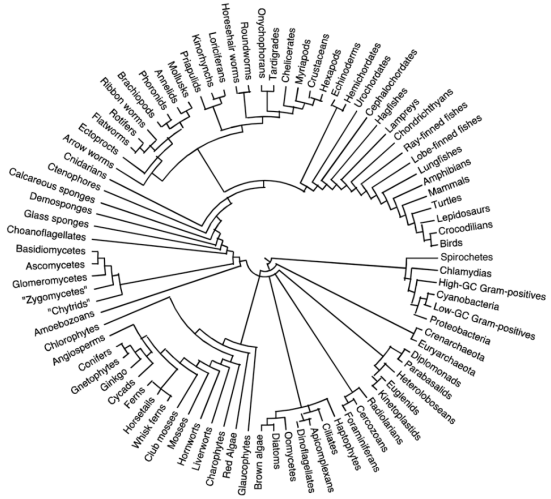
Genomic data for evolutionary inference

Emily Jane McTavish

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

How do you get from

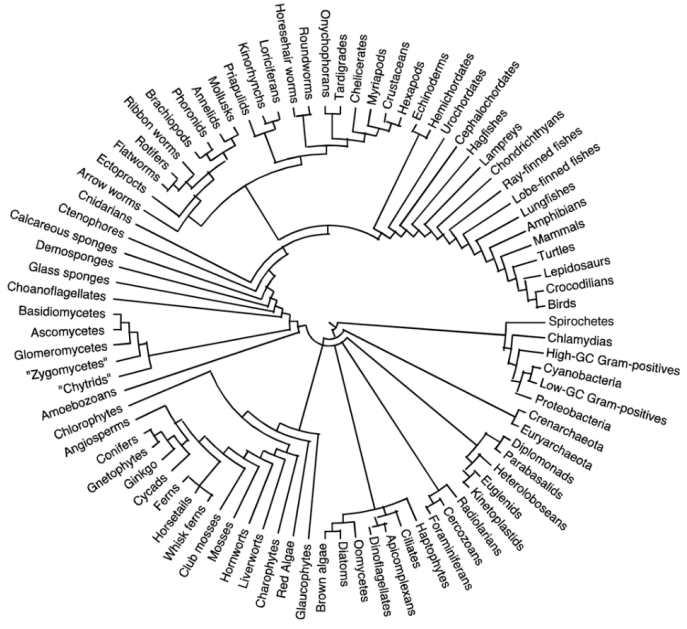
to



?

You've seen a lot about how to get from

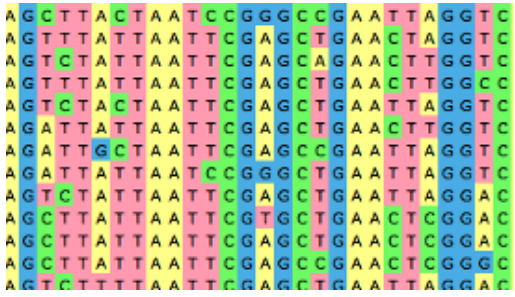
to



I'm going to talk about going from

to

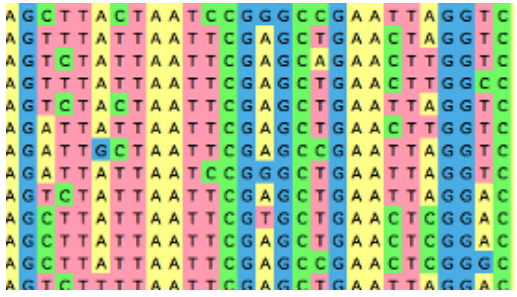
to



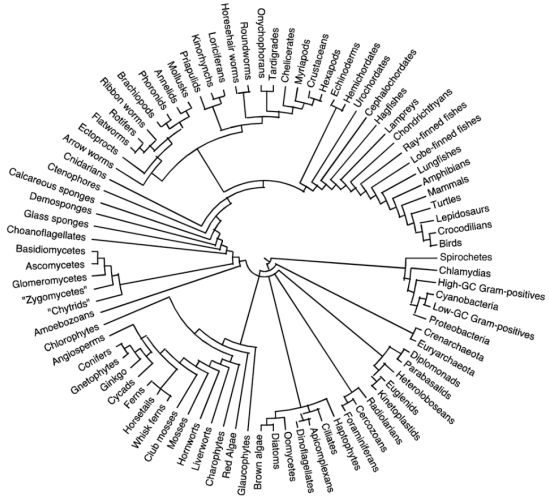
I'm going to talk about going from

to

to



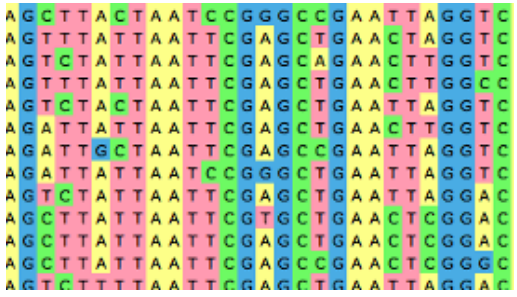
and how those choices can affect



I'm going to talk about going from

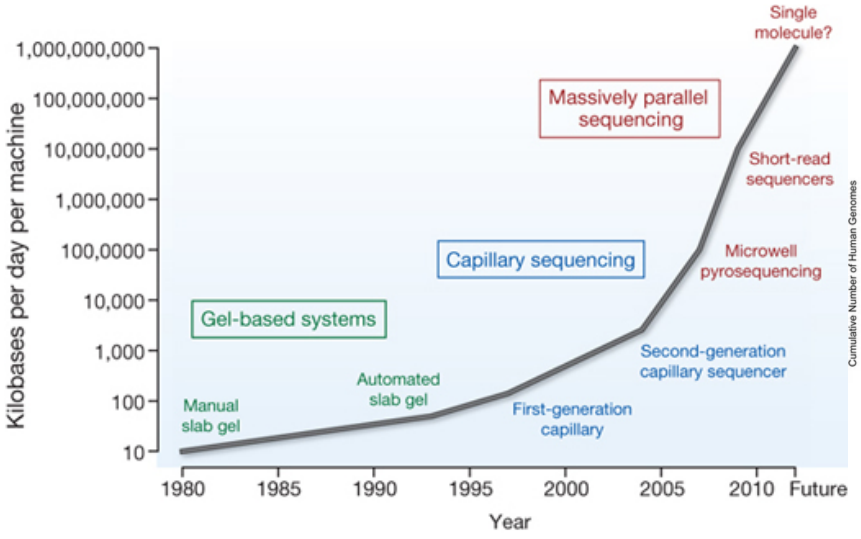
to

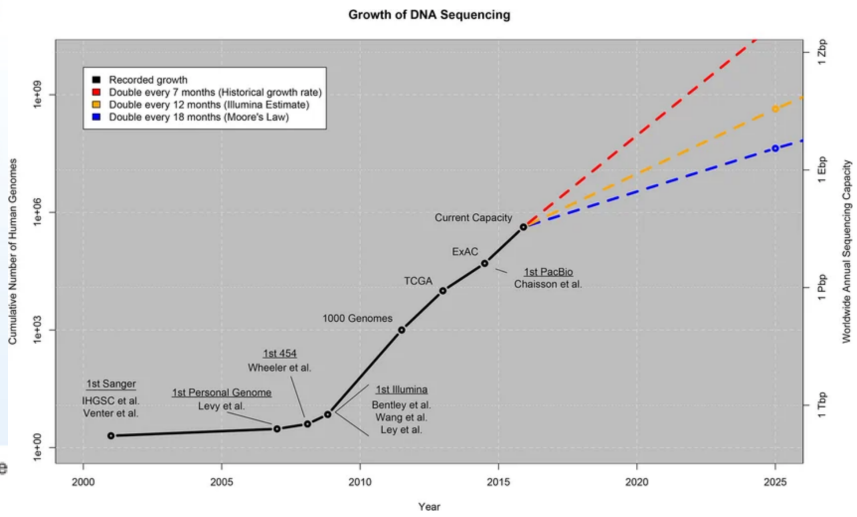
to



and how those choices can affect

The quantity of available sequence data for inferring evolutionary relationships has increased rapidly in recent decades





http://genome.wellcome.ac.uk/

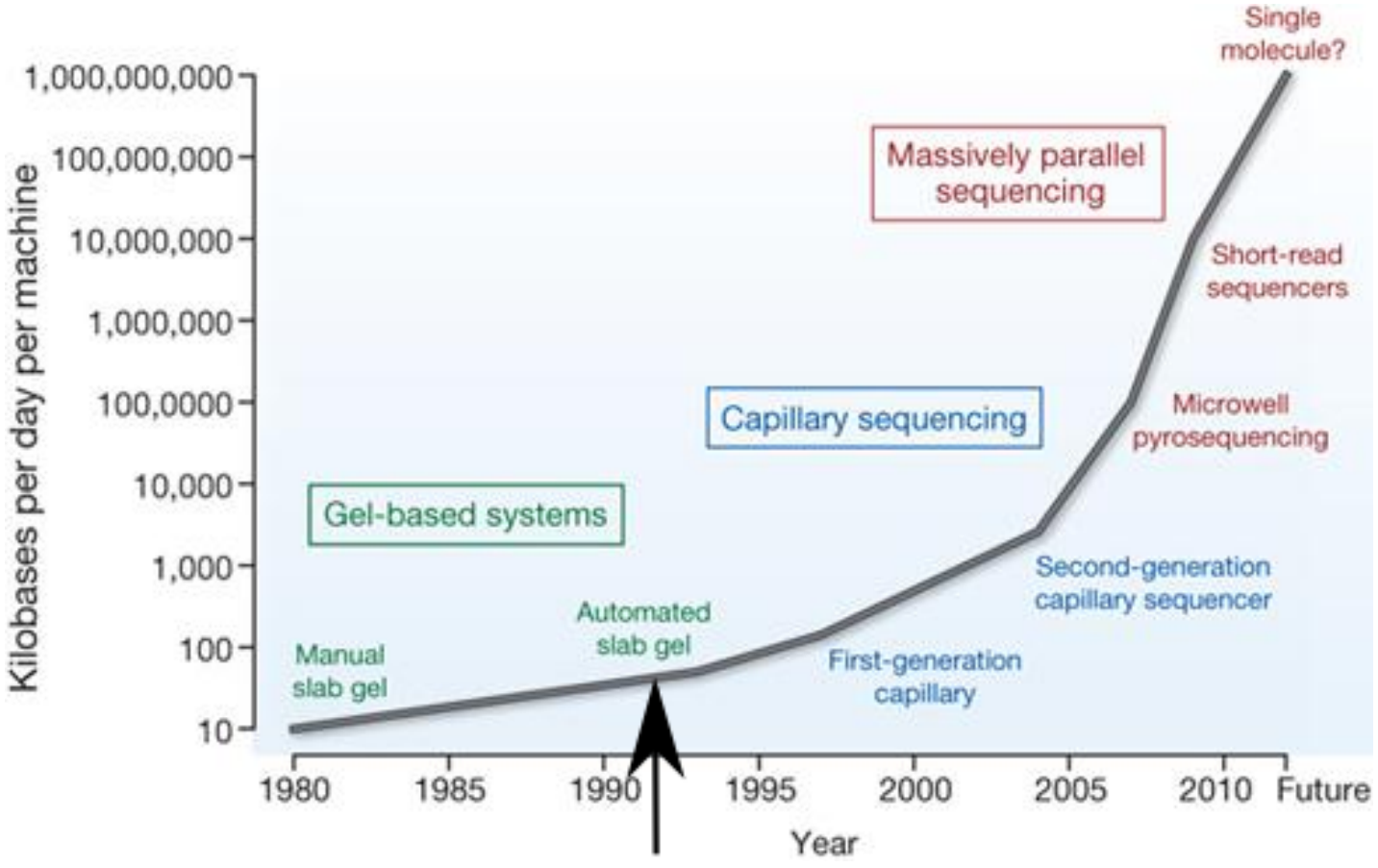
"With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data"

– Jeff Thorne (Evolutionary biologist)

"With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data"

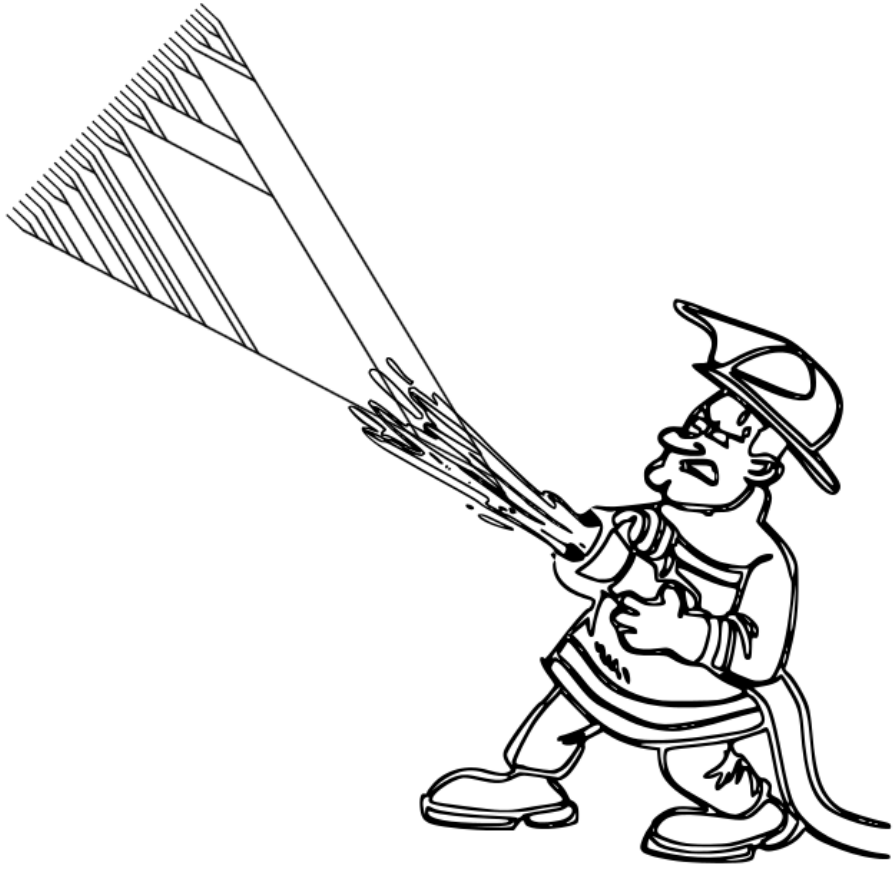
– Jeff Thorne (Evolutionary biologist)

Thorne et al., Journal of Molecular Evolution. **1991**



http://genome.wellcome.ac.uk/

How can we ensure that we are adequately analyzing these huge data sets?



There are a lot of choices to make!

Biological questions

What do you want to know?

What do you already know?

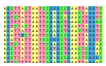
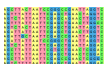
Biological questions

- What do you want to know?
- What do you already know?

Technical questions

- What data is right for our questions?
- Is a closely related reference genome available?
- How should we process and analyze our data?
- What biases may be affecting our inferences?

General approach

- Decide what to sequence ( to )
- Consensus sequence, alignment, locus selection ( to )
- Evolutionary analyses ( to )
- Success!

What to sequence?

to

Survey question!
What kind of data are you using in your research?
PollEv.com/emilyjanemctavish820

Different sequencing approaches enrich the samples for different components of the genome

Enrichment (smallest to largest proportion of genome)

- **Directed PCR** can be efficient, but doesn't scale well or sound fancy (not *omics!).
- **Targeted enrichment, e.g. Rad-tag** need data about the genome to get data
- **Transcriptome** Expression levels, enriched for protein coding genes, will vary based on cell type, environment.
- **Whole genome** Capture all the data, but much of it may not be helpful for phylogenetics

Depending on your questions, any of these could be the best option!

Genomic sequencing

- You have all the data! 👍
- You have to deal with all of the data. 👎

Do you need a whole genome to answer your questions?

Do you need a whole genome to answer your questions?

For phylogenetic and population genetic questions, not necessarily!

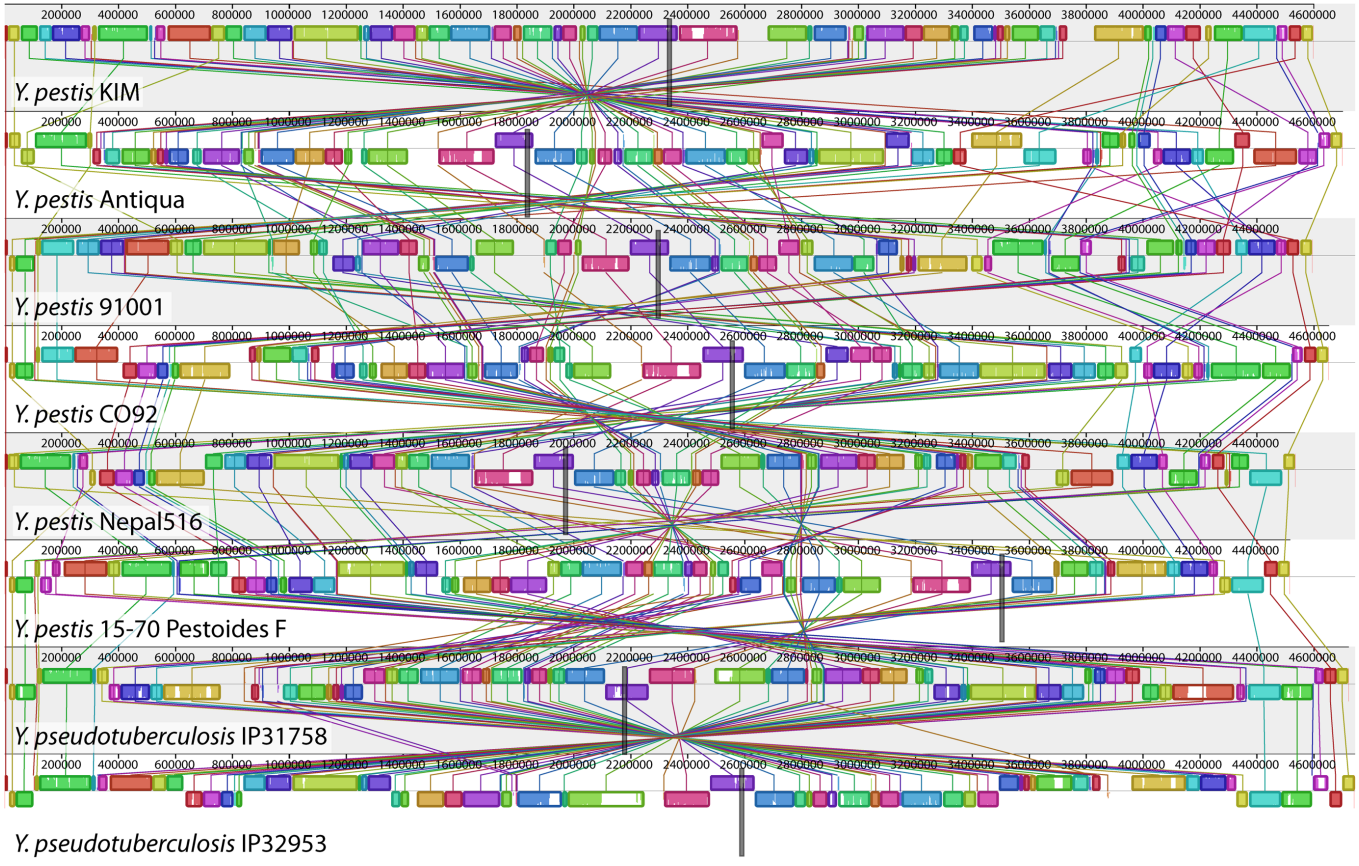
Most phylogenetic methods cannot directly handle whole genome data, but from whole genome sequencing you can get homologous loci, as well as a bunch of other stuff!

To make evolutionary statements, you need to align genomic regions across taxa.

Depending on evolutionary history this can be easy or hard!

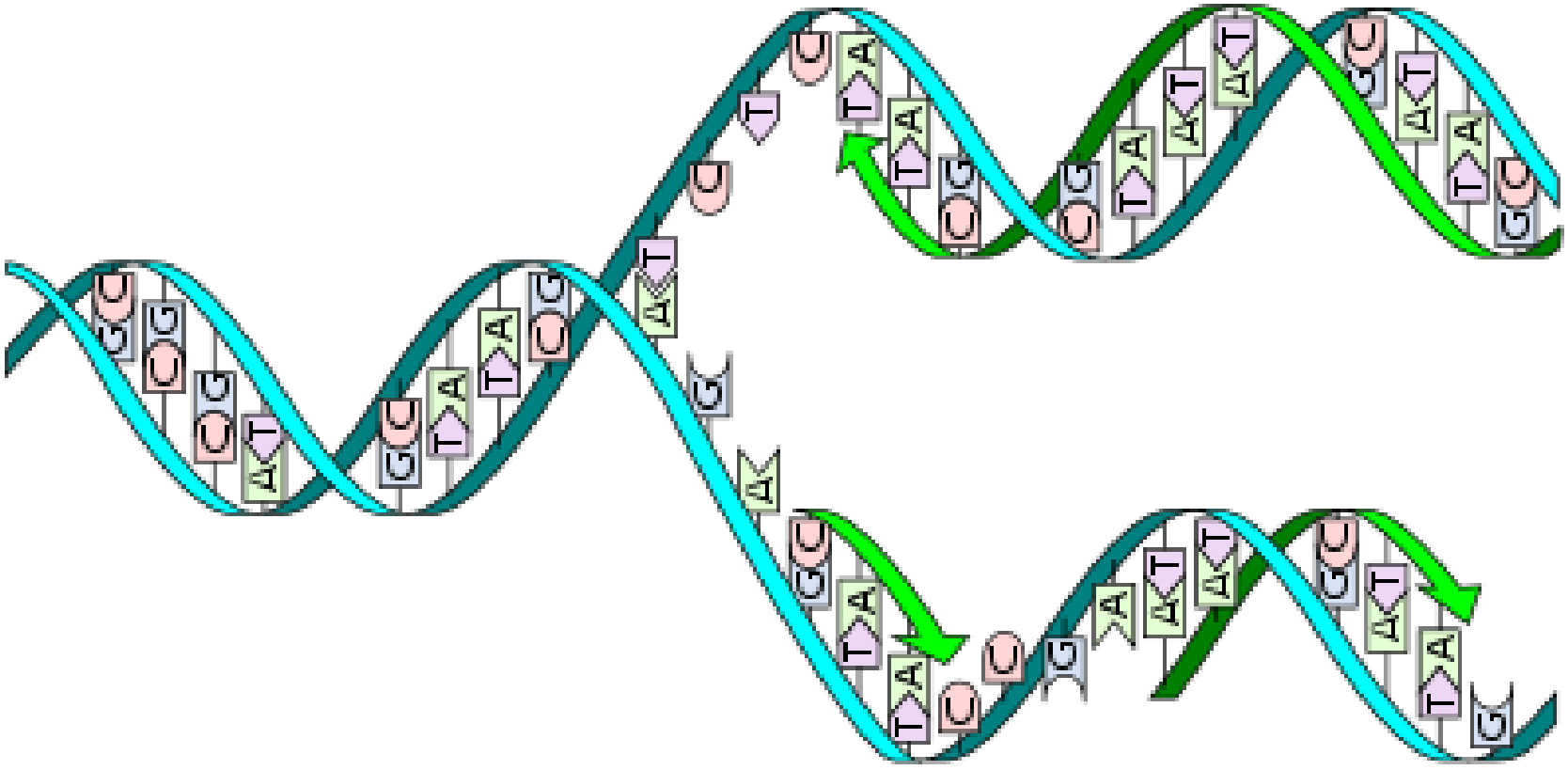
To make evolutionary statements, you need to align genomic regions across taxa.

Depending on evolutionary history this can be easy or hard!



(Darling et al., 2008)

An alignment is a statement of shared ancestry

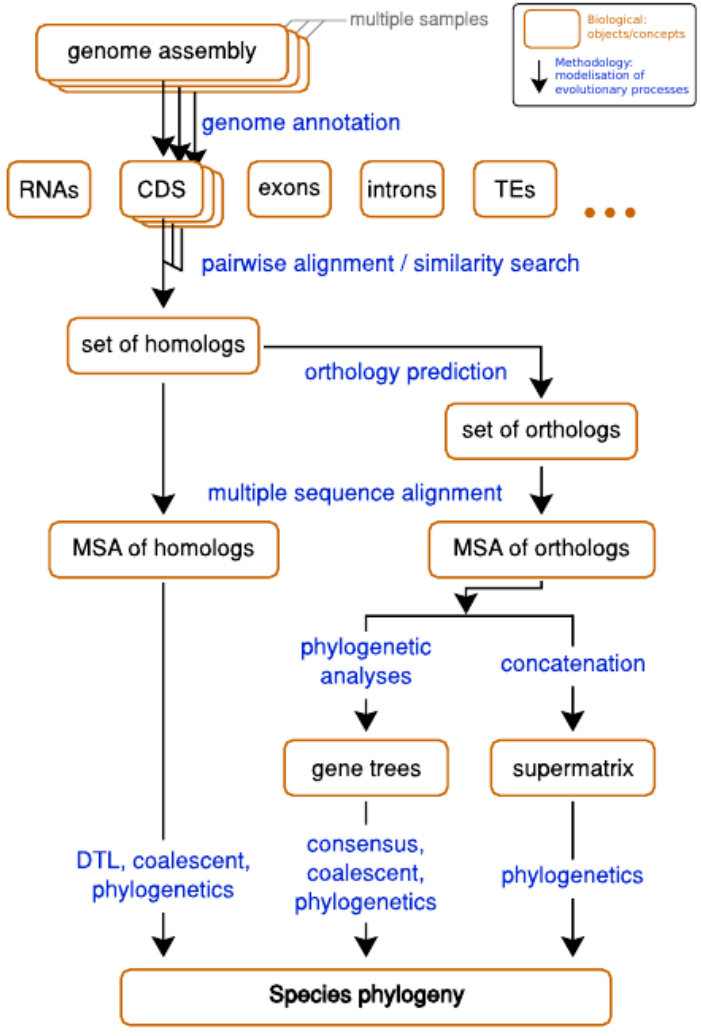


Gene tree (Locus tree)
The ancestry of a homologous region of the genome that has a single evolutionary history (no recombination)

Enrichment methods focus our sequencing efforts on these regions

Free textbook: Phylogenetics in the Genomic Era

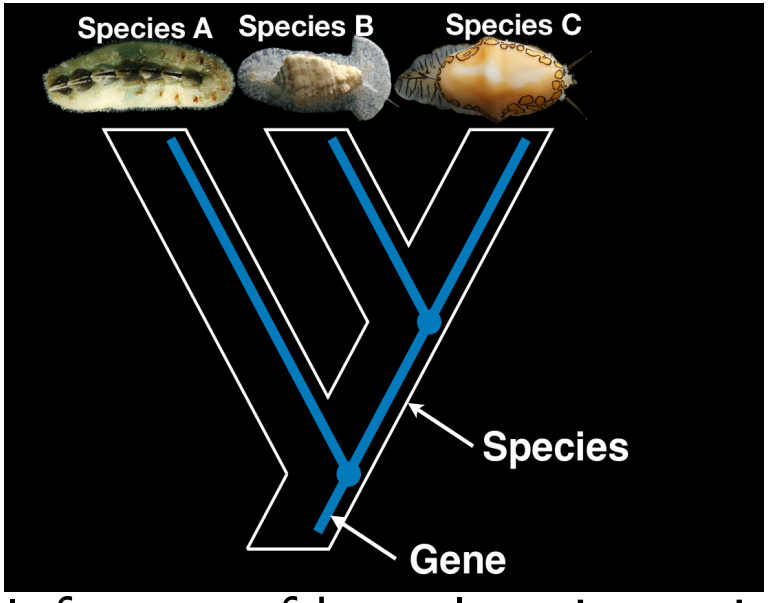
`https://inria.hal.science/PGE/page/table-of-contents`



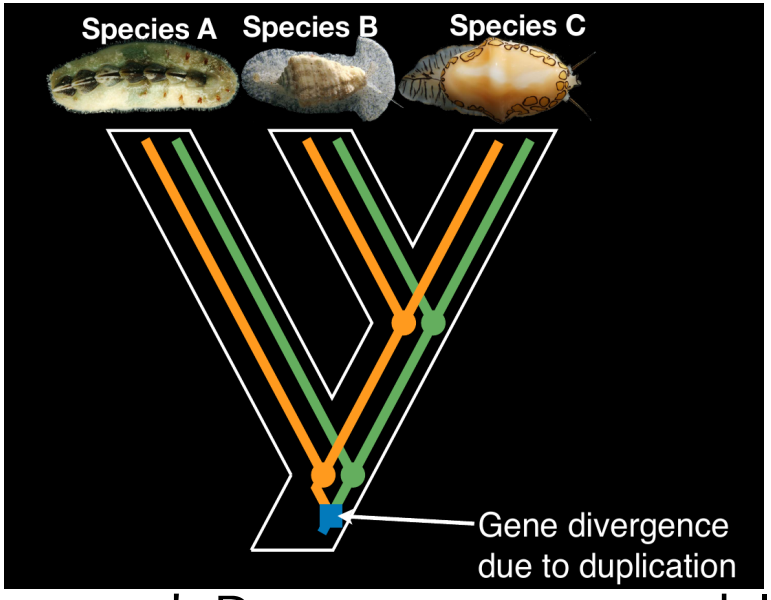
Simion et al. (2020)

Gene duplication and loss

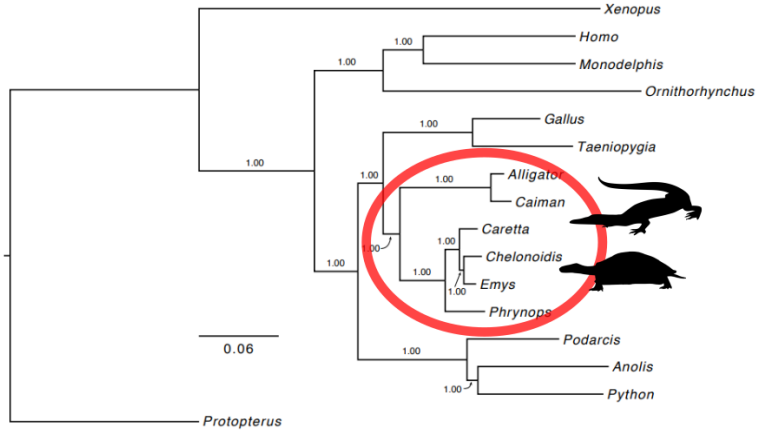
Orthology



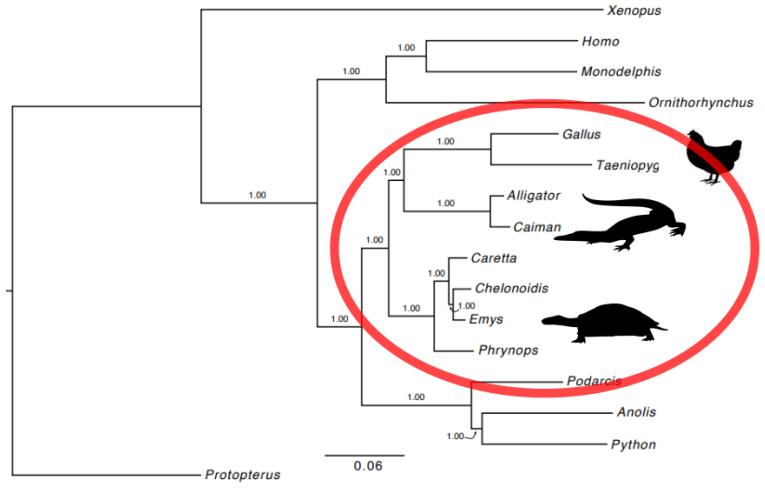
Paralogy



Inference of homology is not incorrect! But our current models are limited. If you treat paralogs as orthologs, you can make incorrect inferences. figure from Casey Dunn



A majority-rule consensus tree from Bayesian phylogenetic analysis of the concatenated dataset of Chiari et al.
248 nuclear genes



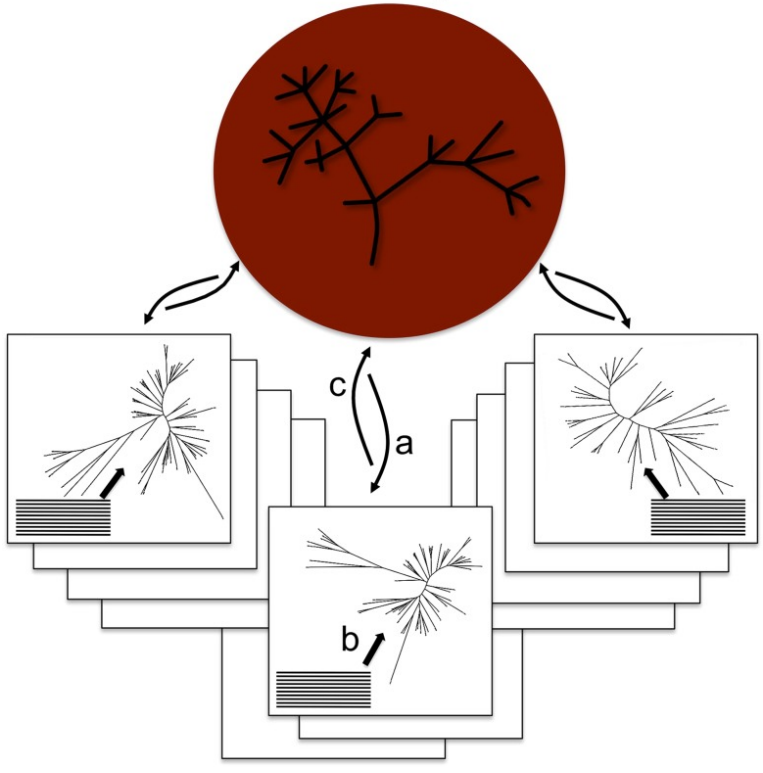
Same analysis of the same dataset
after removal of the two genes with evidence for paralogy.

"investigation of genes with extreme support for turtle placement revealed unappreciated paralogy in a small proportion of alignments (<1%) that had an extraordinary influence on the inferred placement of turtles."
(Brown and Thomson, 2016) (Chiari et al., 2012)

Challenge: The true (unknown) phylogenetic history is needed to assess orthology vs paralogy

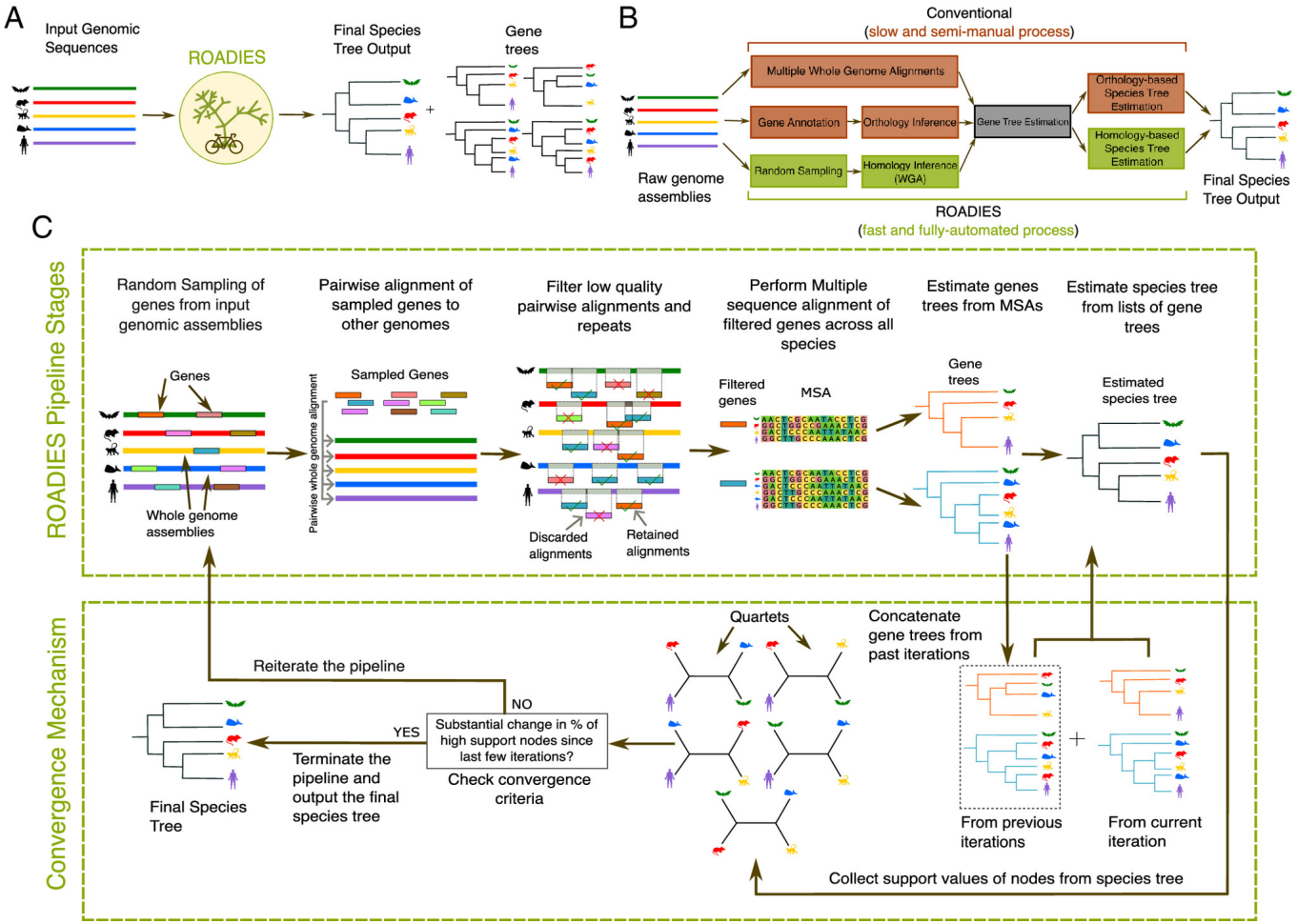
Integrated approaches to Duplication, Transfer, and Loss (DTL) can jointly estimate gene trees and species trees, but are very computationally expensive.

$$L(T, S, N|A) = \prod_{G_i \in \mathcal{G}} L(G_i)$$



Phyldog; (Boussau et al., 2013)

There are several new approaches to automate gathering homologous loci from whole genomes for phylogeentics, e.g. OrthoGarden(Turner et al., 2026), Roadies (Gupta et al., 2025), Read2Tree (Dylus et al., 2024)



(example figure from Roadies paper)

Using all Gene Families Vastly Expands Data Available for Phylogenomic Inference

Megan L. Smith [*,1,2] Dan Vanderpool [1,2] and Matthew W. Hahn [1,2]

"For most subsets of the data and inference methods, using all clusters (i.e. paralogs and orthologs) also results in consistent inferences of species tree topologies. Our results highlight the benefits of using data from all gene families by showing that the amount of data used can be increased by an order of magnitude" **(but there is sensitivity to inference method)**

CellPress Trends in Genetics

Review

New Approaches for Inferring Phylogenies in the Presence of Paralogs

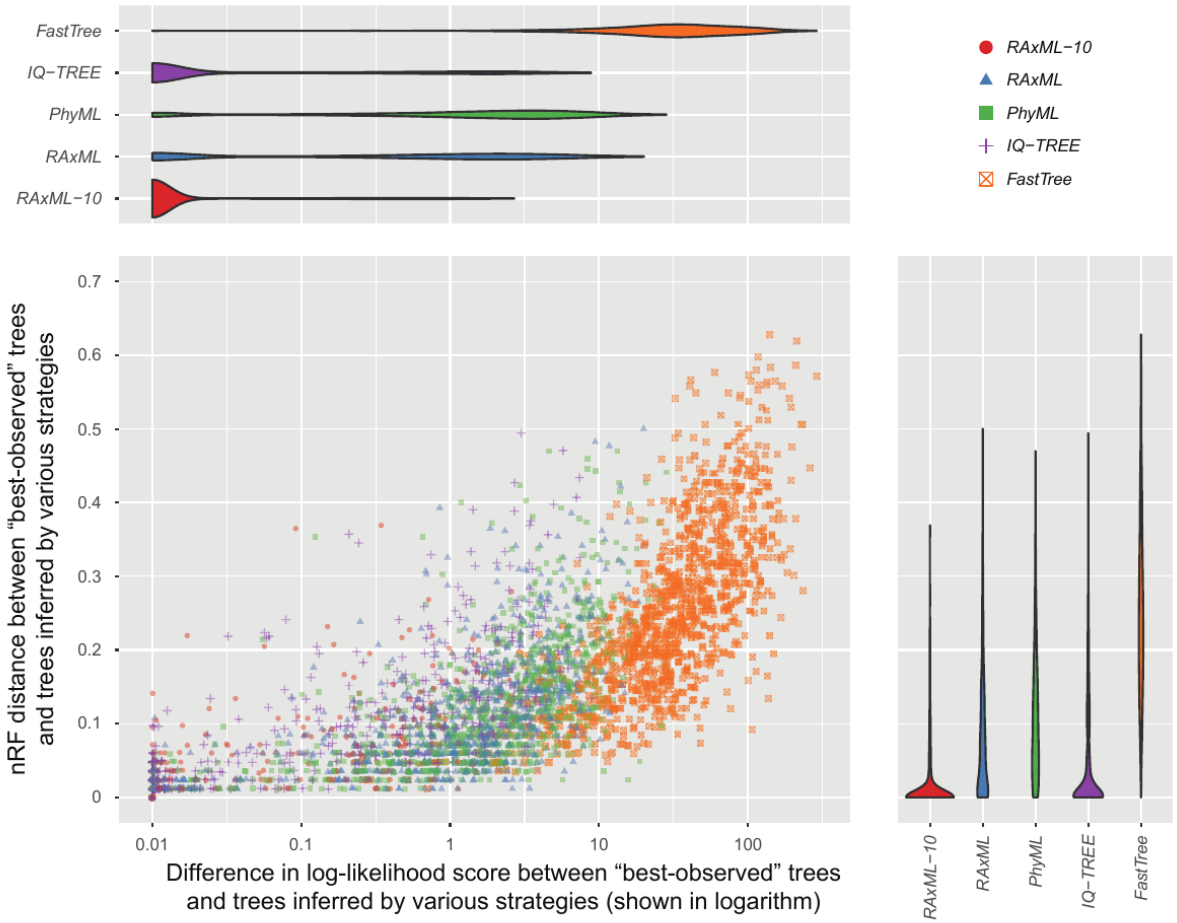
Megan L. Smith [1,*] and Matthew W. Hahn [1]

Smith et al. (2022); Smith and Hahn (2021)

Analyzing genome scale data sets can be SLOW.

Are the tradeoffs of faster methods worth it?

Figure 3

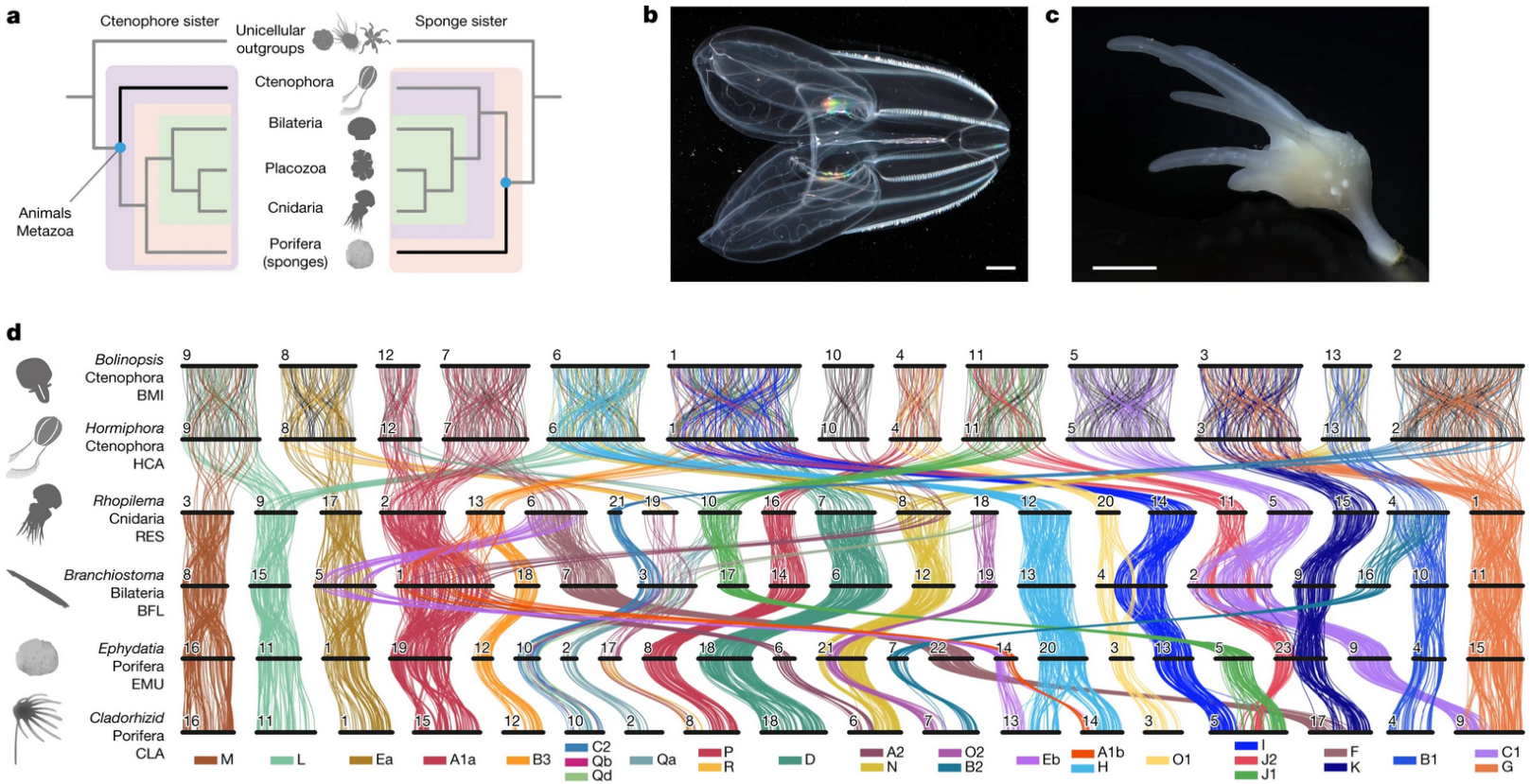


Log-likelihood score differences between inferred trees and "best-observed" trees plotted against topological distances (Zhou et al., 2017).

ML tree inference software:
For VERY large datasets (1000+sequences):

- RAxML/EXaML (Kozlov et al., 2015) is very efficient, especially with multiple runs
- IQ-TREE (Nguyen et al., 2015) also fast and relatively accurate
- FASTTREE(Price et al., 2009) is very fast, but (excessive) tradeoffs with accuracy (per Zhou et al. (2017))

Treating genomes holistically, rather than as a collection of nucleotides, codons, or proteins, may help to answer hard evolutionary questions.



Ancient gene linkages support ctenophores as sister to other animals (Schultz et al., 2023)

Rooting game break! Create a newick tree file in your text editor with the content: `(((C,(D,E)),(F,G),A),B);`
Save it as 'example.tre'.

Re-root the tree. What rootings make the following true? Which cannot be true?

1. A is more closely related to G than it is to C
2. (C,D,E) is sister to (A,B,F,G)
3. (C,D) is sister to (A,B,E,F,G)
4. (C,D,E) is a paraphyletic group
5. (C,D,E) is a monophyletic group
6. (A,B,C) is a monophyletic group

PHYLOGENETICS

Integrative phylogenomics positions sponges at the root of the animal tree

Jacob L. Steenwyk* and Nicole King*

Determining whether sponges or ctenophores root the animal tree has important implications for understanding early animal evolution. Here, we examined support for these competing hypotheses by constructing large and highly informative data matrices containing sequences from sponges, ctenophores, cnidarians, bilaterians, and diverse animal relatives. The new data matrices and 10 published datasets were analyzed in 785 topology tests conducted using integrative phylogenomics, a method that unifies concatenation and coalescence to identify genes with a consistent phylogenetic signal. All 490 statistically significant tests supported the sponge-sister hypothesis and none supported the ctenophore-sister hypothesis; the remaining 295 tests were inconclusive. These results provide compelling evidence for the sponge-sister hypothesis and suggest that integrative phylogenomics provides a robust and powerful approach for disentangling branches in the tree of life.

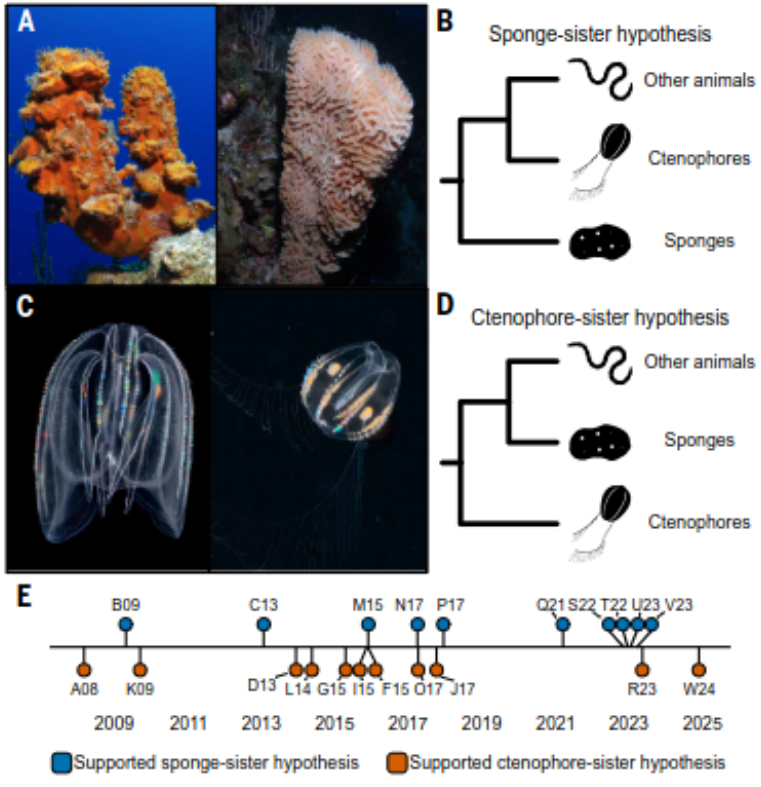


Fig. 1. Major competing hypotheses regarding the root of the animal tree of life. (**A**) Representative images of two sponges (credit: S. Nichols). (**B**) The sponge-sister

Steenwyk and King (2025)

Technical mistakes analysis can lead to incorrect results

Dunn et al. (including Jingxuan Chen!) 2025 comments on Steenwyk and King, 2025

"a procedural error in the use of iqtree. The trees specified with -z should be ... the two maximum likelihood phylogenies inferred under the ctenophore- and sponge-sister constraint trees. ... however, ... the tree file specified with -z was the constraint trees themselves"

`https://www.science.org/doi/10.1126/science.adw9456`

`https://github.com/caseywdunn/sk25`

Interactions between taxon sampling and automating analyses can create problems

comments on Steenwyk and King, 2025

Dunn et al. 2025: "... quartets that contain two sponges and two other animals are systematically incompatible with the collapsed ctenophore-sister tree but will often match the collapsed sponge-sister tree (despite being inherently uninformative about the root)."

Whelan 2025: "(some)...gene trees that "support" sponges-sister are obviously wrong, including errors like Ctenophora sister to a mite "

RESEARCH ARTICLE | PHYLOGENETICS

RETRACTED: Integrative phylogenomics positions sponges at the root of the animal tree

JACOB L. STEENWYK AND NICOLE KING Authors Info & Affiliations

SCIENCE • 13 Nov 2025 • Vol 390, Issue 6774 • pp. 751-756 • DOI: 10.1126/science.adw9456

This article has been retracted.

Retraction of Research Article "Integrative phylogenomics positions sponges at the root of the animal tree"

5 FEB 2026

Lessons from this retraction:

* Taxon sampling changes as you sample across loci
* Automation is both essential, and risky. Whenever possible, examine the trees and alignments themselves.
* Open data makes it possible to catch and address errors through re-analysis

NEWS | 10 March 2026

Keep calm and be transparent: advice from scientists who retracted their papers

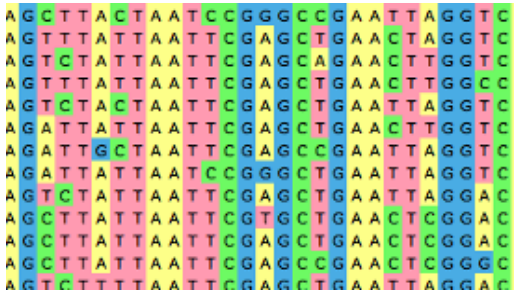
Retractions correct the scientific record, but they have stigma attached to them. Some in the research community want that to change.

By Sofia Caetano Avritzer

Model choice How do the data assembly choices we make in

to

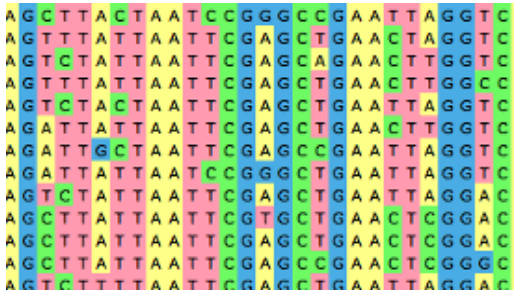
to



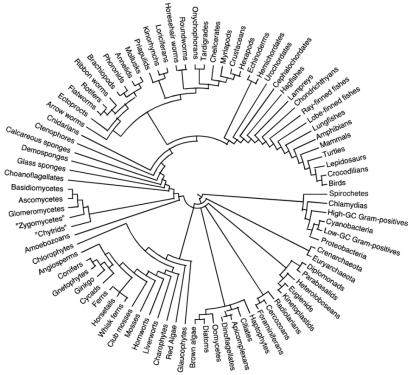
Model choice How do the data assembly choices we make in

to

to



affect

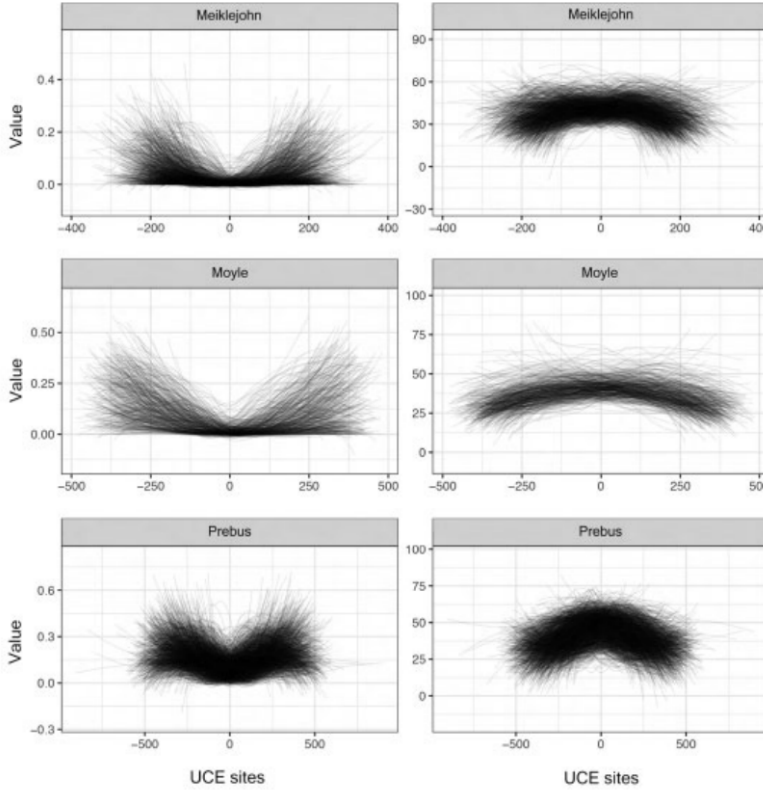


?

It is important to consider what models of evolution are appropriate for your data types

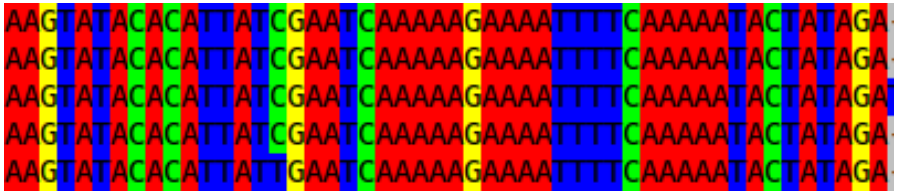
It is important to consider what models of evolution are appropriate for your data types

Entropy (rate proxy), GC content

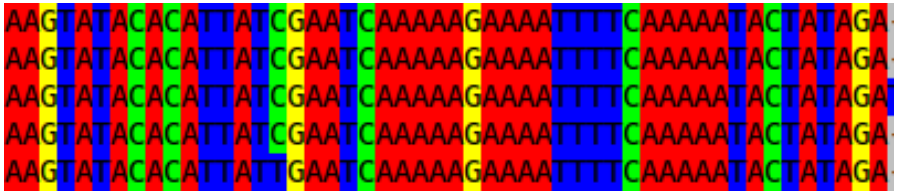


Extreme rate heterogeneity in Ultra Conserved Elements, can be handled with appropriate partitioning
(Tagliacollo and Lanfear, 2018)

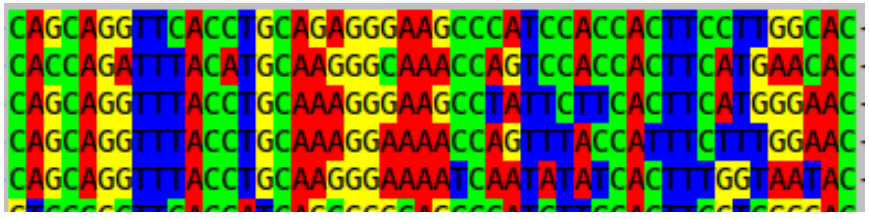
Short Tree


AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATTGAATCAAAAAGAAAATTTTCAAAAATACTATAGA

Short Tree



Long Tree



How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long

How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
but only if we looked for them!

How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
but only if we looked for them!

Can correct analysis of only variable sites (e.g. SNPs) by using appropriate model (implemented in most inference software)
(Lewis, 2001; Felsenstein, 1992)

Pay attention to data *'clean-up'* steps.

e.g.

- Minor allele frequency cutoffs
- Removing non-biallelic sites (multiple hit)
- Filtering out high rate regions
- 'Trimming' alignments

One method's "noise" is another method's data!

"it is possible in many cases to correct the ascertainment bias relatively easily, if reliable information is available regarding the details of the ascertainment scheme." (Nielsen, 2004)

"it is possible in many cases to correct the ascertainment bias relatively easily, if reliable information is available regarding the details of the ascertainment scheme." (Nielsen, 2004)

This information is not always available. Bias can be driven by the true, evolutionary history you are attempting to estimate!

Despite the large volume of data in genomic studies, ascertainment bias is still an issue

~~Despite~~ **because of** the large volume of data in genomic studies, ascertainment bias is still an issue

~~Despite~~ **because of** the large volume of data in genomic studies, ascertainment bias is still an issue

Large data sets can result in very high confidence in wrong answers if the model is not right for the data.

What to do?

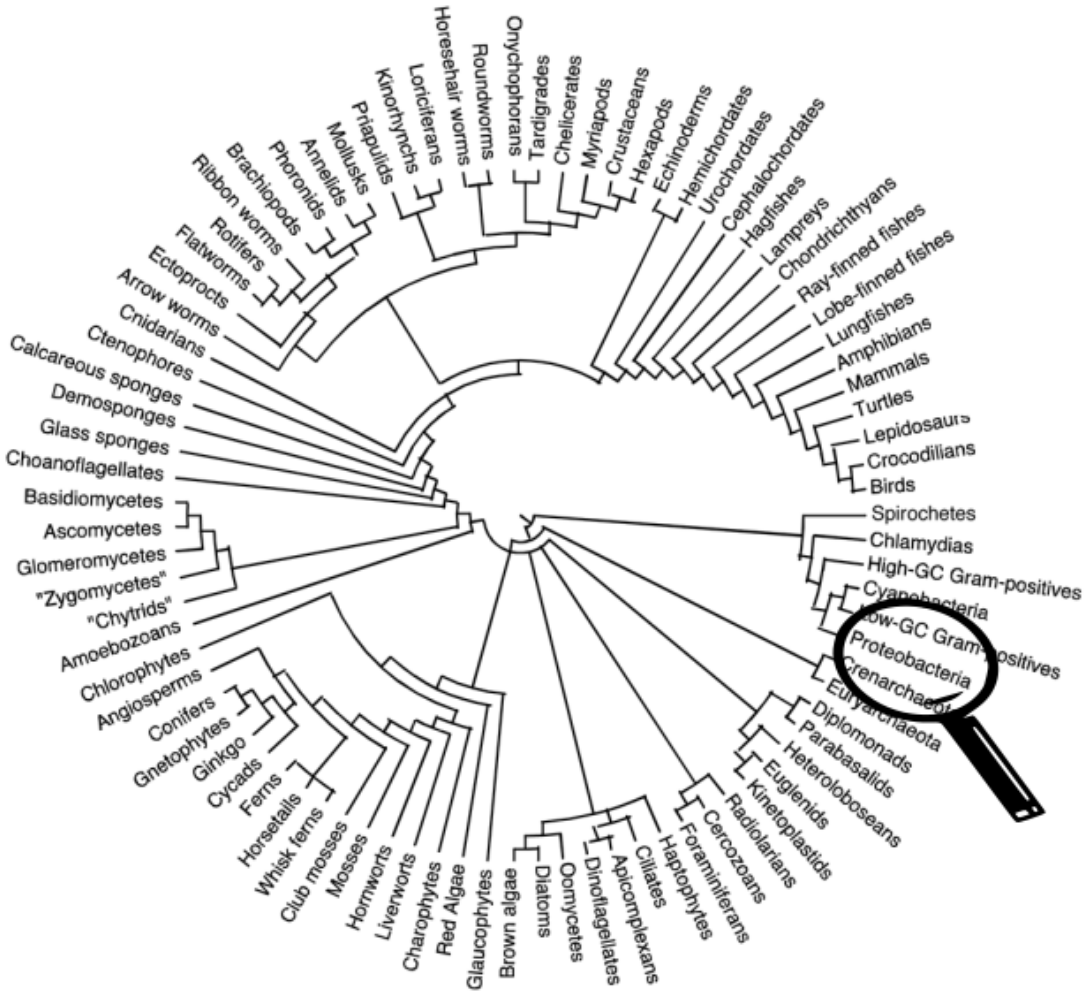
- What data will answer **your** questions?
- Use the most an appropriate available model for your data
- Re-sample your data to test if your key conclusions are robust to choices
- Look closely at your data and your intermediate results to make sure your analyses are doing what you think they are!

"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?

Case study and activity: tracing gonorrhea outbreaks



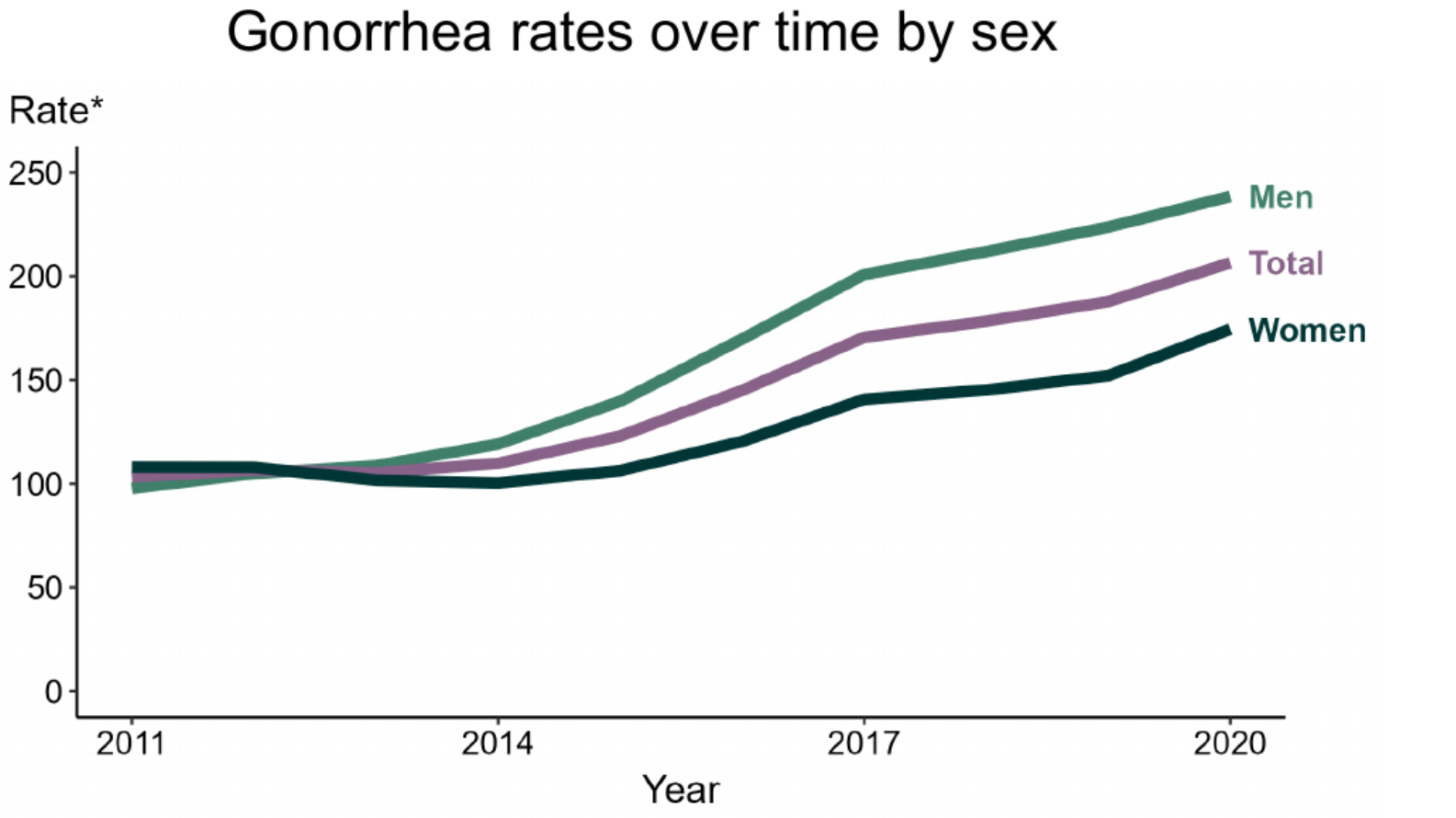
Rapid phylogenetic updating to trace gonorrhea outbreaks

Collaboration with

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

Neisseria gonorrhoeae

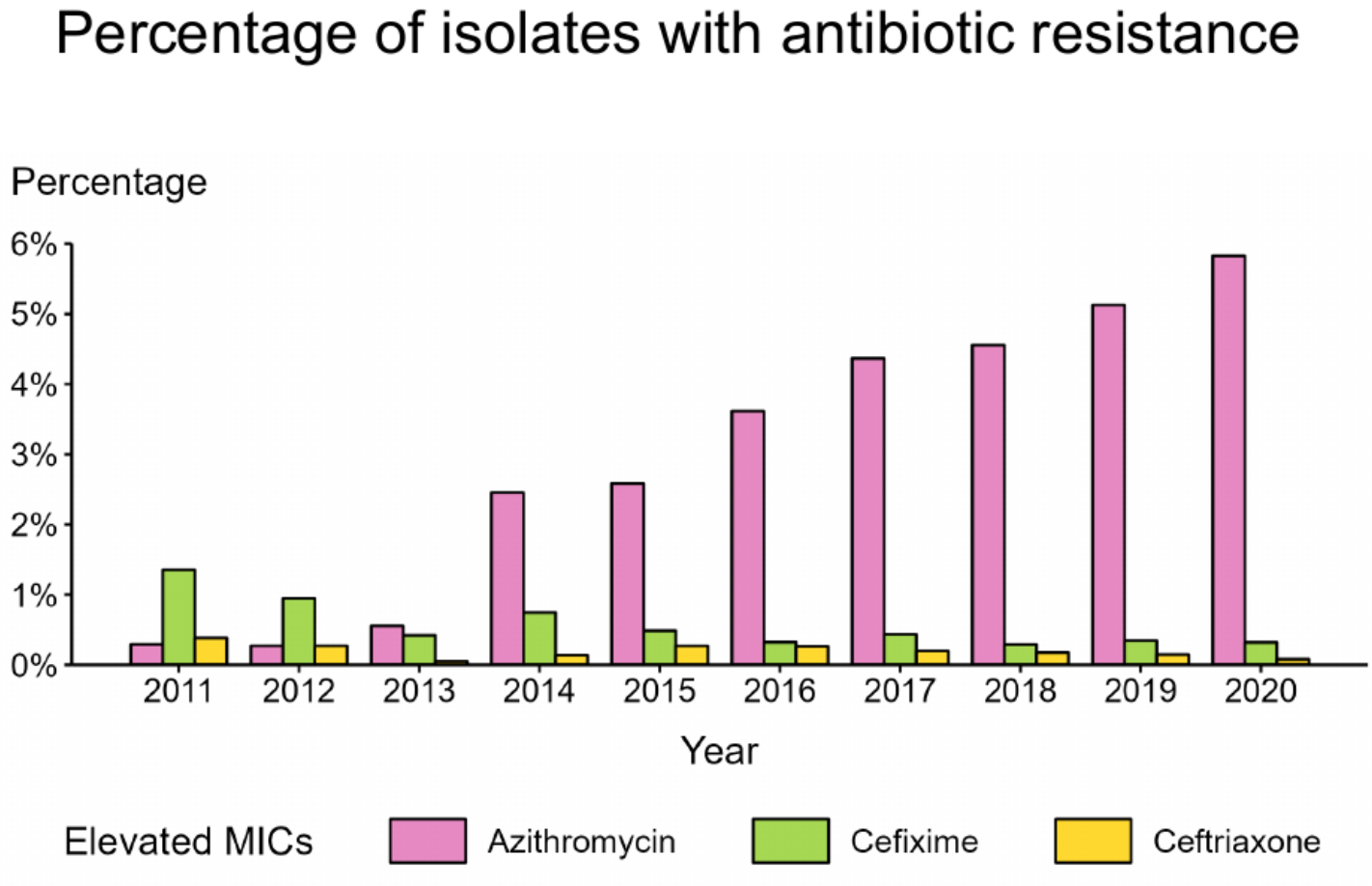
- Gram-negative, diplococci bacteria
- Responsible for the sexually transmitted infection known as gonorrhea
- One of two pathogenic *Neisseria* species known to infect humans
- WHO estimated 82 million new cases among adults worldwide in 2020



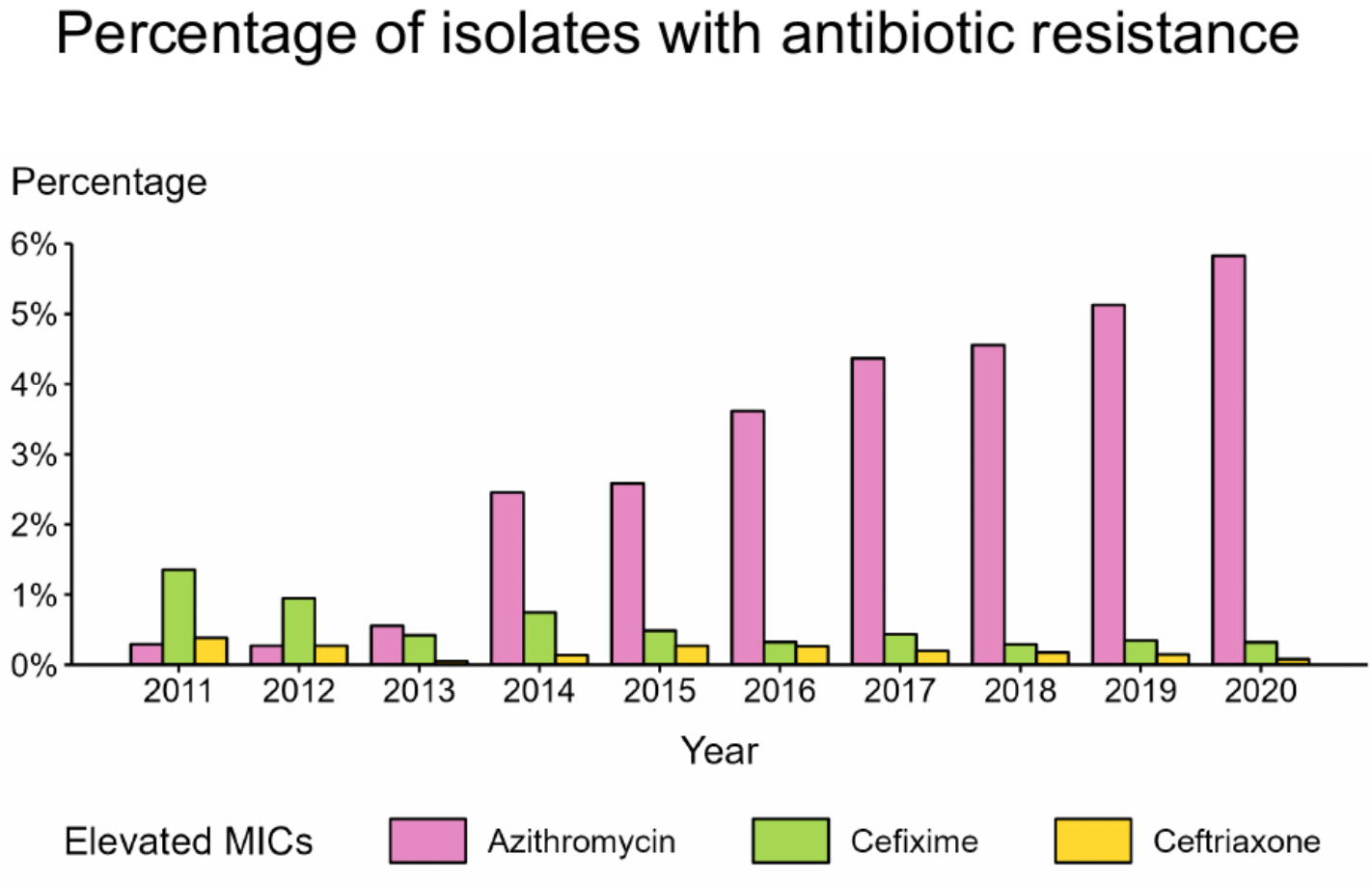
https://www.cdc.gov/std/statistics/2020/figures/GC-2.htm

Recent increase in rates of gonorrhea infections

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



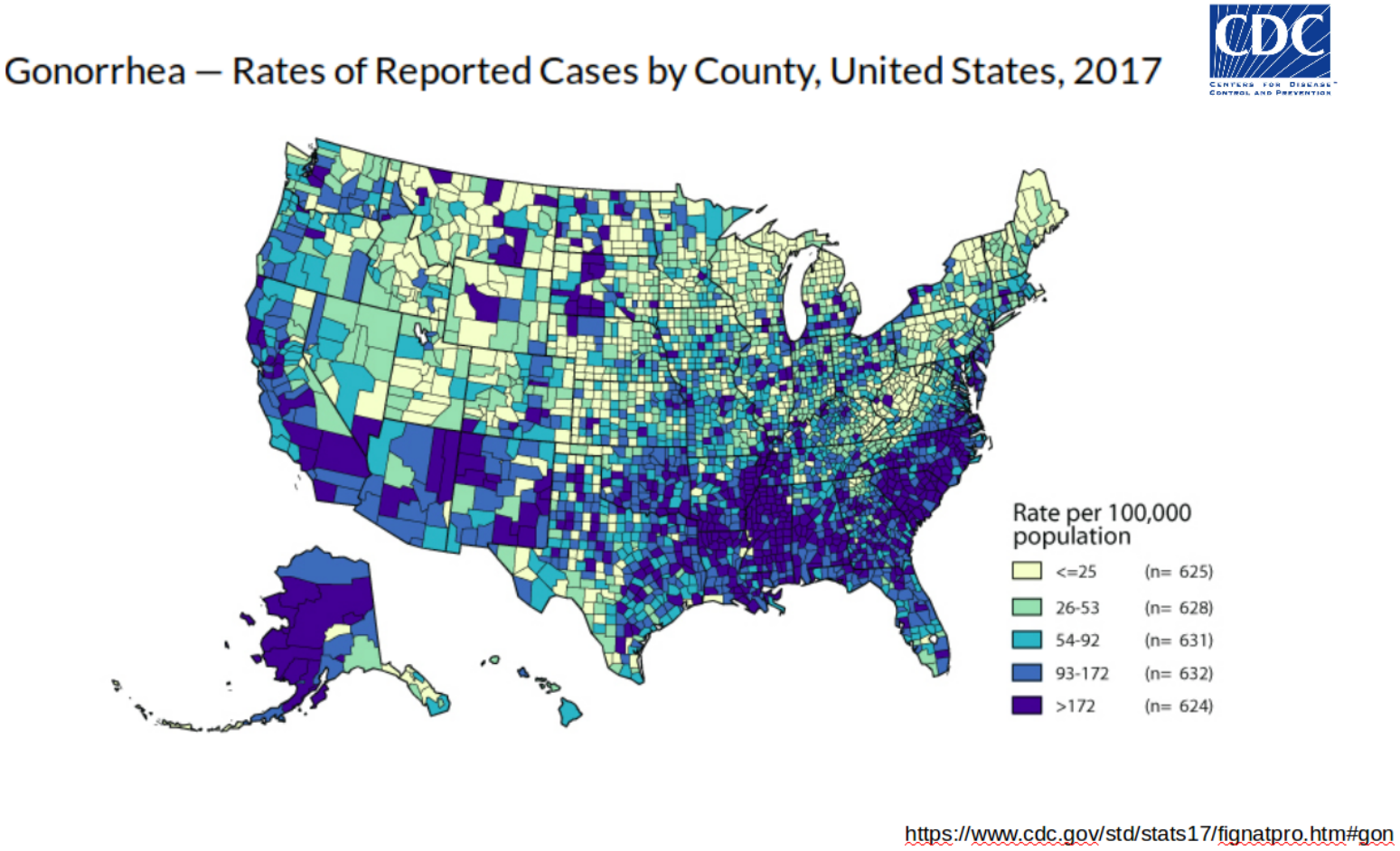
Neisseria gonorrhoeae has progressively developed resistance to each single dose antibiotic.



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.

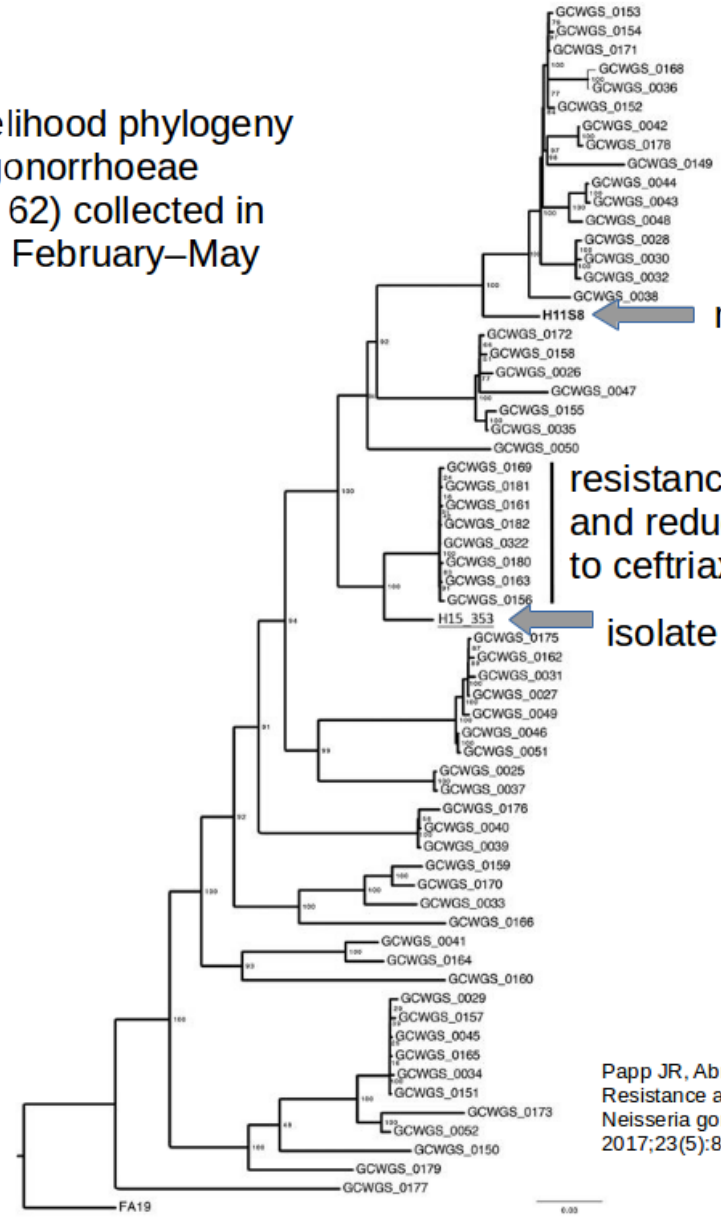


https://www.cdc.gov/std/stats17/figpro.htm#gon

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

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

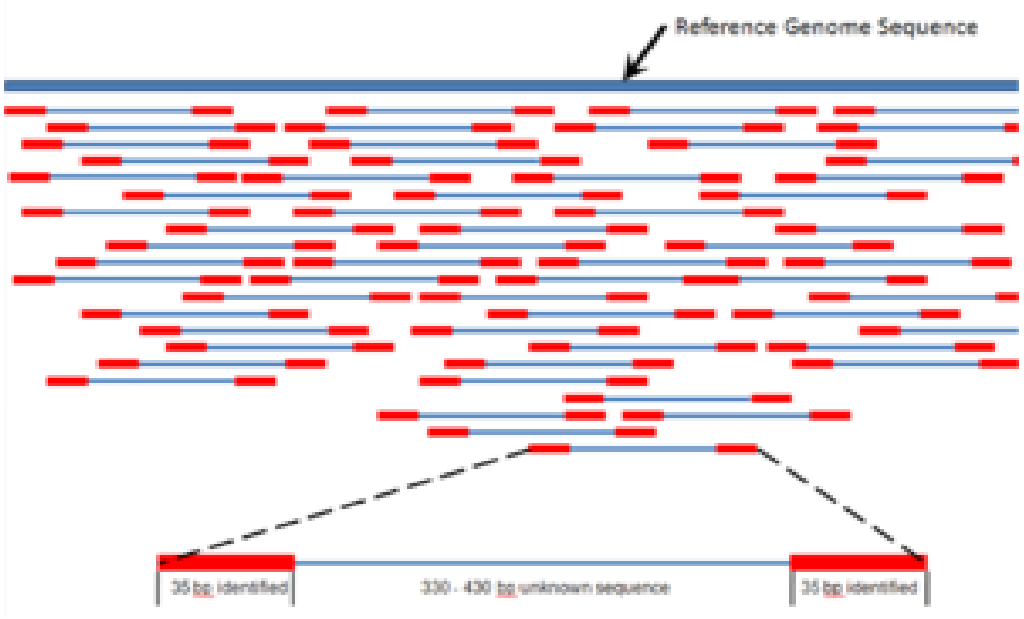
Sequencing errors are likely to be singletons

Will overestimate tip branch lengths

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

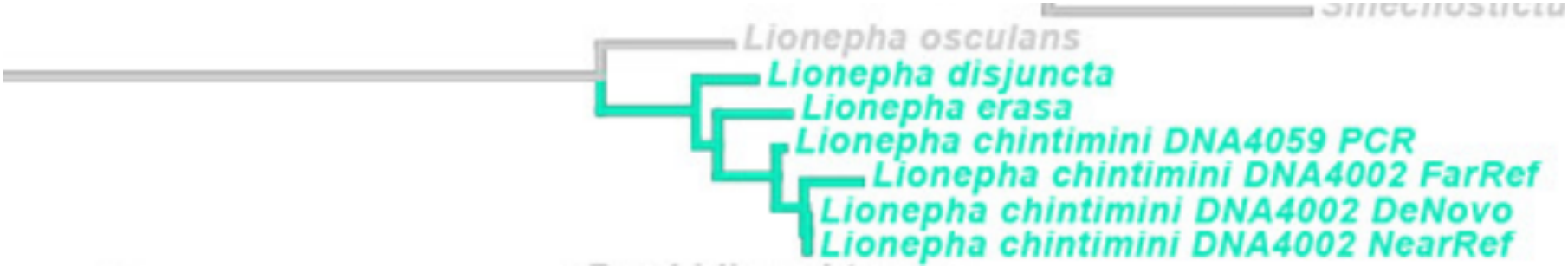
BUT: Reference choice can affect evolutionary inference

- 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

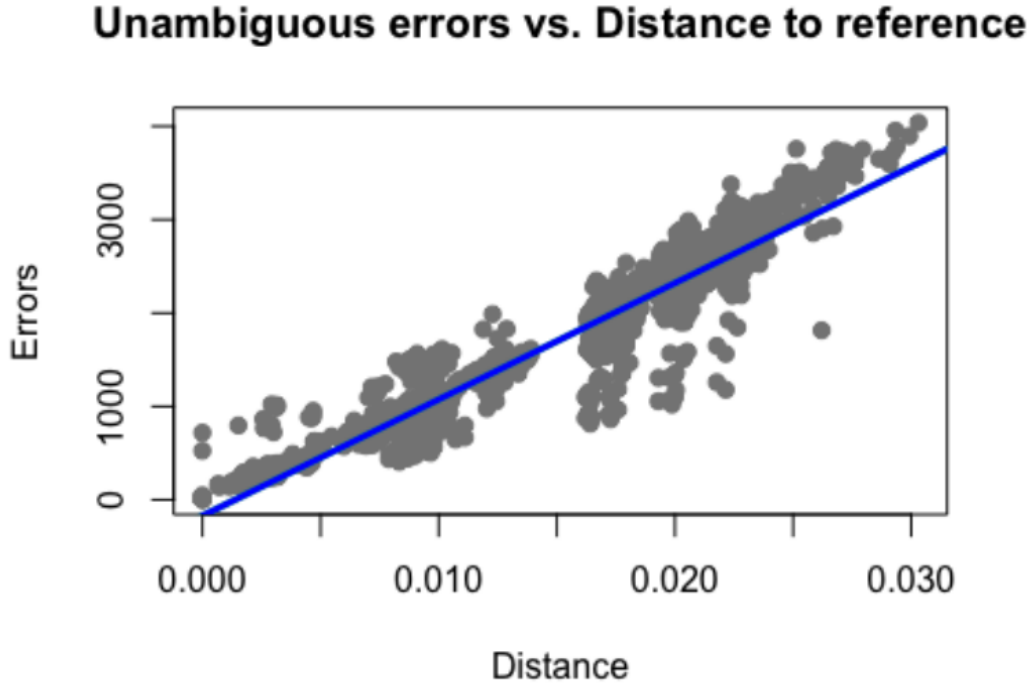
- In humans, in highly polymorphic regions variant calling is biased toward the the reference base (Brandt et al., 2015)

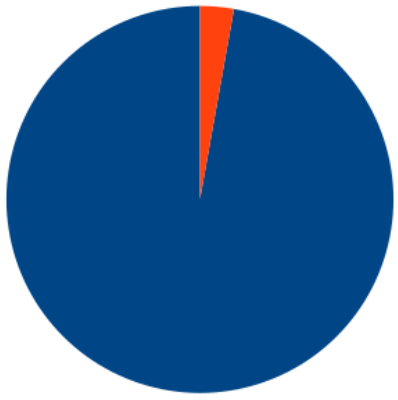
- In fragmented DNA samples from beetles, branch lengths change based on reference choices



(Kanda et al., 2015)

In experimental re-analysis of UCE data, error rate is correlated with distance to reference genome, and errors are strongly biased to the reference base

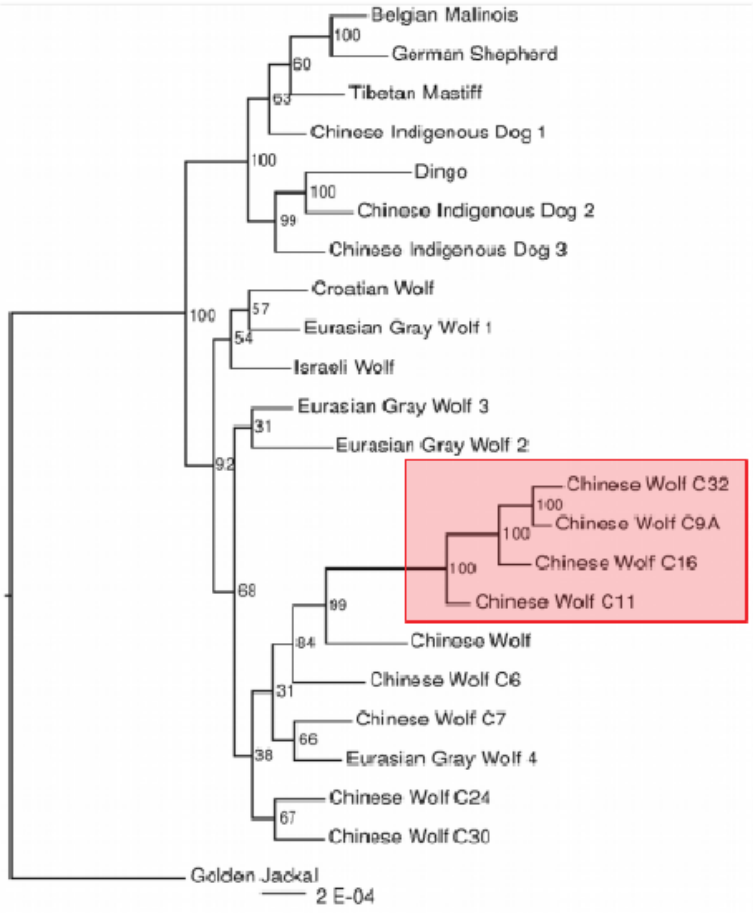




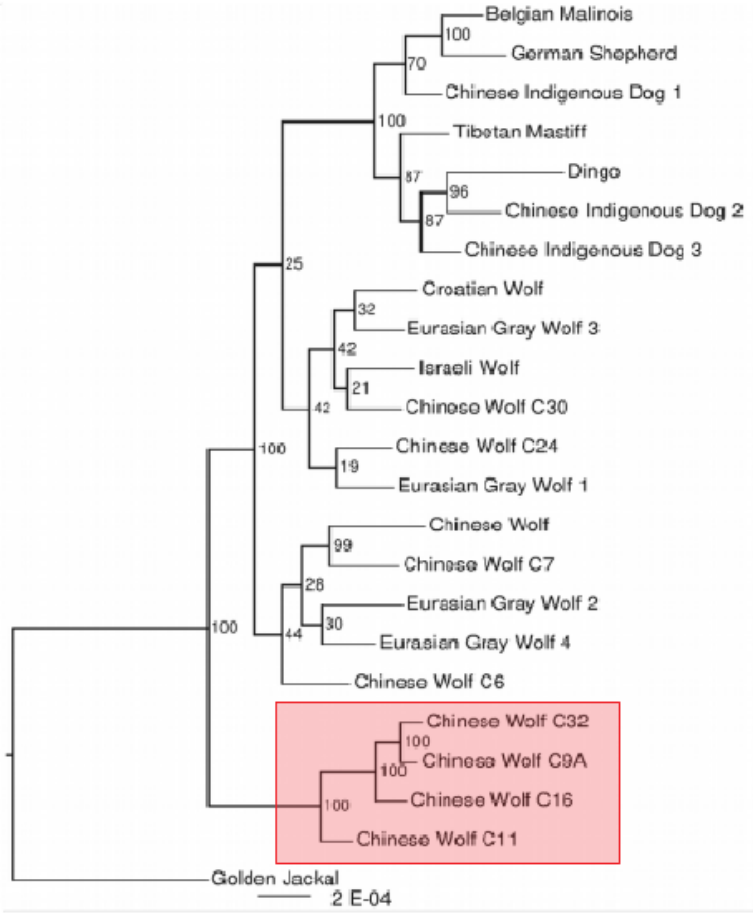
Base call errors match the reference base 97% of the time

(Toscani-Field and McTavish, work in progress)

Reference choice can affect topology



Dog Reference



Wolf Reference

Gopalakrishnan et al. (2017)

Hands on exercise!

How much does what reference you use matter when reconstucting *Neisseria gonorrhoeae* phylogenies?

`https://github.com/snacktavish/TreeUpdatingComparison/blob/master/TreeUpdating.md`

Summary

Bias: Reference choice

Effect on inference:

Errors may be biased towards the sites found in the reference genome used for assembly

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

Mitigation: Use multiple reference genomes, compare results

Conclusions:

When a closely related reference is available, alternatives worsen inference

Sequence calls do change based on choice of reference

BUT phylogenetic conclusions were not affected

Baker, M. (2012). *De novo* genome assembly: what every biologist should know. *Nature Methods*, 9:333–337.

Boussau, B., Szöllosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330. Number: 2.

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.

Brown, J. M. and Thomson, R. C. (2016). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, page syw101.

Chiari, Y., Cahais, V., Galtier, N., and Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biology*, 10(1):65. Number: 1.

Darling, A. E., Miklós, I., and Ragan, M. A. (2008). Dynamics of Genome Rearrangement in Bacterial Populations. *PLoS Genetics*, 4(7). Number: 7.

Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J., and Dessimoz, C. (2024). Inference of phylogenetic trees directly from raw sequencing reads using read2tree. *Nature Biotechnology*, 42(1):139–147.

Felsenstein, J. (1992). Phylogenies from Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46(1):159–173. Number: 1.

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.

Gupta, A., Mirarab, S., and Turakhia, Y. (2025). Accurate, scalable, and fully automated inference of species trees from raw genome assemblies using ROADIES. *Proceedings of the National Academy of Sciences*, 122(19):e2500553122. Publisher: Proceedings of the National Academy of Sciences.

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.

Kozlov, A. M., Aberer, A. J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*, 31(15):2577–2579. Number: 15.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925. Number: 6.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274. Number: 1.

Nielsen, R. (2004). Population genetic analysis of ascertained SNP data. *Human genomics*, 1(3):218–224. Number: 3.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641–1650. Number: 7.

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., and Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, 618(7963):110–117. Number: 7963 Publisher: Nature Publishing Group.

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.

Smith, M. L. and Hahn, M. W. (2021). New Approaches for Inferring Phylogenies in the Presence of Paralogs. *Trends in Genetics*, 37(2):174–187.

Smith, M. L., Vanderpool, D., and Hahn, M. W. (2022). Using all Gene Families Vastly Expands Data Available for Phylogenomic Inference. *Molecular Biology and Evolution*, 39(6):msac112.

Steenwyk, J. L. and King, N. (2025). Retracted: Integrative phylogenomics positions sponges at the root of the animal tree. *Science*, 390(6774):751–756.

Tagliacollo, V. A. and Lanfear, R. (2018). Estimating Improved Partitioning Schemes for Ultraconserved Elements. *Molecular Biology and Evolution*, 35(7):1798–1811. Number: 7.

Turner, J. H., Kuster, R. D., Staton, M. E., and Moulton, J. K. (2026). Orthogarden: a pipeline for propagating phylogenetic trees for non-model organisms from short reads and de novo genome assemblies. *Molecular Biology and Evolution*, 43(3):msag053.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *bioRxiv*, page 142323.