Bayesian Phylogenetic Linguistics

Annika Tjuka, Nataliia Hübler, Olena Shcherbakova

MPI-SHH
SUMMER SCHOOL
2021

Doorway
to Human History

Table of contents

Introduction

Goal, Lexical and Structural Data, Quantitative Approaches, Bayesian Phylogenetic Methods

Activity 1

Poll: Have you ever used Bayesian phylogenetic methods in your research?

A. Yes, for linguistic data.
B. Yes, for other types of data.
C. No.

Goal

Language grouping and dating:

- Categorizing languages into subgroups
- Phylogeny: tracing the history of a language family
- Chronology: defining the date of origin of a language family

Lexical Data

NUMBER	WORD	GROUP
1	Father	2
2	Water	2
3	Night	2
4	Hand	2
5	Woman	2
6	fire	4
7	walk	4
8	mother	4
9	plants	4
10	small	4
11	eat	1
12	water	1
13	hand	1
14	go	1
15	sleep	1

Result of basic vocabulary task (activity 2 from introductory lecture).

Lexical Data

Selected items from the Swadesh 100-list (Swadesh 1955):

Lexical Data

Lexical data (more correctly *cognacy data*) as input for Bayesian phylogenetic analysis:

- Before the lexical data can be used, the lexemes need to be assigned to a cognate set, i.e., lexemes in all languages of a given family.
- Cognate judgments based on the comparative method already established that the languages belong to the same family.
- The input for an analysis is not a list of lexemes, but values of 0 and 1 that represent cognacy relationships.

Structural Data

- phonological and morphological features, for example:
  - two or more contrastive tones
  - distinction of gender in the third-person pronouns

- absence or presence of a feature is coded as 0 or 1, respectively

- grammatical features are often linked so that they can change simultaneously

- some features are more stable than others, for example, inclusive vs. exclusive distinctions and gender distinctions are highly stable (Greenhill et al. 2017)

Structural Data

Grambank (The Grambank Consortium 2021)

- about 200 structural features of 2,000 languages based on reference grammars

- A list of features can be found on the Grambank Wiki page: https://github.com/grambank/grambank/wiki

The age of "*big-ish*" data

- more freely available linguistic data
- not processable by a single person
- need for computational tools to analyze the data

https://glottobank.org/

Quantitative Approaches

- Lexicostatistics for subgrouping
  - the degree of relationship between two languages is based on the number of shared cognates

- Glottochronology for dating
  - the time that lies between the separation of two languages can be measured by the degree of their similarity and a constant “clock” rate

- Both approaches were rejected, based on the problem that languages vary in their rate of change.

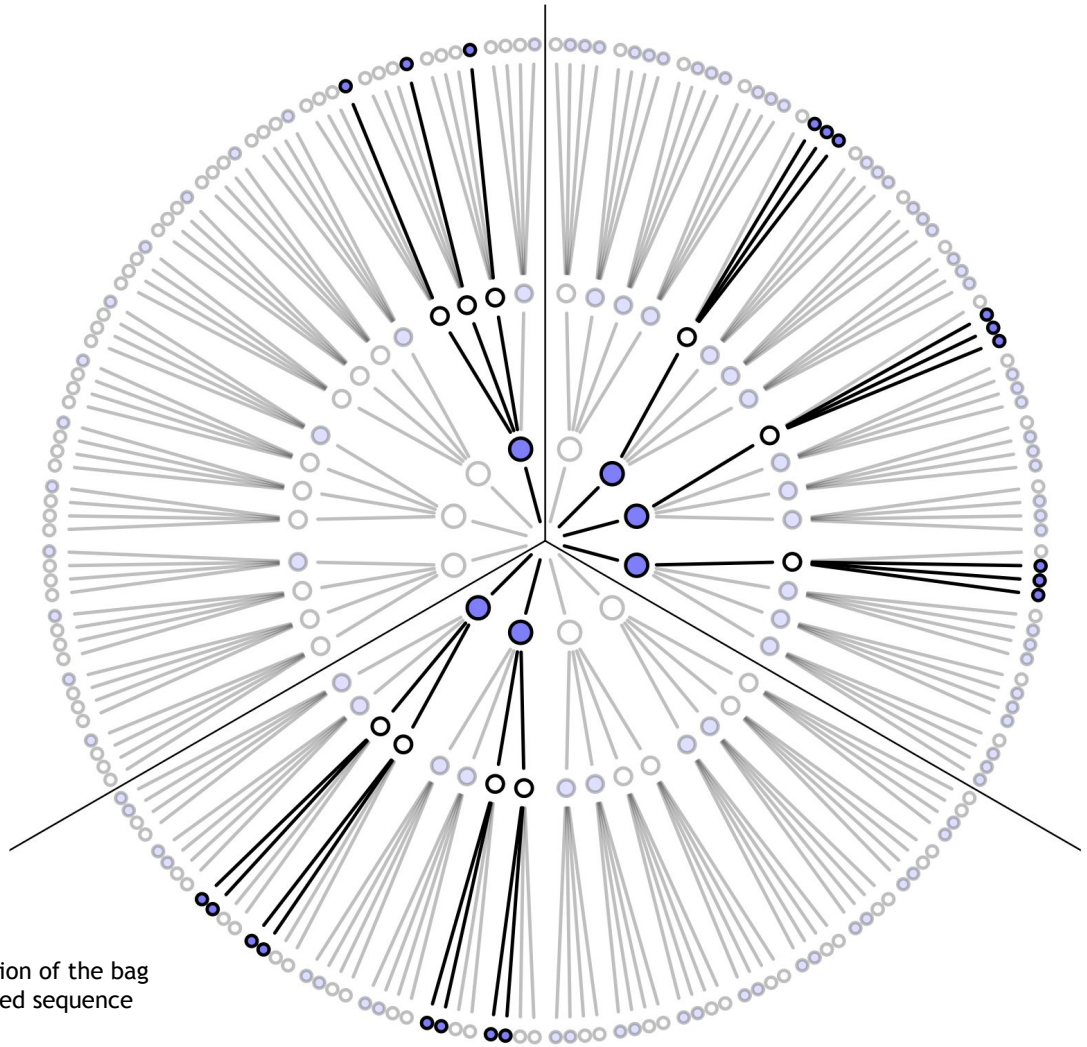
Quantitative Approaches

- Biologists faced similar problems and developed methods that account for variation in rates.

- Bayesian phylogenetic methods from evolutionary biology make up for the shortcomings in the old approaches and have been applied in an abundance of studies over the past decades.

- They are “a useful *supplement* to the comparative method” (Greenhill et al. 2020).

Bayesian Phylogenetic Methods



Each possible composition of the bag that lead to the observed sequence (McElreath 2020).

Bayesian Phylogenetic Methods

- Testing which set of trees best fit the data based on the model
- Finding the most probable tree**S** (not a single one!)
- Making inferences by using Bayes' Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

  - A, B = events, i.e., *tree*, *data*
  - P(A|B) = probability of A given B is true
  - P(B|A) = probability of B given A is true
  - P(A), P(B) = the independent probabilities of A and B

Bayesian Phylogenetic Methods

LEXICAL DATA →

Fijian	taba-na
Marquesan	peheu
Hawaiian	ēheu
Tahitian	pererau
Maori	parirau

COGNATE CODED DATA →

Fijian	A	100
Marquesan	B	010
Hawaiian	B	010
Tahitian	C	001
Maori	C	001

PRIOR INFORMATION →

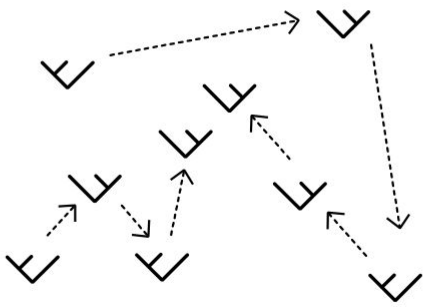
- Tree prior
- Branch length
- Calibrations
- Constraints
- Model parameters

MODEL SELECTION →

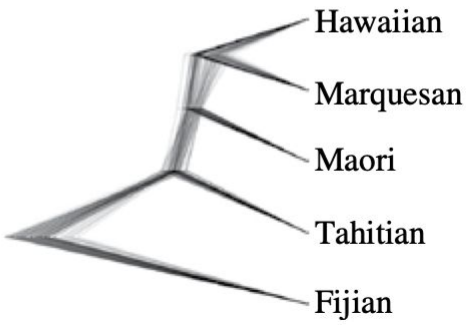
- Simple Binary
- Covarion
- Stochastic Dollo
- Multistate/Binary
- Strict/Relaxed Clock

Markov chain Monte Carlo

MCMC SEARCH →



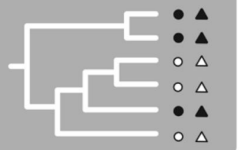
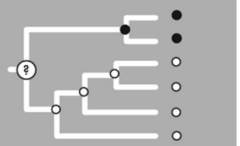
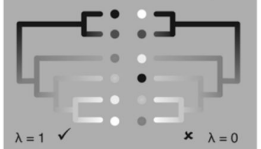
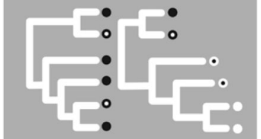
POSTERIOR DISTRIBUTION OF TREES →



Steps in a Bayesian phylogenetic analysis (Greenhill et al. 2020).

Bayesian Phylogenetic Methods

Overview of five types of evolutionary questions that can be answered using phylogenetic comparative methods (Jordan 2013).

RESEARCH QUESTION	COMPONENTS	EXAMPLE
Correlated evolution *Are two traits changing together?* 	**Data** Two discrete presence/absence traits or Two continuously varying traits **Tree** Any fully-bifurcating phylogeny(s) **Outcomes** Can build up pathways of correlated changes; combined with ancestral states can infer direction of change	**Cattle lead to loss of matriliny in Bantu-speaking societies** Lexical tree of 68 Bantu languages Data: descent and pastoralism Dependent model of coevolution more likely than one where traits evolved independently Pastoralism changed before matriliny Holden & Mace 2003
Ancestral states *What was the earlier form of a trait?* 	**Data** A trait with 2+ categorical states or A continuously varying trait **Tree** Any fully-bifurcating phylogeny(s) **Outcomes** Can test models of sequential change; Can "fossilise" ancestral nodes if known; Can test competing hypoth-eses about ancestral states	**Matrilocal residence is ancestral in Austronesian** Lexical trees of 135 AN languages Data: postmarital residence Matrilocality inferred for PAN and PMP Switches to matrilocality less likely than to other forms of residence Jordan et al 2009
Phylogenetic signal *Does a trait track history?*  λ = 1 ✓ × λ = 0	**Data** Any continuously varying trait **Tree** Any fully-bifurcating phylogeny(s) **Outcomes** Can estimate degree of phylogenetic signal; Can test if signal is significant and therefore must be controlled for	**Population size and the rate of lexical evolution** Lexical tree of 351 AN languages Data: population size, amount of lexical change Population size and density have lambda (λ) values close to one, indicating strong historical signal Jordan & Currie submitted
Mode of change *Is change gradual or punctual?* 	**Data** Measures of branch (path) lengths or Any continuous varying trait **Tree** Any fully-bifurcating phylogeny(s) with meaningful branch lengths **Outcomes** Can use kappa statistic to quantify degree of punctual/gradual evolution in any one character	**Languages evolve in punctuational bursts** Lexical trees of AN, Bantu & IE Data: path lengths, number of nodes Relationship between path length and nodes suggests splitting events cause more lexical evolution Atkinson et al 2008
Rate of change *How fast do traits change?* 	**Data** Any discrete presence/absence traits **Tree** Any fully-bifurcating phylogeny(s) with meaningful branch lengths **Outcomes** Can determine rate of change of a trait; Combined with known time-depth of phylogeny can infer dates	**Similar rates of evolution for lexical & typological features** Lexical trees of AN & IE languages Data: typological features Estimate of evolutionary rates was equivalent across both language families and both types of features Greenhill et al 2010

Data, Trees and Networks

Bayesian tree-building

Networks

Admixture

Overview

Two case studies with two types of data:

- Dravidian languages with lexical data
- Sahul languages with typological data

- Bayesian tree-building
  - Data preparation
  - Priors and models
  - Running the analysis
  - Tree topology(ies)
- NeighbourNet
- Admixture

What is our goal?

- we have an idea of the relationships between the languages and want to date the branches and the root -> Bayesian phylogenetic methods
- we want to get an initial picture of the relationships between languages and see whether the data is tree-like -> Neighbour-Net
- we want to see the proportions of borrowings per language and let the model decide on the number of groups in the data -> STRUCTURE

Data: Why basic vocabulary?

- not all languages have a rich system of morphological markers (e.g. Sino-Tibetan)
- often basic vocabulary is all we have (for poorly documented or extinct languages)
- tokens of regular sound changes are found mostly in cognates
- cognates are concentrated basic vocabulary, which is less prone to replacement
- “lexical data” ≠ “cognacy data”

Data collection: basic vocabulary

- Collect items for a 100, 200 Swadesh list or a Leipzig-Jakarta list
- code the items for cognacy
  - follow the principles of the comparative method
  - this coding can turn your 100 item list into a 1000 item list
  - the most time-consuming step

Data collection: structural features

Akin-mi:	*min-du:*	*tuŋa-βa*	*ɲami:-βa*	*ani:-ra-n.*
father–POSS.1SG	1SG-DAT	five–ACC	female.deer–ACC	give–AOR–3SG

'My father gave me five female deer.'

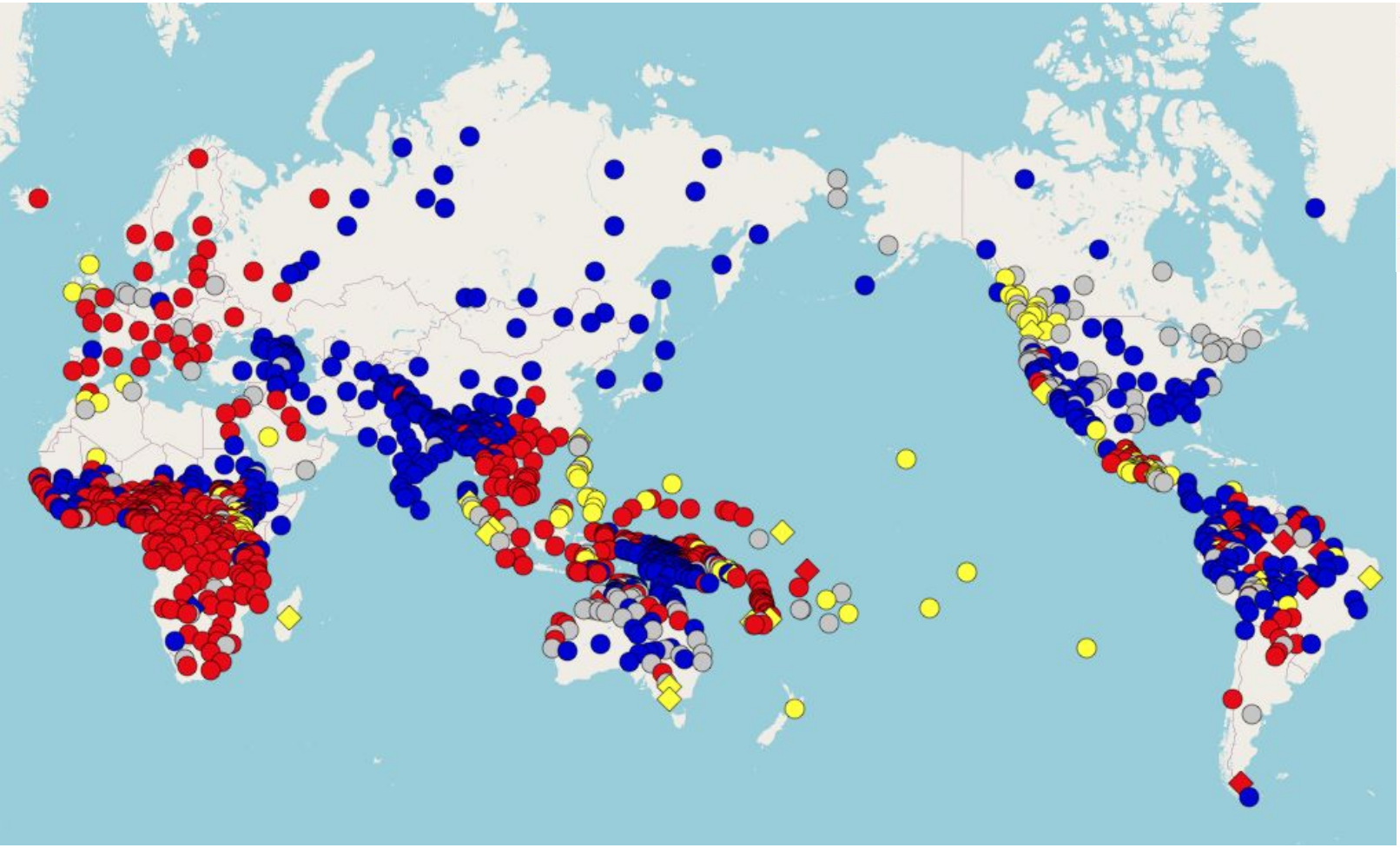
- Is the verb at the end of the sentence? (word order: father deer give) - Yes, 1
- Do we see on the noun, whom it belongs to? (father-my)? - Yes, 1
- Does the numeral come before the noun? (five deer) - Yes, 1
- Do we see on the verb, who performs the action? (3SG - i.e. the father) - Yes, 1
- Are there cases for nouns? And for pronouns? (at least dative and accusative) - Yes and yes, 1 and 1

Activity 2

What is the position of the verb in one of the languages you speak?

1. At the beginning of the sentence (Eat an apple I).
2. In the middle of the sentence (I eat an apple.)
3. At the end of the sentence (I an apple eat.)

Word order



https://wals.info

Data: two types of matrices

	“head 1”	“head 2”	“head 3”
language 1	1	1	0
language 2	0	1	1
language 3	0	0	0
language 4	1	0	1

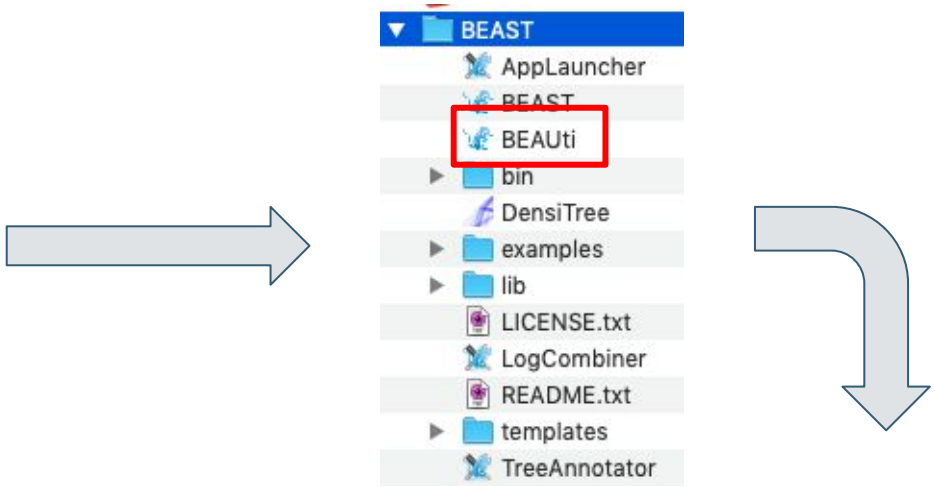
binary

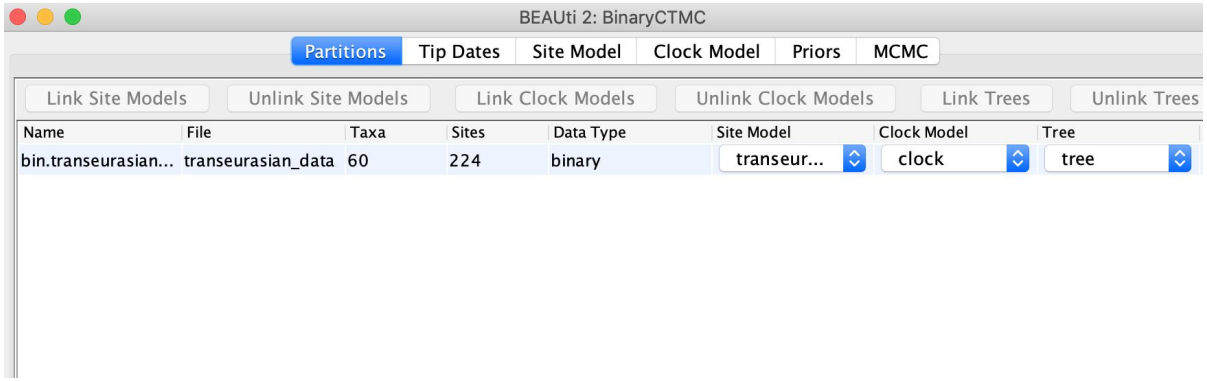
	"head"	"hand"	"one"
language 1	1	1	1
language 2	1	2	2
language 3	2	2	2
language 4	3	3	1

multistate

Data formatting and import

```
#NEXUS
BEGIN DATA;
    DIMENSIONS NTAX=60 NCHAR=224;
    FORMAT MISSING=- SYMBOLS="01";
MATRIX
Turkish
001010000100000001111100000000011110
010000000000100000100010000100100010
111101100101010101011101001100
Bashkir
0000-1000000000001111100000000011110
-100000000001-000010101010010010010
10101110010101-10101110----100
BeryozovkaEven
00001-100000010001----000000100111
--0000000-00100000------10010010010
101010110011000101011 10--00100
Bonan
00000-1000010000111111000000000111
0101100000-0--0000-0110000000010000
111111101110101110--100100101
```





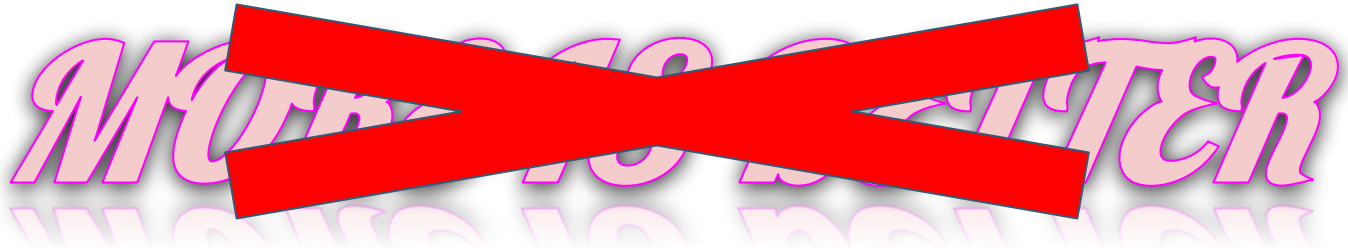
Software and tutorials:
http://www.beast2.org
https://beast.community
https://taming-the-beast.org

Prior information

- calibrate some nodes: historical events, ancient texts, archaeological data
- set monophyletic constraints: we know that some languages belong together based on phonological innovations
- risk of overparameterization: run the analysis with and without data to see the impact of the priors

Prior information

- calibrate some nodes: historical events, ancient texts, archaeological data
- set monophyletic constraints: we know that some languages belong together based on phonological innovations
- risk of overparameterization: run the analysis with and without data to see the impact of the priors



Models of character change

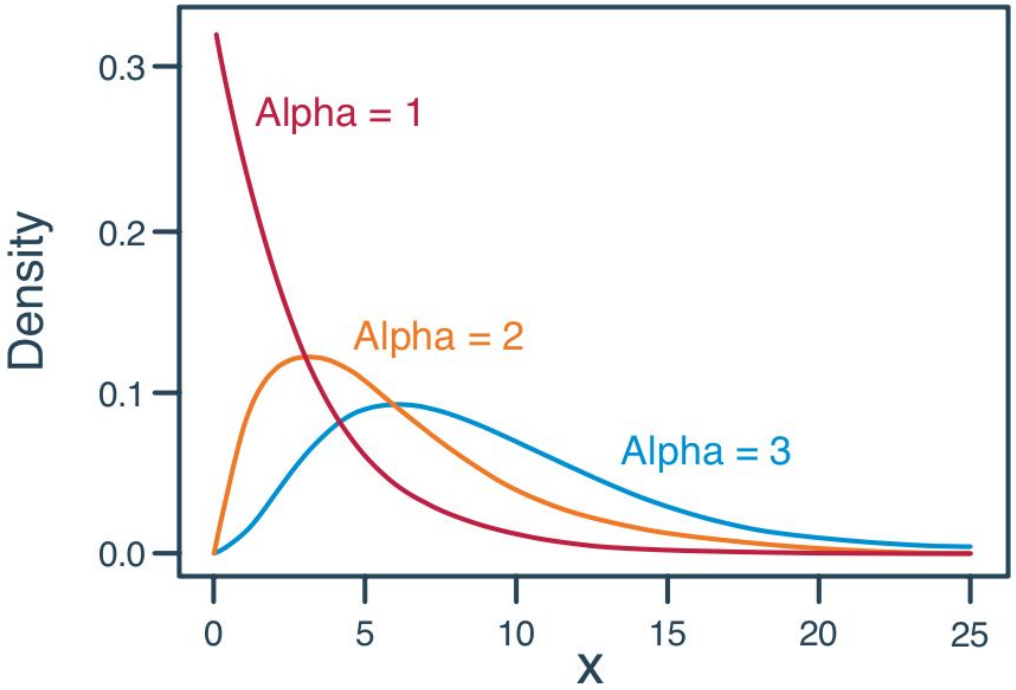
- Continuous Time Markov Chain model: one rate of cognate gain and loss
- covarion model: sites can switch between fast and slow rates on different parts of the tree
  - different items on the Swadesh list can have different replacement rates
  - e.g. in Indo-European, “name” evolves slower than “dirty”
- stochastic Dollo model: cognates can only be gained once, but get lost multiple times
- … new models are developed and adapted to language data all the time

Allowing rates to vary

- gamma
- covarion
- partitioning the data into subsets and allowing each subset to evolve at its own rate
  - e.g. partition all the cognates from each word, so each word has its own rate
  - only works if you have a lot of data as it's adding one (or more) parameter per partition

Gamma rate heterogeneity

- a common way to handle rate variation (especially in the CTMC model)
- each cognate gets a rate sampled from a gamma distribution



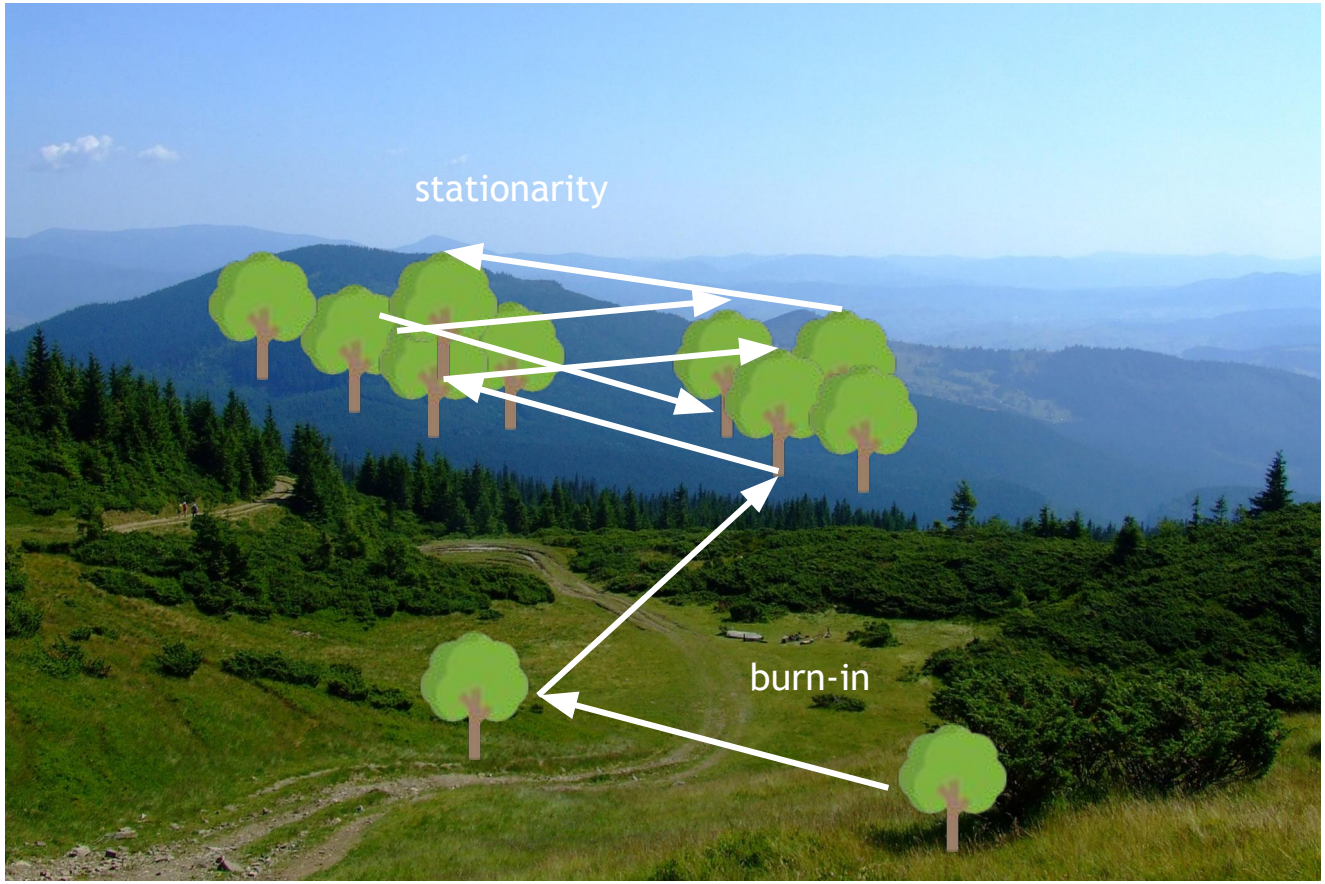
see more on gamma distributions here

Strict vs. relaxed clock

- strict clock ≈ glottochronology
- each branch is allowed to have its own rate
- data can inform the analysis of the rate variation

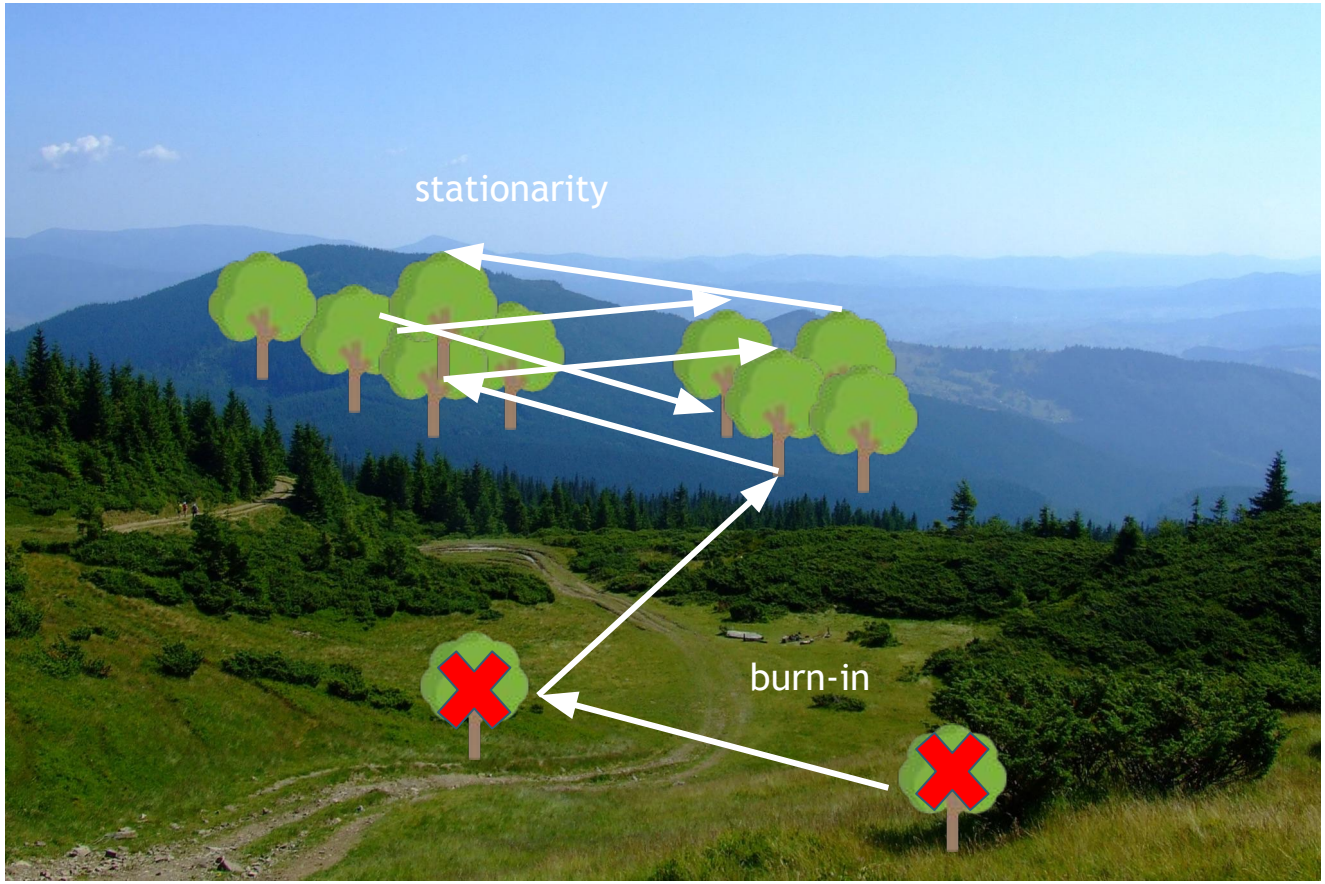
Image from Simon Greenhill's (2017) lecture on Phylogenetics and Tree-Thinking

Trees and MCMC search



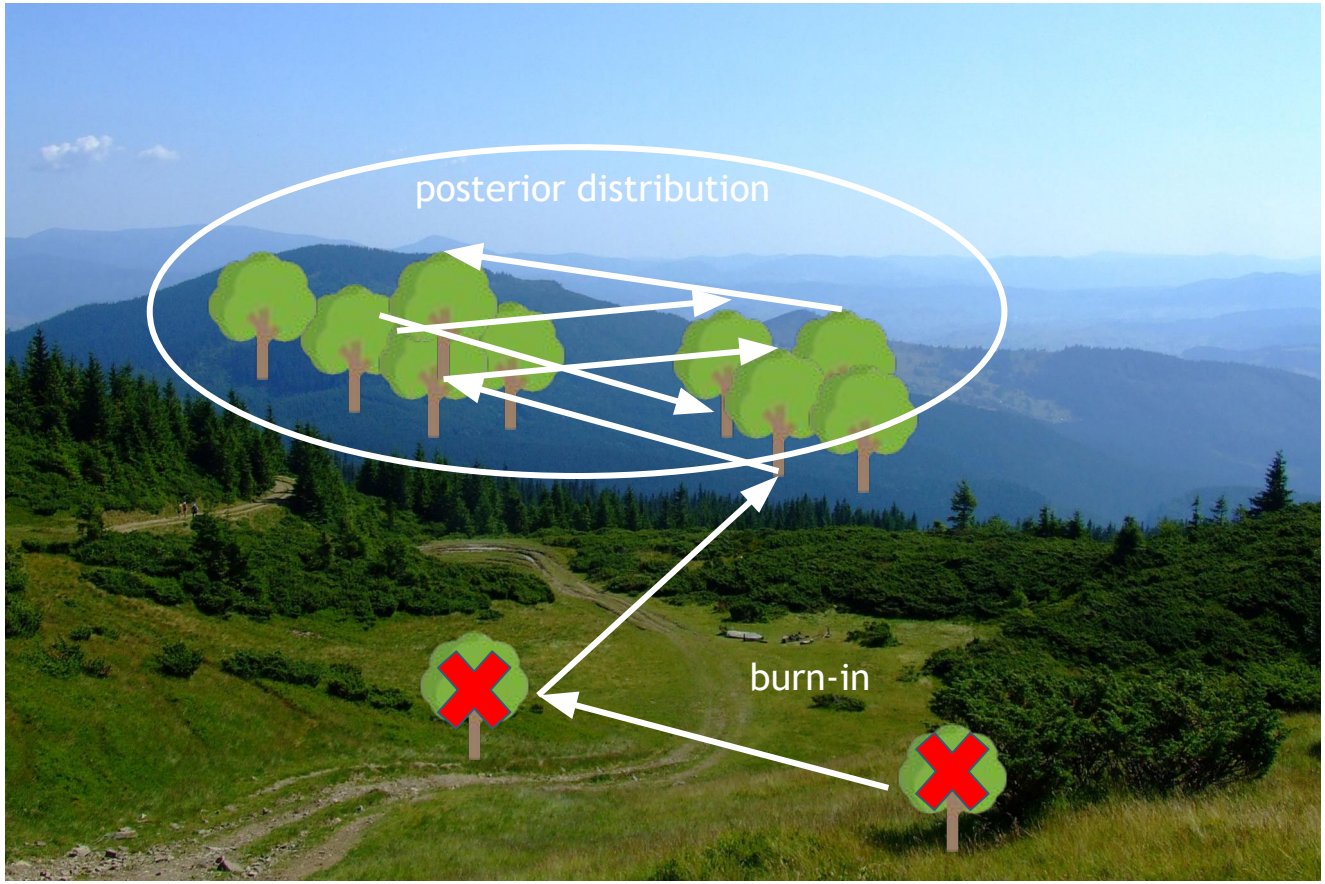
Background image: Kleiner Funken - own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=48963487
Tree icon: www.iconpacks.net

Trees and MCMC search



Background image: Kleiner Funken - own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=48963487
Tree icon: www.iconpacks.net

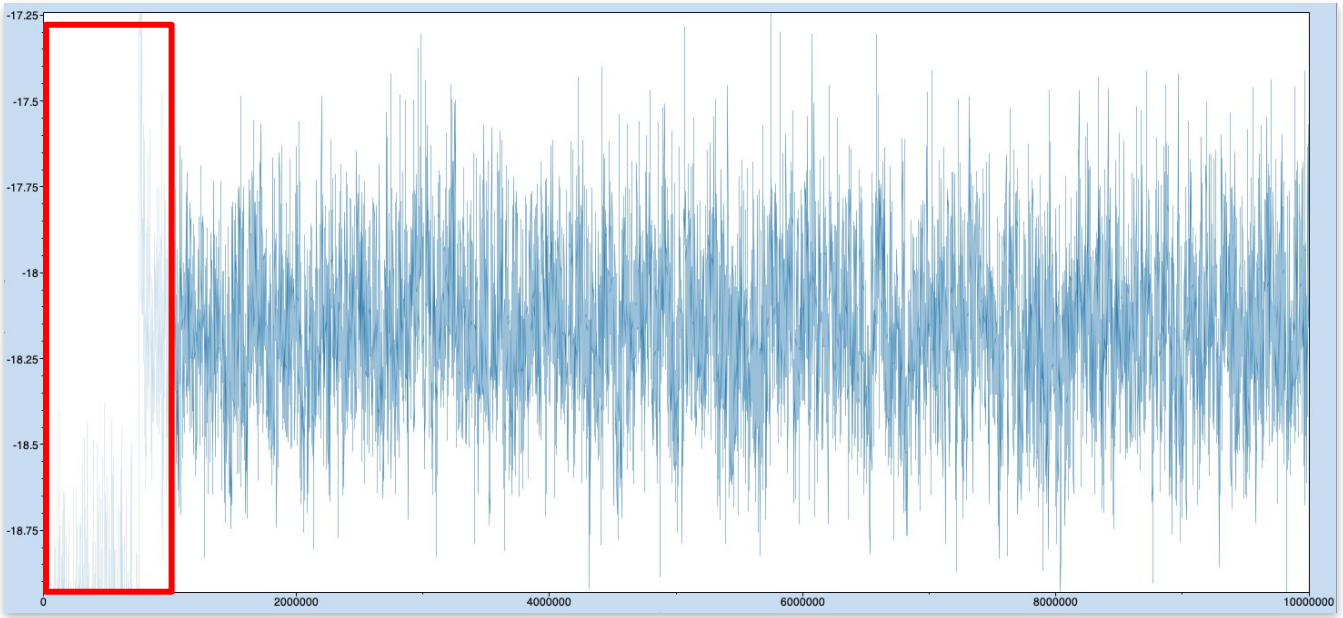
Trees and MCMC search



Background image: Kleiner Funken - own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=48963487
Tree icon: www.iconpacks.net

Running the analysis

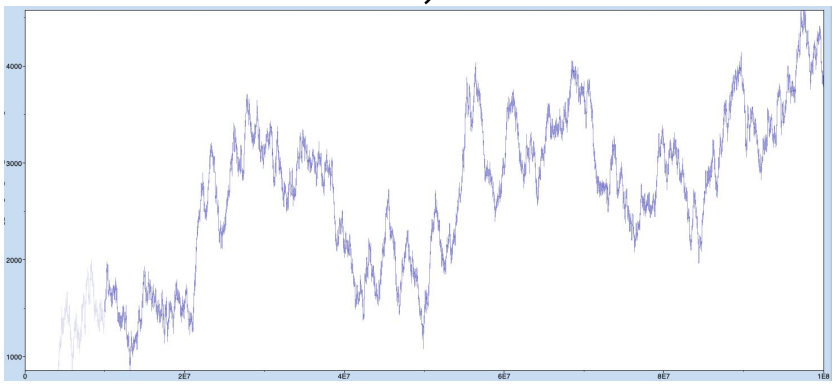
- run the analysis long enough to achieve convergence (e.g. chain length: 10 000)
- remember the burn-in, where the estimations are unstable and have to be removed before the analysis



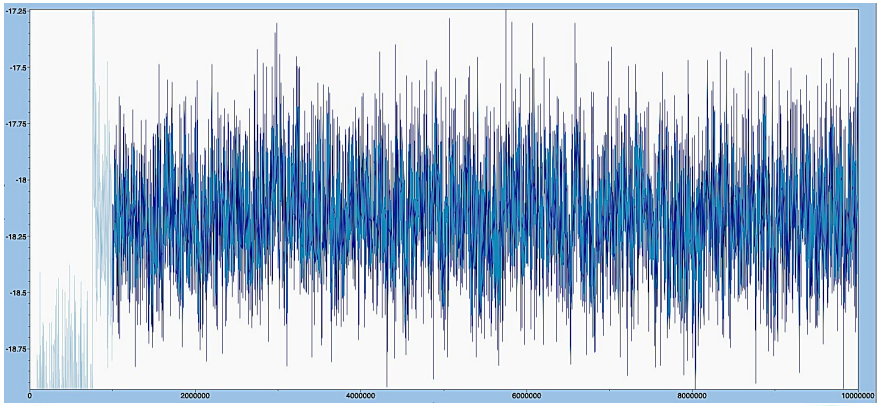
find out more about Tracer here

Does the chain do its job?

- this is a bad, misbehaved chain



- this is a good, well-behaved chain, means good convergence



Dravidian language family

some of the authors:

80 varieties, spoken in India, Nepal, Pakistan and Afghanistan

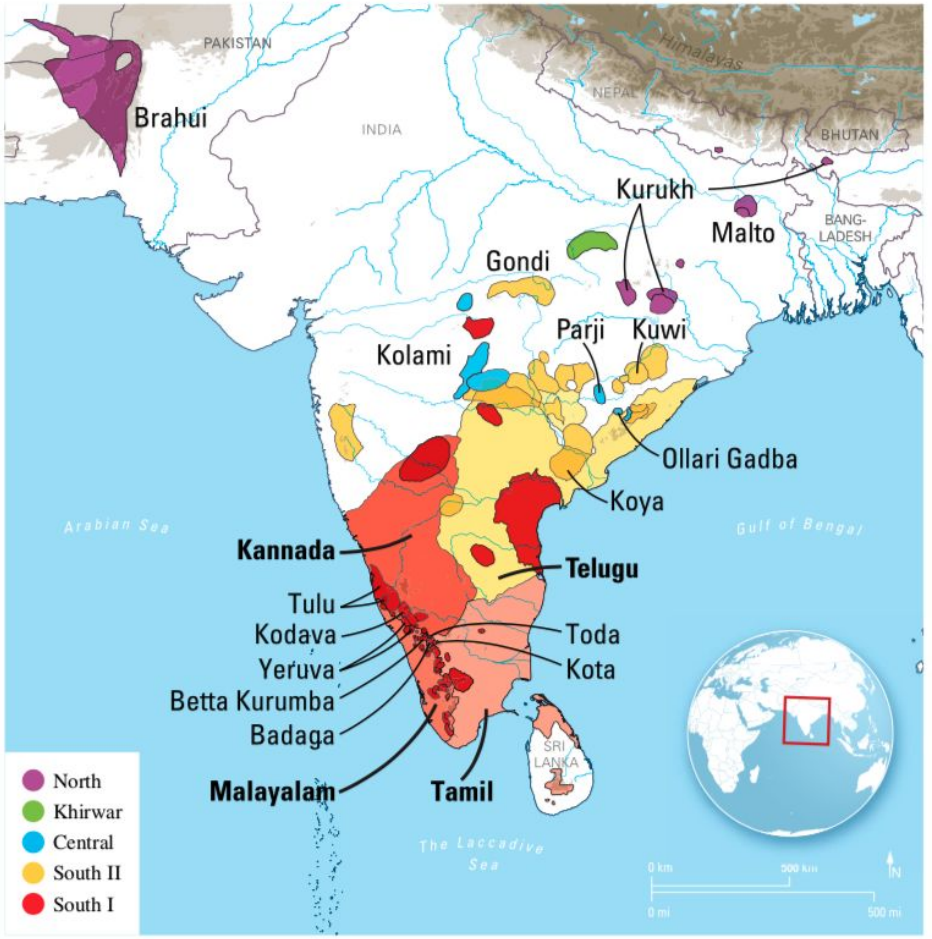
spoken by over 200 million people

Aim: geographical origin, dispersal through time, internal relationships

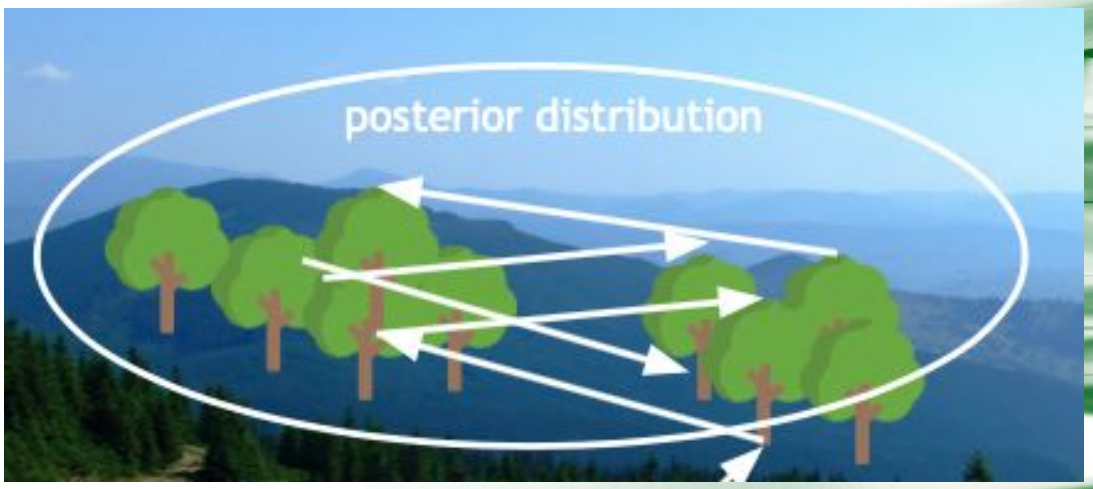
Data: 100 items Swadesh list, 20 languages

Study: Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., & Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. Royal Society open science, 5(3), 171504.

link to the paper



Bayesian forest



Tulu
Tamil
Malayalam
Kannada
Yeruva
Kodava
Badga
Toda
Kota
Betta_Kurumba
Kolami
Gondi
Kuwi
Telugu
Koya
Parji
Ollari_Gadba
Brahui
Kurukh
Malto

South I
(prior)

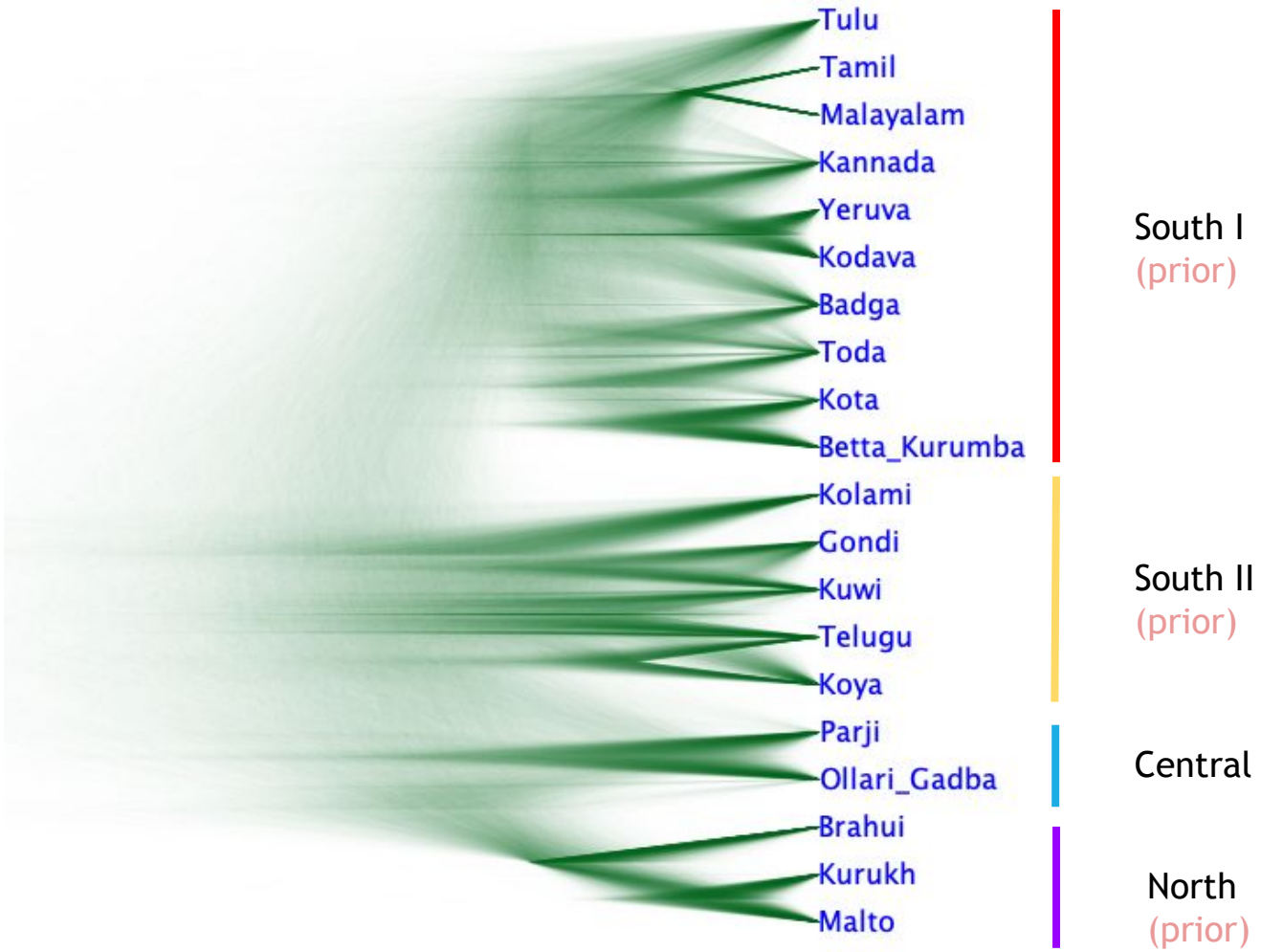
South II
(prior)

Central

North
(prior)

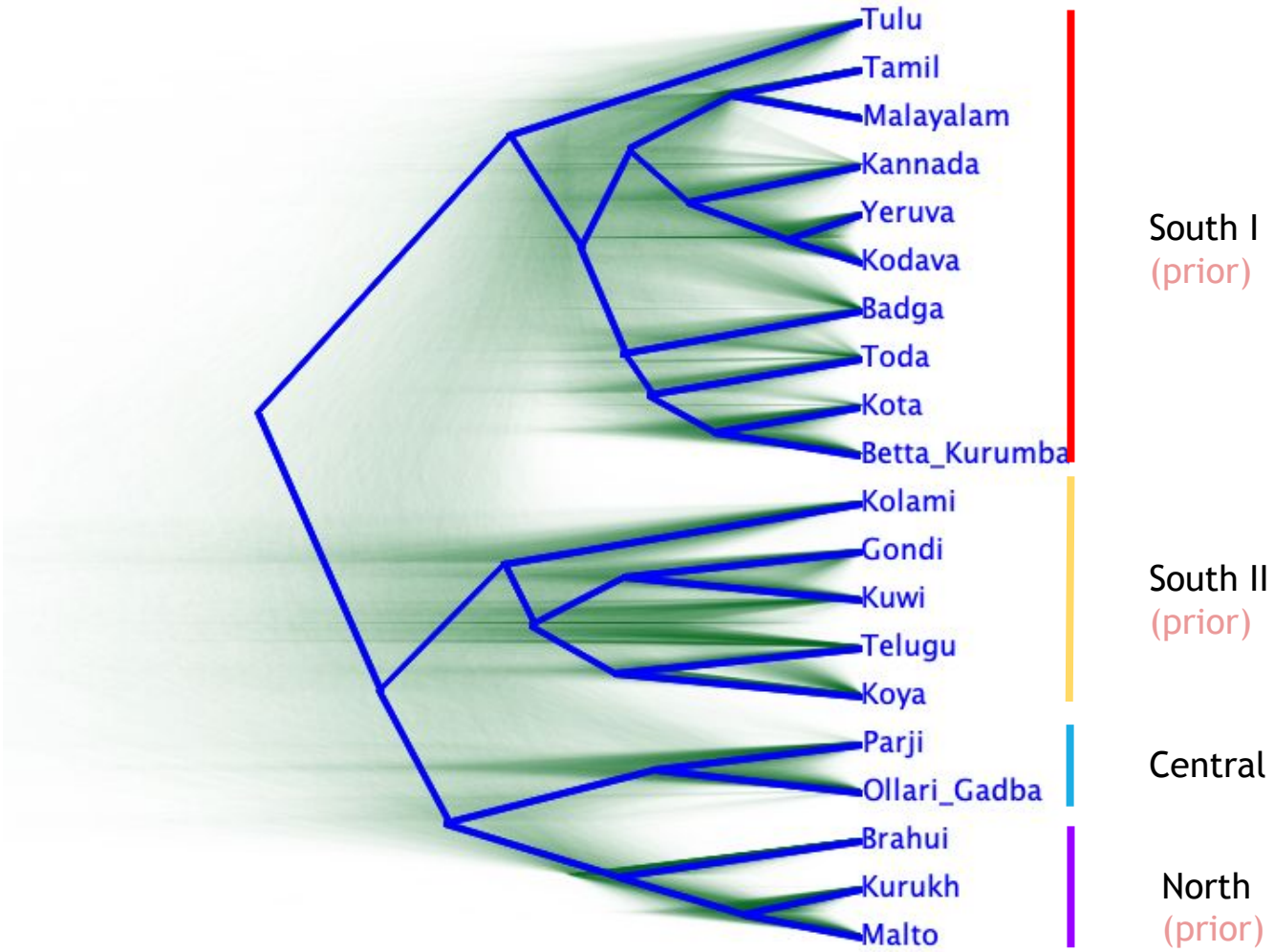
link to the software
DensiTree here

Bayesian forest

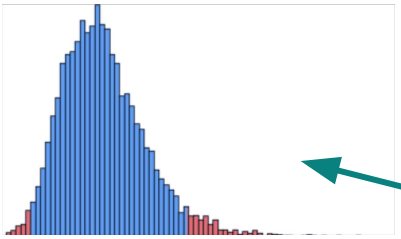


link to the software DensiTree here

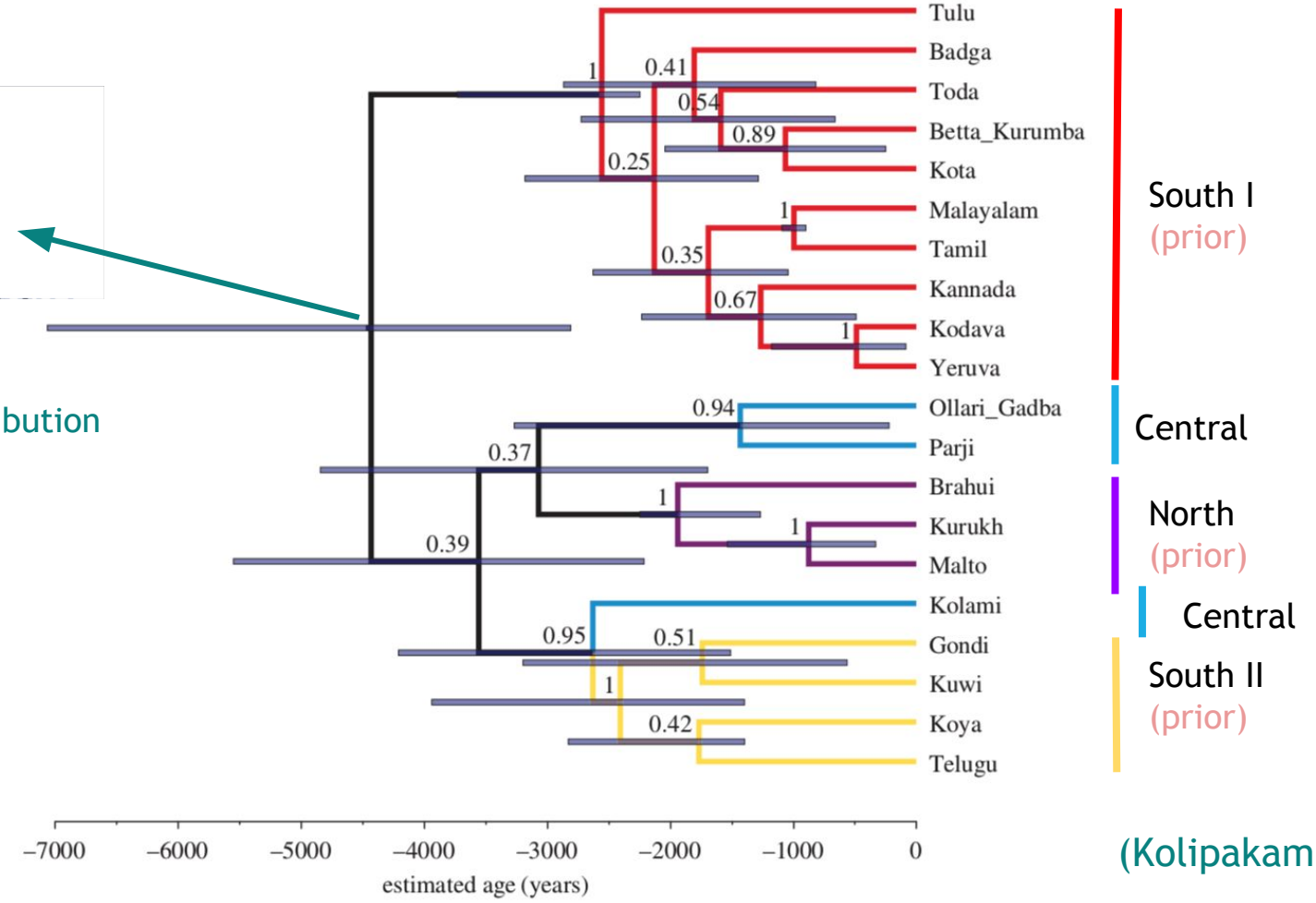
Bayesian forest



Maximum clade credibility tree



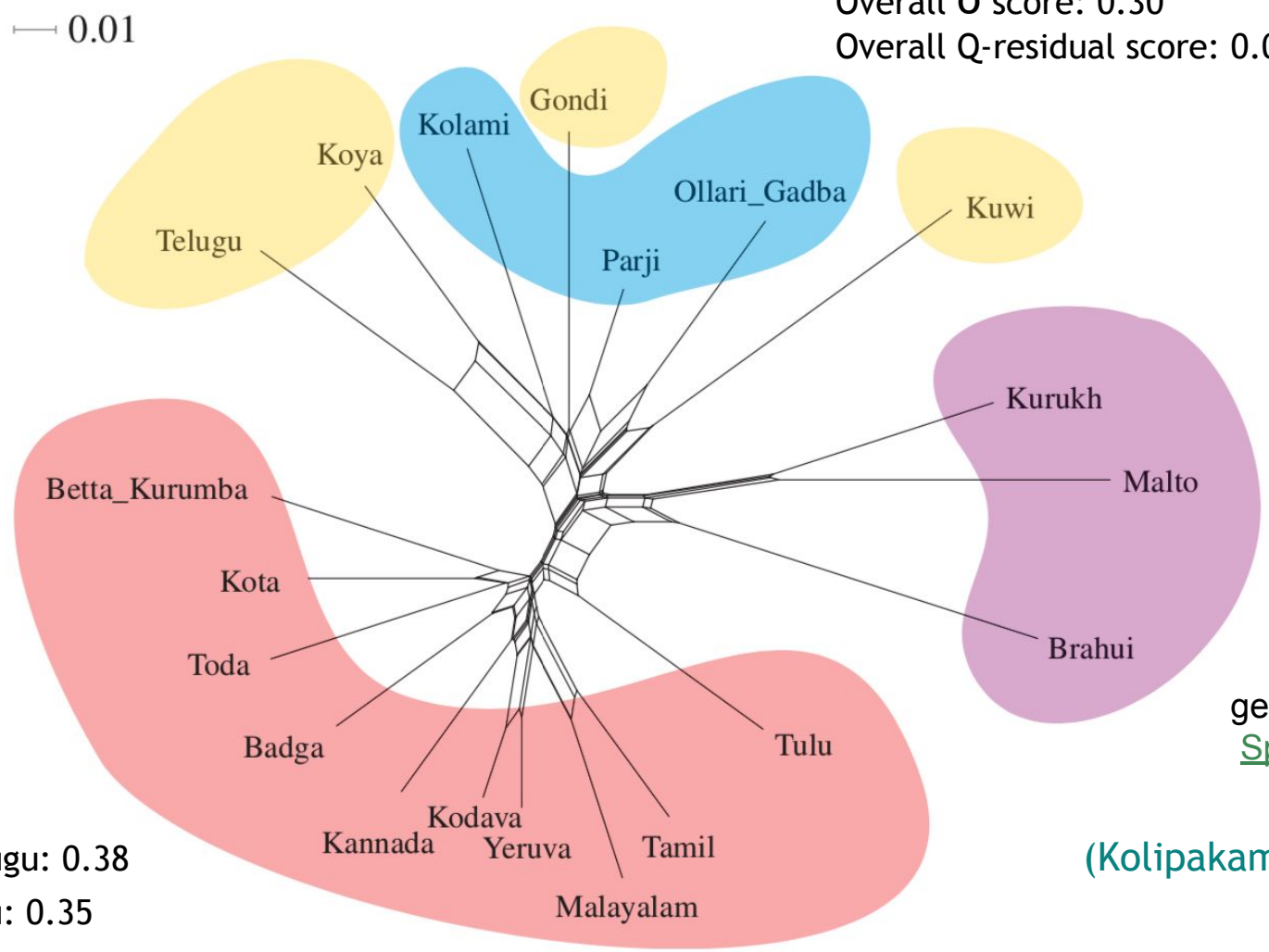
the root age is a probability distribution and not a point estimate



(Kolipakam et al. 2018)

link to the software FigTree here

Neighbour-Net



Overall δ score: 0.30
Overall Q-residual score: 0.0069

get the software
SplitsTree

δ score for Telugu: 0.38
δ score for Tulu: 0.35

(Kolipakam et al. 2018)

Measures of tree-likeness

- Delta-score and Q-residual score: each taxon is scored from 0 to 1 according to how much it is involved in conflicting signal
- E.g. δ=0: distances between taxa exactly fit the tree

	Delta-score	Q-residual score
Indo-European	0.22	0.002
Dravidian	0.30	0.0069
Austronesian	0.33	0.002
Polynesian	0.41	0.02

Kolipakam et al. (2018)
Gray et al. (2010)

What influences delta and Q-residual scores?

- language contact due to various socio-linguistic factors, such as multilingualism (e.g. Dravidian languages)
- dialect chain break-up (e.g. Japonic languages)
- rapid language spread and diversification (e.g. Pacific)
- language family age: the older, the more tree like: "Over time networks get pruned by language extinction to appear more tree-like" (e.g. Indo-European)
- noise in the data

Kolipakam et al. (2018)
Gray et al. (2010)

Admixture analysis with STRUCTURE

- a Bayesian algorithm originally developed to discover populations on the basis of recombining genetic markers
- individuals/languages are probabilistically assigned to one or several populations
- visualises sources of admixture: we can attribute different layers of ancestry to different populations/language families and determine the alleles/features that contribute to each layer
- we get the most likely number of populations/families and the most likely contribution of each population/family to each language

(Reesink et al. 2009)

Languages of Sahul

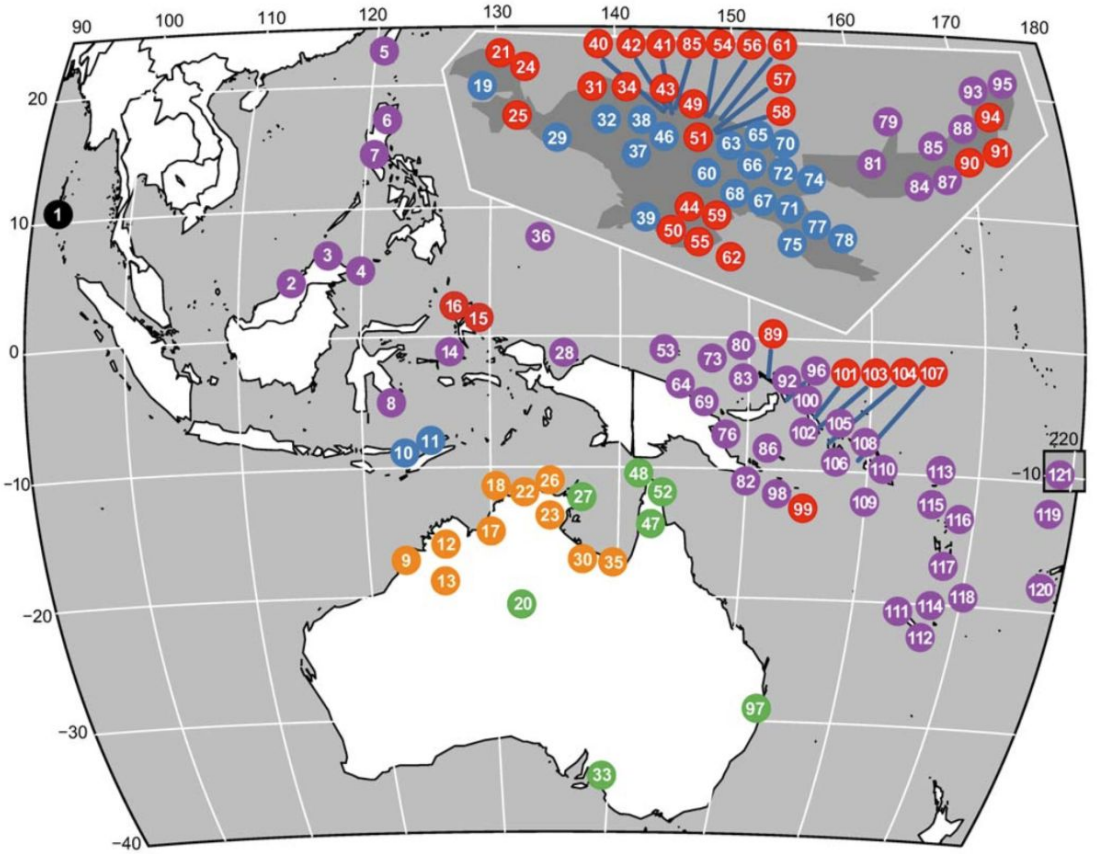
some of the authors:

Aim: ancient connections between the languages in Australia, New Guinea and surrounding islands, not spotted by the comparative method

Data: 160 features, 121 languages (Papuan, Trans New Guinea, Pama-Nyungan, Austronesian, Andamese and other)

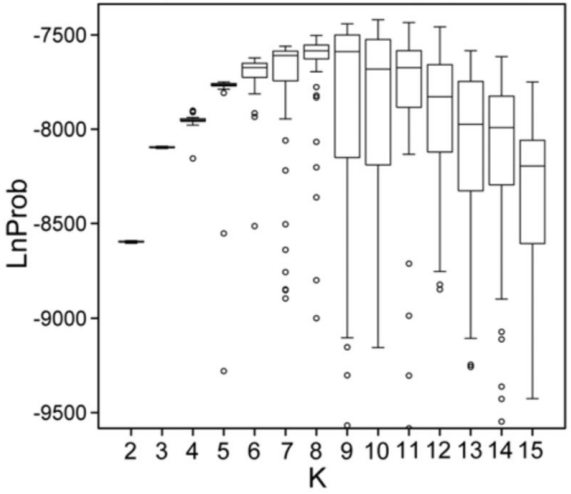
Study: Reesink, G., Singer, R., & Dunn, M. (2009). Explaining the linguistic diversity of Sahul using population models. PLoS biology, 7(11), e1000241.

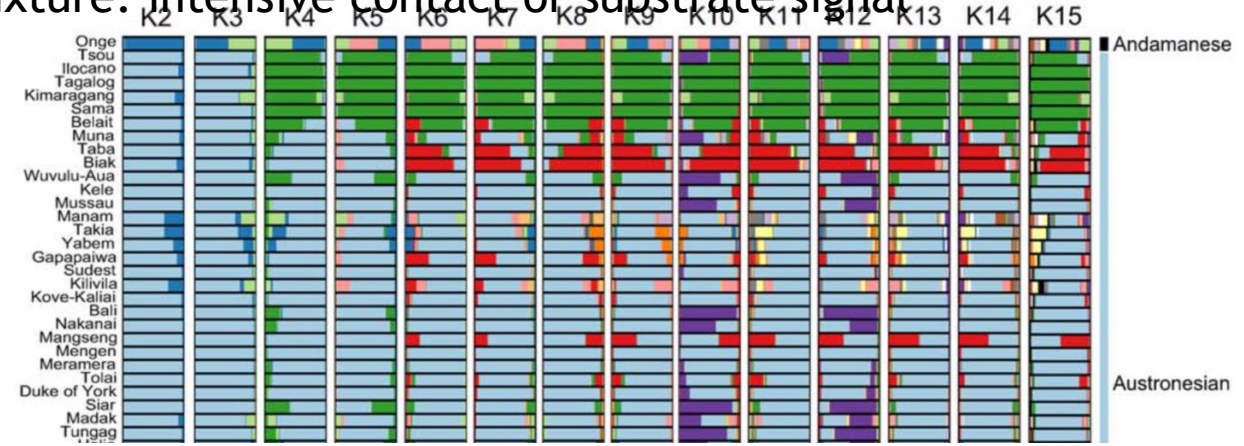
link to the paper



Admixture analysis with STRUCTURE

- 10 populations: some align well with known language families, some do not
- traces of early dispersals: ancient connections between Australian languages and some Papuan groups (supports some long-standing hypotheses)
- high levels of admixture: intensive contact or substrate signal

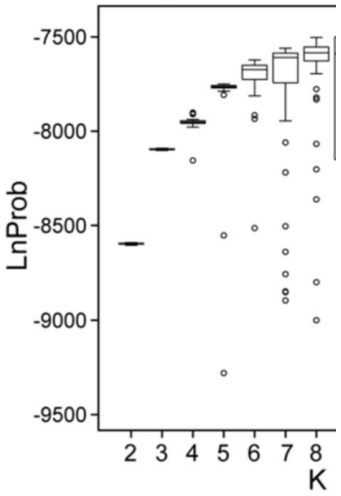


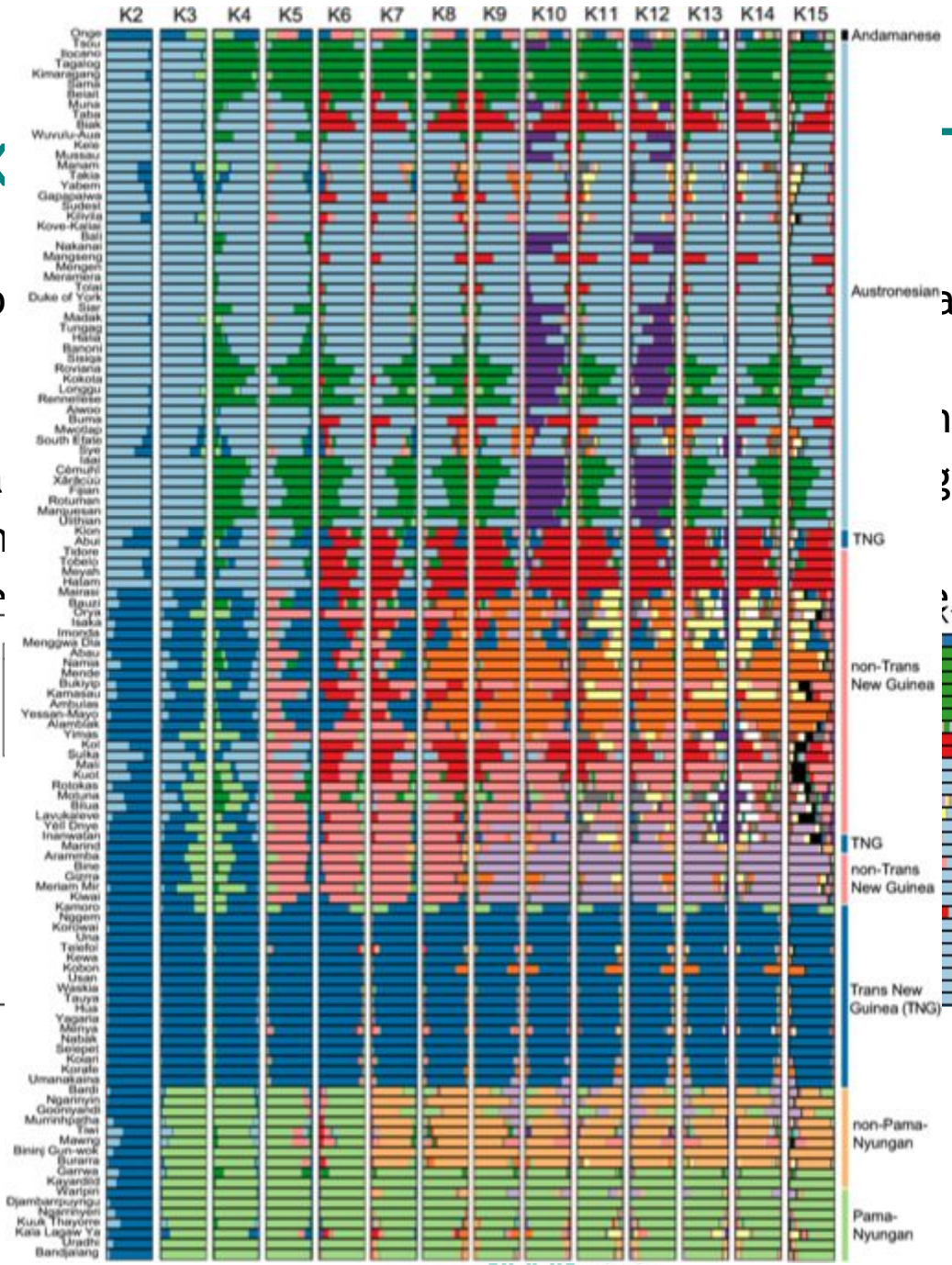


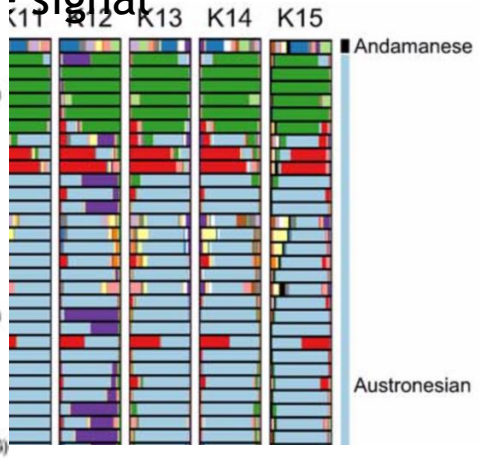
(Reesink et al. 2009)

Admix TURE

- 10 pop amilies, some do not
- traces n Australian langua g-standing hypoth
- high le signal







(Reesink et al. 2021)

Admixture analysis with STRUCTURE

- well applicable to language structures, but also maybe to other kinds of data
- in linguistics, we can
  - verify the existing language families or detect language contact events
  - test hypotheses about migrations of speaker groups
  - detect substratum in case of population replacement and language shift
- a drawback: results heavily dependent on the sample (we can be misguided by excluding particular languages)

Take-away

- Bayesian tree-building does not replace the comparative method, it complements it
- Bayesian tree-building does replace glottochronology: allows rate variation thanks to the relaxed clock model
- NeighbourNet and STRUCTURE work well with even relatively small data sets, at high or unknown levels of borrowing and do not have requirements on the established relationships between languages

Activity 3

Language data: Think of how some equivalents or comparable concepts in genetics and archaeology (any parts of the genome?)

Do you think the application of methods from genetics to language data is justified?

Put your thoughts in the chat, share them after the lecture during the Q&A session or bring them to the informal discussion on gathertown.

Phylogenetic Comparative Methods (PCM)

What?

Why?

How?

Toolkit and case studies

SUMMER SCHOOL 2021

Doorway to Human History

What?

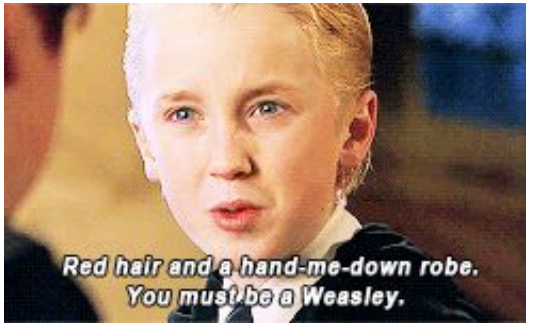
- “Phylogenetic methods encompass a broad family of mathematical approaches that can be used to construct, analyse and incorporate phylogenies” (Evans et al. 2021)
  - a **synchronic** aspect: the investigation of a feature in contemporaneous languages/cultures
  - a **diachronic** aspect: the descent relationships of languages/cultures through time
  - “the synchronic variation cannot be properly understood without the control for evolutionary relatedness that is provided by the ancestor-descendant relationships of the diachronic phylogenetic tree” (Jordan 2013)

Why?

Incorporating phylogenetic information from trees into analysis allows us:

- to control for **common ancestry/shared inheritance/genealogical relatedness**:
  - i.e. ensure that we are not treating languages **related by descent** as if they were independent

1

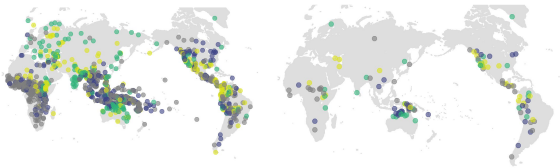


Why?

PCM alternative	Limitations
Using standard correlational methods	• the assumption of independence of observations is violated ⇨ Spanish & Portuguese = Spanish & Chinese

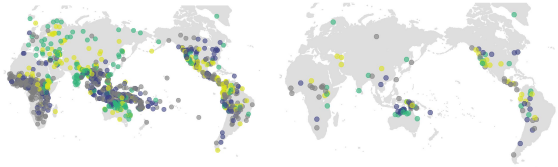
Why?

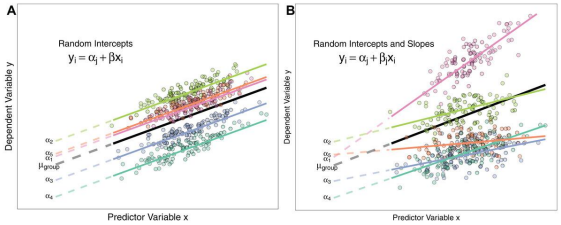
PCM alternative	Limitations
Using standard correlational methods	• the assumption of independence of observations is violated Spanish & Portuguese = Spanish & Chinese
Sampling languages from different families & areas	• the loss of valuable datapoints • the remaining datapoints may still be non-independent (Eff 2004) and inaccurately reflect large-scale patterns



Why?

PCM alternative	Limitations
Using standard correlational methods	• the assumption of independence of observations is violated ⇨ Spanish & Portuguese = Spanish & Chinese
Sampling languages from different families & areas	• the loss of valuable datapoints • the remaining datapoints may still be non-independent (Eff 2004) and inaccurately reflect large-scale patterns
Fitting models with random effects of language families	• the simplification of the relationships between languages within families ⇨ Spanish & Portuguese = Spanish & Russian





graphs from Harrison et al. 2018

Why?

Incorporating phylogenetic information from trees into analysis allows us:

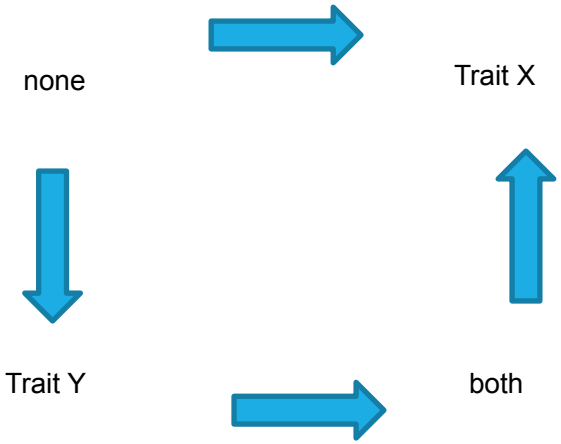
- to statistically assess support for different models and test several competing hypotheses
- to examine correlation in the evolution of traits
- to determine the direction of causal relationships between traits:
  - a variety of diachronic processes could have given rise to the synchronic distribution of traits among languages/cultures
  - PCM help to establish those diachronic processes that were indeed responsible for the synchronic patterns and discard irrelevant ones

Why?

Distribution patterns of two binary traits

	Trait X absent	Trait X present
Trait Y absent	678	243
Trait Y present	301	237

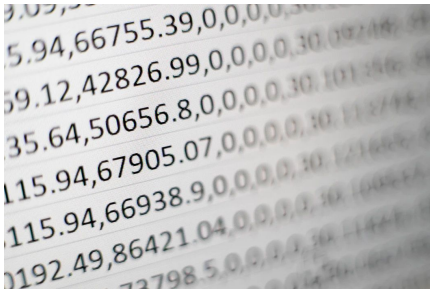
Causal relationships between two binary traits



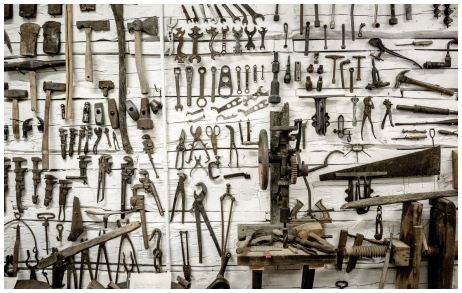
How?

phylogeny

data of interest



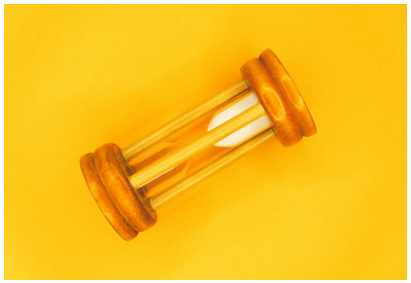
appropriate methods



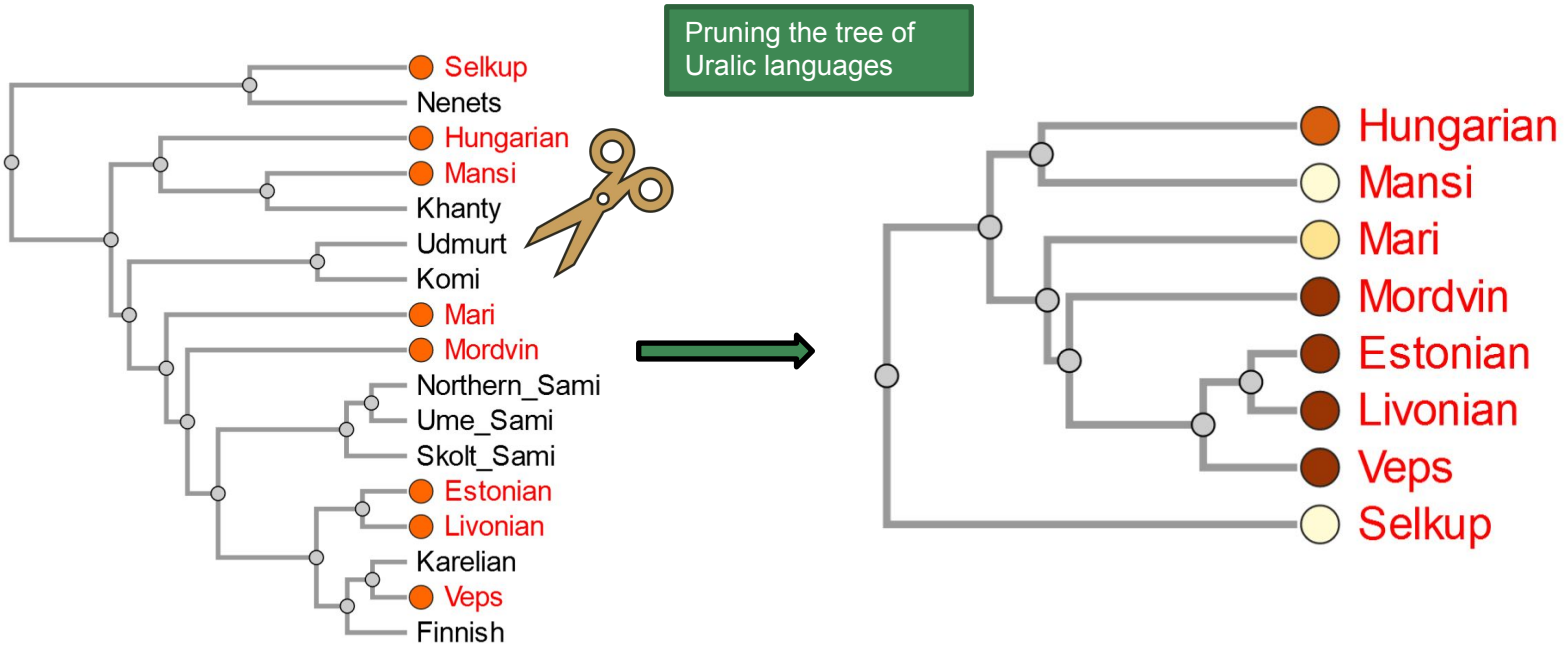
1

2

asking big-picture questions and testing hypotheses to investigate history of languages and cultures around the world



How?

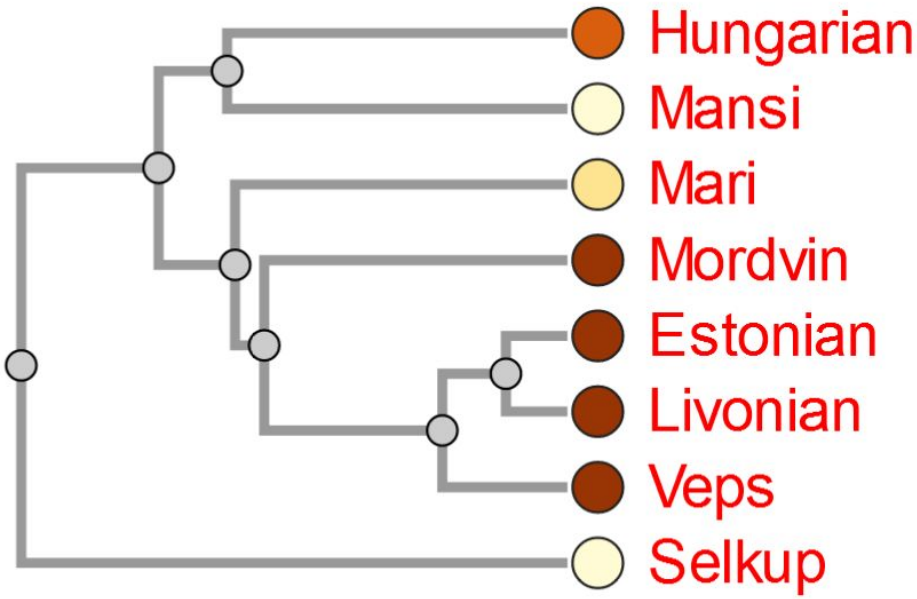


Honkola et al. 2013

Kirby et al. 2016

D-PLACE

How?



Mapping political complexity data on the tree:

- acephalous
- simple chiefdom
- complex chiefdom
- state
- large state

Our toolkit for today

Choosing appropriate methods based on the data and research question

BayesTraits

- a computer package used for analyzing trait evolution among species for which a phylogeny (phylogenies) is available (Meade & Pagel 2016)
- analysis of 1 or 2 variables (traits) that either vary continuously or take several discrete states

Phylogenetic path analysis (*phylopath*)

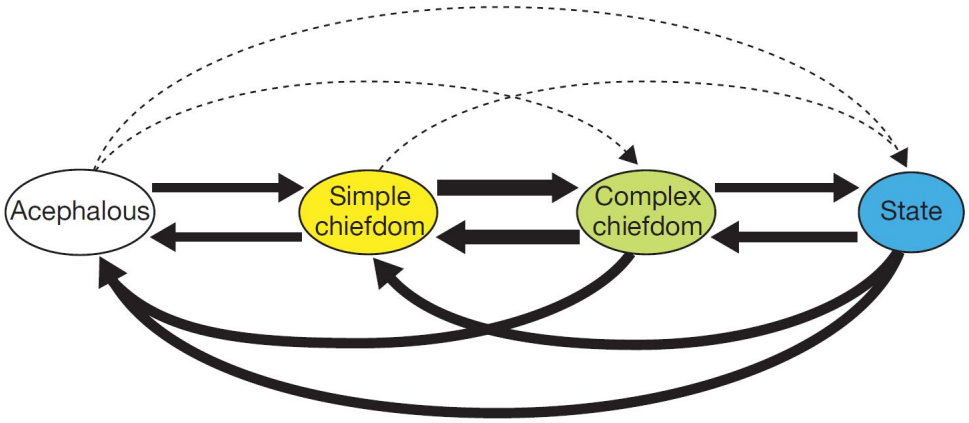
- a method used for comparing causal models and described in von Hardenberg & Gonzalez-Voyer (2013)
- implemented in R package *phylopath* (van der Bijl 2018)
- analysis of several variables of different types

Methods within BayesTraits

- *MultiState* is used for:
  - ancestral states reconstruction & testing models of trait evolution of a *trait with two or more discrete states* (e.g. 0, 1, 2, & 3; A, B, C)
- *Discrete* is used for:
  - testing for coevolution between pairs of *discrete binary traits* (e.g. 0 & 1; “low” & “high”)
- *Continuous* is used for:
  - testing for coevolution between pairs of *continuously varying traits (*Meade & Pagel 2016)

Method: MultiState

- *MultiState* is used for:
  - ancestral states reconstruction & testing models of trait evolution of a *trait with two or more discrete states* (e.g. 0, 1, 2, & 3; A, B, C)



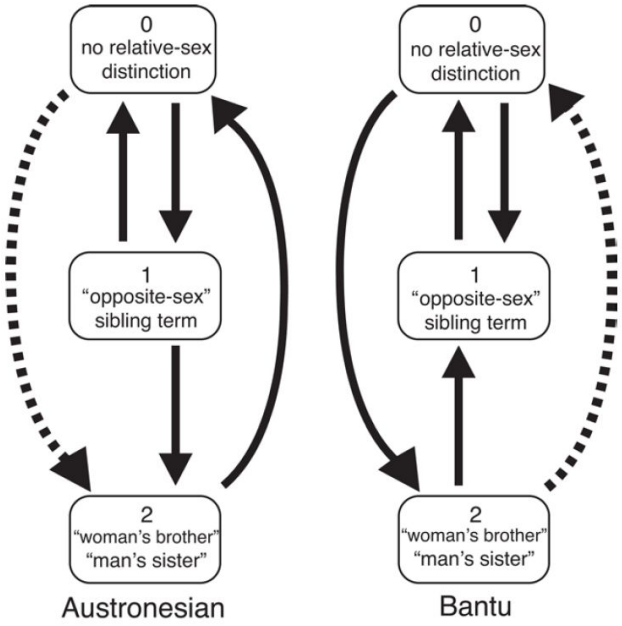
Currie et al. 2010

- 84 Austronesian languages (Gray et al. 2009)
- 1 “multistate” trait with 4 possible states

Method: MultiState

sibling

brother/sister



- 208 Austronesian languages (Gray et al. 2009)
- 73 Bantu languages (Aktinson et al. 2008)
- 1 "multistate" trait with 3 possible states

Jordan 2013

Method: Discrete

- *Discrete* is used for:
  - testing for coevolution between pairs of *discrete binary traits* (e.g. 0 & 1; “low” & “high”)

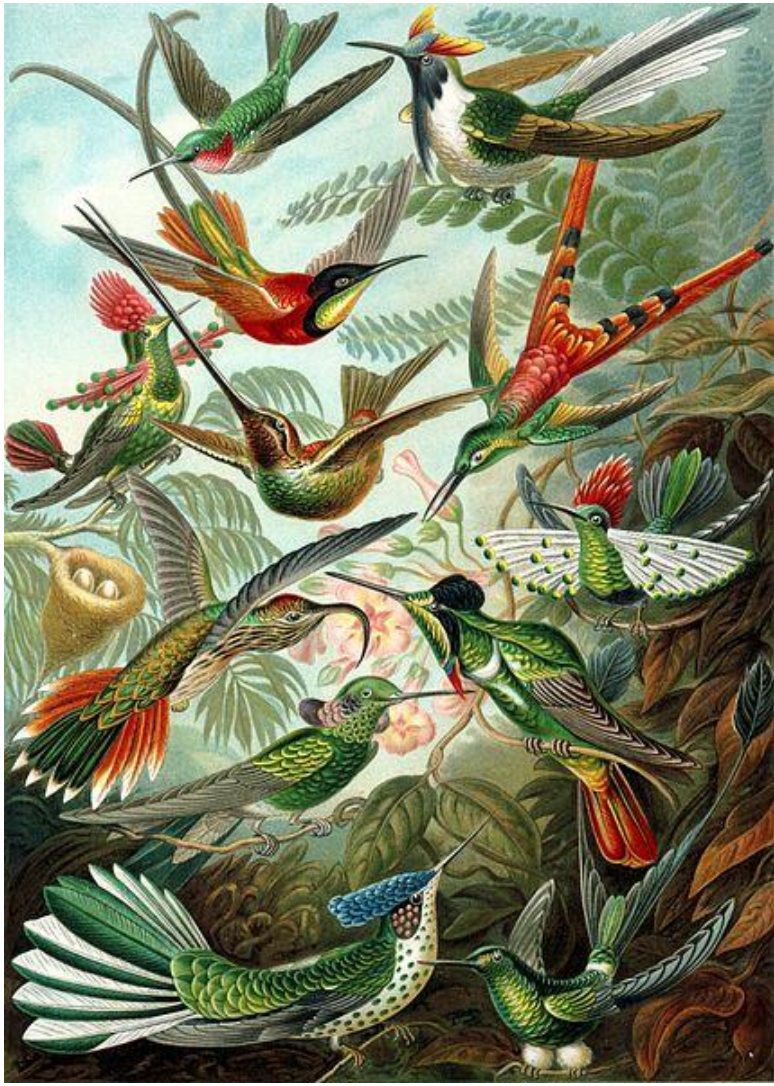
Coevolution

How to test for coevolution:

1) Fitting independent and dependent models of evolution onto phylogenies

- Under independent model, two traits of interest evolve separately and their changes are independent from each other
- Under dependent model, the changes in one trait dependent in changes in the other trait

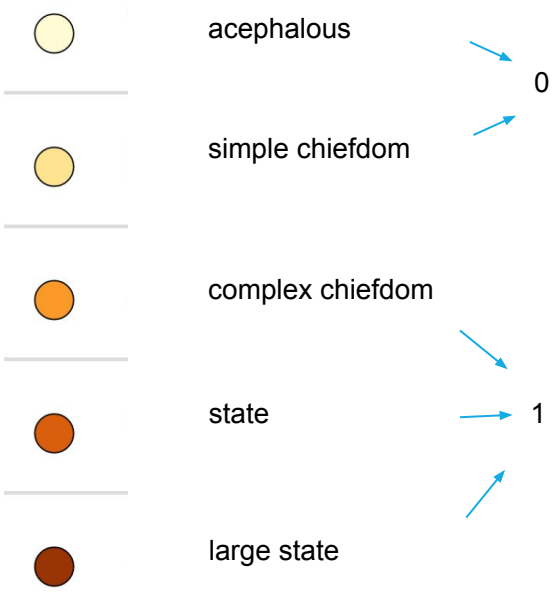
2) Statistically evaluating which model describes our data best

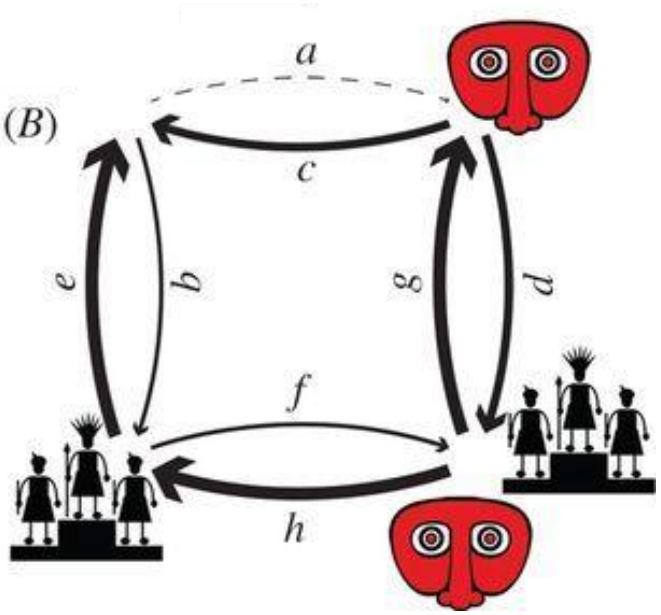


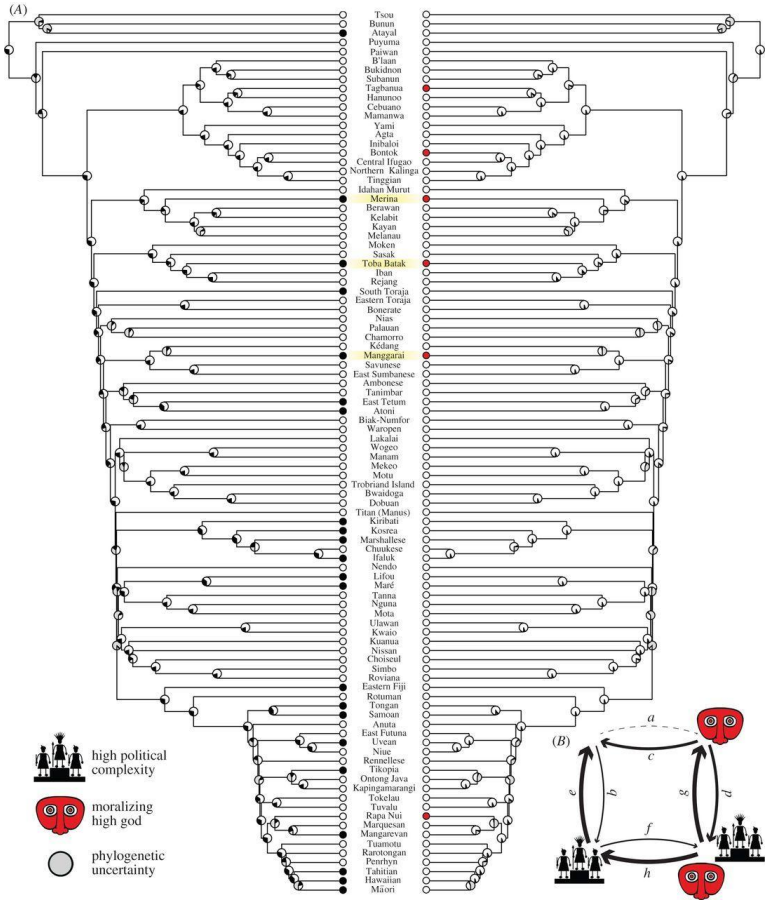
Ernst Haeckel's 1899 *Kunstformen der Natur*

Method: Discrete

- *Discrete* is used for:
  - testing for coevolution between pairs of *discrete binary traits*







Watts et al. 2015

Method: Discrete

- *Discrete* is used for:
  - testing for coevolution between pairs of *discrete binary traits*

Preposition (***in*** *the house*)

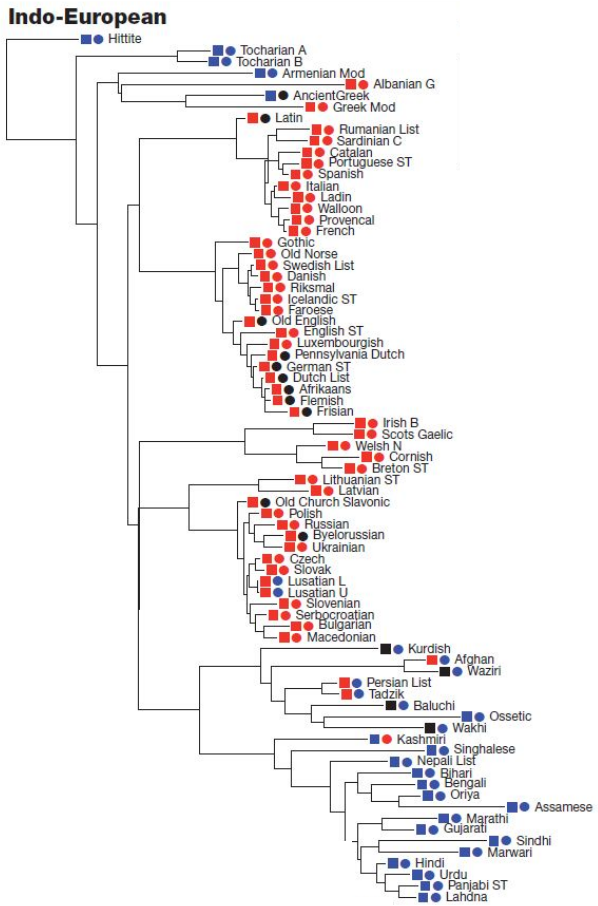
Postposition (*the house* ***in***)

Both (polymorphic state)

Order: verb–object (*eat an apple*)

Order: object–verb (*an apple eat*)

Both (polymorphic state)



Dunn et al. 2011

Preposition + order: verb–object

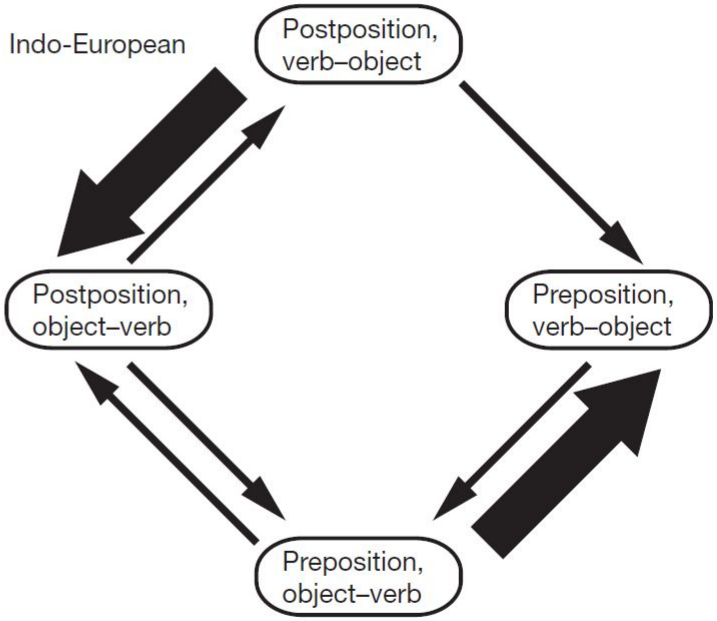
Postposition + order: object–verb

Preposition + order: object–verb

Postposition + order: verb–object

Preposition & postposition are possible

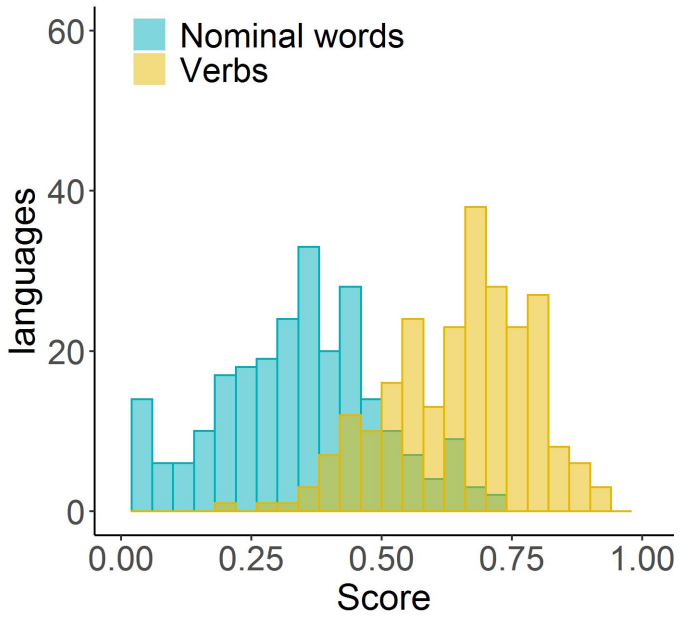
Orders: object–verb & verb–object are possible



79 Indo-European languages
(Gray & Atkinson 2003)

Dunn et al. 2011

Method: Continuous

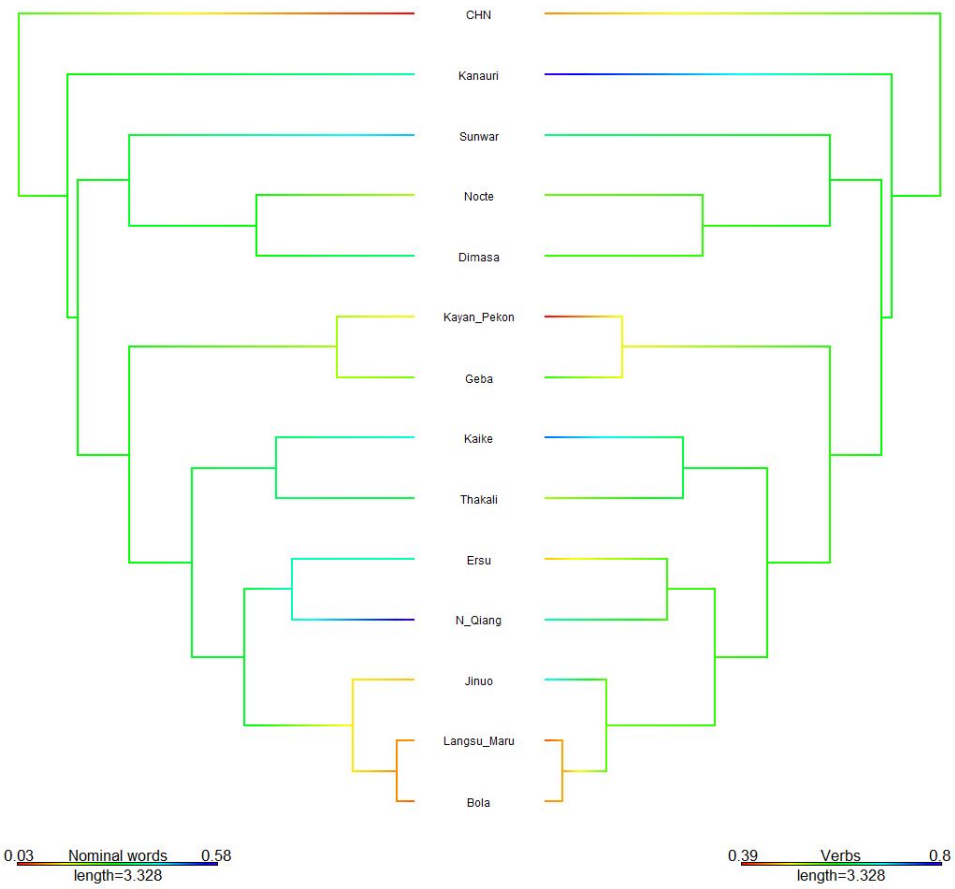


Shcherbakova et al. (submitted)

- *Continuous* is used for:
  - testing for coevolution between pairs of *continuously varying traits*

Amounts of grammatical coding in nominal words (nouns, pronouns, adjectives…) and verbs, with typological features information obtained from Grambank database

The Grambank Consortium 2021



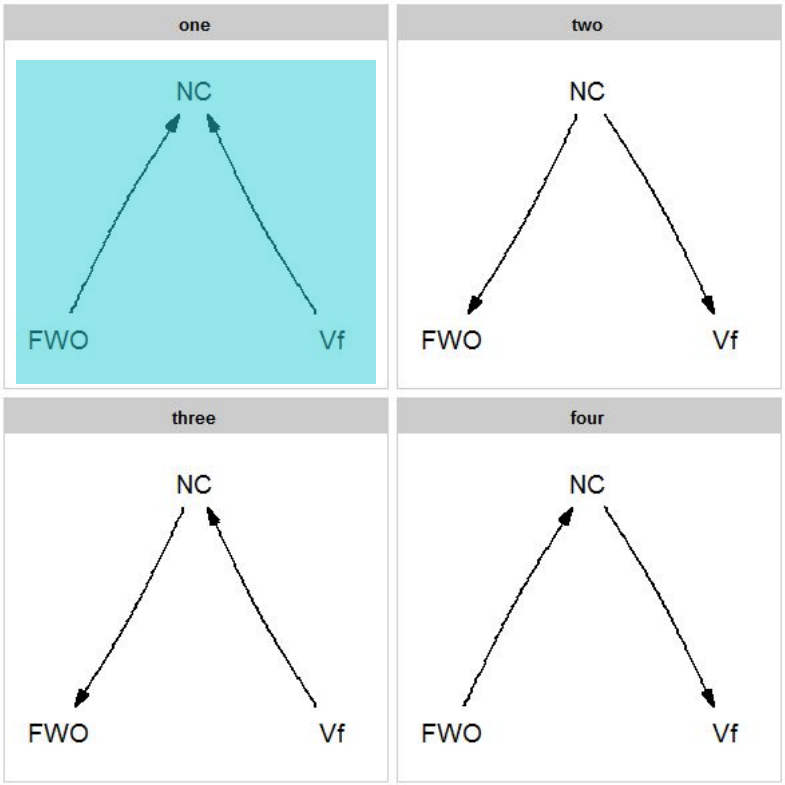
14 Sino-Tibetan languages (Zhang et al. 2020)

Shcherbakova et al. (submitted)

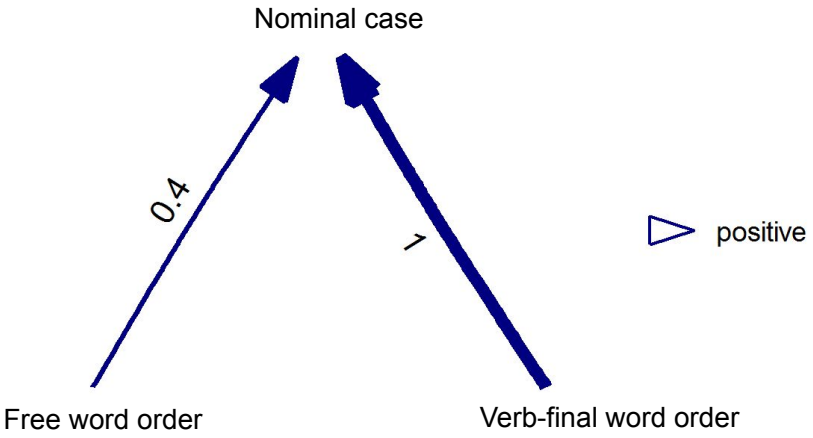
Method: phylogenetic path analysis

- a method used for comparing causal models and described in von Hardenberg & Gonzalez-Voyer (2013)
- implemented in R package *phylopath* (van der Bijl 2018)
- analysis of several variables of different types

Method: phylogenetic path analysis



Method: phylogenetic path analysis



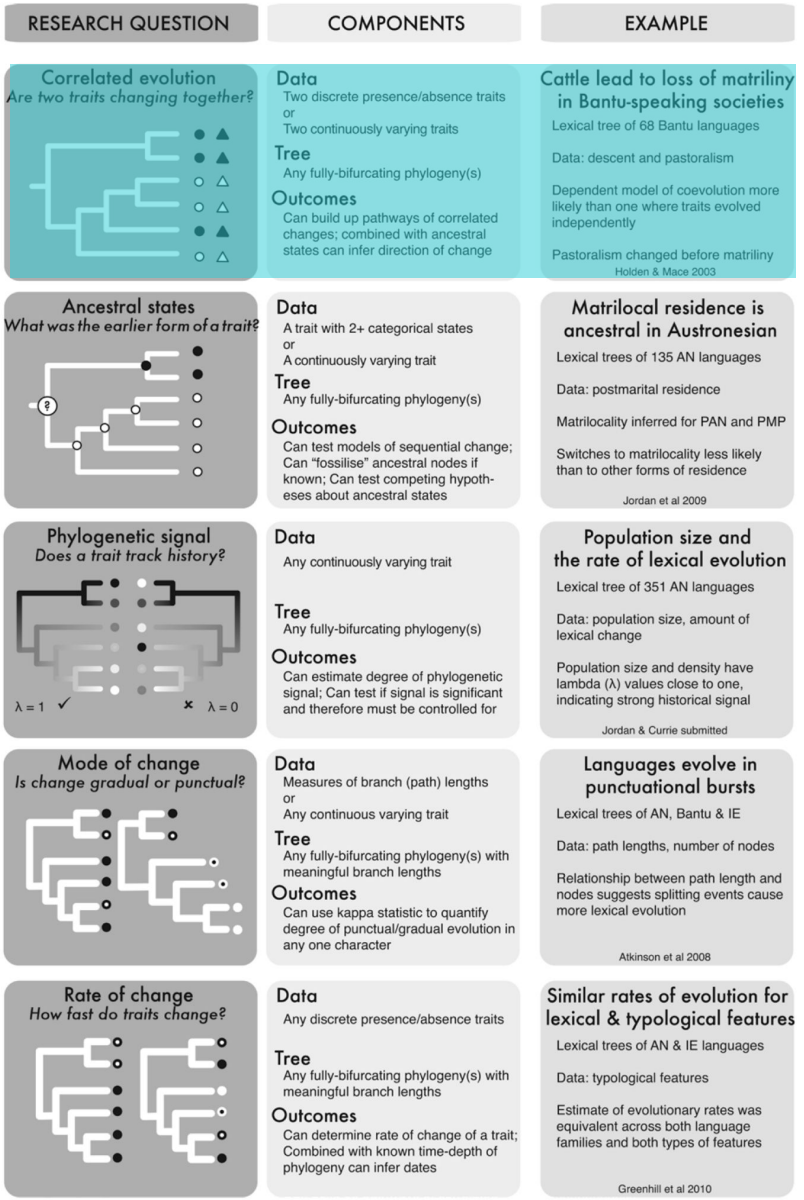
Shcherbakova et al. 2021 (conference paper)

Outlook

- Combining 1 binary & 1 continuous variable:
  - threshold model (Felsenstein 2012) implemented within R package *phytools* (Revell 2012)
- Incorporating control for geographic non-independence on top of controlling for relatedness:
  - e.g. brms (Bürkner 2018), spatiophylogenetic modelling (Dinnage et al. 2020)

Bayesian Phylogenetic Methods

Overview of five types of evolutionary questions that can be answered using phylogenetic comparative methods (Jordan 2013).



Questions?

References

Evans, Cara L., Simon J. Greenhill, Joseph Watts, Johann-Mattis List, Carlos A. Botero, Russell D. Gray & Kathryn R. Kirby. 2021. The uses and abuses of tree thinking in cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1828). 1–12. https://doi.org/10.1098/rstb.2020.0056.

Felsenstein, J. 1978. The number of evolutionary trees. Systematic zoology, 27(1), 27-33.

Greenhill, Simon J., Paul Heggarty & Russell D. Gray. 2020. Bayesian Phylolinguistics. In *The Handbook of Historical Linguistics*, 226–253. New York, USA: John Wiley & Sons. https://doi.org/10.1002/9781118732168.ch11.

Gray, R. D., Bryant, D., & Greenhill, S. J. 2010. On the shape and fabric of human history. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1559), 3923-3933.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., & Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. Royal Society open science, 5(3), 171504.

Jordan, Fiona M. 2013. Comparative phylogenetic methods and the study of pattern and process in kinship. In Patrick McConvell, Ian Keen & Rachel Hendery (eds.), Kinship systems: change and reconstruction, 43–58. Salt Lake City: University of Utah Press.

Reesink, G., Singer, R., & Dunn, M. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS biology*, *7*(11), e1000241.

Swadesh, Morris. 1964. K voprosy o povyshenii toçnosti v leksikostatistiçeskom datirovanii. Novoe v Lingvistike 1. 53-87.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. International Journal of American Linguistics 21. 121-137.

The Grambank Consortium (eds.) 2021. Grambank. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://grambank.clld.org, Accessed on 2021-08-11.)

McElreath, R., 2020. Statistical rethinking: A Bayesian course with examples in R and Stan. 2nd edition. https://doi.org/10.1201/9780429029608

Further Readings

Jacques, Guillaume & Johann-Mattis List. 2019. Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics* 9(1). 128–167. https://doi.org/10.1075/jhl.17008.mat.

Felsenstein, Joseph. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology* 22(3). 240–249. https://doi.org/10.1093/sysbio/22.3.240.

Greenhill, S. J., & Gray, R. D. (2009). Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. Austronesian historical linguistics and culture history: a festschrift for Robert Blust. Canberra: Pacific Linguistics, 375-397.

Pearl, Judea. 2000. *Causality*. Cambridge, Massachusetts: Cambridge University Press.

Verkerk, Annemarie. 2019. Detecting non-tree-like signal using multiple tree topologies. Journal of Historical Linguistics 9(1). 9–69. https://doi.org/10.1075/jhl.17009.ver.

McElreath, R., 2020. Statistical rethinking: A Bayesian course with examples in R and Stan. 2nd edition. https://doi.org/10.1201/9780429029608

Nelson-Sathi, S., List, J. M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W., & Dagan, T. (2011). Networks uncover hidden lexical borrowing in Indo-European language evolution. Proceedings of the Royal Society B: Biological Sciences, 278(1713), 1794-1803.

Further Readings

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill & Mark Pagel. Languages evolve in punctuational bursts. *Science* 319.5863 (2008): 588-588.

Bijl, Wouter van der. 2018. Phylopath: easy phylogenetic path analysis in R. *PeerJ* 6: e4718.

Bürkner, Paul-Christian. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10(1): 395–411.

Currie, Thomas E., Simon J. Greenhill, Russell D. Gray, Toshikazu Hasegawa, and Ruth Mace. 2010. Rise and fall of political complexity in Island South-East Asia and the Pacific. Nature 467 (7317): 801–4.

Dinnage, Russell, Alexander Skeels, and Marcel Cardillo. Spatiophylogenetic modelling of extinction risk reveals evolutionary distinctiveness and brief flowering period as threats in a hotspot plant genus. *Proceedings of the Royal Society B* 287.1926 (2020): 20192817.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473 (7345): 79–82.

Eff, E. Anthon. Does Mr. Galton still have a problem? Autocorrelation in the standard cross-cultural sample. *World Cultures* 15.2 (2004): 153-170.

Felsenstein, Joseph. A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist* 179.2 (2012): 145-156.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323 (5913): 479–83.

Gray, Russell D., and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426 (6965): 435–39.

Harrison, Xavier A., Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E.D. Goodwin, Beth S. Robinson, David J. Hodgson & Richard Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6 (2018): e4794.

Honkola Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen & Niklas Wahlberg. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *Journal of Evolutionary Biology*, 26(6):1244–1253.

Further Readings

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale & Michael C. Gavinet. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PloS one* 11.7 (2016): e0158391.

Meade, Andrew, and Mark Pagel. BayesTraits V3 manual. See http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/Files/BayesTraitsV3.Manual.pdf (2016).

Revell, Liam J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution* 3.2 (2012): 217-223.

Shcherbakova, Olena, Damián E. Blasi, Volker Gast, Russell D. Gray & Simon J. Greenhill. 2021. Evolution of case systems. 54th Annual Meeting of the Societas Linguistica Europaea, Athenes, Greece, 30 August-3 September, pp. 215-216.

Shcherbakova, Olena, Volker Gast, Damián E. Blasi, Hedvig Skirgård, Russell D. Gray, and Simon J. Greenhill. (submitted) Coevolution of nominal and verbal coding. In "Measuring Language Complexity", eds. Ehret, Katharina, Alice Blumenthal-Dramé, Christian Bentz & Aleksandrs Berdicevskis. Special Issue, *Linguistics Vanguard*.

Von Hardenberg, Achaz & Alejandro Gonzalez-Voyer. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution: International Journal of Organic Evolution* 67.2 (2013): 378-387.

Watts, Joseph, Simon J. Greenhill , Quentin D. Atkinson , Thomas E. Currie , Joseph Bulbulia & Russell D. Gray. Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proceedings of the Royal Society B: Biological Sciences* 282.1804 (2015): 20142556.

Zhang, Hanzhi, Ting Ji, Mark Pagel, and Ruth Mace. 2020. Dated phylogeny suggests early Neolithic origin of Sino-Tibetan languages. *Scientific Reports* 10 (1): 20792.

Images

https://emojipedia.org/whatsapp/2.19.352/see-no-evil-monkey/

https://omg-imatotalmess.tumblr.com/post/174538436737/imagine-draco-malfoy

https://unsplash.com/photos/tGTVxeOr_Rs

https://unsplash.com/photos/YVT5aF2QM7M

https://unsplash.com/photos/Wpnoqo2plFA

https://unsplash.com/photos/dyXoMMxDplY