Phylogenetic Biology
Week 7

Biology 1425

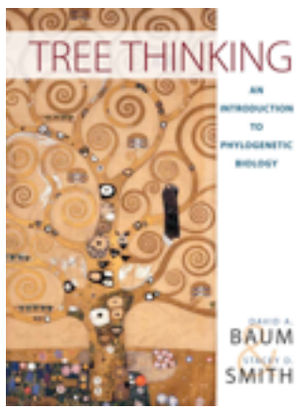
Professor: Casey Dunn, dunnlab.org
Brown University

2013

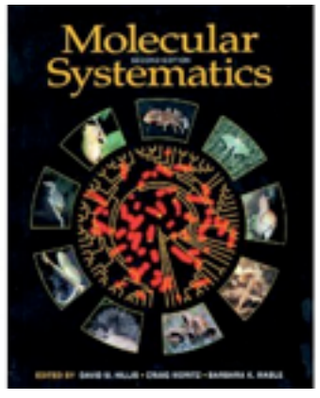
Front matter...

Sources

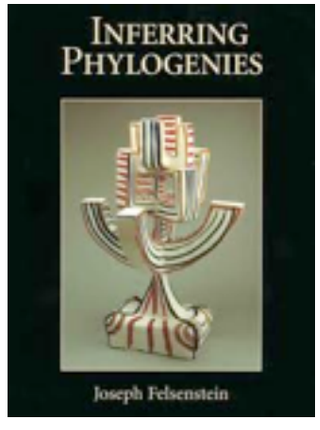
Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165



Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In: Molecular Systematics, Second Edition. eds: D. M. Hillis, C Moritz, & B. K. Mable. Sinauer Associates. ISBN 9780878932825



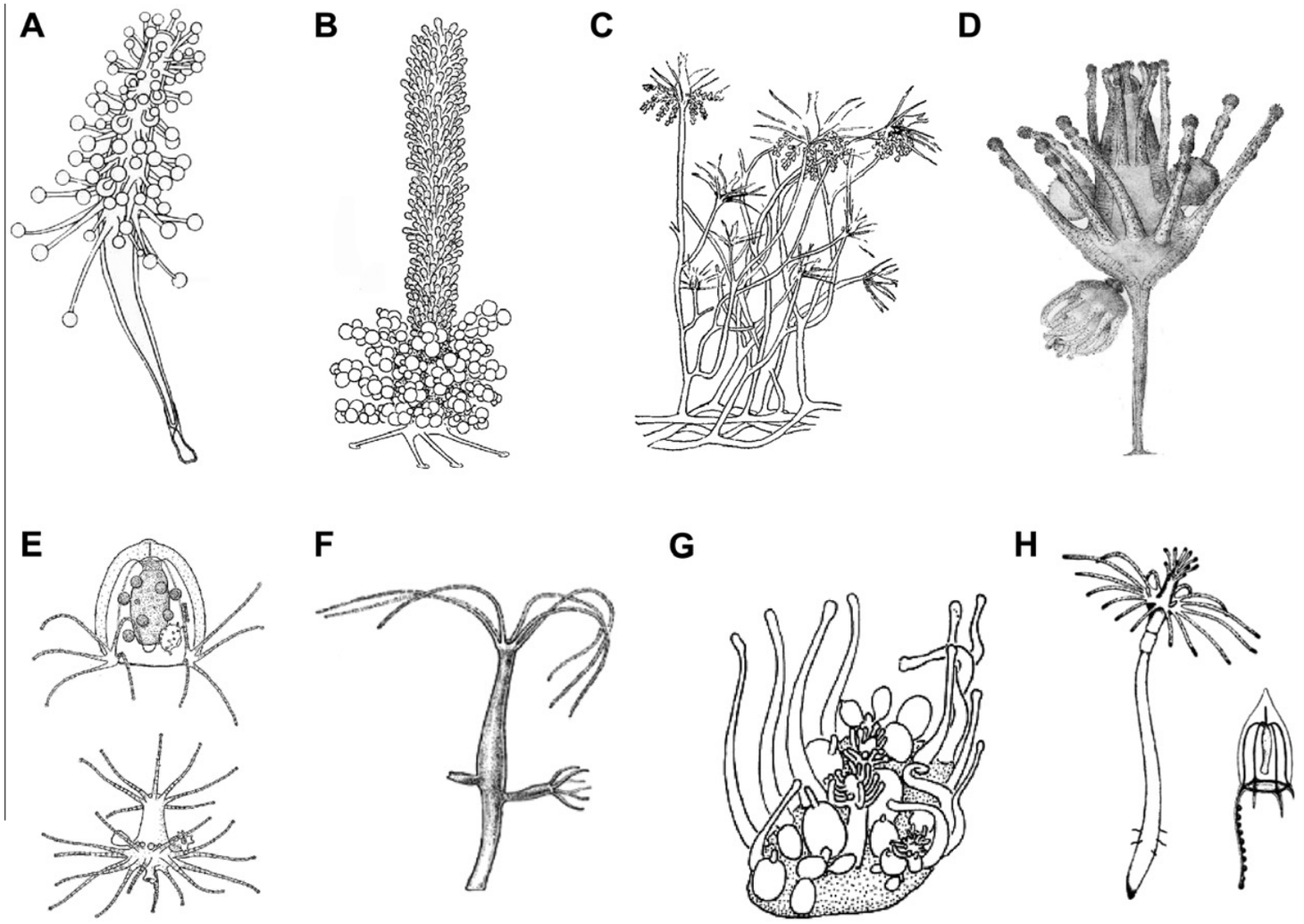
Felsenstein, J. (2003) Inferring Phylogenies. Sinauer Associates. ISBN 978-0878931774

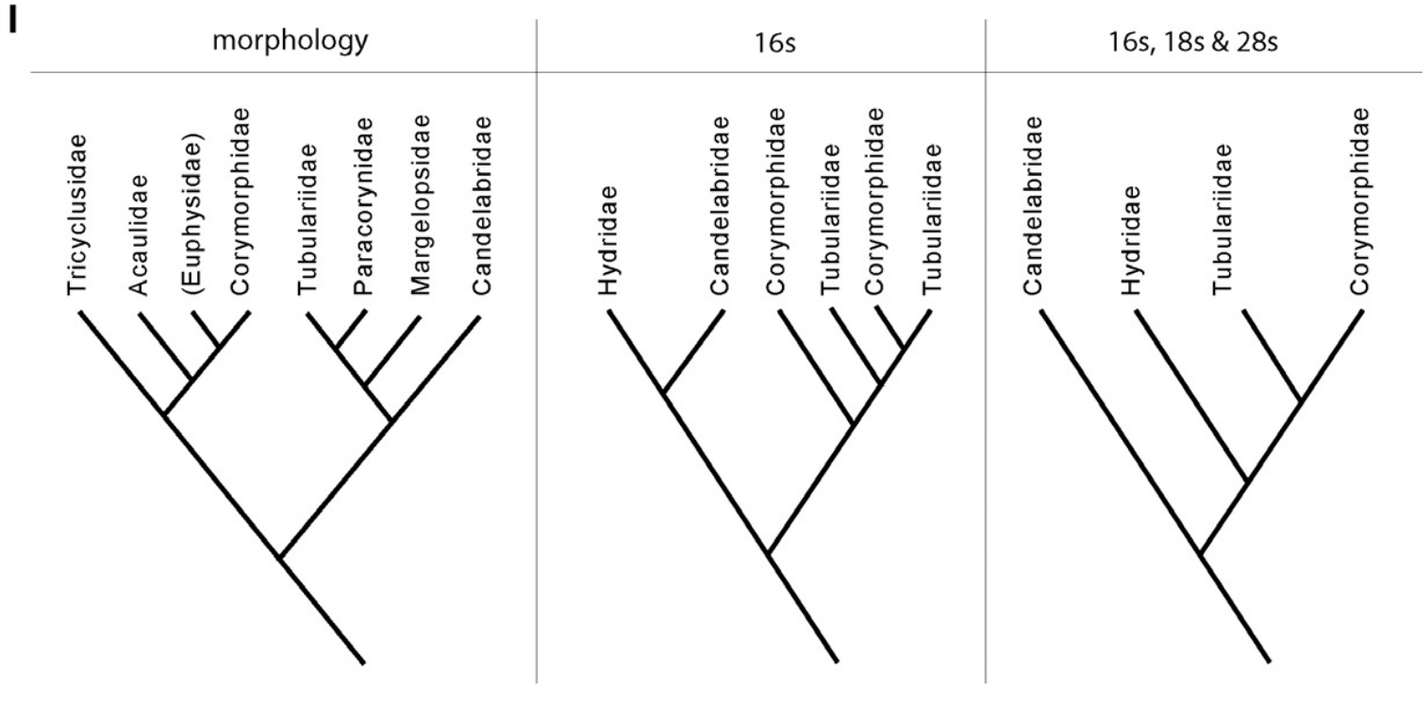
Other non-original content is referenced by url.

Phylogenetics in the wild- Project design and the challenges that arise when executing a study

An example study

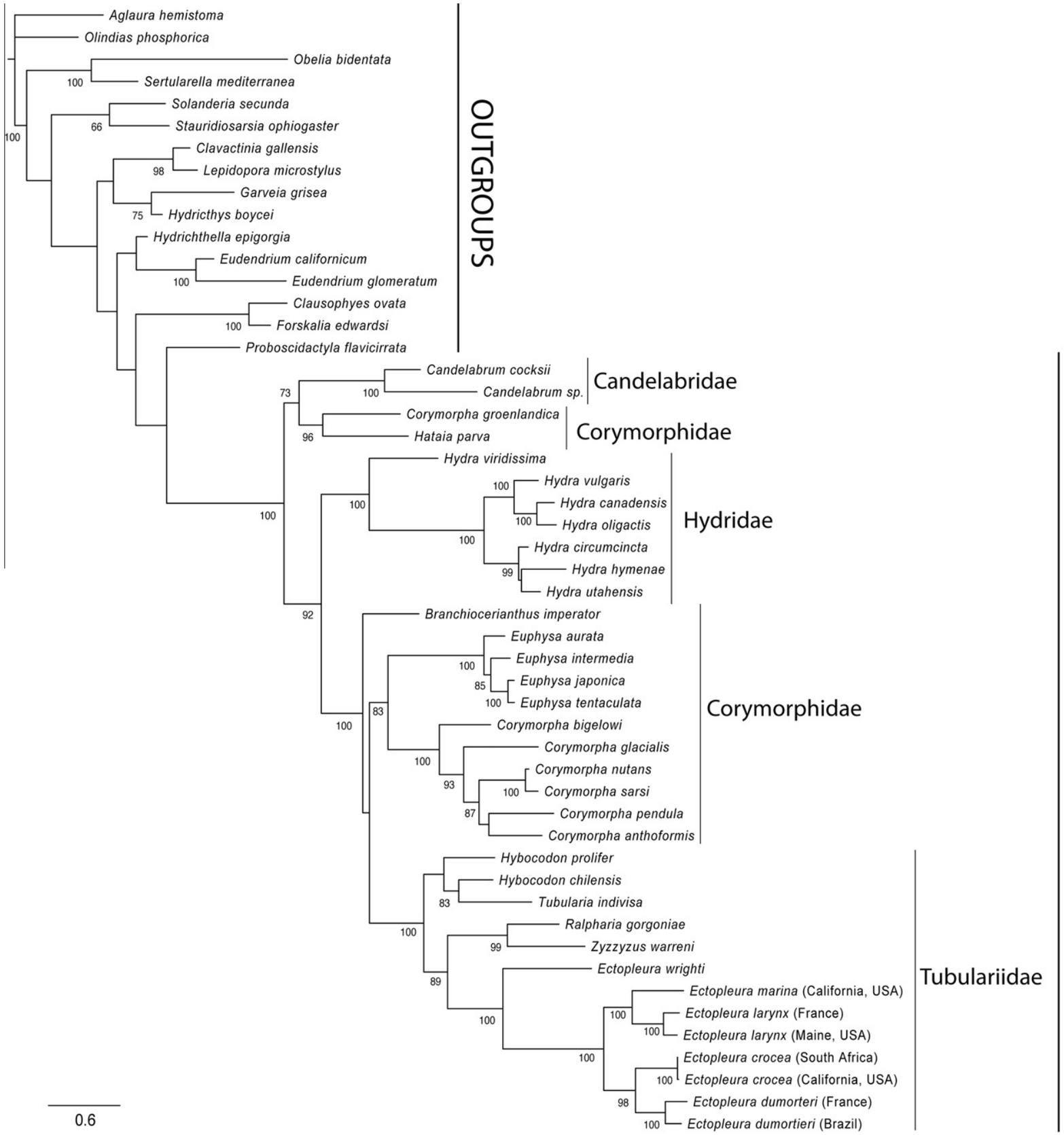
Nawrocki, A. M., Collins, A. G., Hirano, Y. M., Schuchert, P. & Cartwright, P. **Phylogenetic placement of Hydra and relationships within Aplanulata (Cnidaria: Hydrozoa)**. Molecular Phylogenetics and Evolution 67, 60–71 (2013). http://dx.doi.org/10.1016/j.ympev.2012.12.016





Nawrocki et al 2013, Figure 1

Bootstrap, all markers



Nawrocki et al 2013, Figure 3

How important is taxon sampling?

How important is taxon sampling?

The better your taxon sampling, the more questions you can address.

How important is taxon sampling?

The better your taxon sampling, the more questions you can address.

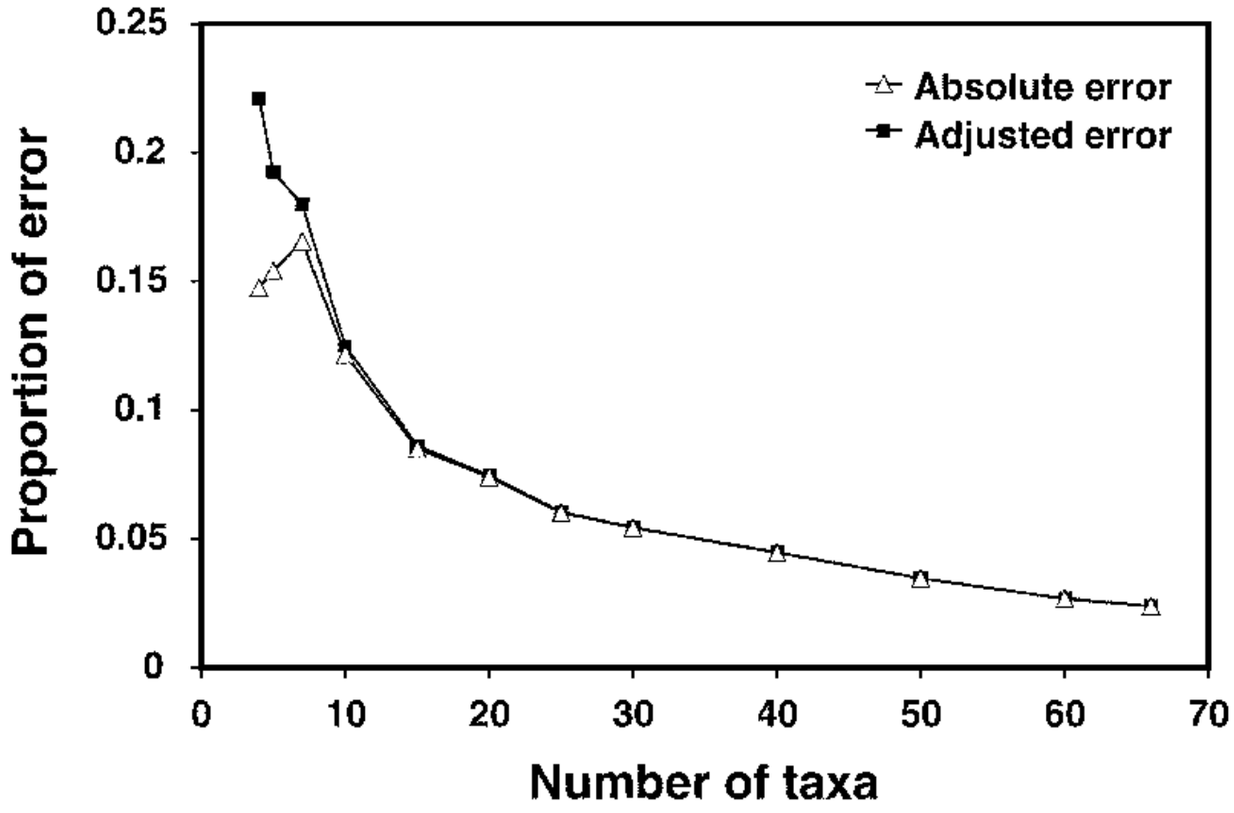
Better taxon sampling can improve accuracy

Syst. Biol. 51(4):588–598, 2002
DOI: 10.1080/10635150290102339

Increased Taxon Sampling Greatly Reduces Phylogenetic Error

DERRICK J. ZWICKL AND DAVID M. HILLIS

Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, Texas 78712, USA; E-mail: zwickl@mail.utexas.edu and dhillis@mail.utexas.edu



Zwickl and Hillis 2002,
http://dx.doi.org/10.1080/10635150290102339

How important are outgroups?

How important are outgroups?

Poor outgroup sampling is one of the biggest rookie mistakes in phylogenetic analyses

How important is character sampling?

How important is character sampling?

Better character sampling can improve accuracy

How important is character sampling?

Better character sampling can improve accuracy

Better character sampling can provide more rigorous evaluation of hypotheses

How does missing data impact phylogenetic analyses?

Missing data leads to loss of signal.

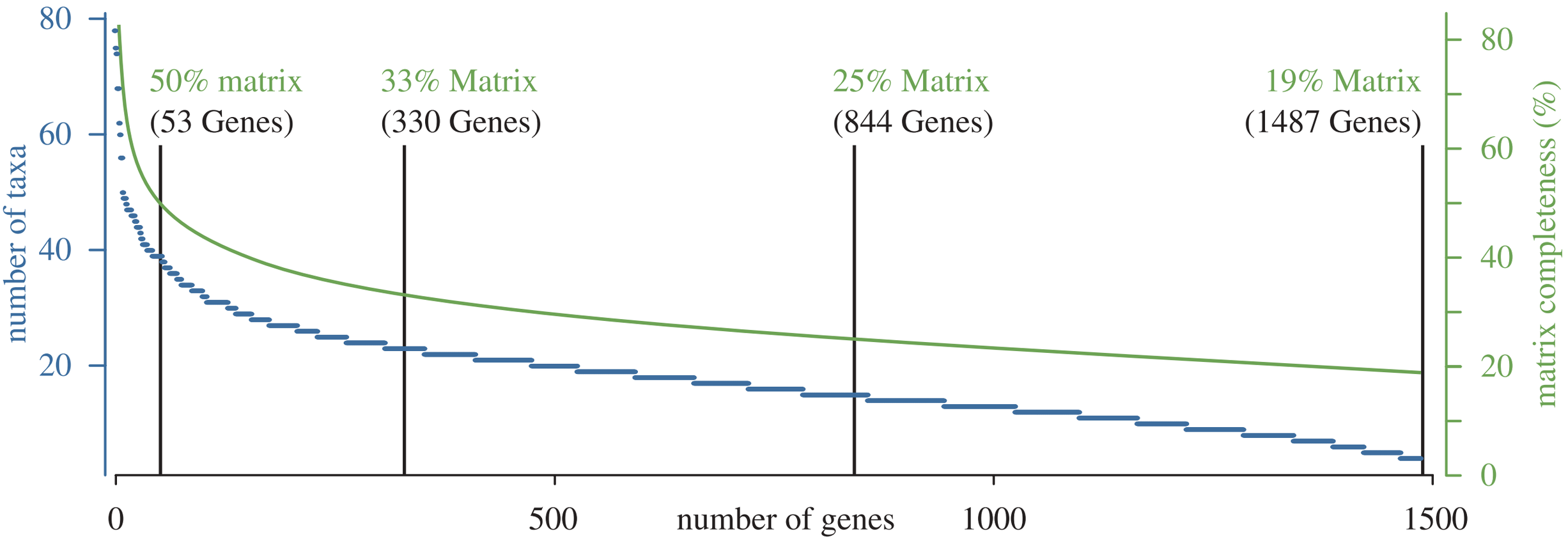
Missing data leads to loss of signal.

Can it also lead to inconsistency?

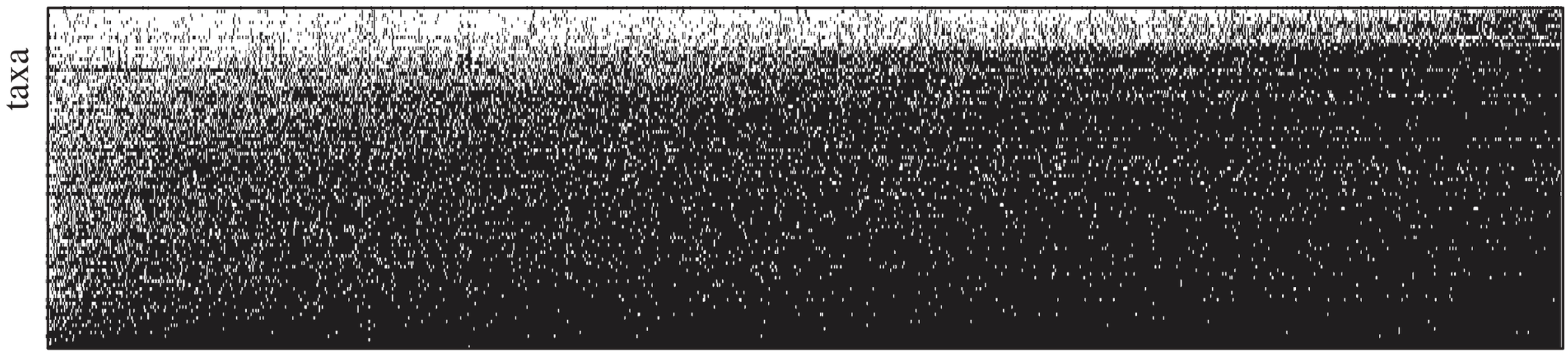
Missing data in phylogenetic analyses are rarely distributed randomly.

Some genes are better sampled than others, some taxa are better sampled than others.

(*a*)



(*b*)



Hejnol et al. 2009, http://dx.doi.org/10.1098/rspb.2009.0896

Should you add poorly sampled genes and taxa to an analysis, or have a smaller but more complete analysis?

Amount of missing data:

less	more
Eliminate poorly sampled taxa	Broader taxon sampling
Eliminate poorly sampled genes	Broader gene sampling
Additional expense	Reduced expense
	Could be impacted by inconsistency

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 39 (2006) 34–42

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Methodological Review

Missing data and the design of phylogenetic analyses

John J. Wiens *

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA

Received 5 March 2005

Available online 6 May 2005

Taxon

	1	2	3	4	5	6	7	8	9	10
Character 1	A	G	A	C	G	A	T	A	C	G
Character 2	G	G	A	C	G	T	T	A	C	G
Character 3	G	?	?	?	G	A	?	C	?	T
Character 4	G	?	?	?	C	A	?	C	?	T

Analyze data matrix with missing data cells

Taxon

	1	5	6	8	10
Character 1	A	G	A	A	G
Character 2	G	G	T	A	G
Character 3	G	G	A	C	T
Character 4	G	C	A	C	T

Eliminate missing data by excluding taxa

Taxon

	1	2	3	4	5	6	7	8	9	10
Character 1	A	G	A	C	G	T	T	A	C	G
Character 2	G	G	A	C	G	A	T	A	C	G

Eliminate missing data by excluding characters

Wiens 2006, http://dx.doi.org/10.1016/j.jbi.2005.04.001

“Recent simulations suggest that highly incomplete taxa can be accurately placed in phylogenies, as long as many characters have been sampled overall. Furthermore, adding incomplete taxa can dramatically improve results in some cases by subdividing misleading long branches. Adding characters with missing data can also improve accuracy, although there is a risk of long-branch attraction in some cases.”

Wiens 2006, http://dx.doi.org/10.1016/j.jbi.2005.04.001

Syst. Biol. 58(1):130–145, 2009

DOI:10.1093/sysbio/syp017
Advance Access publication on May 21, 2009

The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference

ALAN R. LEMMON[1,2,3,*], JEREMY M. BROWN[1], KATHRIN STANGER-HALL[4], AND EMILY MORIARTY LEMMON[1,3]

[1]*Section of Integrative Biology, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA;*
[2]*Present address: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120, USA;*
[3]*Present address: Department of Biological Science, Florida State University, Tallahassee, FL 32306, USA;*
[4]*Plant Biology Department, University of Georgia, 403 Biosciences Building, Athens, GA 30602, USA;*
**Correspondence to be sent to: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120, USA; E-mail: alemmon@evotutor.org.*

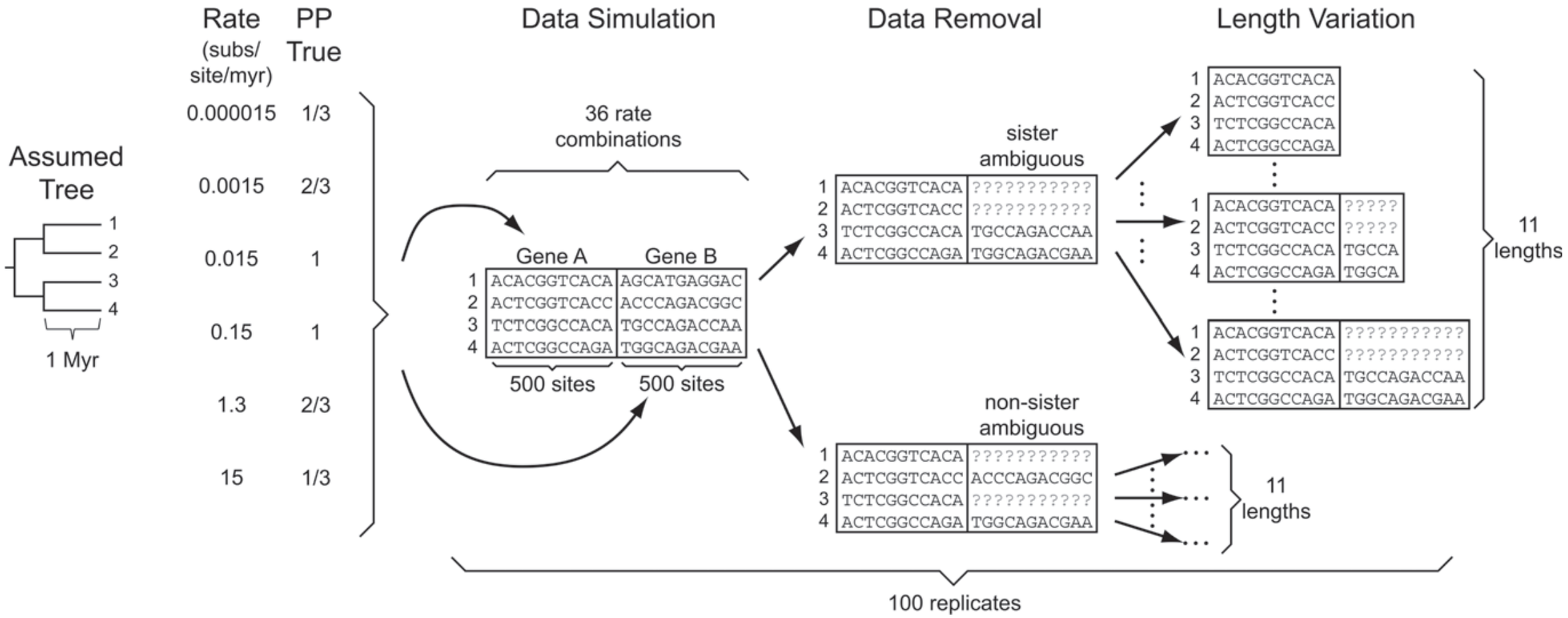


FIGURE 1. Simulation design. Among-site rate variation was simulated using 6 rates of evolution (chosen to produce the desired PP for the true tree with 500 sites) combined across 2 genes to form 36 rate combinations. Gene A contained unambiguous sites, whereas Gene B contained ambiguous sites. Ambiguous characters were present for either sister or nonsister taxa. Although Gene A always contained 500 sites, the length of Gene B varied from 0 to 500 sites. Note that Gene B contained no topological information, regardless of the rate of evolution. PP = posterior probabilities.

“We find that in both ML and Bayesian frameworks, among-site rate variation can interact with ambiguous data to produce misleading estimates of topology and branch lengths.”

Lemmon et al. 2009, http://dx.doi.org/10.1093/sysbio/syp017

BMC Evolutionary Biology

Research article

Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment?

Stefanie Hartmann[1,2] and Todd J Vision*[1]

Address: [1]Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA and [2]Institute for Biochemistry and Biology, Karl-Liebknecht-Strasse 24-25, University of Potsdam, 14476 Potsdam, Germany

Email: Stefanie Hartmann - stefanie.hartmann@uni-potsdam.de; Todd J Vision* - tjv@bio.unc.edu

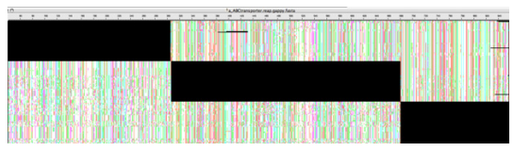
* Corresponding author

Published: 26 March 2008

BMC Evolutionary Biology 2008, **8**:95 doi:10.1186/1471-2148-8-95

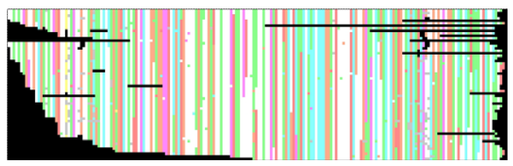
Received: 27 October 2007
Accepted: 26 March 2008

A

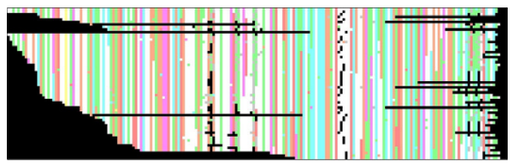


B

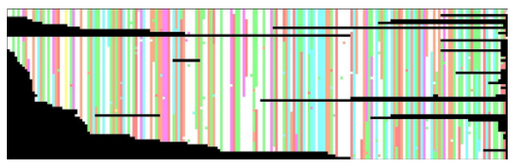
1



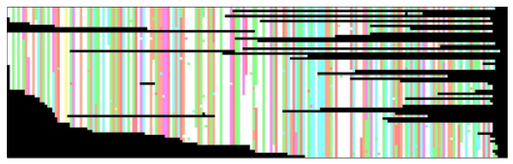
2



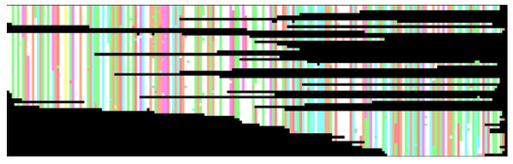
3



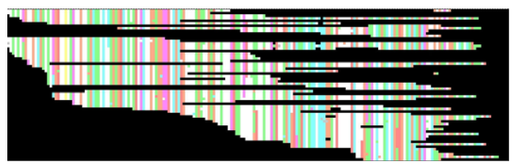
4



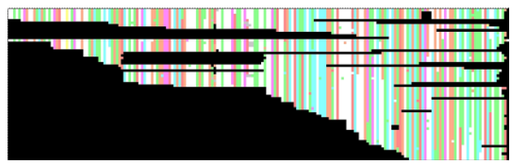
5



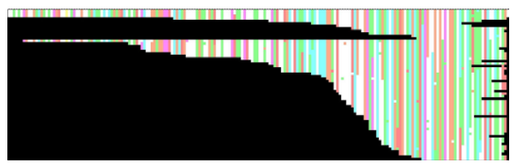
6



7

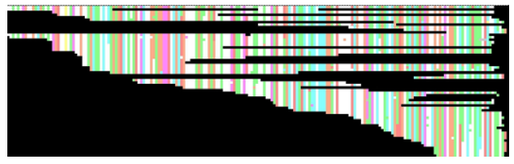


8

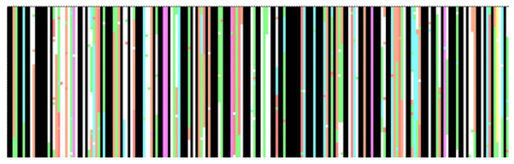


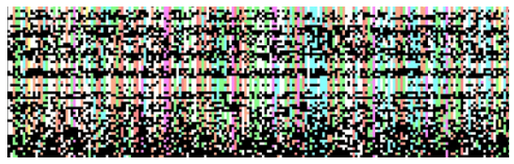
9

10



C





Hartmann et al. 2008, http://dx.doi.org/10.1186/1471-2148-8-95

A

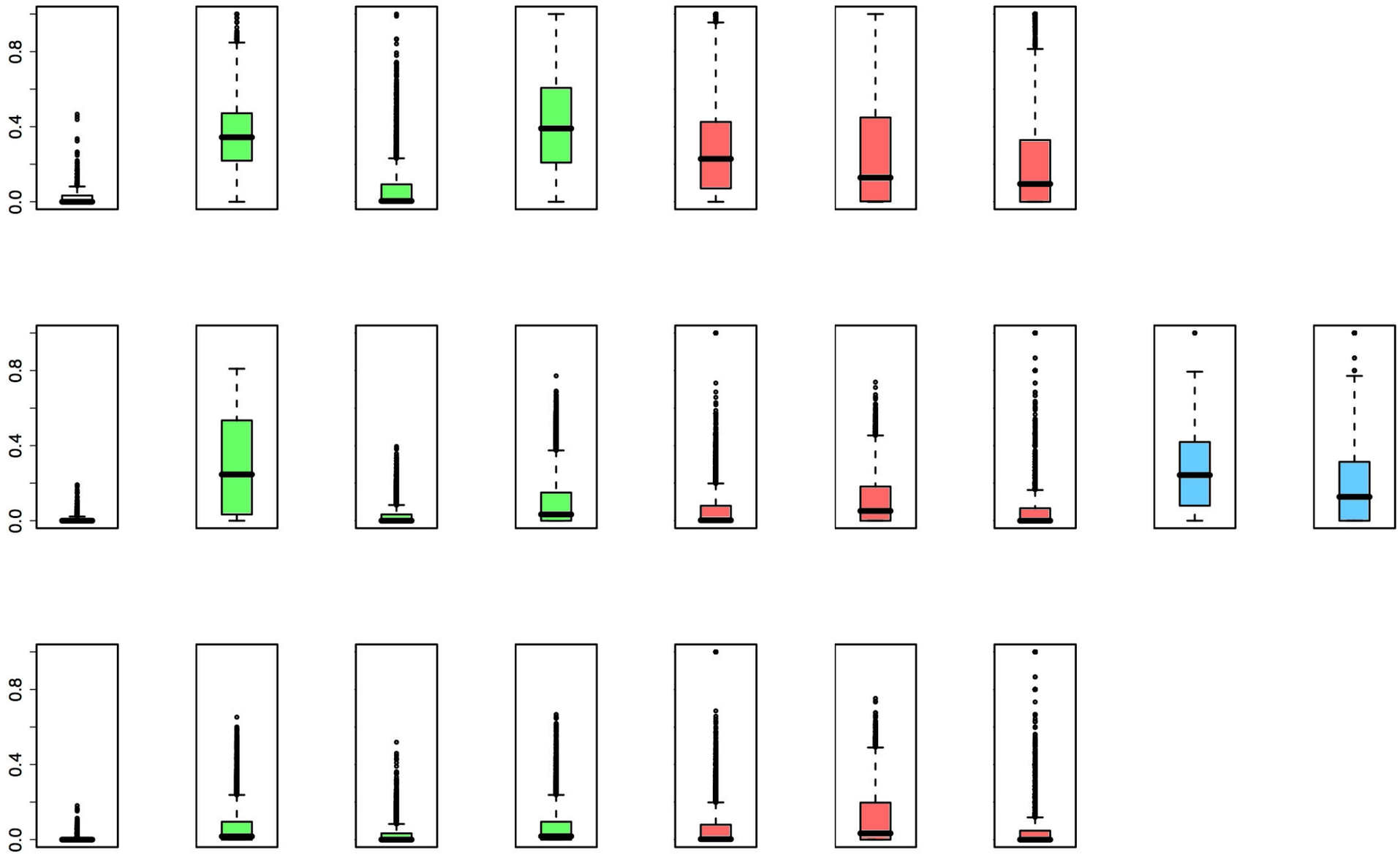
no treatment | alignment masking | SIA approach

full | ESTlike | column-deleted | random-deleted | ESTlike | column-deleted | random-deleted | (full) ESTlike | (pruned) ESTlike

MP stQD

NJ stQD

ML stQD



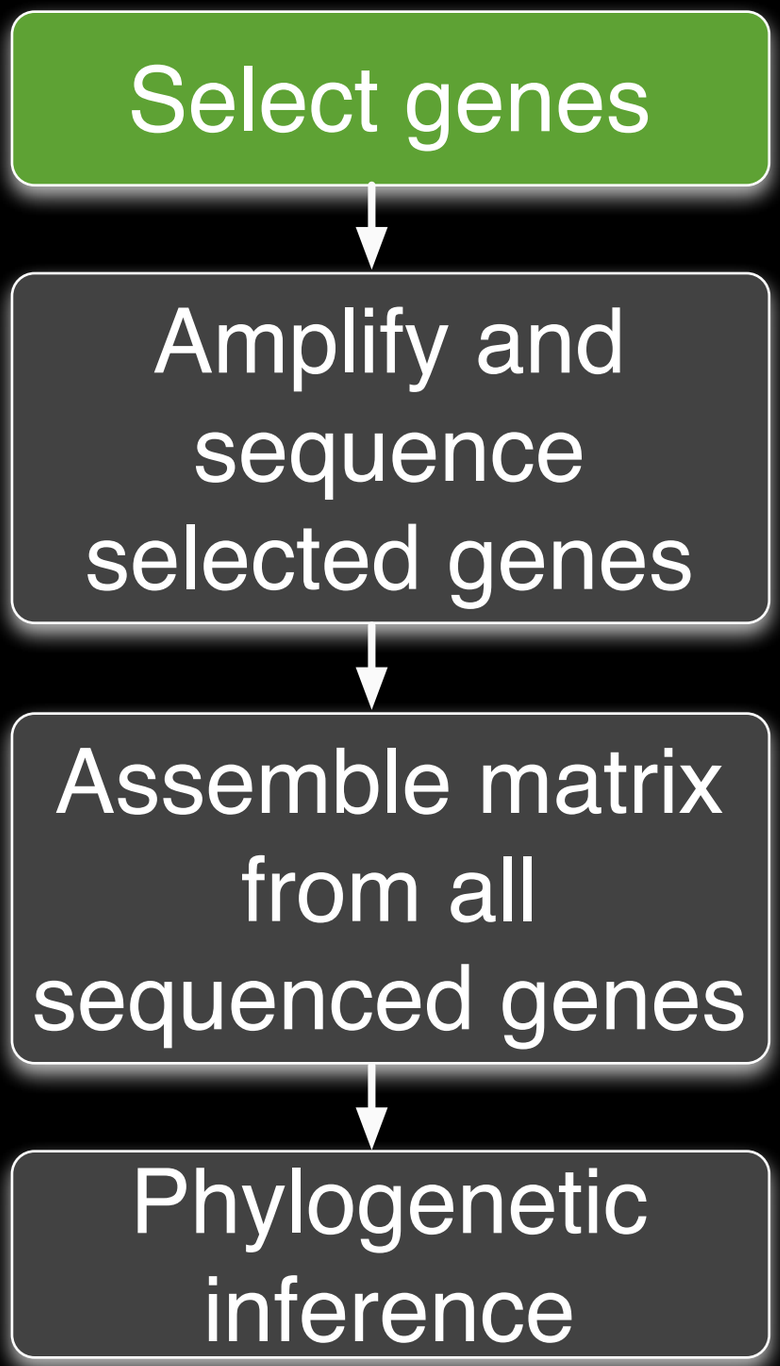
Hartmann et al. 2008, http://dx.doi.org/10.1186/1471-2148-8-95

How to collect your molecular data?

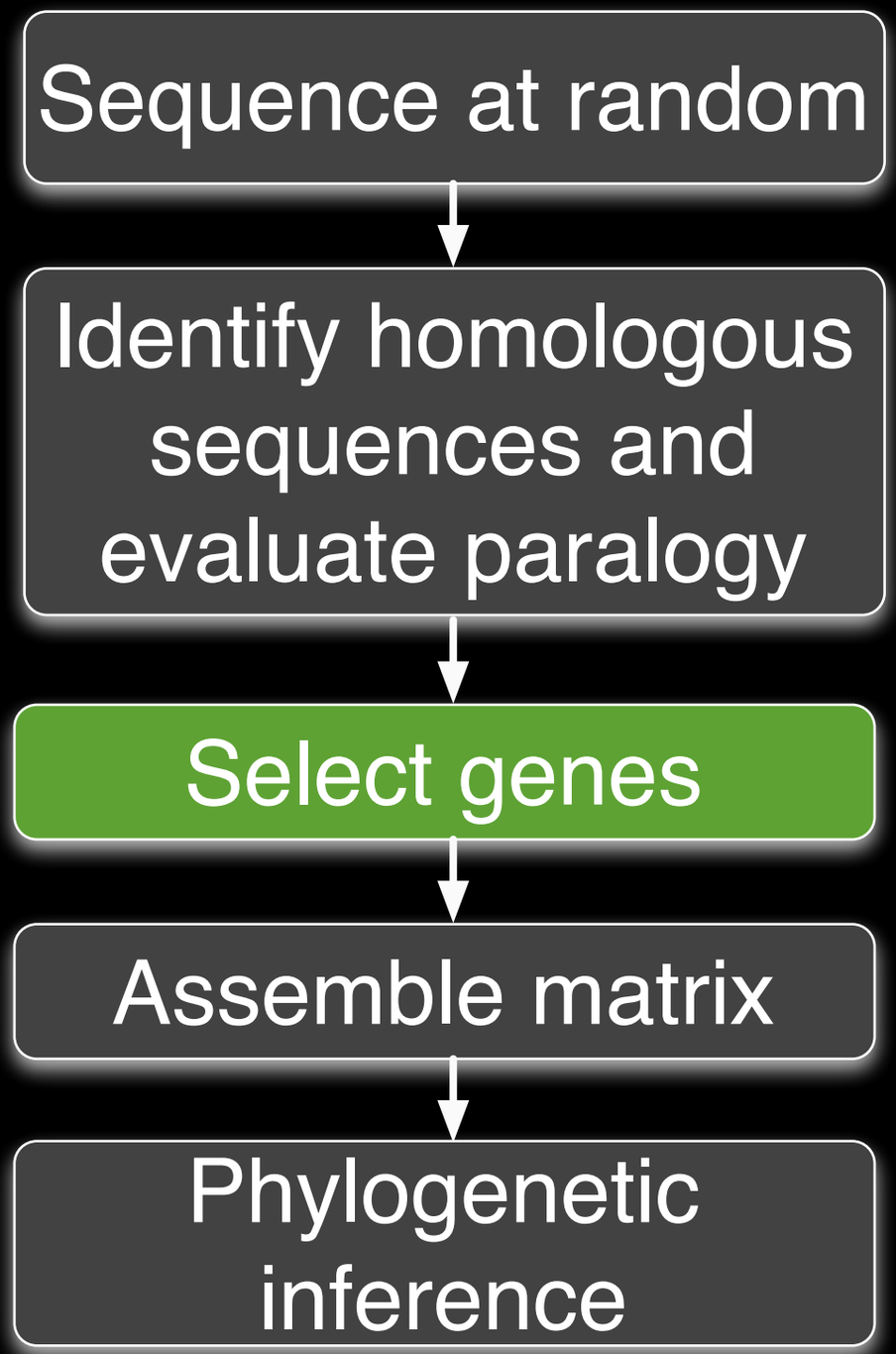
Directed sequencing - pick genes first, then sequence just them

Phylogenomic analyses - use high throughput transcriptome and genome data to broadly sample many genes

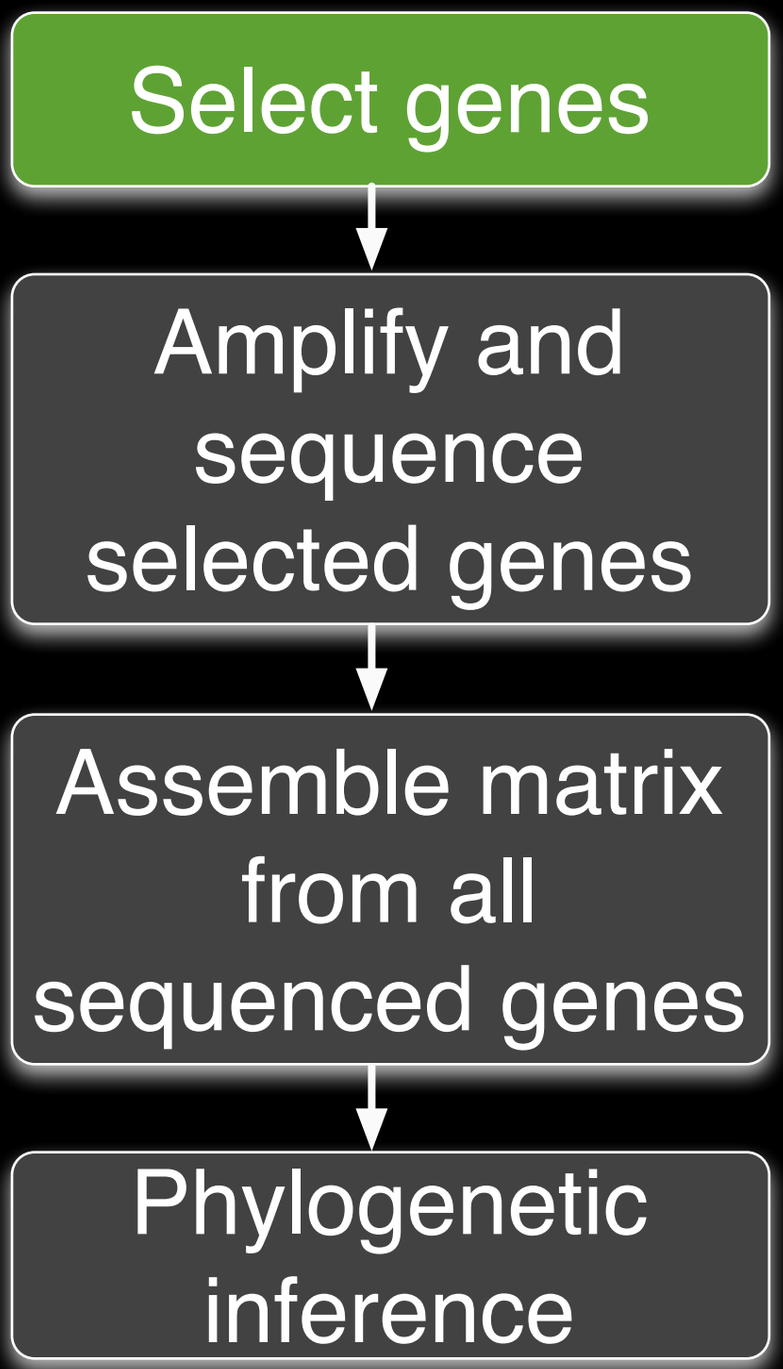
Gene selection as part of project design



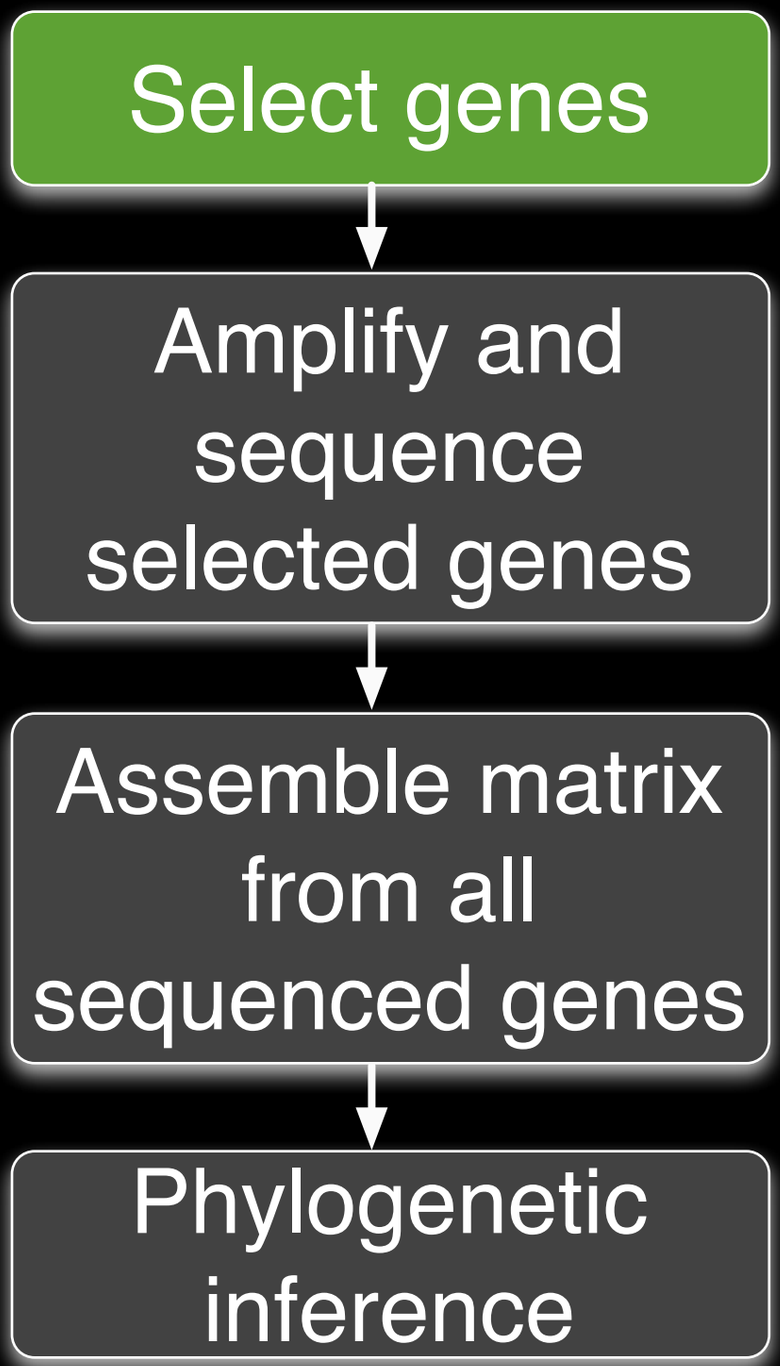
Gene selection as part of analysis



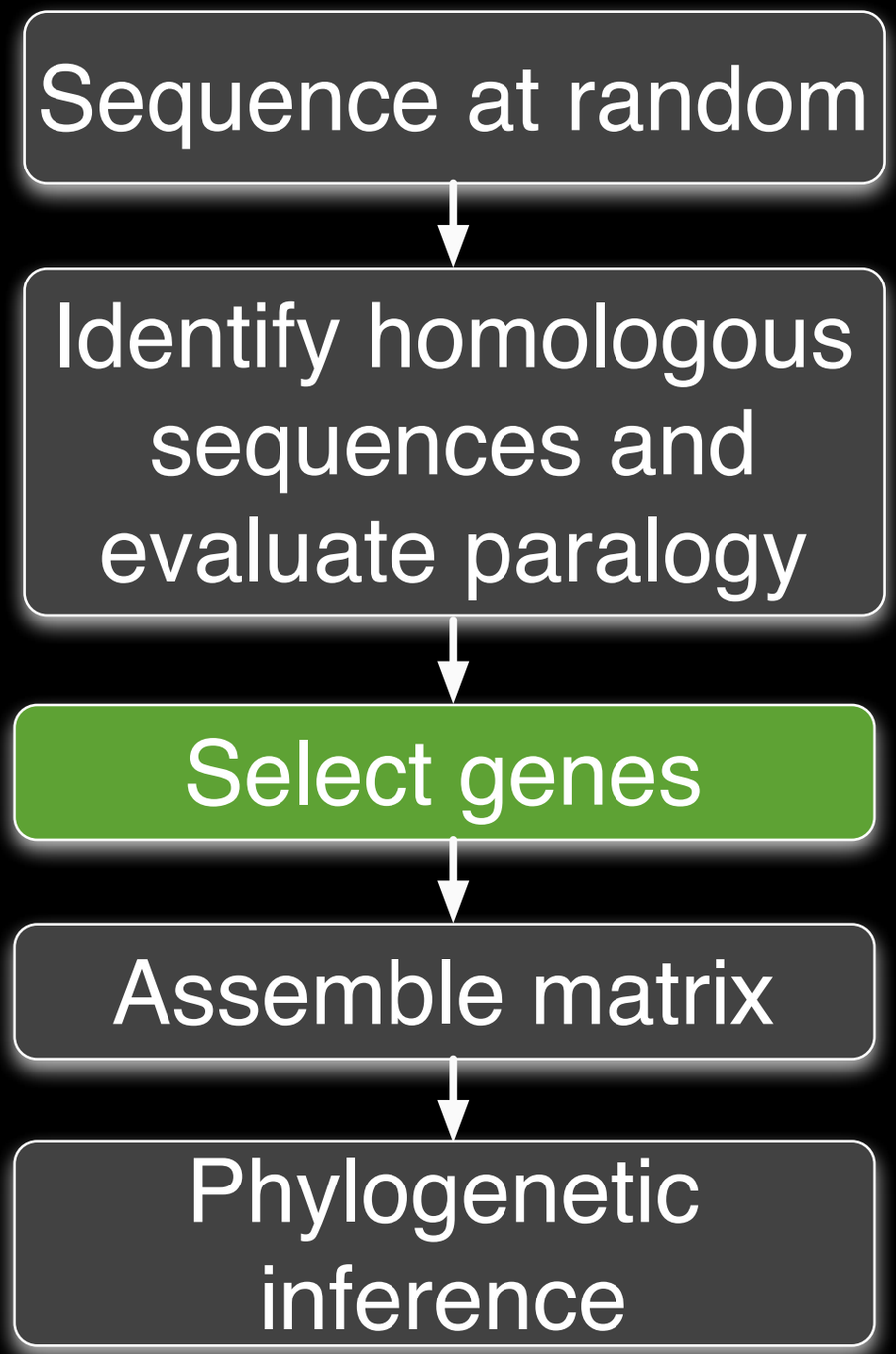
Gene selection as part of project design



Gene selection as part of project design



Gene selection as part of analysis



Should you include morphological data in your analysis?

Should you include morphological data in your analysis?

If you are analyzing fossils, it is essential.

Common tools for inferring trees:

Parsimony	**Likelihood**	**Bayesian**
PAUP*	PAUP*	MrBayes
Poy	RAxML	PhyloBayes
TNT	GARLI	
	Phylip	