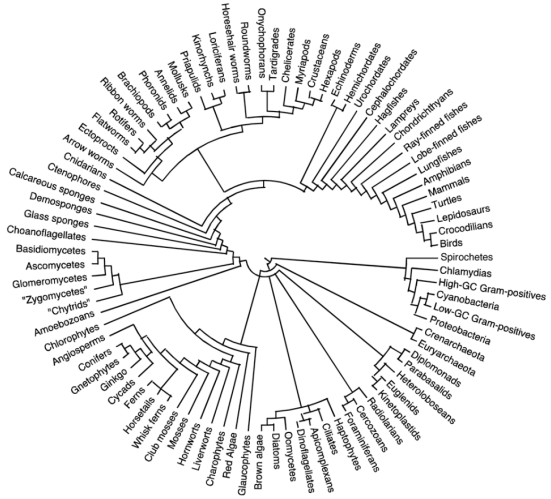
Data set assembly 1

Emily Jane McTavish

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

How do you get from

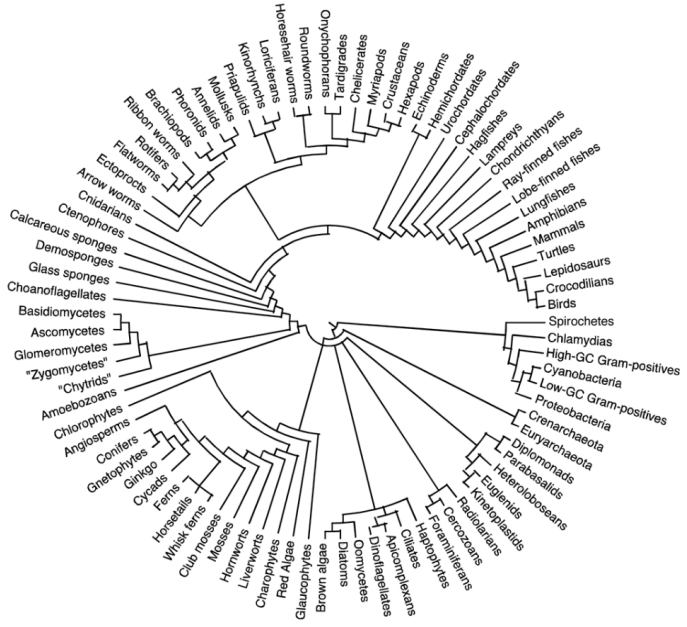
to



?

You've seen (and will see) a lot about how to get from

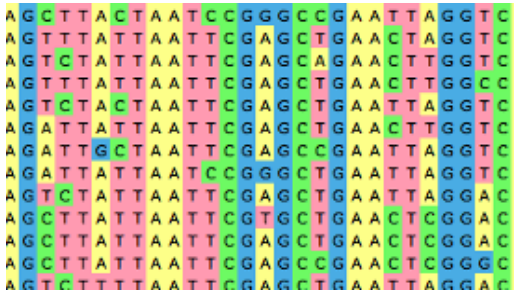
to



I'm going to talk about going from

to

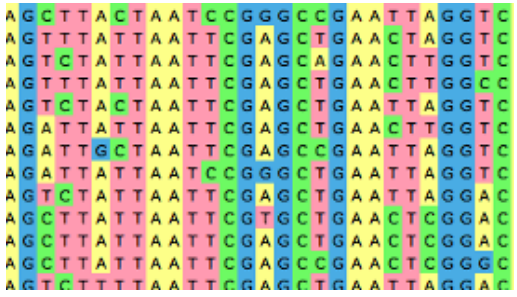
to



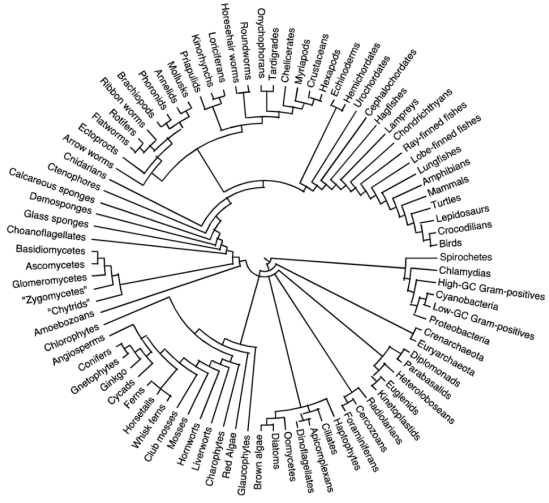
I'm going to talk about going from

to

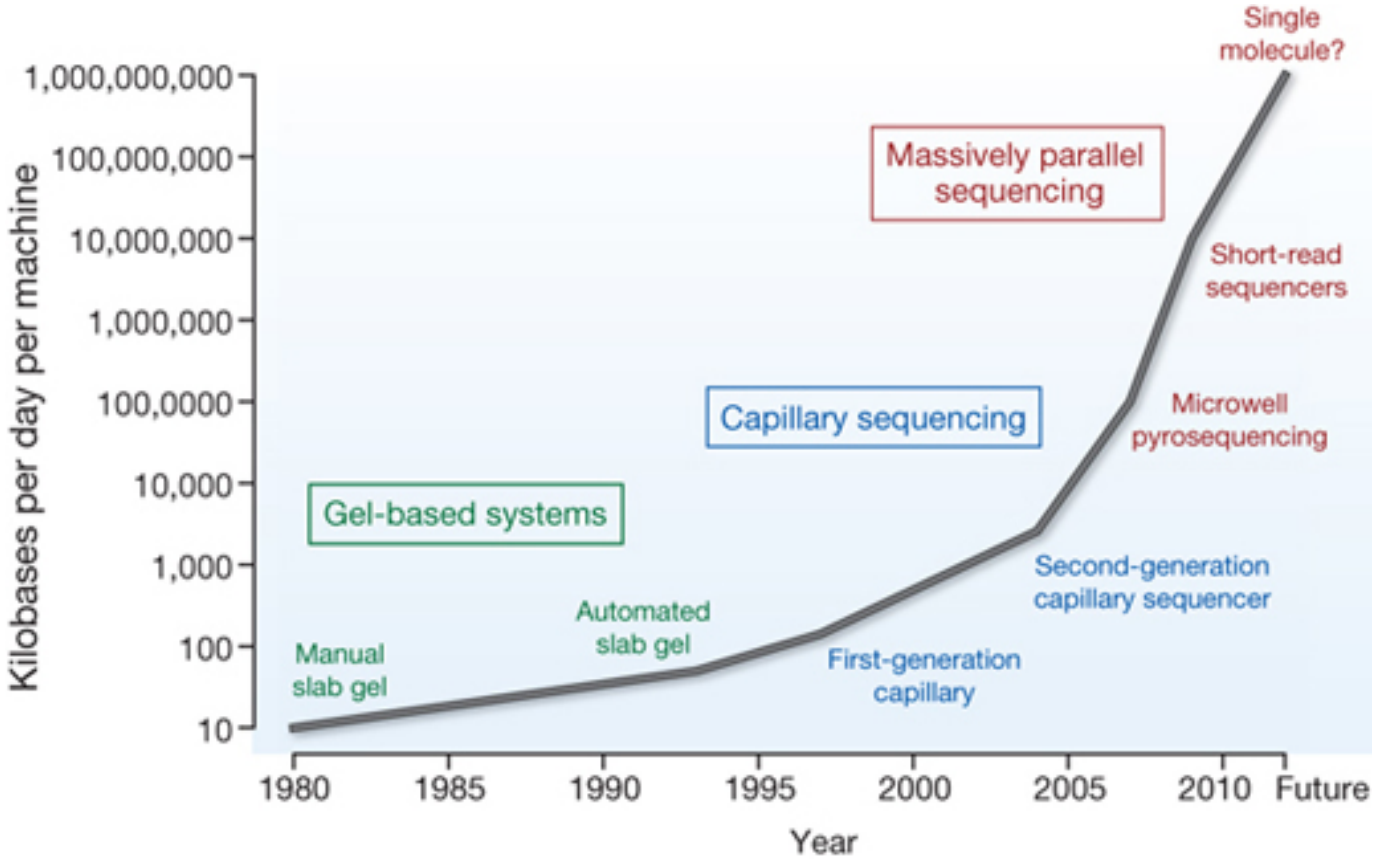
to



and how those choices can affect



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



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

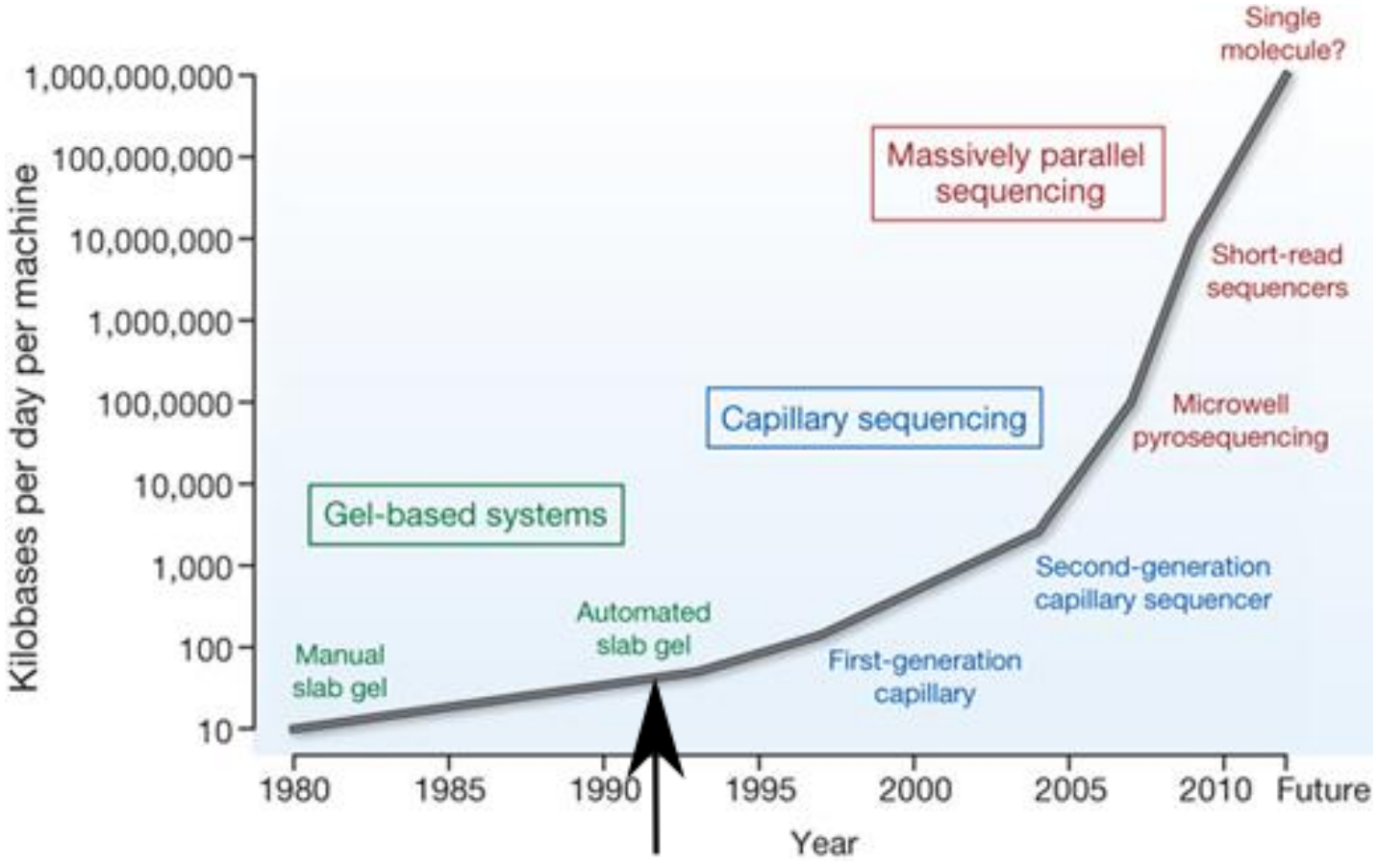
"With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data"

– Jeff Thorne

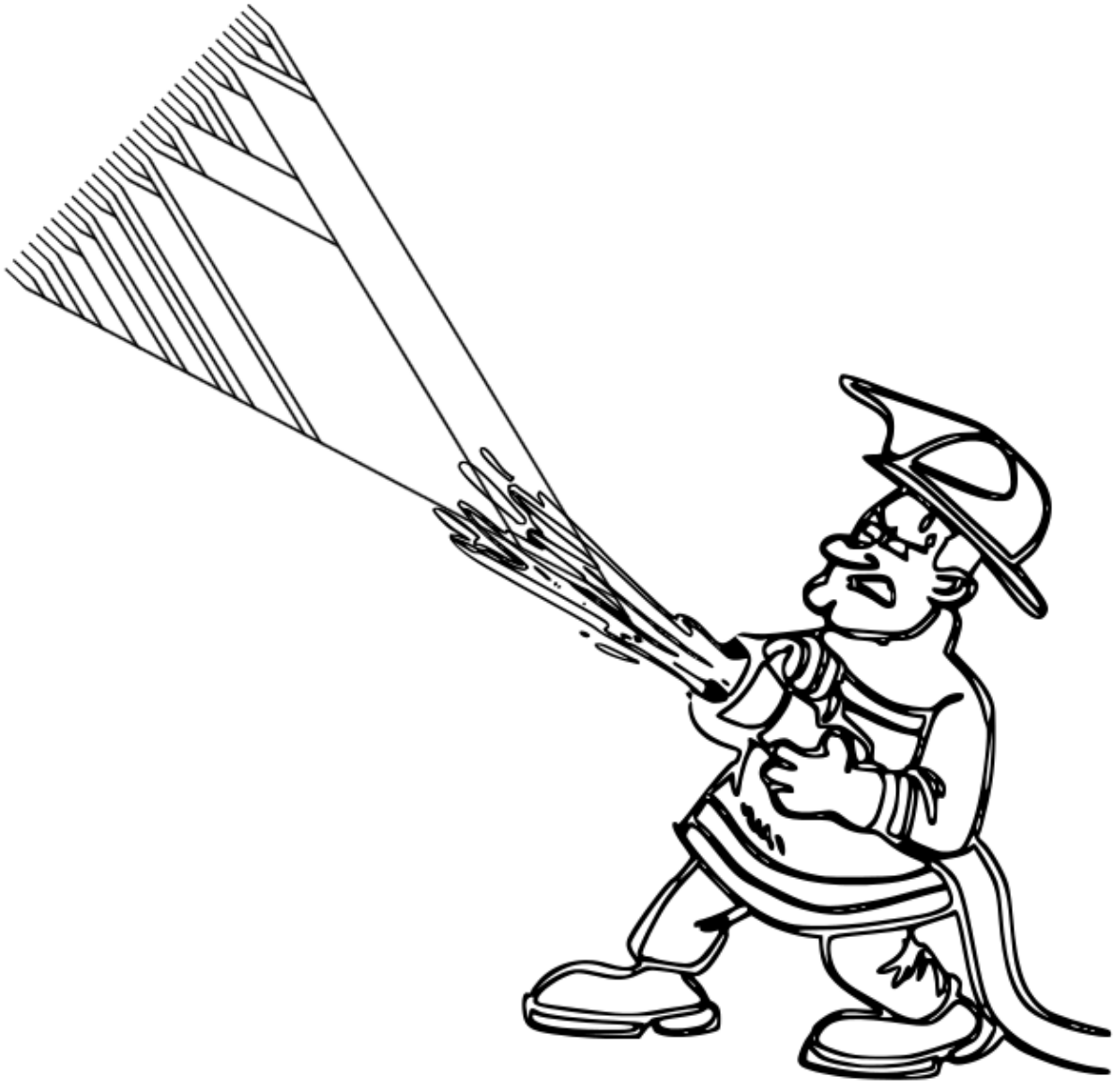
"With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data"

– Jeff Thorne

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?

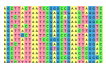
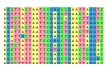
Biological questions

- 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!

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

If sequencing short reads, you need to put the pieces back together!

to

Genomic sequencing

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

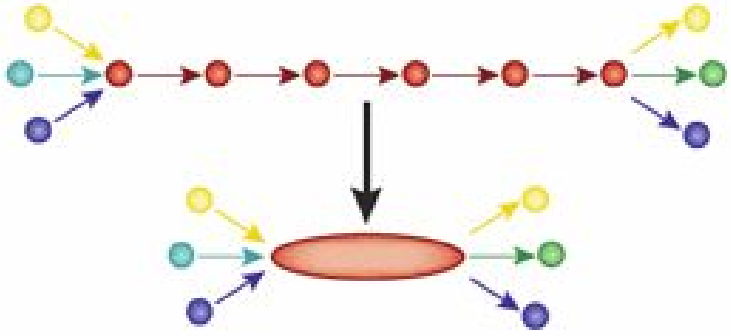
De novo assembly

1. Fragment DNA and sequence

2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

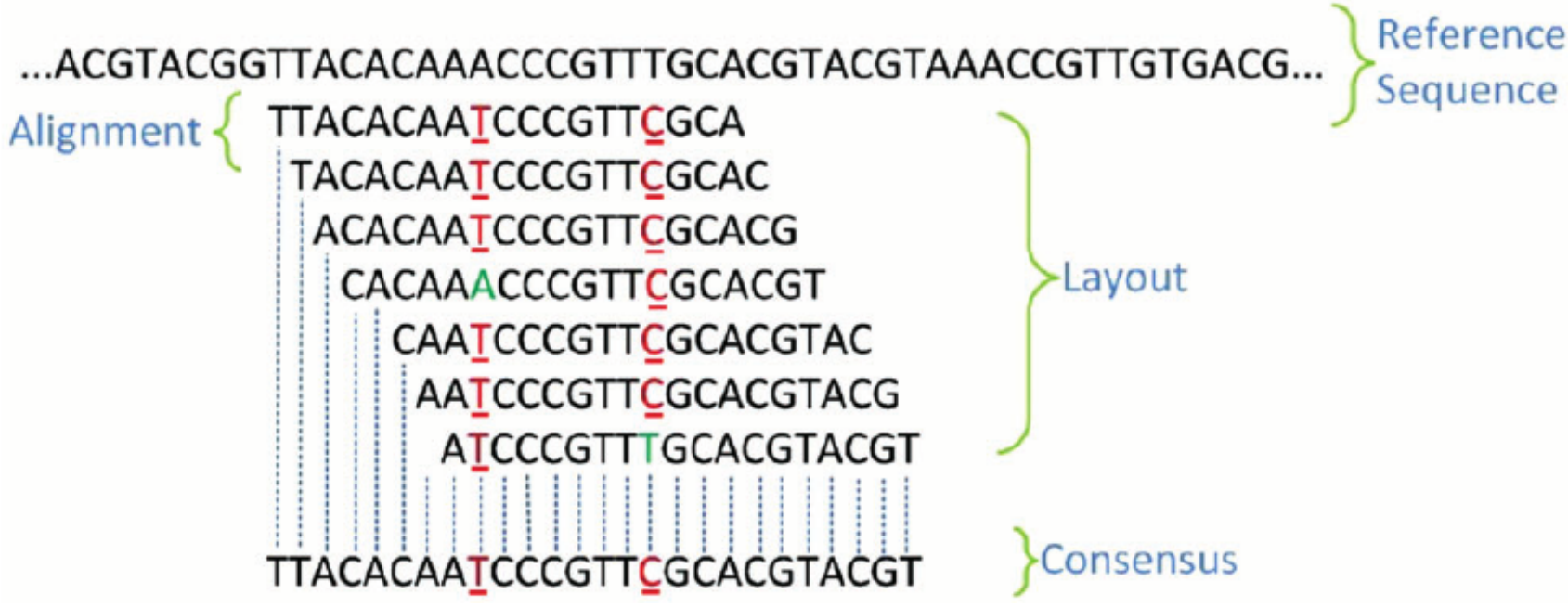
3. Assemble overlaps into contigs



4. Assemble contigs into scaffolds

(Baker, 2012)

Mapping to a reference genome



Long read data

Can help with repeat regions, and assembly

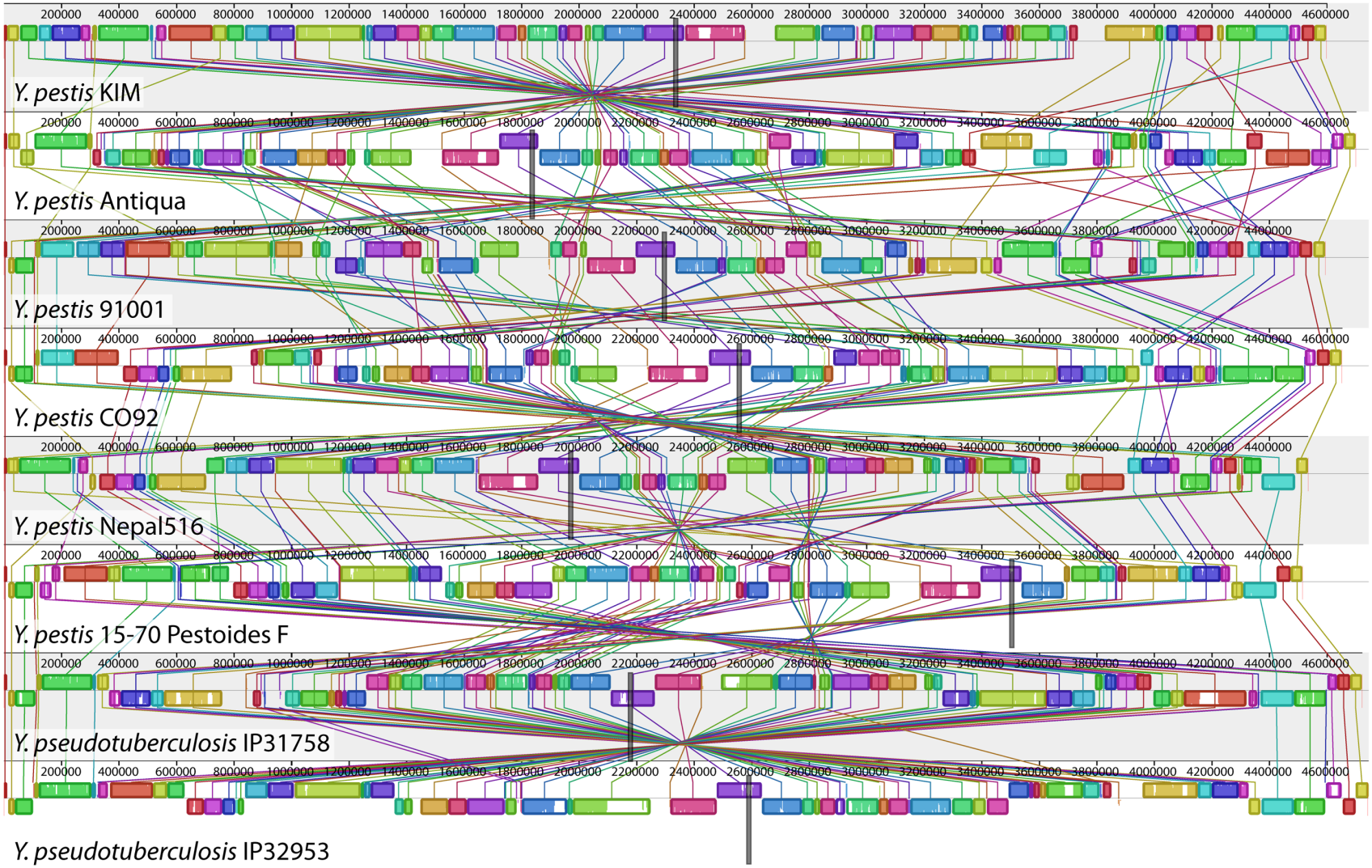
Error rates were high, but are coming down

Very good for capturing structural rearrangements and creating long contigs. (so you)

To make evolutionary statements, you need to align genomic regions across taxa.
Depending on evolutionary history this can be easy or hard!

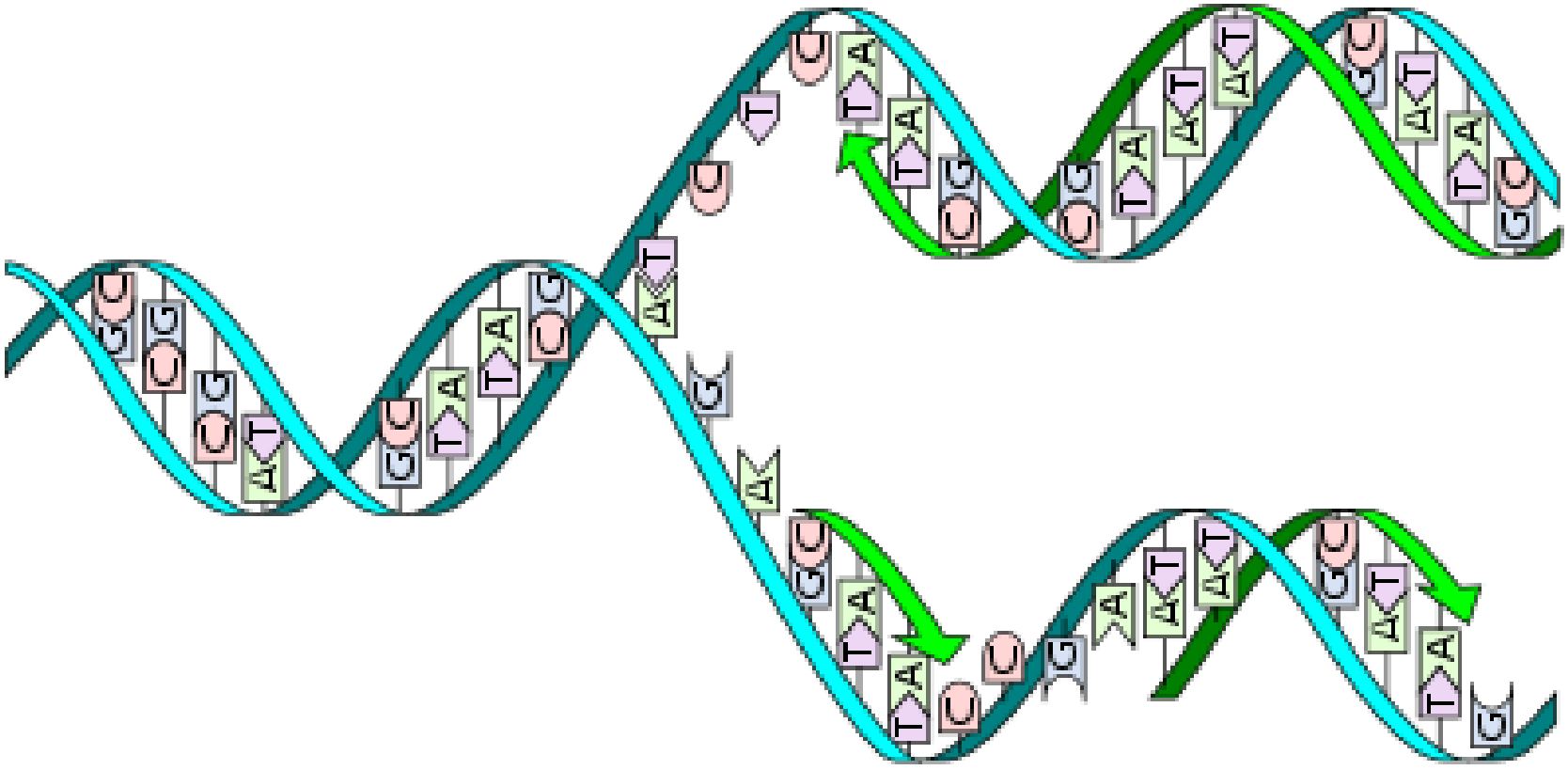
To make evolutionary statements, you need to align genomic regions across taxa.

Depending on evolutionary history this can be easy or hard!



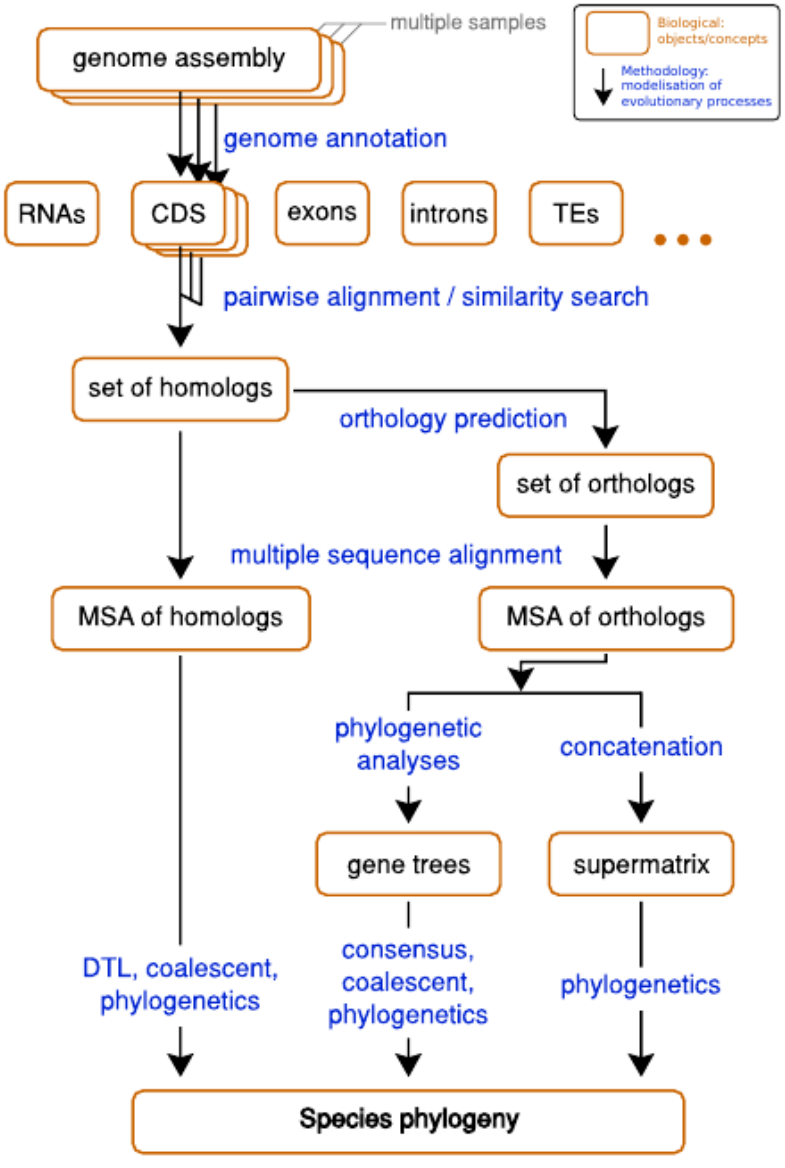
(Darling et al., 2008)

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)

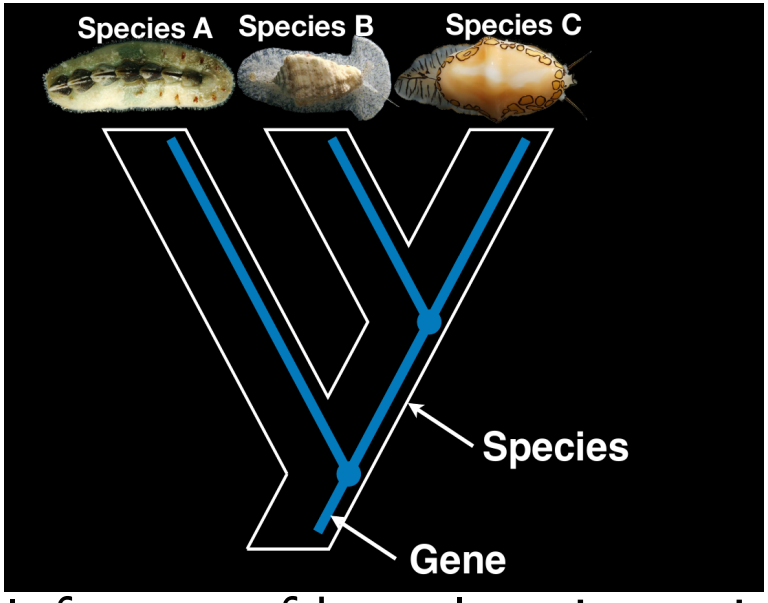
Enrichment methods focus our sequencing efforts on these regions



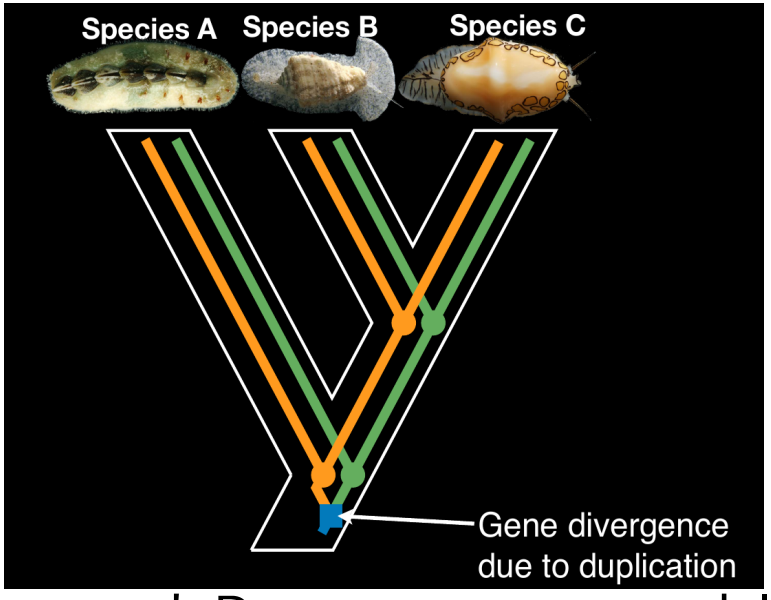
Simion et al. (2020)

Gene duplication and loss

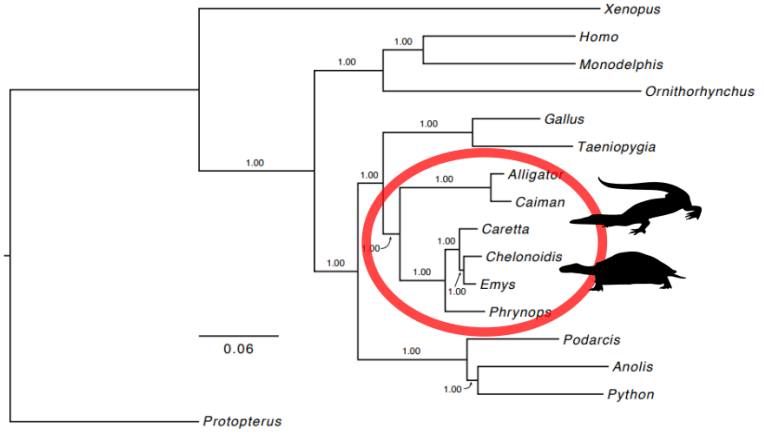
Orthology



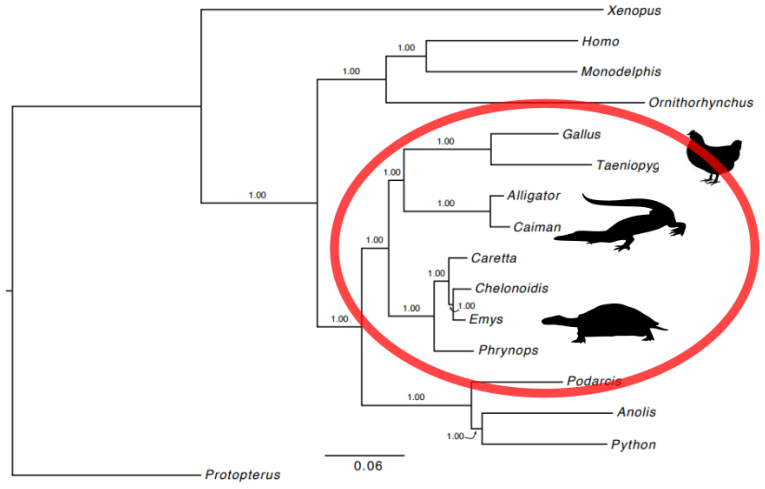
Paralogy



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



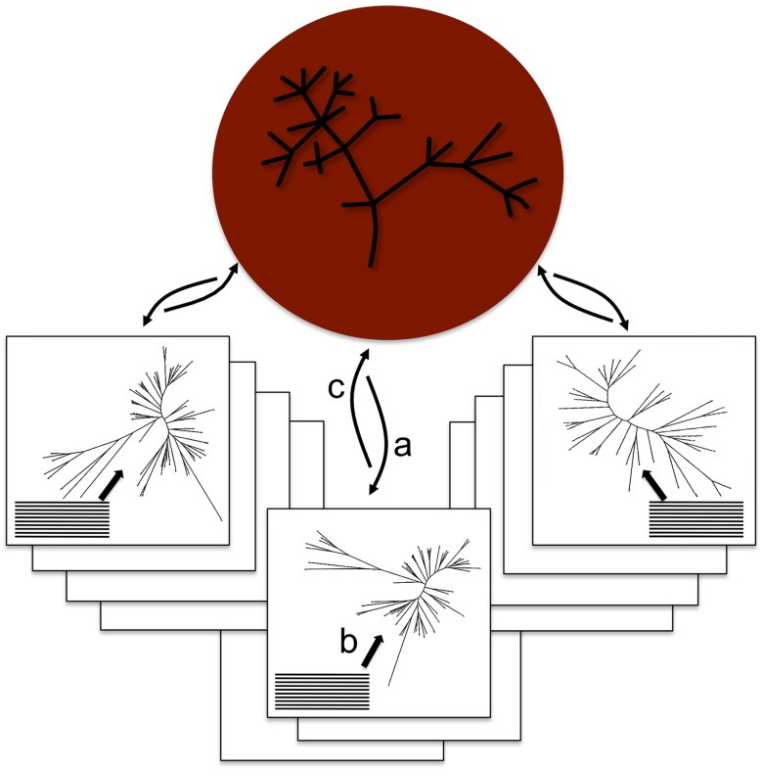
Same analysis of the same dataset
after removal of the two genes with evidence for paralogy.

"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 computationally expensive.

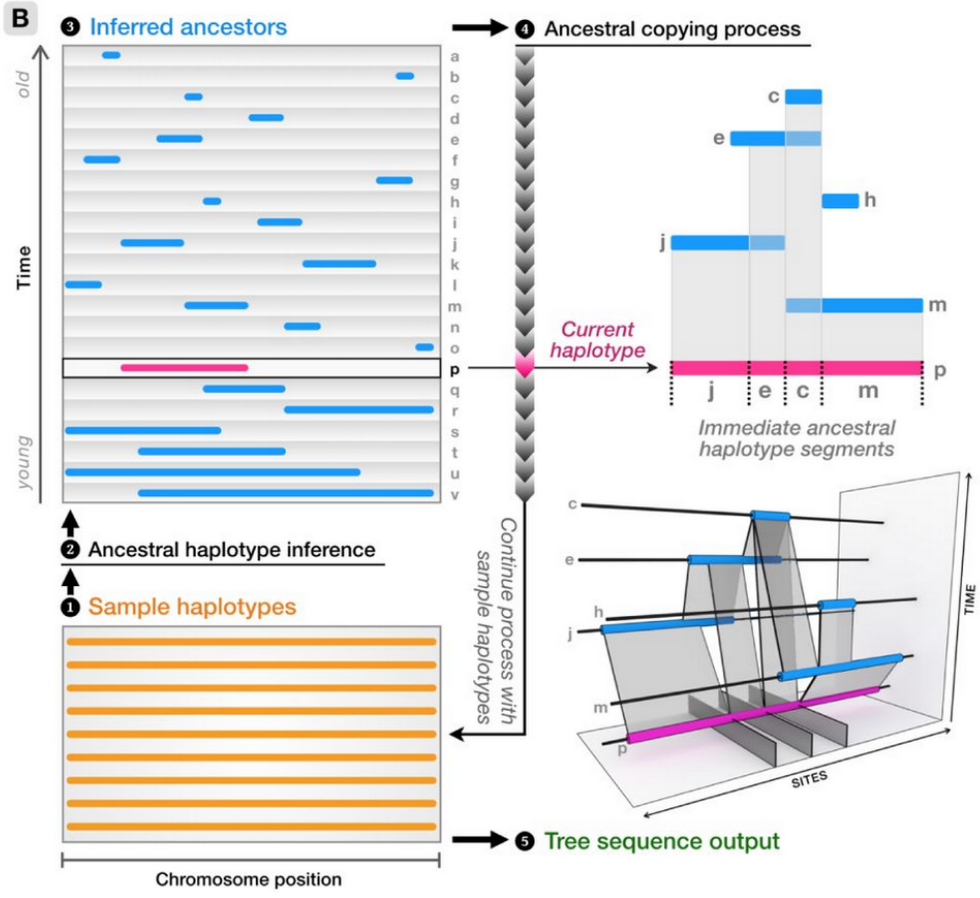
$$L(T, S, N|A) = \prod_{G_i \in \mathcal{G}} L(G_i)$$



Talk to Bastien Bousseau!
Phyldog; (Boussau et al., 2013)

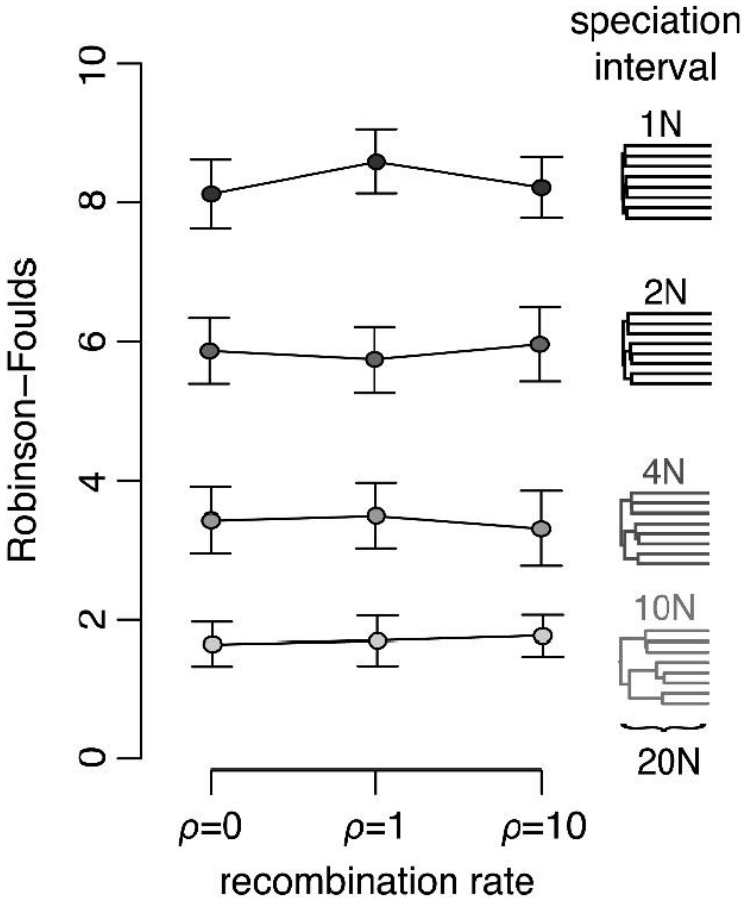
As we densely sample genomes and taxa, the size of a ‘locus’ or ‘un-recombined region’ will get smaller.

It is possible to jointly estimate recombination and ancestries along genomes



But do you really need to? (Kelleher et al., 2018)

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

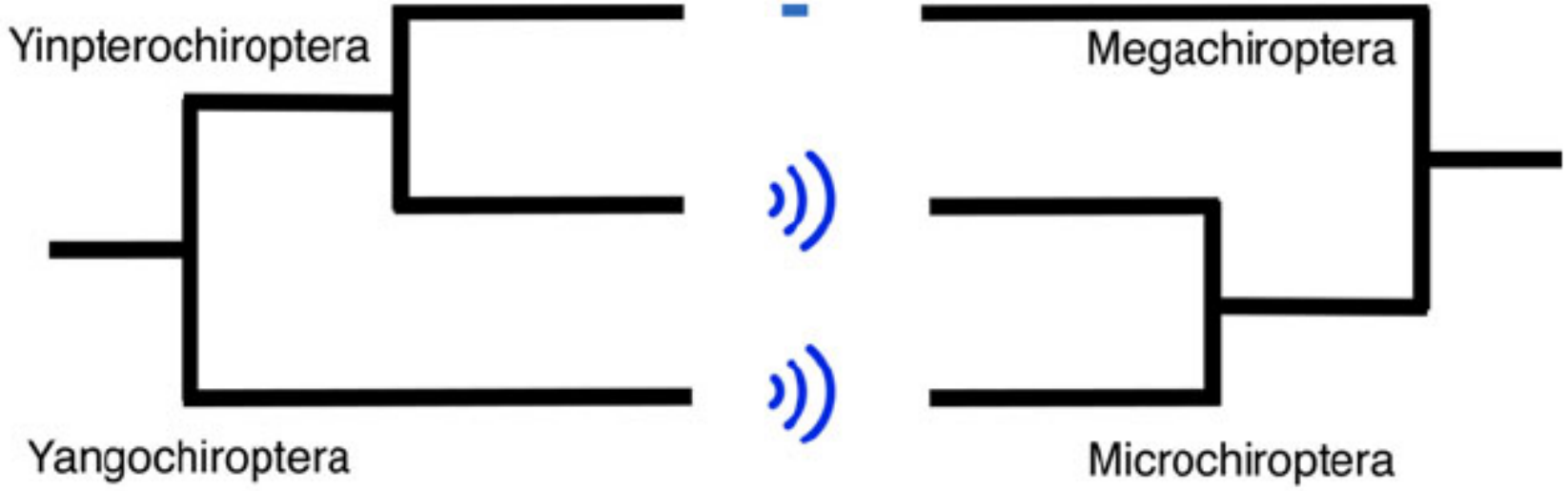


(Lanier and Knowles, 2012)

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.

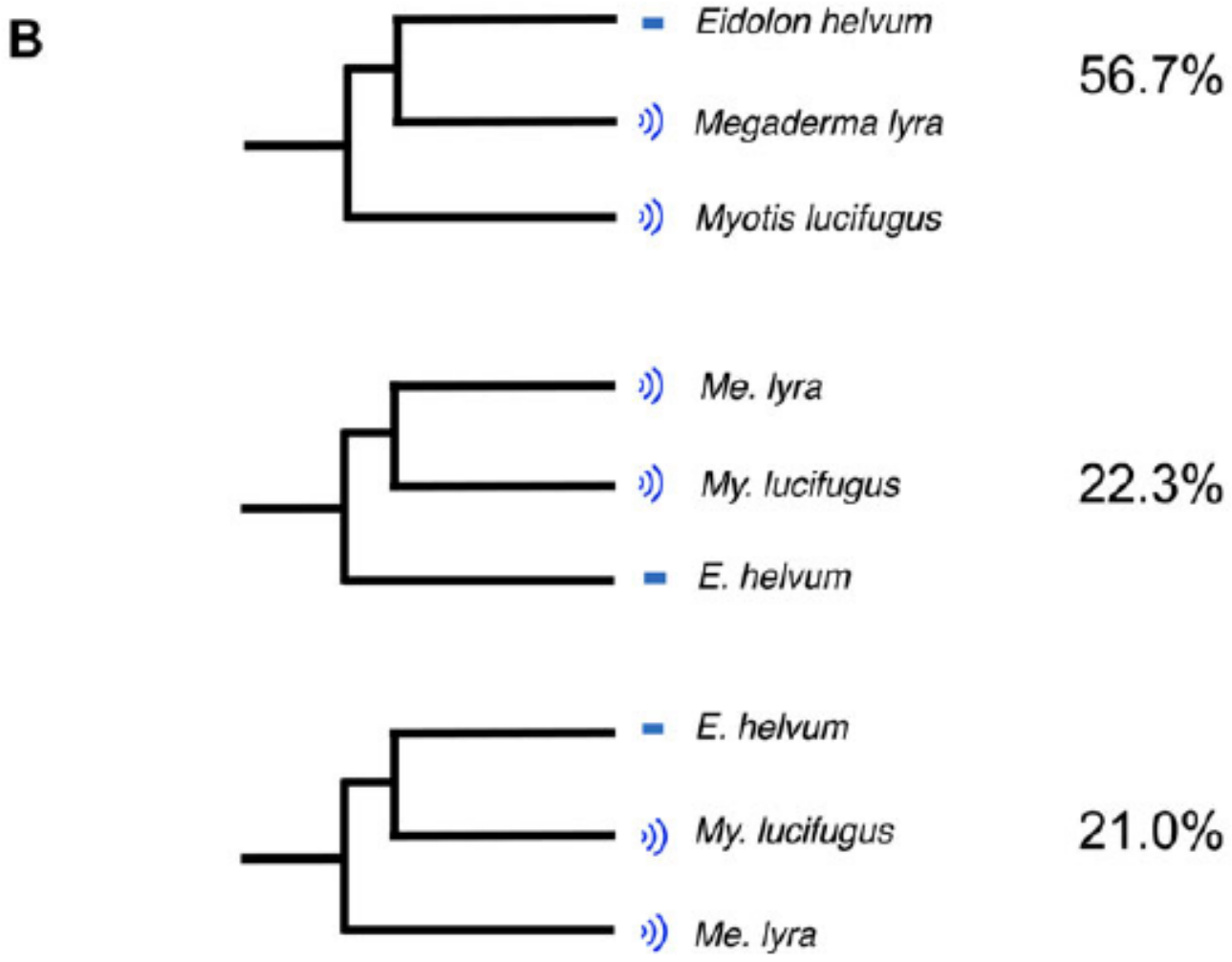
Is the species tree even what you want?

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)

Do you even need the species tree?



(Hahn and Nakhleh, 2016)

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!

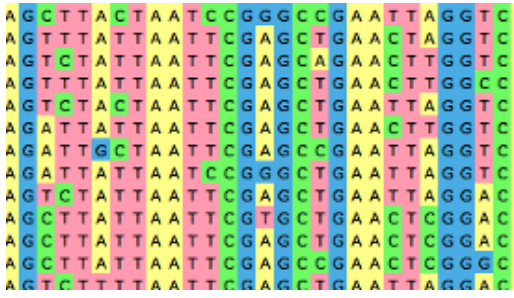
Most phylogenetic methods cannot directly handle whole genome data, but from whole genome sequencing you can get homologous loci, as well as a bunch of other stuff!

Data processing/ascertainment bias

How do the choices we make in

to

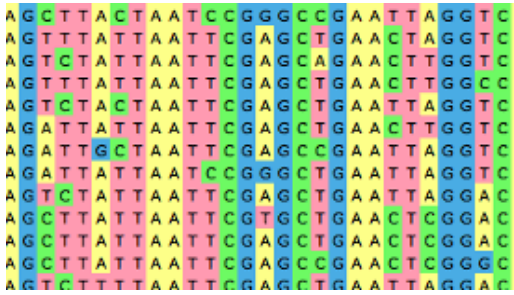
to



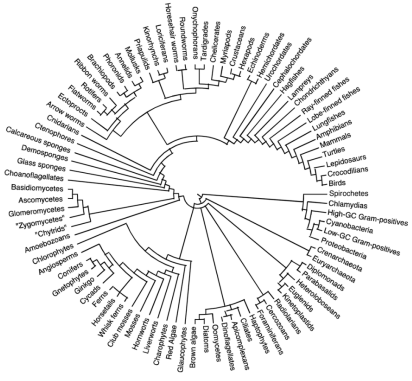
How do the choices we make in

to

to



affect



?

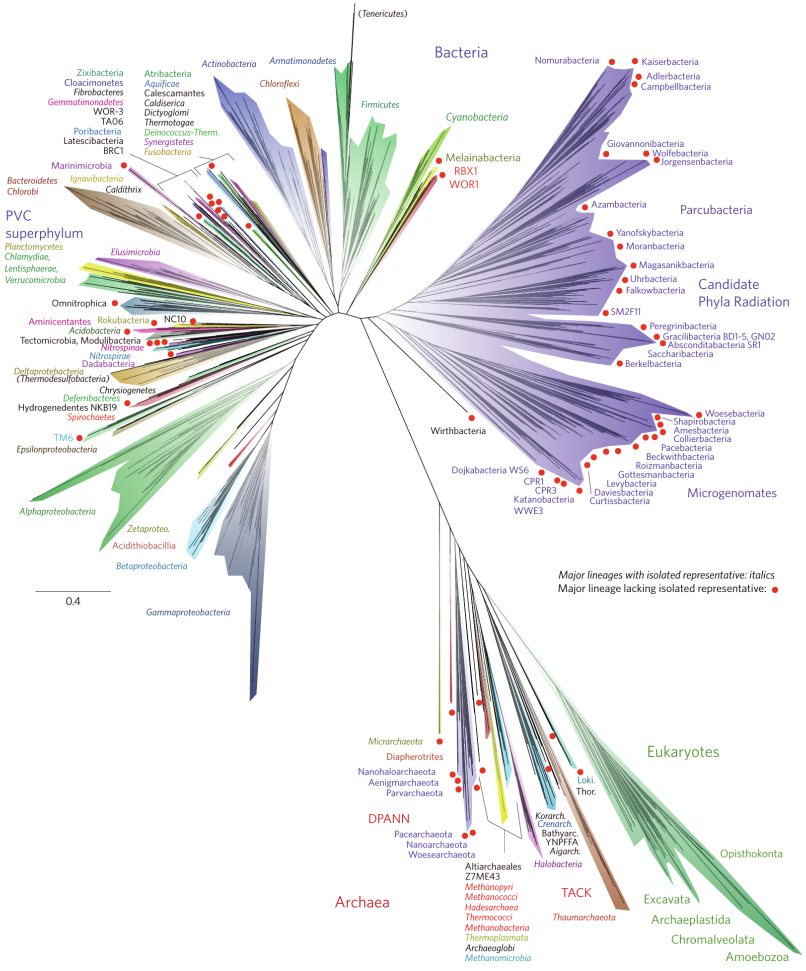
Ascertainment bias

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

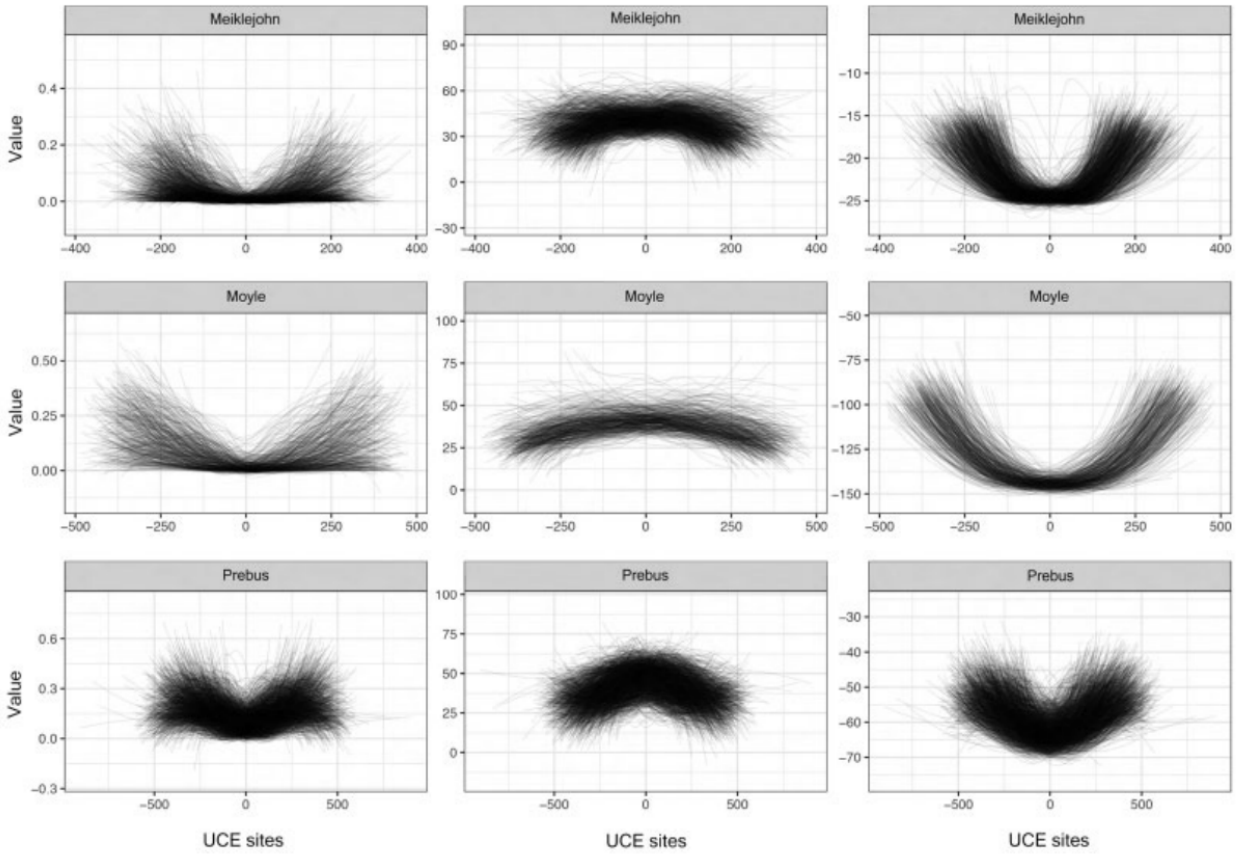


(Hug et al., 2016)

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, 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)

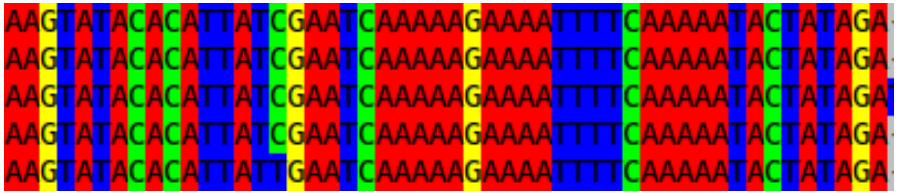
Analyzing only variable sites (e.g. Single Nucleotide Polymorphism (SNP) analyses)

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

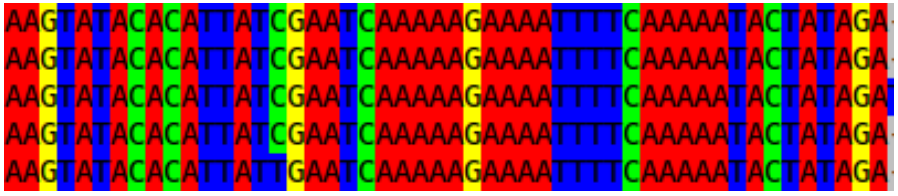
Lewis (2001) developed a likelihood model for estimating phylogeny morphological character data, which can be conditioned on all characters being

Based on correction for problem of not counting un-observed restriction sites in

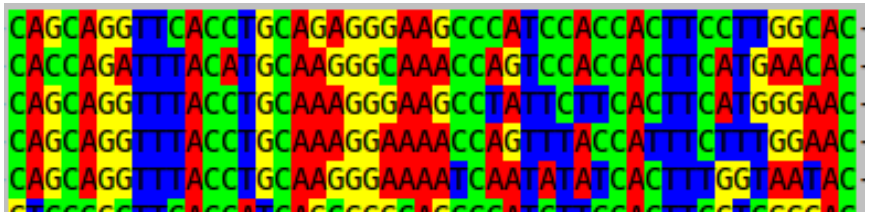
Short Tree


AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATCGAATCAAAAAGAAAATTTTCAAAAATACTATAGA
AAGTATACACATTATTGAATCAAAAAGAAAATTTTCAAAAATACTATAGA

Short Tree

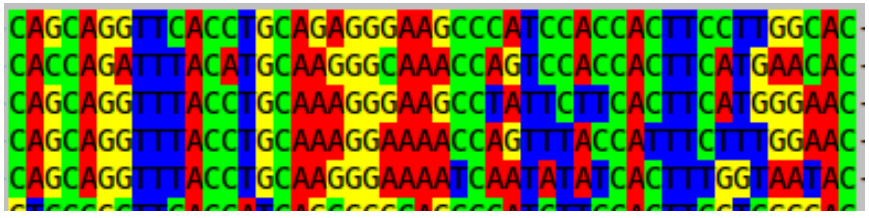


Long Tree

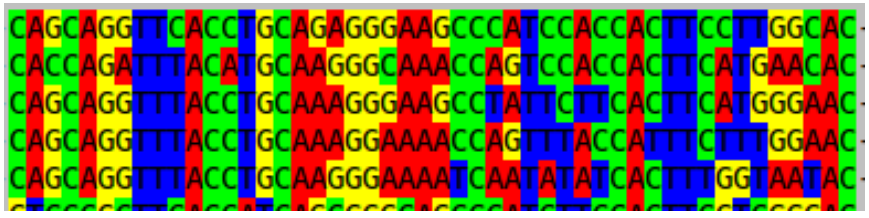


How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long

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 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

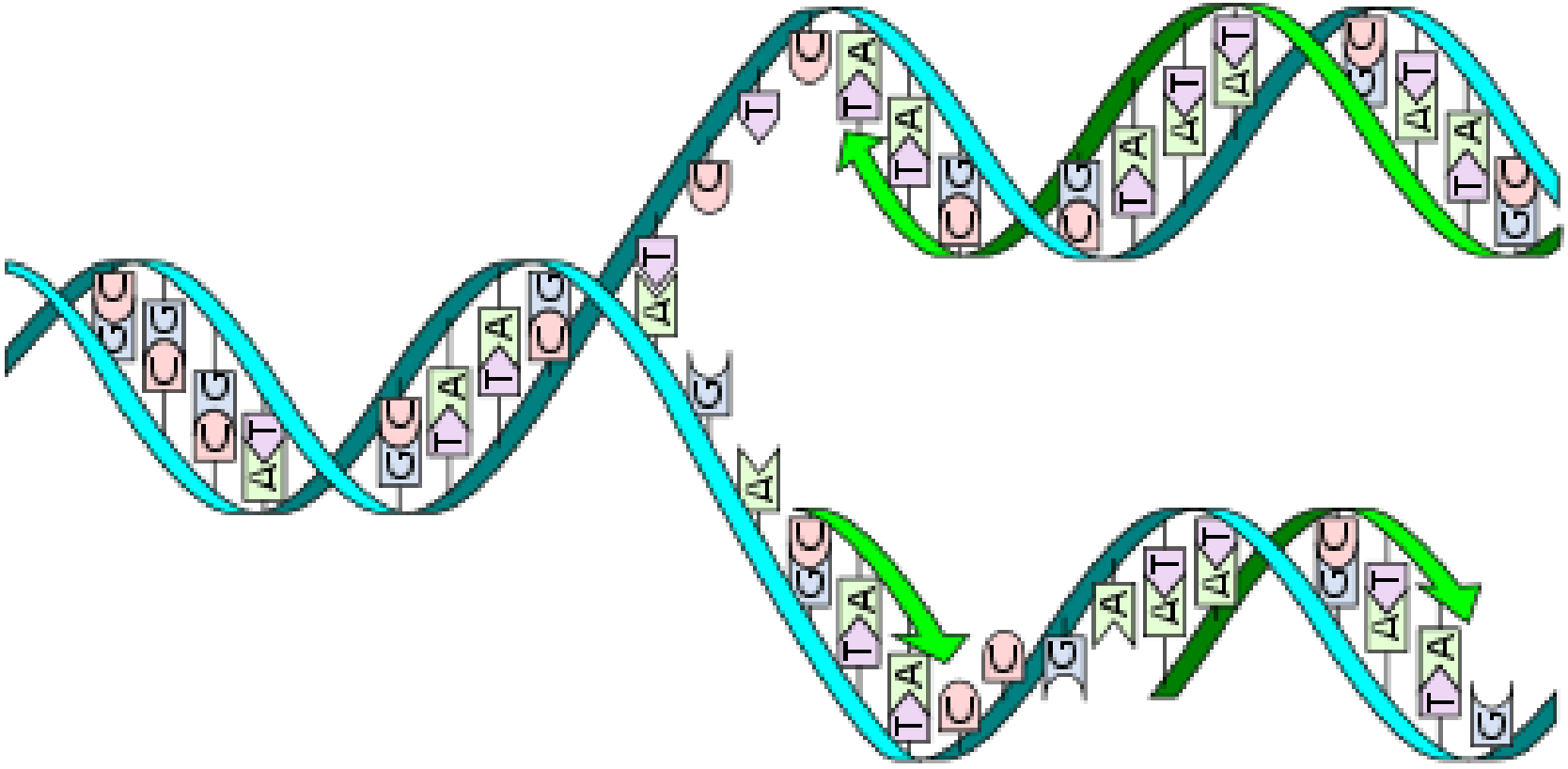
~~Despite~~ **because of** the large volume of data in genomic studies, ascertainment bias is still an issue

Ascertainment Bias Exercise

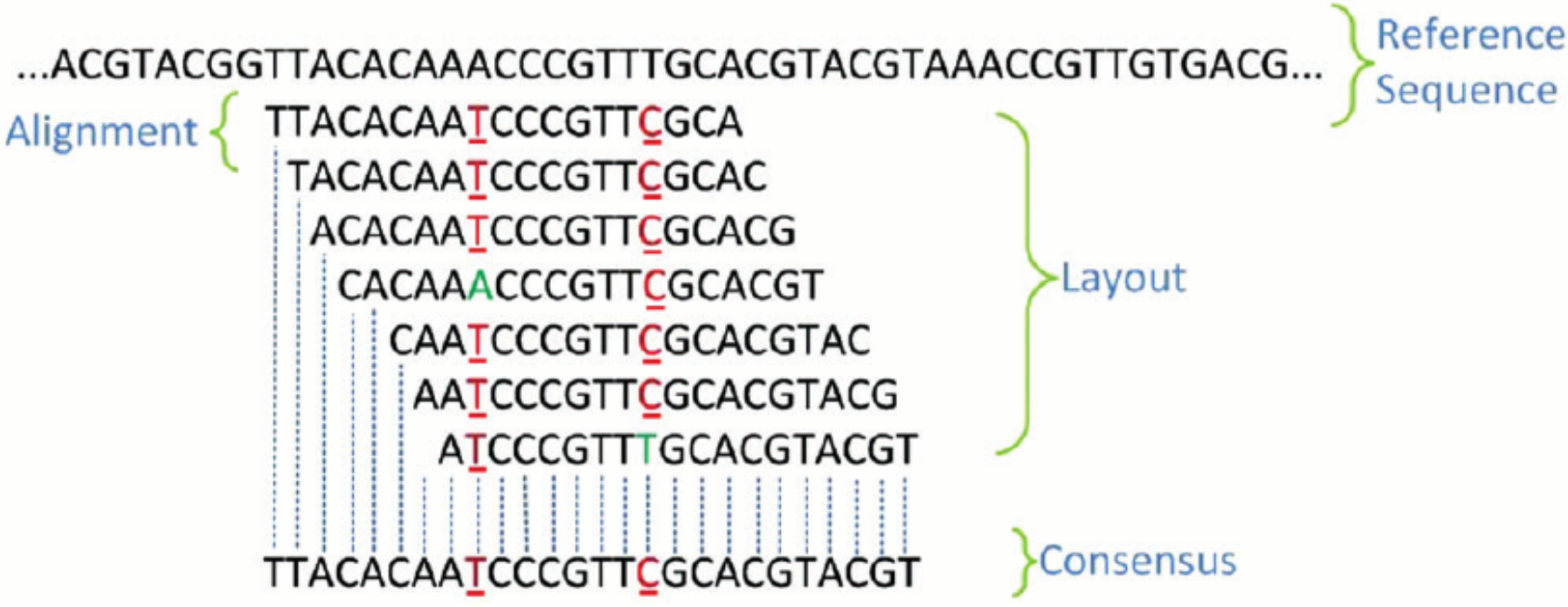
Part 1

https://github.com/snacktavish/sequence_data_exercise/blob/main/AscertainmentBias.md

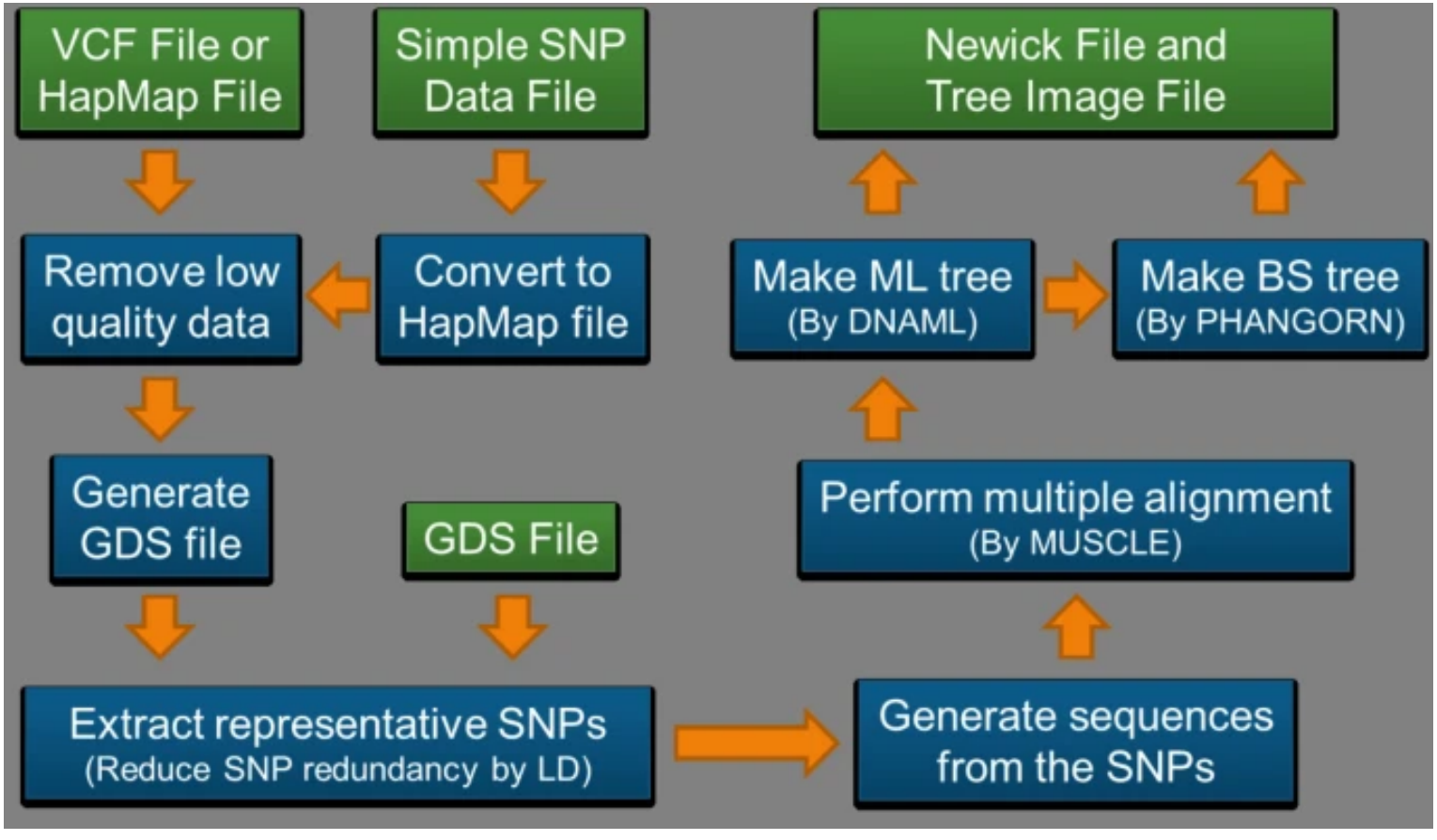
Exploring homology statements An alignment is a statement of shared ancestry (homology)



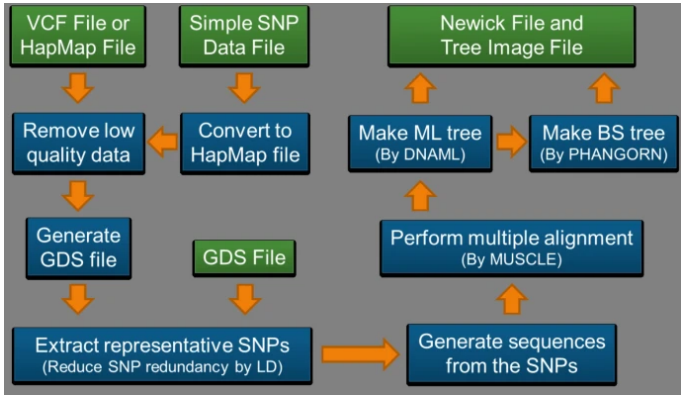
When in our analysis pipelines are we making homology statements?


...ACGTACGGTTACACAAACCCGTTTGCACGTACGTAAACCGTTGTGACG...
Reference Sequence
Alignment
TTACACAATCCCGTTCGCA
TACACAATCCCGTTCGCAC
ACACAATCCCGTTCGCACG
CACAAACCCGTTCGCACGT
CAATCCCGTTCGCACGTAC
AATCCCGTTCGCACGTACG
ATCCCGTTTGCACGTACGT
Layout
TTACACAATCCCGTTCGCACGTACGT
Consensus

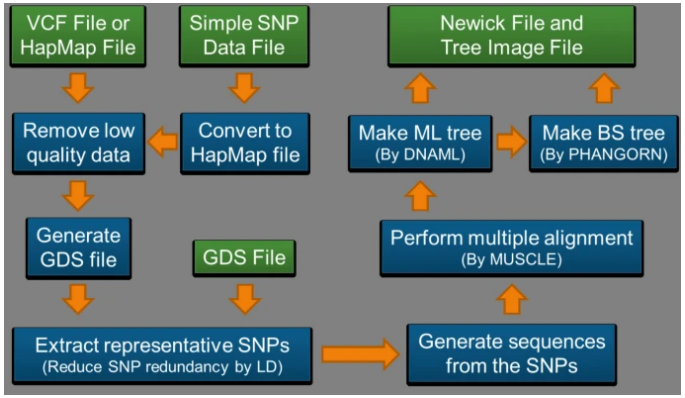
Where are the homology statements in this pipeline?



(Lee et al., 2014)



SNPhylo (Lee et al., 2014) has around 350 citations...



SNPhylo (Lee et al., 2014) has around 350 citations...

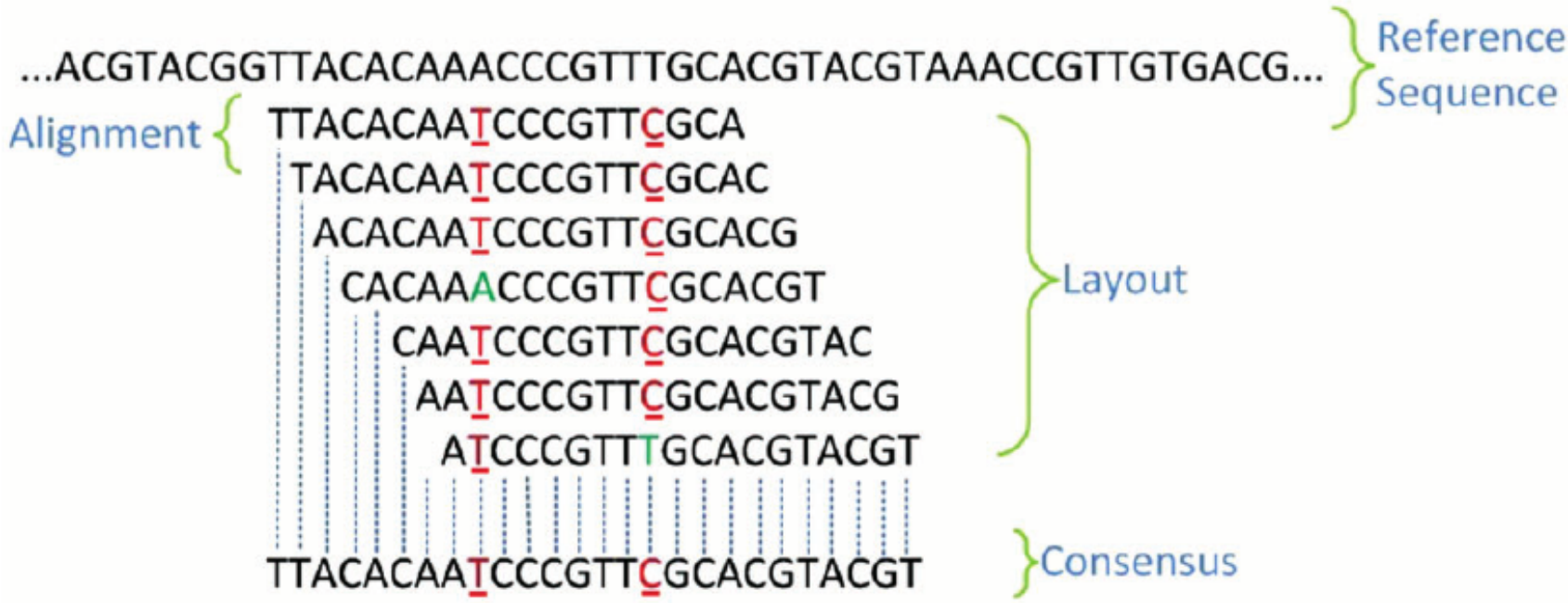
"Users of software pipelines that automatically assemble RAD loci and generate phylogenies (Bertels et al. 2014; Lee et al. 2014) should be careful to verify that the proper models are being used for phylogeny estimation, since default settings may not be appropriate for data sets composed entirely of SNPs." (Leaché et al., 2015)

Ascertainment Bias Exercise

Part 2

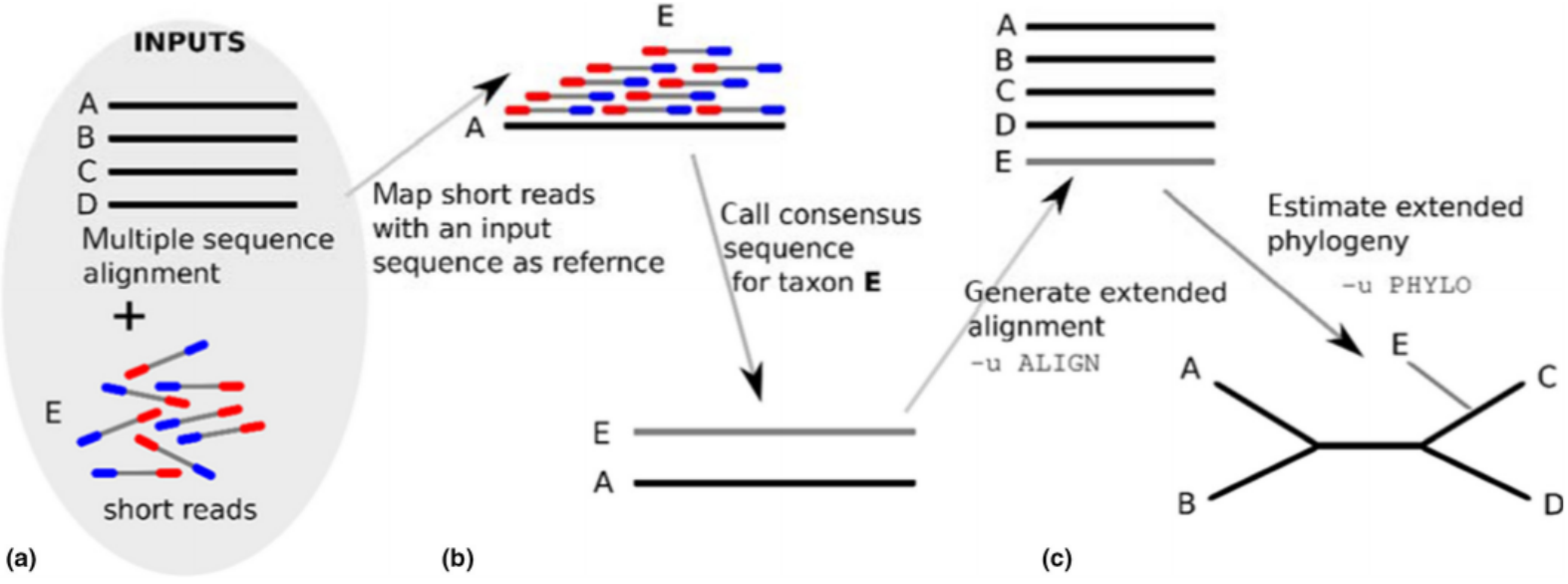
https://github.com/snacktavish/sequence_data_exercise/blob/main/AscertainmentBias.md

Assembly



Alignment

Phylogenetically informed phylogenomic updating approach: Extensiphy



- Assembles only homologous regions of interest
- Maintains homology from assembly to MSA
- Can use multiple references to generate consensus sequence
- Tree search speed up due to starting tree

`github.com/mctavishlab/extensiphypipeline`

Toscani-Field et al. (2022)

What to do?

- What data will answer **your** questions?
- Are there existing data you want to be able integrate with?
- Use the most an appropriate available model for your data

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.

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.

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

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17. Number: 1.

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.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):e1006581. Publisher: Public Library of Science.

Lanier, H. C. and Knowles, L. L. (2012). Is Recombination a Problem for Species-Tree Analyses? *Systematic Biology*, 61(4):691–701. Number: 4.

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.

Lee, T.-H., Guo, H., Wang, X., Kim, C., and Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, 15(1):162.

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

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

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.

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

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