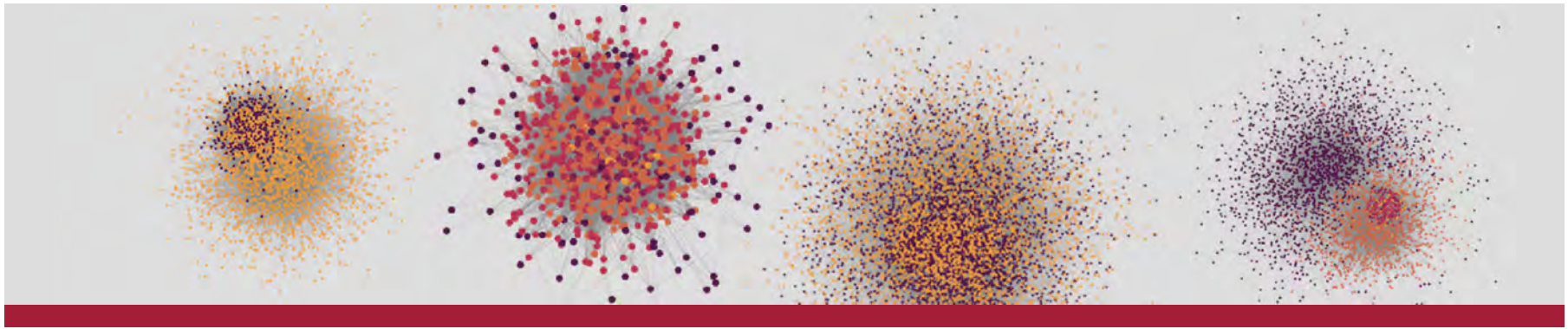
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
Seminar für Sprachwissenschaft

ERC Project CrossLingference



The effect of priors on tree topologies

26. February 2021

European Research Council
Established by the European Commission

The Phylogenetic Party

Linguistics: Inferring Trees

- **Inferring Language Family Trees**
  Grollemund et al. (2015); Bowern and Atkinson (2012)
- **Dating Language Families**
  Rama (2018); Chang, Cathcart, Hall, and Garrett (2015); Gray and Atkinson (2003)
- **Spread of Languages**
  Bouckaert et al. (2012)

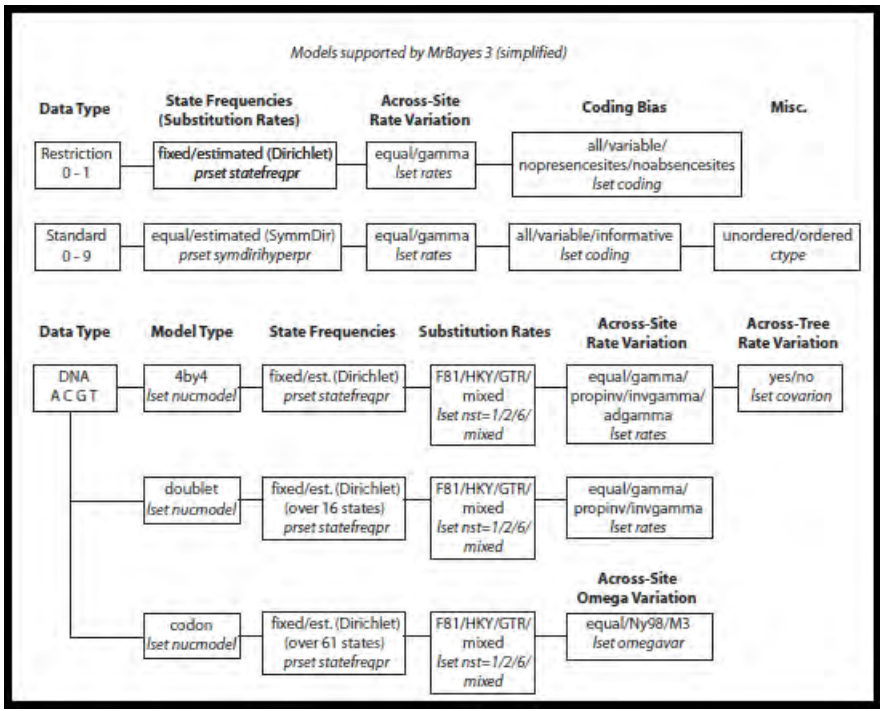
Linguistics: Using Trees

- **Lexical Change**
  Greenhill et al. (2017)
- **Reconstruction**
  Bouchard-Côté, Hall, Griffiths, and Klein (2013); Jäger and List (2018)
- **Comparative Studies**
  Calude and Verkerk (2016); Dunn, Greenhill, Levinson, and Gray (2011); Cathcart, Hölzl, Jäger, Widmer, and Bickel (2020)
- **Language Diversity**
  Bentz, Dediu, Verkerk, and Jäger (2018)

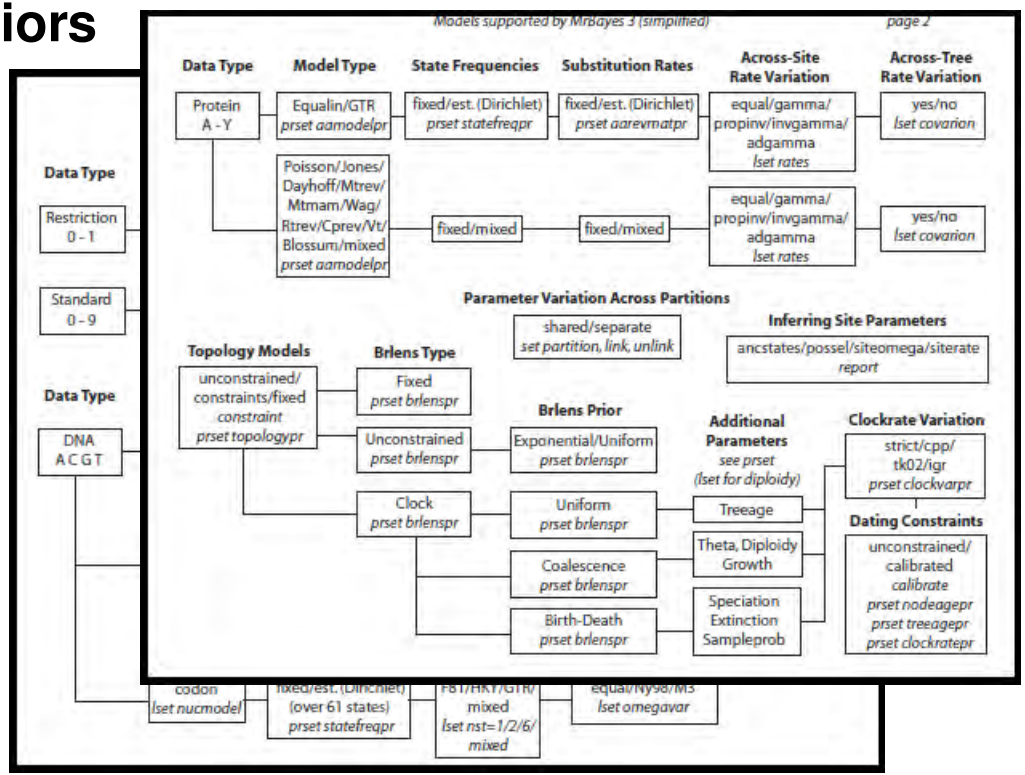
Checking Models!

- **Models and Data**
  Yanovich (2018); Rama, List, Wahle, and Jäger (2018)
- **The effect of Priors**
  Rama (2018)
- **Probability of “false positive” results**
  Rönchen & Wiklund (2021, MaEiQCL)

Priors over Priors



Priors over Priors



Different Priors

- Accross-Site Rate Variation
  - Equal Rates
  - Gamma Distributed
- Topology Priors
  - unconstrained
  - Birth Death Process + strict clock
  - Birth Death Process + Relaxed clock (independent gamma)
  - Uniform + strict clock
  - Uniform + Relaxed clock (independent gamma)

Data (c.f. Rama et al. (2018))

Dataset	# Meanings	# Languages
Austronesian (Greenhill, Blust, & Gray, 2008)	210	45
Austro-Asiatic (Sidwell, 2015)	200	58
Indo-European (Dunn, 2012)	208	42
Pama-Nyungan (Bowern & Atkinson, 2012)	183	67
Sino-Tibetan (Peiros, 2004)	110	64

Methods

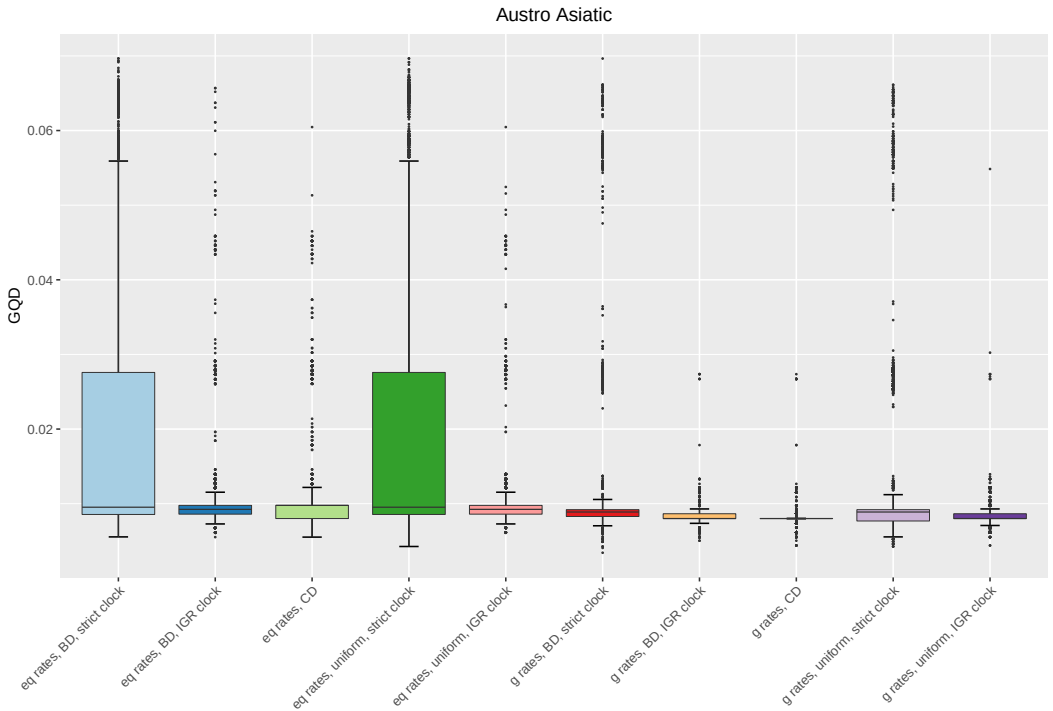
- Mr. Bayes (Ronquist & Huelsenbeck, 2003) + BEAGLE (Suchard & Rambaut, 2009)
  - 2 Runs, 4×10^6 generations, 25% burn-in, sample every 5000 generations
- spr-space (Whidden & Matsen, 2015) + Cytoscape (Shannon et al., 2003)

Analysis

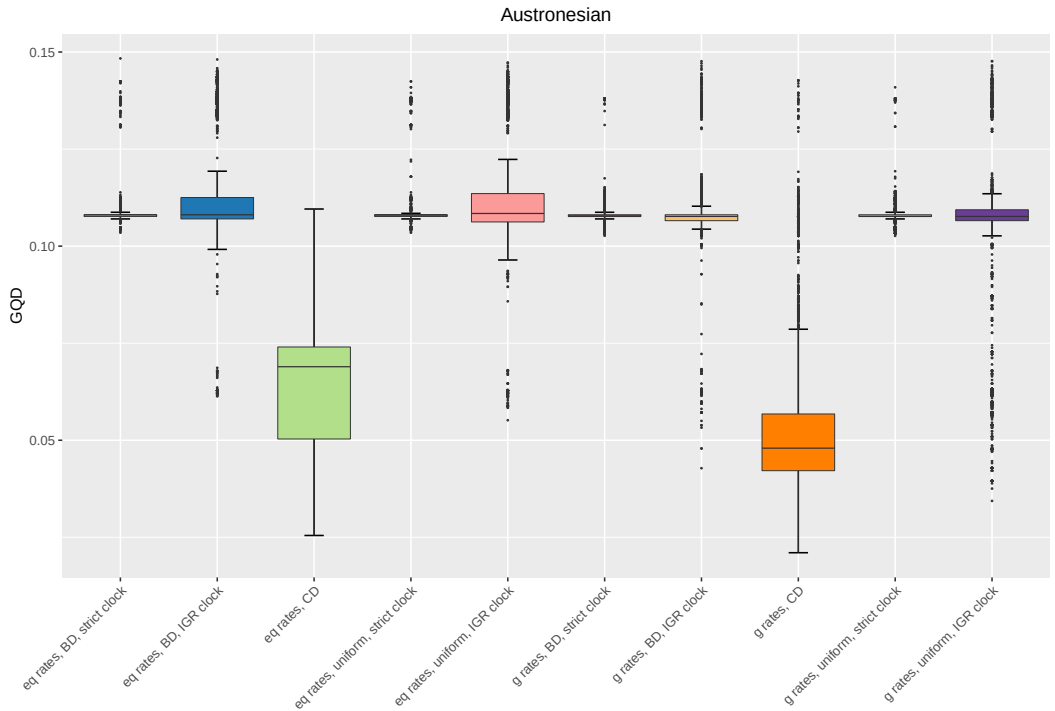
	Austroasiatic	Austronesian	Indo-European	Pamanungian	Sino-Tibetean
eq rates, BD, strict clock	0.02	0.11	0.02	0.13	0.06
eq rates, BD, IGR clock	0.01	0.11	0.03	0.14	0.07
eq rates, CD	0.01	0.06	0.04	0.14	0.08
eq rates, uniform, strict clock	0.02	0.11	0.02	0.13	0.07
eq rates, uniform, IGR clock	0.01	0.11	0.03	0.14	0.07
γ rates, BD, strict clock	0.01	0.11	0.02	0.13	0.06
γ rates, BD, IGR clock	0.01	0.11	0.03	0.14	0.05
γ rates, CD	0.01	0.05	0.03	0.14	0.06
γ rates, uniform, strict clock	0.01	0.11	0.02	0.13	0.06
γ rates, uniform, IGR clock	0.01	0.11	0.03	0.14	0.06

Mean Generalized quartet distances between trees in the posterior distribution and the gold standard tree. (Pompei, Loreto, & Tria, 2011)

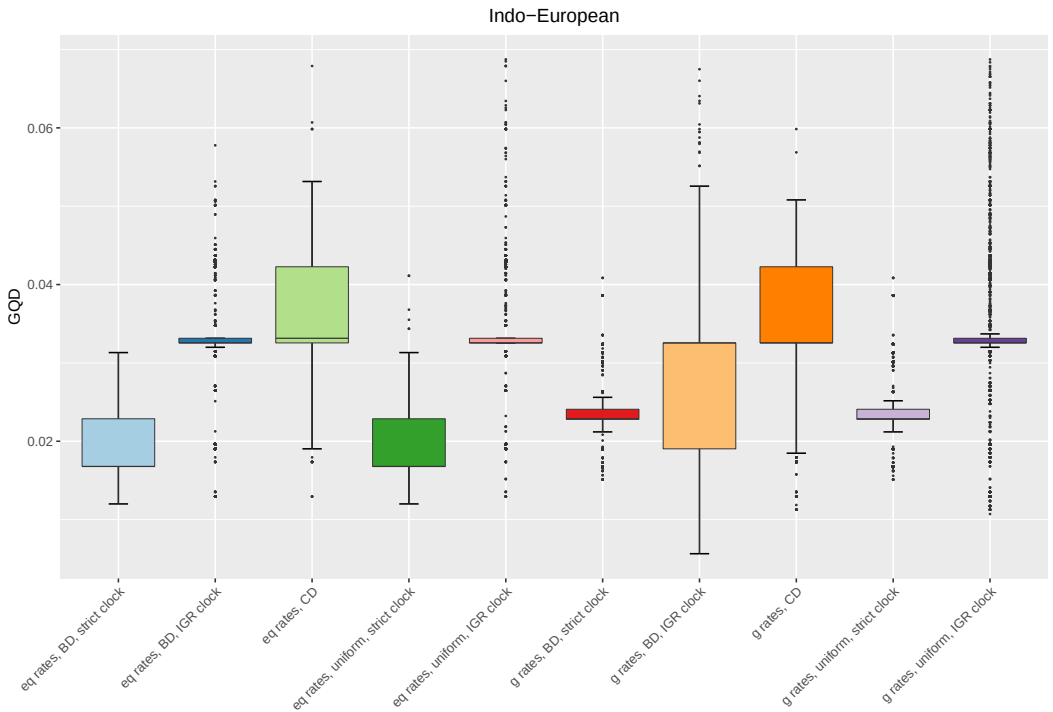
GQDs


Austro Asiatic
GQD
0.06
0.04
0.02
eq rates, BD, strict clock
eq rates, BD, IGR clock
eq rates, CD
eq rates, uniform, strict clock
eq rates, uniform, IGR clock
g rates, BD, strict clock
g rates, BD, IGR clock
g rates, CD
g rates, uniform, strict clock
g rates, uniform, IGR clock

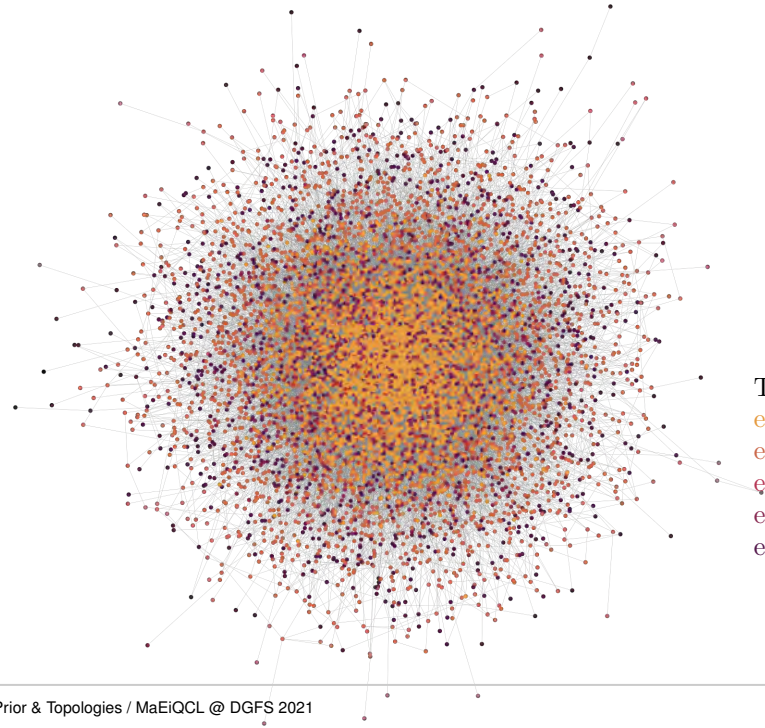
GQDs


Austronesian
GQD
0.15
0.10
0.05
eq rates, BD, strict clock
eq rates, BD, IGR clock
eq rates, CD
eq rates, uniform, strict clock
eq rates, uniform, IGR clock
g rates, BD, strict clock
g rates, BD, IGR clock
g rates, CD
g rates, uniform, strict clock
g rates, uniform, IGR clock

GQDs

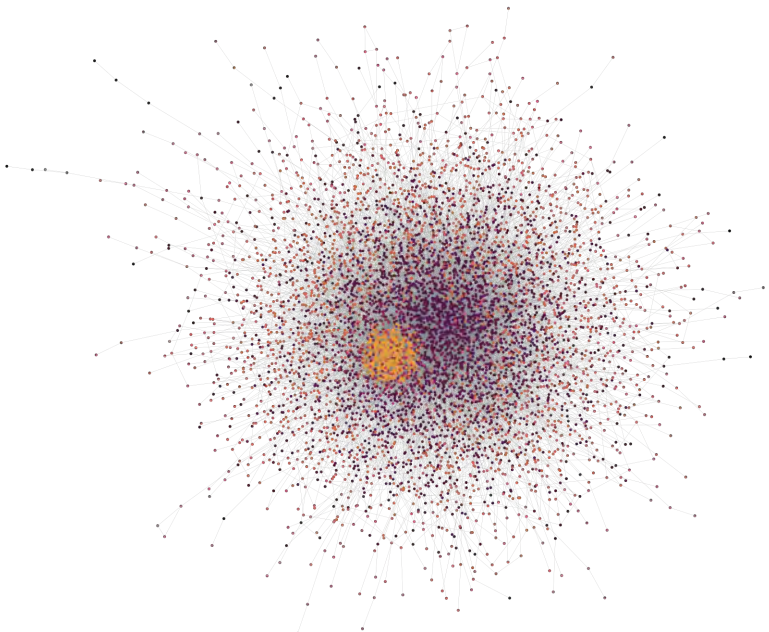

Indo−European
GQD
0.06
0.04
0.02
eq rates, BD, strict clock
eq rates, BD, IGR clock
eq rates, CD
eq rates, uniform, strict clock
eq rates, uniform, IGR clock
g rates, BD, strict clock
g rates, BD, IGR clock
g rates, CD
g rates, uniform, strict clock
g rates, uniform, IGR clock

SPR spaces - Austro Asiatic



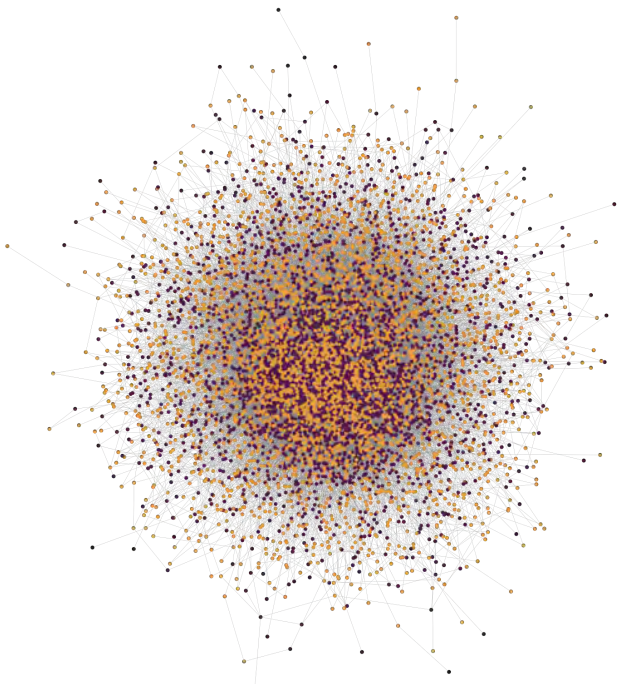
Topologies
eq rates, BD, strict clock: 2781
eq rates, BD, IGR clock: 6698
eq rates, CD: 3016
eq rates, uniform, strict clock: 3933
eq rates, uniform, IGR clock: 7706

SPR spaces - Austro Asiatic



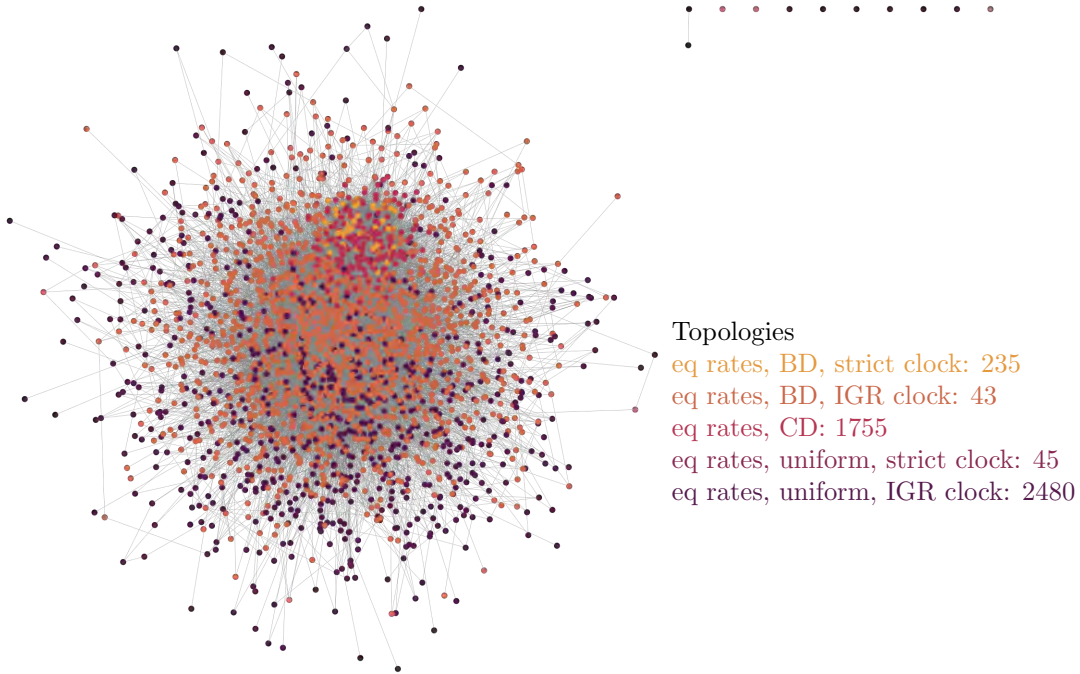
Topologies
γ rates, BD, strict clock: 388
γ rates, BD, IGR clock: 2795
γ rates, CD: 475
γ rates, uniform, strict clock: 672
γ rates, uniform, IGR clock: 3881

SPR spaces - Austro Asiatic

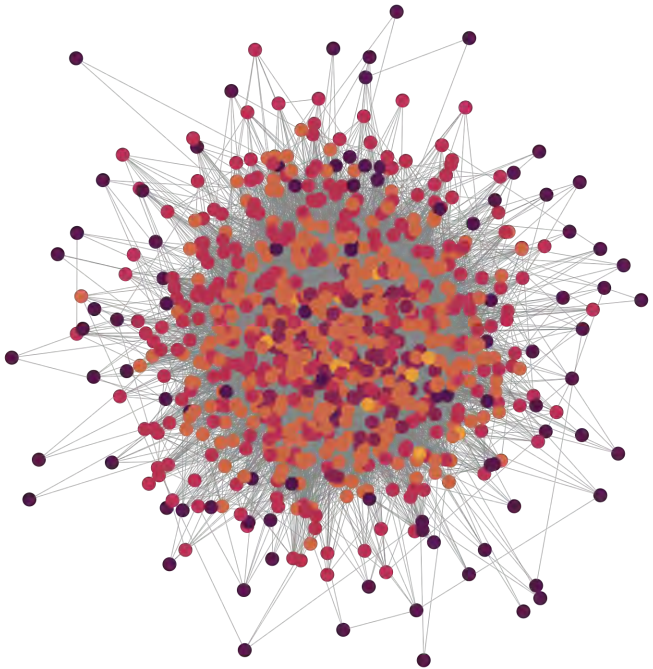


Topologies
eq rates: 7706
γ rates: 6698

SPR spaces - Austronesian

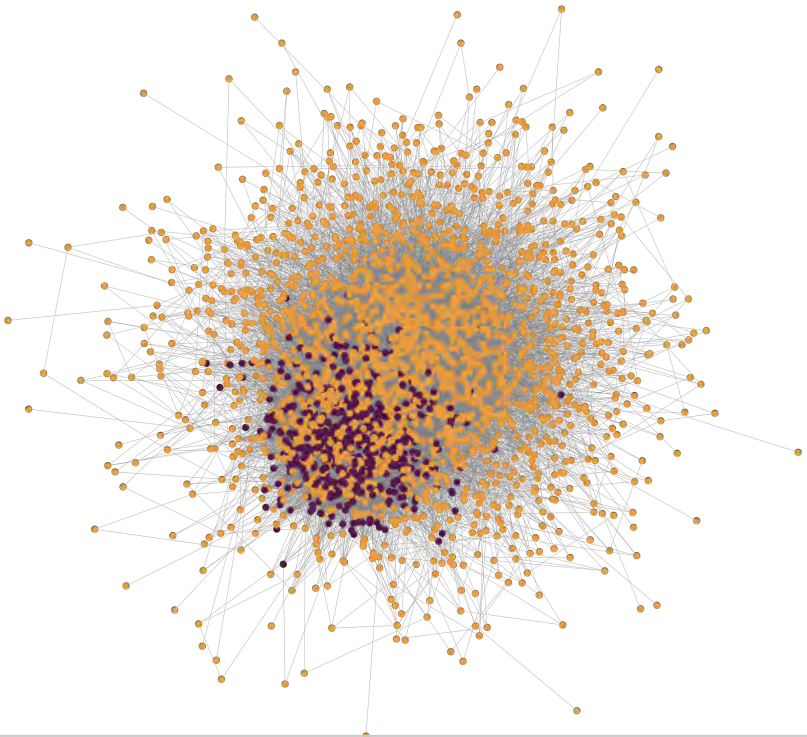


SPR spaces - Austronesian



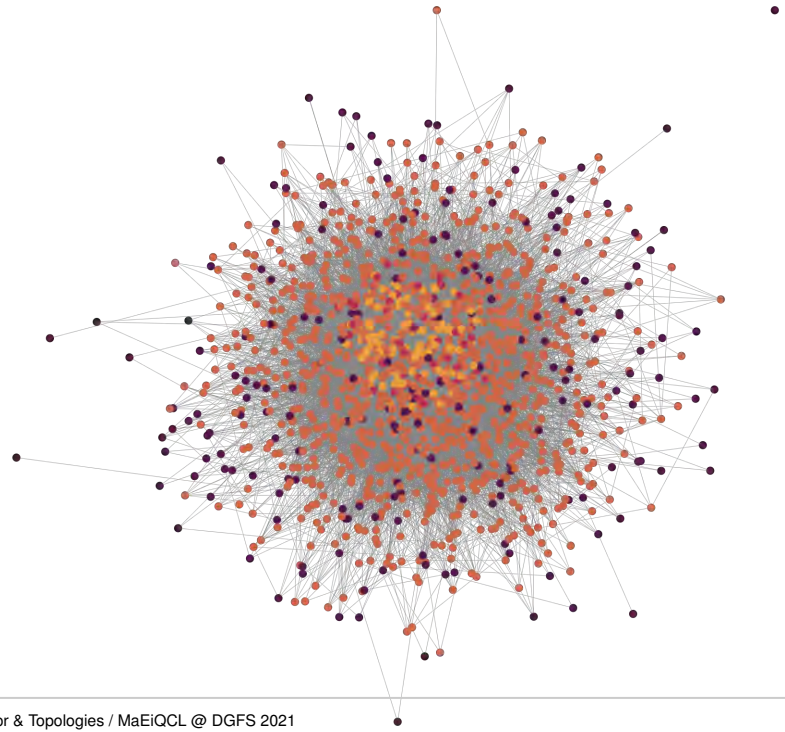
Topologies
γ rates, BD, strict clock: 100
γ rates, BD, IGR clock: 448
γ rates, CD: 758
γ rates, uniform, strict clock: 88
γ rates, uniform, IGR clock: 854

SPR spaces - Austronesian



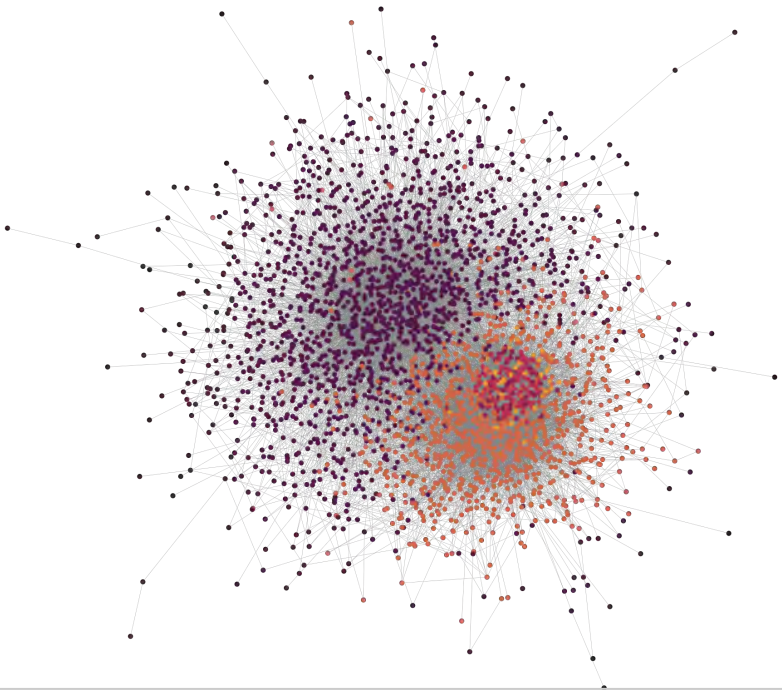
Topologies
eq rates: 2480
γ rates: 854

SPR spaces - Indo European



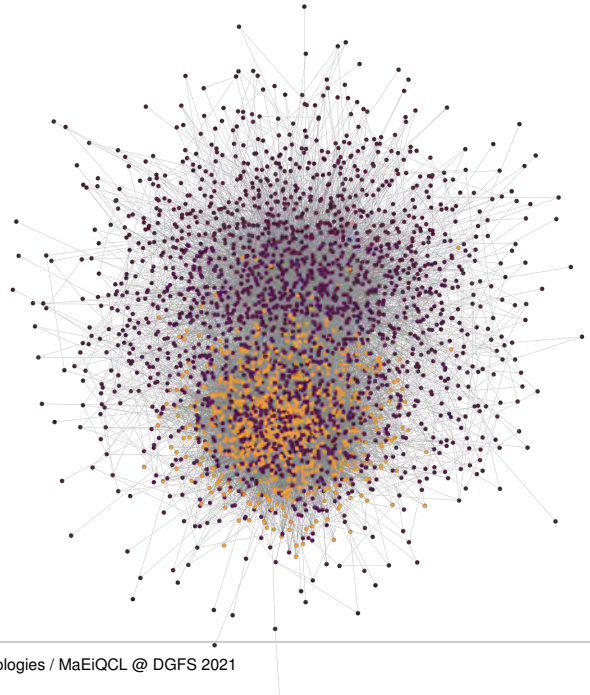
Topologies
eq rates, BD, strict clock: 105
eq rates, BD, IGR clock: 1230
eq rates, CD: 156
eq rates, uniform, strict clock: 107
eq rates, uniform, IGR clock: 1426

SPR spaces - Indo European



Topologies
γ rates, BD, strict clock: 256
γ rates, BD, IGR clock: 1194
γ rates, CD: 138
γ rates, uniform, strict clock: 210
γ rates, uniform, IGR clock: 2957

SPR spaces - Indo European



Topologies
eq rates: 1426
γ rates: 2957

Summary

- In terms of the GQD (almost) all models/priors perform equally well
- However the posterior distributions of tree topologies differ
  - Caveat: Relaxed Clock (IGR) models explore the most tree topologies
  - Across Site rate variation shows a mixed picture
- We can investigate the posterior distribution of trees and detect pathological behavior or regions

References I

Bentz, C., Dediu, D., Verkerk, A., & Jäger, G. (2018, nov). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, *2*(11), 816–821. doi: 10.1038/s41562-018-0457-6

Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013, feb). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, *110*(11), 4224–4229. doi: 10.1073/pnas.1204678110

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., … Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960.

Bowern, C., & Atkinson, Q. D. (2012). Computational phylogenetics of the internal structure of pama-nguyan. *Language*, *88*(4), 817–845. doi: 10.1353/lan.2012.0081

Calude, A. S., & Verkerk, A. (2016, apr). The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution*, *1*(2), 91–108. doi: 10.1093/jole/lzw003

Cathcart, C. A., Hölzl, A., Jäger, G., Widmer, P., & Bickel, B. (2020, oct). Numeral classifiers and number marking in indo-iranian. *Language Dynamics and Change*, 1–53. doi: 10.1163/22105832-bja10013

Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis. *Language*, *91*(1), 194–244.

Dunn, M. (2012). *Indo-european lexical cognacy database (ielex).*

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011, apr). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82. doi: 10.1038/nature09923

Gray, R. D., & Atkinson, Q. D. (2003, November). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, *426*(6965), 435-439.

Greenhill, S. J., Blust, R., & Gray, R. D. (2008, November). The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evolutionary bioinformatics online*, *4*, 271–283.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*. Retrieved from `https://www.pnas.org/content/early/2017/10/02/1700388114` doi: 10.1073/pnas.1700388114

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., & Pagel, M. (2015, sep). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, *112*(43), 13296–13301. doi: 10.1073/pnas.1503793112

Jäger, G., & List, J.-M. (2018, jun). Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change*, *8*(1), 22–54. doi: 10.1163/22105832-00801002

Peiros, I. (2004). *Dataset on sino-tibetan languages encoded in starling.*

Pompei, S., Loreto, V., & Tria, F. (2011). On the accuracy of language trees. *PloS one*, *6*(6), e20109.

Rama, T. (2018). Three tree priors and five datasets. *Language Dynamics and Change*, *8*(2), 182–218. doi: 10.1163/22105832-00802005

Rama, T., List, J.-M., Wahle, J., & Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the north american chapter of the association of computational linguistics* (p. 393-400). Retrieved from `https://aclanthology.coli.uni-saarland.de/papers/N18-2063/n18-2063`

Ronquist, F., & Huelsenbeck, J. P. (2003, August). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*, *19*, 1572–1574.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003, nov). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. doi: 10.1101/gr.1239303

Sidwell, P. (2015). Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies (Notes, Reviews, Data-Papers)*, *44*, lxviii–ccclvii.

Suchard, M. A., & Rambaut, A. (2009, apr). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, *25*(11), 1370–1376. doi: 10.1093/bioinformatics/btp244

Whidden, C., & Matsen, F. A. (2015, jan). Quantifying MCMC exploration of phylogenetic tree space. *Systematic Biology*, *64*(3), 472–491. doi: 10.1093/sysbio/syv006

Yanovich, I. (2018, jun). The effect of dictionary omissions on phylogenies computationally inferred from lexical data. *Language Dynamics and Change*, *8*(1), 78–107. doi: 10.1163/22105832-00801007