

Genomic data for evolutionary inference

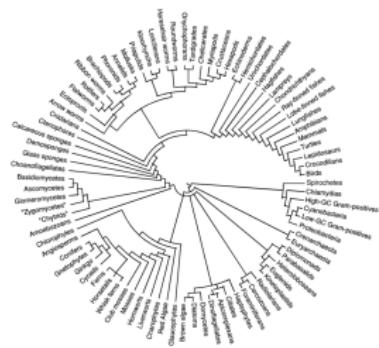
Emily Jane McTavish

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

How do you get from



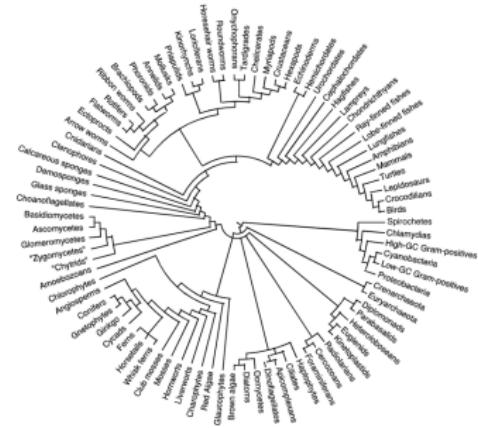
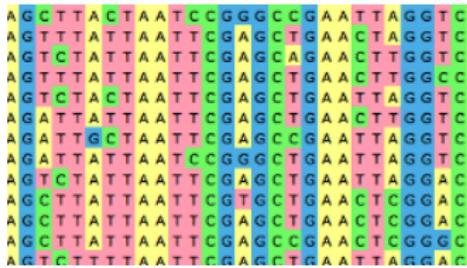
to



?

You've seen a lot about how to get from

to



I'm going to talk about going from



to



to

```

A G C T T A C T C T A A T C C G G G C C G A A T T A G G T C
A G T T T A T T A A T T C G A C T G T G A A C T A G G C C
A G T C T A T T A A T T C G A G C A G A A C T T G G G C C
A G T T T A T T A A T T C G A G C T G G A A T T A G G T C
A G T C T A C T A A T T C G A G C T G G A A T T A G G T C
A G A T T A T T A A T T C G A G C T G G A A C T T G G G C C
A G A T T G C T A A T T C G A G C C G A A T T A G G T C
A G A T T A T T A A T T C G G G G C C G A A T T A G G T C
A G T C T A T T A A T T C G A C T G T G A A T T A G G T C
A G G T T A T T A A T T C G T T G A C T G T G A A T T A G G T C
A G G T T A T T A A T T C G A C T G T G A A C T T G G G A C
A G G T T A T T A A T T C G A C T G C G A C T G T G A A T T A G G T C
A G G T T A T T A A T T C G A C T G C G A C T G T G A A T T A G G T C
A G G T T A T T A A T T C G A C T G C G A C T G T G A A T T A G G T C

```

I'm going to talk about going from



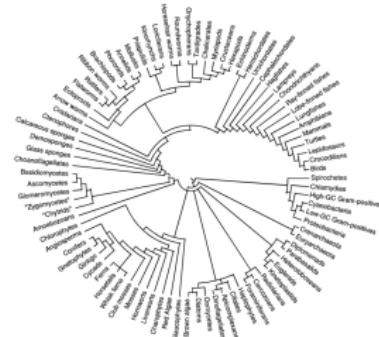
to



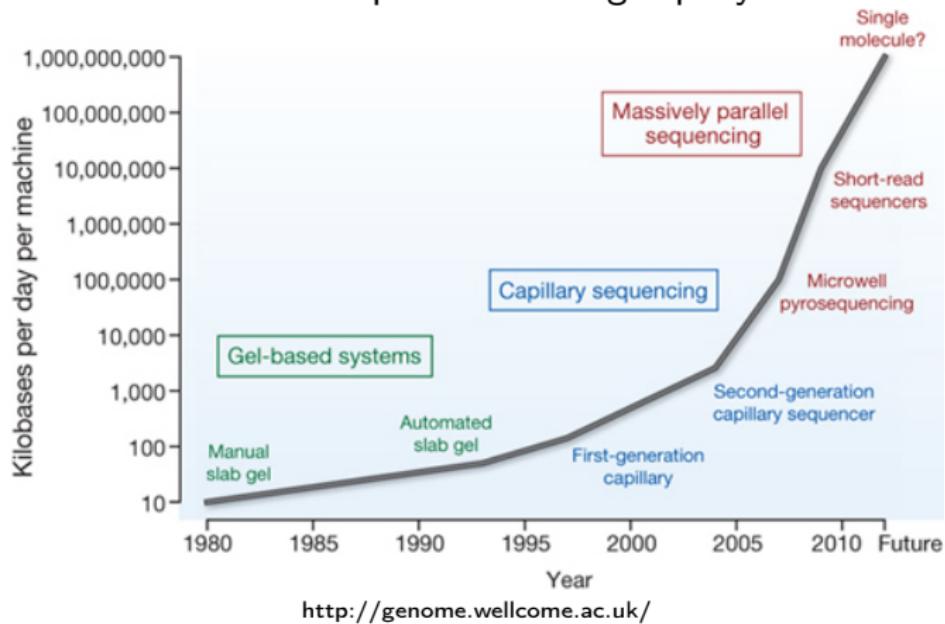
to

A grid of DNA sequence data. The grid has 10 columns and 10 rows. Each cell contains a four-letter sequence (e.g., ACGT) in a color-coded scheme where A is green, C is pink, G is blue, and T is yellow. The sequences represent a single row of a genome.

and how those choices can affect



The quantity of available sequence data for inferring evolutionary relationships is increasing rapidly



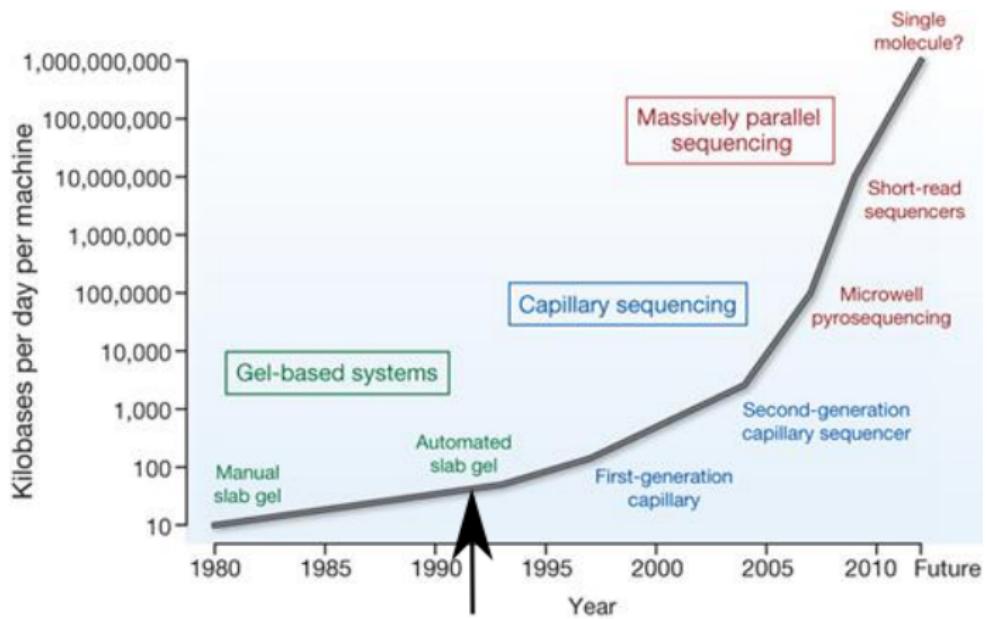
“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”

– Jeff Thorne (Evolutionary biologist)

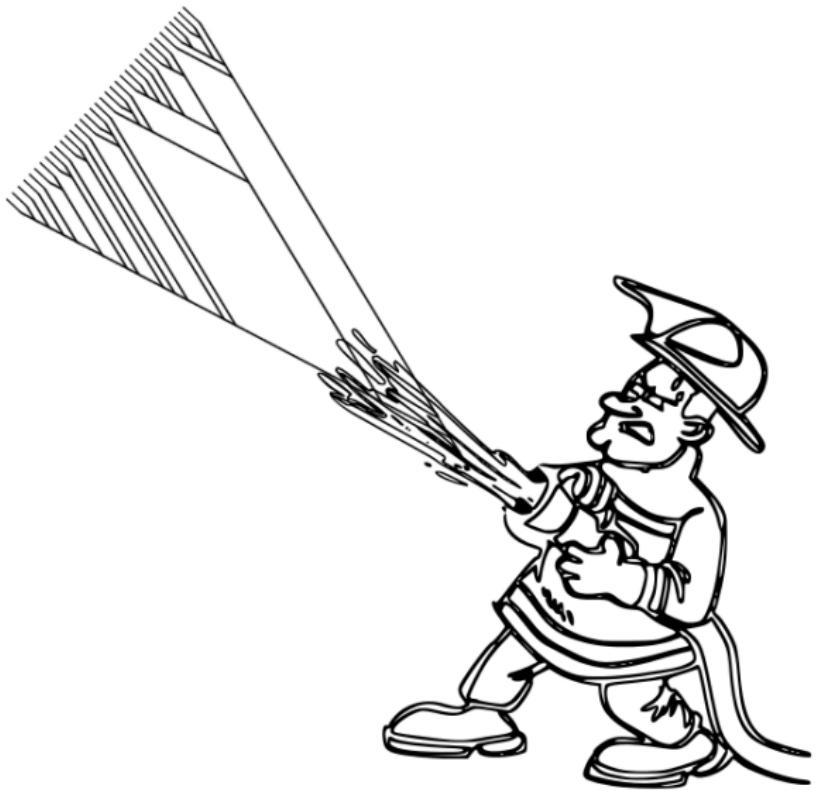
“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”

– Jeff Thorne (Evolutionary biologist)

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



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



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



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

Enrichment (smallest to largest proportion of genome)

- * Directed PCR
- * Targeted enrichment, Rad-tag etc
- * Transcriptome
- * Whole genome

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

Survey question! PollEv.com/emilyjanemctavish820

Directed PCR

Simple and cheap for a small number of genes

Doesn't scale so well to many genes

Doesn't sound fancy

Targeted enrichment (e.g. Ultra-conserved elements, probes for orthologous single copy genes, etc.)

- Use hybridization to enrich particular regions

- Works well even on degraded DNA

- Need to synthesize probes specific to each region
 - need data to get data!

- Data sets can be combined across projects if same probe set applied

Non-targeted enrichment (RAD-tag, ddRAD etc.)

Select randomly distributed, but consistent, genome regions

Comparable across closely related taxa, but not more distant taxa

Each locus has very few variable sites (not good for generating gene trees)

Whole transcriptome

Enriched for expressed protein coding genes

Content will vary based on cell type,
environment, etc.

Provides expression level data

Whole genome sequencing

Capture all the data

In a phylogenetic context, often only cost effective for small genomes

Annotation is hard! Often need transcriptome to get genes

Mapping or assembly can be slow,

Need to put the pieces back together (usually)!



to

A	G	C	T	T	A	C	T	A	A	T	C	G	G	G	C	C	G	A	A	T	T	A	G	G	T	C		
A	G	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	A	G	G	T	C	
A	G	T	C	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	A	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C
A	G	A	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C
A	G	T	C	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	C	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	C	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	C	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	
A	G	T	C	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C	

Genomic sequencing

You have all the data! 

You have to deal with all of the data. 

De novo assembly

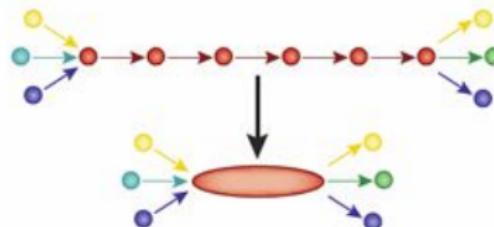
1. Fragment DNA and sequence



2. Find overlaps between reads

...AGCCTAGACCTACA**GGATGCGCGACACGT**
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

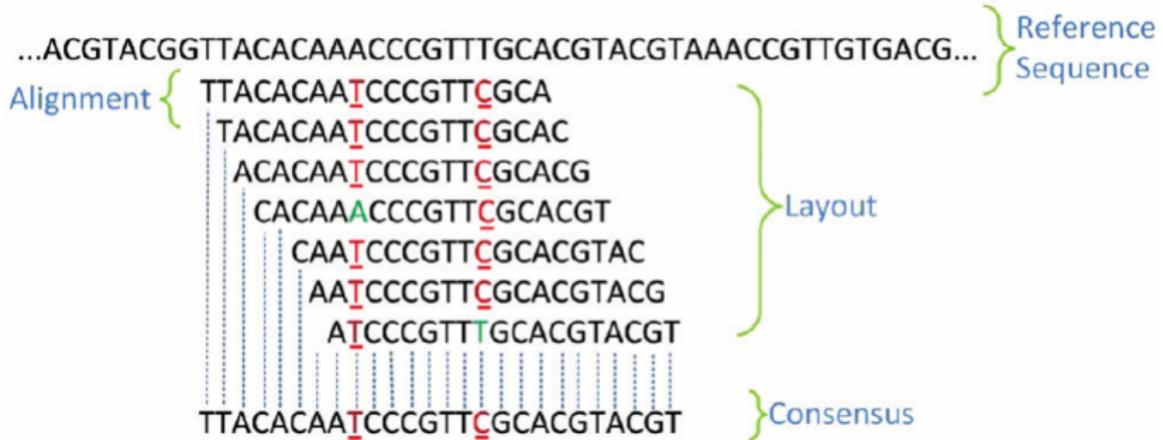


4. Assemble contigs into scaffolds



(Baker, 2012)

Mapping to a reference genome

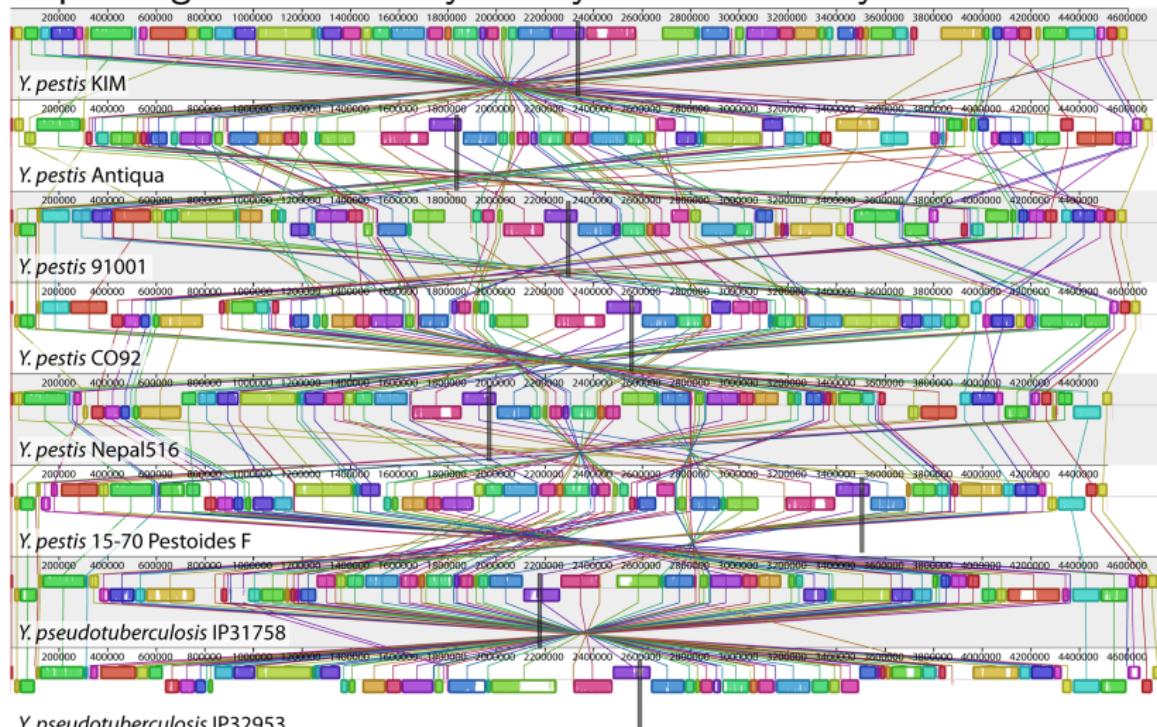


To make evolutionary statements, you need to align genomic regions across taxa.

Depending on evolutionary history this can be easy or hard!

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

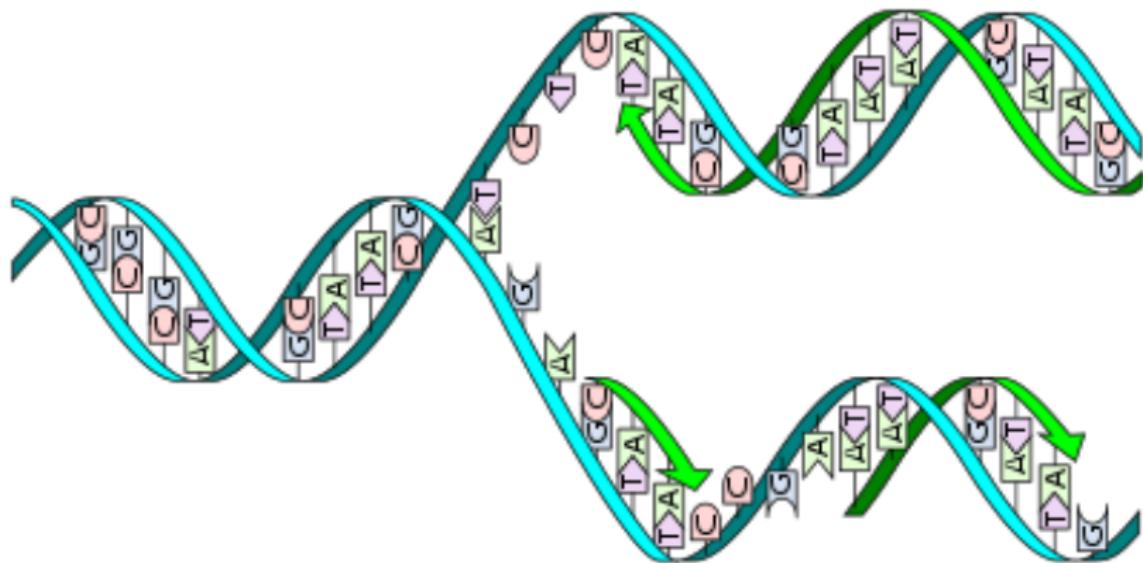
Depending on evolutionary history this can be easy or hard!



Y. pseudotuberculosis IP32953

(Darling et al., 2008)

An alignment is a statement of shared ancestry



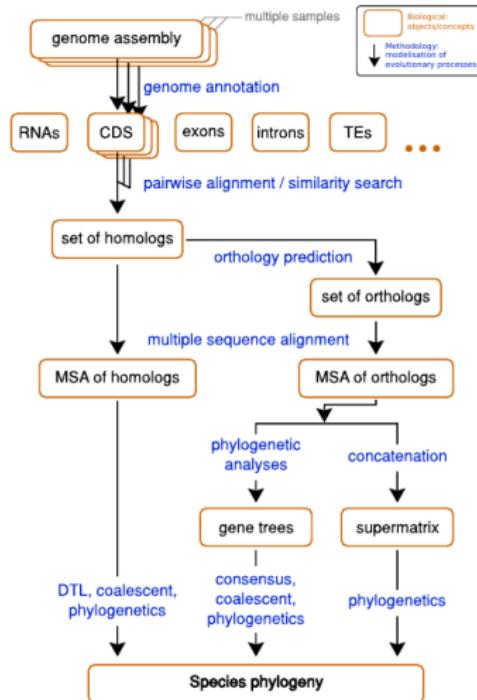
Gene tree (Locus tree)

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

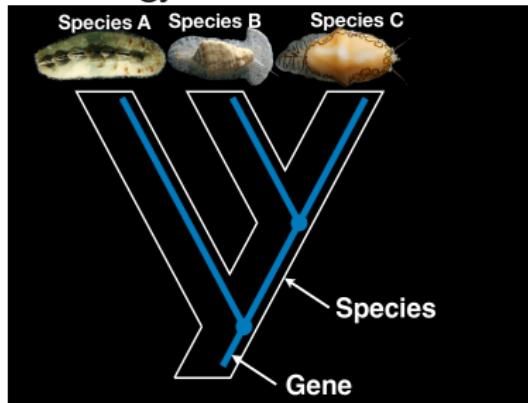
<https://inria.hal.science/PGE/page/table-of-contents>



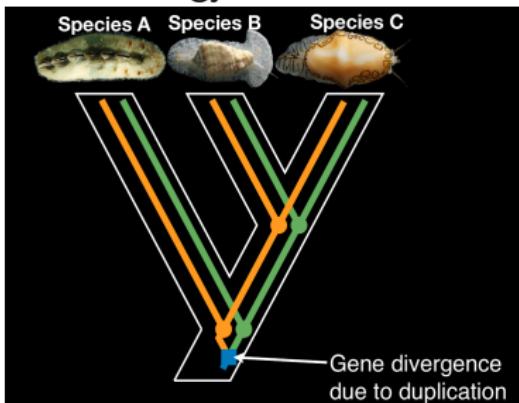
Simion et al. (2020)

Gene duplication and loss

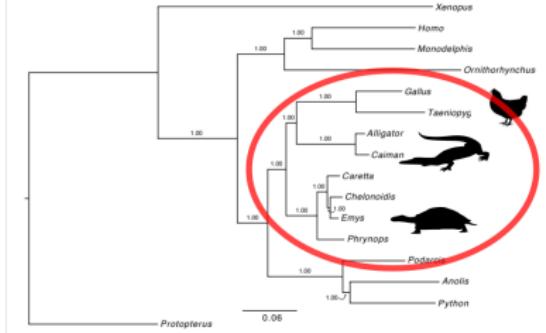
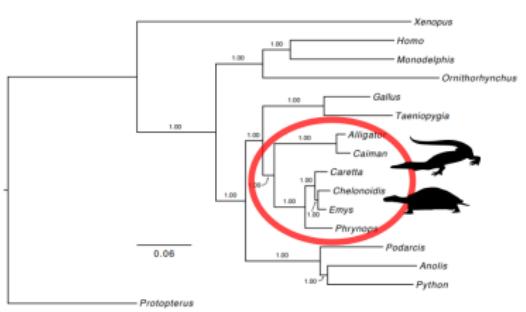
Orthology



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

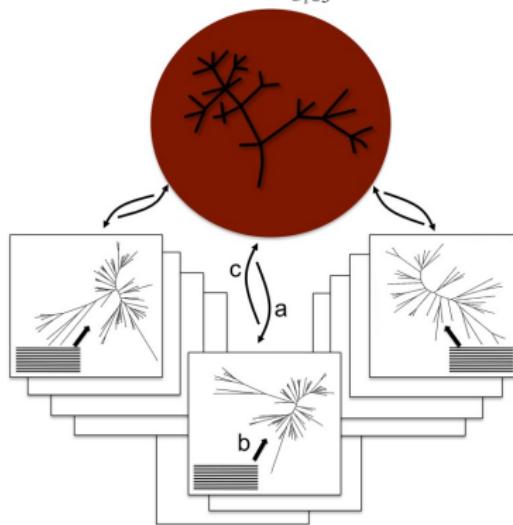


“investigation of genes with extreme support for turtle placement revealed unappreciated paralogy in a small proportion of alignments (<1%) that had an extraordinary influence on the inferred placement of turtles.”
(Brown and Thomson, 2016) (Chiari et al., 2012)

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)

Using all Gene Families Vastly Expands Data Available for Phylogenomic Inference

Megan L. Smith ,^{1,2} Dan Vanderpool  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

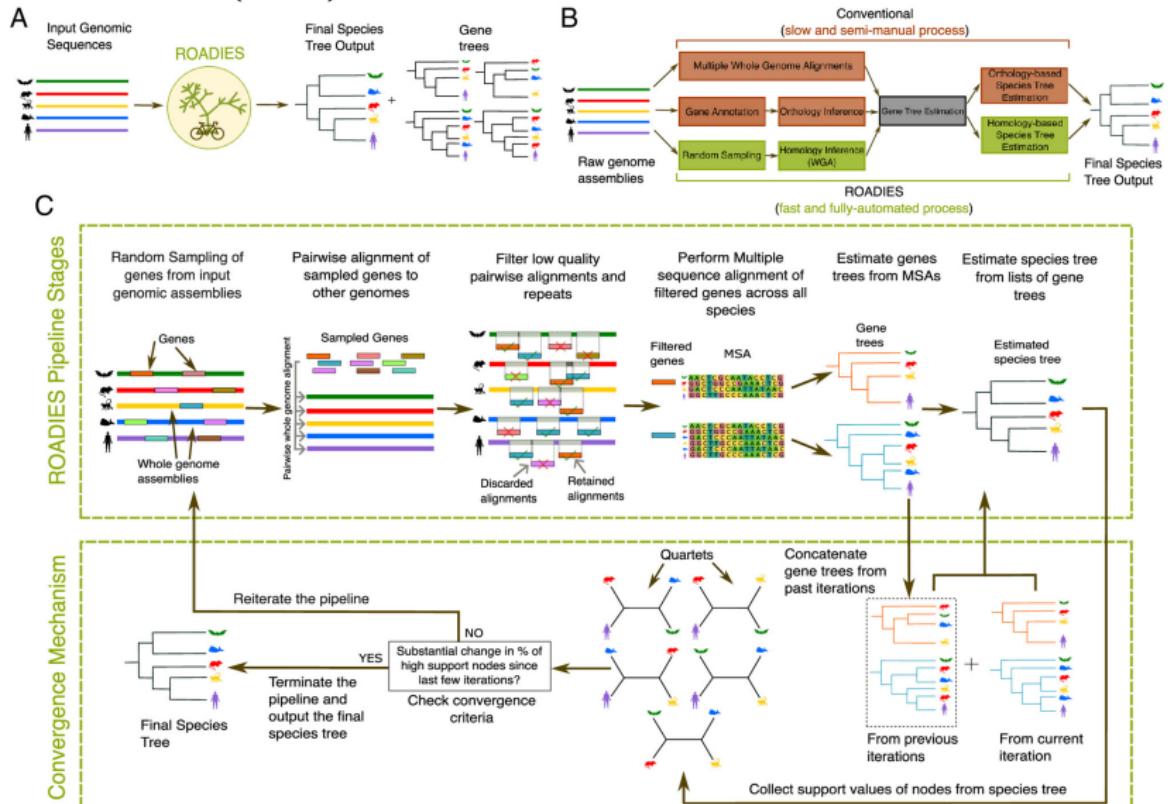
Review

New Approaches for Inferring Phylogenies in the Presence of Paralogs

Megan L. Smith^{1,*} and Matthew W. Hahn¹

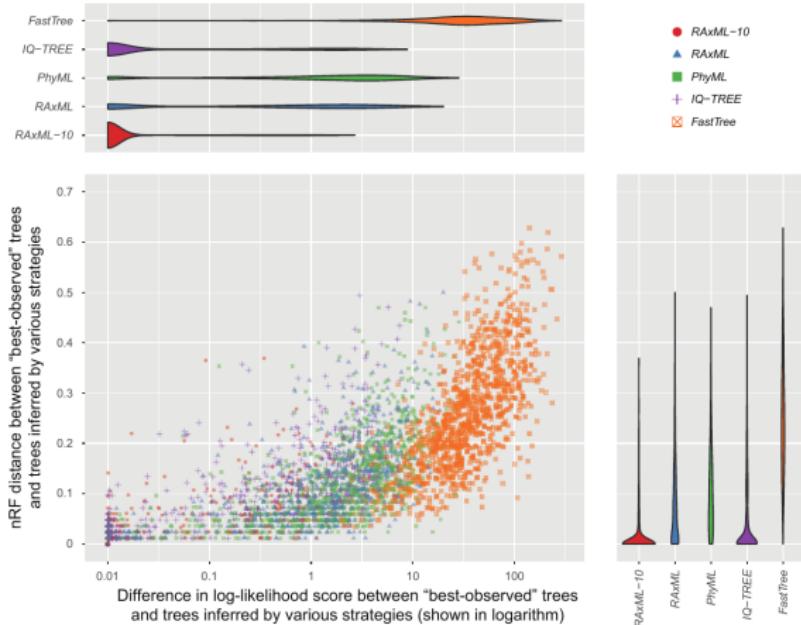
Smith et al. (2022); Smith and Hahn (2021)

New this month: ROADIES - “Reference-free, Orthology-free, Annotation-free, Discordance-aware Estimation of Species Trees,” Gupta et al. (2025)



Analyzing genome scale data sets can be SLOW.
Are the tradeoffs of faster methods worth it?

Figure 3



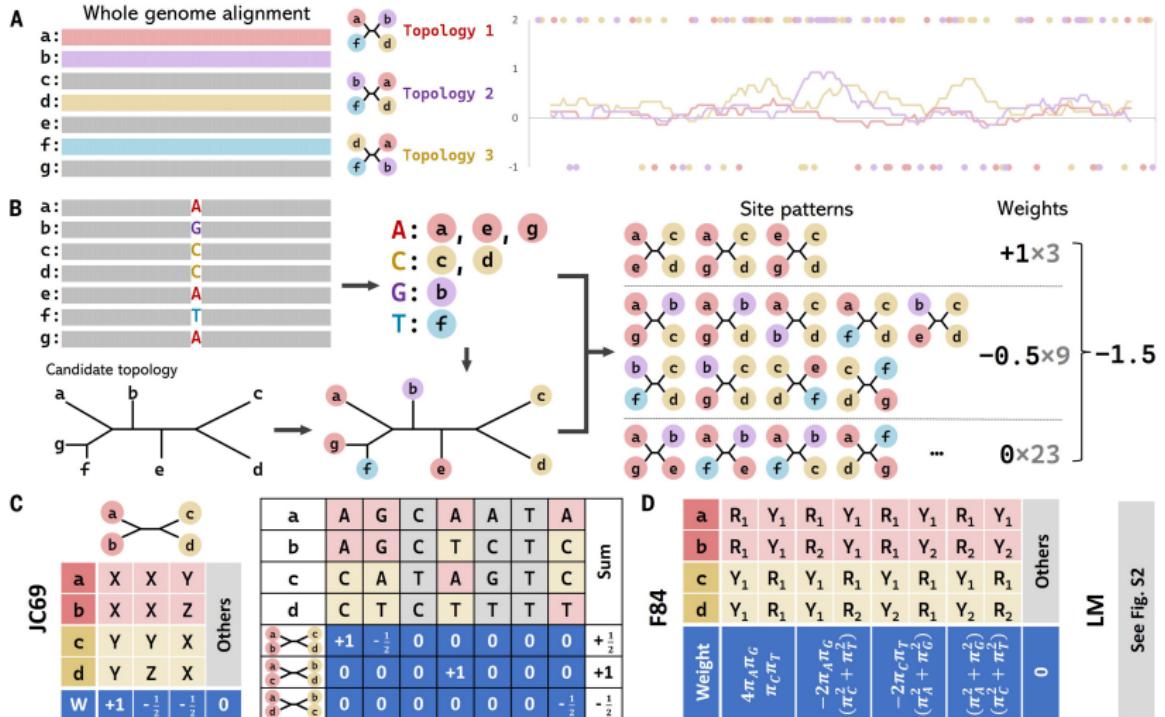
Log-likelihood score differences between inferred trees and “best-observed” trees plotted against topological distances (Zhou et al., 2017).

ML tree inference software:

For VERY large datasets (1000+sequences):

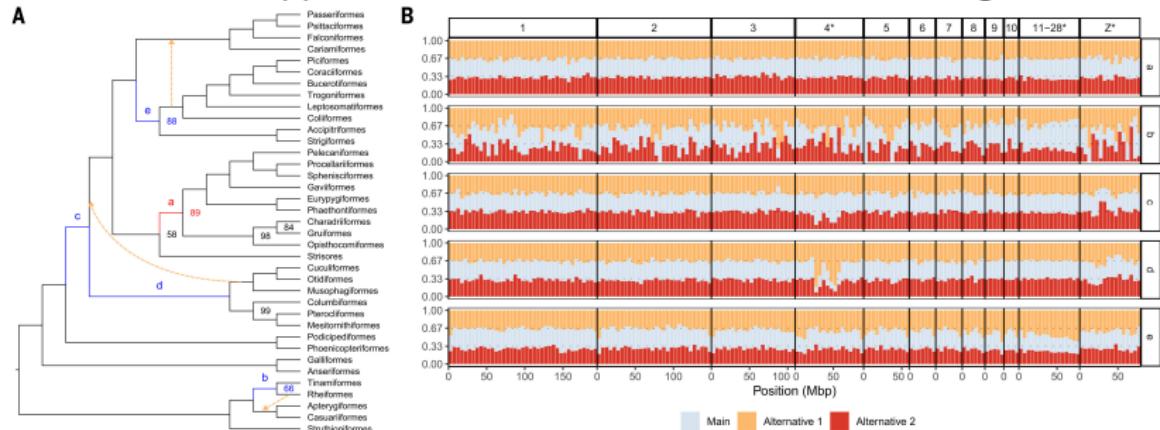
- RAxML/EXaML (Kozlov et al., 2015) is very efficient, especially with multiple runs
- IQ-TREE (Nguyen et al., 2015) also fast and relatively accurate
- FASTTREE(Price et al., 2009) is very fast, but (excessive) tradeoffs with accuracy (per Zhou et al. (2017))

New this year: CASTER - site based quartet method



Zhang et al. (2025)

Can visualize support for different resolutions across the genome



Zhang et al. (2025)

Incorporates substitution models - but not really model based. No branch lengths.

a	A	G	C	A	A	T	A	Sum
b	A	G	C	T	C	T	C	
c	C	A	T	A	G	T	C	
d	C	T	C	T	T	T	T	
(a b) >> c d	+1	- $\frac{1}{2}$	0	0	0	0	0	+ $\frac{1}{2}$
(a c) >> b d	0	0	0	+1	0	0	0	+1
(a d) >> b c	0	0	0	0	0	0	- $\frac{1}{2}$	- $\frac{1}{2}$

How well CASTER performs on hard problems (like ones that violate the simple conditions of simulations) is not clear!

Very fast! 11 hours for 140 genes across 1,100 taxa on a data set we tried, but the tree did not align well with our expectations/accepted taxonomy.

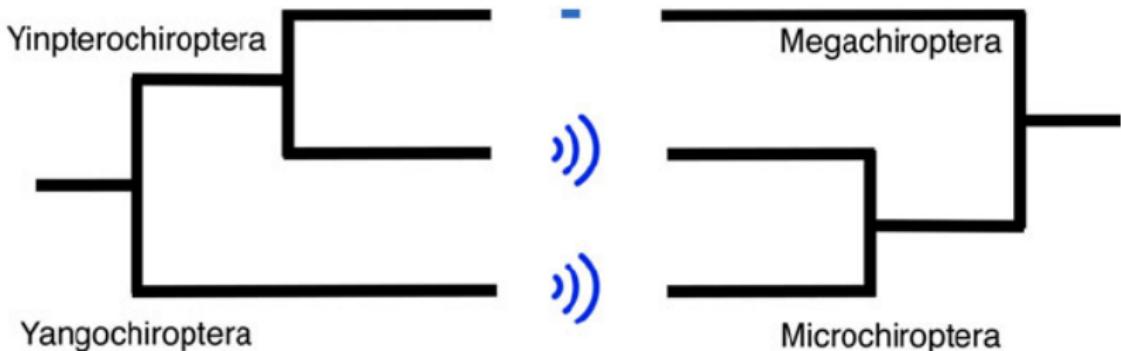
How well CASTER performs on hard problems (like ones that violate the simple conditions of simulations) is not clear!

Very fast! 11 hours for 140 genes across 1,100 taxa on a data set we tried, but the tree did not align well with our expectations/accepted taxonomy.

If you have to use “quick and dirty” or black box methods in order to be able to analyze large data sets - more data may result in WORSE answers

Is the species tree even what you want?

Different gene trees can drive different conclusions

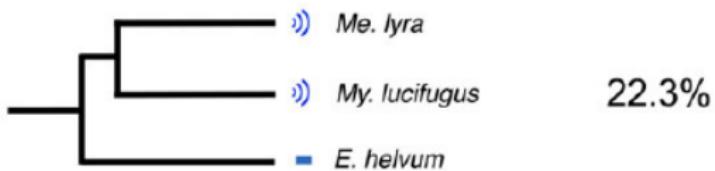
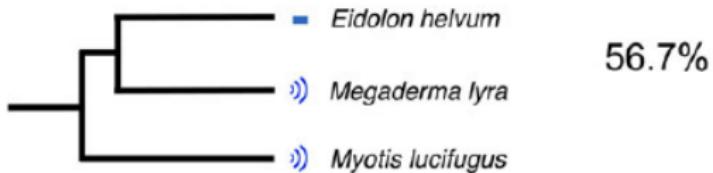


Species relationships between echolocating and nonecholocating bats (after Teeling 2009). Left: inferences from DNA sequence data.

Right: traditional species relationships inferred from morphological characters (and limited sequence data). (Hahn and Nakhleh, 2016)

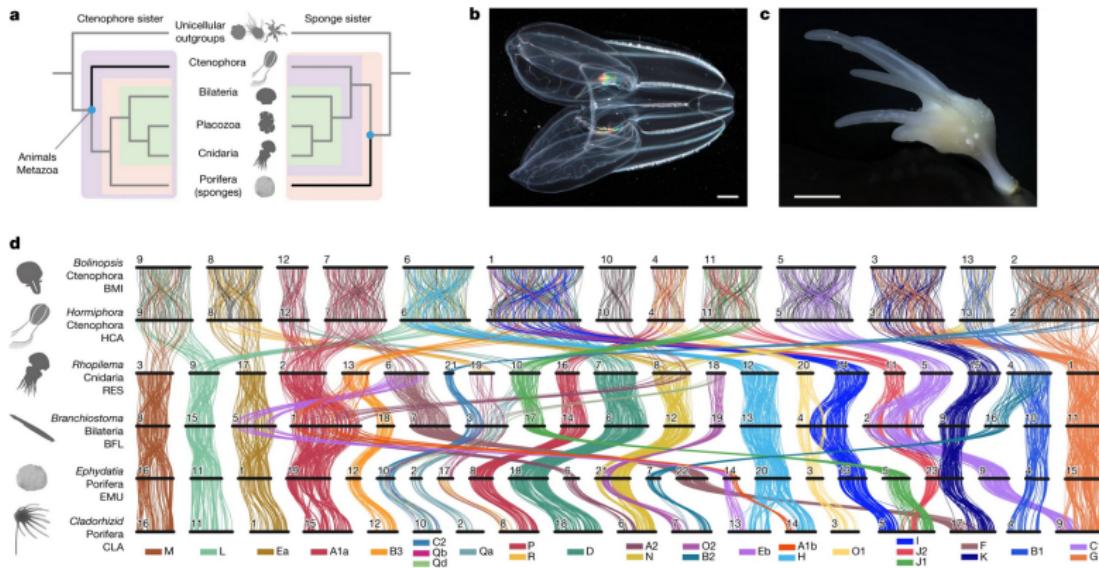
If you are interested in a trait controlled by one or a few genes, the species tree may not describe the evolutionary history.

B



(Hahn and Nakhleh, 2016)

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)

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!

Data processing/ascertainment bias

How do the choices we make in



to



to

A grid of DNA sequence data represented by colored boxes. The columns are labeled with the four bases: A (green), T (red), C (blue), and G (yellow). The rows show different sequence contexts, such as AT, TA, CT, CG, AG, GC, etc. The sequence reads from top to bottom as follows:

A	G	C	T	T	A	C	T	A	A	T	C	C	G	G	C	G	A	A	T	T	A	G	G	T	C
A	G	T	T	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	A	G	G	T	C		
A	G	T	C	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C		
A	G	T	C	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	C	C		
A	G	A	T	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C		
A	G	A	T	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C		
A	G	C	T	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	G	C		
A	G	C	T	T	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	G	G		

How do the choices we make in



to



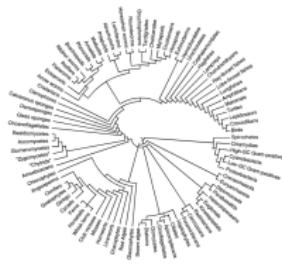
to

affect

```

A G C T T A C T A A T C C G G G C C G A A T T A G G T C
A G T T T A T T A A T T C C G A G C T G A A C T T A G G T C
A G T C T A T T A A T T C C G A G C T G A A C T T T G G G C
A G T T T A T T A A T T C C G A G C T G A A C T T T G G G C
A G T C T A C T A A T T C C G A G C T G A A C T T A G G T C
A G T T T A T T A A T T C C G A G C T G A A C T T T G G G C
A G A T T A T T A A T T C C G A G C G G A A T T A G G T C
A G A T T C T G A T T A A T T C C G A G C G G A A T T A G G T C
A G T T T A T T A A T T C C G G G C C G A A T T A G G T C
A G T C T A T T A A T T C C G A G C T G A A C T T A G G T C
A G T T T A T T A A T T C C G T G C T G A A C T T C T G G G A C
A G C T T A T T A A T T C C G A G C T G A A C T T C T G G G A C
A G C T T T T T A A T T C C G A G C G G A A T T A G G T C
A T C T T T T A A T T C C G A G C T G A A C T T A G G T C

```



?

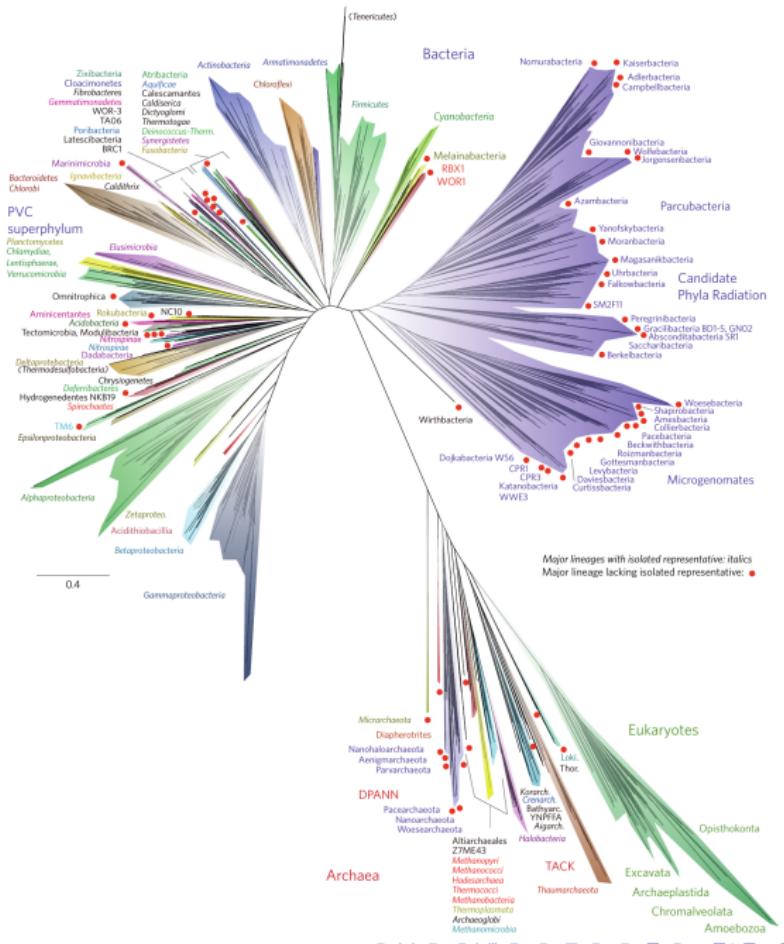
Ascertainment bias

A bias in parameter estimation or testing caused by non-random sampling of the data.
(also sometimes overlapping with 'selection bias' or 'acquisition bias')

Ascertainment bias is ubiquitous!

- Surveying volunteers
- Studying undergraduates
- Sampling across 'species'
- Discarding rare outliers

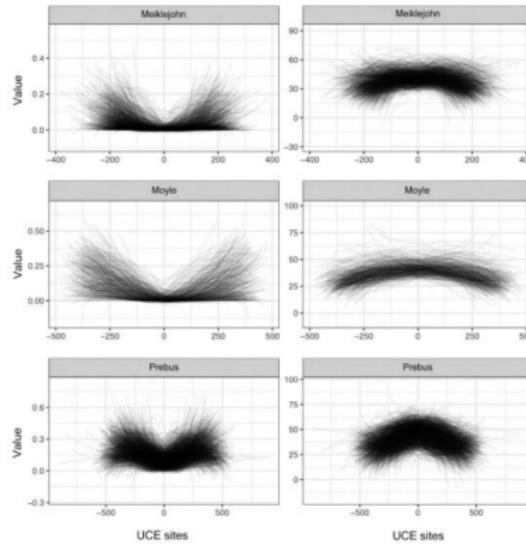
Sampling across the tree of life (Hug et al., 2016)



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)

Should you trim your alignments?

Short Tree

AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA
AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA
AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA
AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA
AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA

The sequence logo displays the consensus sequence AAGTATACACATTATCGAAATCAAAAGAAAATTTCAAAAAATCTATAGA. The letters are color-coded: A (green), T (yellow), C (blue), and G (red). The height of each letter at a position indicates its frequency of occurrence at that position across the aligned sequences.

Short Tree

AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA
AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA
AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA
AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA
AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA
AAGTATACACATTATCGAACTAAAAAGAAAATTTCCTAAAAATCTATAGA

Long Tree

CAGCAGGGTTTACCTTGCAAGGGCAAAACAGTCCACCACTTCCTGGCAC
CACCAAGATTTTACATGCAAGGGCAAAACAGTCCACCACTTCATGAACAC
CAGCAGGGTTTACCTTGCAAGGGAGCCATTTCCTTCACTTCAGGGAAC
CAGCAGGGTTTACCTTGCAAGGGAAAAACAGTTTACCATTCCTGGAAC
CAGCAGGGTTTACCTTGCAAGGGAAAAACAAATATAACACTTGGTAATAC

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

Case studies:

Phylogenetics of *Penstemon* using RADseq data
Tracing gonorrhea outbreaks

Phylogenetics of Penstemon using RADseq data

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



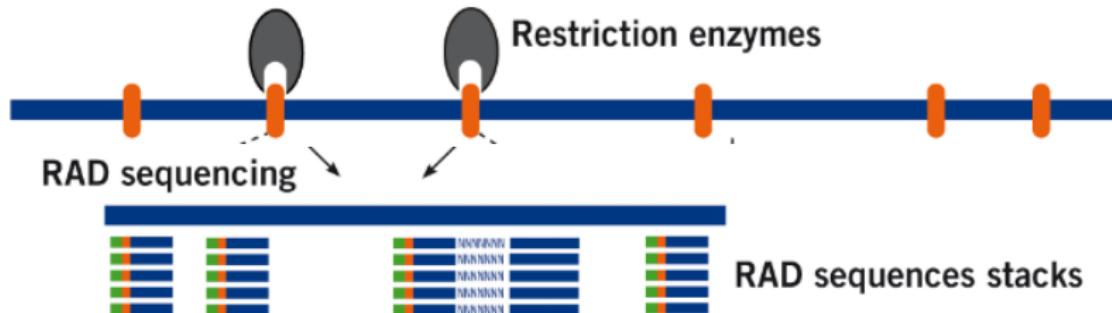
Data:

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

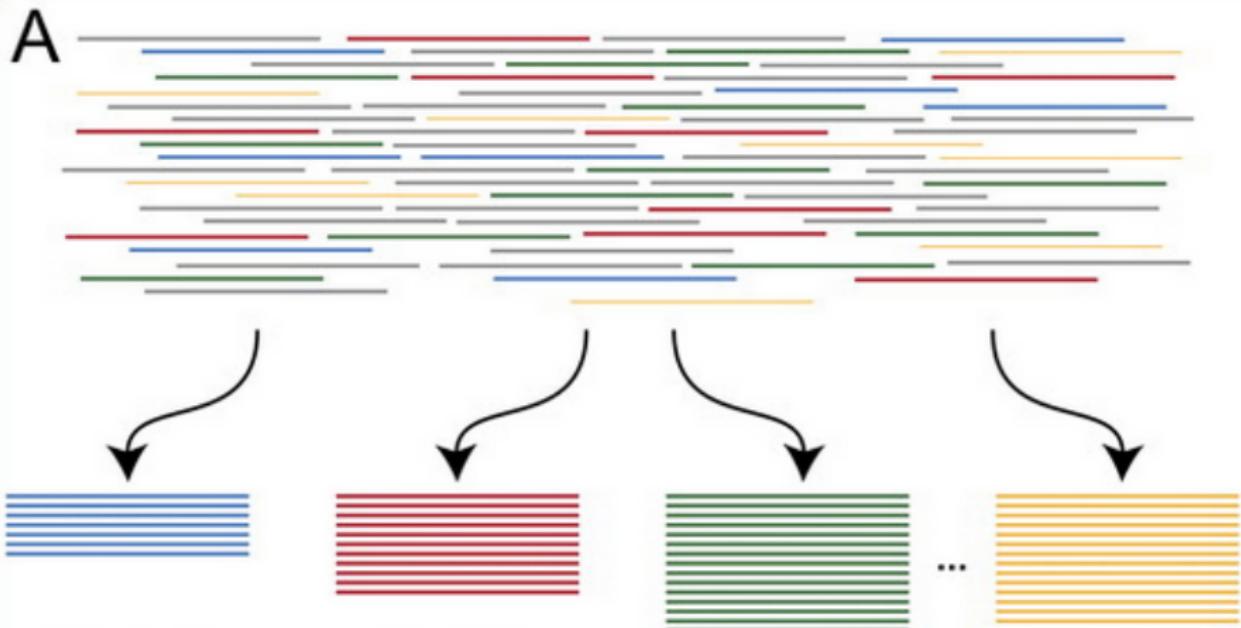


In comparison: Shotgun Sequencing



(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 (Catchen et al., 2011)

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

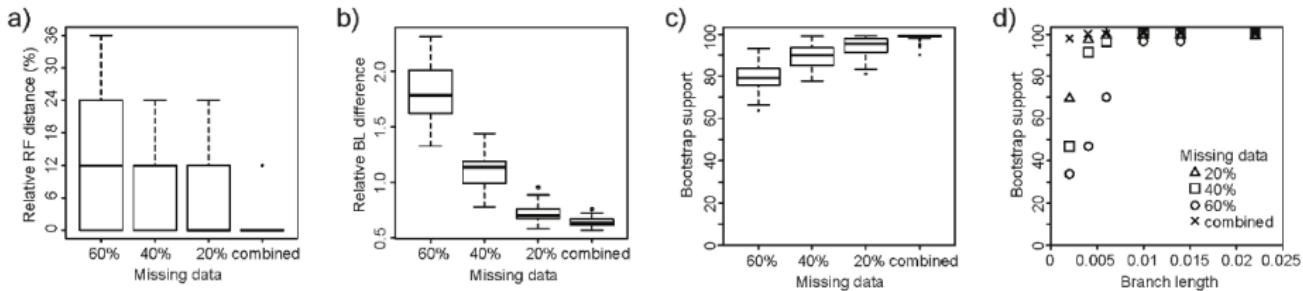


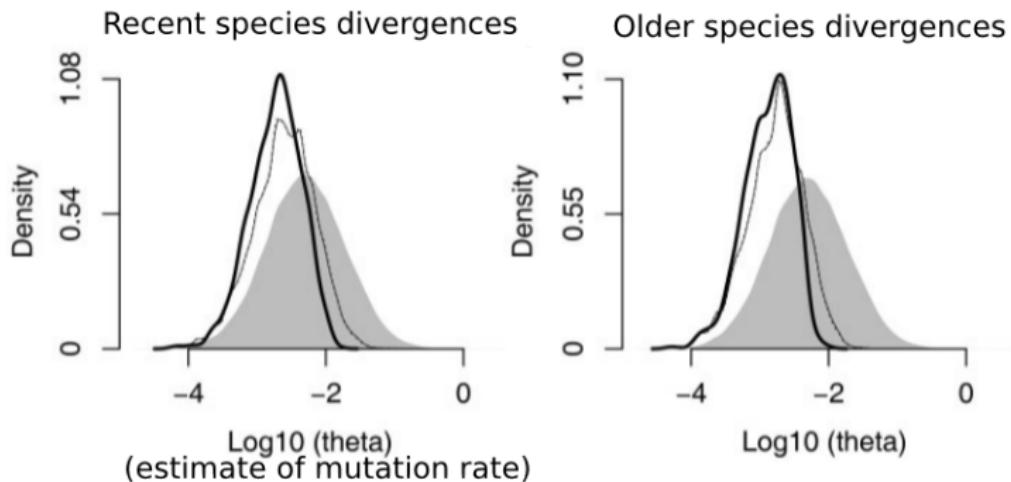
Figure 4: Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support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



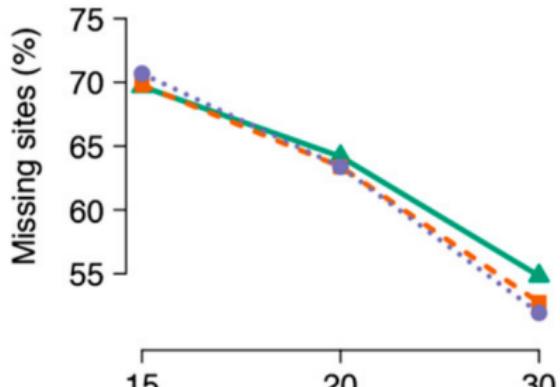
Tradeoffs:

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

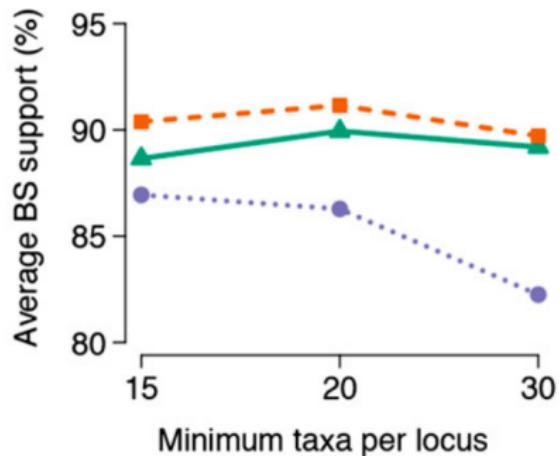


▲ $W_{\text{clust}} = 0.80$

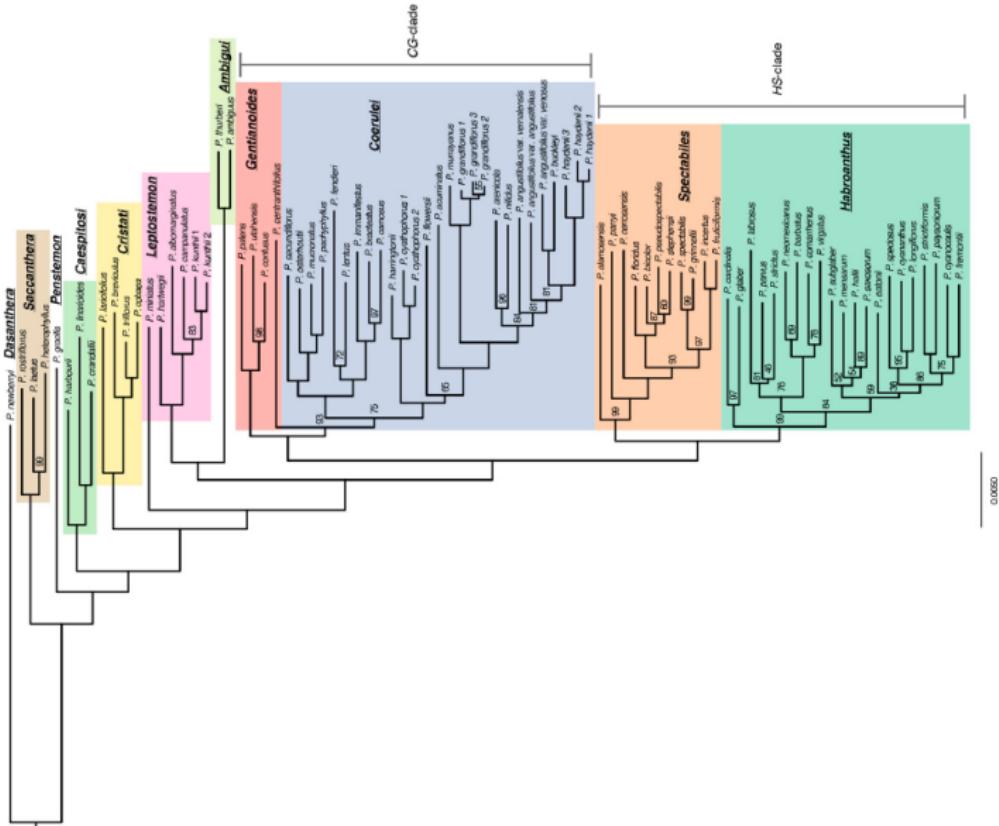
■ $W_{\text{clust}} = 0.90$

● $W_{\text{clust}} = 0.95$

(Wessinger et al., 2016)



Missing data is phylogenetically biased

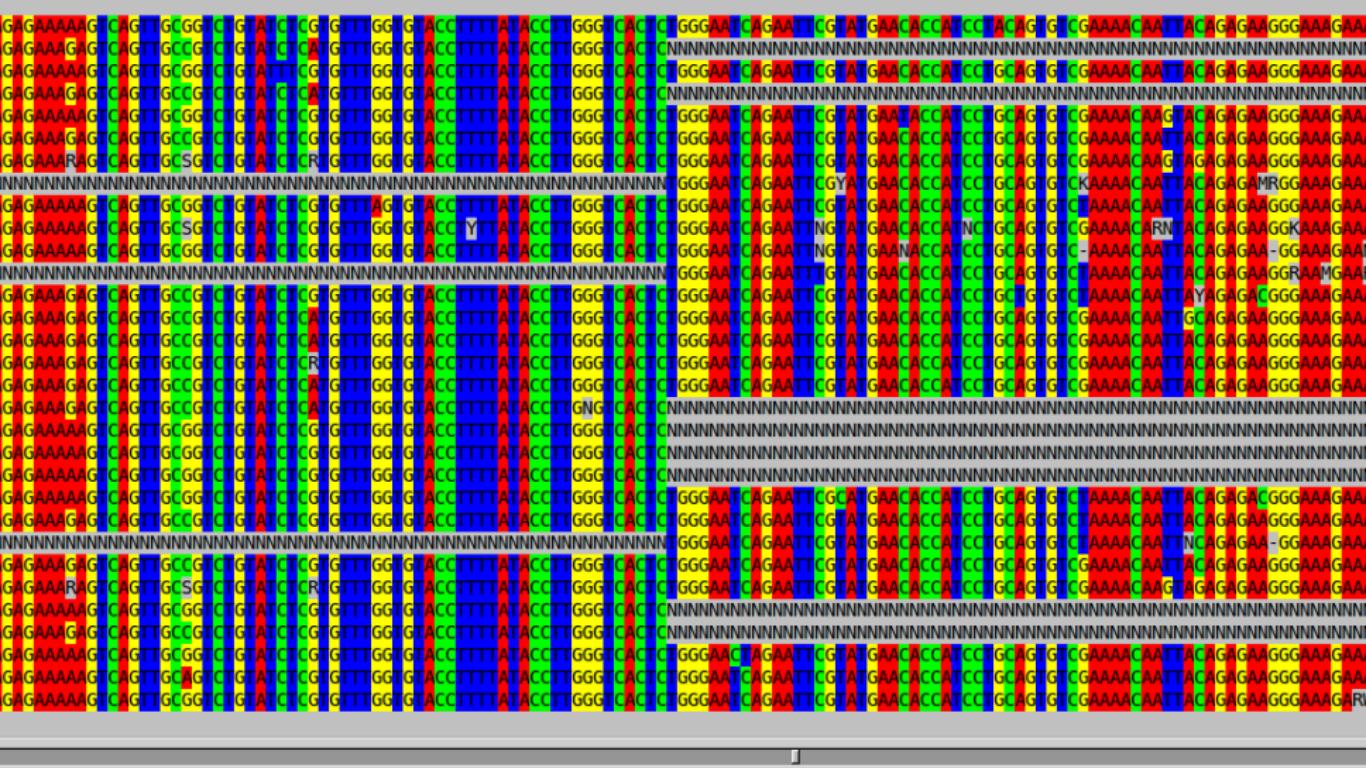


Across full dataset, many loci are only found in one of the major clades

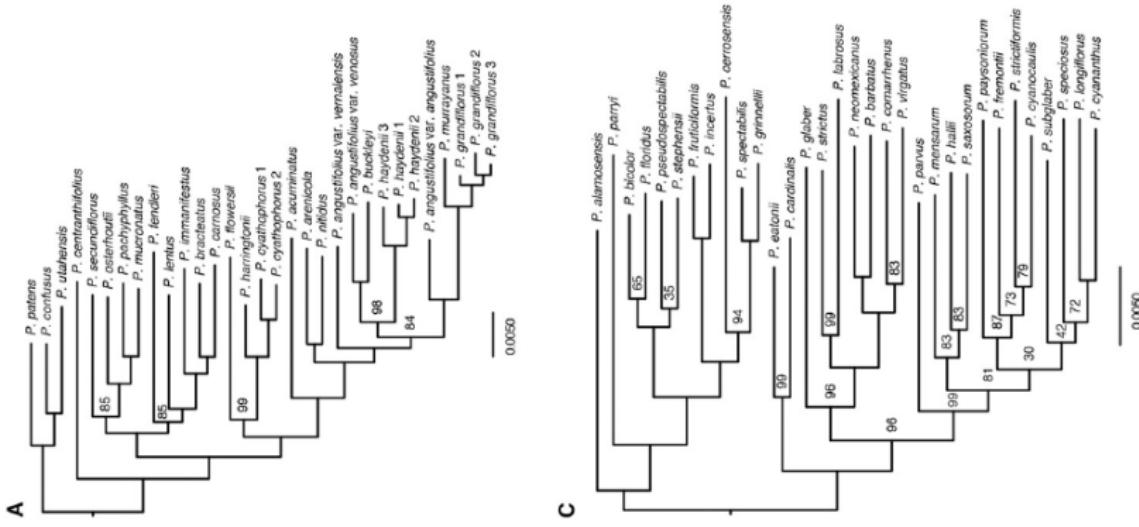
ees Search: Goto: Help

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

Trees | Search: | Goto: | Help |

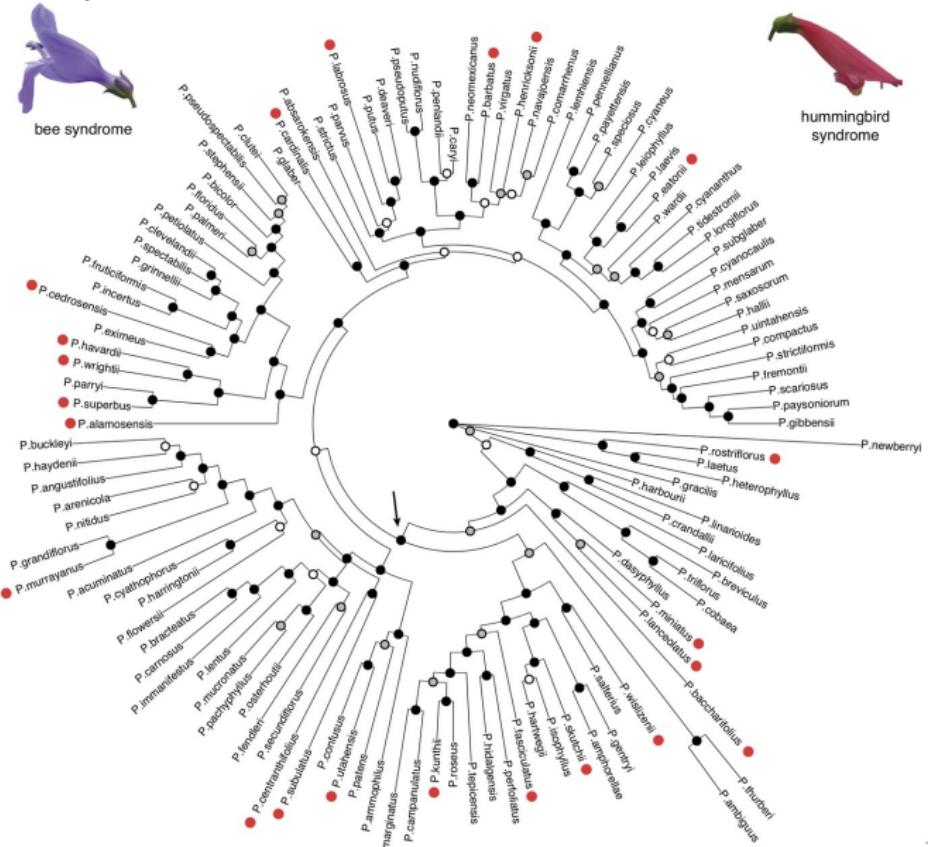


Build (and report!) multiple trees using different filtering parameters



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

Transitions to hummingbird pollination have occurred many times from a bee-pollinated ancestor.



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

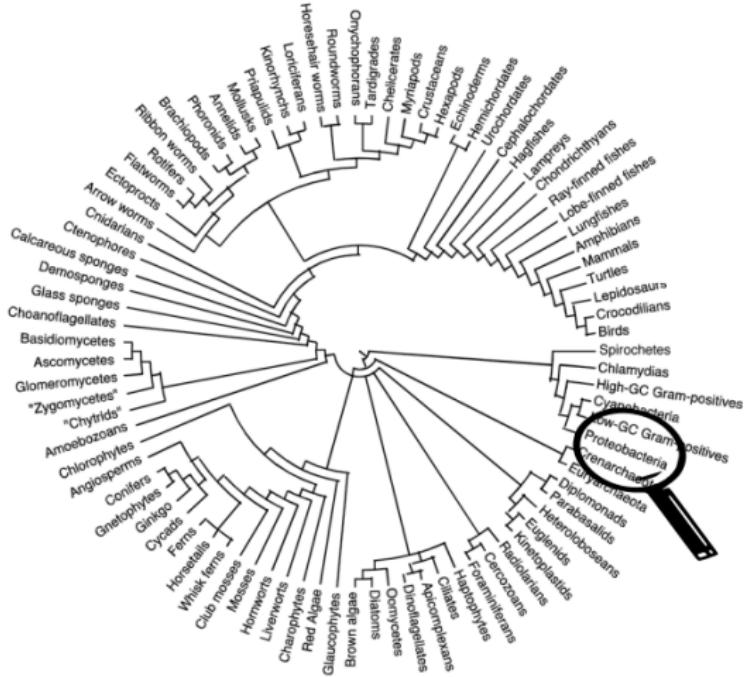
Conclusions:

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



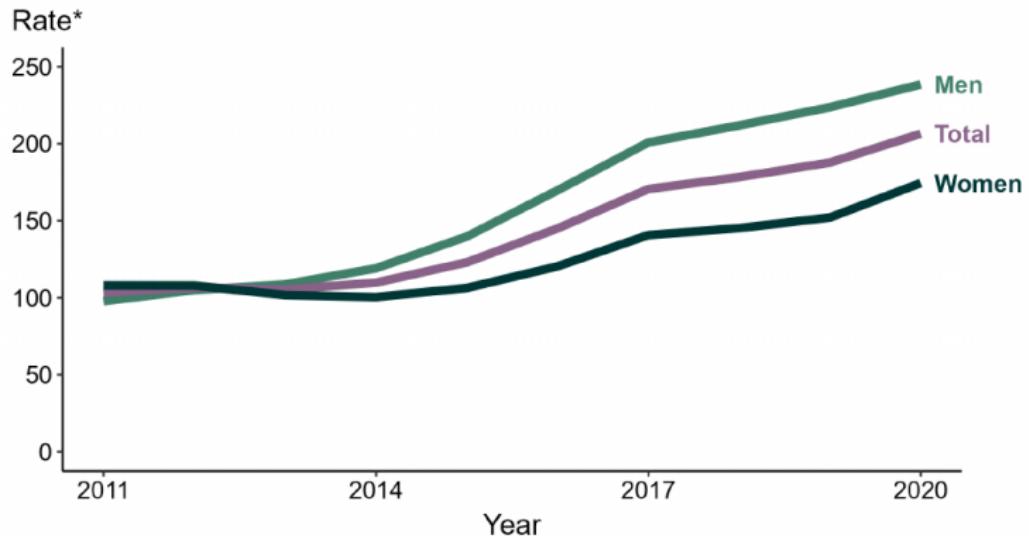
Collaboration with
Jack Cartee , Jeanine Abrams-McLean , and Jasper Toscani Field
(former PhD student, UC Merced)

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



Gonorrhea rates over time by sex

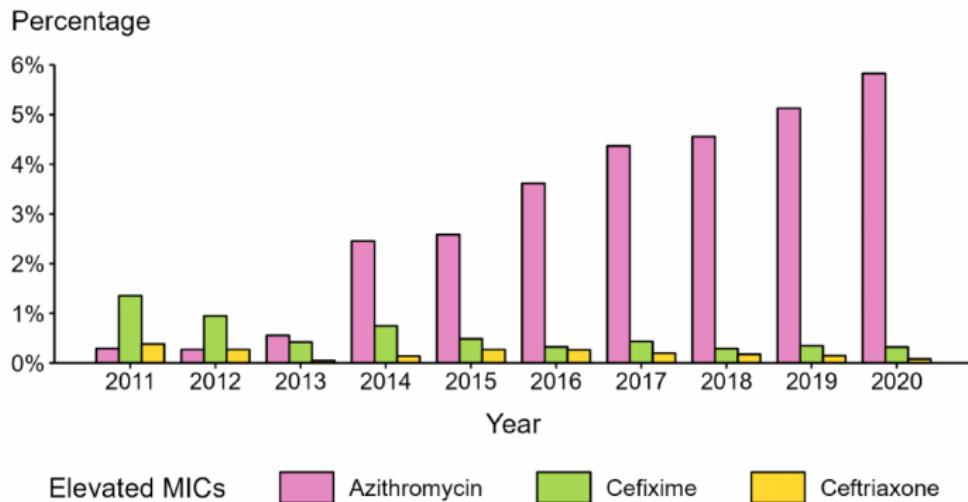


<https://www.cdc.gov/std/statistics/2020/figures/GC-2.htm>

Recent increase in rates of gonorrhea infections

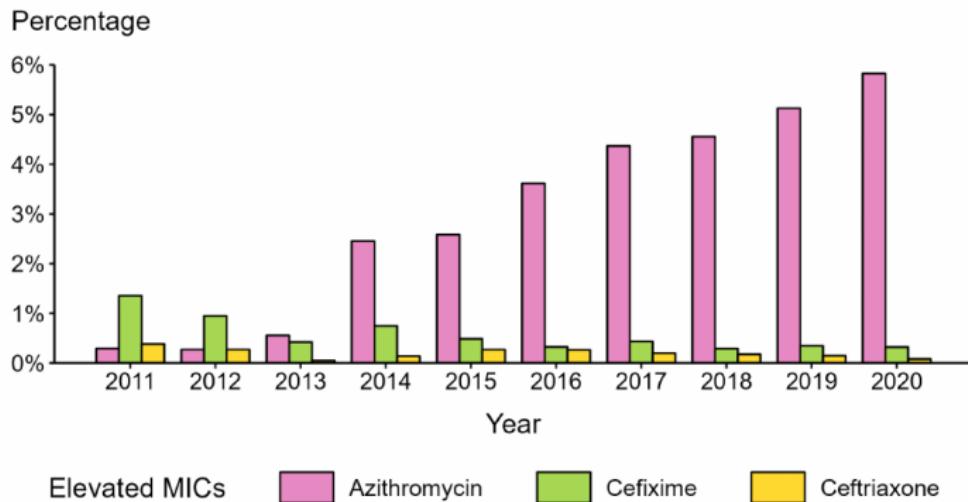
Neisseria gonorrhoeae has progressively developed resistance to each single dose antibiotic.

Percentage of isolates with antibiotic resistance



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

Percentage of isolates with antibiotic resistance

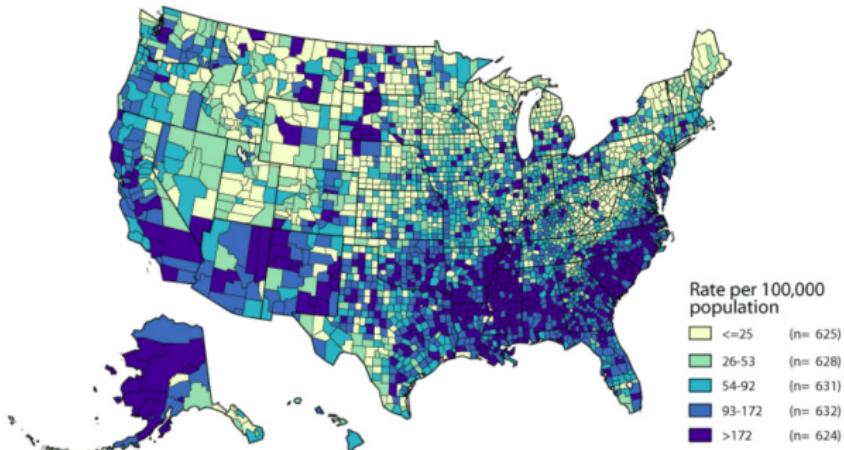


Only remaining recommended treatment option is dual therapy with a ceftriaxone plus azithromycin

"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.

Gonorrhea – Rates of Reported Cases by County, United States, 2017



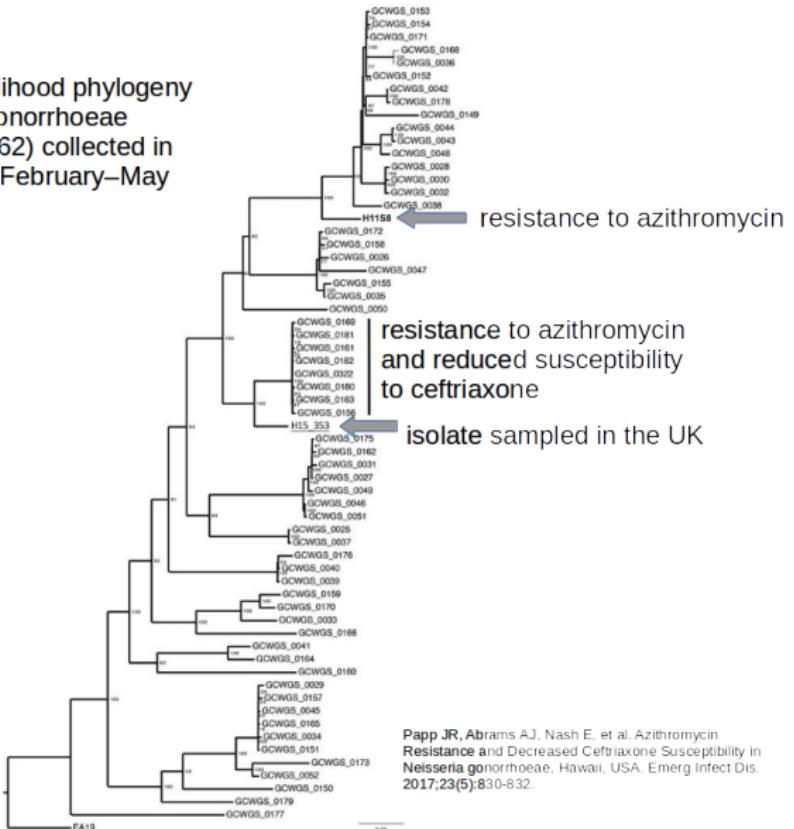
<https://www.cdc.gov/std/stats17/fignatpro.htm#gon>

Approach:

Whole genomic sequencing of *Neisseria gonorrhoea* 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



Challenges:

Thousands of samples; new isolates sequenced every day

Speed from sampling → 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)

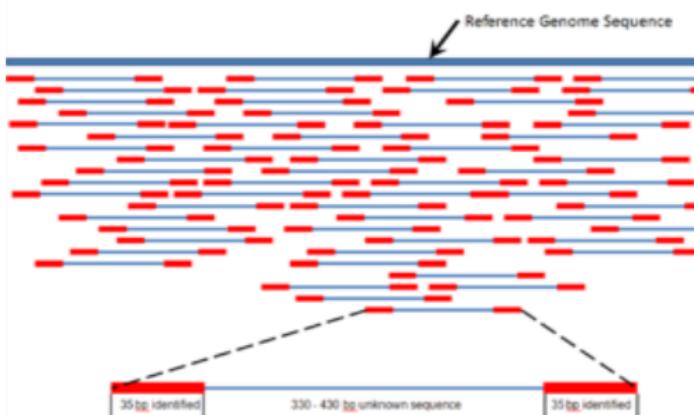


Not currently implemented in phylogenetic likelihood models

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.



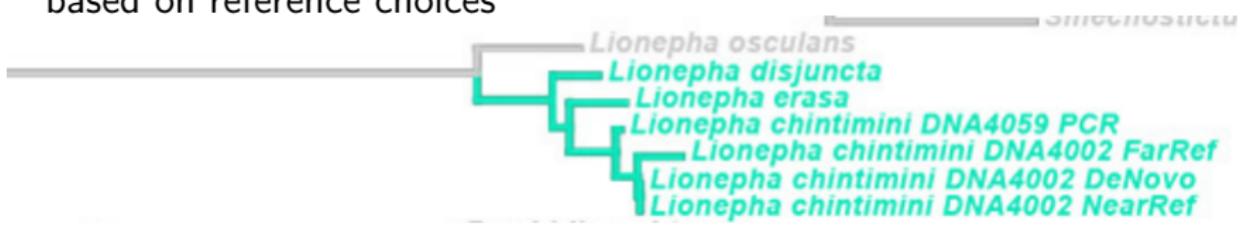
BUT: Reference choice can affect evolutionary inference

BUT: Reference choice can affect evolutionary inference

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

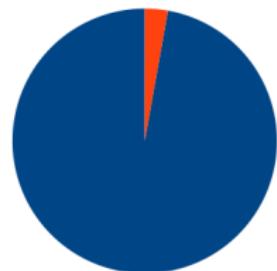
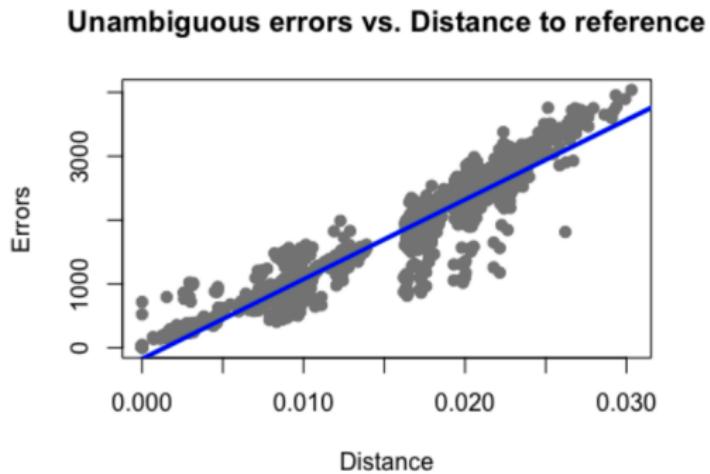
BUT: Reference choice can affect evolutionary inference

- In humans, in highly polymorphic regions variant calling is biased toward the reference base (Brandt et al., 2015)
- In fragmented DNA samples from beetles, branch lengths change based on reference choices



(Kanda et al., 2015)

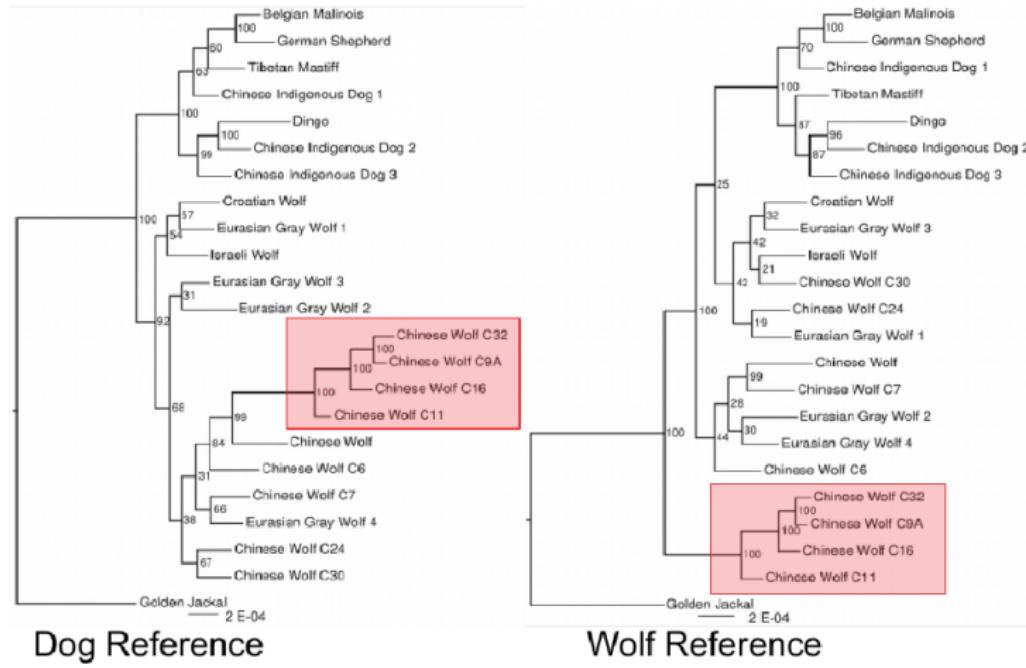
In experimental re-analysis of UCE data, error rate is correlated with distance to reference genome, and errors are strongly biased to the reference base



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

(Toscani-Field and McTavish, work in progress)

Reference choice can affect topology



Gopalakrishnan et al. (2017)

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

Hands on exercise! <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

Big picture

All data sets are biased, genome scale data is no exception

Careful project planning helps

Interrogate potential biases in data sets AND methods

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 can 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?

- 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.
- Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, 1(3):171–182. Number: 3.

- 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.
- 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.
- Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17. Number: 1.
- 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.
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1:16048.

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

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

- Wessinger, C. A., Rausher, M. D., and Hileman, L. C. (2019). Adaptation to hummingbird pollination is associated with reduced diversification in *Penstemon*. *Evolution Letters*, 3(5):521–533.
- Zhang, C., Nielsen, R., and Mirarab, S. (2025). CASTER: Direct species tree inference from whole-genome alignments. *Science*, 387(6737):eadk9688. Publisher: American Association for the Advancement of Science.
- 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.