Ascertainment bias in genomic phylogenetic inference

Emily Jane McTavish

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

How do our choices in data collection affect our inferences?

Intro

How have we

- designed our sampling?
- chosen our data type?
- processed and filtered our data?

and how do those choices affect our results?

Intro

Ascertainment bias

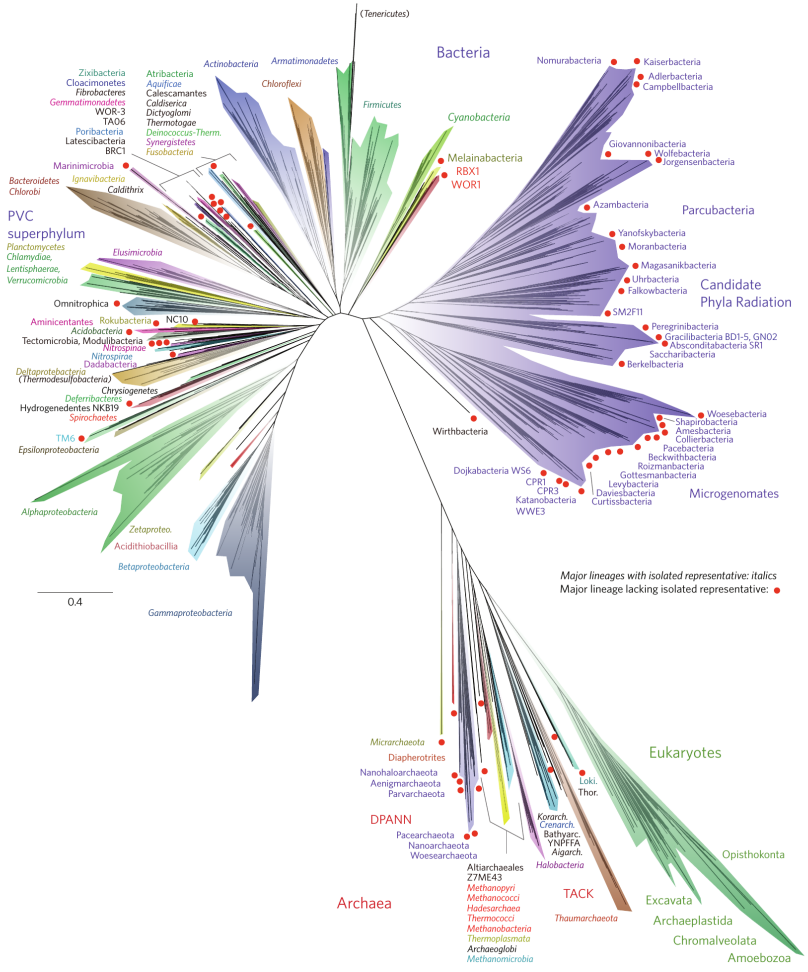
A bias in parameter estimation or testing caused by non-random sampling of the data.

'Ascertainment bias' in a broad sense, covers 'selection bias' or 'acquisition bias', all three terms have been used for overlapping issues.

Ascertainment bias is ubiquitous!

- Surveying volunteers
- Studying undergraduates
- Sampling across ‘species’
- Discarding rare outliers

Sampling across the tree of life



Hug et al. (2016)

Outline

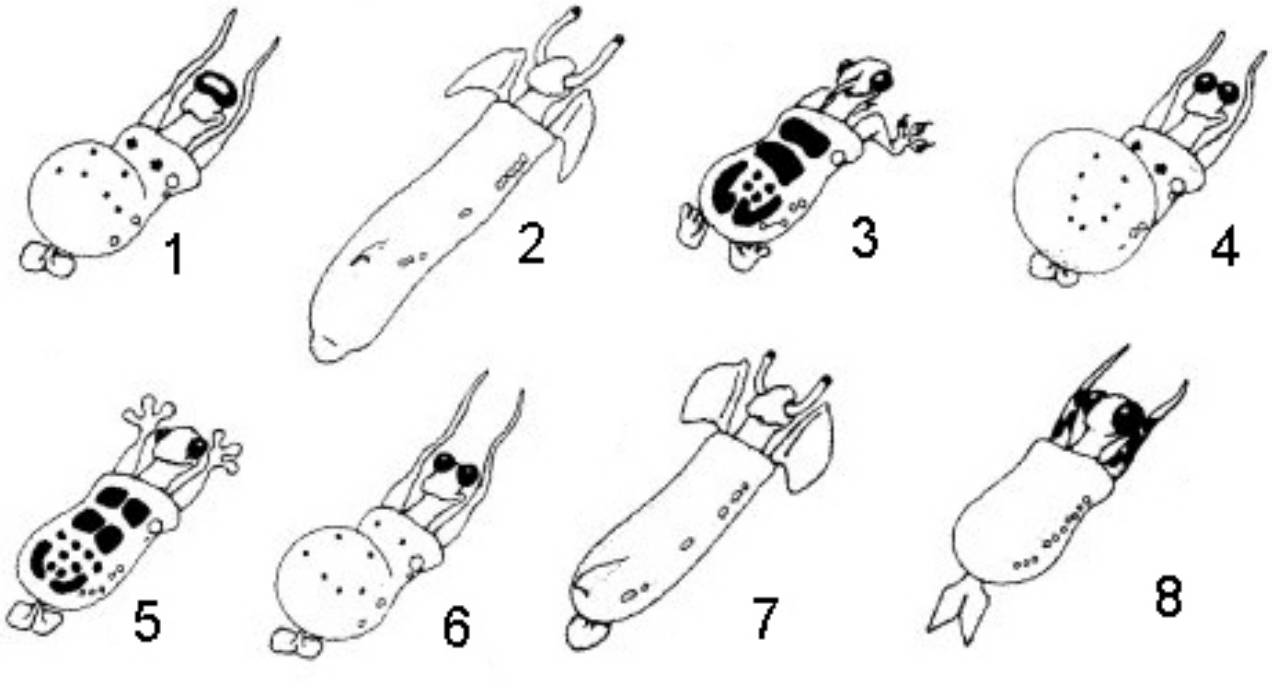
Introduction

Case studies of ascertainment bias in analyses of genomic data sets

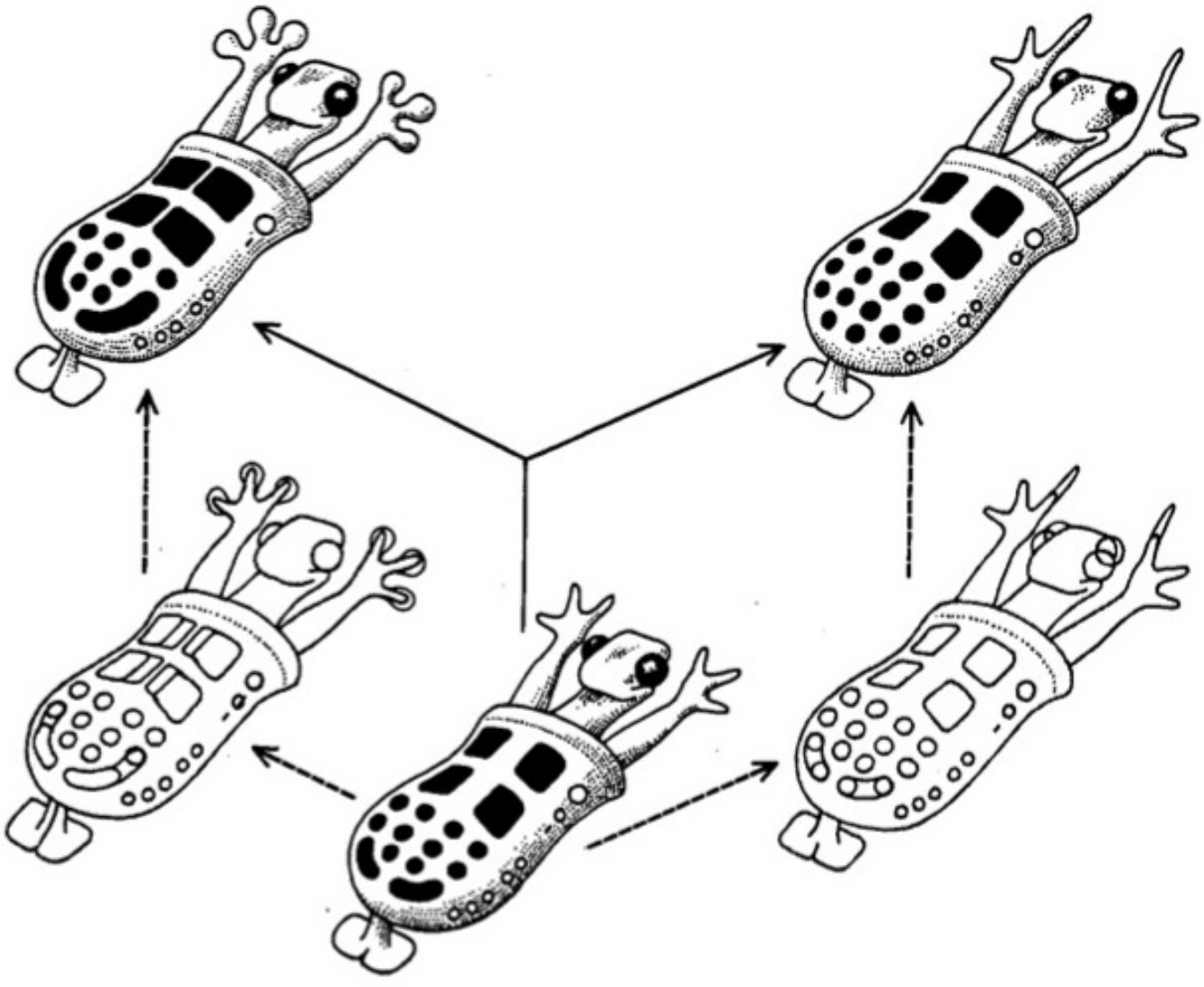
- Inferring cattle global evolutionary history using SNP chip data
- Phylogenetics of Penstemon using RADseq data
- Tracing a pathogenic *Salmonella* outbreak using whole genome sequencing

What to do?

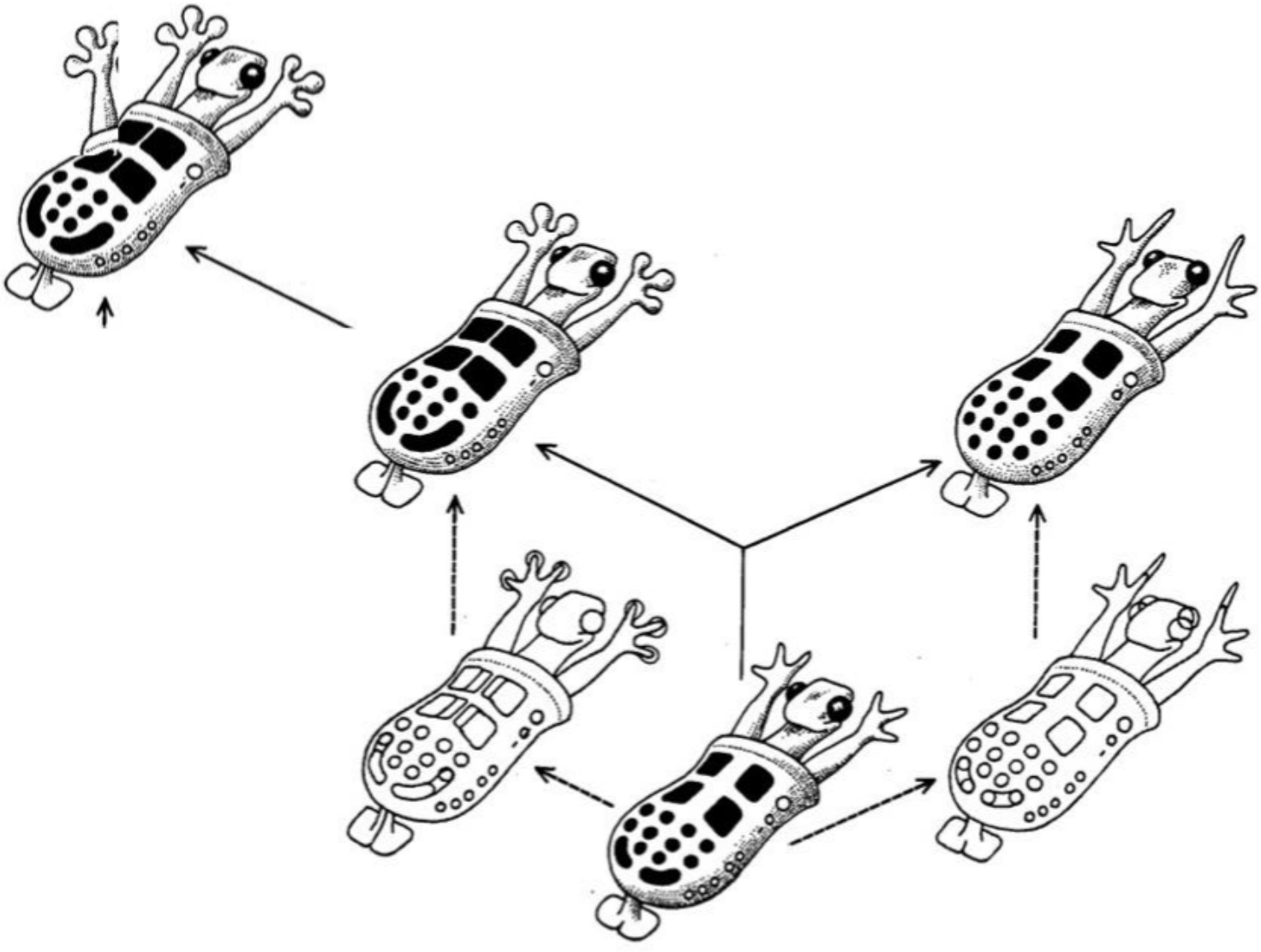
Caminalcules



What characters would you use to build a phylogeny of these animals? (critters made by Joseph H. Camin as an early form of simulated phylogenetic data)



Focus on ‘informative’ i.e. variable traits



Bias due to inclusion of only variable sites

Analagous to including only variable sites in sequence aanalysis

Bias due to inclusion of only variable sites

Analagous to including only variable sites in sequence aanalysis
Instead of 'Anything Can Happen Now'

Bias due to inclusion of only variable sites

Analagous to including only variable sites in sequence aanalysis
Instead of 'Anything Can Happen Now'
'Something Definitely Happened Once!'

Bias due to inclusion of only variable sites

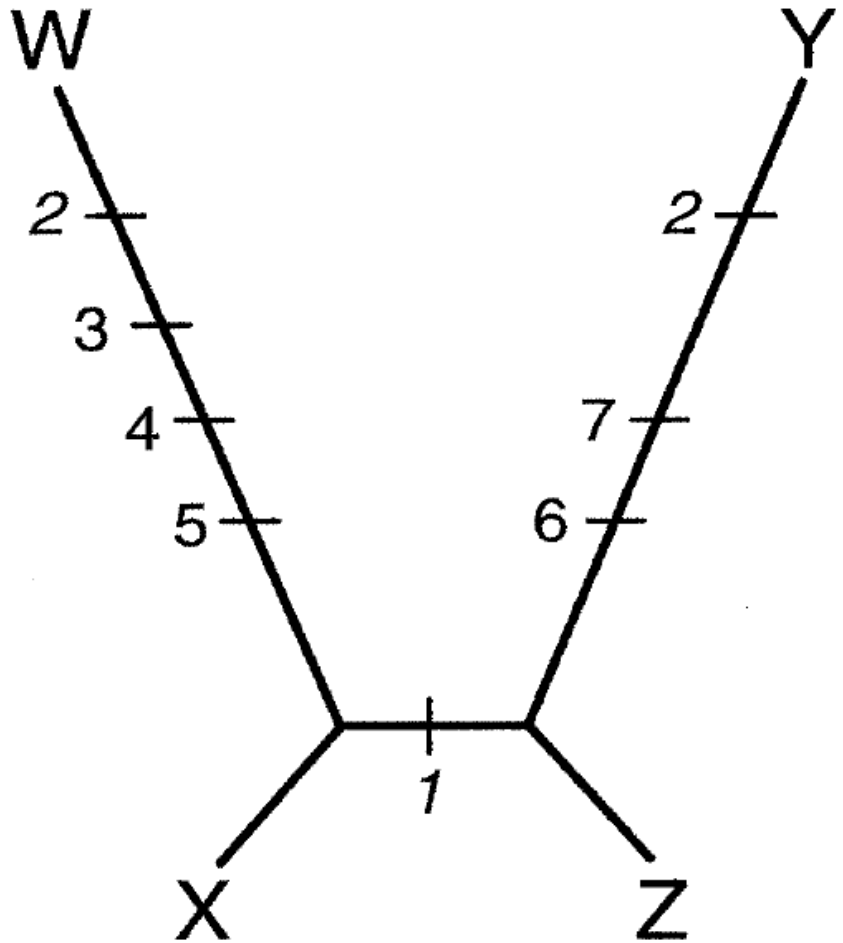
Analagous to including only variable sites in sequence aanalysis
Instead of 'Anything Can Happen Now'
'Something Definitely Happened Once!'

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

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

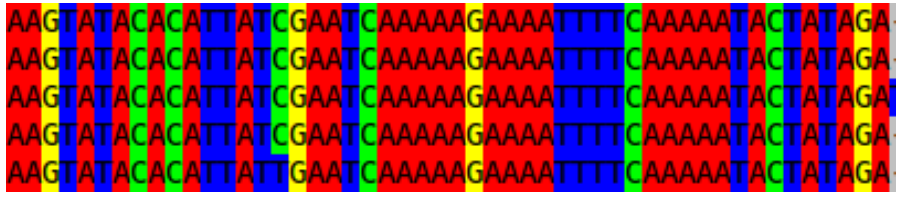
Based on correction for problem of not counting un-observed restriction sites in Felsenstein (1992)

Markov model for morphological data allows for likelihood based branch lengths. Lewis (2001)

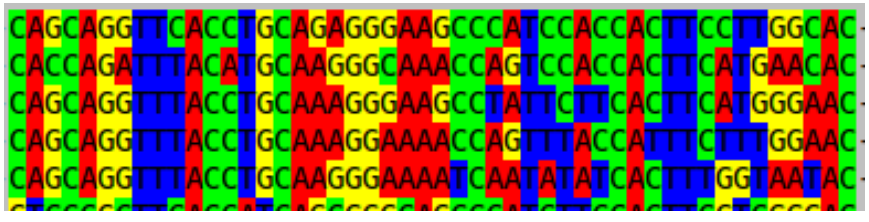


Lewis (2001) correction for only including variable sites applies to sequence data as well.

Short Tree



Long Tree



How surprised should we be?

Not seeing any invariant sites is very surprising unless branches are very long

How surprised should we be?

Not seeing any invariant sites is very surprising unless branches are very long

but only if we looked for them!

How surprised should we be?
Not seeing any invariant sites is very surprising unless branches are very long
but only if we looked for them!

Extension of Lewis (2001) model for analysis of only variable sites implemented in RAxML (Leaché et al., 2015)

"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

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 Study I

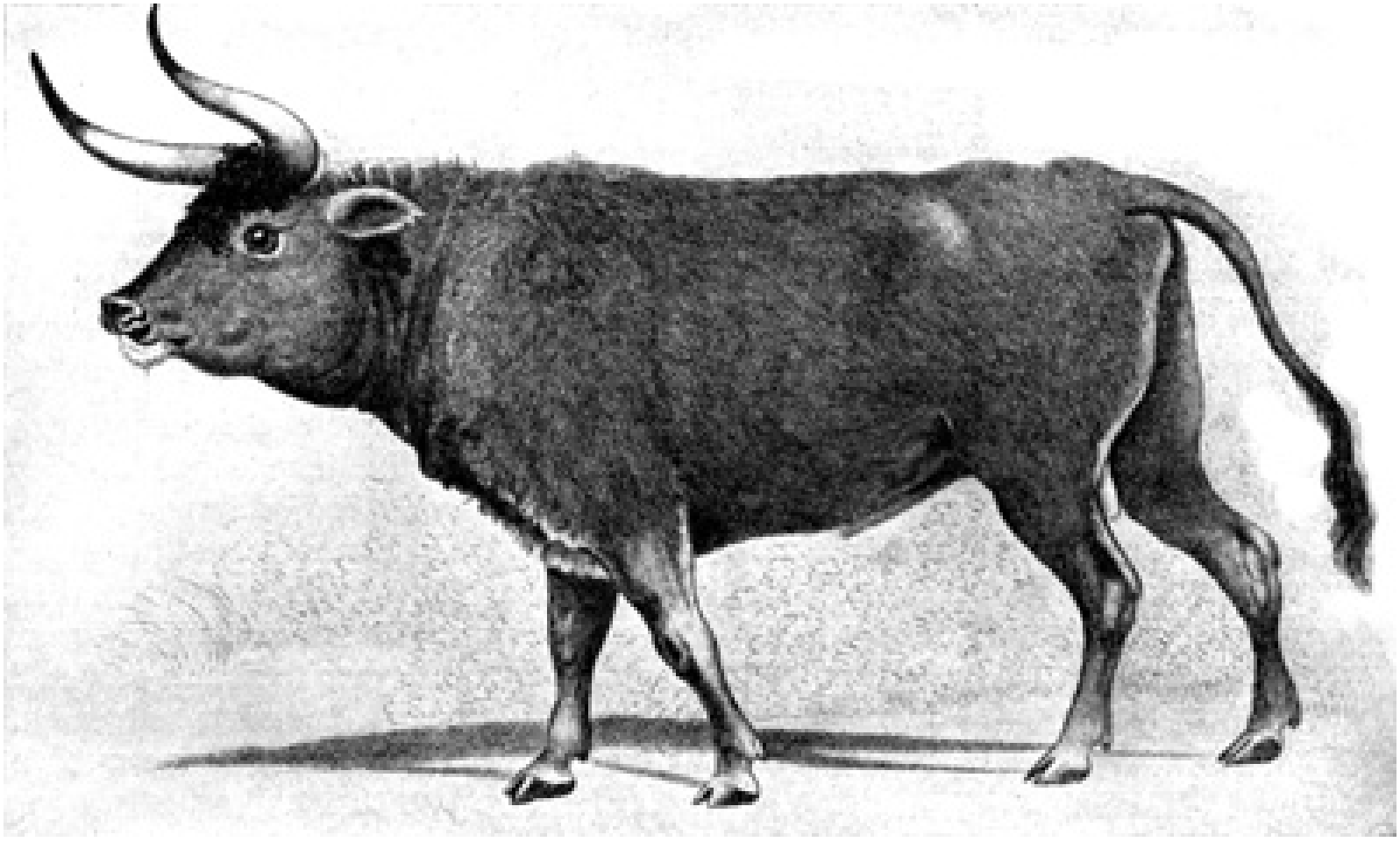
Inferring cattle global evolutionary history using SNP chip data

Question

What is the global evolutionary history of domesticated cattle?

A brief history of cattle

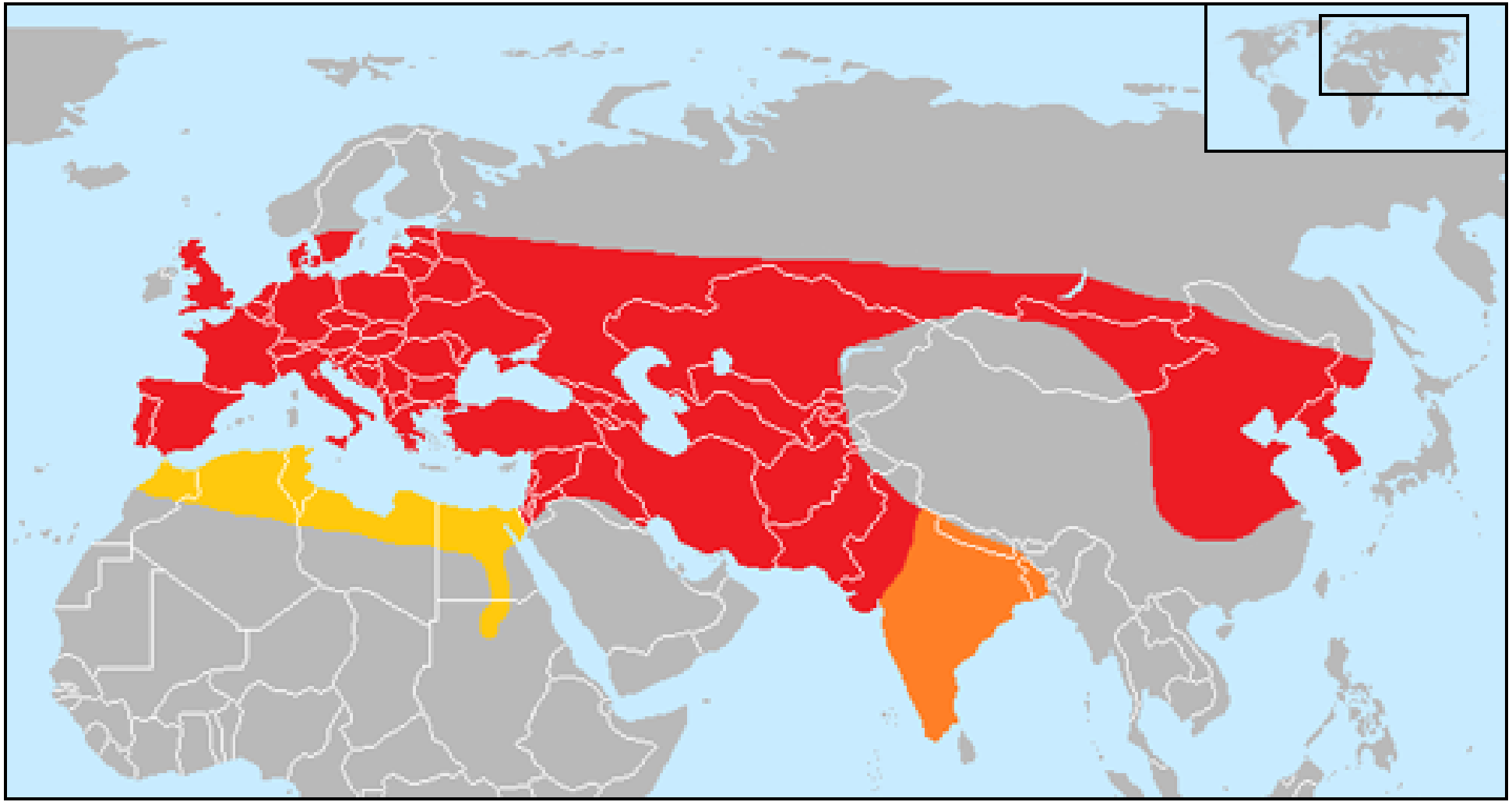
Domesticated from the aurochs *(Bos primigenius)*
Last aurochs died in 1627



19th century copy of an earlier painting

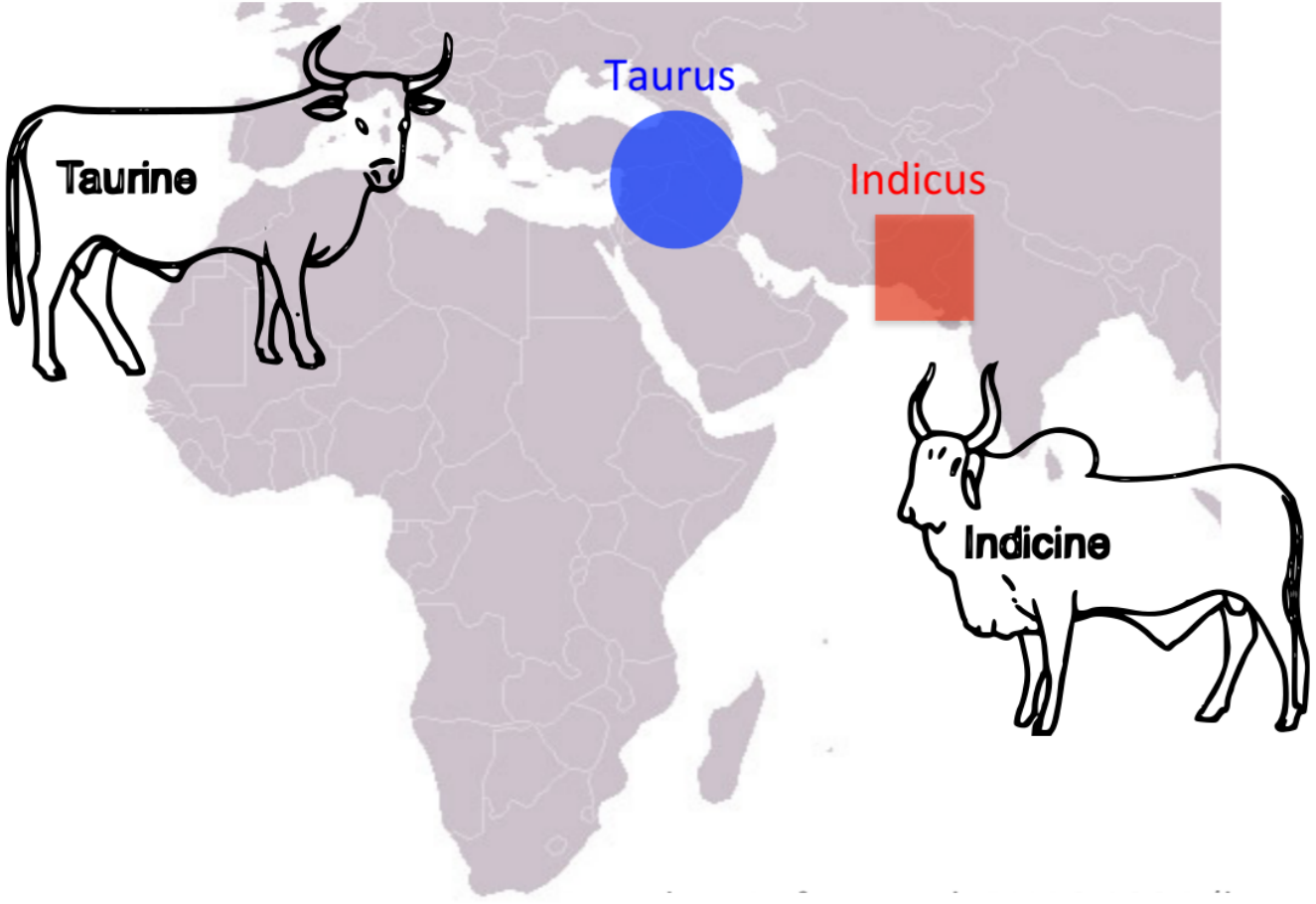
Distributon of Aurochs

Colors represent subspecies

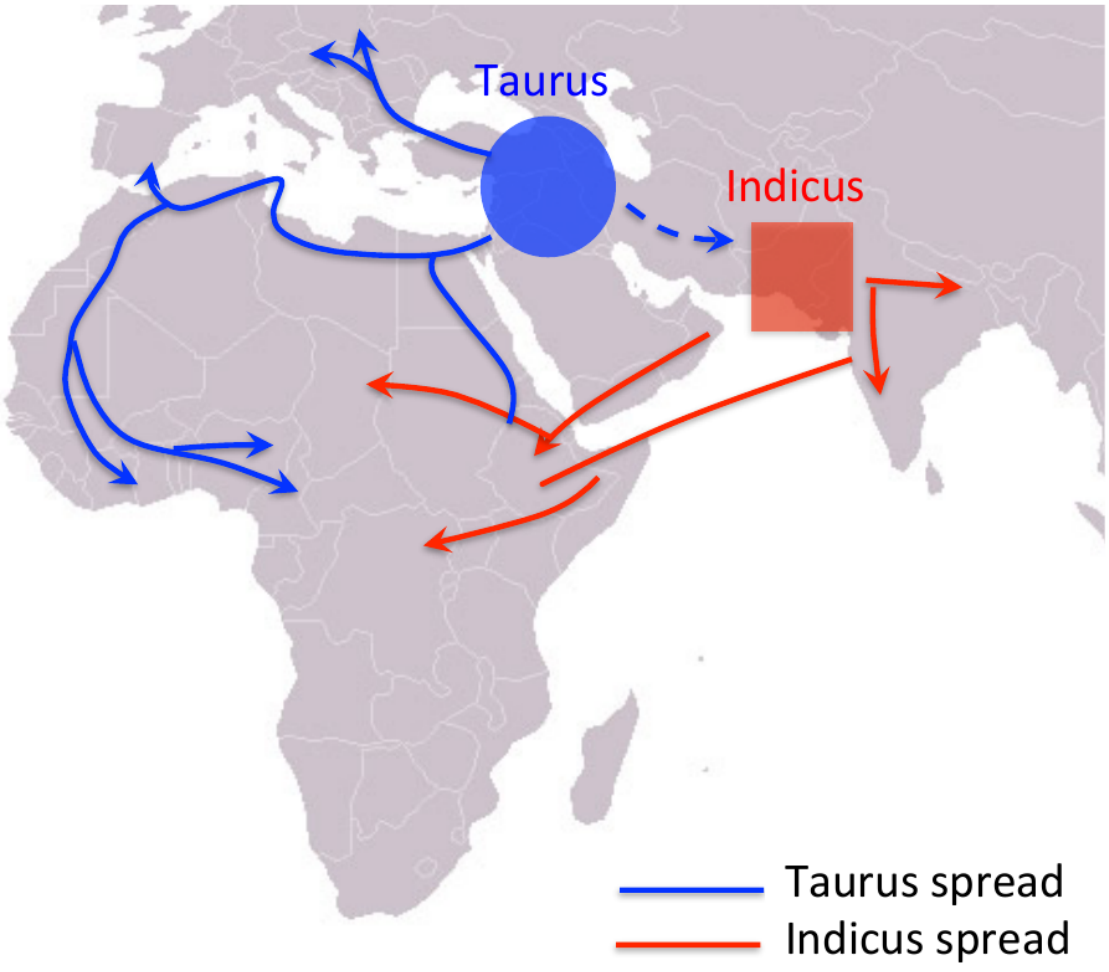


C.T. van Vuure, Retracing the Aurochs. 2005

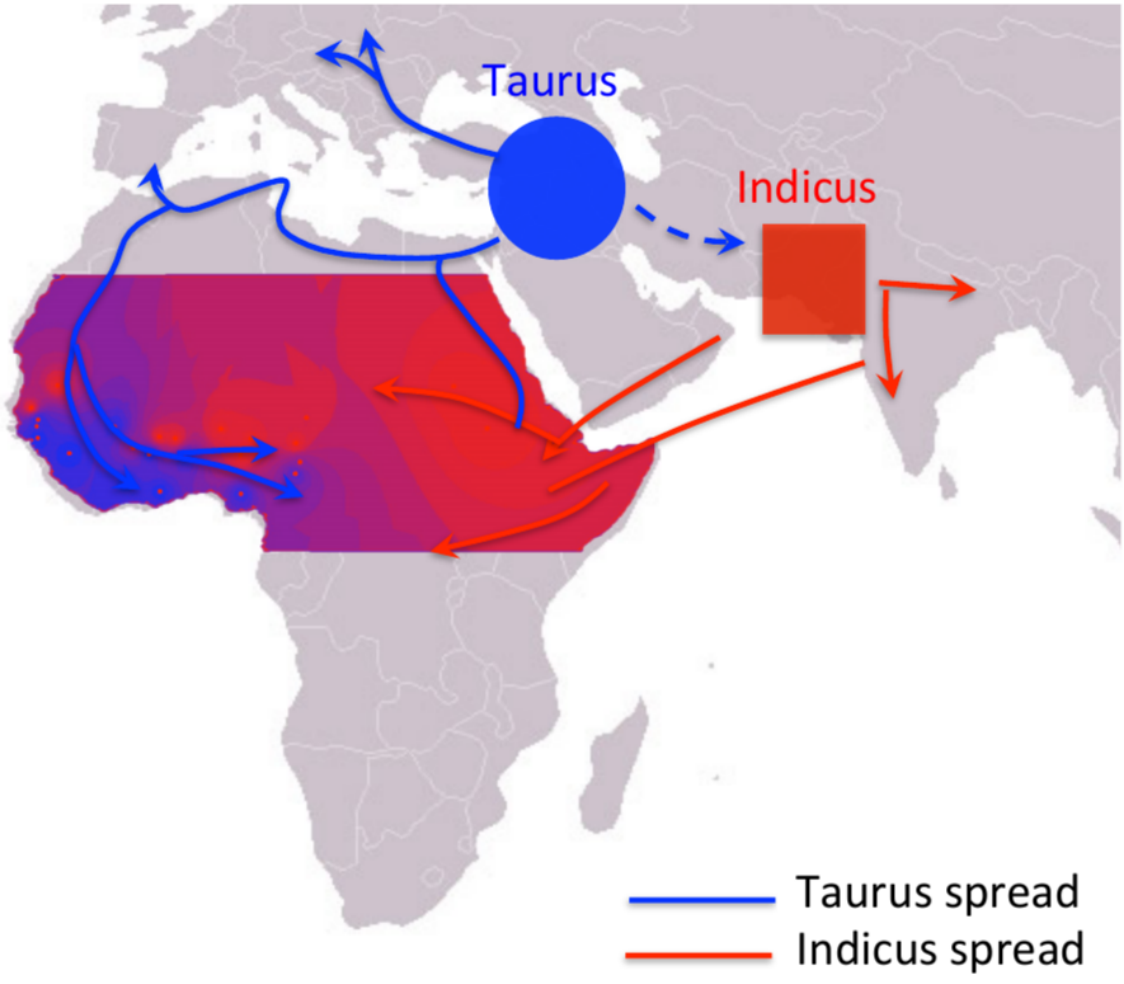
Two key domestication events



Subsequent spread

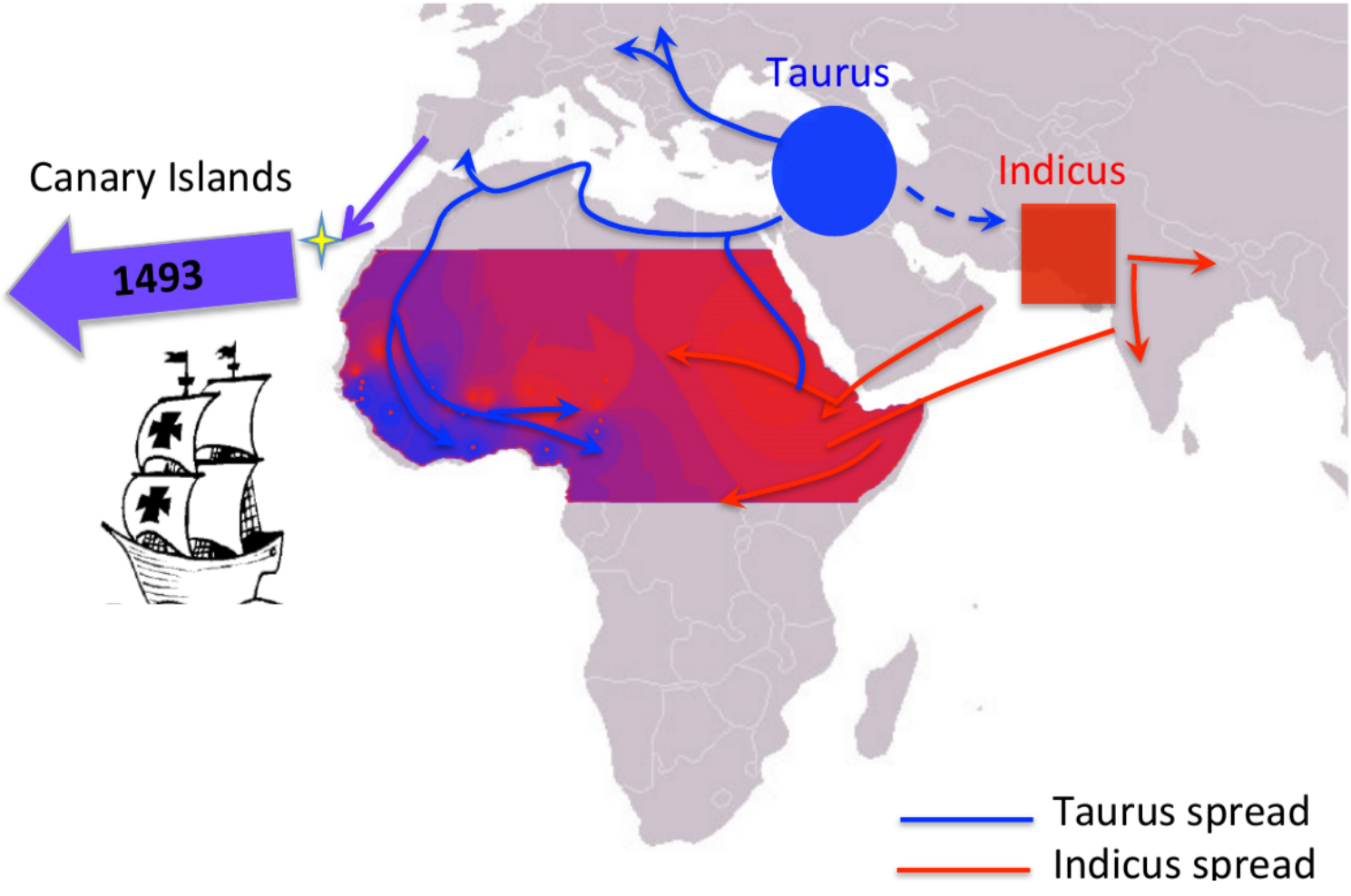


Hybridization cline across Africa



Freedman et al., Molecular Ecology. 2004

Introduction to the New World



First cattle in the Caribbean - 1493
Introduced to mainland - 1521
Descendants of these cattle include Texas Longhorns

Genomic data

Data set

- 54,609 single nucleotide loci, mapped to assembled *taurine* genome
- 1428 individuals from 58 breeds
- 963 taurus (including 71 from New World breeds)
- 273 indicus
- 191 hybrid origin

Potential for bias

Data set

- 50K single nucleotide polymorphism (SNPs) selected from re-sequencing data
- Biased selection of sites with high minor allele frequencies
- Only single nucleotides sequenced (SNP chip)
- Geographic bias in panel

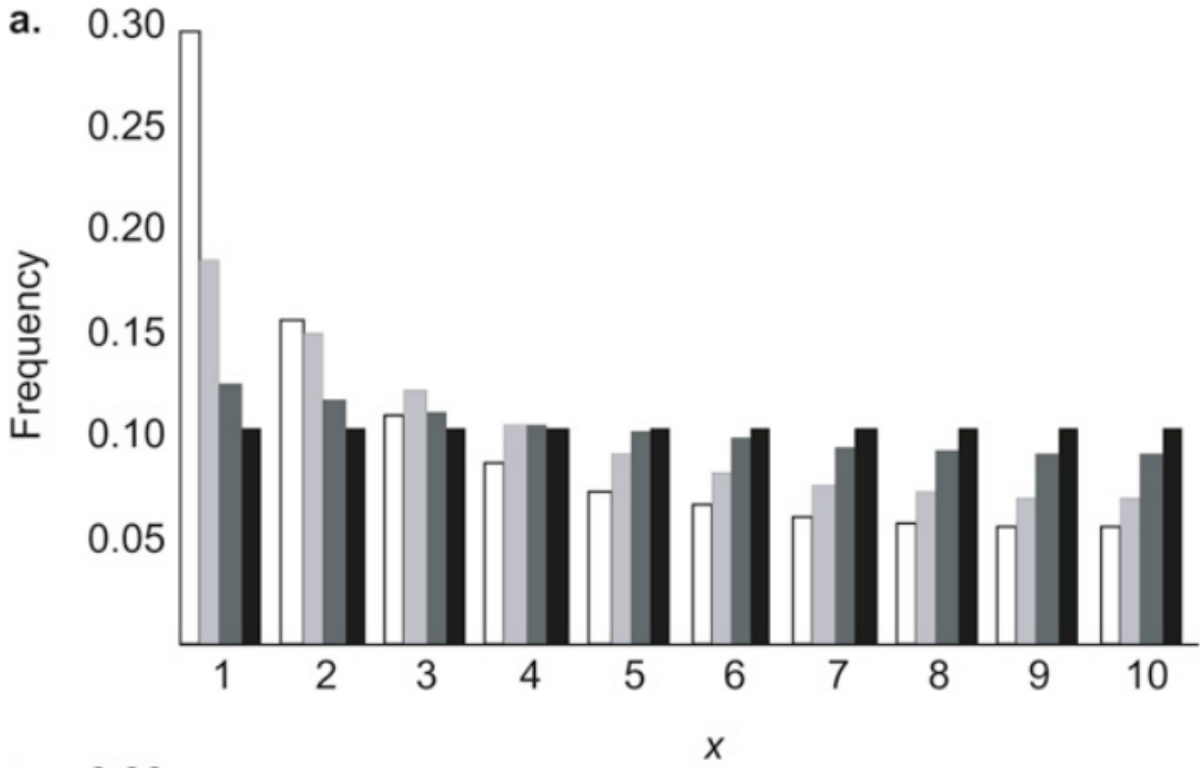
Potential for bias

Data set

- 50K single nucleotide polymorphism (SNPs) selected from re-sequencing data
- Biased selection of sites with high minor allele frequencies
- Only single nucleotides sequenced (SNP chip)
- Geographic bias in panel

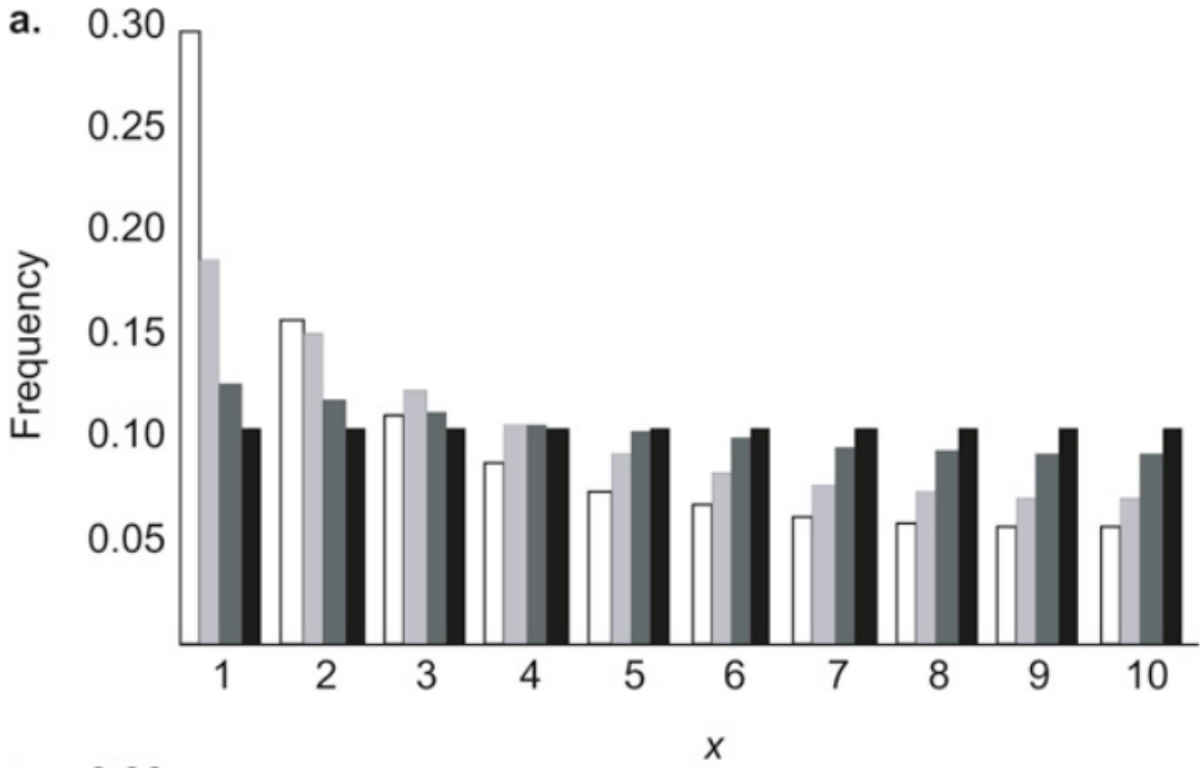
Where $$$ and project planning collide!

Effect of sample size on site frequency spectrum



The expected frequency of observed alleles of frequency X, using sample size of d = 2 (black), d = 5 (darkgrey), d = 10 (light grey) and d = 20 (white; no ascertainment bias). (Nielsen, 2004)

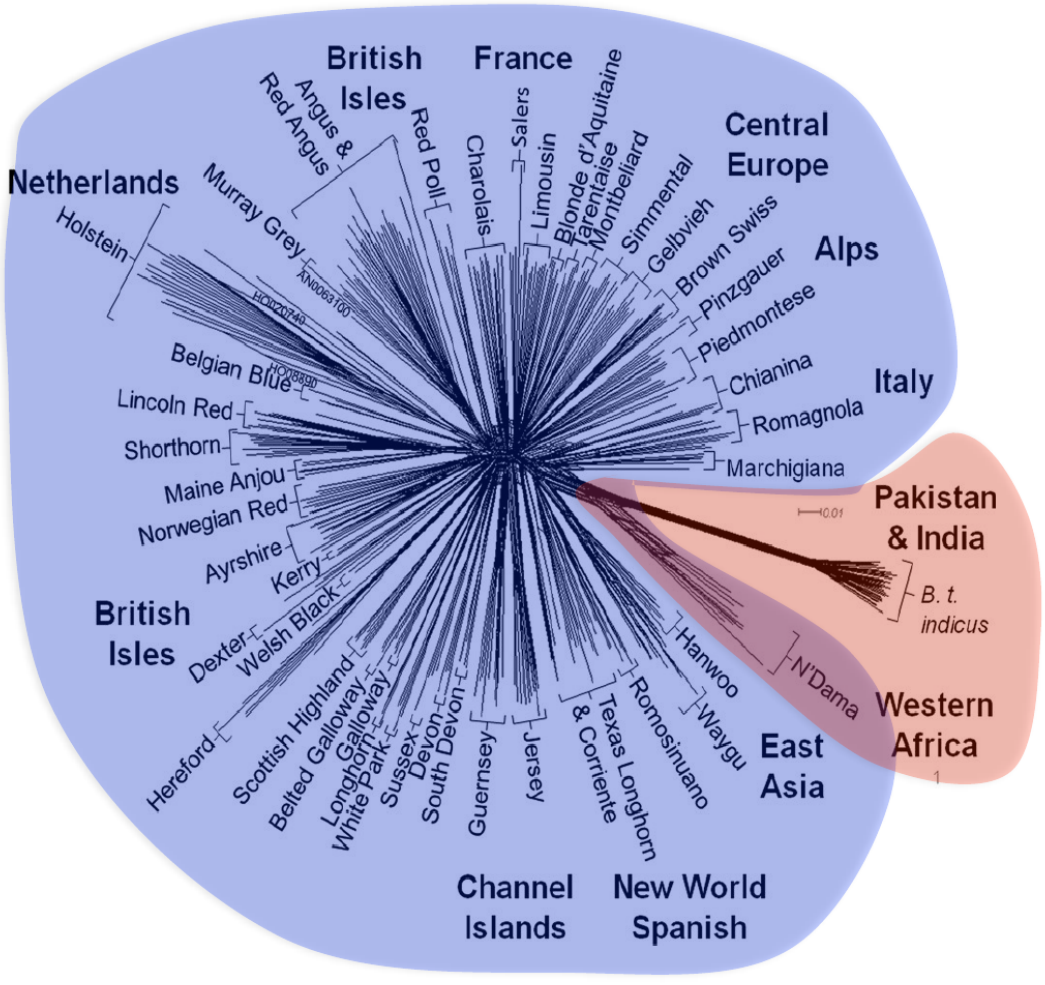
Effect of sample size on site frequency spectrum



The expected frequency of observed alleles of frequency X, using sample size of d = 2 (black), d = 5 (darkgrey), d = 10 (light grey) and d = 20 (white; no ascertainment bias). (Nielsen, 2004)

Singleton mutations are very important for estimating population sizes

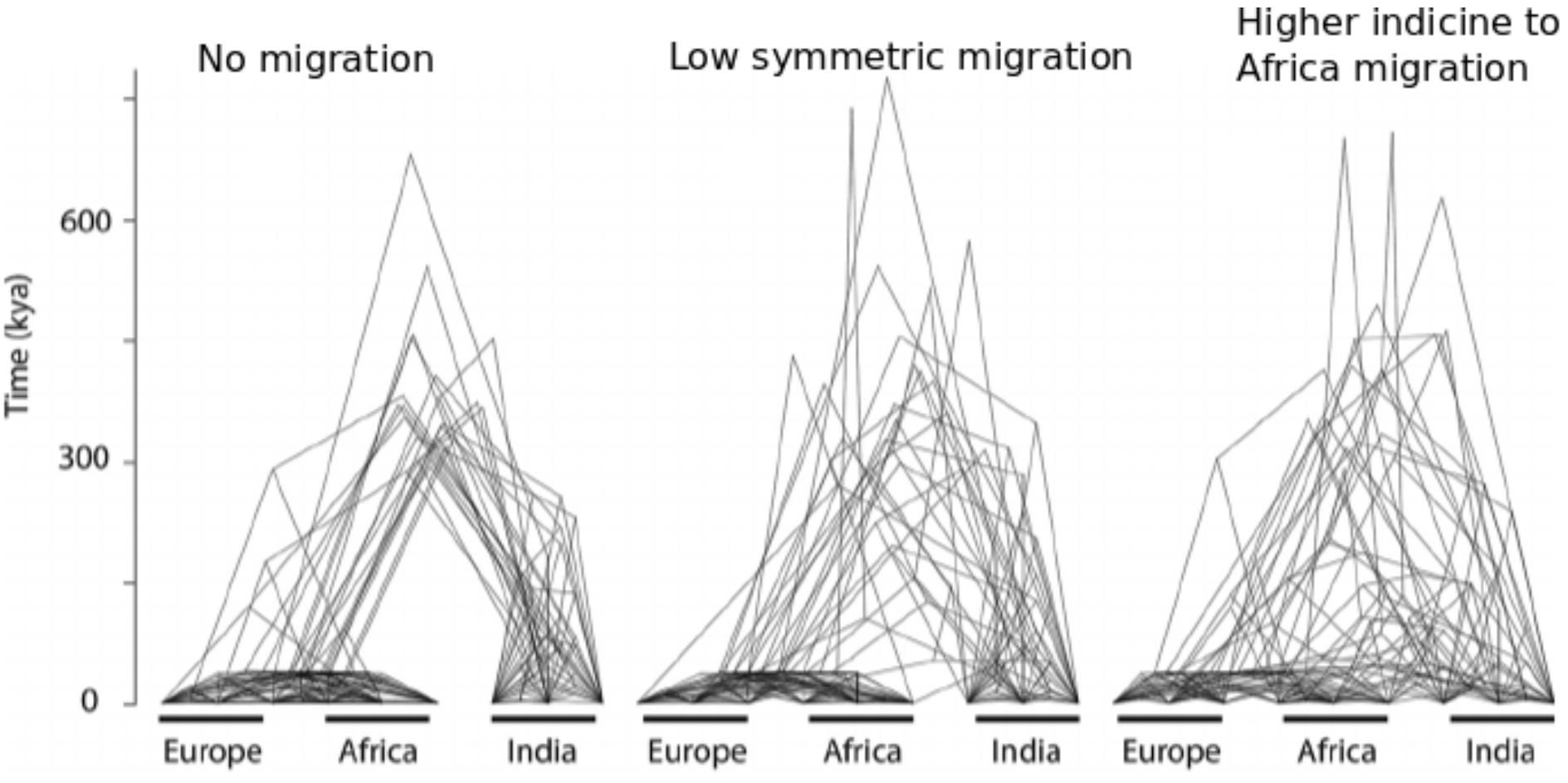
Subpopulation bias



Decker et al., 2009

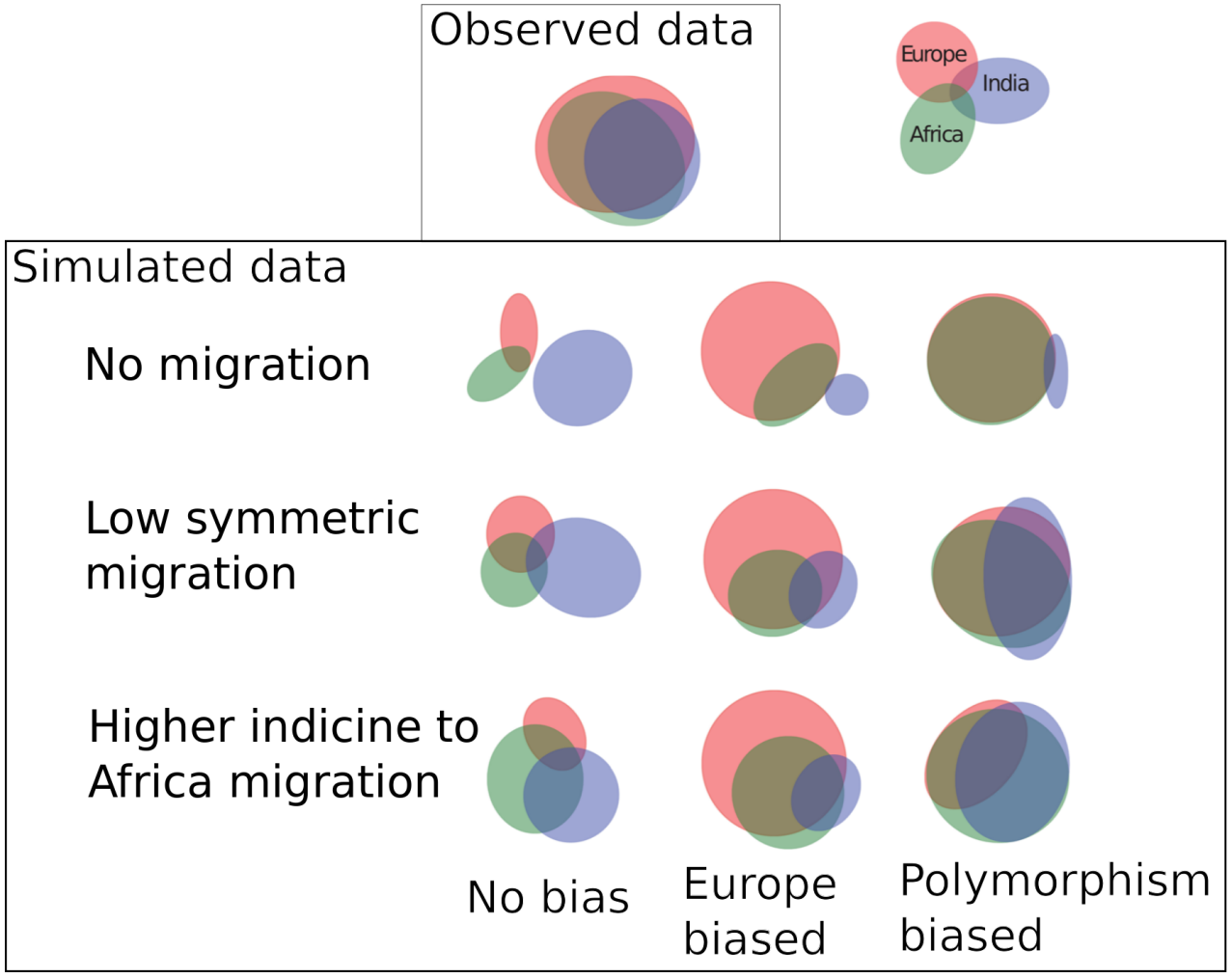
Approach

Simulations to test model fit based on population divergence and population size parameters from less biased data



McTavish and Hillis (2015)

SNP data is strongly biased toward shared polymorphisms

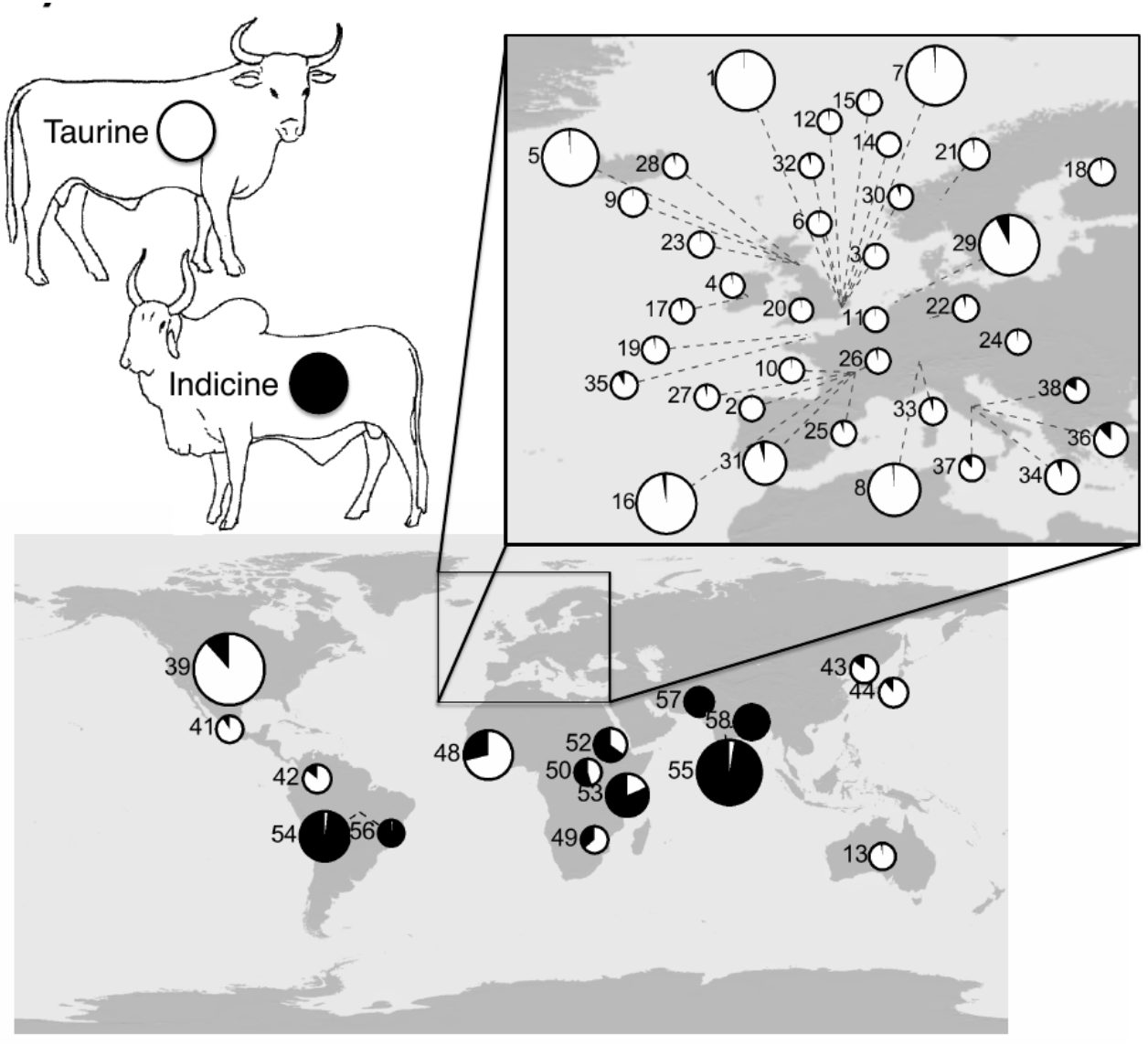


Approach

Performed analyses across simulated, biased data sets

- Ascertainment sample enriched for deep polymorphisms
- Population size estimates and branch lengths highly affected by bias
- Inference of hybridization using principal components analysis was not significantly affected, consistent with theory in McVean (2009)

Global hybridization patterns



McTavish et al. (2013)

Summary

Bias:

- Only variable sites sequenced
- Sites selected based on being variable in a subset of total population
- Uneven geographic sampling

Potential effect on inference:

- Many! Especially diversity and branch lengths under represented for some groups

Mitigation:

- Use parametric simulation and applying bias to simulated data to test effects on inference

Conclusions:

- Even strongly biased data can provide reliable estimates of some parameters

Case study II

Phylogenetics of Penstemon using RADseq data

Question

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

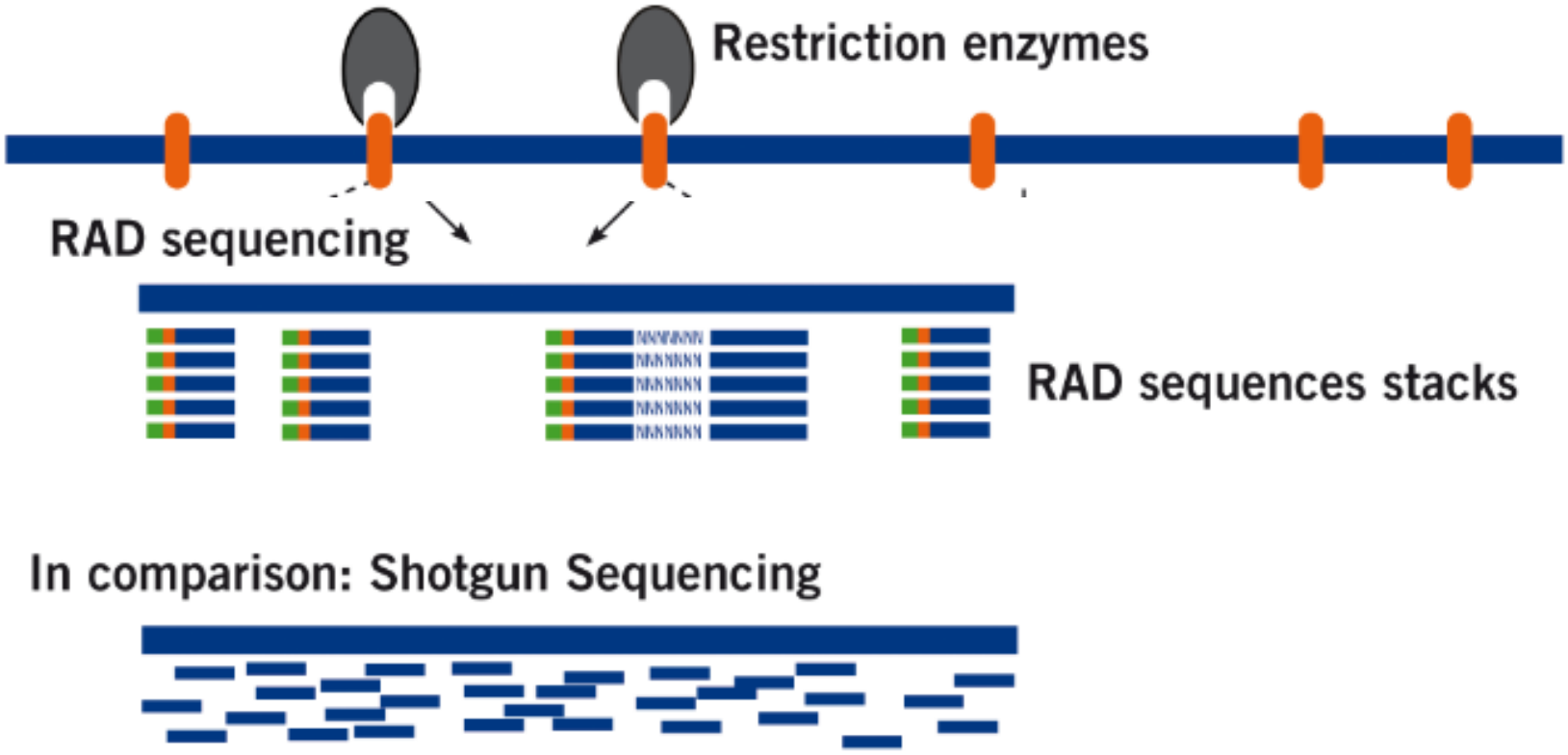
Data:

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

RADseq

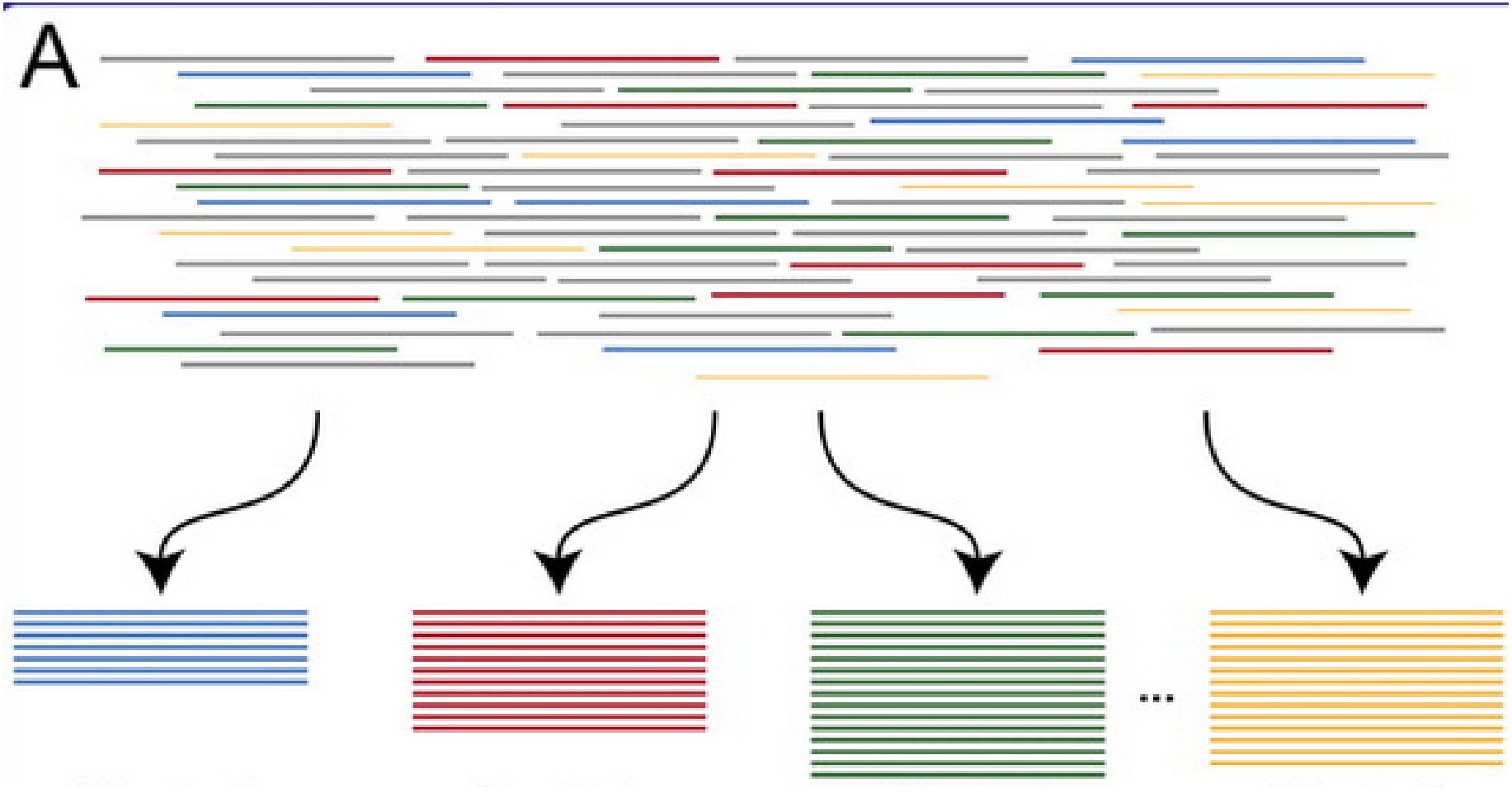
Uses restriction enzymes to fragment DNA

Targets sequencing to the same regions across taxa



(figures from floragenex.com)

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



Clustered using Stacks 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

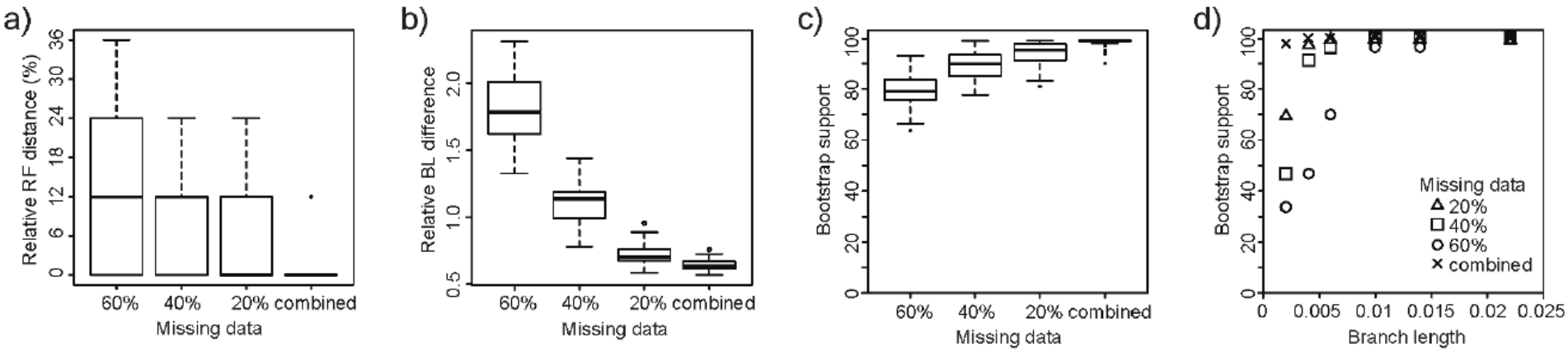


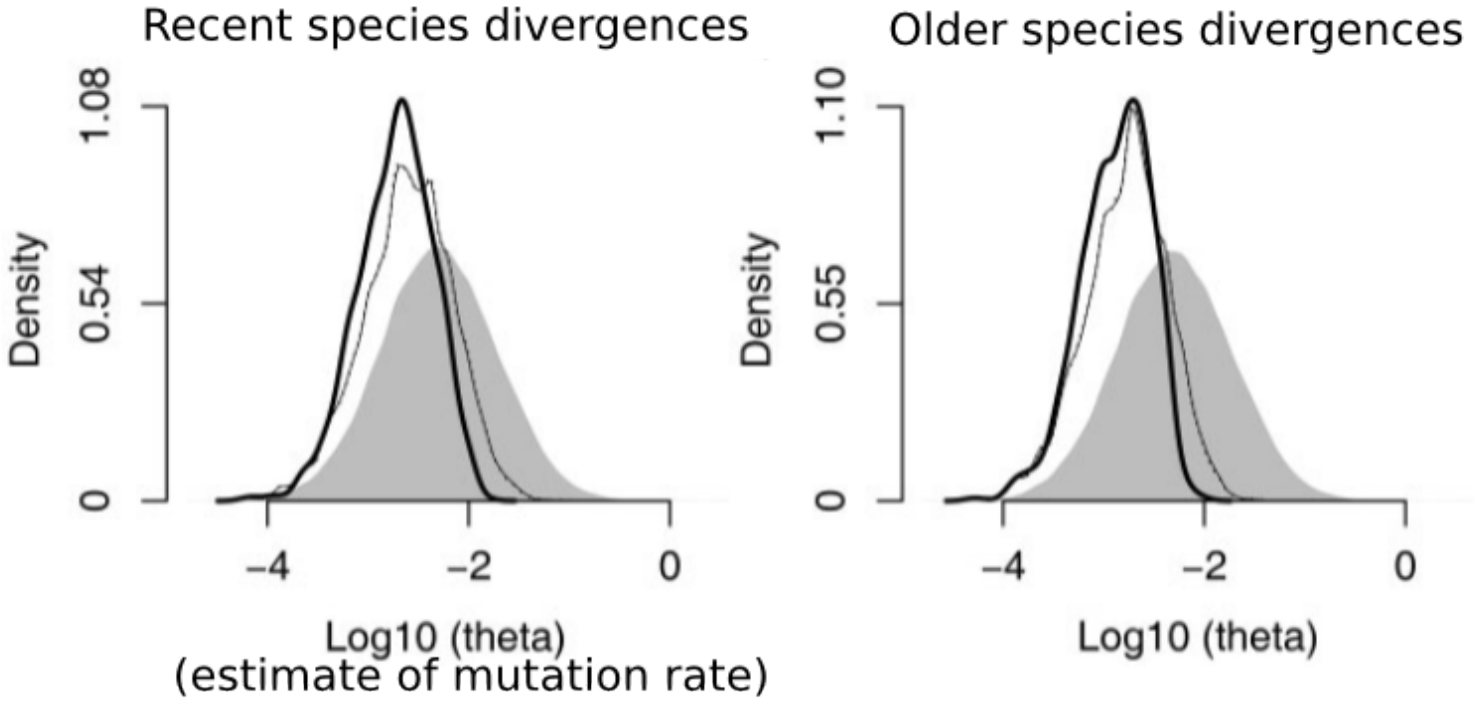
Figure 4: Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support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

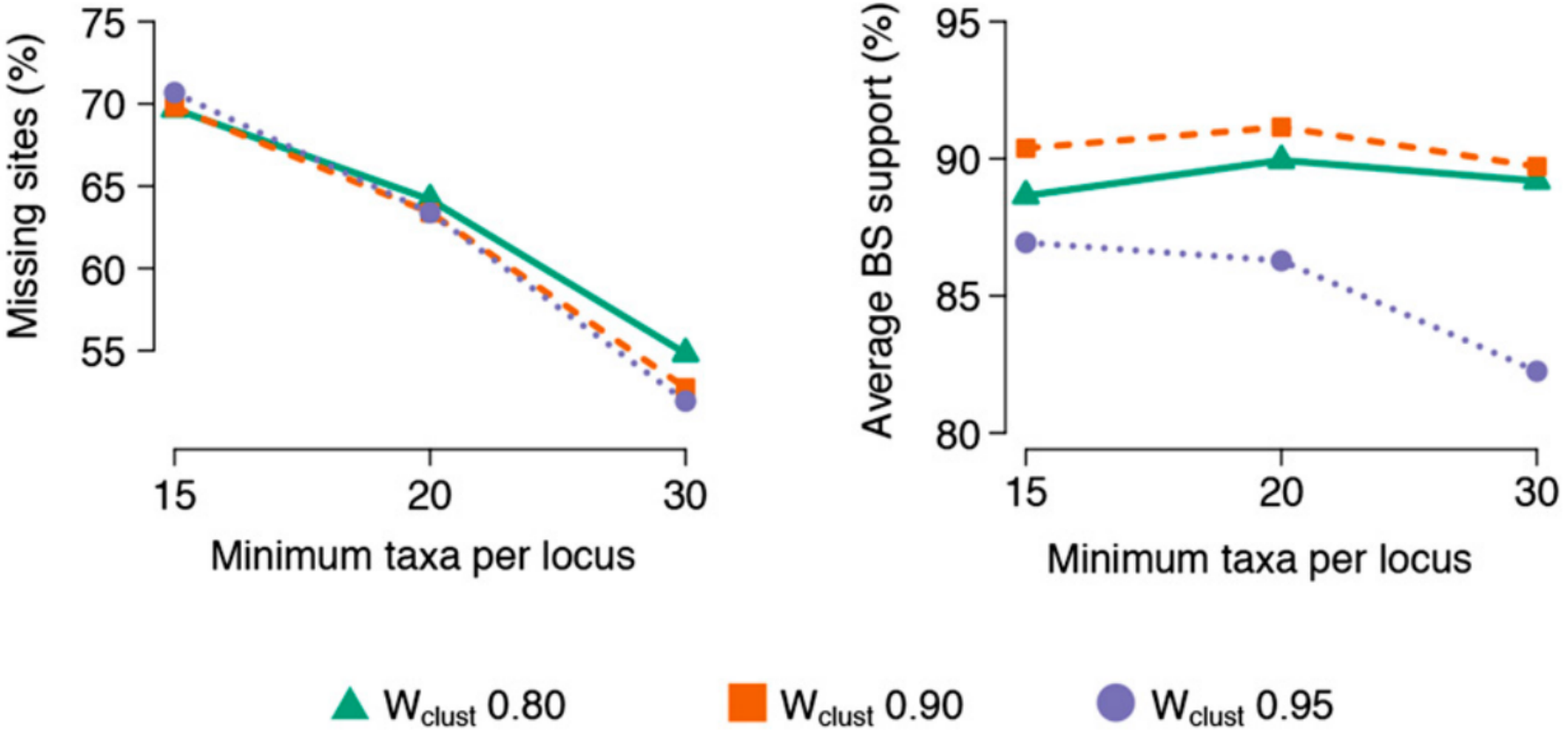
Huang and Knowles (2014)

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

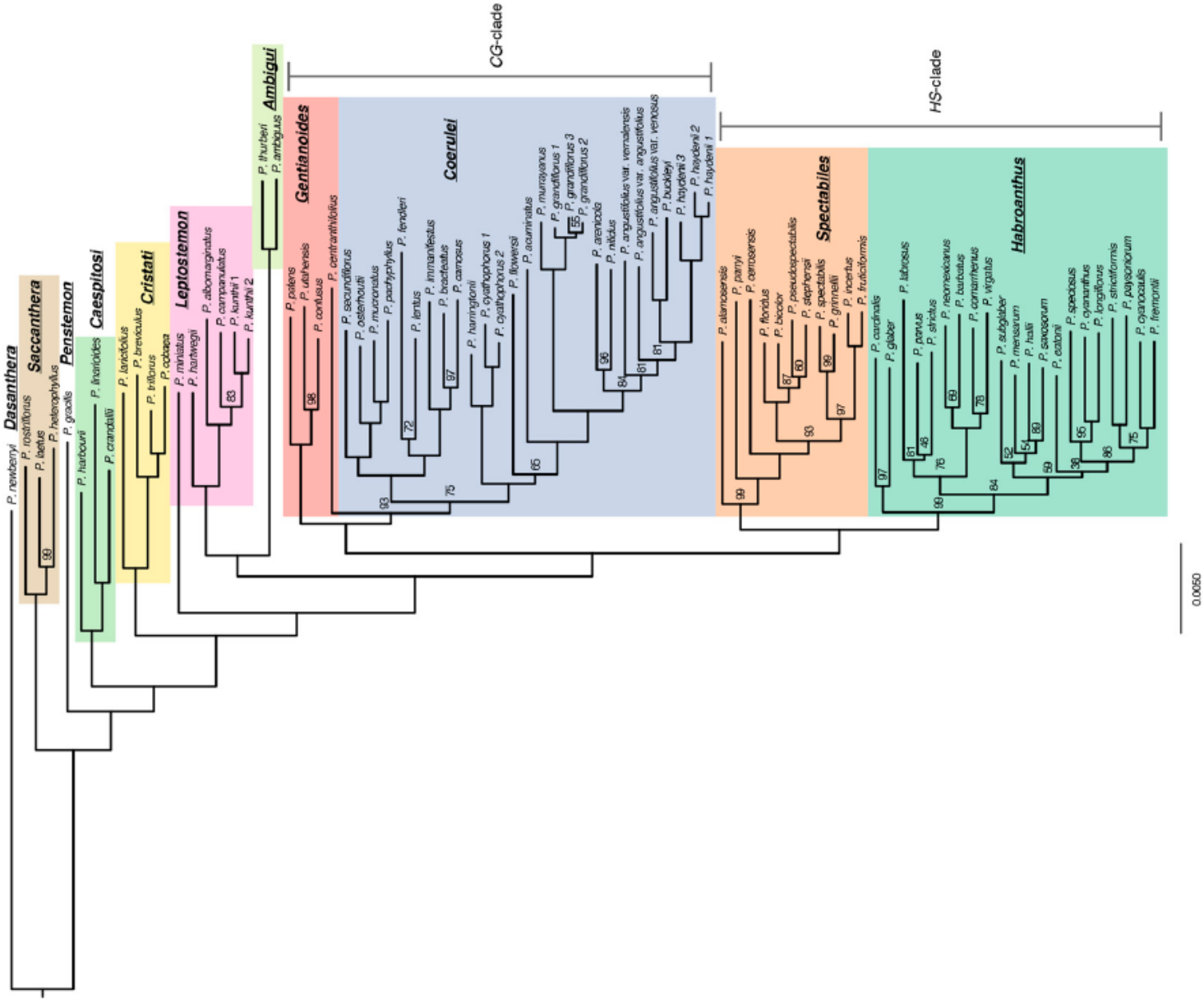
Approach

Investigate a range of parameters

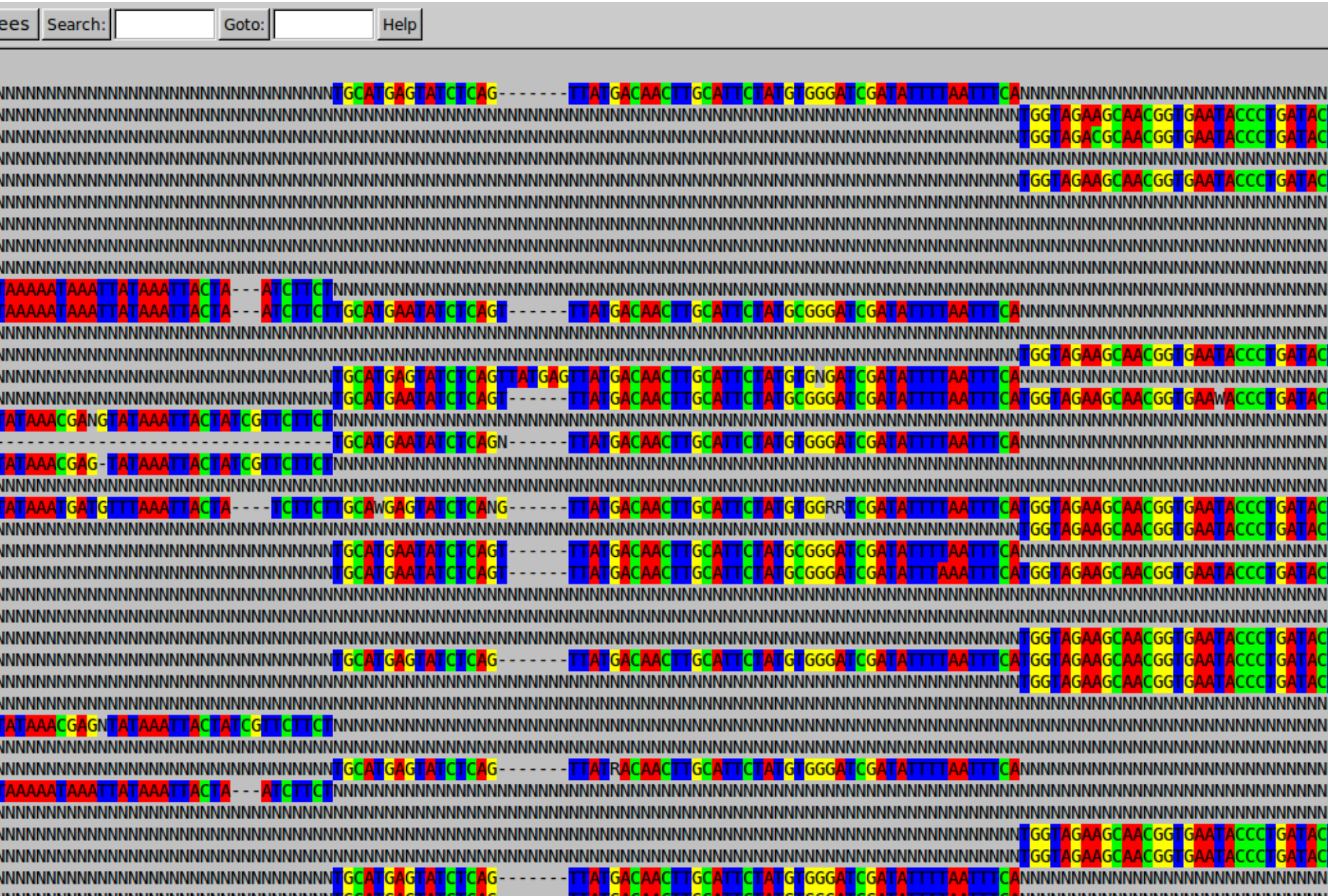


Wessinger et al. (2016)

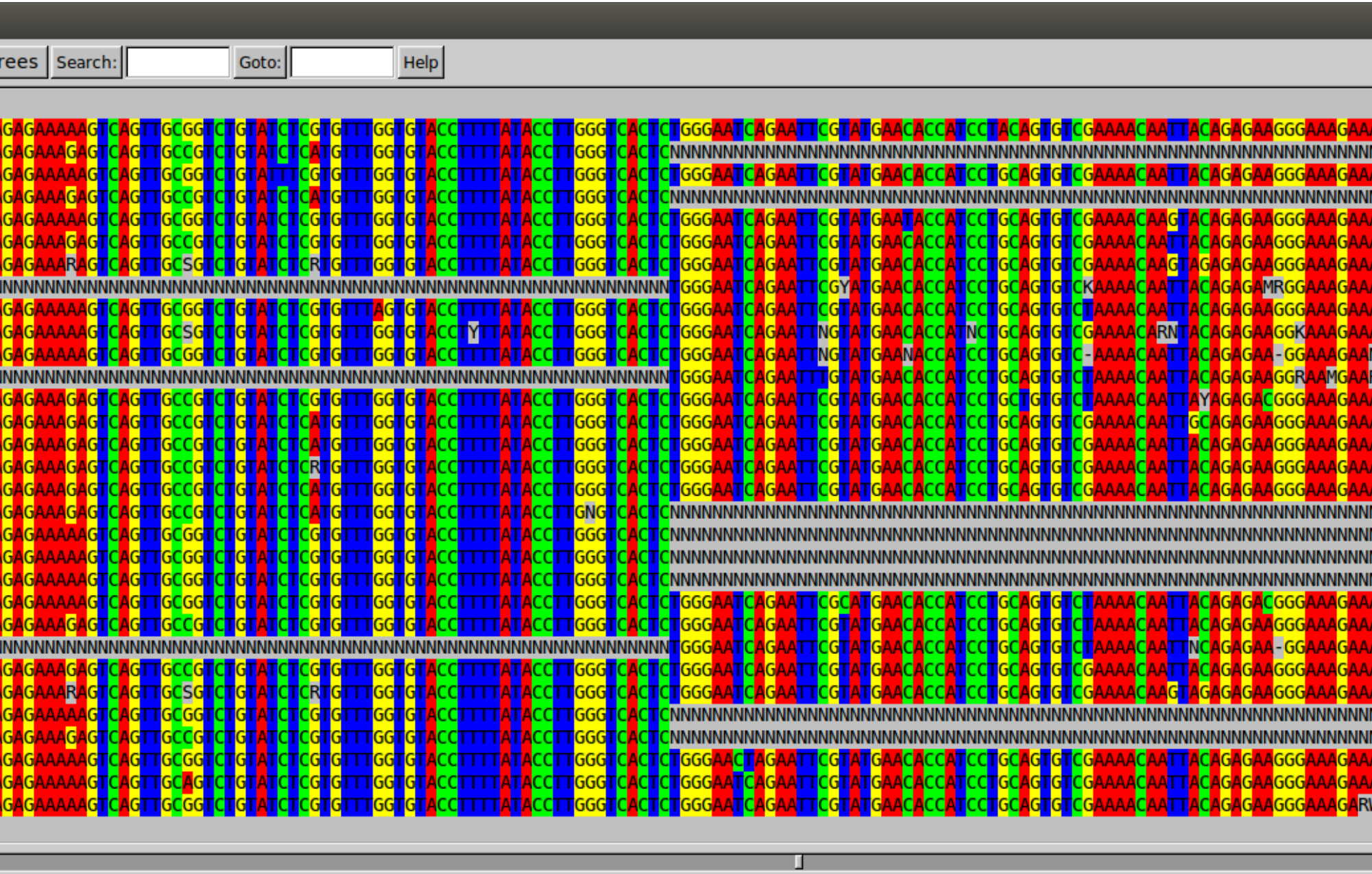
Missing data is phylogenetically biased



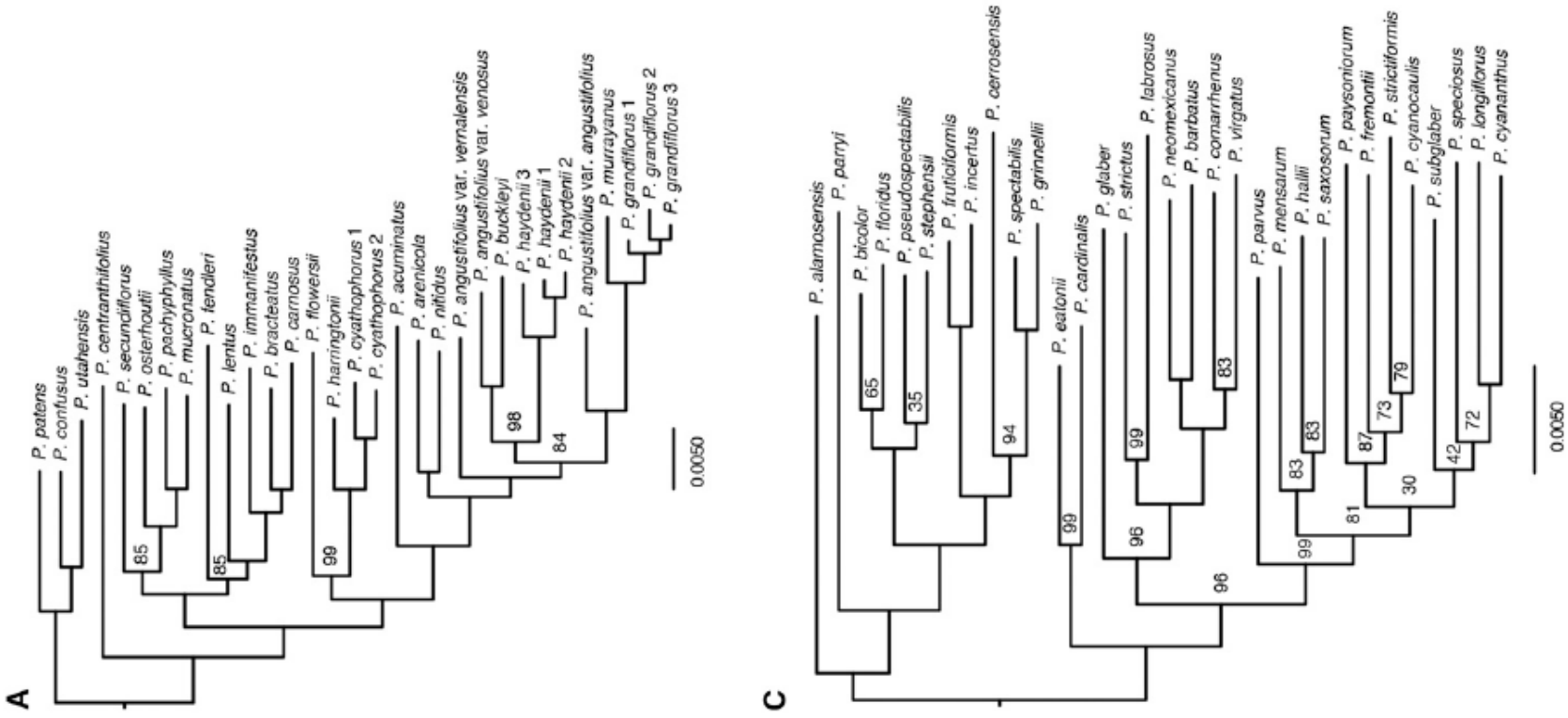
Across full dataset, many loci are only found in one of the major clades



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



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



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

Summary:

Bias:

Clustering parameters drive non-random missing data

Potential effect on inference:

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

Mitigation:

Estimate relationships under a range of filtering parameters

Conclusions:

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

Case study III

Whole genome sequencing of bacterial lineages

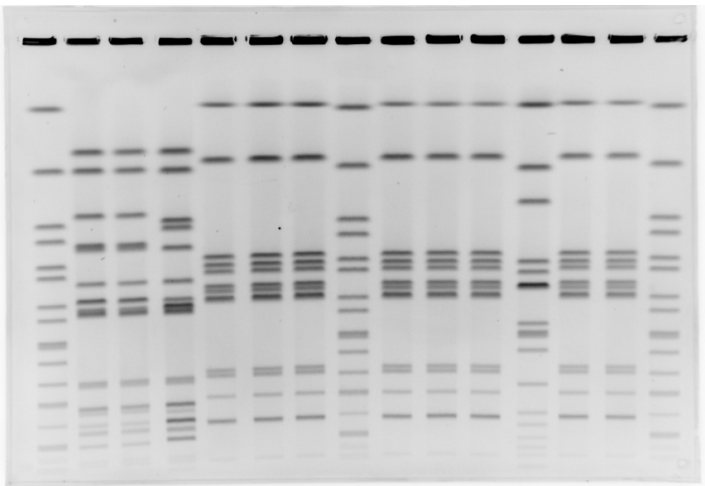
Question

How can we track and control food-borne pathogens?

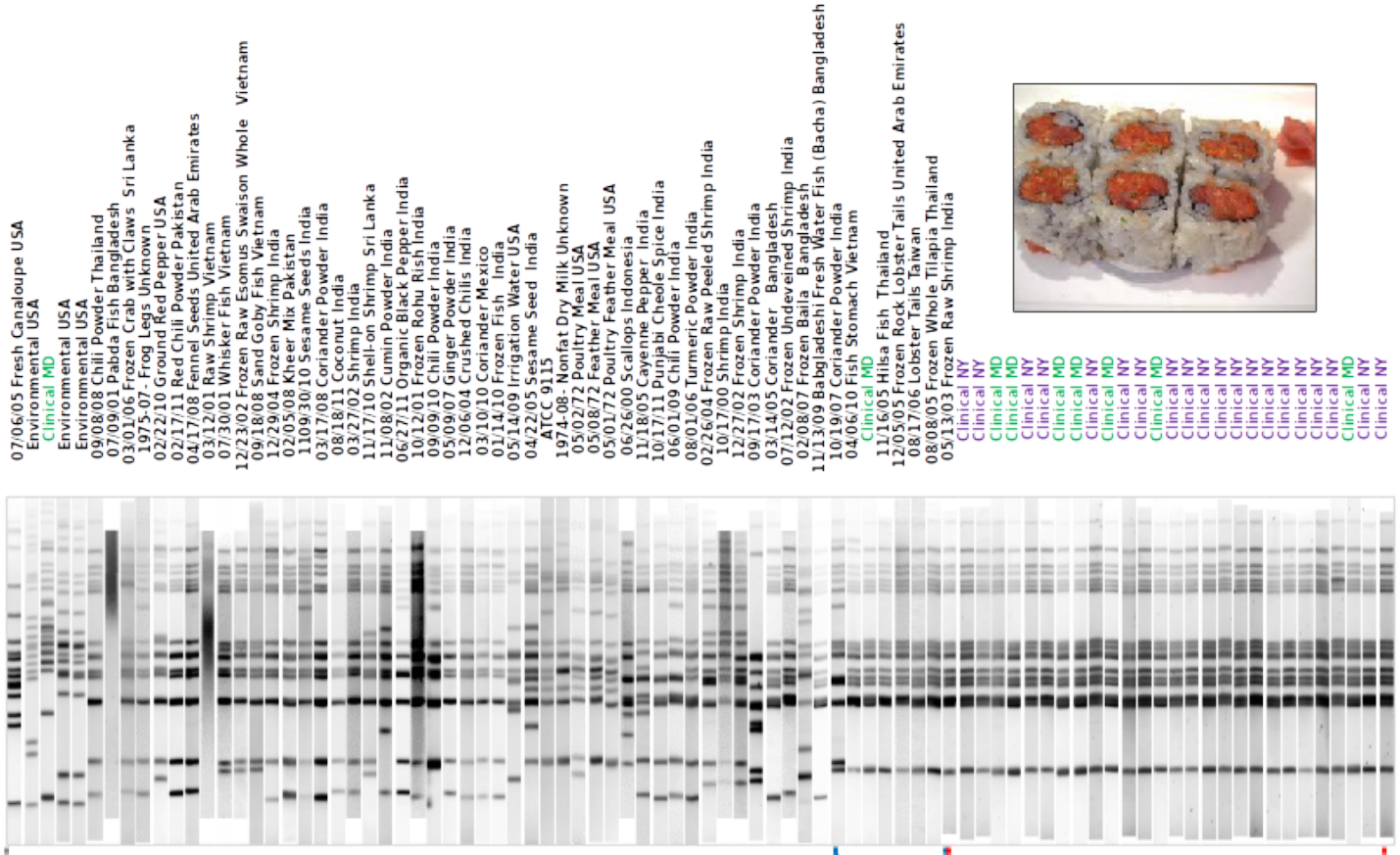
Traditional Food and Drug Administration approach to pathogen identification

Antigens are screened to identify serovar (bacterial variant)

PFGE: whole genome restriction digest + pulse field gel electrophoresis for discrimination within serovar



Salmonella Bareilly Outbreak April-June 2012



PFGE identical in red

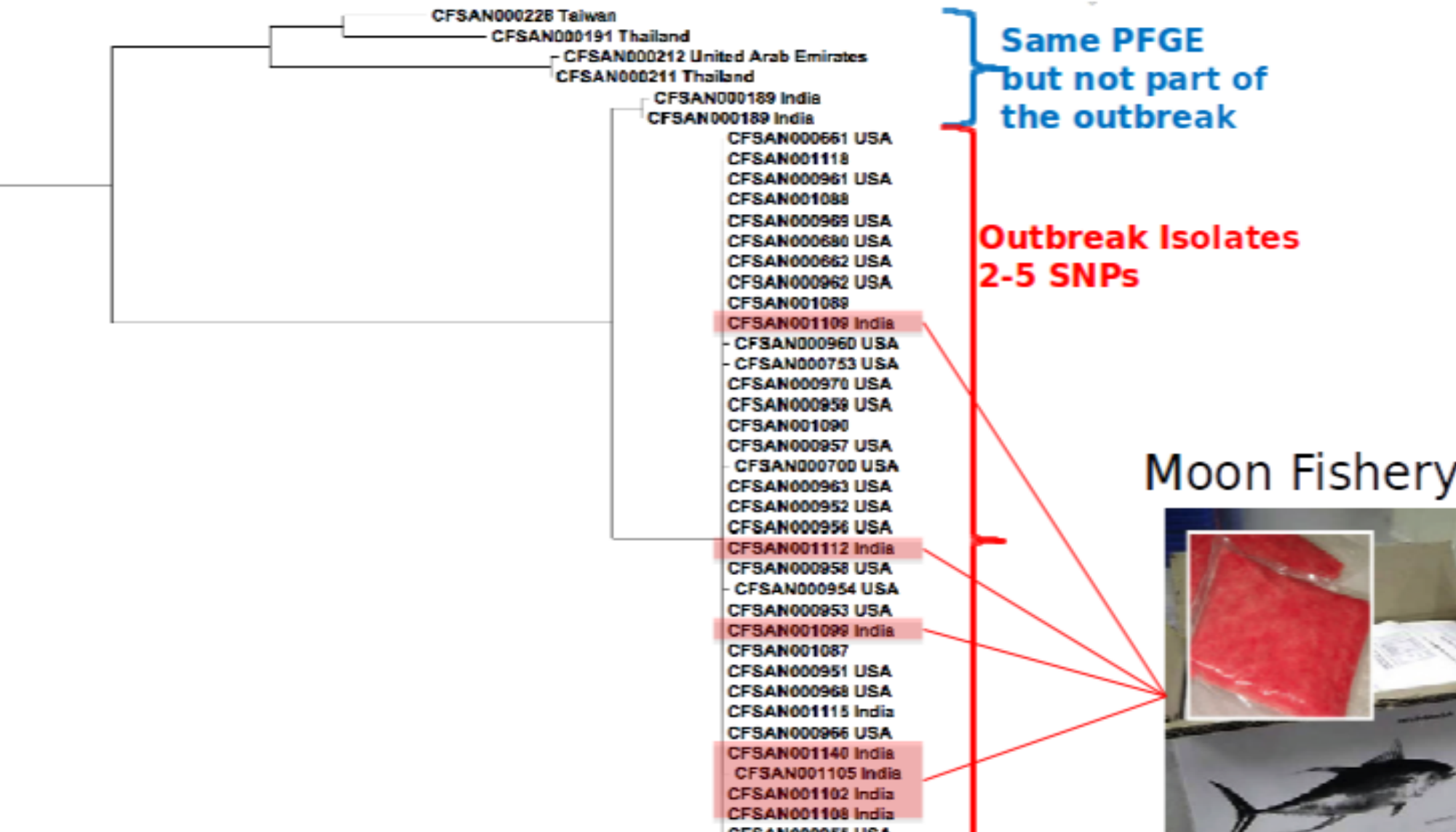
NGS distinguishes geographical structure among closely related *Salmonella* Bareilly strains

OFFICE OF REGULATORY SCIENCE

Data:

- Whole genome shotgun sequencing reads
- Clinical and test samples associated with *Salmonella* Bareilly outbreak
- Several options for reference genomes to map reads

Whole genome sequence data resolves these relationships



Concerns:

Speed from sampling $\rightarrow$ phylogeny important

Concerns:

Speed from sampling $\rightarrow$ phylogeny important

Need to rely on phylogenies for legal action (requires high confidence)

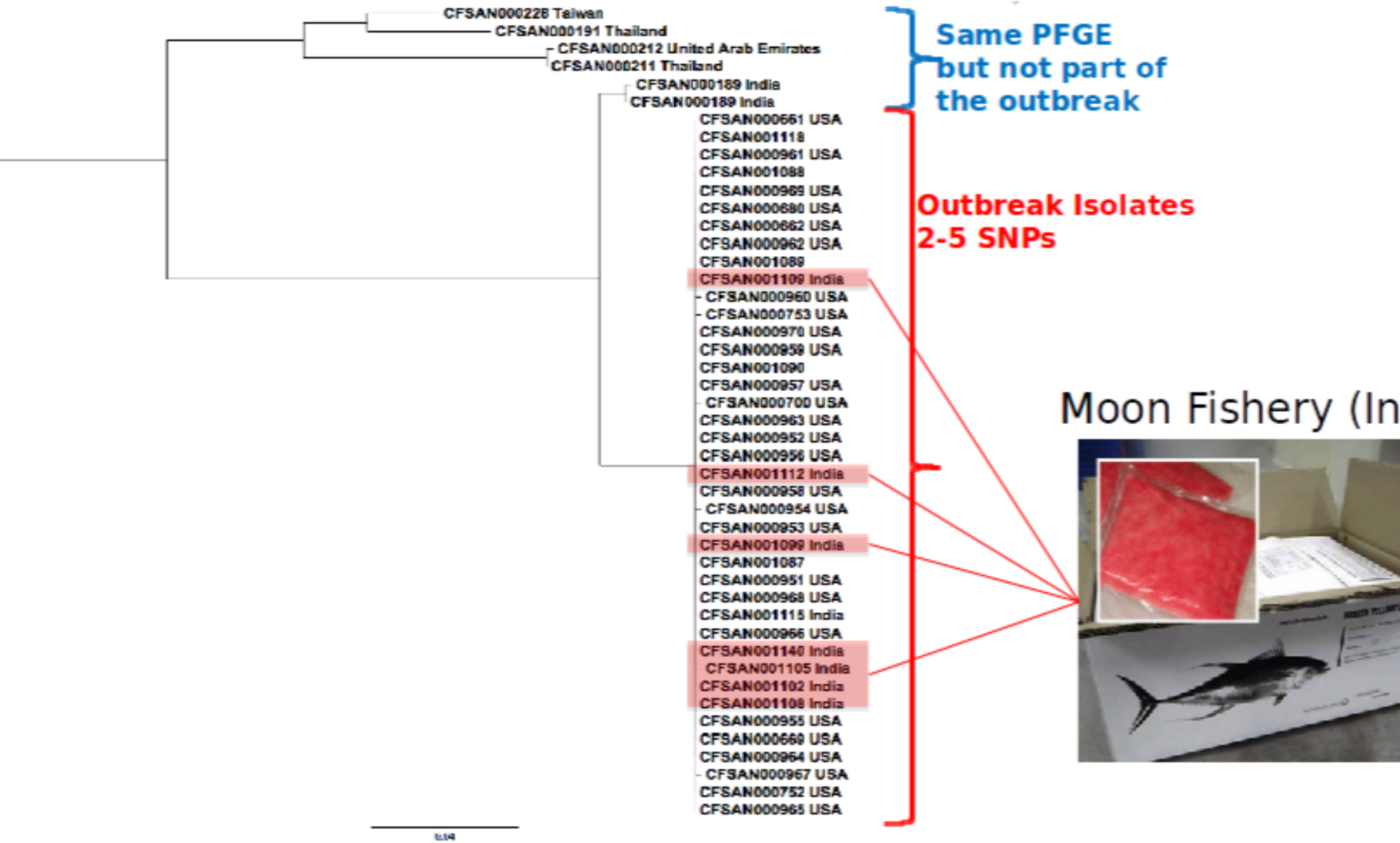
Concerns:

- Speed from sampling $\rightarrow$ phylogeny important
- Need to rely on phylogenies for legal action (requires high confidence)
- No/little *a priori* taxonomic information, reference genomes not necessarily available

Concerns:

- Speed from sampling $\rightarrow$ phylogeny important
- Need to rely on phylogenies for legal action (requires high confidence)
- No/little *a priori* taxonomic information, reference genomes not necessarily available
- Often very little variability (e.g. 20-200 variable sites in a 4.5 Mb genome) among outbreak strains and closely related non-pathogenic strains - BUT variable sites clustered in genome

20 mutations across a 4,730,612 bp base pair genome



Potential issues:

- Effect of choice of reference genome
- Sequencing error

Reference choice can affect inference

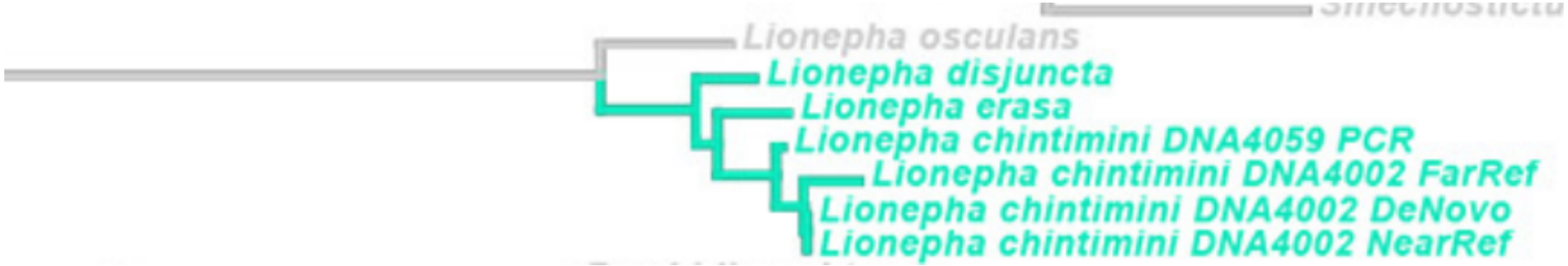
Reference choice can affect inference

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

Reference choice can affect inference

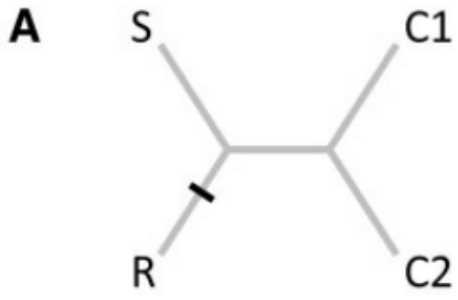
In humans, in highly polymorphic regions variant calling is biased toward the the reference base (Brandt et al., 2015)

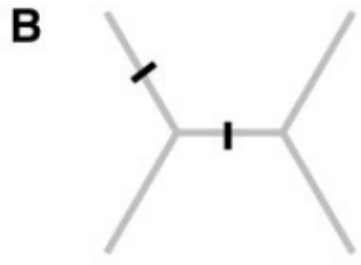
Changes branch lengths when assembling fragmented DNA from preserved beetles

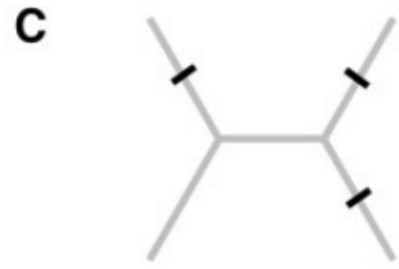


Kanda et al. (2015)

Mapping sequencing reads to reference genomes also requires similarity cutoffs that generate biased missing data





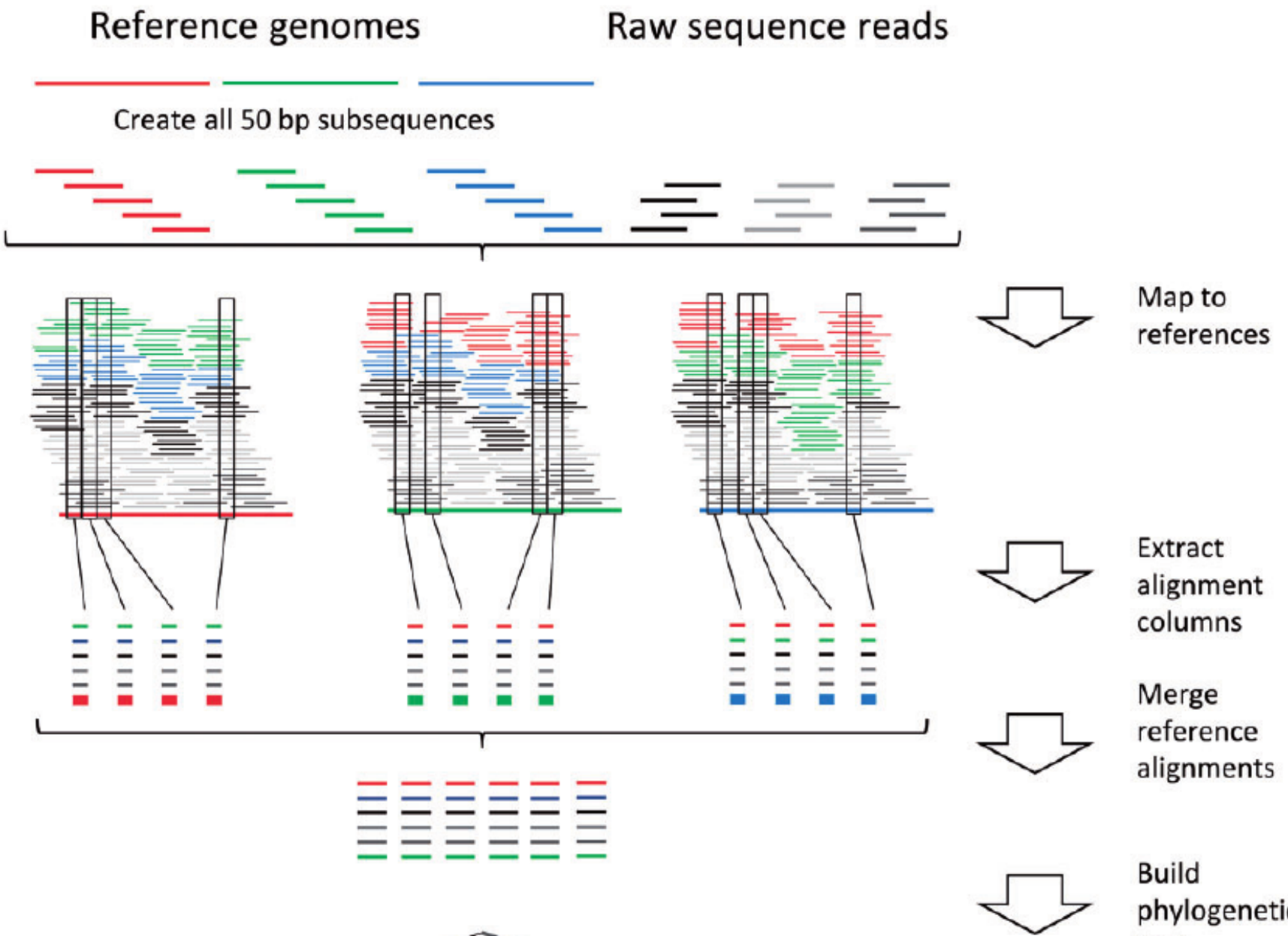



```
Reference: ATATGCAGTTACAGTACACTA   ATATGCAGTAACAGTACACTA   ATATGCAGTAACAGTACACTA
Sister:    ATATGCAGTAACAGTACACTA   ATATGCAGTTACAGTACACTA   ATATGCAGTTACAGTACACTA
Cousin 1:  ATATGCAGTAACAGTACATAG   ATATGCAGTAACAGAACACTA   ATATGCAGTAAGAGTACACTA
Cousin 2:  ATATGCAGTAACAGTACATAG   ATATGCAGTAACAGAACACTA   ATATCCAGTAACAGTACACTA
```

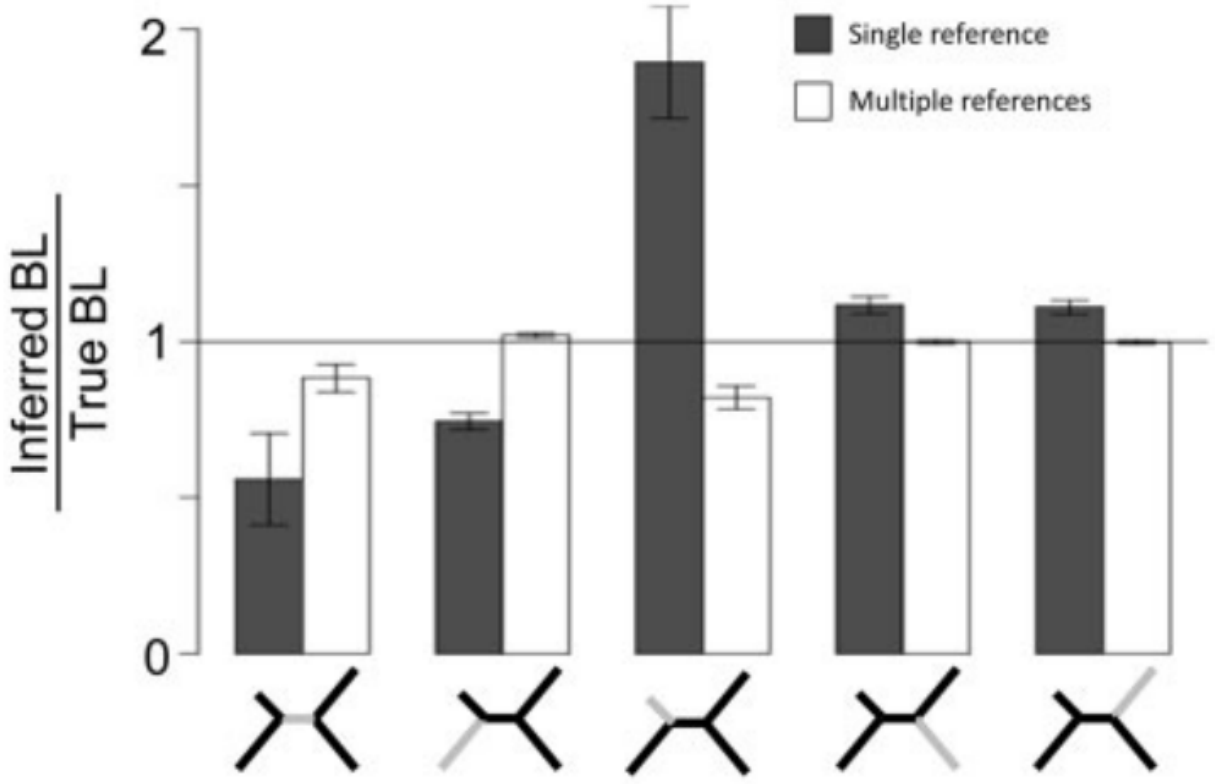
Reads with more mutations than the cutoff will not map, and will be discarded

Bertels et al. (2014)

Bertels et al. (2014) suggest ameliorating reference bias by using multiple reference genomes to map reads



Using multiple references improves branch length inference



(but these are simulated on trees with very long and very skewed branch lengths) Bertels et al. (2014)

Sequencing error

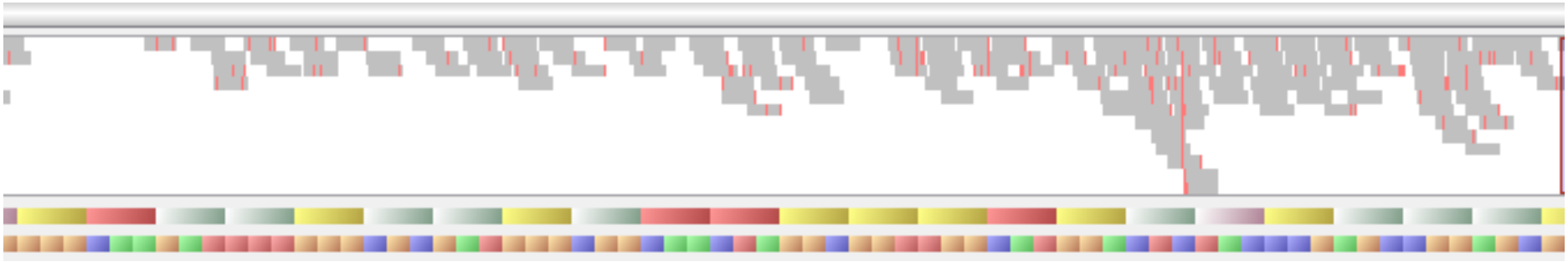
Potentially problematic when real variable sites are rare

Sequencing errors are likely to be singletons

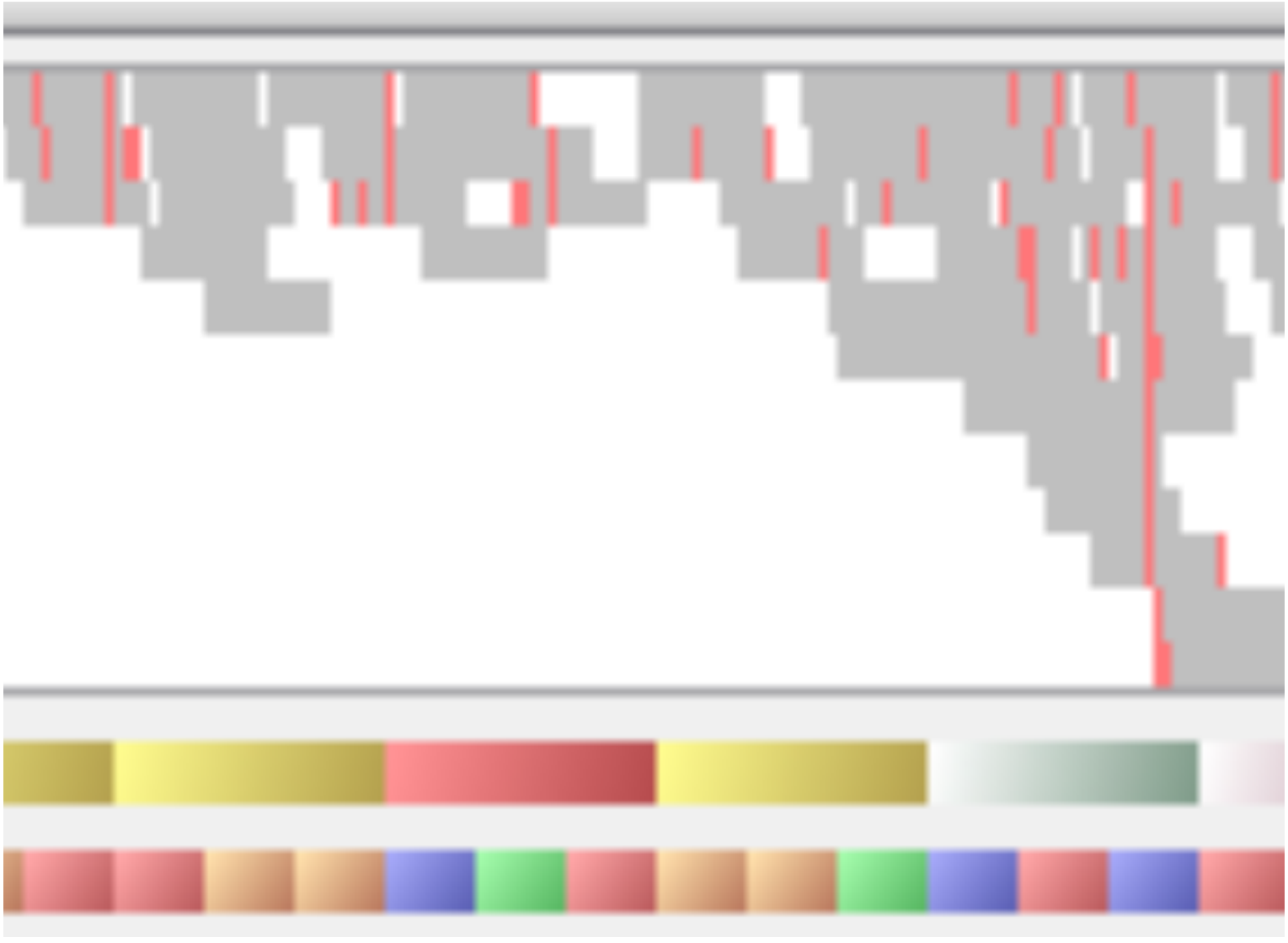
Will overestimate tip branch lengths

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

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



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



Not currently implemented in phylogenetic likelihood models

Approach

Mapped reads to several reference options (both sequentially and using REALPHY)

Simulated short reads on observed phylogeny and tested if outbreak clade could be recovered

- Phylogeny and model of evolution (using Seq-Gen)
- Distribution of mutations across the genome
- Read coverage
- Observed sequencing error profiles (Using ART)
- Read simulation pipeline on GitHub
  `https://github.com/snacktavish/TreeToReads`

Results

In *Salmonella* Barielly test case, using 3 reference genomes, REALPHY found no variable sites in outbreak and related lineages

Simulated data mapped to reference within outbreak clade recovered the correct phylogeny reliably at coverages as low as 5x

No incorrect base calls due to sequencing error, even at low coverage

Summary

Bias: Reference, sequencing error

Effect on inference:

- Not mapping reads on lineages more distant from reference genome will decrease those branch lengths
- Sequencing error can increase terminal branch lengths relative to internal branches

Mitigation: Used multiple reference genomes, and parametric bootstrapping to assess robustness

Conclusions:

- When a closely related reference in available, alternatives worsen inference
- At high (around 40x) coverage all mutations are confidently recovered
- Even at lower coverage (around 5x) high confidence in monophyly of outbreak clade

Conclusions

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

Careful project planning helps

Interrogate potential biases in data sets

What to do?

How do you determine if ascertainment biases are affecting your inference?

- 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?

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.

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.

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.

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

Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., Muruvanda, T., Wang, C., Kastanis, G., Keys, C., Janies, D., Senturk, I. F., Catalyurek, U. V., Wang, H., Hammack, T. S., Wolfgang, W. J., Schoonmaker-Bopp, D., Chu, A., Myers, R., Haendiges, J., Evans, P. S., Meng, J., Strain, E. A., Allard, M. W., and Brown, E. W. (2016). Tracing Origins of the Salmonella Bareilly Strain Causing a Food-borne Outbreak in the United States. *Journal of Infectious Diseases*, 213(4):502–508.

Huang, H. and Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology*, page syu046.

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*, page 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.

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

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., Decker, J. E., Schnabel, R. D., Taylor, J. F., and Hillis, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences*, 110(15):E1398–E1406.

McTavish, E. J. and Hillis, D. M. (2015). How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics*, 16(1):266.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686.

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.

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.