

# Genomics for evolutionary inference

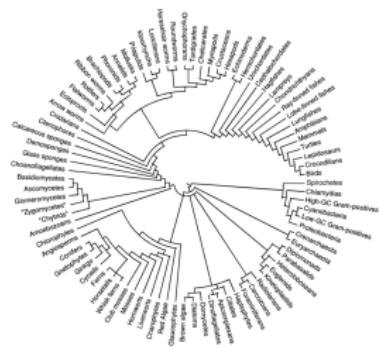
Emily Jane McTavish

Life and Environmental Sciences  
University of California, Merced  
[ejmctavish@ucmerced.edu](mailto:ejmctavish@ucmerced.edu), [twitter:snacktavish](#)

## How do you get from



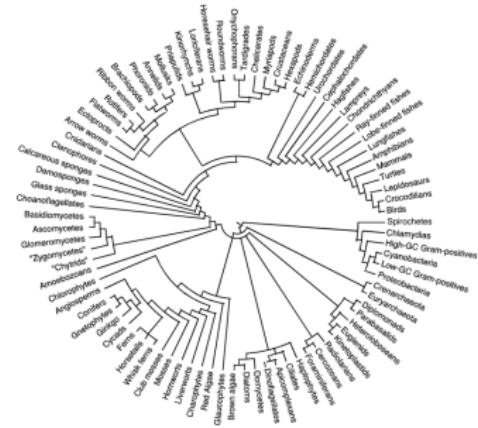
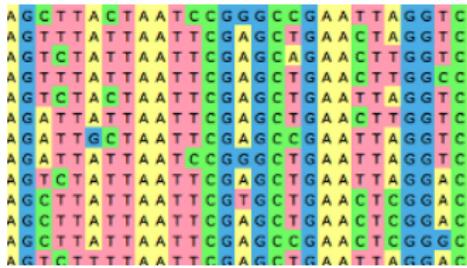
to



?

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

```

I'm going to talk about going from



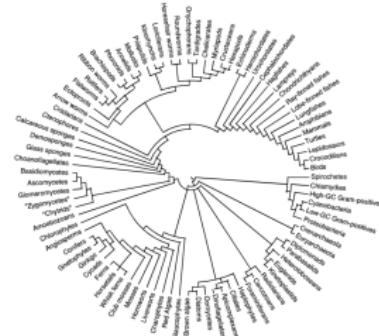
to



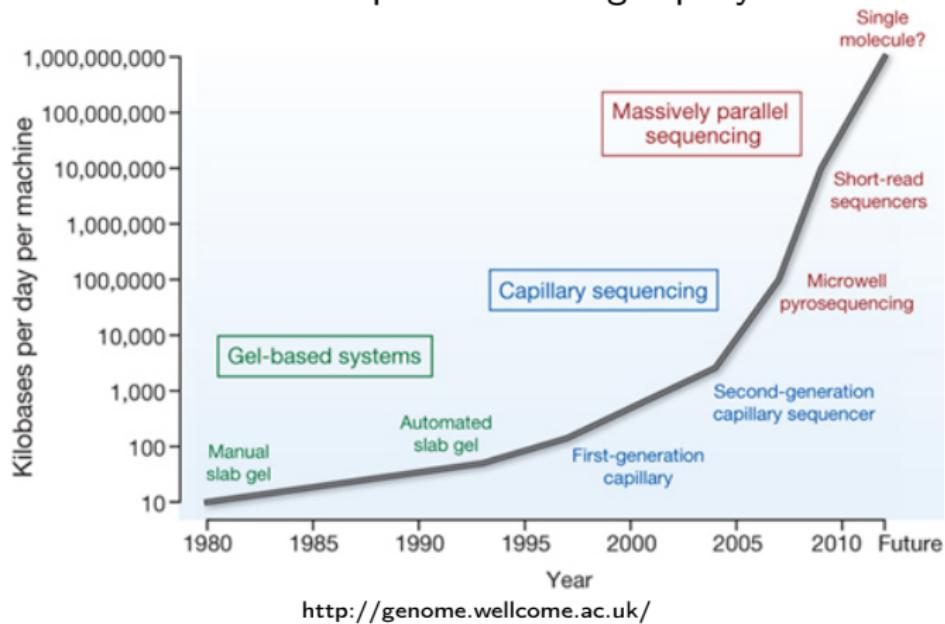
to

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 show some variation across the grid.

and how those choices can affect



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



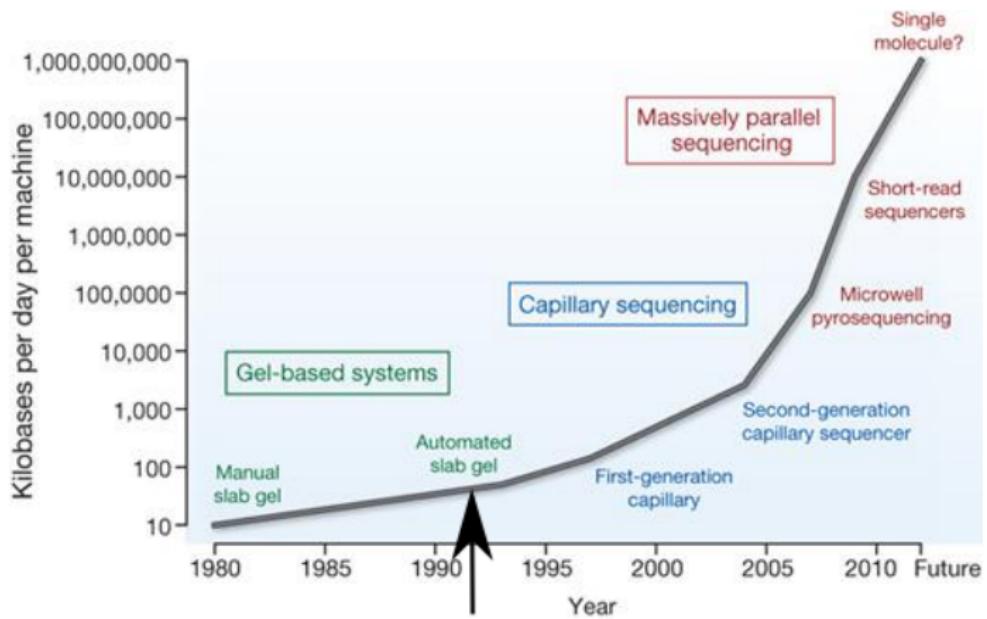
*“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”*

– Jeff Thorne (Evolutionary biologist)

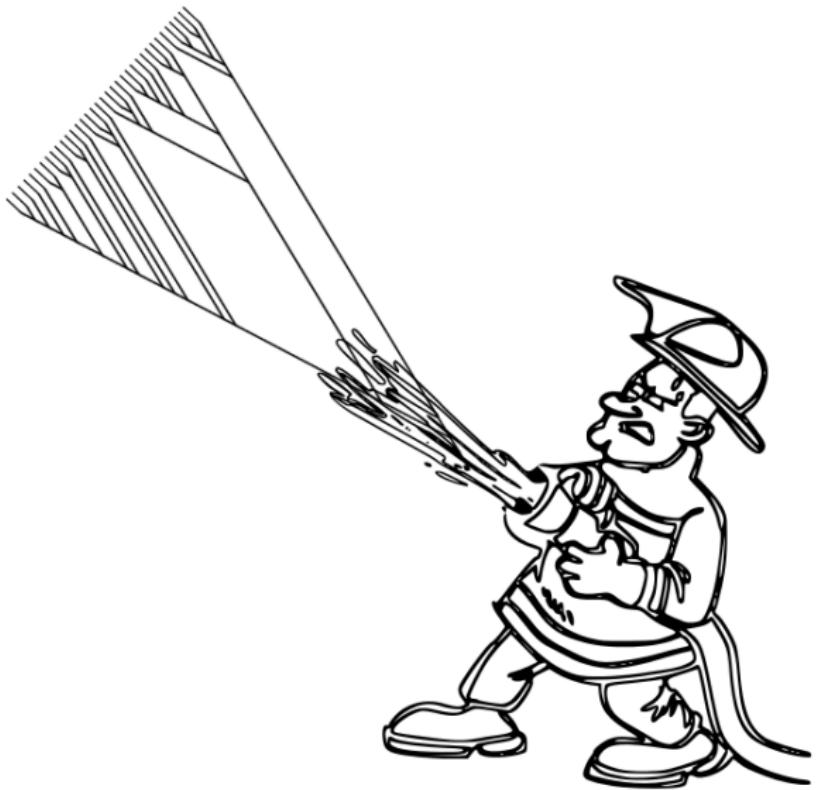
*“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”*

– 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?

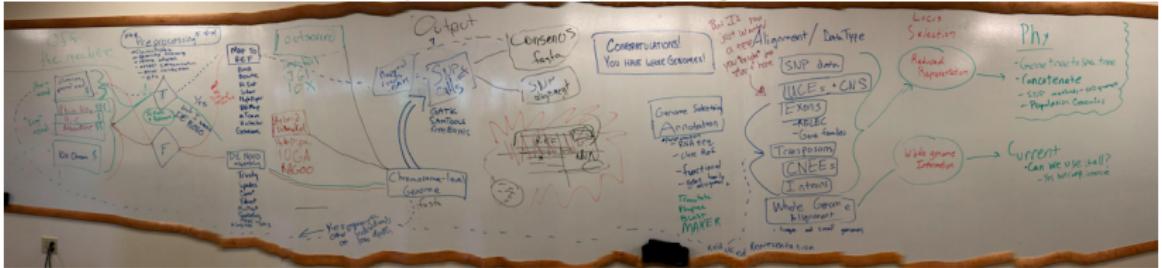
How should we process and analyze our data?

What biases may be affecting our inferences?

## General approach

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

There are a lot of possible paths you can take!



Flowchart capturing genomic analysis pipelines used by participants at Trees in the Desert workshop, May 2019

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!

## 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, currently 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!



to

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

## Genomic sequencing

You have all the data! 

You have to deal with all of the data. 

## *De novo* assembly

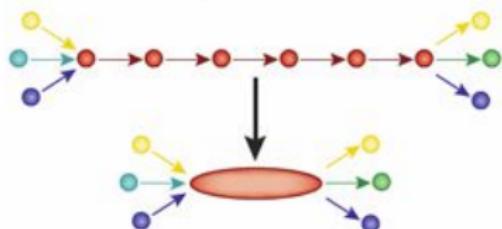
1. Fragment DNA and sequence



2. Find overlaps between reads

...AGCCTAGACCTACA**GGATGCGCGACACGT**  
**GGATGCGCGACACGT**CGCATATCCGGT...

3. Assemble overlaps into contigs

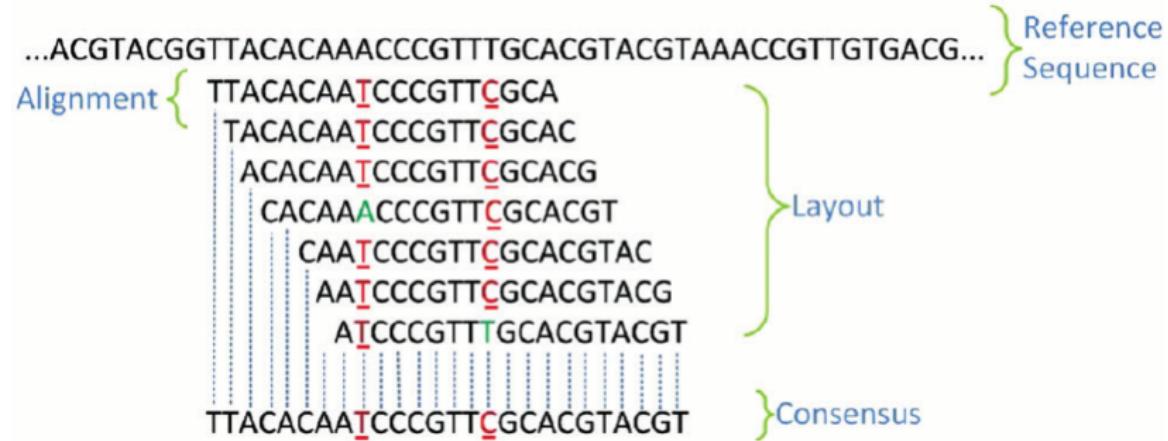


4. Assemble contigs into scaffolds



(Baker, 2012)

## Mapping to a reference genome

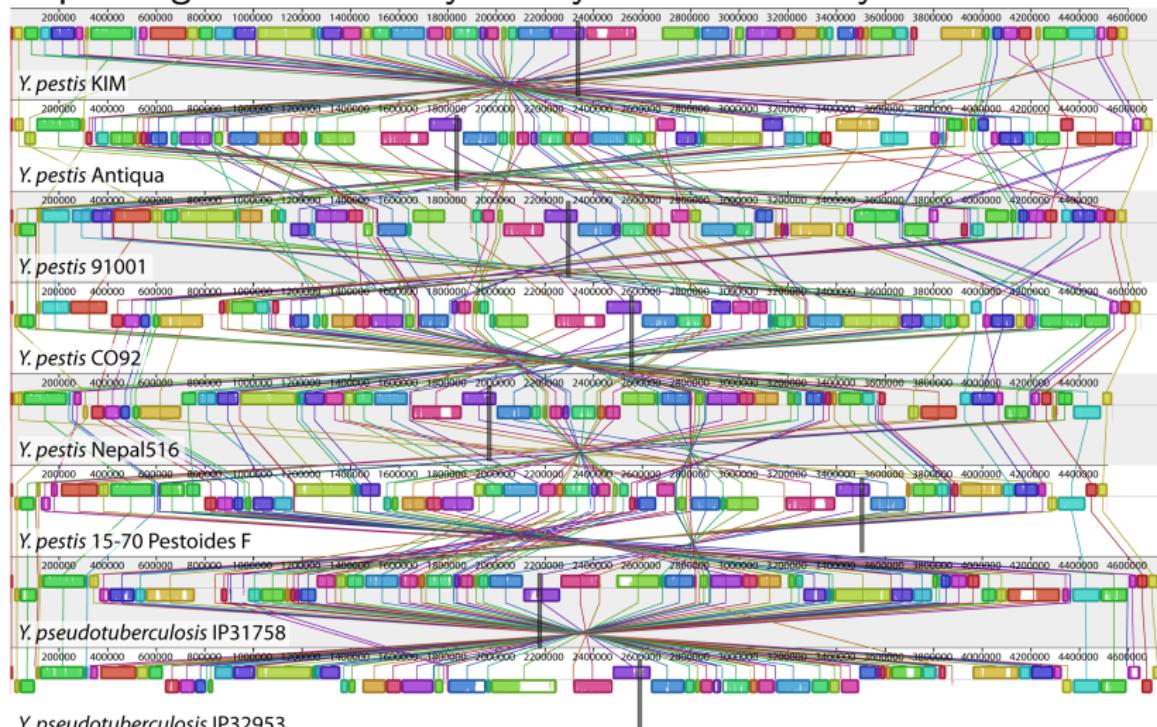


To make evolutionary statements, you need to align genomic regions across taxa.

Depending on evolutionary history this can be easy or hard!

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

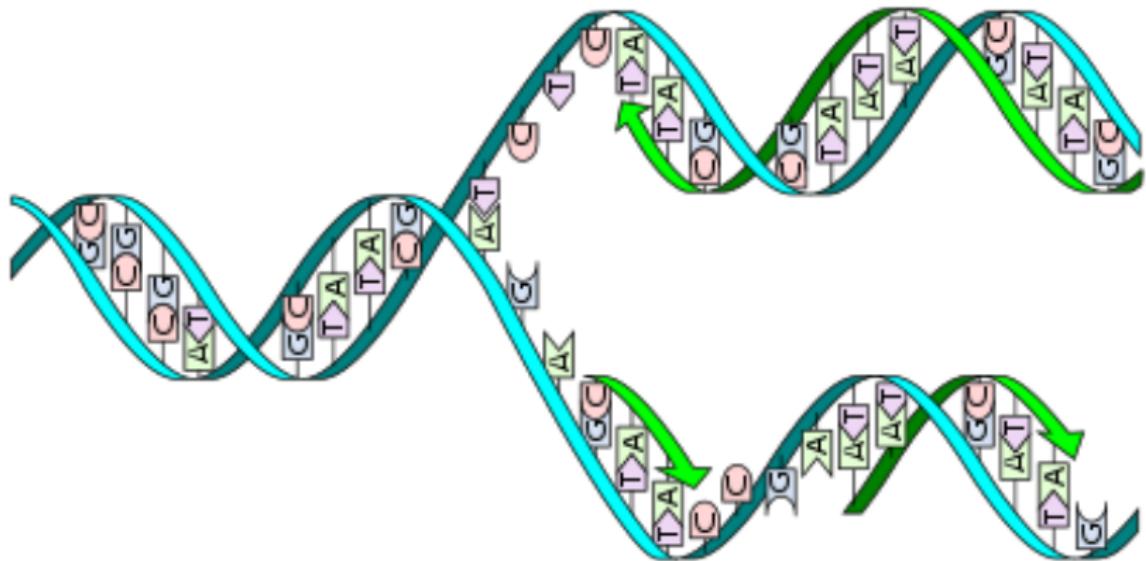
Depending on evolutionary history this can be easy or hard!



*Y. pseudotuberculosis* IP32953

(Darling et al., 2008)

An alignment is a statement of shared ancestry



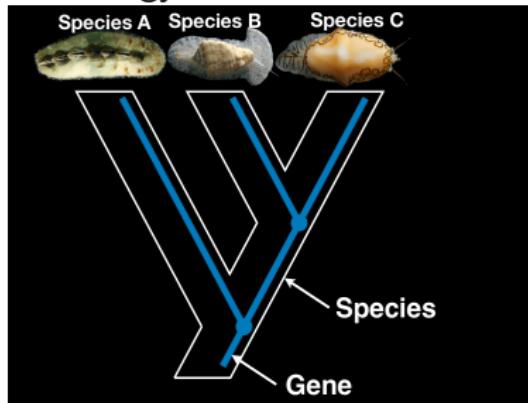
## **Gene tree** (Locus tree)

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

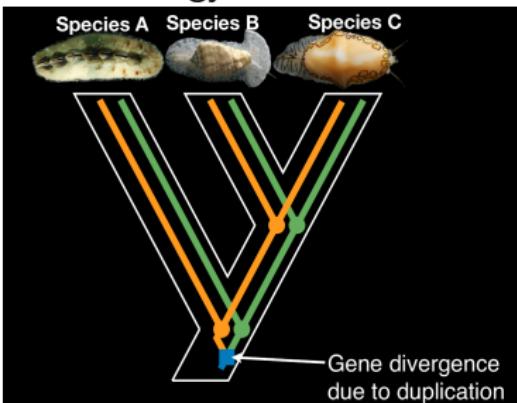
Enrichment methods focus our sequencing efforts on these regions

## Gene duplication and loss

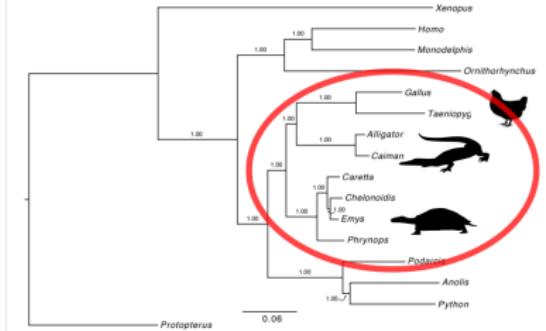
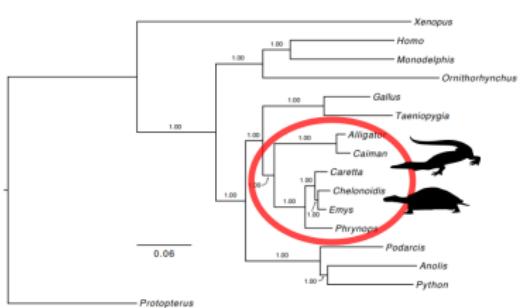
### Orthology



### 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)

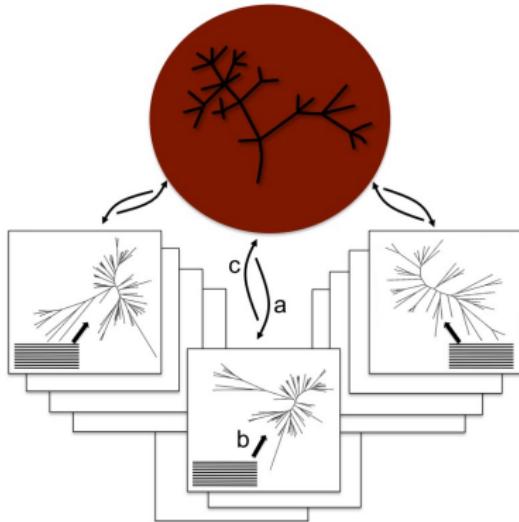
**Challenge:** The true (unknown) phylogenetic history is needed to assess orthology vs paralogy

"We used to be so blissful, back in the day. We just had data for one gene."

Paul Lewis

Integrated approaches can jointly estimate gene trees and species trees

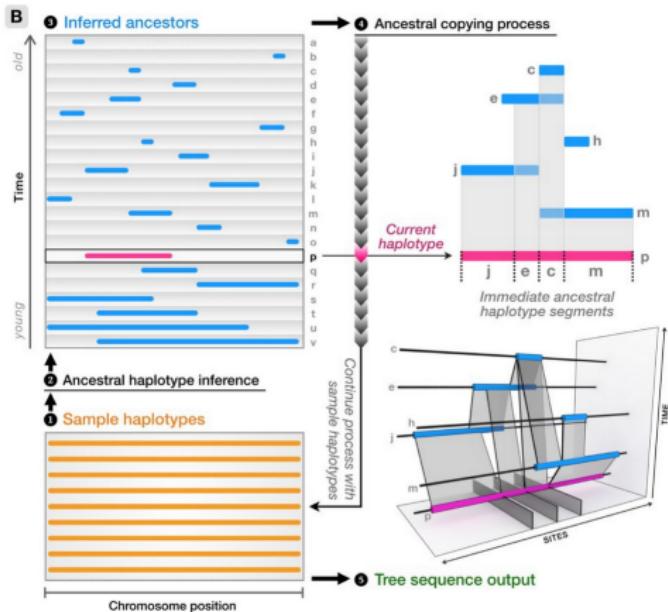
$$L(T, S, N | A) = \prod_{G_i \in \mathcal{G}} L(G_i)$$



(Boussau et al., 2013) But are computationally expensive.

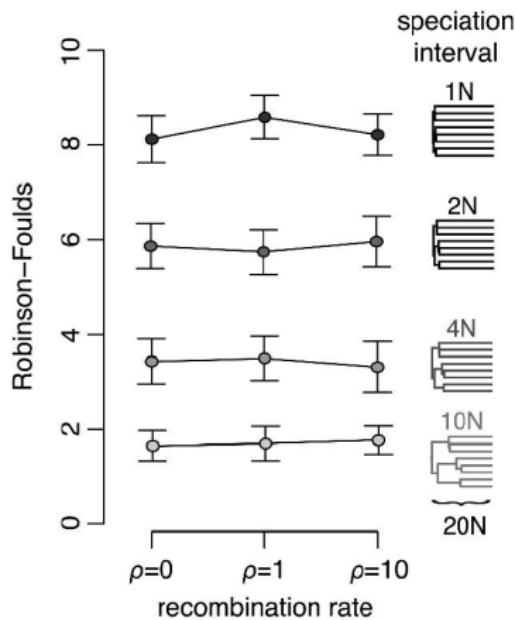
As we densely sample genomes and taxa, the size of a 'locus' or 'un-recombined region' will get smaller.

# Jointly estimate recombination and ancestries along genomes



Currently only applied within humans! (Kelleher et al., 2018)

## Species tree methods are robust to intra-locus recombination (based on analyses of simulated data)



Robinson Foulds (RF)  
distance: the symmetric  
difference between trees -  
the number of branches in  
tree 1 and not in tree 2 +  
the number of branches in  
tree 2 and not in tree 1.

(Lanier and Knowles, 2012)

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!

## How do the choices we make in



to



to

```

ACGCTTACTAACTCGGGCCGAAATTAGGTG
TTTTATTAATTCGAGCTAATTTAGGTC
GTTCTATTAAATTCCGAGCAAACTTGGGT
GTTTTATTAATTCGAGCTAACCTGGGC
TCTCATTTAAATTCCGAGCTAAATTTAGGT
GATTTATTAATTCGAGCTAACCTGGGT
GAAATTGCGAAATTCCGAGCTAAATTTAGGT
GATTTATTAATTCGAGCTAACCTGGGT
TCTCATTTAAATTCCGAGCTAAATTTAGGT
GATTTATTAATTCGAGCTAACCTGGGT
GAAATTGCGAAATTCCGAGCTAAATTTAGGT
GATTTATTAATTCGAGCTAACCTGGGT
TCTCATTTAAATTCCGAGCTAAATTTAGGT
GATTTATTAATTCGAGCTAACCTGGGT
GCTTTATTAATTCGAGCTAACCTGGGT
GCTTTATTAATTCGAGCTAACCTGGGT

```

## How do the choices we make in



to



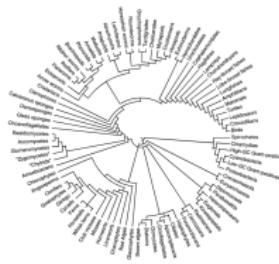
to

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

```



?

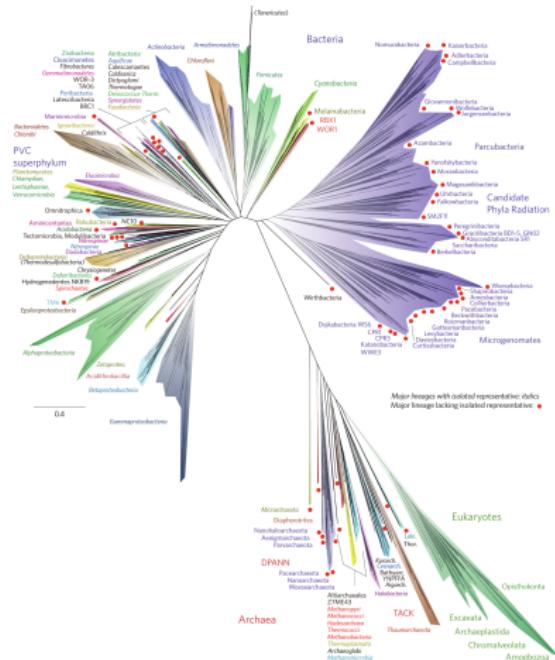
## **Ascertainment bias**

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



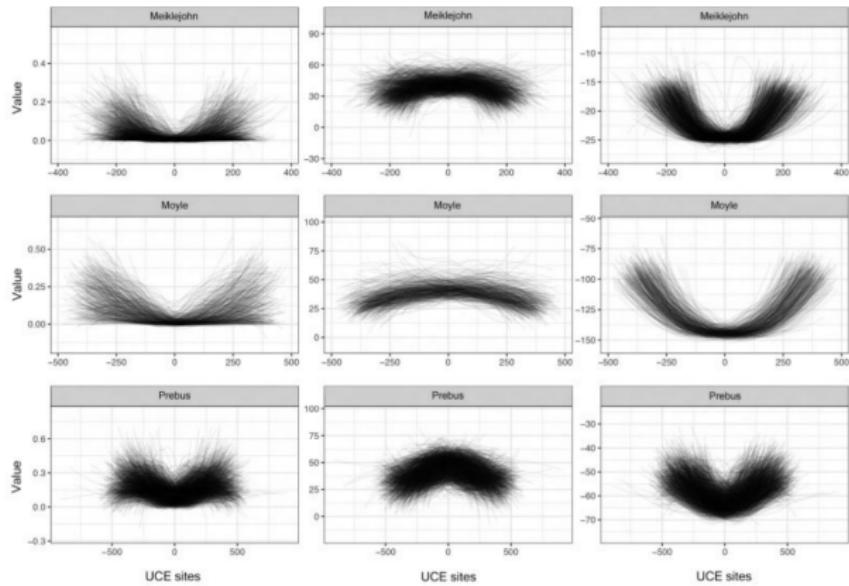
(Hug et al., 2016)

# Ascertainment bias in genomic data

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, Multinomial likelihood**



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

**Analyzing only variable sites** (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of 'Anything Can Happen Now'

**Analyzing only variable sites** (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of 'Anything Can Happen Now'  
**'Something Definitely Happened Once!'**

**Analyzing only variable sites** (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of 'Anything Can Happen Now'  
**'Something Definitely Happened Once!'**

This affects our ability to estimate branch lengths using likelihood  
Intuitively, will increase inferred branch lengths  
can also affect tree topology

## 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 by applying Lewis (2001) model for analysis of only variable sites implemented inference software

Based on correction for problem of not counting un-observed restriction sites in (Felsenstein, 1992)

“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

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*?



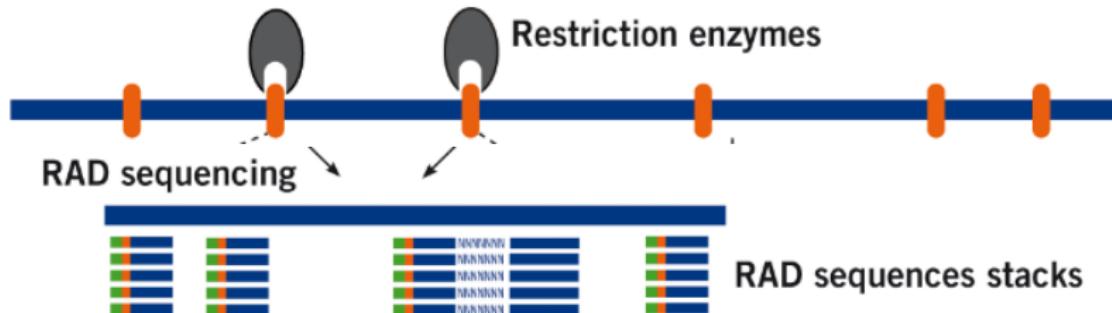
## Data:

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

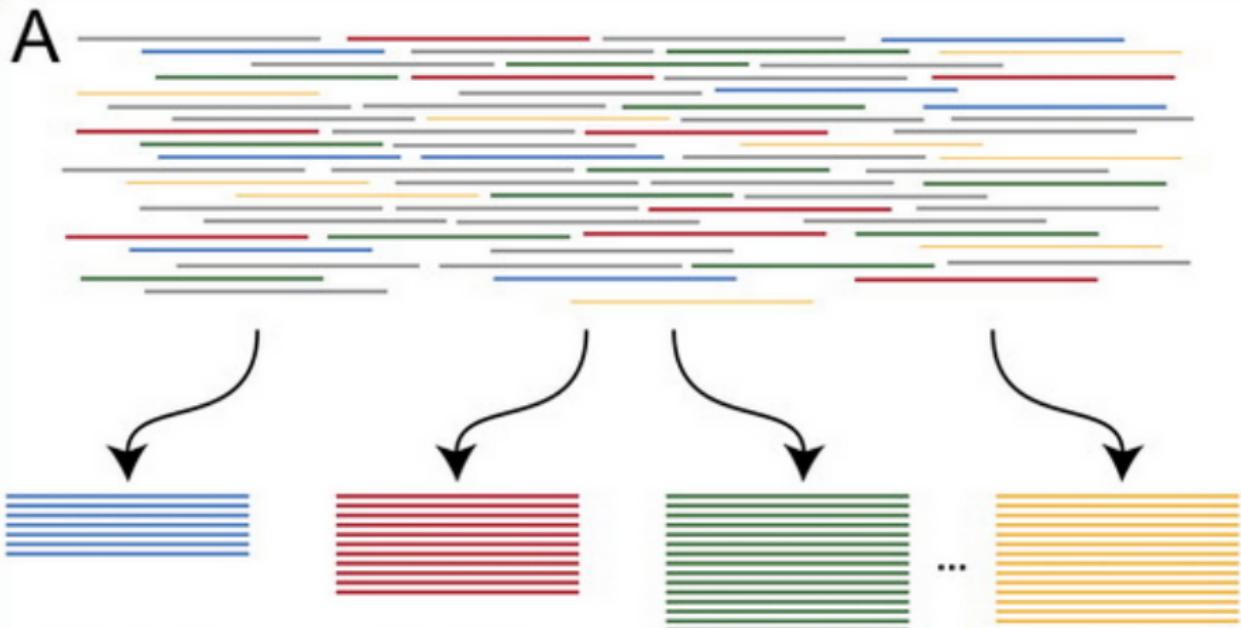


In comparison: Shotgun Sequencing



(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

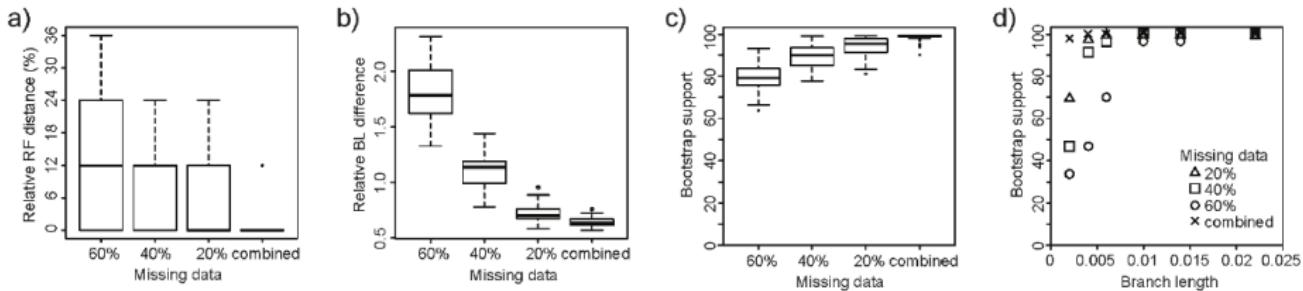


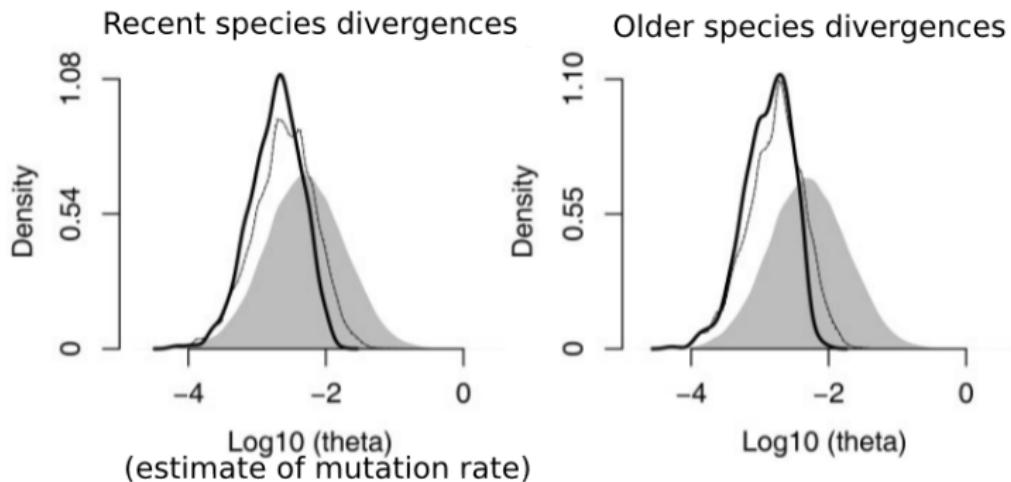
Figure 4: Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support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



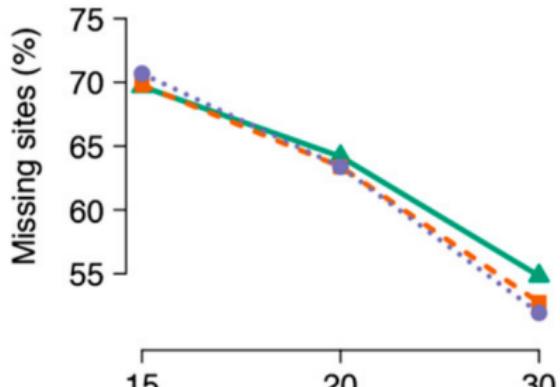
## Tradeoffs:

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

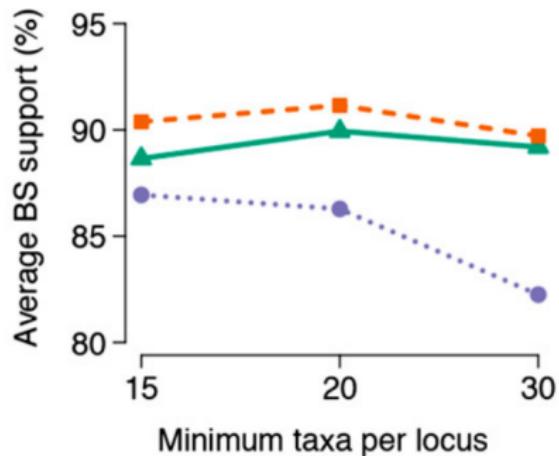


▲  $W_{\text{clust}} = 0.80$

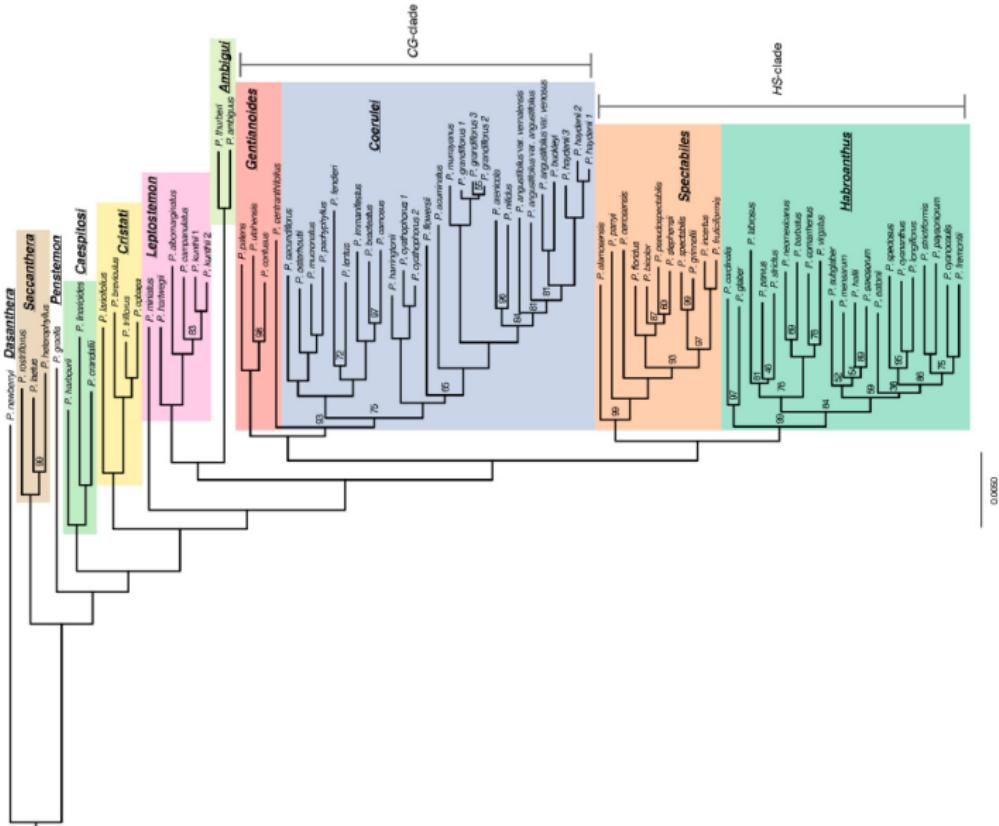
■  $W_{\text{clust}} = 0.90$

●  $W_{\text{clust}} = 0.95$

(Wessinger et al., 2016)



Missing data is phylogenetically biased

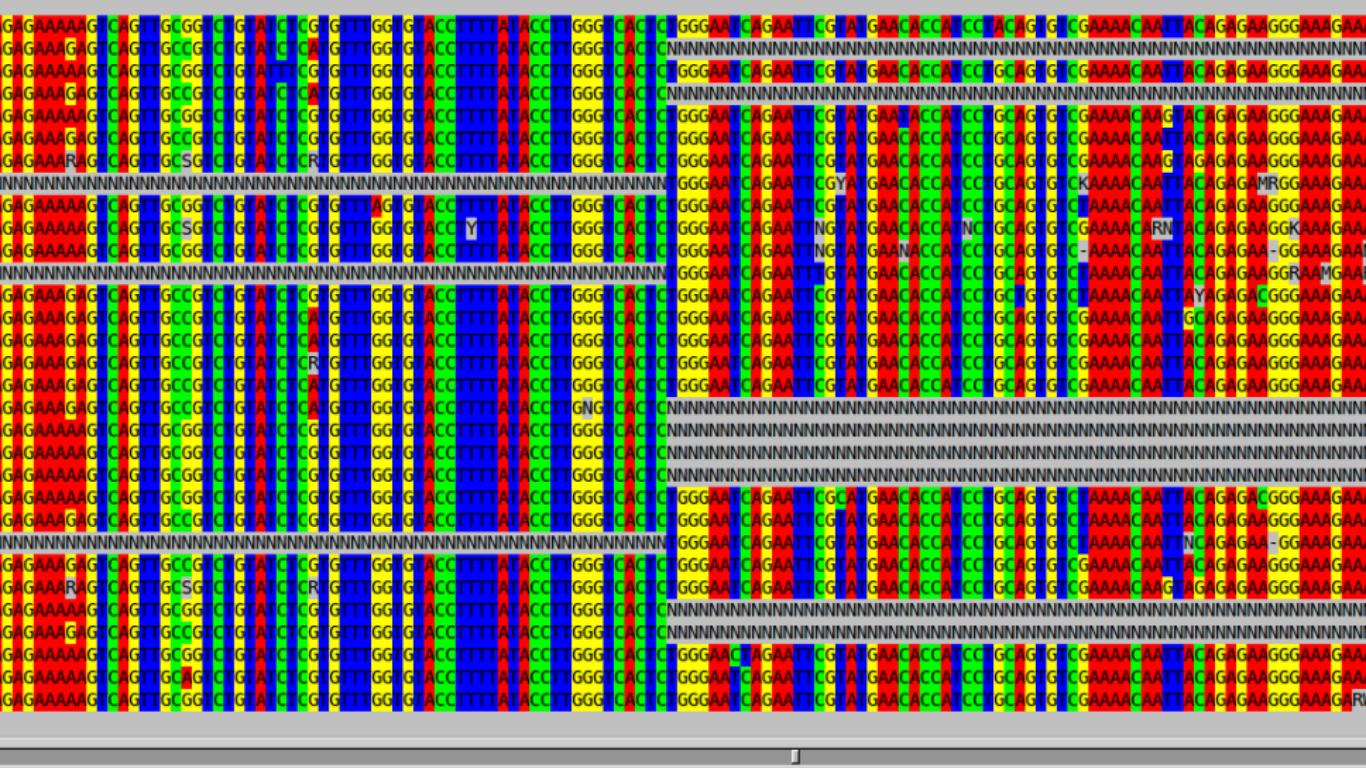


Across full dataset, many loci are only found in one of the major clades

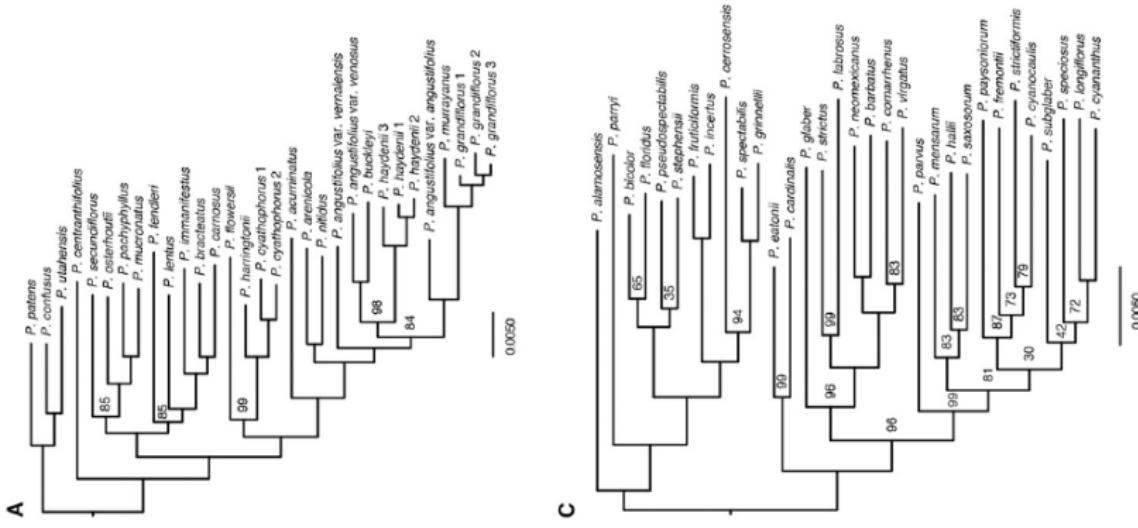
ees Search: Goto: Help

# 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)

## 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

## 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

Questions?

- Baker, M. (2012). *De novo* genome assembly: what every biologist should know. *Nature Methods*, 9:333–337.
- 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.
- 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.
- 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.
- 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.
- 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.
- Darling, A. E., Miklós, I., and Ragan, M. A. (2008). Dynamics of Genome Rearrangement in Bacterial Populations. *PLoS Genetics*, 4(7).
- Escalona, M., Rocha, S., and Posada, D. (2018). NGSphy: phylogenomic simulation of next-generation sequencing data. *Bioinformatics (Oxford, England)*, 34(14):2506–2507.
- Felsenstein, J. (1992). Phylogenies from Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46(1):159–173.

- 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.
- 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.
- Kelleher, J., Wong, Y., Albers, P. K., Wohns, A. W., and McVean, G. (2018). Inferring the ancestry of everyone. *bioRxiv*.

- Kuhner, M. K. and McGill, J. (2014). Correcting for Sequencing Error in Maximum Likelihood Phylogeny Inference. *G3: Genes/Genomes/Genetics*, 4(12):2545–2552.
- Lanier, H. C. and Knowles, L. L. (2012). Is Recombination a Problem for Species-Tree Analyses? *Systematic Biology*, 61(4):691–701.
- 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.
- Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925.

- 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. (2004). Population genetic analysis of ascertained SNP data. *Human genomics*, 1(3):218–224.
- 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.
- 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.
- Tagliacollo, V. A. and Lanfear, R. (2018). Estimating Improved Partitioning Schemes for Ultraconserved Elements. *Molecular Biology and Evolution*, 35(7):1798–1811.

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.