhner and McGill (2014) developed a correction for sequencing error in maximum likelihood phylogenetic inference.
Uses a constant expected error per site

Could use a “genotype likelihood”, capturing coverage and read quality (Nielsen et al., 2011)

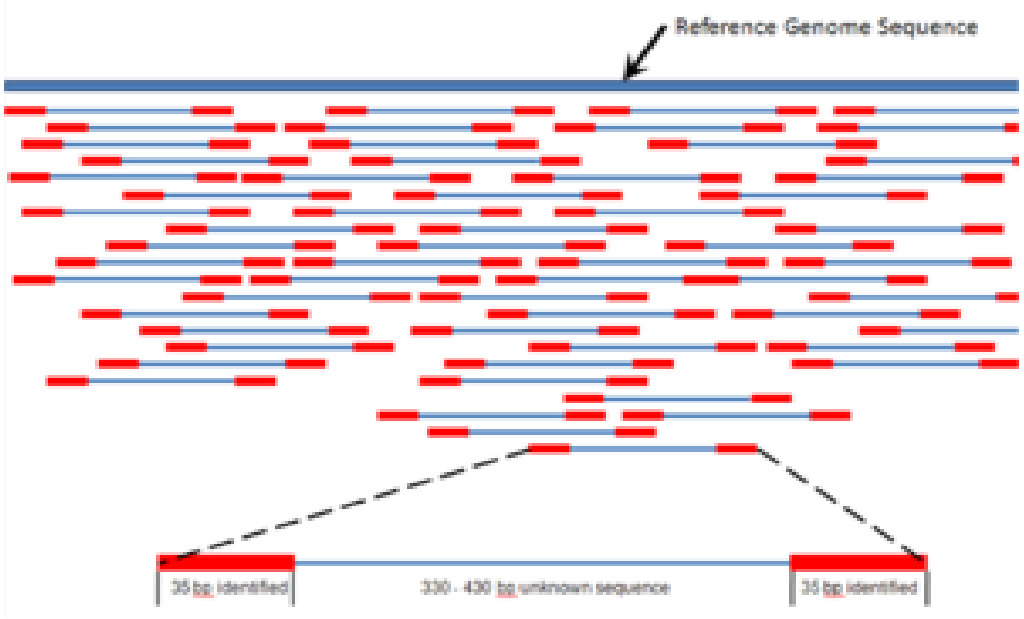


Not currently implemented in phylogenetic likelihood models

At high coverage, effect of sequencing error is likely low!

Effect of reference choice

Reference based mapping of short reads can speed up generating a consensus sequence.



BUT: Reference choice can affect evolutionary inference

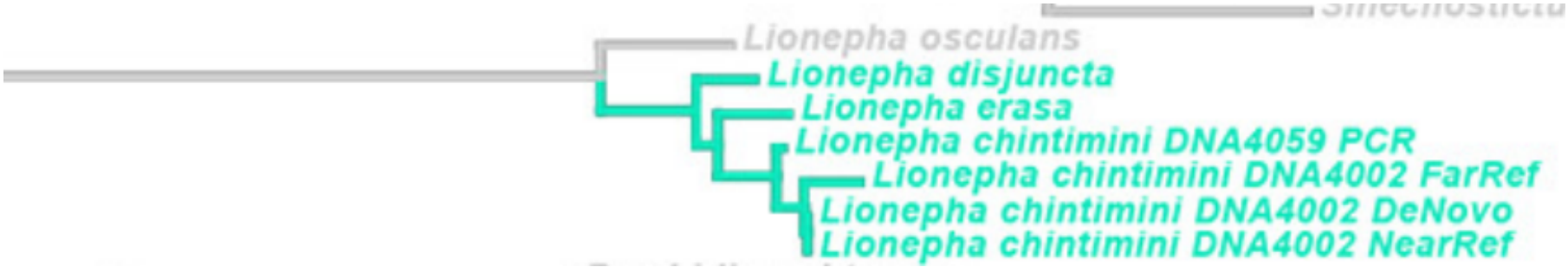
BUT: Reference choice can affect evolutionary inference

In humans, in highly polymorphic regions variant calling is biased toward the the reference base (Brandt et al., 2015)

BUT: Reference choice can affect evolutionary inference

In humans, in highly polymorphic regions variant calling is biased toward the the reference base (Brandt et al., 2015)

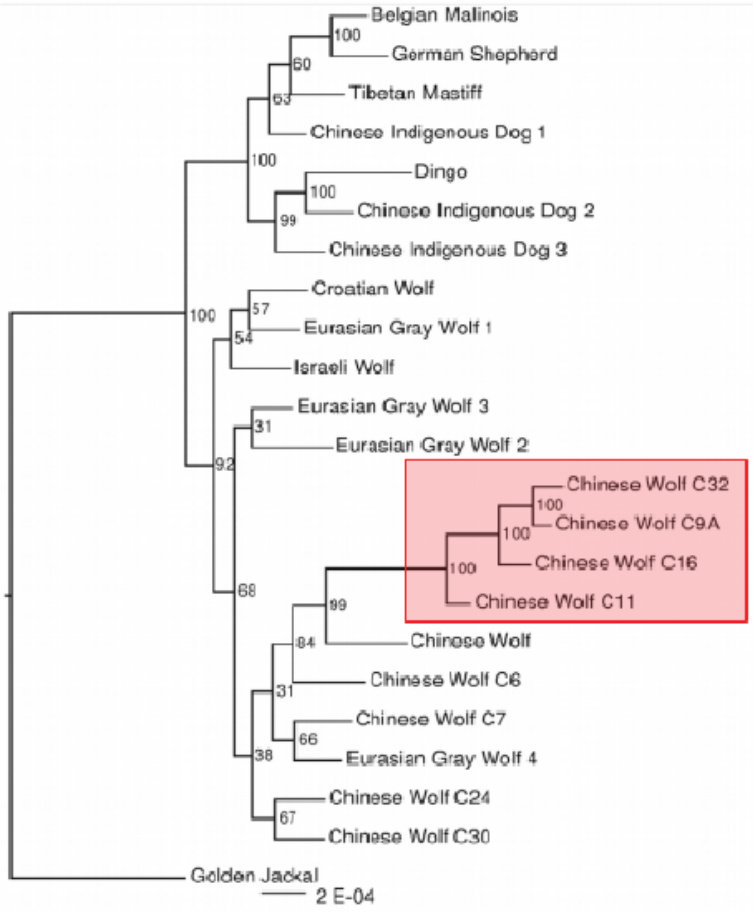
In fragmented DNA samples from beetles, branch lengths change based on reference choices



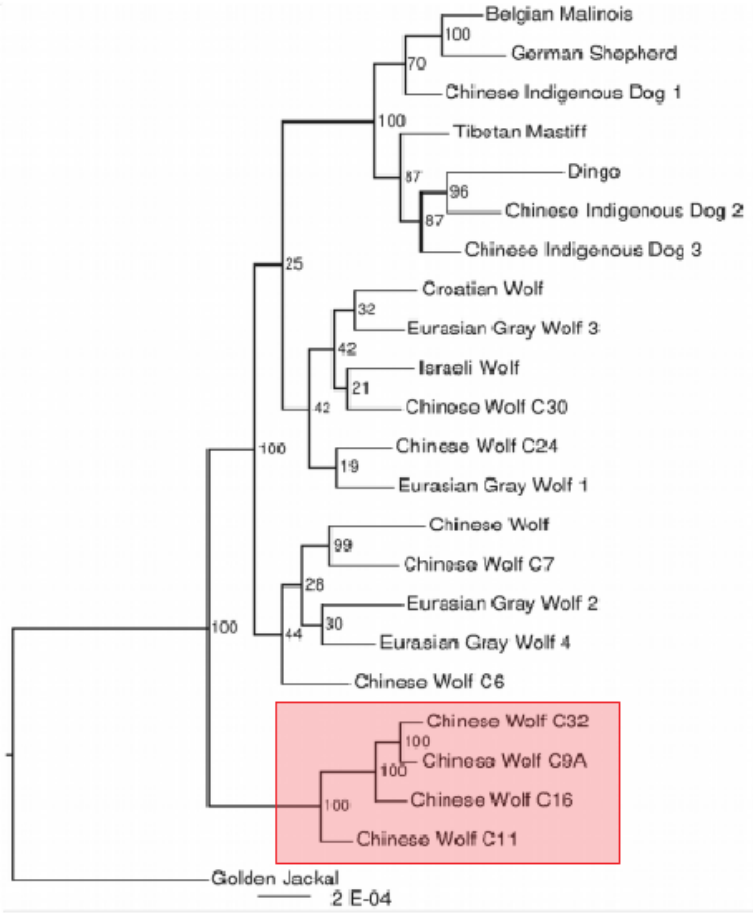
(Kanda et al., 2015)

A reference mapping based approach will discard information about structural variants not found in the reference

Reference choice can affect topology



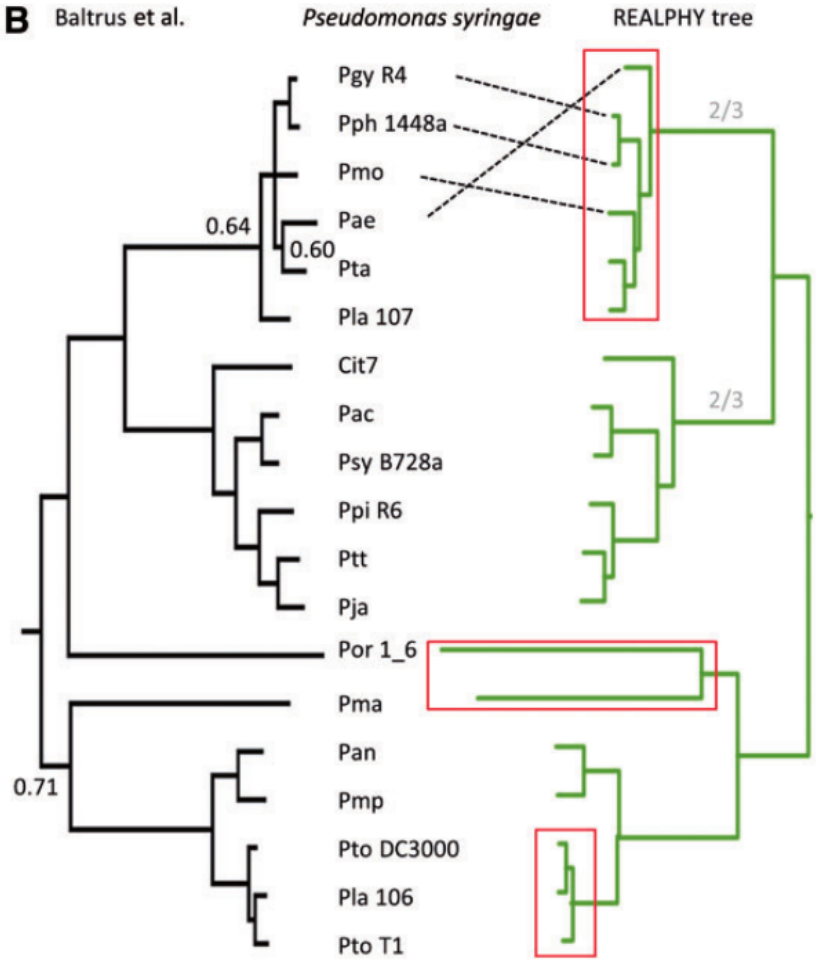
Dog Reference



Wolf Reference

Gopalakrishnan et al. (2017)

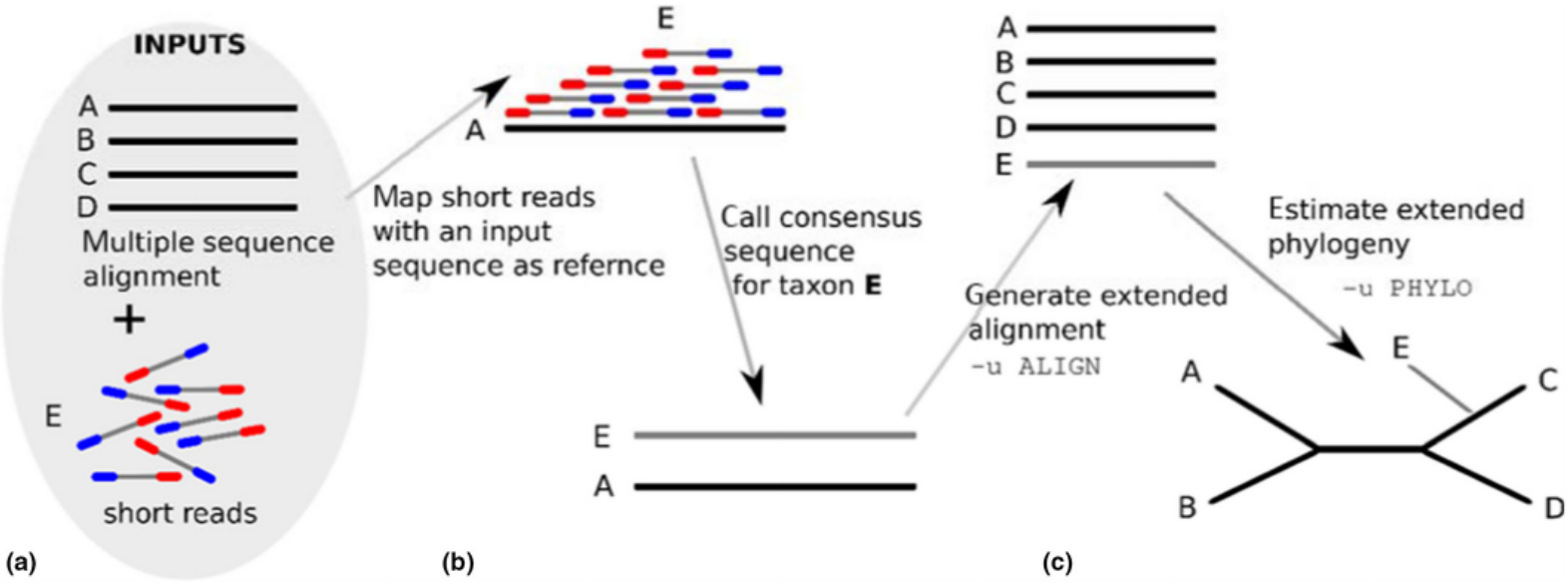
Reference choice can affect topology inference



Mapping sequencing reads to reference genomes requires similarity cutoffs that generate biased missing data (Bertels et al., 2014)

Problem: The true (unknown) phylogenetic history will affect how reads map across the genome.

Phylogenetically informed phylogenomic updating approach:



Assembles only homologous regions of interest

Can use multiple references to generate consensus sequence

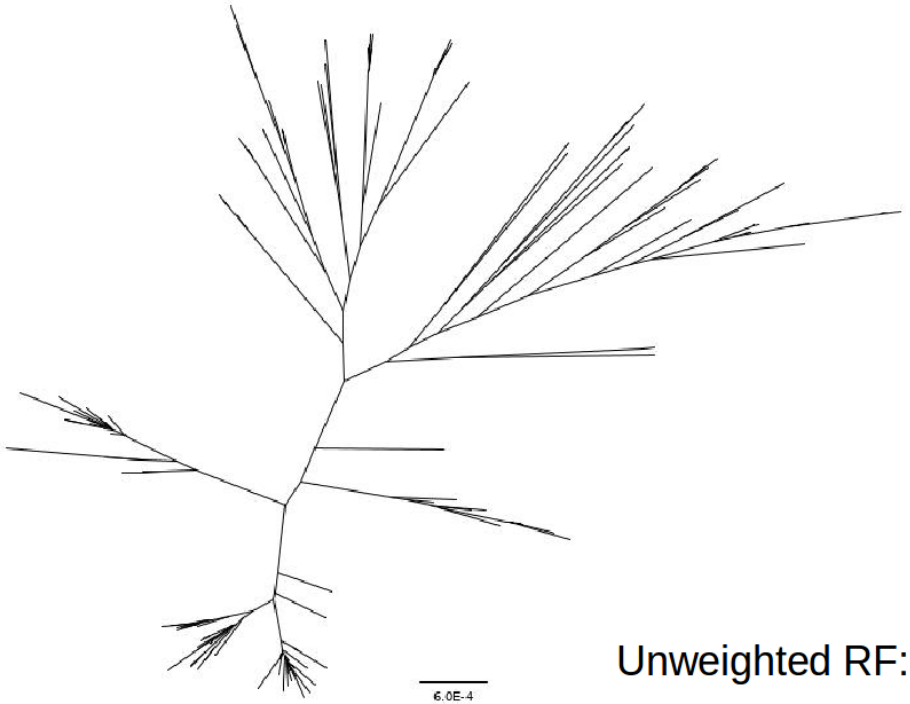
Tree search speed up due to starting tree

`github.com/mctavishlab/extensiphypipeline`

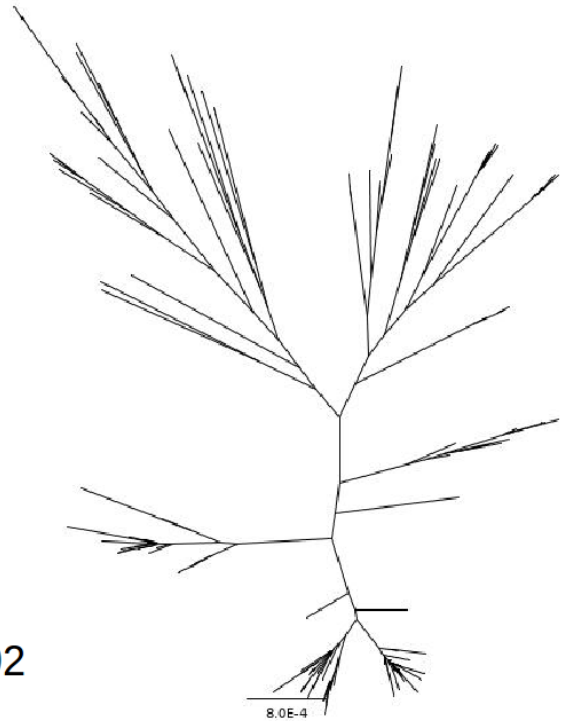
Toscani-Field et al. (2022)

Tree from traditional method

Updated tree

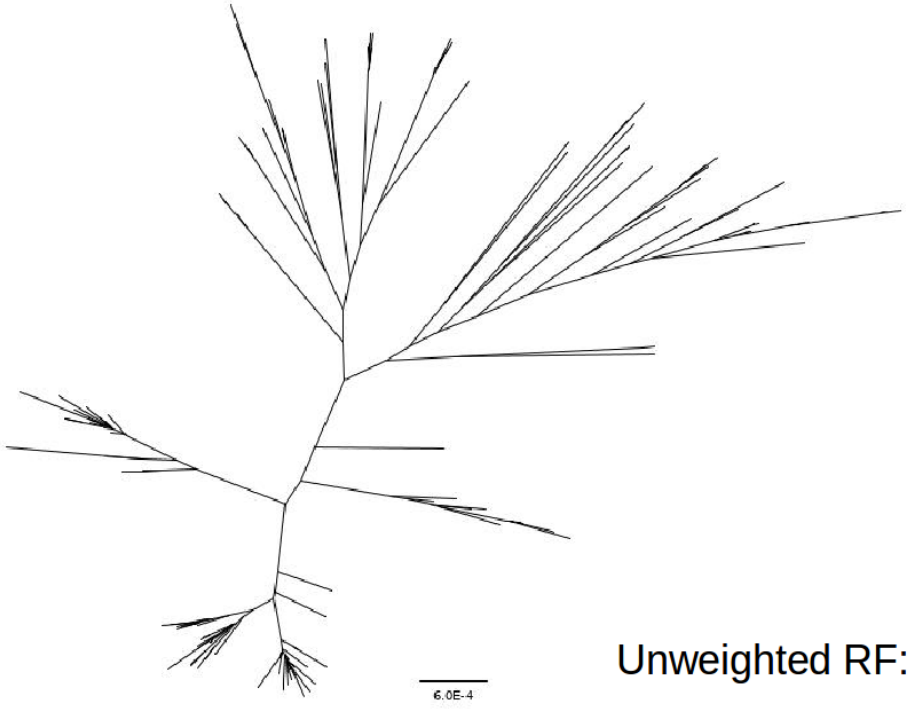


Unweighted RF: 92

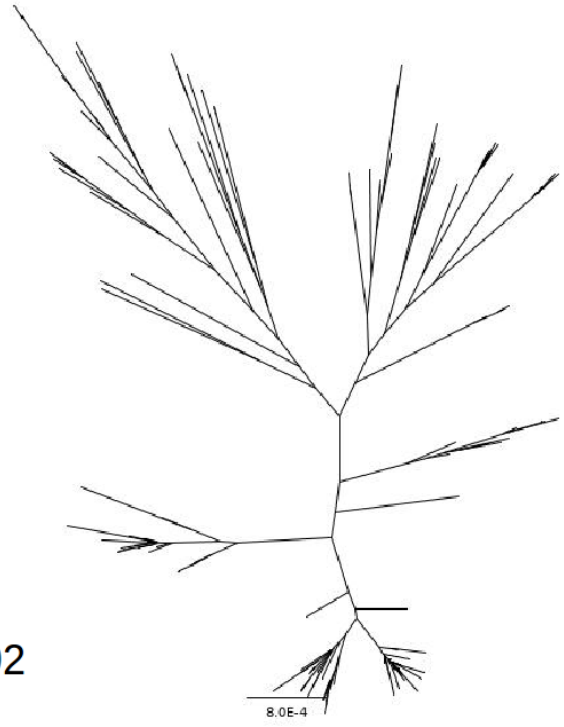


Results:

Ok... 🤔 the tree is different! but is it better or worse?



Unweighted RF: 92

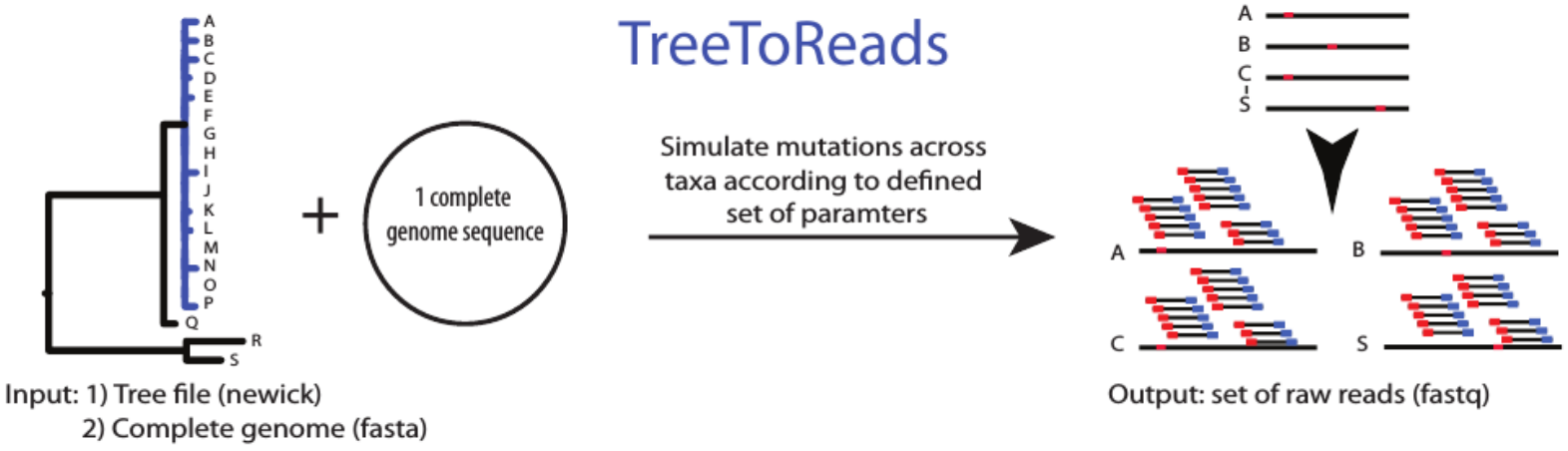


Testing the approach using simulations:
TreeToReads

Takes into account:

- Phylogeny and model of evolution
- Insertions and deletions
- Distribution of mutations across the genome
- Read coverage
- Sequencing error profiles (observed or estimated)

Generates short read data with which to test assembly, alignment and inference pipelines.

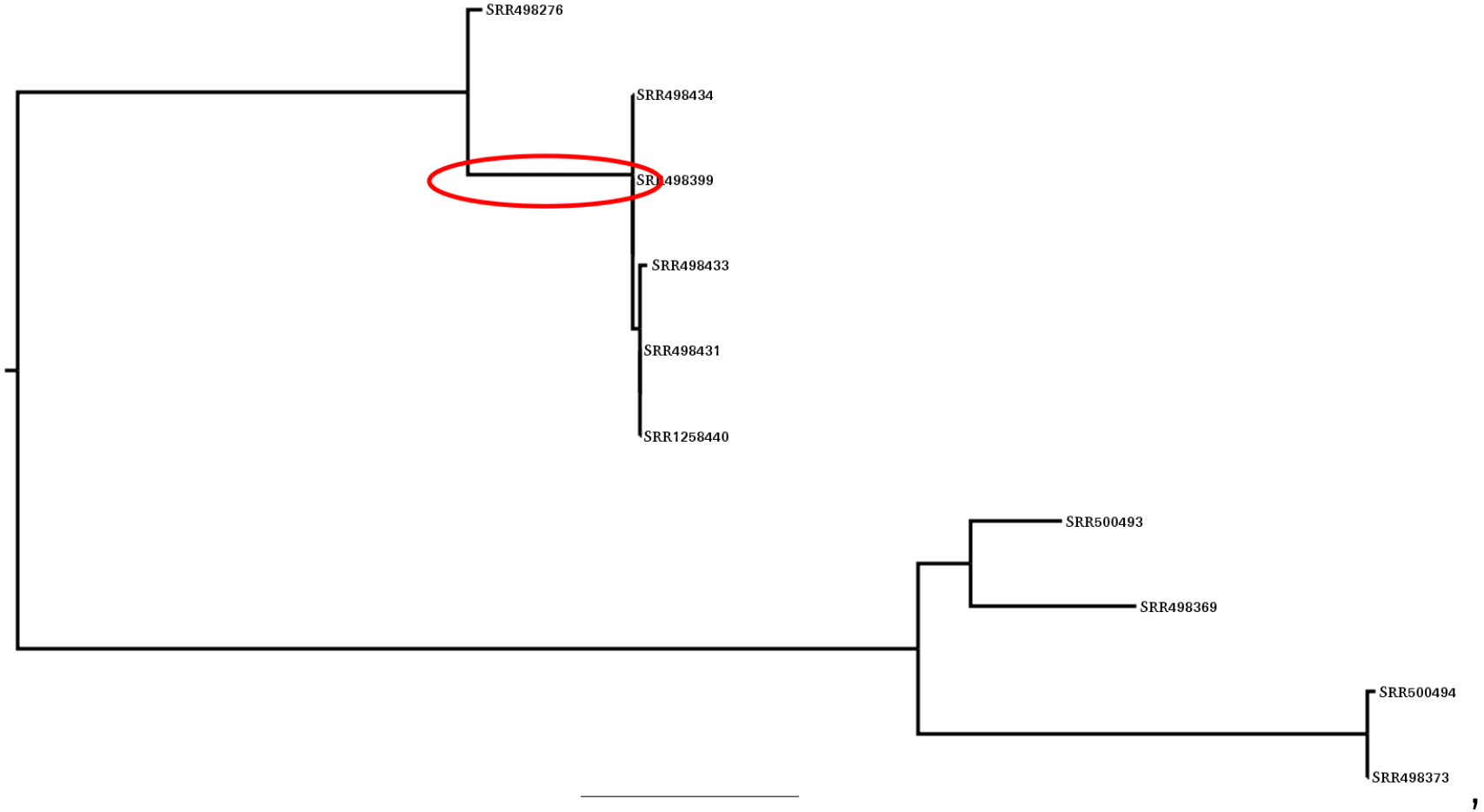


Input genome for simulation is a tip on simulated tree
Can test alignment to other empirically observed genomes (McTavish et al., 2017)

`github.com/snacktavish/treetoreads`

Other new approaches for generating reads from phylogenies: *NGSphy* (Escalona et al., 2018), *Jackalope* (R package) (Nell, 2019)

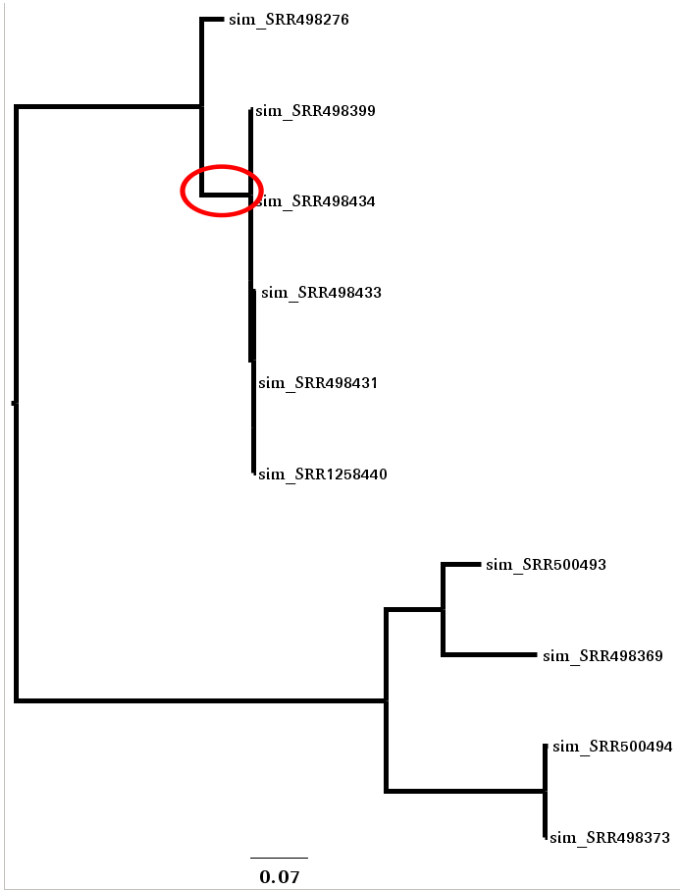
Take observed outbreak tree



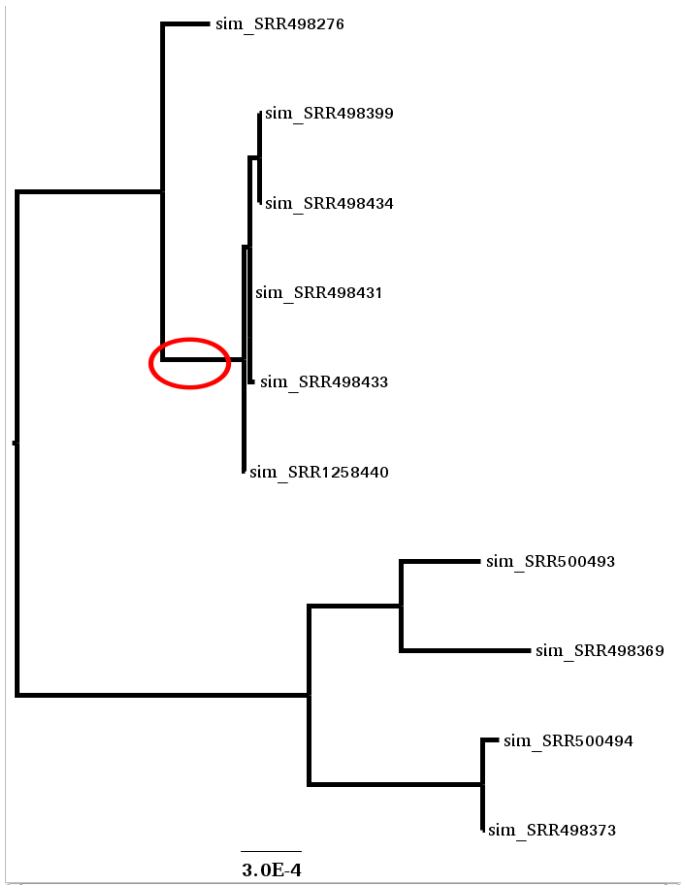
Simulate reads using empirical parameters

Infer trees from reads using two different reference genomes.

Reference within outbreak reference



Distant (1% sequence divergence)



Simulation summary

- In this example, even distant reference genome did not affect parameter of interest (monophyly of outbreak), although it did affect branch lengths
- Effects of read mapping parameters and reference genome choice are likely to be idiosyncratic
- By using empirical estimates for evolutionary model, can investigate effects on parameters of interest
- Currently applying this approach to test gonorrhea phylogenetic updating procedure

Summary

Bias: Sequencing error, reference choice

Effect on inference:

Sequencing error can increase terminal branch lengths relative to internal branches

Not mapping reads on lineages more distant from reference genome will decrease those branch lengths

Mitigation: Use multiple reference genomes, simulation based tests to assess accuracy

Conclusions:

When a closely related reference is available, alternatives worsen inference

At high (around 40x) coverage all mutations are confidently recovered

Even at lower coverage (around 5x) high confidence in monophyly of outbreak clade

Gonorrhea reference bias exercise
https://github.com/snacktavish/sequence_data_exercise/blob/main/DataSetAssembly.md

Big picture

- All data sets are biased, genome scale data is no exception
- Careful project planning helps
- Interrogate potential biases in data sets

What to do?

- What data will answer **your** questions?
- Are there existing data you want to be able integrate with?
- Consider in which direction biases are likely to sway results
- Use the most an appropriate available model for your data
- Re-sample your data to test if your key conclusions are robust to choices
- Simulation approaches to test if parameters of interest are affected by sampling and ascertainment schemes

"The phylogenomic approach is, despite its flaws, surprisingly robust, as most pipelines will lead to the recovery of a similar species tree topology.

This can be explained by the sheer quantity of phylogenetic signal accumulated when thousands of molecular markers are combined."
Simion et al. (2020)

Questions?

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., and Nimwegen, E. v. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution*, 31(5):1077–1088. Number: 5.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3: Genes|Genomes|Genetics*, 5(5):931–941. Number: 5.

Escalona, M., Rocha, S., and Posada, D. (2018). NGSphy: phylogenomic simulation of next-generation sequencing data. *Bioinformatics*.

Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M.-H. S., Kuderna, L. F. K., Räikkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T., Hansen, A. J., Dalén, L., and Gilbert, M. T. P. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC Genomics*, 18(1):495.

Huang, H. and Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology*, 65(3):357–365. Number: 3.

Kanda, K., Pflug, J. M., Sproul, J. S., Dasenko, M. A., and Maddison, D. R. (2015). Successful Recovery of Nuclear Protein-Coding Genes from Small Insects in Museums Using Illumina Sequencing. *PLOS ONE*, 10(12):e0143929. Number: 12.

Kuhner, M. K. and McGill, J. (2014). Correcting for Sequencing Error in Maximum Likelihood Phylogeny Inference. *G3: Genes|Genomes|Genetics*, 4(12):2545–2552. Number: 12.

Leaché, A. D., Banbury, B. L., Felsenstein, J., Oca, A. N.-M. d., and Stamatakis, A. (2015). Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, page syv053.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., and Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology*, 58(1):130–145. Number: 1.

McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M., and Timme, R. E. (2017). TreeToReads - a pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*, 18:178.

Nell, L. A. (2019). jackalope: a swift, versatile phylogenomic and high-throughput sequencing simulator. *bioRxiv*.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6):443–451. Number: 6.

Roure, B., Baurain, D., and Philippe, H. (2013). Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 30(1):197–214. Number: 1.

Simion, P., Delsuc, F., and Philippe, H. (2020). To What Extent Current Limits of Phylogenomics Can Be Overcome? page 2.1:1. Publisher: No commercial publisher | Authors open access book.

Toscani-Field, J., Abrams, A. J., Cartee, J. C., and McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*, 13(3):682–693. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13790.

Wessinger, C. A., Freeman, C. C., Mort, M. E., Rausher, M. D., and Hileman, L. C. (2016). Multiplexed shotgun genotyping resolves species relationships within the North American genus Penstemon. *American Journal of Botany*, 103(5):912–922. Number: 5.