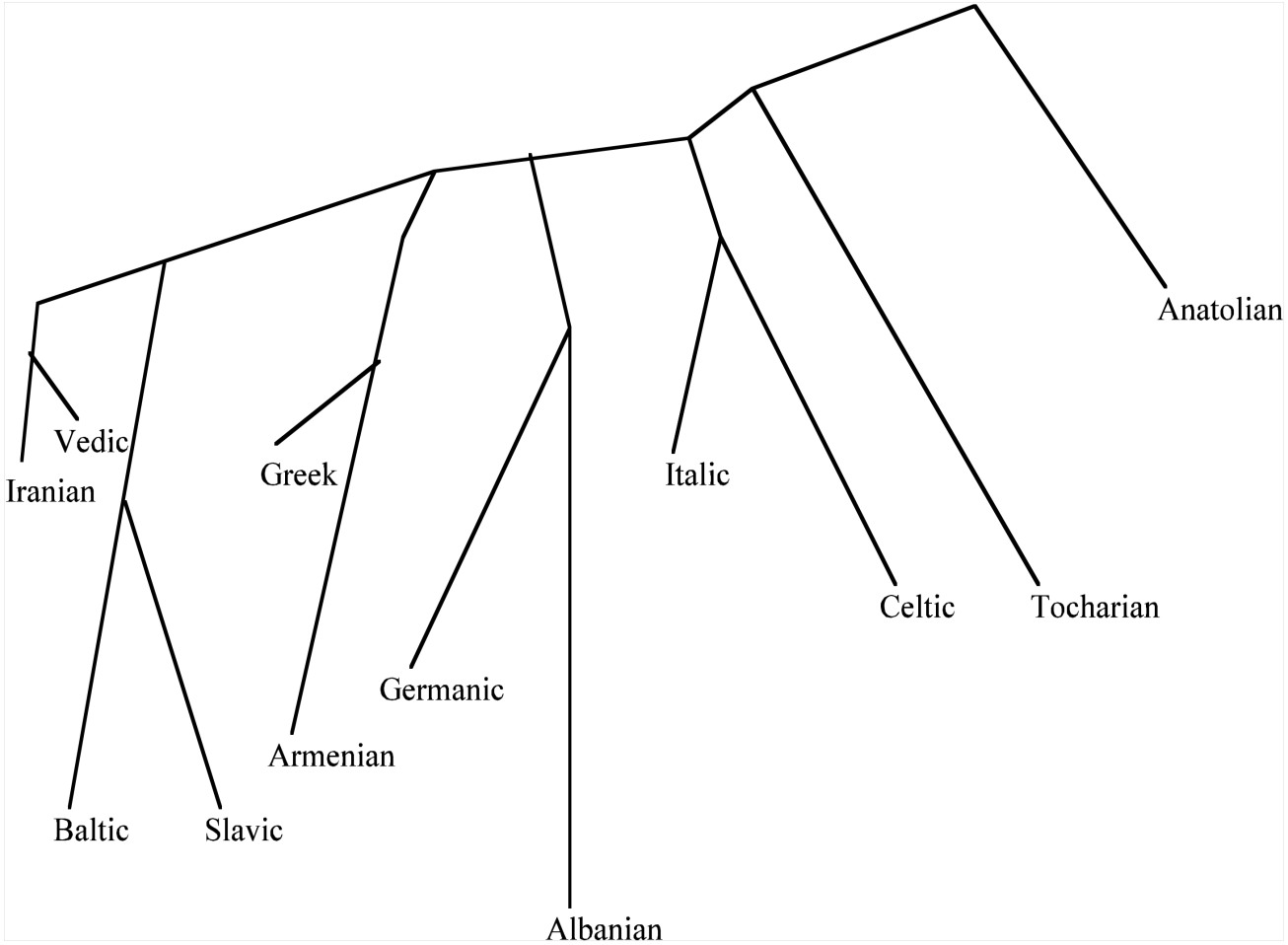
A simulation study comparing phylogeny reconstruction methods for linguistics

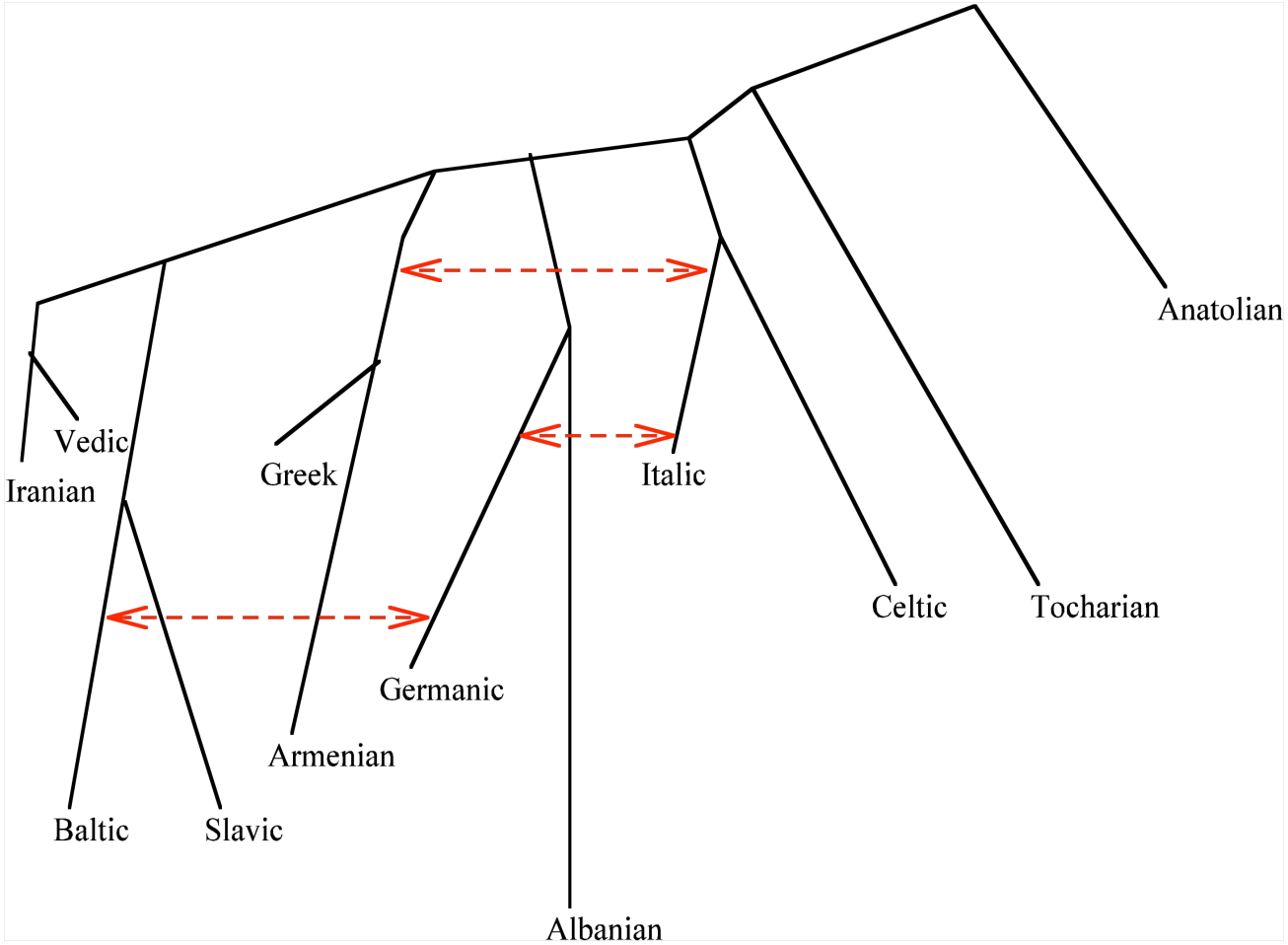
Tandy Warnow
The University of Texas at Austin

Collaborators: Francois Barbancon,
Don Ringe, Luay Nakhleh, and Steve Evans

Possible Indo-European tree (Ringe, Warnow and Taylor 2000)



Possible phylogenetic network for IE
Nakhleh *et al*., Language 2005



Controversies for IE history

- Subgrouping: Other than the 10 major subgroups, what is likely to be true? In particular, what about
  - Italo-Celtic
  - Greco-Armenian
  - Anatolian + Tocharian
  - Satem Core (Indo-Iranian and Balto-Slavic)
  - Location of Germanic
- Dates?
- How tree-like is IE?

Controversies for IE history

- Subgrouping: Other than the 10 major subgroups, what is likely to be true? In particular, what about
  - Italo-Celtic
  - Greco-Armenian
  - Anatolian + Tocharian
  - Satem Core (Indo-Iranian and Balto-Slavic)
  - Location of Germanic
- Dates?
- How tree-like is IE?

Note: many reconstructions of IE have been done, but produce different histories. Most have been done on the Dyen et al. lexical database.

The performance of methods on an IE data set

(Transactions of the Philological Soc., Nakhleh et al. 2005)

Observation: Different datasets (not just different methods) can give different reconstructed phylogenies.

Objective: Explore the differences in reconstructions as a function of data (lexical alone versus lexical, morphological, and phonological), screening (to remove obviously homoplastic characters), and methods.

Better datasets

- Ringe & Taylor (differs from Dyen, Kruskal, Black in using earliest well-attested data instead of modern data)
  - The screened full dataset of 294 characters (259 lexical, 13 morphological, 22 phonological)
  - The unscreened full dataset of 336 characters (297 lexical, 17 morphological, 22 phonological)
  - The screened lexical dataset of 259 characters.
  - The unscreened lexical dataset of 297 characters.

Differences between different characters

- **Lexical**: most easily borrowed (most borrowings detectable), and homoplasy relatively frequent (we estimate about 25-30% overall for our wordlist, but a much smaller percentage for basic vocabulary).
- **Phonological**: can still be borrowed but much less likely than lexical. Complex phonological characters are infrequently (if ever) homoplastic, although simple phonological characters very often homoplastic.
- **Morphological**: least easily borrowed, least likely to be homoplastic.

Table 1: The 24 IE languages analyzed.

Language	Abbreviation	Language	Abbreviation
Hittite	HI	Old English	OE
Luvian	LU	Old High German	OG
Lycian	LY	Classical Armenian	AR
Vedic	VE	Tocharian A	TA
Avestan	AV	Tocharian B	TB
Old Persian	PE	Old Irish	OI
Ancient Greek	GK	Welsh	WE
Latin	LA	Old Church Slavonic	OC
Oscan	OS	Old Prussian	PR
Umbrian	UM	Lithuanian	LI
Gothic	GO	Latvian	LT
Old Norse	ON	Albanian	AL

Phylogeny reconstruction methods

- Neighbor joining
- UPGMA (technique in glottochronology)
- Maximum parsimony
- Maximum compatibility (weighted and unweighted)
- Gray and Atkinson (Bayesian estimation based upon presence/absence of cognates)

Some observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).

Some observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).
- Other than UPGMA, all methods reconstruct the ten major subgroups, as well as **Anatolian + Tocharian** and **Greco-Armenian** - *but are otherwise quite different*!

Some observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).
- Other than UPGMA, all methods reconstruct the ten major subgroups, as well as **Anatolian + Tocharian** and **Greco-Armenian** - *but are otherwise quite different*!
- The Satem Core (Indo-Iranian plus Balto-Slavic) is not always reconstructed.

Some observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).
- Other than UPGMA, all methods reconstruct the ten major subgroups, as well as **Anatolian + Tocharian** and **Greco-Armenian** - *but are otherwise quite different*!
- The Satem Core (Indo-Iranian plus Balto-Slavic) is not always reconstructed.
- Almost all analyses put Italic, Celtic, and Germanic together. (The only exception is weighted maximum compatibility on datasets that include morphological characters.)

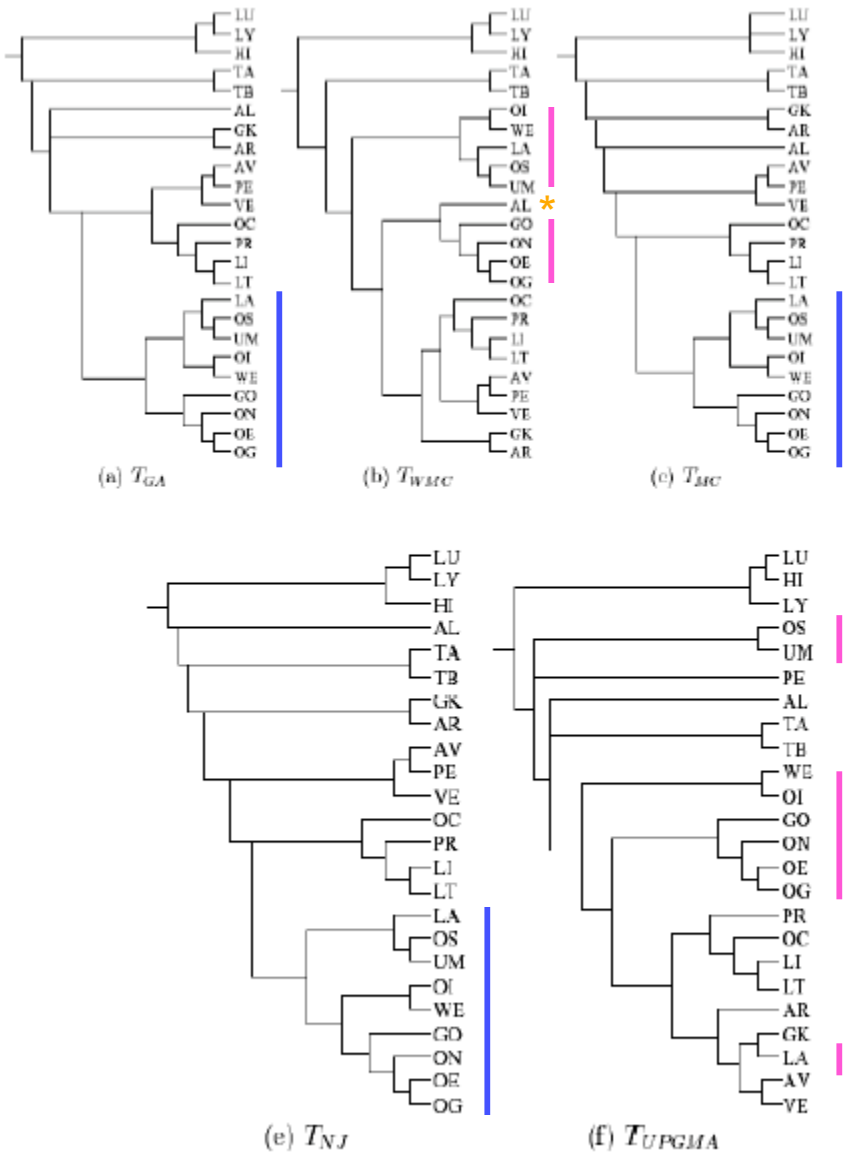


Figure 1. Five trees inferred on the screened full dataset

GA = Gray+Atkinson Bayesian MCMC method

WMC = weighted maximum compatibility

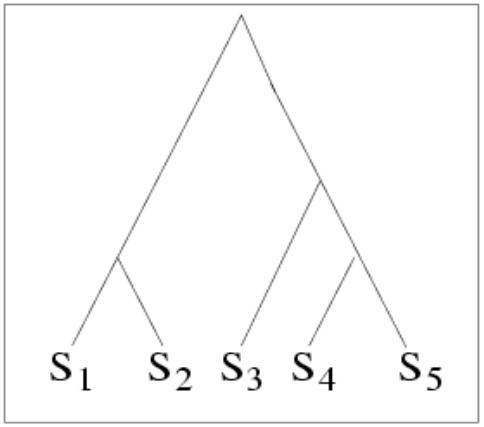
MC = maximum compatibility (identical to maximum parsimony on this dataset)

NJ = neighbor joining (distance-based method, based upon corrected distance)

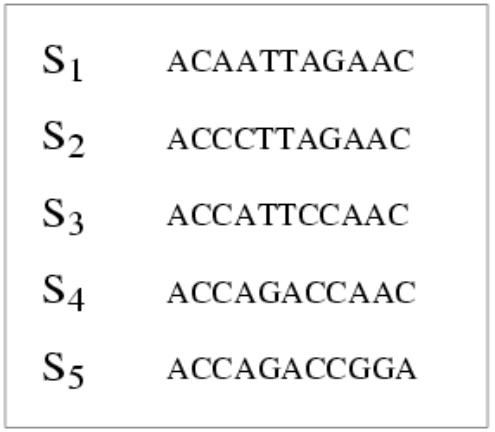
UPGMA = agglomerative clustering technique used in glottochronology.

Different methods/data
give different answers.
We don't know
which answer is correct.
Which method(s)/data
should we use?

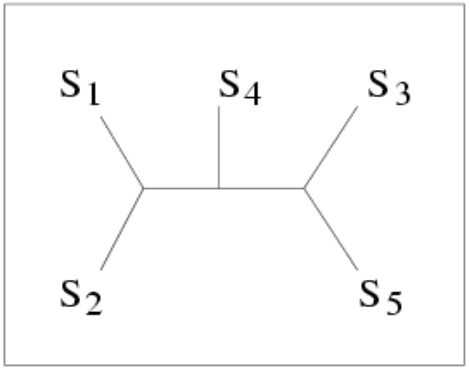
Simulation study (cartoon)



TRUE TREE

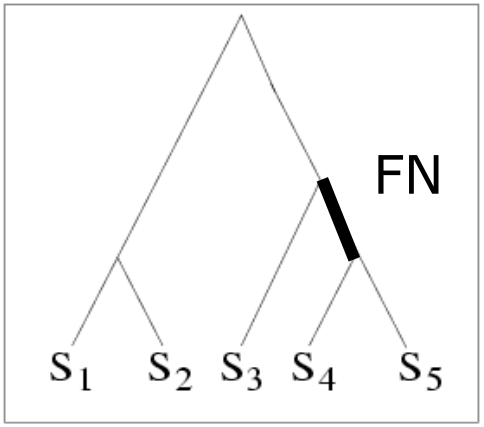


DNA SEQUENCES

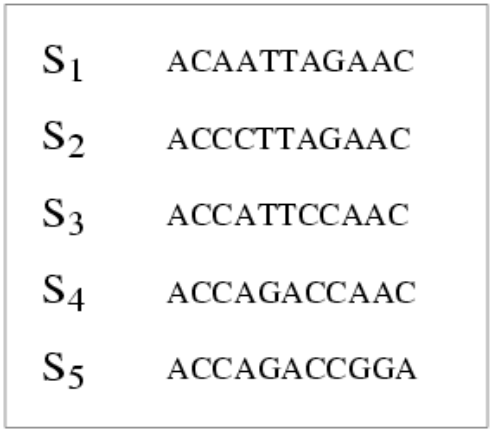


INFERRED TREE

Simulation study (cartoon)



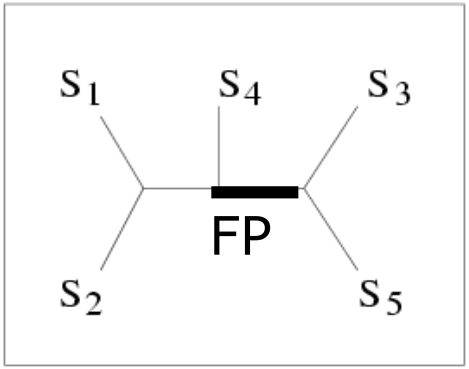
TRUE TREE



DNA SEQUENCES

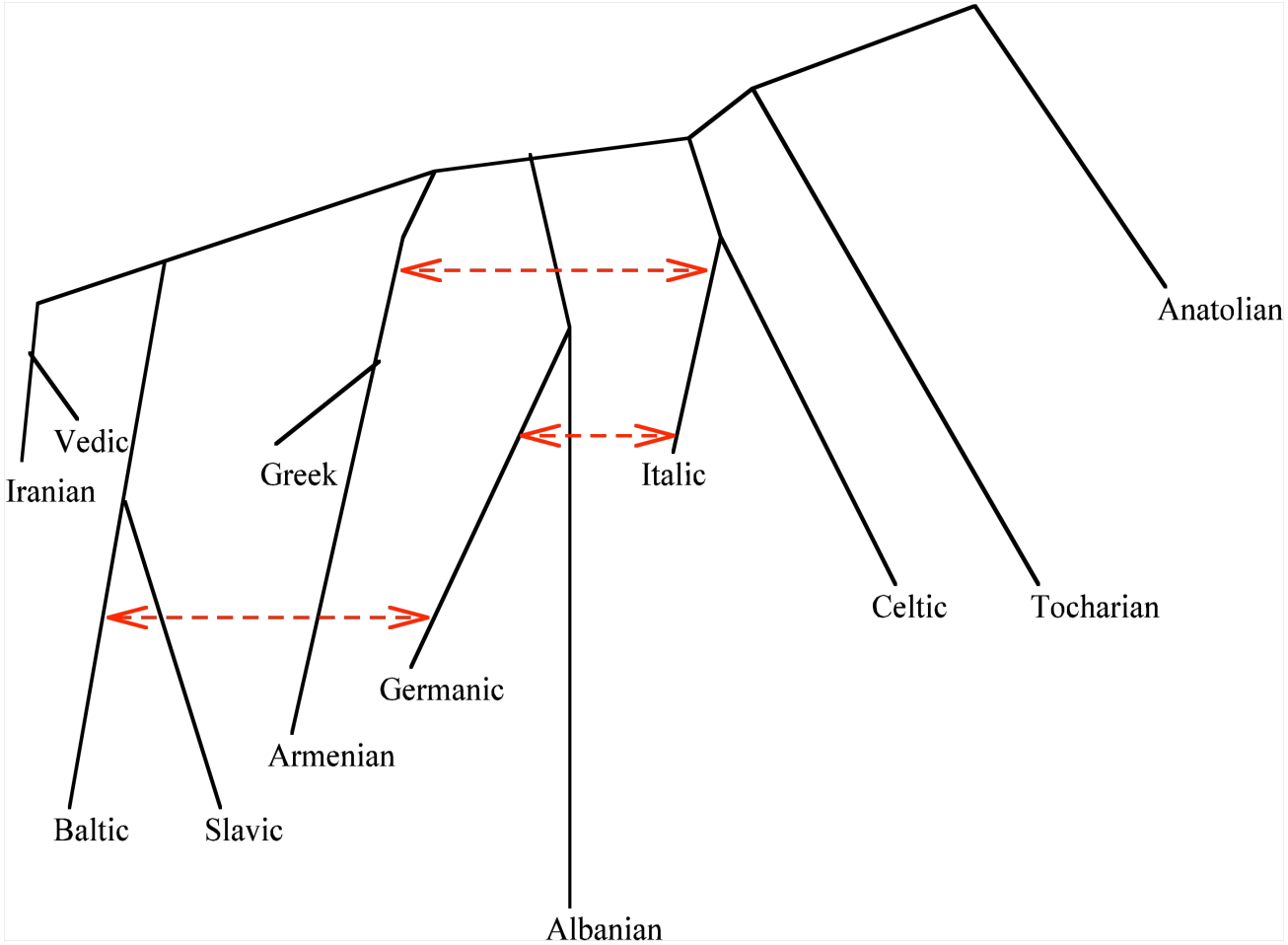
FN: false negative
(**missing edge**)
FP: false positive
(incorrect edge)

50% error rate

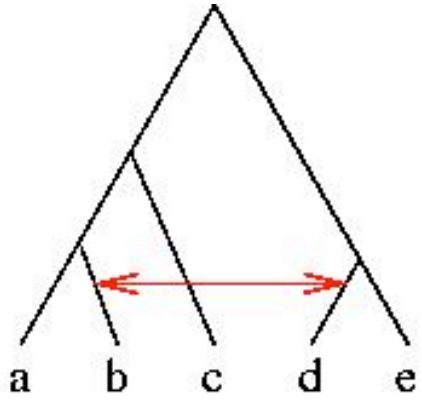


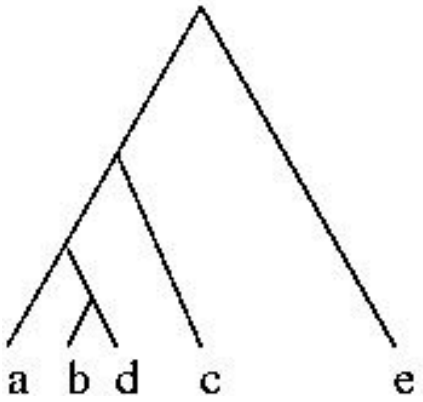
INFERRED TREE

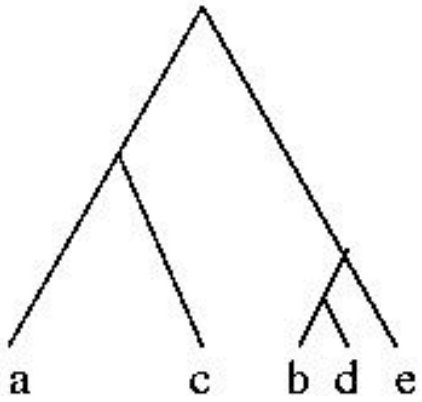
Phylogenetic Network Evolution

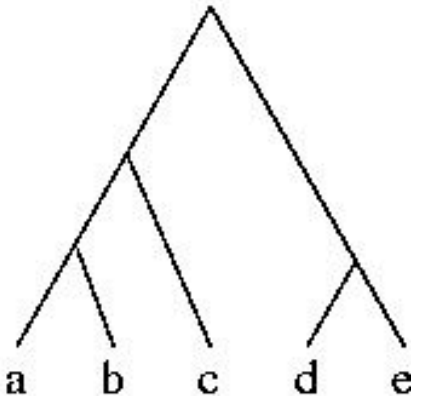


Modelling borrowing: Networks and Trees within Networks

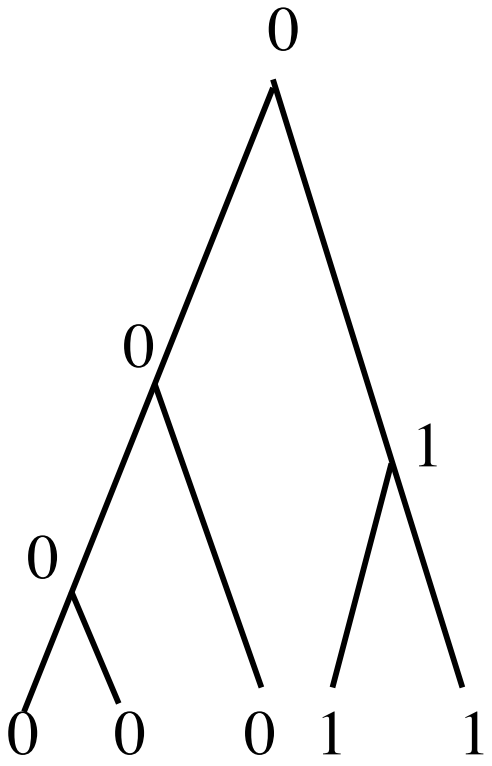




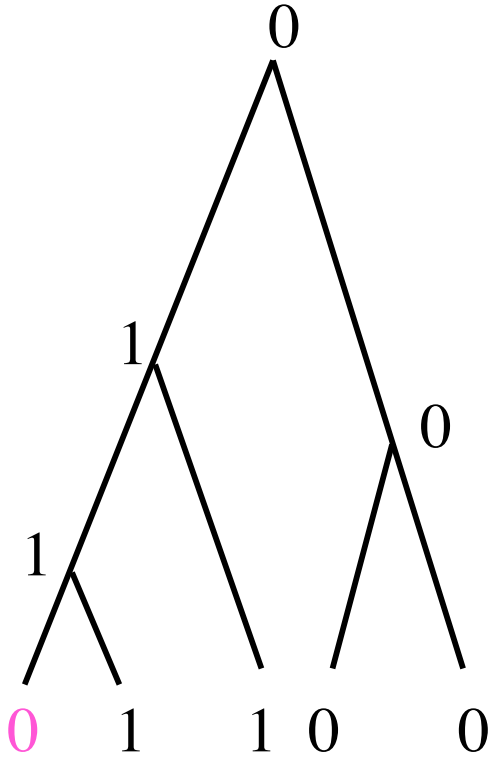




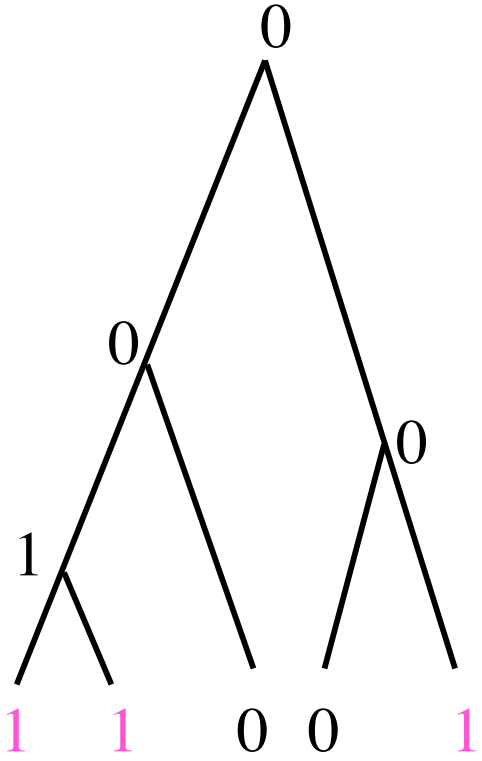
Some useful terminology: homoplasy



no homoplasy

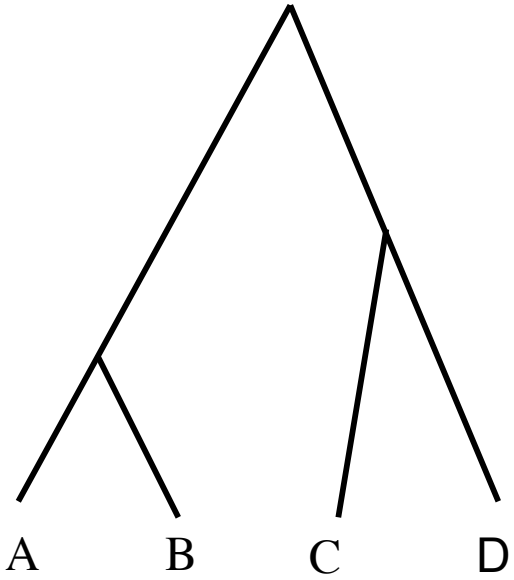


back-mutation

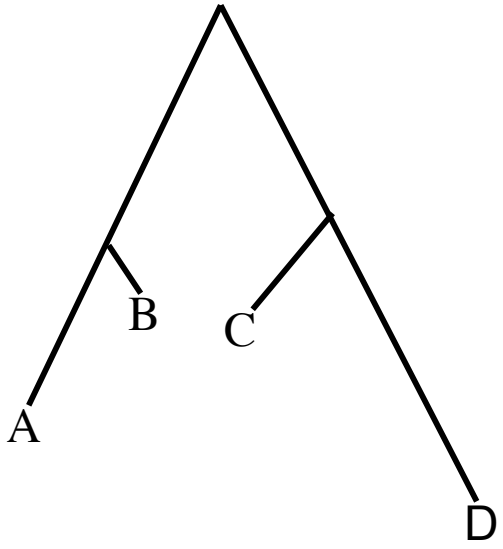


parallel evolution

Some useful terminology: lexical clock



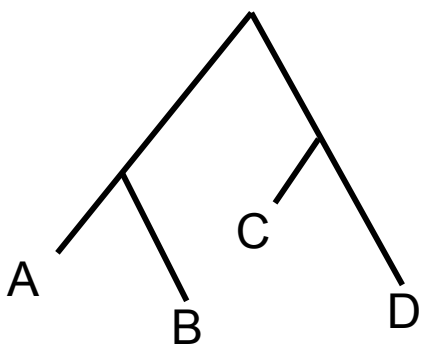
lexical clock

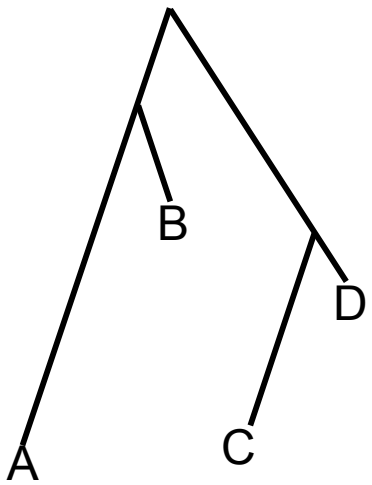


no lexical clock

edge lengths represent expected numbers of substitutions

Heterotachy = departure from rates-across-sites





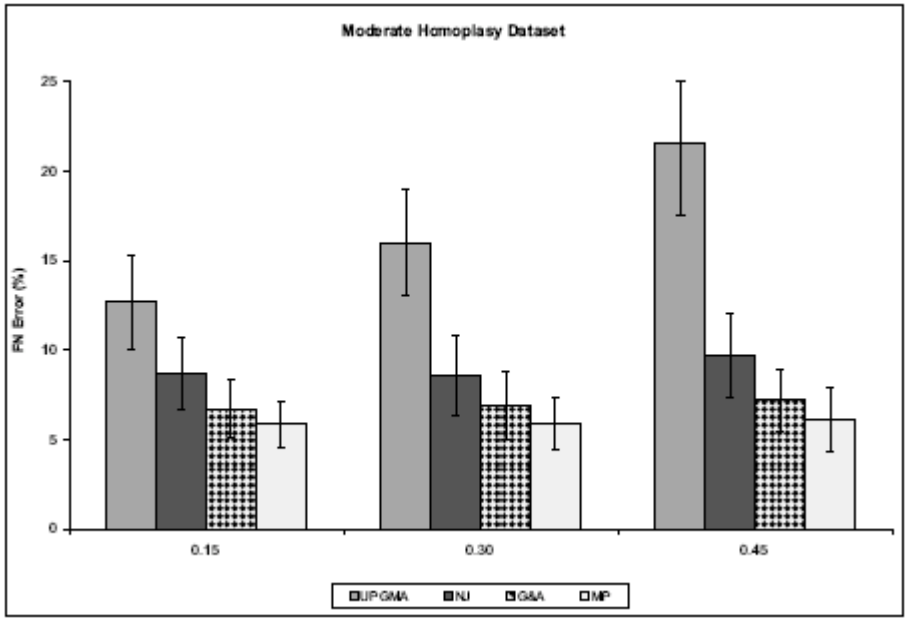
The underlying tree is fixed, but there are no constraints on edge length variations between characters.

Our simulation study

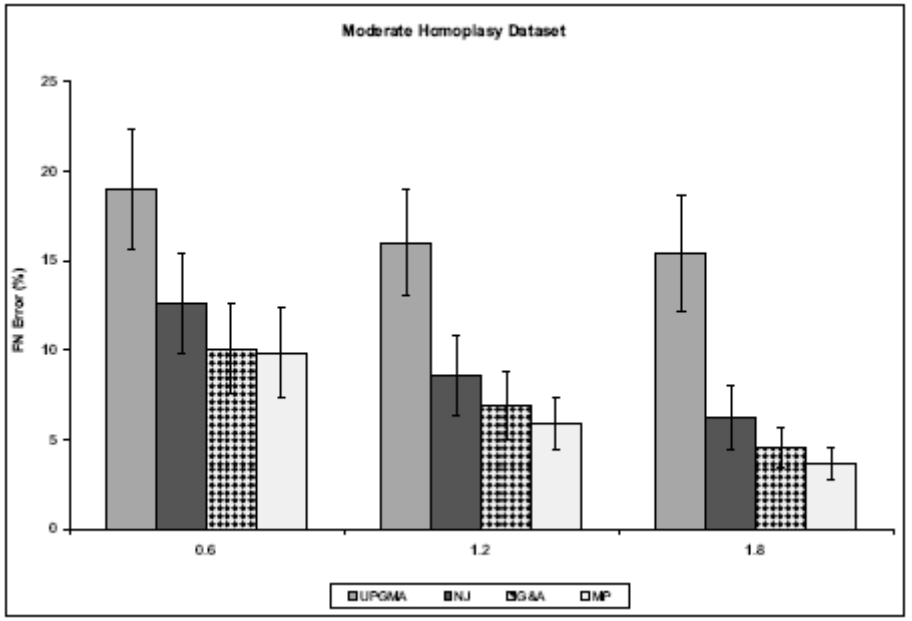
- Model phylogenetic networks: each had 30 leaves and up to three contact edges, and varied in the ***deviation from a lexical clock.***
- Multi-state characters: 360 lexical and 60 morphological, with varied rates of **homoplasy** and **borrowing** (to reflect empirically estimated rates). We also varied the degree of **heterotachy**.
- Performance metric: We compared estimated trees to the "**genetic tree**" with respect to the missing edge rate.

Observations

1. Choice of reconstruction method does matter.
2. Relative performance between methods is quite stable (distance-based methods worse than character-based methods).
3. Choice of data does matter (good idea to add morphological characters).
4. Accuracy only slightly lessened with small increases in homoplasy, borrowing, or deviation from the lexical clock.
5. Some amount of heterotachy helps!



(i)



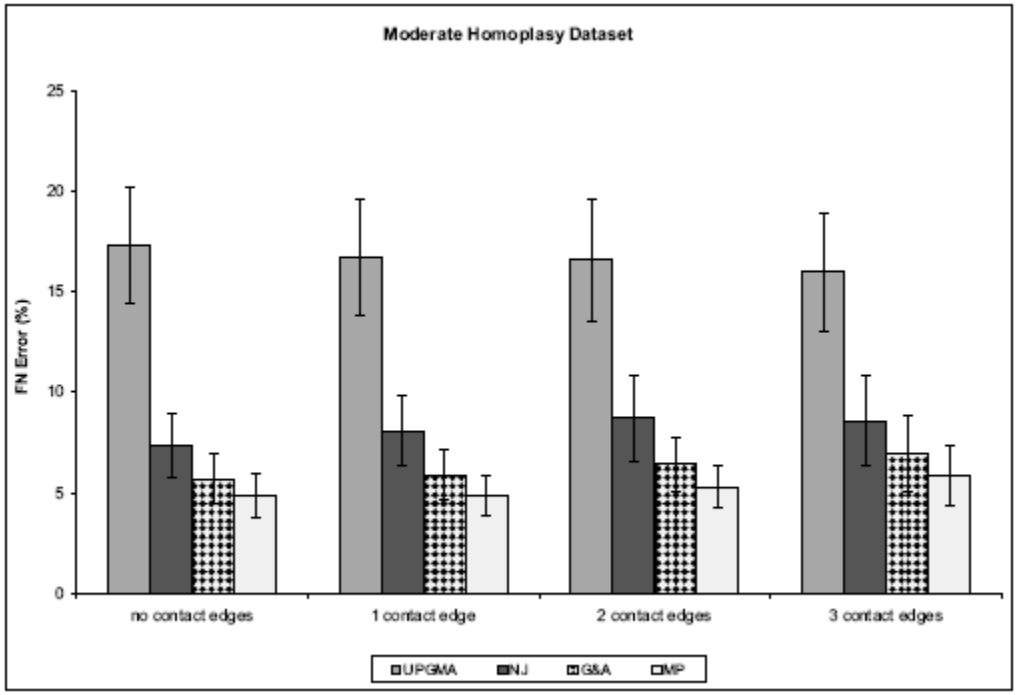
(ii)

Relative performance of methods **(UPGMA, NJ, G&A, MP)** on moderate homoplasy datasets under various model conditions:

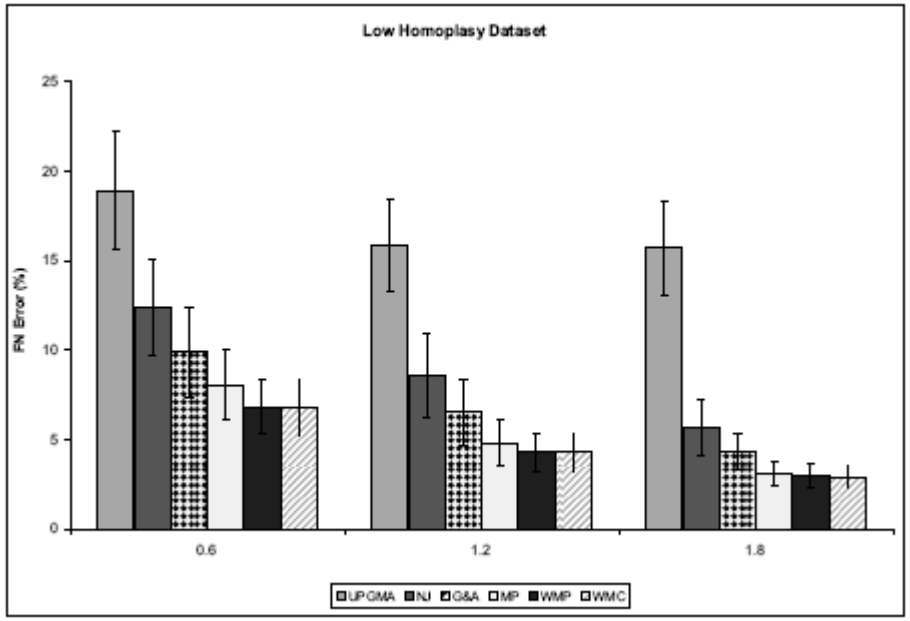
(i) varying the deviation from the lexical clock,

(ii) varying heterotachy, and

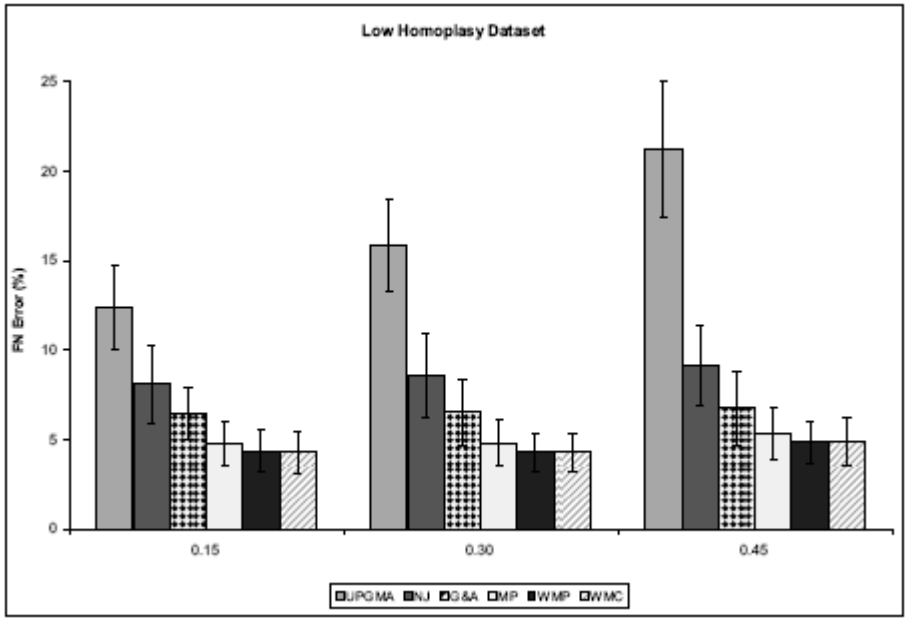
(iii) varying the number of contact edges.



(iii)



(i)



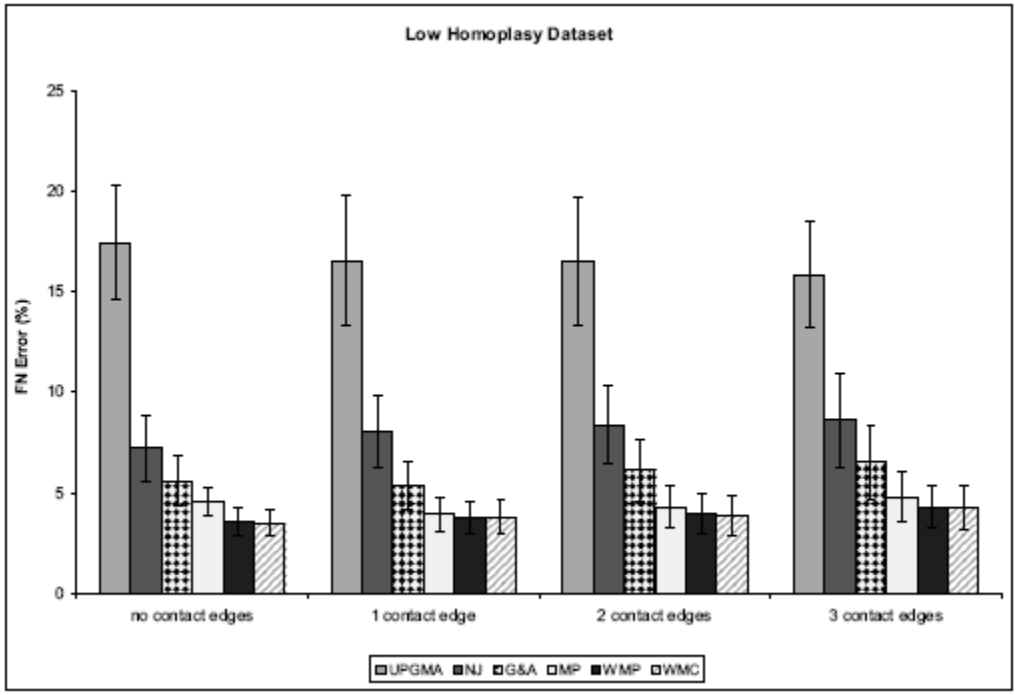
(ii)

Relative performance of methods **(UPGMA, NJ, G&A, MP, WMP, WMC)** for low homoplasy datasets under various model conditions:

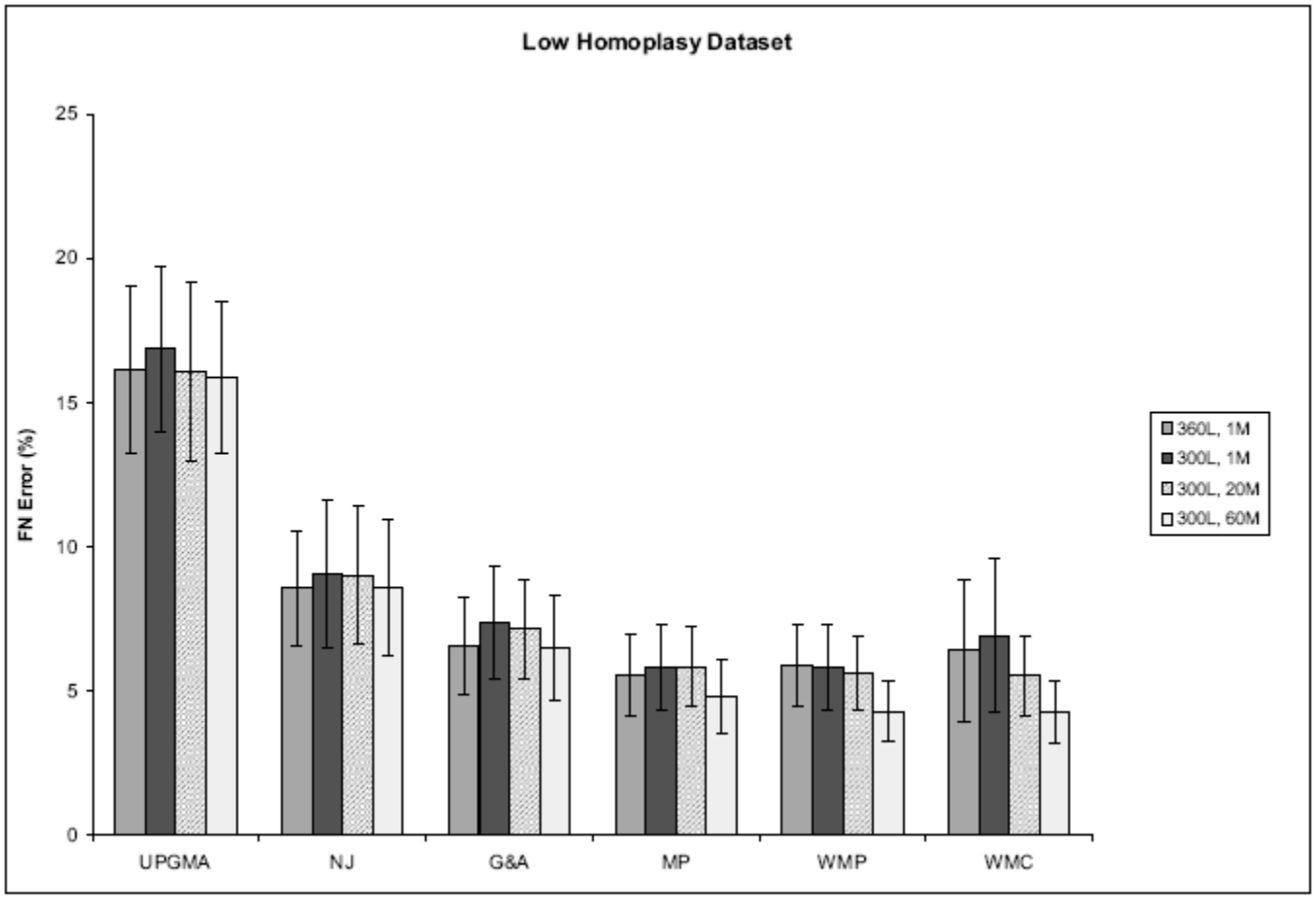
(i) varying the heterotachy,

(ii) varying the deviation from the lexical clock, and

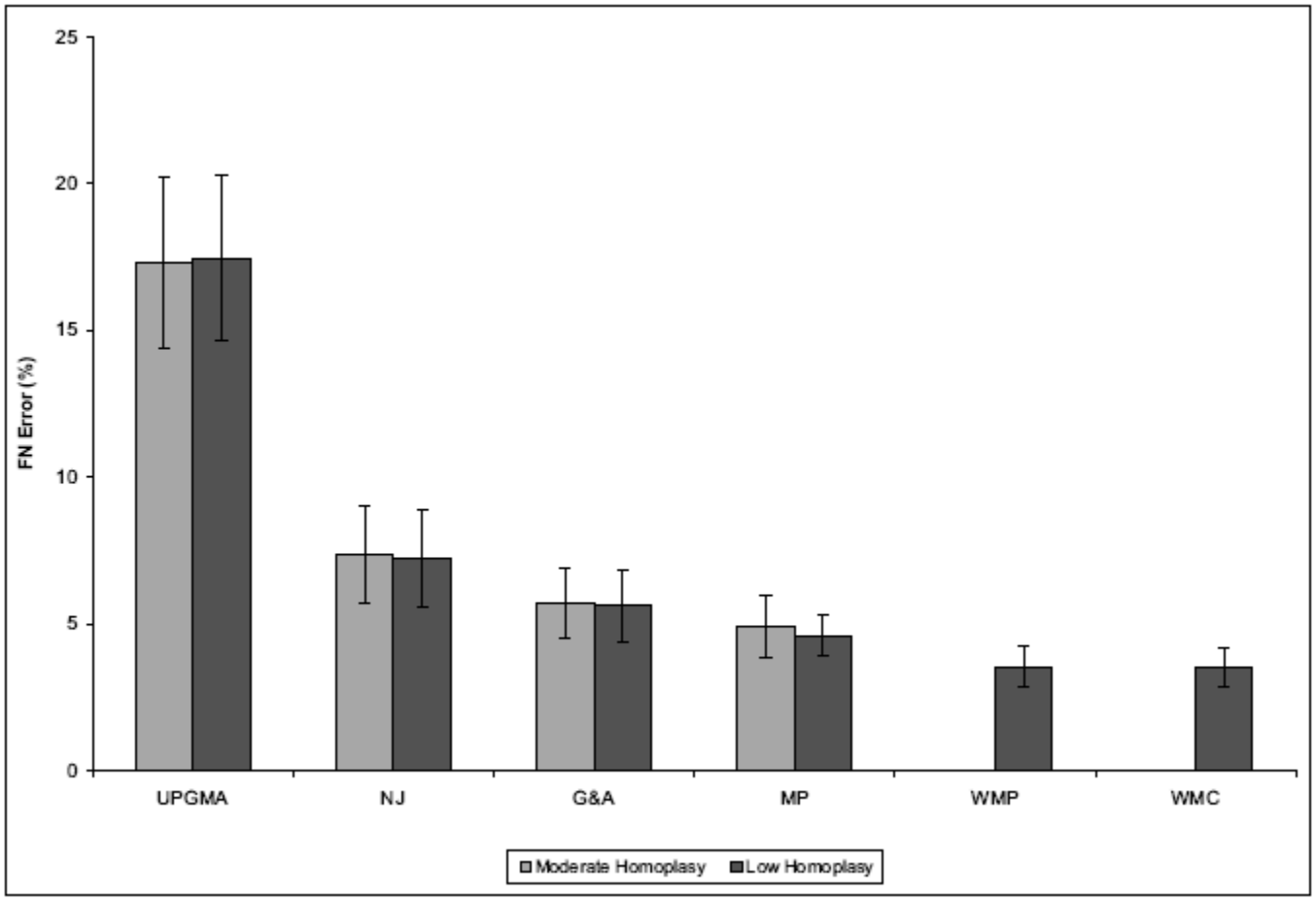
(iii) varying the number of contact edges.



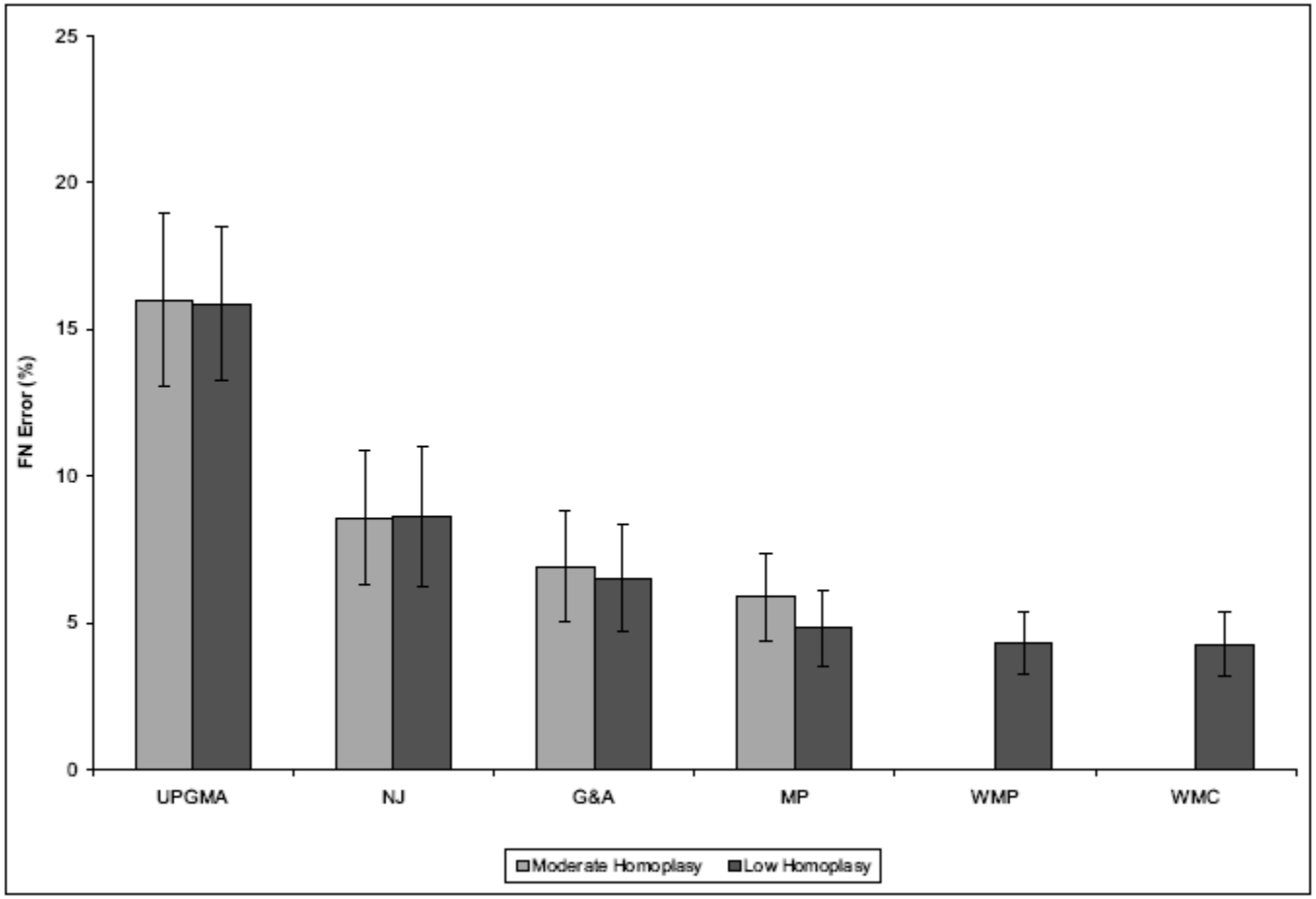
(iii)



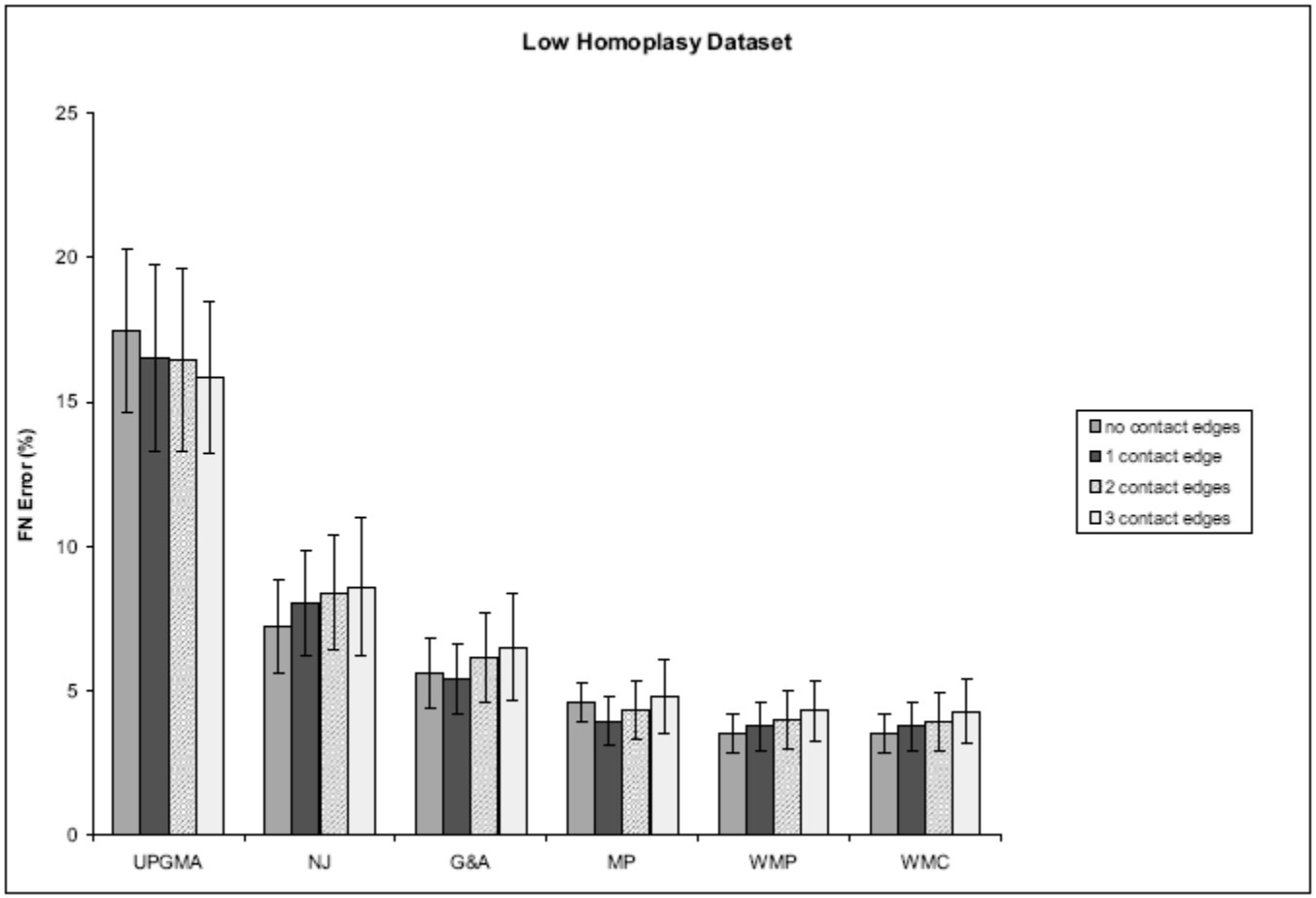
Impact of data selection for characters evolved down a network with three contact edges, under low homoplasy (“screened data"), moderate deviation from a lexical clock, and moderate heterotachy.



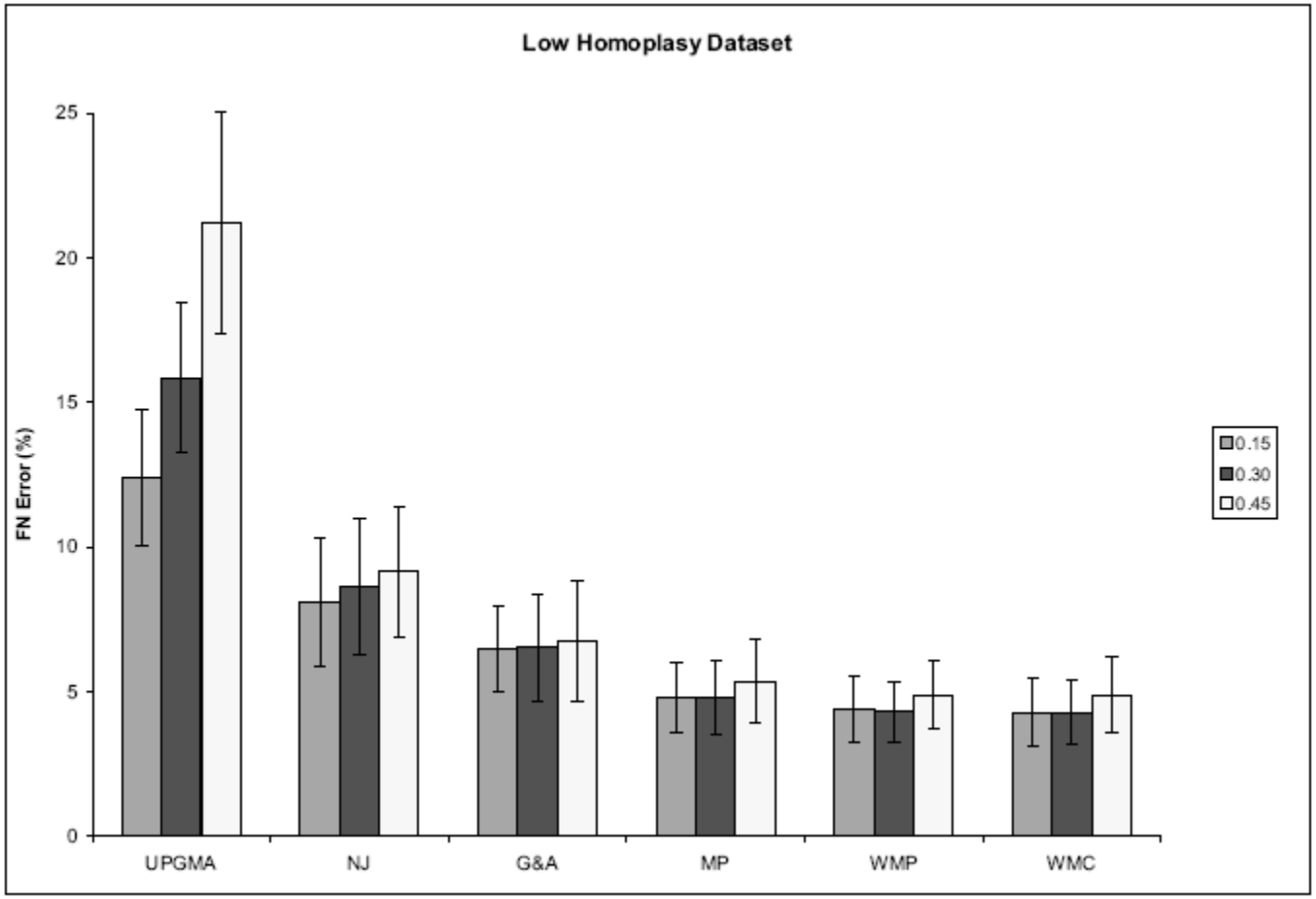
Impact of homoplasy for characters evolved down a tree under a moderate deviation from a lexical clock and moderate heterotachy. Our weighting is inappropriate for "unscreened" data.



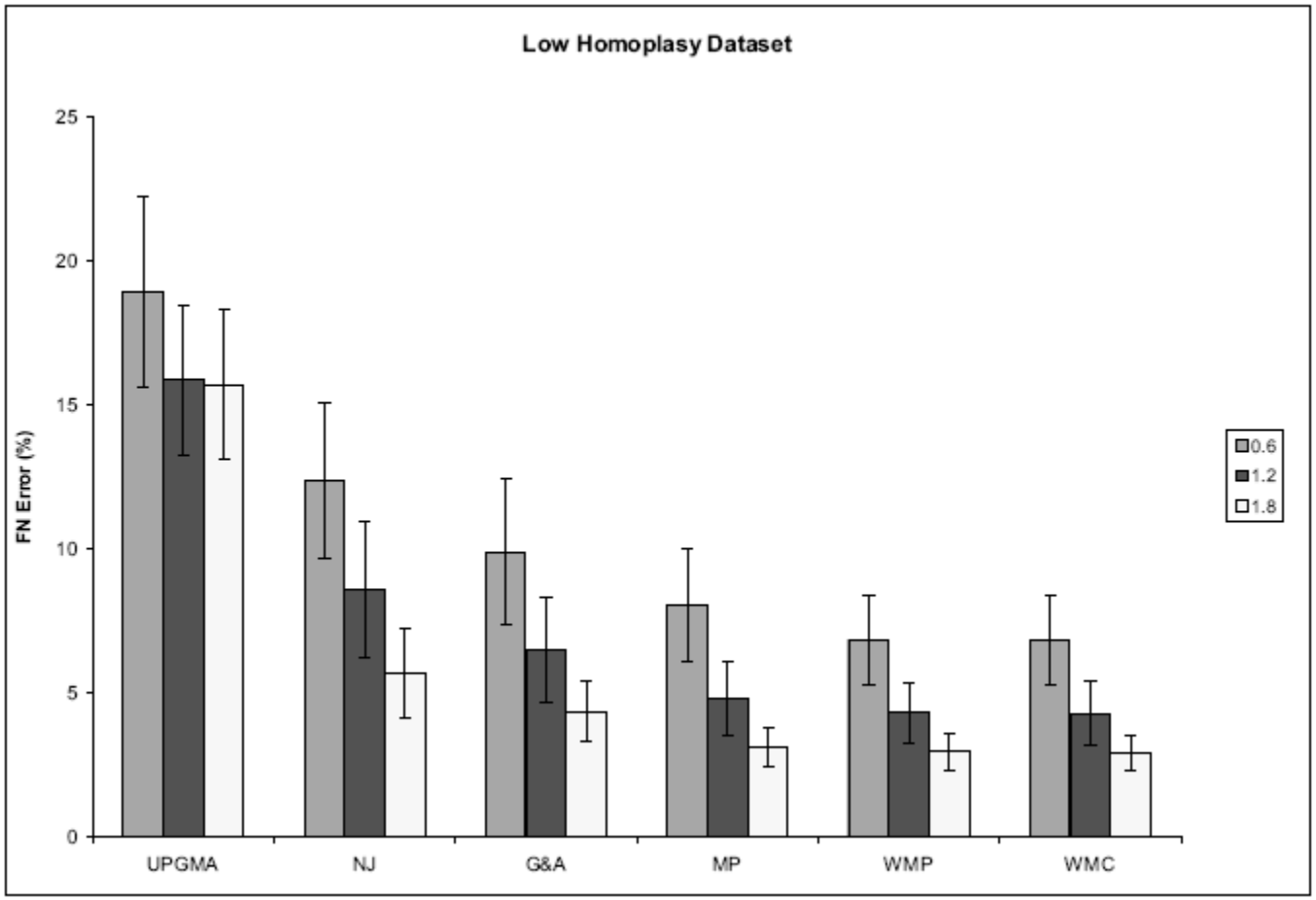
Impact of homoplasy for characters evolved down a network with three contact edges under a moderate deviation from the lexical clock and moderate heterotachy.



Impact of the number of contact edges for characters evolved under low homoplasy, moderate deviation from a lexical clock, and moderate heterotachy.



Impact of the deviation from a lexical clock for characters evolved down a network with three contact edges under low levels of homoplasy and with moderate heterotachy. We vary the deviation from a lexical clock from low to moderate.



Impact of heterotachy for characters evolved down a network with three contact edges, with low homoplasy, and with moderate deviation from a lexical clock. Heterotachy increases with the parameter.

Conclusions and comments

- Choice of reconstruction method does matter.
- Relative performance between methods is quite stable (distance-based methods worse than character-based methods).
- Choice of data does matter (good idea to add morphological characters).
- Accuracy only slightly lessened with small increases in homoplasy, borrowing, or deviation from the lexical clock.
- Some amount of heterotachy helps!

Future research

- Should we screen? The simulation uses low homoplasy as a proxy for screening, but real screening throws away data and may introduce bias.
- How do we detect/reconstruct borrowing?
- How do we handle missing data in methods based on stochastic models?
- How do we handle polymorphism?

Future research (cont.)

- We need more investigation of methods based on stochastic models (Bayesian beyond G+A, maximum likelihood, NJ with better distance corrections), as these are now the methods of choice in biology. This requires ***better models of linguistic evolution*** and hence ***input from linguists***!

Acknowledgements

- Funding: The Isaac Newton Institute, The US National Science Foundation, the David and Lucile Packard Foundation, the Radcliffe Institute for Advanced Studies, The Program for Evolutionary Dynamics at Harvard, and the Institute for Cellular and Molecular Biology at UT-Austin.

- Collaborators: Don Ringe, Steve Evans, Luay Nakhleh, and Francois Barbancon.

For more information

- Please see the Computational Phylogenetics for Historical Linguistics web site for papers, data, and additional material **http://www.cs.rice.edu/~nakhleh/CPHL**