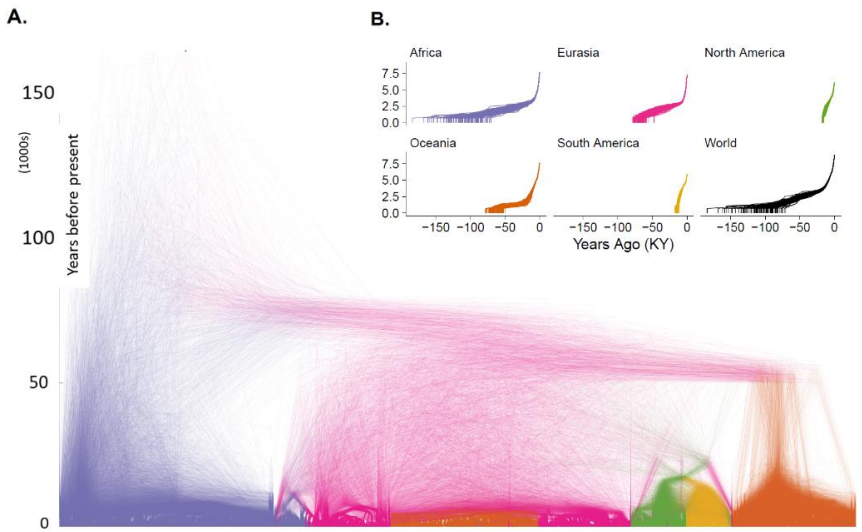
The forest or the supertree:

does the world tree make the same predictions about linguistic change as family-specific phylogenies?

SARAH BABINSKI[1], ALEXANDRU CRAEVSCHI[1], SIRUI CHENG[2], CHUNDRA CATHCART[1]

(1) ISLE, UNIVERSITY OF ZURICH
(2) CORPUS CHRISTI COLLEGE, UNIVERSITY OF OXFORD

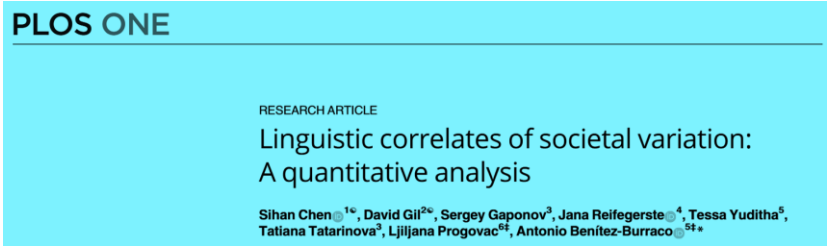
New Advances in Phylolinguistics
11-12 June 2025



The world tree

- Bouckaert et al. (2022)
- Global phylogeny of over 6000 languages
- Family-level relationships more defined, with very long branches connecting these
- Benefit: inferences can be made on a global scale, including small families and isolates

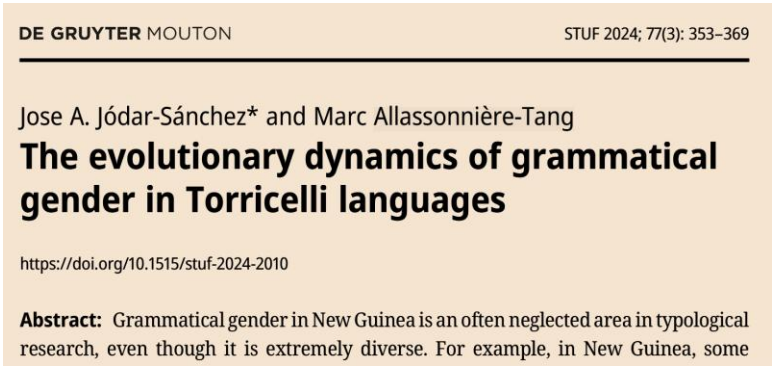
Bouckaert et al. (2022). Global language diversification is linked to socio-ecology and threat status.


PLOS ONE

RESEARCH ARTICLE

Linguistic correlates of societal variation: A quantitative analysis

Sihan Chen[1☯], David Gil[2☯], Sergey Gaponov[3], Jana Reifegerste[4], Tessa Yuditha[5], Tatiana Tatarinova[3], Ljiljana Progovac[6‡], Antonio Benítez-Burraco[5‡*]


DE GRUYTER MOUTON STUF 2024; 77(3): 353–369

Jose A. Jódar-Sánchez* and Marc Allassonnière-Tang

The evolutionary dynamics of grammatical gender in Torricelli languages

https://doi.org/10.1515/stuf-2024-2010

Abstract: Grammatical gender in New Guinea is an often neglected area in typological research, even though it is extremely diverse. For example, in New Guinea, some


Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02506-z OPEN

Check for updates

Early humans out of Africa had only base-initial numerals

One-Soon Her[1,2], Yung-Ping Liang[2], Eugene Chan[3], Hung-Hsin Hsu[2,4], Anthony Chi-Pin Hsu[1] & Marc Allassonnière-Tang[5✉]

The world tree

- This tree has already been used to make big claims
  - Connections of language + societal structure (Chen et al. 2024)
  - Change in a small family (Jódar-Sánchez & Allassonnière-Tang 2024)
  - Inferring ancestral features of early human language (Her et al. 2024)

The world tree

Unanswered questions:

1. Does a global phylogeny make the same predictions as those made using many single-lineage trees?

2. Are there any systematic differences in rate inference due to the long branch lengths the world tree requires?

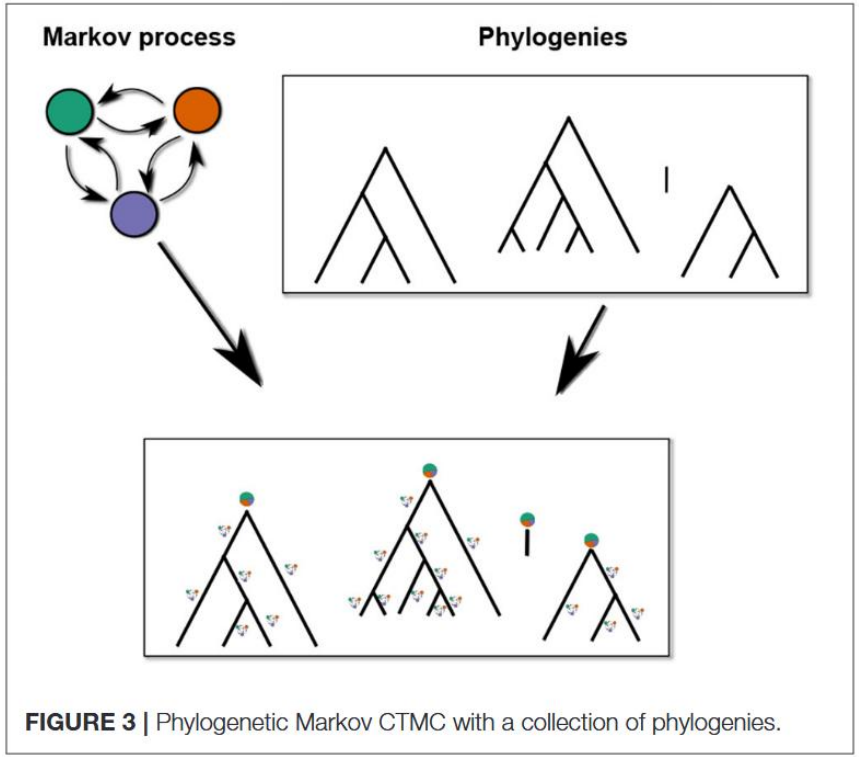


FIGURE 3 | Phylogenetic Markov CTMC with a collection of phylogenies.

An alternative method

Jäger & Wahle (2021): “independent but identical” CTMC processes run on each branch of each phylogeny

- Also accounts for small families and isolates by using a joint likelihood across phylogenies
- Does not have the ‘long branch problem’
- Is overall more computationally efficient

Jäger, G., & Wahle, J. (2021). Phylogenetic Typology. *Frontiers in Psychology*.

The current study

- Testing rate inference on Bouckaert et al. (2022)'s global phylogeny against single-lineage subtrees of this phylogeny using Jäger & Wahle (2021)'s method
- Inferring *rates of change* of GramBank features across these two regimes
- Comparing *correlated evolution* for some likely-correlated and likely-uncorrelated feature pairs

Methods

- GramBank: 195 features, most are binary-coded
  - Around 2,500 language varieties
- Two models of evolution:
  - ‘Non-covarion’ CTMC
  - Covarion: ‘hot’ and ‘cold’ regimes within each CTMC
- Two types of phylogeny:
  - **Global:** Bouckaert et al. (2022)’s world tree
  - **Local:** Separate lineages, extracted from the global phylogeny

Methods

- Method from Jäger & Wahle (2021):
  - “independent but identical” CTMC processes run on each branch of each phylogeny
  - Joint likelihood across phylogenies determined using Felsenstein’s (1981) pruning algorithm
- Models were run in Stan
- For the global phylogeny, basic setup is similar but there is only one tree

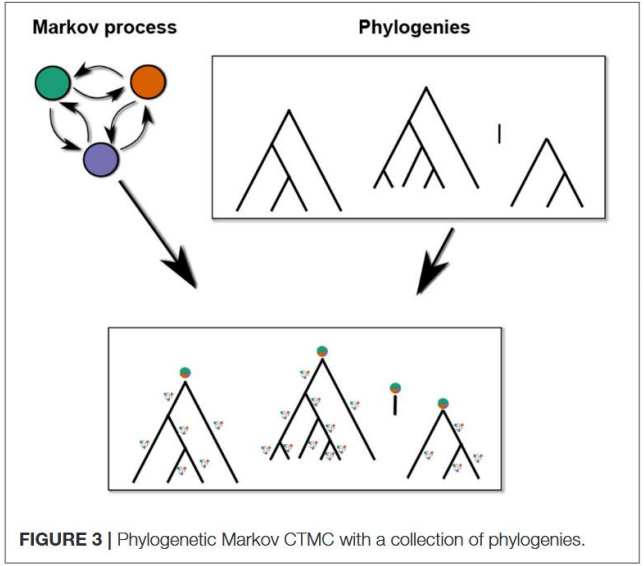


FIGURE 3 | Phylogenetic Markov CTMC with a collection of phylogenies.

Jäger, G., & Wahle, J. (2021). Phylogenetic Typology. *Frontiers in Psychology.*
Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters. *Evolution.*

Correlated evolution

- **Do these phylogeny types make the same predictions about correlated evolution?**
- 10 pairs of likely-correlated features, and 10 pairs likely-uncorrelated features
  - Determined in part by mutual information scores
- Determine Q matrices using Pagel's (1994) method
- Calculate *phi coefficient* of correlation
  - On a scale from -1 to 1
  - Also from Jäger & Wahle (2021)

$$\frac{\mathrm{cov}(f_1, f_2)}{\sqrt{\mathrm{var}(f_1)\mathrm{var}(f_2)}} = \frac{p_{00}p_{11} - p_{10}p_{01}}{\sqrt{(p_{00}+p_{01})(p_{10}+p_{11})(p_{00}+p_{10})(p_{10}+p_{11})}}$$

Pagel, M. (1994). Detecting correlated evolution on phylogenies. *Proceedings of the Royal Society of London*.

Correlated Features

Feature 1	Feature 2
37 Is there a postposed complementizer in complements of verbs of thinking and/or knowing?	64 Is there a preposed complementizer in complements of verbs of thinking and/or knowing?
92 Is there grammatical marking of direct evidence (perceived with the senses)?	93 Is there grammatical marking of indirect evidence (hearsay, inference, etc.)?
139 Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?	169 Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass/exceed' meaning?
31 Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?	173 Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?
61 Is there a non-bound comparative degree marker modifying the property word in a comparative construction?	166 Is there a bound comparative degree marker on the property word in a comparative construction?
169 Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass/exceed' meaning?	178 Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?
139 Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?	178 Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?
65 Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?	167 Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?
3 Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?	173 Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?
3 Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?	31 Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?

Non-correlated Features

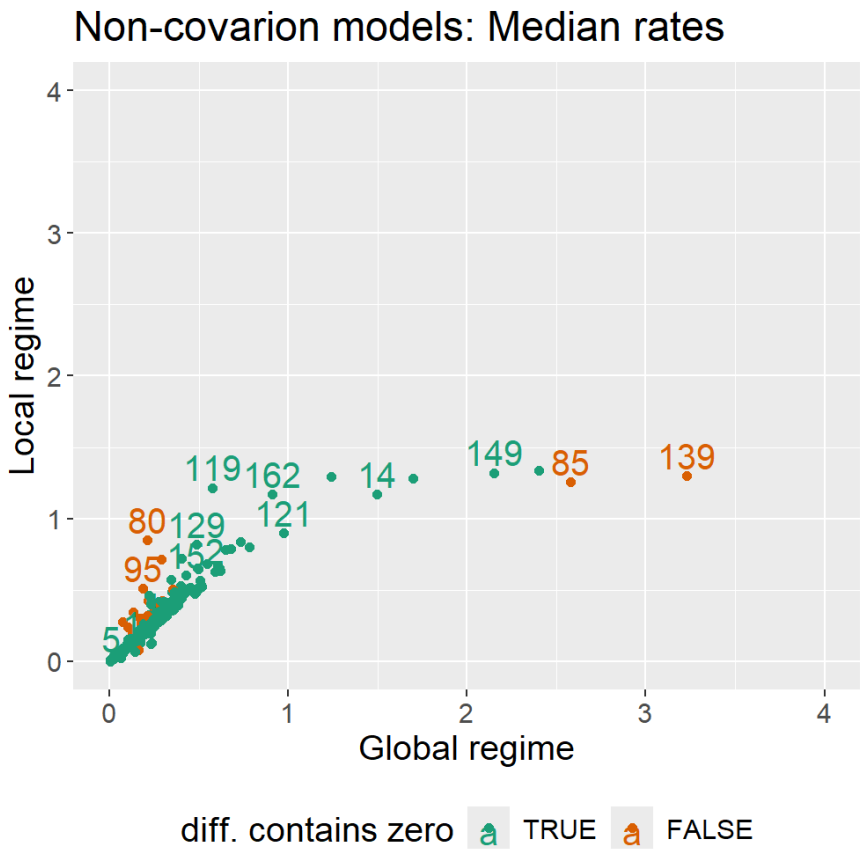
Feature 1	Feature 2
7 Do indefinite nominals commonly have indefinite articles?	74 Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferable from context ("pro-drop" or "null anaphora")?
7 Do indefinite nominals commonly have indefinite articles?	179 Can adnominal possession be marked by a suffix on the possessor?
7 Do indefinite nominals commonly have indefinite articles?	97 Can adnominal possession be marked by a prefix on the possessed noun?
7 Do indefinite nominals commonly have indefinite articles?	56 Can adnominal possession be marked by a suffix on the possessed noun?
56 Can adnominal possession be marked by a suffix on the possessed noun?	183 Is there a morphologically marked inverse on verbs?
7 Do indefinite nominals commonly have indefinite articles?	177 Is there any ergative alignment of flagging?
4 Can adnominal possession be marked by a prefix on the possessor?	183 Is there a morphologically marked inverse on verbs?
7 Do indefinite nominals commonly have indefinite articles?	102 Is there any accusative alignment of flagging?
4 Can adnominal possession be marked by a prefix on the possessor?	7 Do indefinite nominals commonly have indefinite articles?
7 Do indefinite nominals commonly have indefinite articles?	21 Is there any neutral alignment of flagging?

Results

RATE INFERENCE

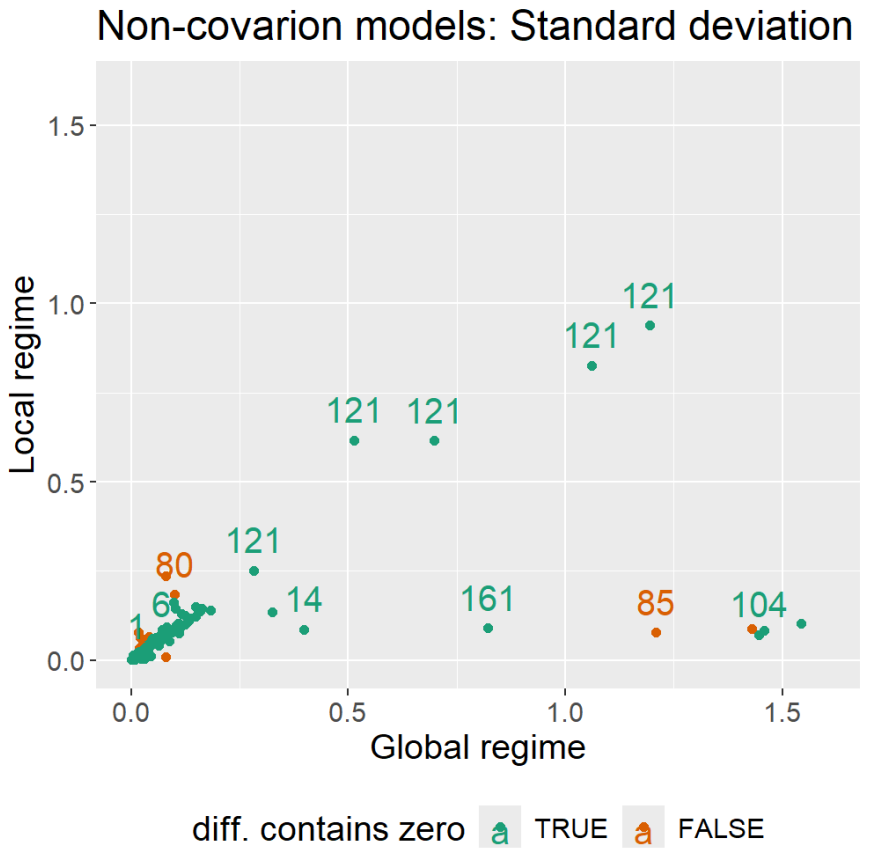
Non-covarion models

- Median rates are indistinguishable from one another most of the time
- Some outliers, most of which have a higher global median compared to local
  - (but there are outliers in the other direction too)



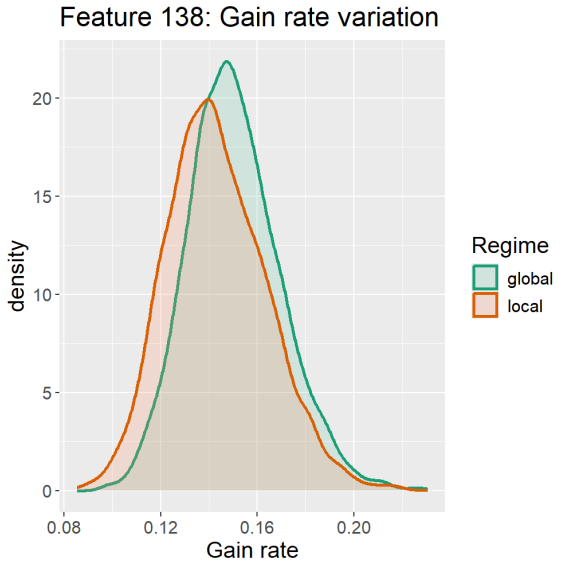
Non-covarion models

- Standard deviations are also very similar in most cases
- But some outliers exist, usually with higher SD in the global regime



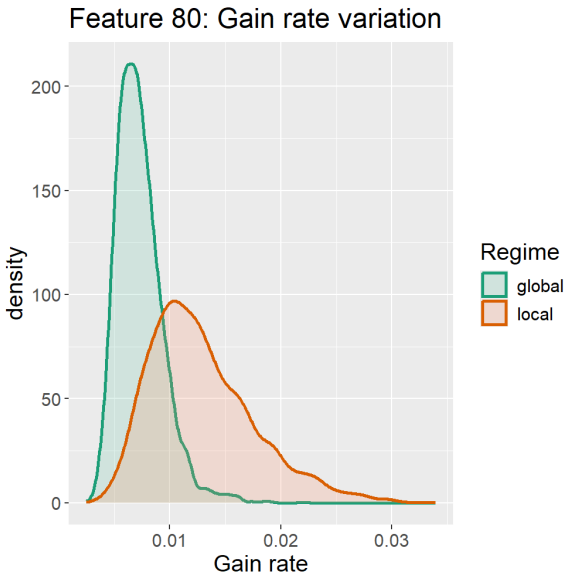
Rate variation: Similar distributions

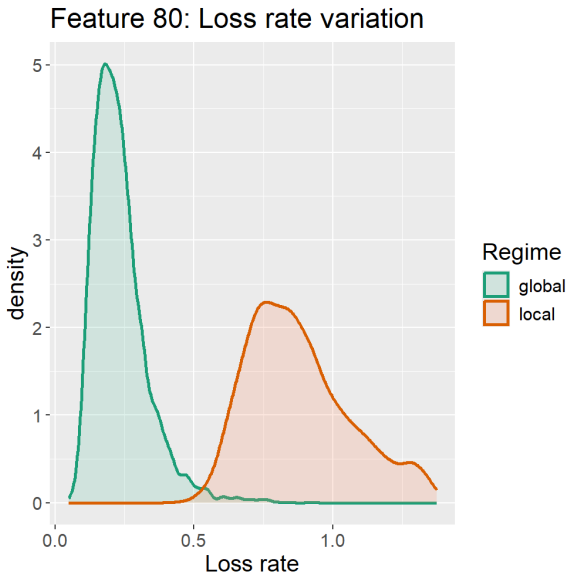
- Usual case: rate distributions are very similar in both median and SD



Rate variation: Local median > global median

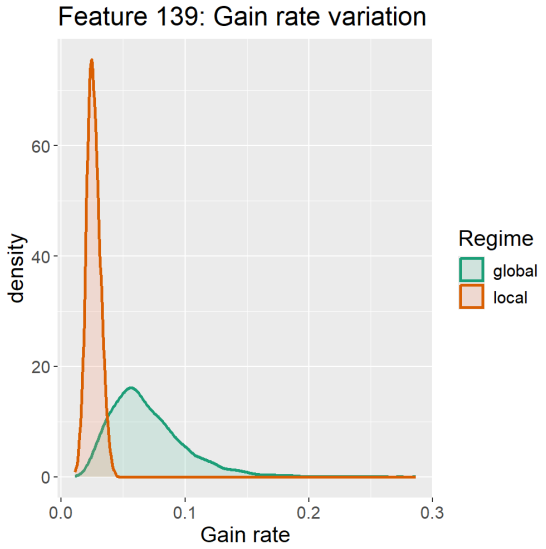
- Some outliers show higher local median with a more diffuse distribution

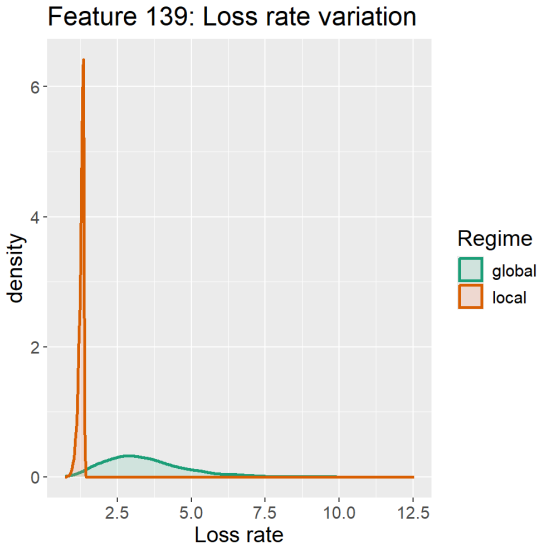




Rate variation: Global median > local median

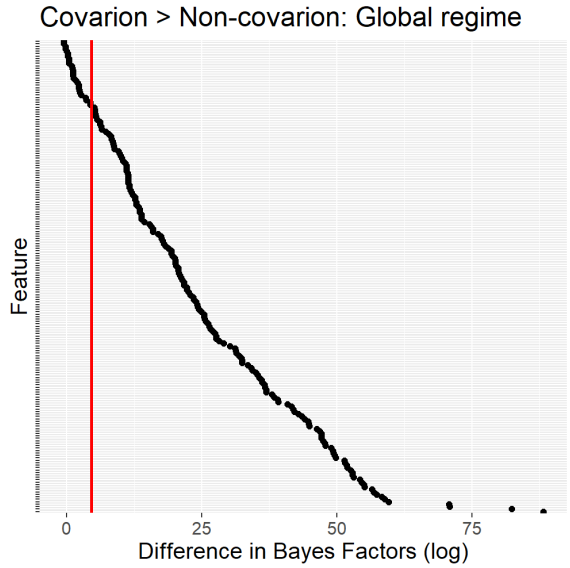
- Other outliers have higher global median with a more diffuse distribution

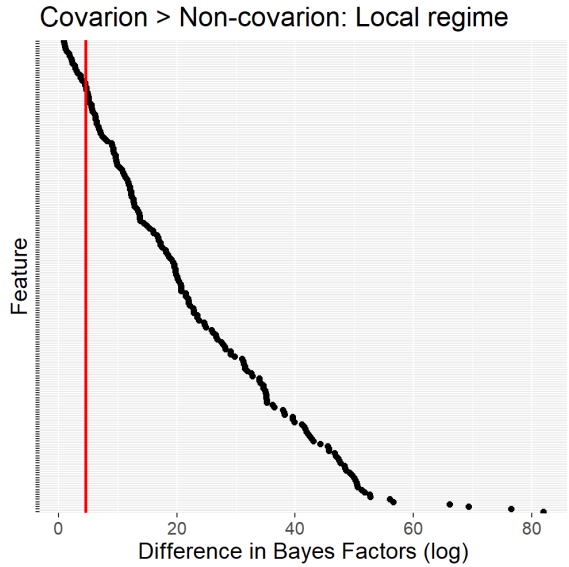




Covarion > Non-covarion

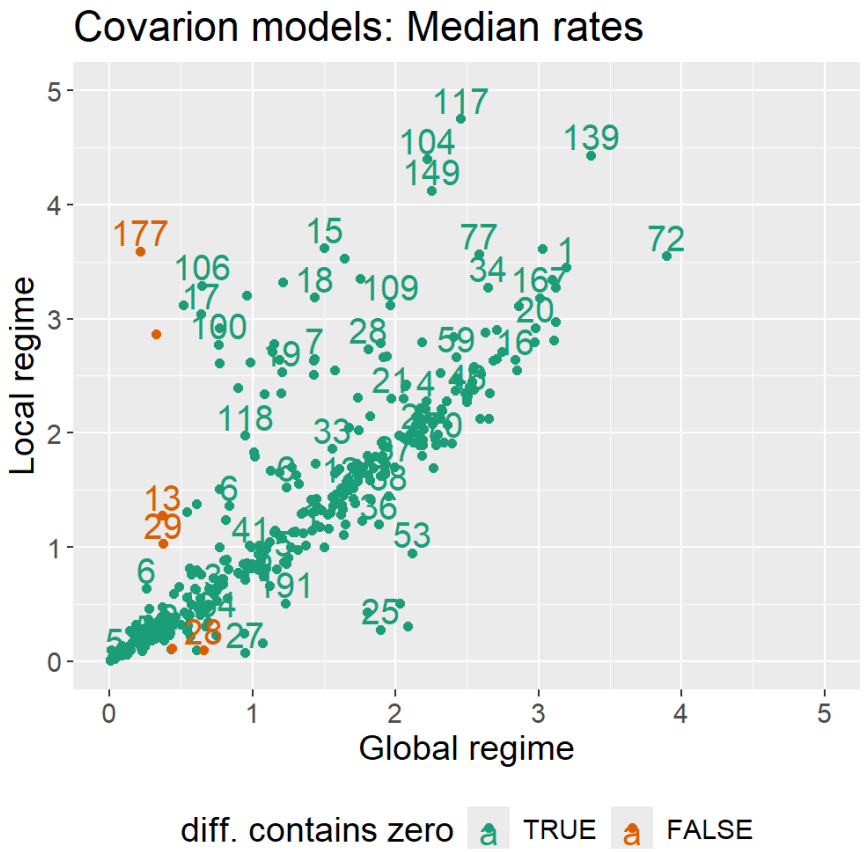
- Covarion models had consistently better fit to the data than non-covarion models





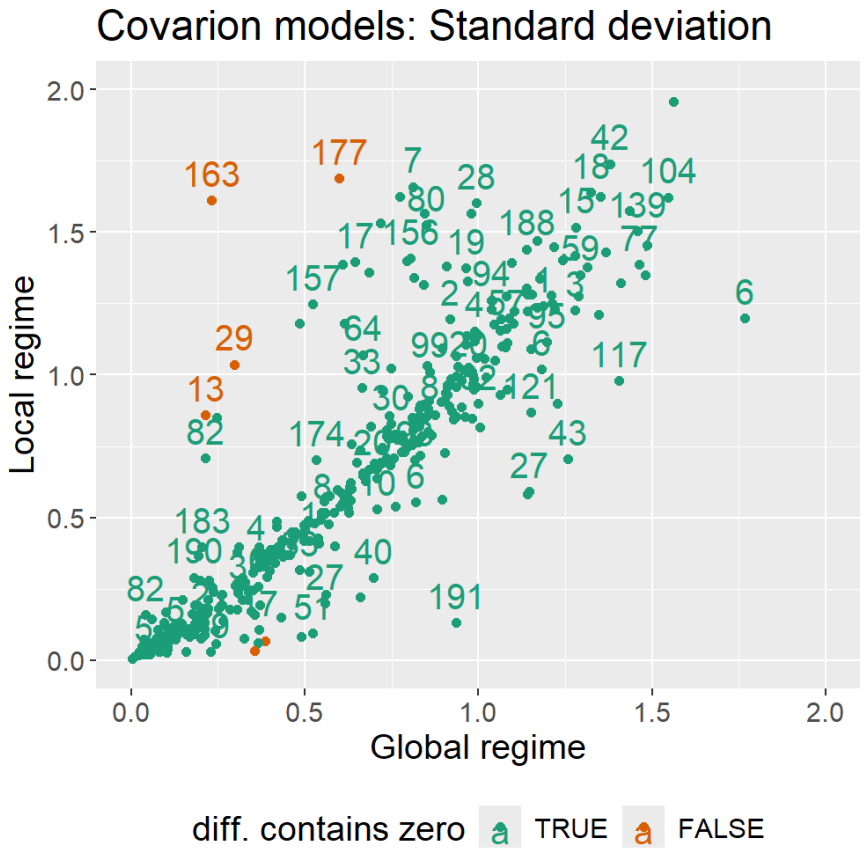
Covarion models

- Overall distribution of rate medians differs from non-covarion medians
- When rate medians differ, usually the global median is lower than the local



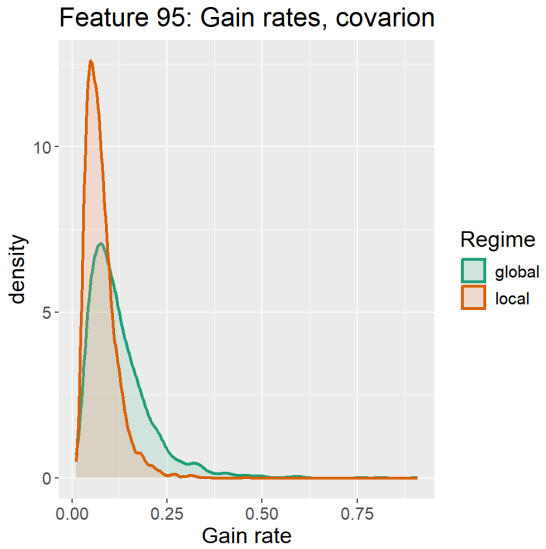
Covarion models

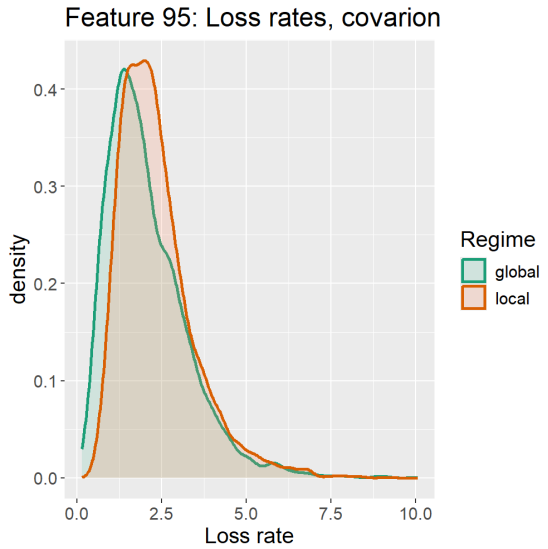
The global regime is also more likely to have a smaller SD than the local regime



Rate variation: similar distributions

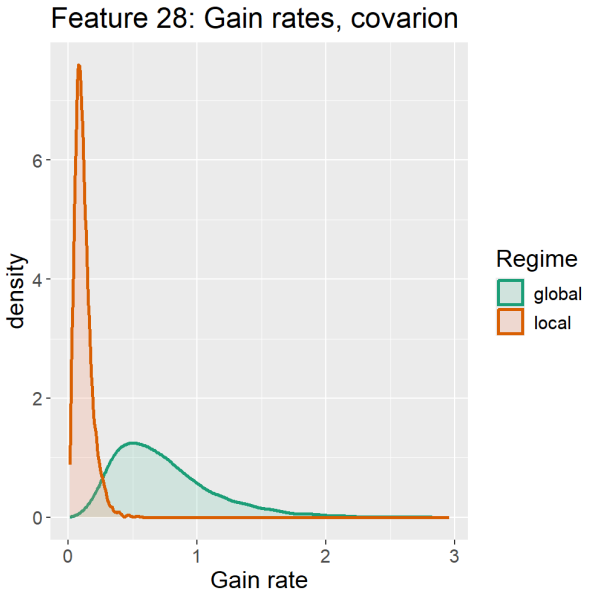
- Usual case: both gain and loss rates are similar across regimes

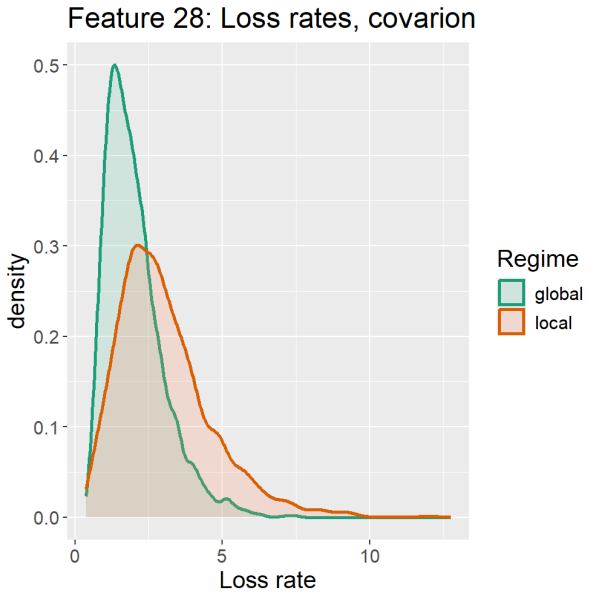




Rate variation: global > local

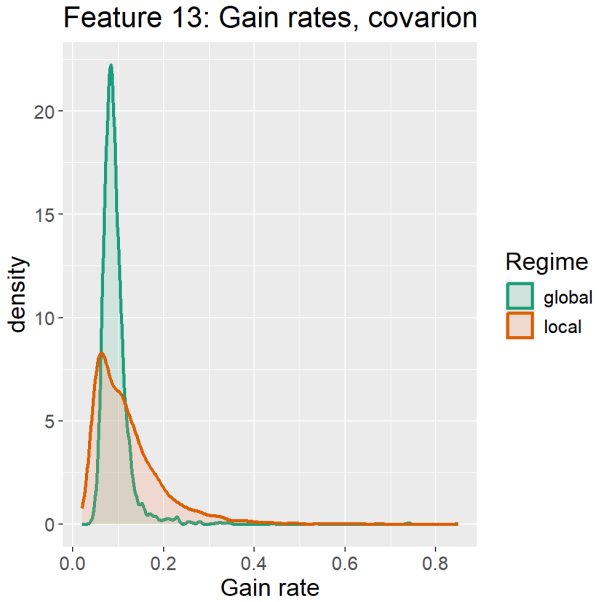
- Global rates are higher in **gain rates**

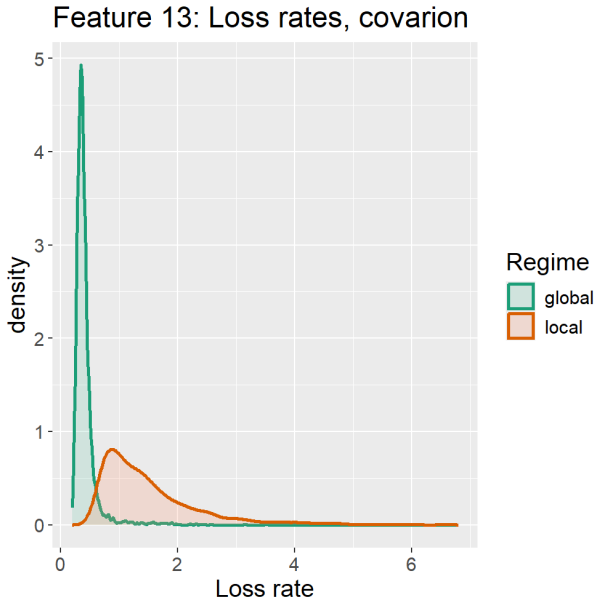




Rate variation: local > global

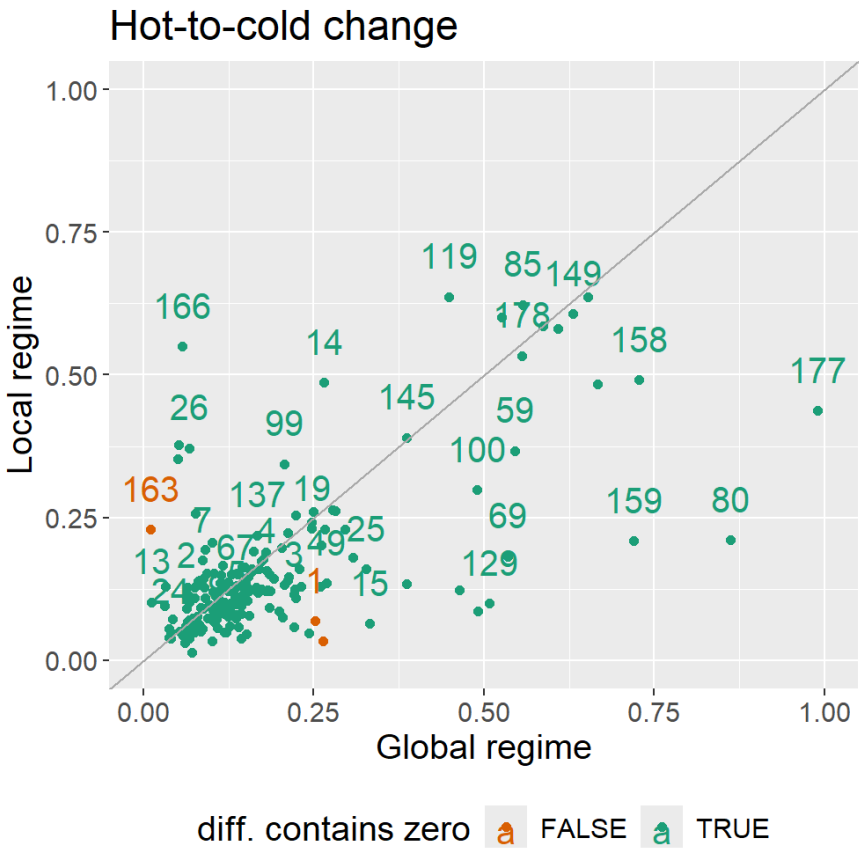
- Local rates are higher in **loss rates**





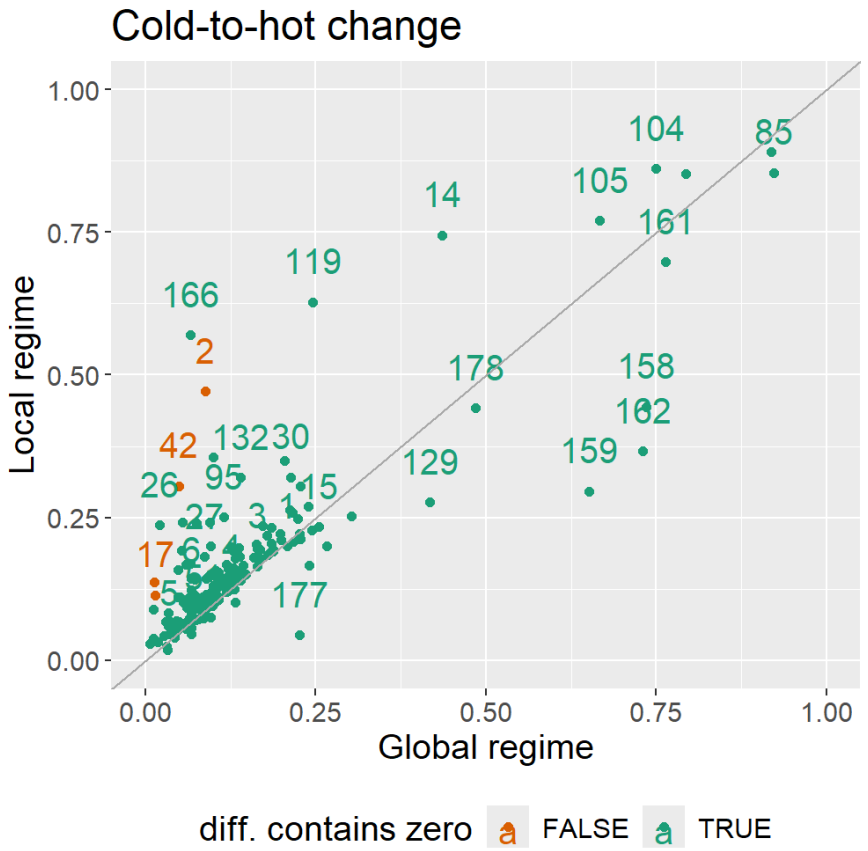
Covarion models

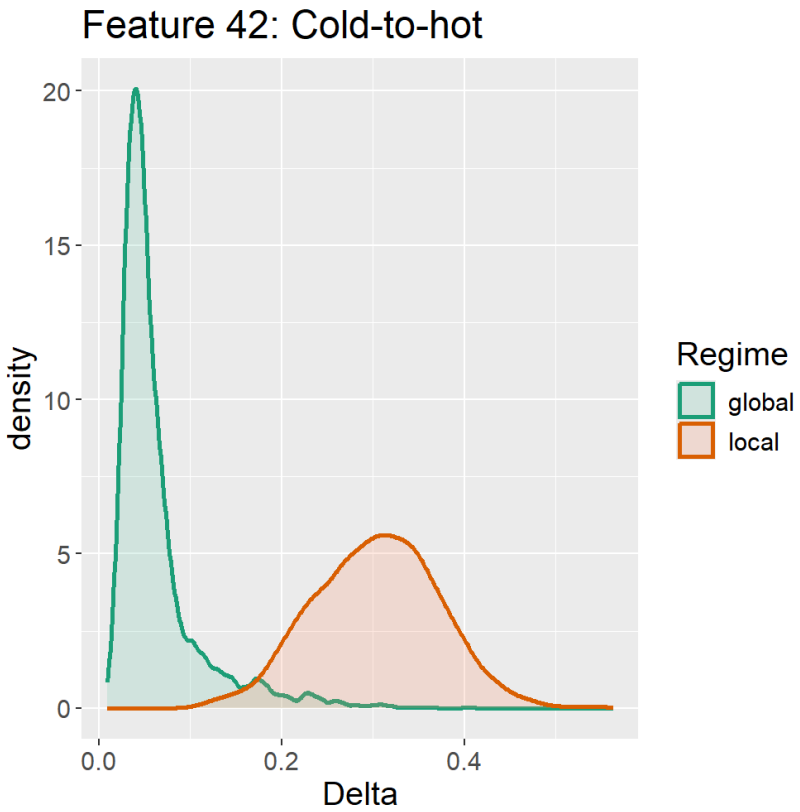
- Change rate from ‘hot’ to ‘cold’ regimes is variable, but not in any particular direction
  - Sometimes global rates are higher, sometimes local ones are



Covarion models

- Change rates from the ‘cold’ to ‘hot’ regime tend toward global being more conservative
  - i.e. the global regime tends to stay in the ‘cold’ regime longer





Cold-to-hot switch rate: local > global

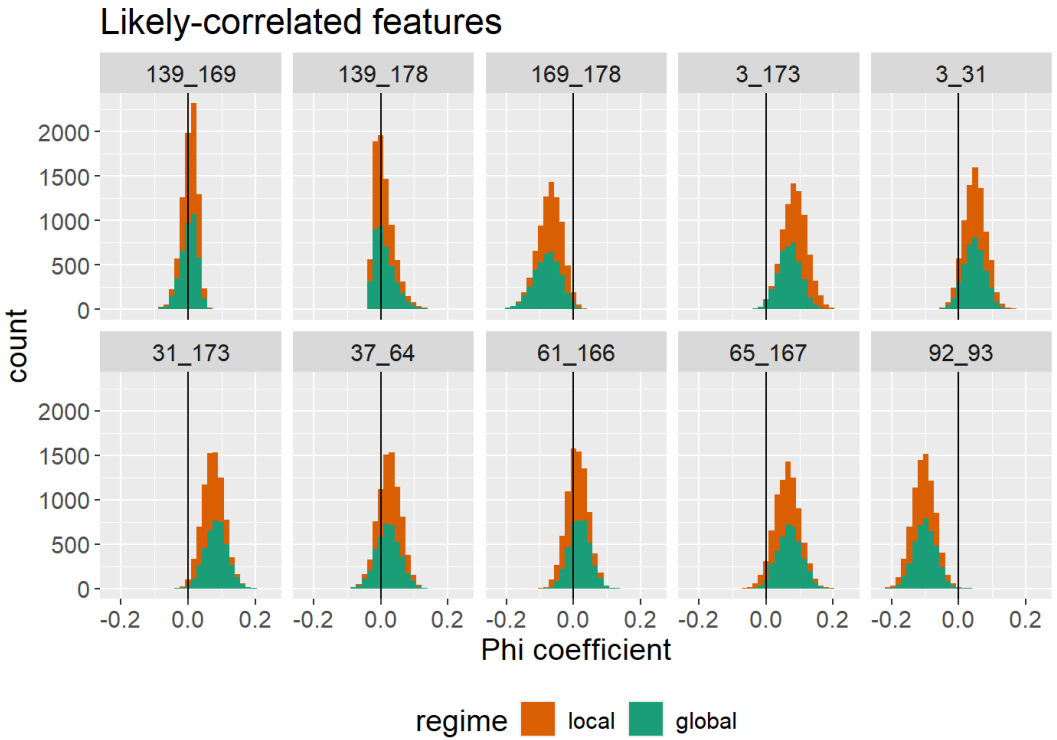
- The global regime tends to have lower switch rates from the ‘cold’ to the ‘hot’ regime
- This indicates an overall inference of slower change on the global phylogeny
- Rates in the global regime are less likely to leave the ‘cold’ regime

Results

CORRELATED EVOLUTION

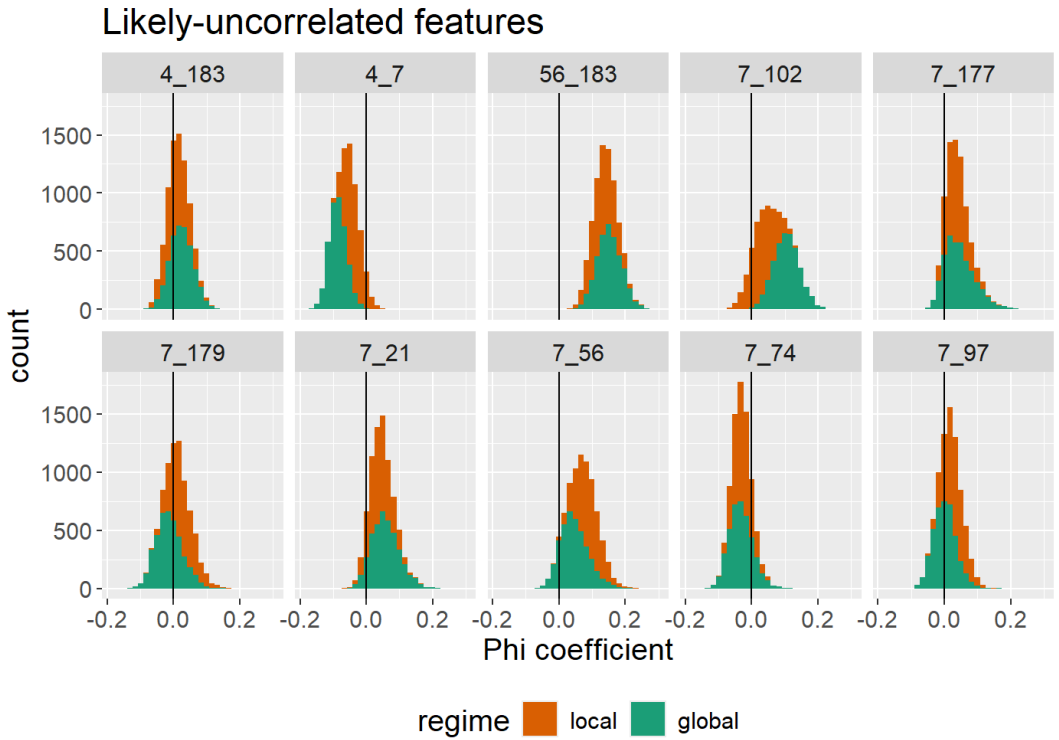
Likely-correlated features

- Correlations overall are not very strong
- Crucially: local and global regimes are not hugely different from one another



Likely-uncorrelated features

- Strange result: the strongest correlation across these 20 pairs is 56 & 183, which should not be correlated at all
- Again, not much difference between local and global regimes
- Will need additional support from Bayes Factors to determine if this correlation is meaningful



Summary

- Overall, rate inference may not be so different across global and local approaches to strongly prefer one over the other
- However, there may be a tendency for features to remain in the ‘cold’ regime in a covarion model for longer when using the global approach
  - Likely caused by long branches near the root
- Estimates of correlated evolution do not seem hugely different across regimes
- The local approach of Jäger & Wahle (2021) is more computationally efficient (and can be parallelized) so may be preferable
- Still to consider: Bayes Factors

References

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022). *Global language diversification is linked to socio-ecology and threat status*. https://doi.org/10.31235/osf.io/f8tr6

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., Tatarinova, T., Progovac, L., & Benítez-Burraco, A. (2024). Linguistic correlates of societal variation: A quantitative analysis. *PLOS ONE*, *19*(4), e0300838. https://doi.org/10.1371/journal.pone.0300838

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, 1229-1242.

Her, O.-S., Liang, Y.-P., Chan, E., Hsu, H.-H., Hsu, A. C.-P., & Allassonnière-Tang, M. (2024). Early humans out of Africa had only base-initial numerals. *Humanities and Social Sciences Communications*, *11*(1), 254. https://doi.org/10.1057/s41599-023-02506-z

Jäger, G., & Wahle, J. (2021). Phylogenetic Typology. *Frontiers in Psychology*, *12*, 682132. https://doi.org/10.3389/fpsyg.2021.682132

Jódar-Sánchez, J. A., & Allassonnière-Tang, M. (2024). The evolutionary dynamics of grammatical gender in Torricelli languages. *STUF - Language Typology and Universals*, *77*(3), 353–369. https://doi.org/10.1515/stuf-2024-2010

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *255*(1342), 37-45.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye 葉婧婷, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray (2023) Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16), eadg6175.