Complexity clines out of Africa

Gerhard Jäger
joint work with Christian Fröhlich

Tübingen University

Workshop *Interaction and the Evolution of Linguistic Complexity*

Bayes Centre, Edinburgh, June 18, 2019

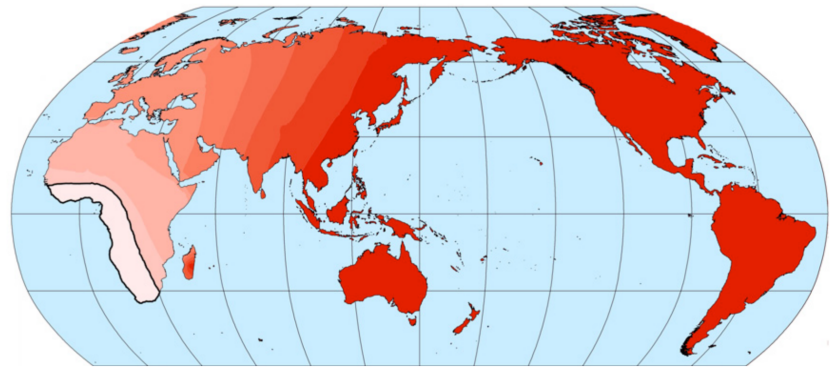
WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

DFG

Atkinson's observation

- Quentin Atkinson 2011 (Science, *Phonemic diversity supports a serial founder effect model of language expansion from Africa*):
  - phonemic diversity of languages decreases with distance from Africa
  - consistent with a **serial founder effect**
  - evidence for origin of modern humans from Africa



(Atkinson, 2011, Figure S7)

- correlations are only informative about real-world processes if there is a plausible causal story
- Atkinson:

The number of phonemes—perceptually distinct units of sound that differentiate words—in a language is positively correlated with the size of its speaker population (*1*) in such a way that small populations have fewer phonemes. Languages continually gain and lose phonemes because of stochastic processes (*2*, *3*). If phoneme distinctions are more likely to be lost in small founder populations, then a succession of founder events during range expansion should progressively reduce phonemic diversity with increasing distance from the point of origin, paralleling the serial founder effect observed in population genetics (*4*–*9*). A founder effect has already been

- causal mechanism rests on assumed correlation of inventory size with population size
- latter assumption debunked by Moran et al. (2012)

Looking at the data

Data used in this talk

- Phoible (Moran and McCloy, 2019)
  - phoneme inventories of 2,186 doculects
- ASJP (Wichmann et al., 2018)
  - 40-item Swadesh lists of 7,655 doculects
- intersection (linked via ISO code; only one doculect per ISO): 1,455 languages

Inventory Scots (Northern) (EA 2505)

Segment list | IPA chart

Showing 1 to 37 of 37 entries

← Previous 1 Next →

Segment class	Segment	Marginal	Allophones	Representation
--any--	Search	--any--	Search	Search
consonant	m	False		2914 (96%)
vowel	l	False		2779 (92%)
consonant	j	False		2716 (90%)
vowel	u	False		2646 (88%)
vowel	a	False		2600 (86%)
consonant	w	False		2483 (82%)
consonant	n	False		2349 (78%)
consonant	s	False		2020 (67%)
consonant	b	False		1906 (63%)
consonant	ŋ	False		1897 (63%)
vowel	e	False		1841 (61%)
vowel	o	False		1826 (60%)
consonant	g	False		1712 (57%)
consonant	h	False		1703 (56%)
consonant	d	False		1376 (46%)
consonant	f	False		1329 (44%)
consonant	tʃ	False		1218 (40%)
vowel	ɛ	False		1129 (37%)
consonant	ʃ	False		1104 (37%)
vowel	ɔ	False		1070 (35%)
consonant	z	False		893 (30%)
consonant	v	False		816 (27%)
consonant	r	False		774 (26%)
consonant	kʰ	False		605 (20%)
consonant	pʰ	False		592 (20%)
consonant	x	False		576 (19%)
consonant	ʒ	False		478 (16%)
vowel	ɪ	False		444 (15%)
consonant	tʰ	False		403 (13%)
vowel	ʌ	False		133 (4%)
consonant	θ	False		123 (4%)
consonant	ʍ	False		40 (1%)
consonant	ɬ	False		31 (1%)
vowel	əi	False		18 (1%)
vowel	əu	False		17 (1%)
vowel	ae	False		8 (0%)
vowel	ɔe	False		3 (0%)

Showing 1 to 37 of 37 entries

← Previous 1 Next →

Contributor

Dmitry Nikolaev cite

Sources

Millar 2007

Scots

Map

Coordinates WGS84 55°53'N, 3°43'W
55.89, -3.71

Links

Wordlist Scots

Compiled by André Müller and Helen Geyer

Showing 1 to 59 of 59 entries

← Previous 1 Next →

No.	Meaning	Word	Loan
Search	Search	Search	--any--
1	I	a	False
1	I	3	False
2	you	yI	False
2	you	yu	False
3	we	wI	False
3	we	w3	False
11	one	en	False
11	one	e	False
12	two	tw~o	False
18	person	bodI	False
19	fish	fiS	False
21	dog	dog	False
21	dog	doug	False
22	louse	lus	False
22	louse	pulI	False
23	tree	trI	False
25	leaf	lIf	False
25	leaf	blad	False
28	skin	skin	False
30	blood	blId	False
31	bone	ben	False
34	horn	horn	False
39	ear	Ir	False
40	eye	I	False
40	eye	klk3r	False
41	nose	noz	False
41	nose	nEb	False
43	tooth	te8	False
43	tooth	tI8	False
44	tongue	toN	False
47	knee	nI	False
47	knee	knI	False
48	hand	hond	False
48	hand	han	False
51	breast	birzk3t	False
51	breast	borzk3t	False
54	drink	drINk~	False
54	drink	bEb	False
57	see	sI	False
58	hear	hIr	False
61	die	dI	False
66	come	kam	False
66	come	k3m	False
72	sun	son	False
72	sun	sIn	False
74	star	starn	False
74	star	blENk~3r	False
75	water	wat3r	False
77	stone	sten	False
82	fire	f3I3r	False
82	fire	IN3I	False
85	path	pad	False
85	path	p38	False
86	mountain	bEn	False
86	mountain	montIn	False
92	night	nIt	False
95	full	fu	False
96	new	ny~u	False
100	name	nem	False

Showing 1 to 59 of 59 entries

← Previous 1 Next →

Glottocode: scot1243 ISO 639-3: sco

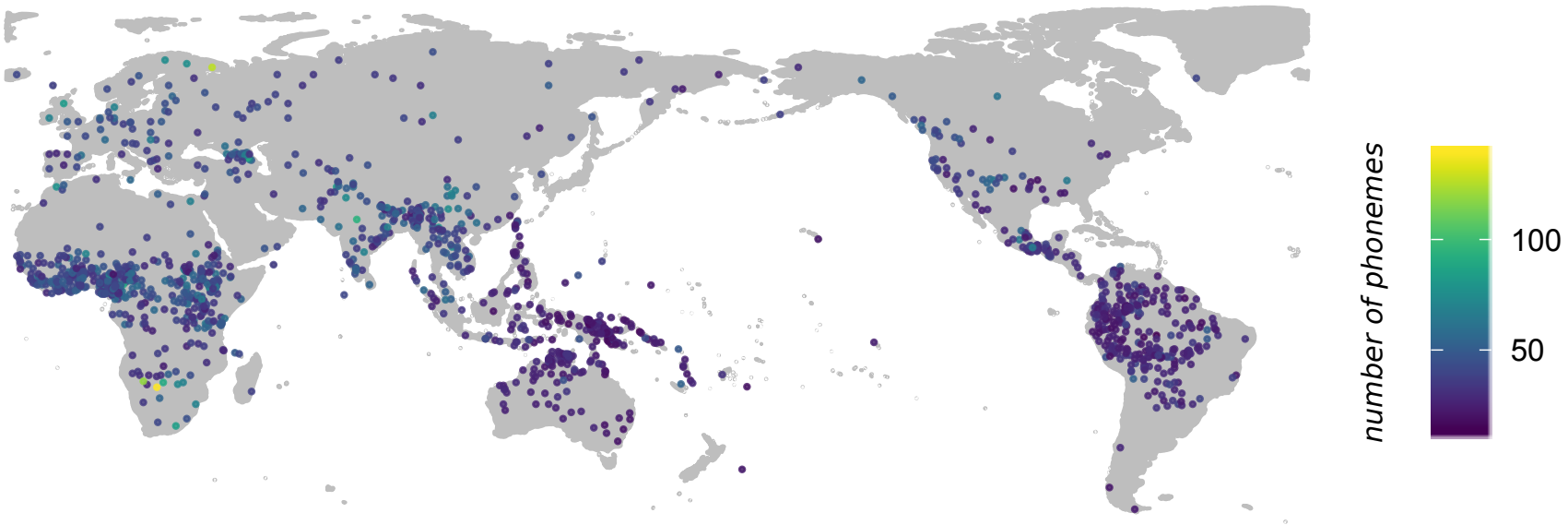
Coordinates WGS84	55°52'N, 4°16'W 55.86, -4.26
number of speakers	99,200
status	alive

Classification

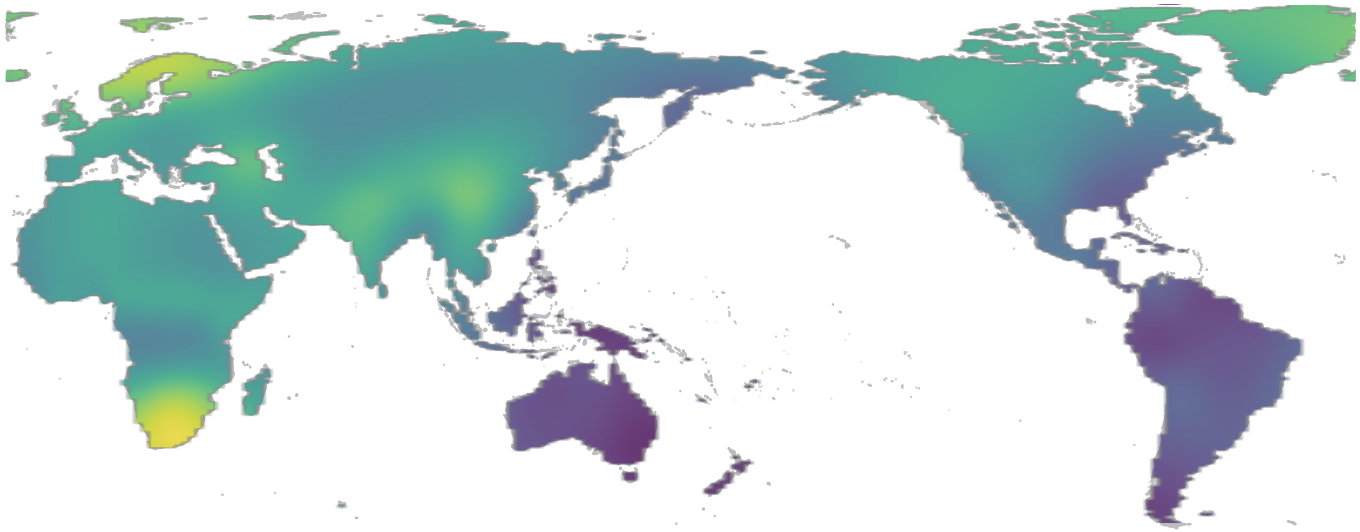
WALS
IE › Germanic
Glottolog
Indo European › Germanic › Northwestgermanic › Westgermanic › Northseagermanic › Anglo Frisian › Anglian
Ethnologue
Indo European › Germanic › West › English

Sources

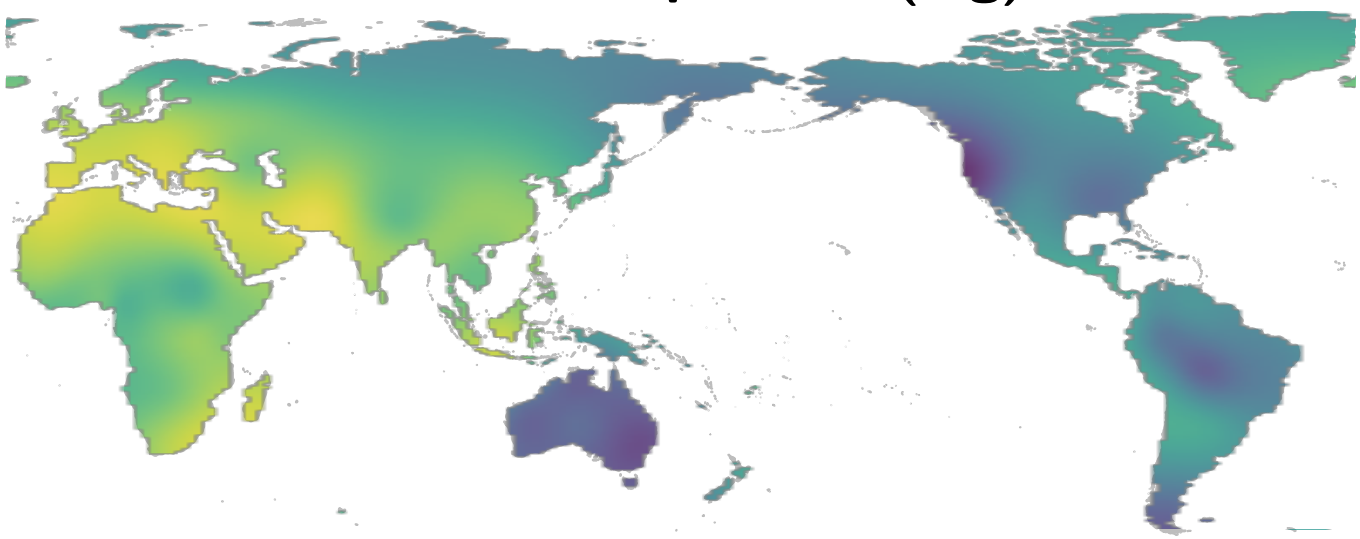
n.a. n.d.
http://www.scots-online.org/dictionary/engscots.asp


number of phonemes
100
50

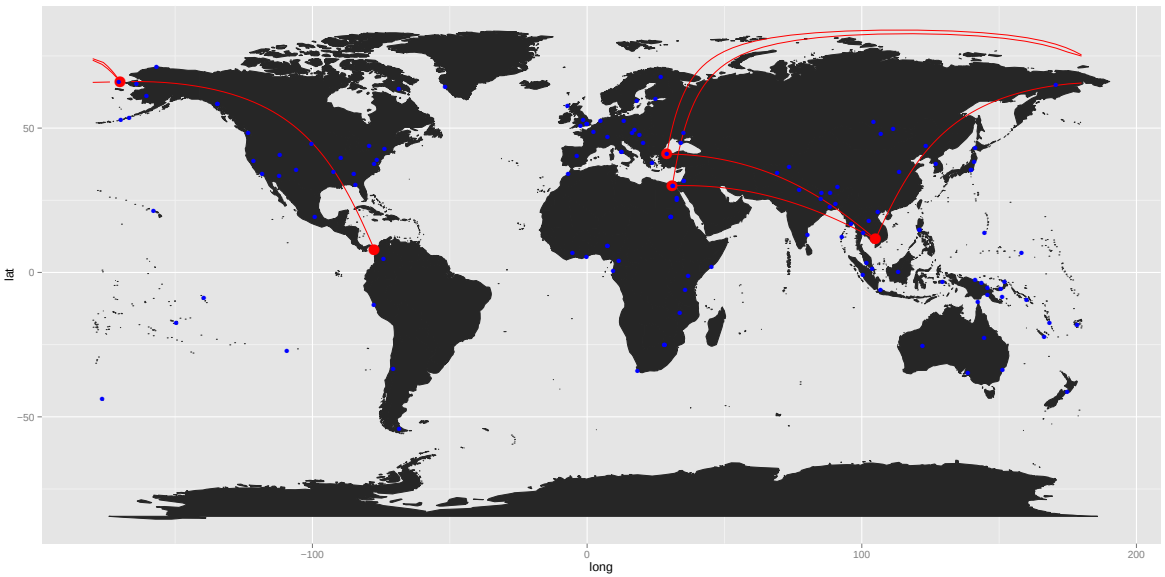
number of phonemes



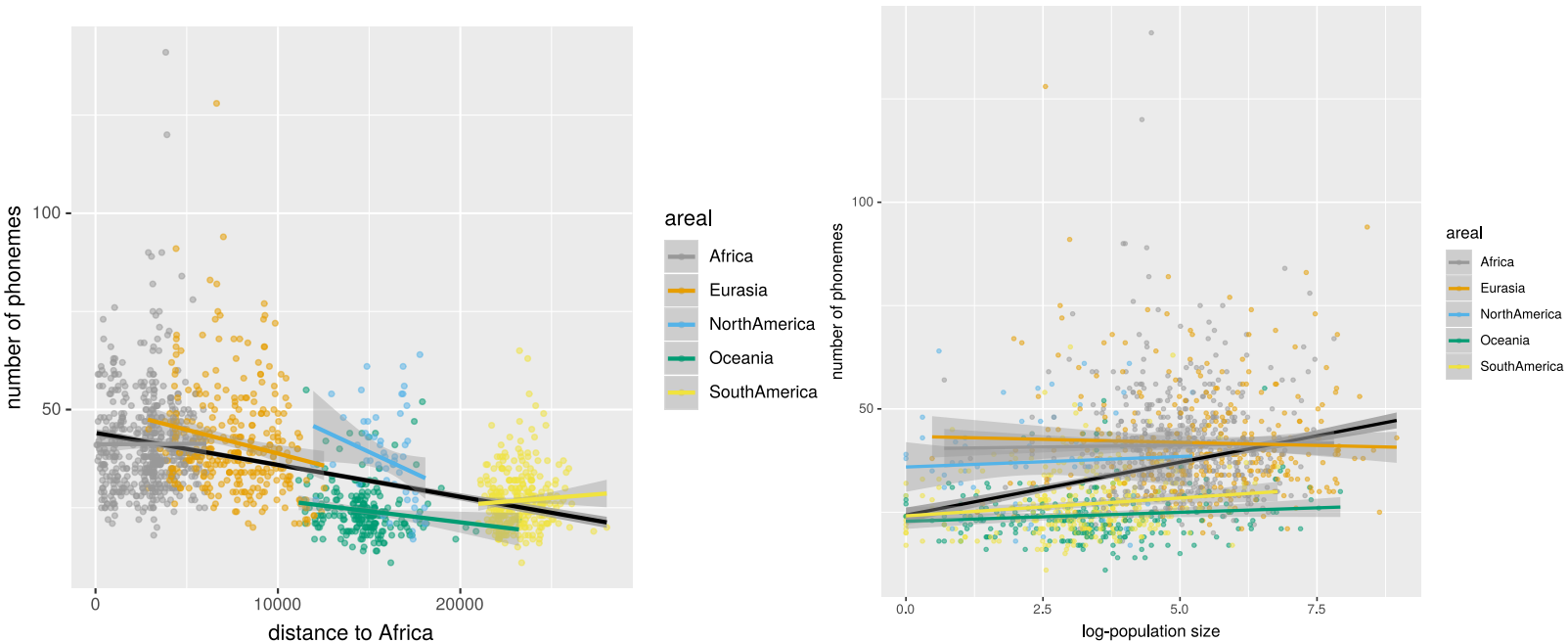
number of speakers (log)



Land-based distances

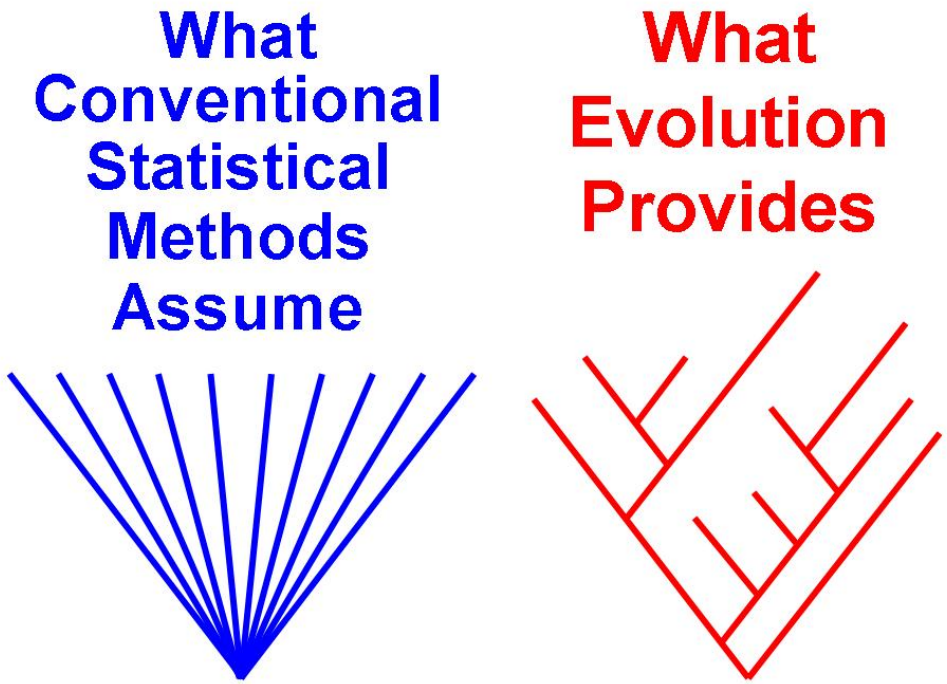


- following Atkinson 2011:
  - Africa/Asia: *Cairo*
  - Asia/Europ: *Istanbul*
  - Asia/Oceania: *Phnom Phen*
  - Asia/North America: *Bering Strait*
  - North America/South America: *Panama*
- following Wichmann et al. (2011), I use Addis Ababa as putative place of origin here


number of phonemes
100
50
0
10000
20000
distance to Africa
areal
Africa
Eurasia
NorthAmerica
Oceania
SouthAmerica
number of phonemes
100
50
0.0
2.5
5.0
7.5
log-population size
areal
Africa
Eurasia
NorthAmerica
Oceania
SouthAmerica

Controlling for phylogenetic non-independence

Spurious correlations in cross-lingustic data

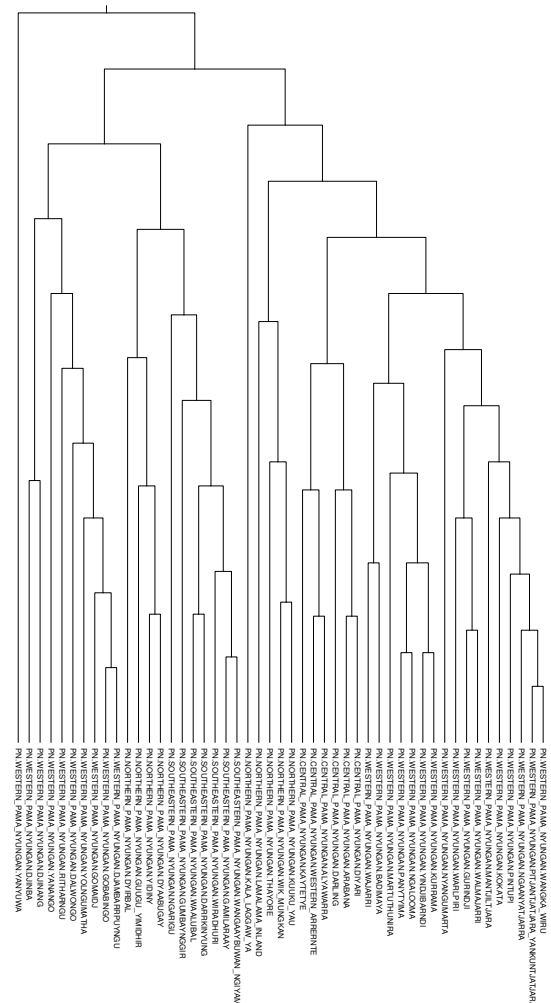


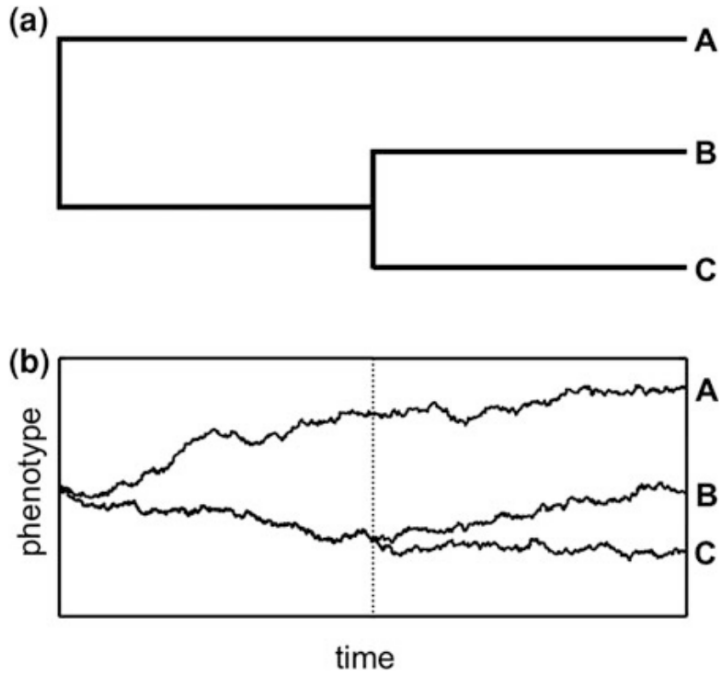
Treating different languages as independent samples creates many spurious correlations!

- mixed-effects models, with genealogical units and/or linguistic macro-areas as random factors, mitigate this problem somewhat
- genealogical non-independence is no binary distinction though; Scots is related to Hindi, but it much closer related to English
- degree of relatedness can be estimated via **phylogenetic inference** from lexical data

The phylogenetic comparative method

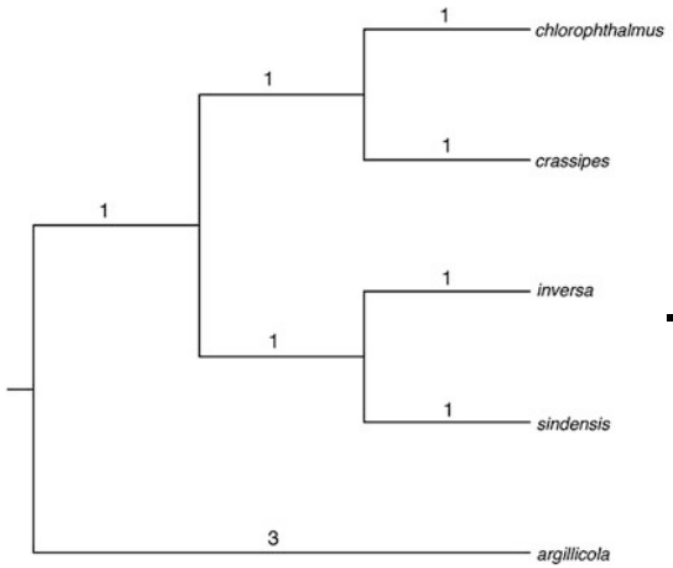
- data-generating process is governed by a **phylogenetic tree**
- simplest model for continuous variables:
  - variable evolves according to **Brownian motion** from root to leaves
  - when branch splits, to independent copies keep evolving randomly





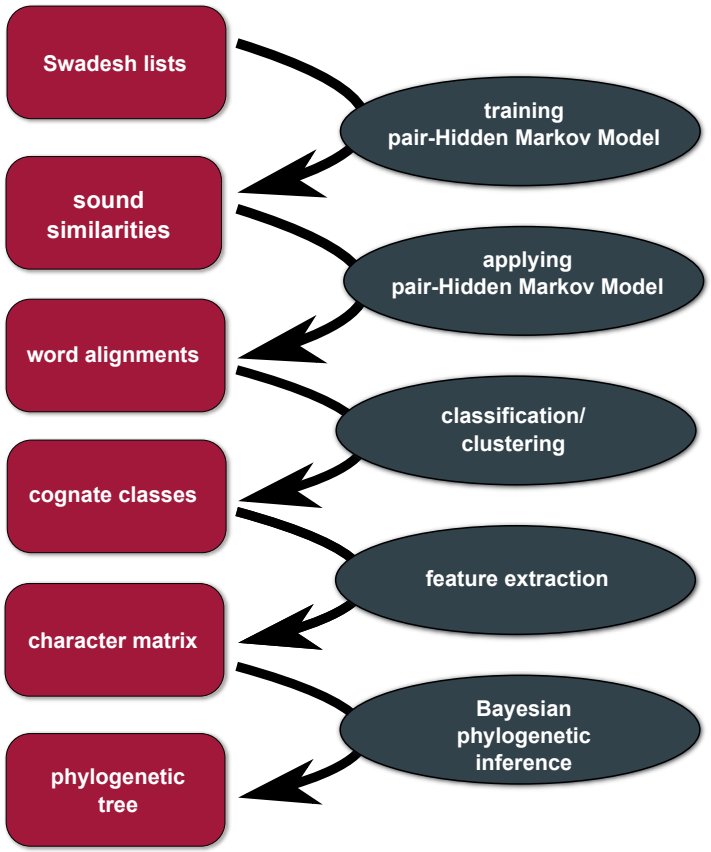
- covariance between two leaves $\propto$ time traveled together

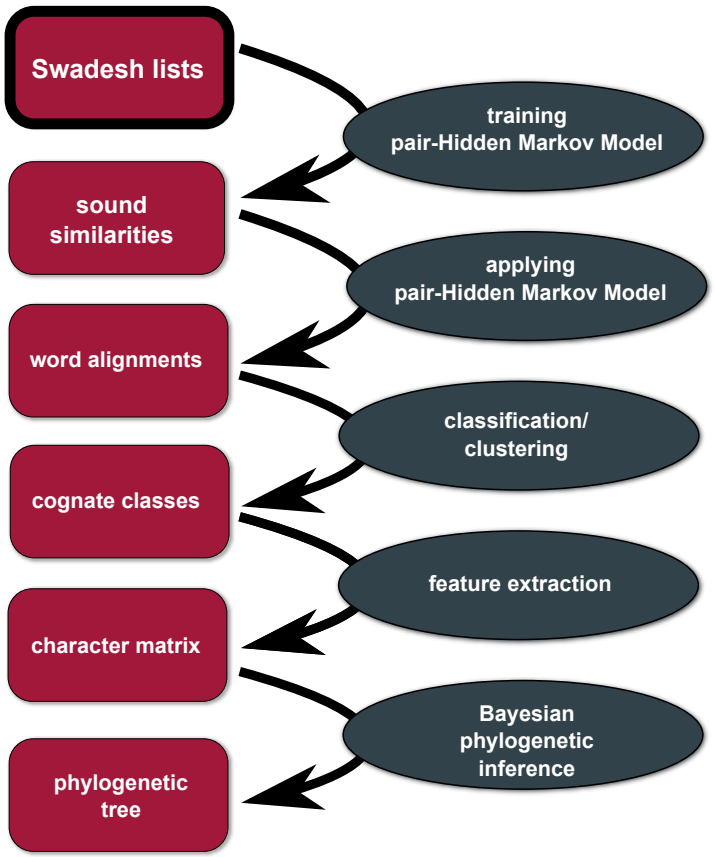
- measurements at the leaves are distributed according to multivariate normal distribution
- variance-covariance matrix is determined by tree



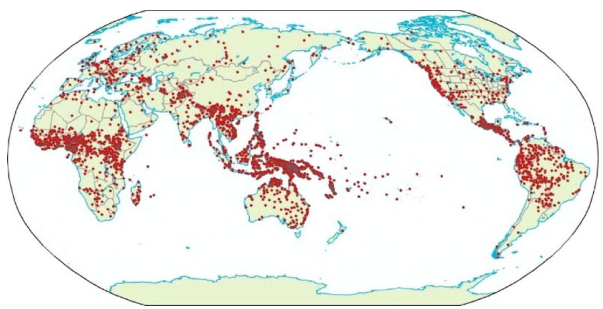
$$\mathbf{C}_{\mathbf{phyl}} = \begin{bmatrix} 3 & 2 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 & 0 \\ 1 & 1 & 3 & 2 & 0 \\ 1 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

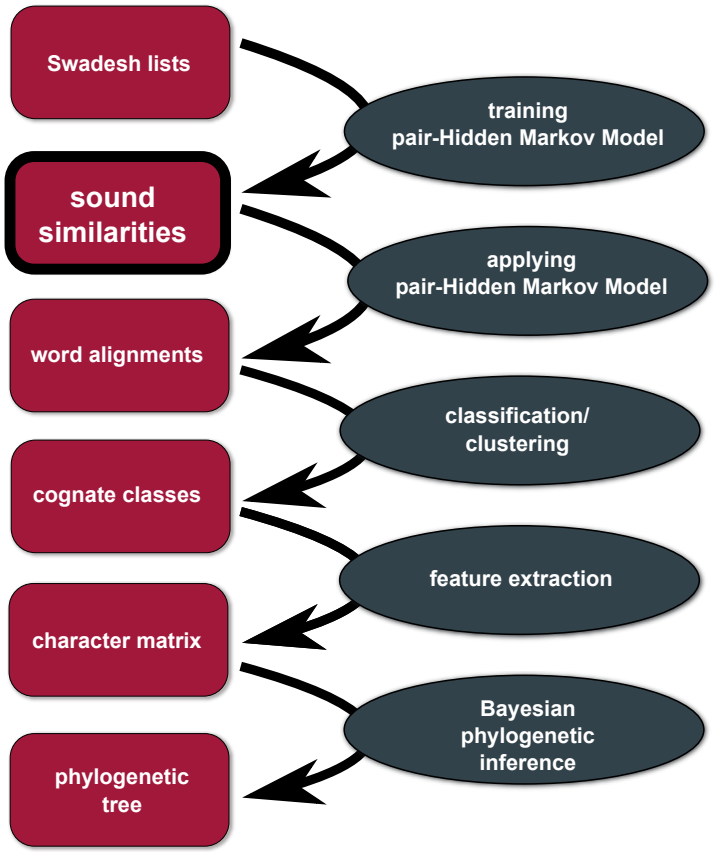
Inferring a language tree

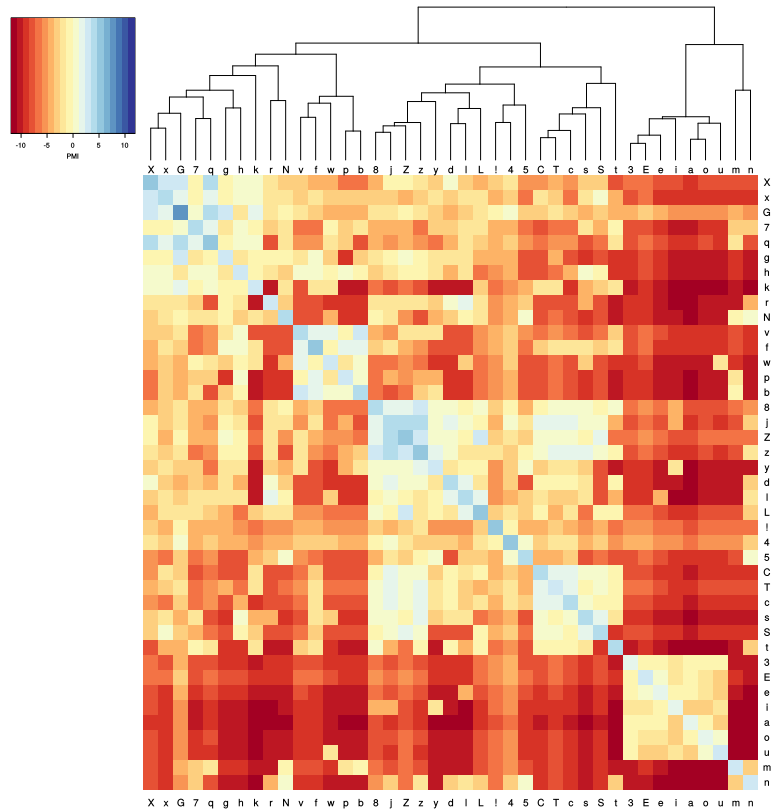

Swadesh lists
sound similarities
word alignments
cognate classes
character matrix
phylogenetic tree
training pair-Hidden Markov Model
applying pair-Hidden Markov Model
classification/ clustering
feature extraction
Bayesian phylogenetic inference

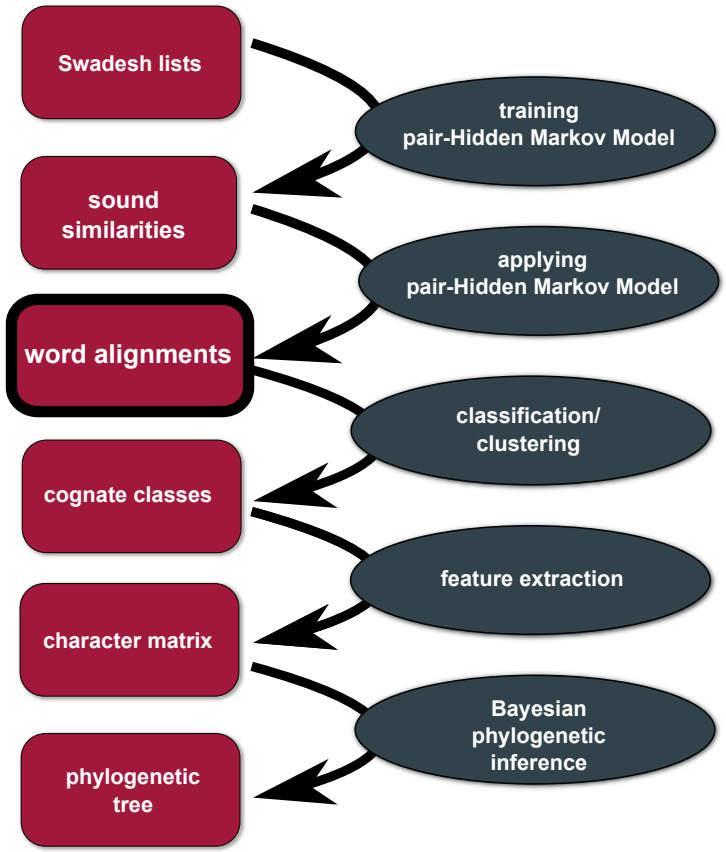


concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

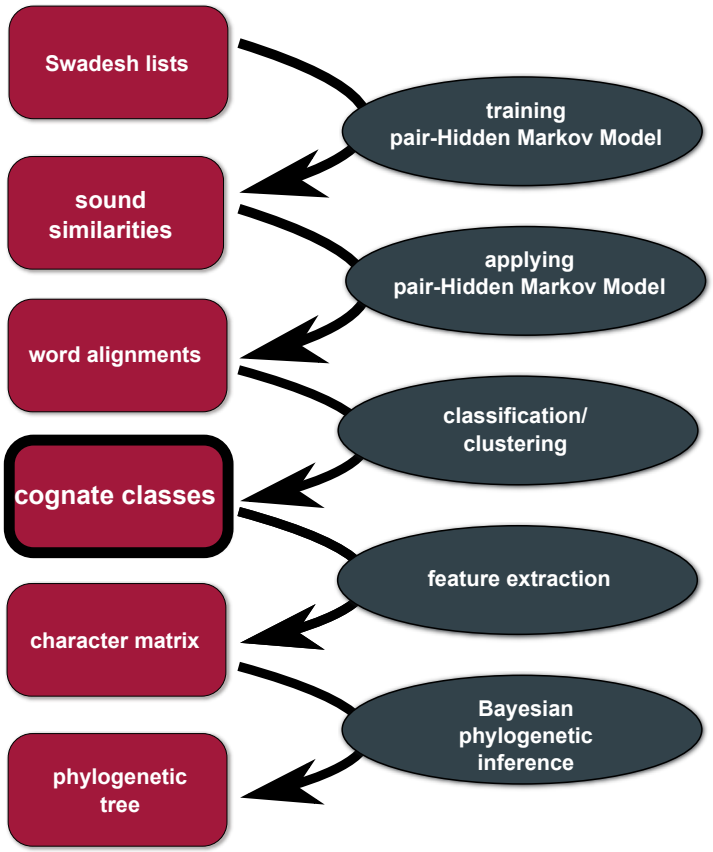


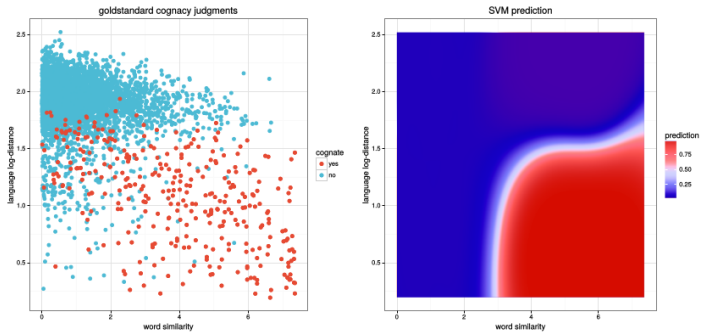

Swadesh lists
sound similarities
word alignments
cognate classes
character matrix
phylogenetic tree
training pair-Hidden Markov Model
applying pair-Hidden Markov Model
classification/ clustering
feature extraction
Bayesian phylogenetic inference


PMI
X x G 7 q g h k r N v f w p b 8 j Z z y d l L ! 4 5 C T c s S t 3 E e i a o u m n

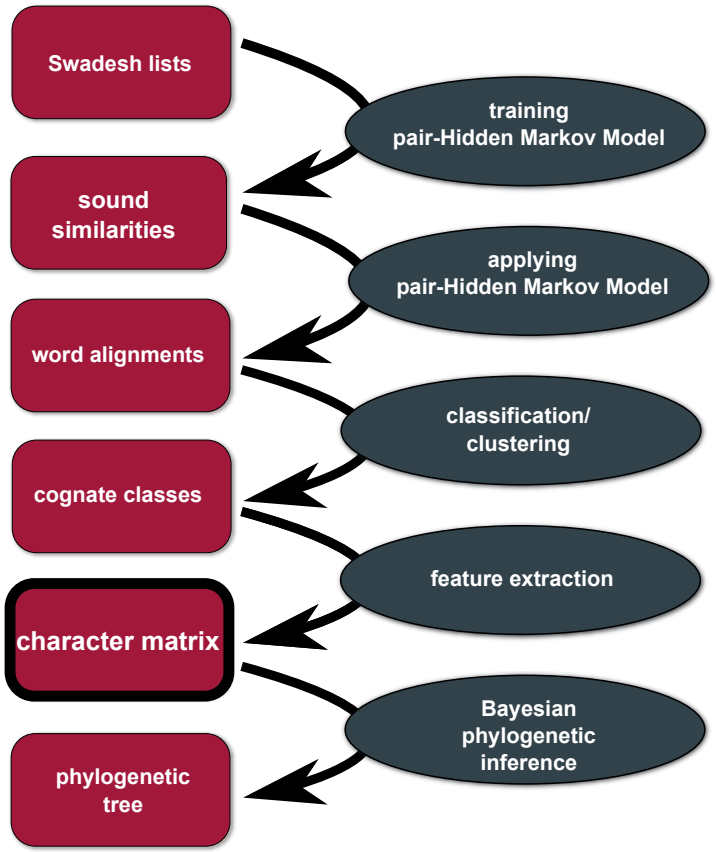


Language	*fish:z*	*tongue:1*	*smoke:1*
Abui-Atangmelang	-af-u		
Abui-Fuimelang	-af-u	tal-i-fi--	
Adang	aab--	tal-E-b---	awai--b-a-n-o-7o-
Blagar-Bakalang	-ab--	--j-e-bur-	--ad--b-a-n-aNka-
Blagar-Bama	aab--	teg-e-bur-	------b-e-n-a-xa-
Blagar-Kulijahi	-ab--	tej-e-bur-	------b-e-n-aNka-
Blagar-Nule	aab--	tej-e-bur-	--ad--b-e-n-aNka-
Blagar-Tuntuli	aab--	tej-e-bur-	a-adgeb-a-n-a-q--
Blagar-Warsalelang	-ab--	tel-e-bur-	a-ad--b-a-n-a-x--
Bunaq			------b-o-t-o-h--
Deing	haf--		------buu-n------
Hamap	7ab--	nar-ø-buN-	------b-a-n-o-7--
Kabola	hab--	tal-e-b---	awal--b-e-n-e-7o-
Kaera-Padangsul	-ab--	talee-b---	a-ad--b-e-naa-x--
Kafoa	-afUi	tal-i-p---	------f-o-n-a----
Kamang	-ap-i	nal---pu--	------p-u-n----a-
Kiraman	-Eb--	nal-i-bar-	--ar--b-a-n-o-kan
Klon	-eb-i	gel-E-b---	--ed-ab-o-n------
Kui	-eb--	tal-i-ber-	--ar--b-o-n-o-k--
Kula	-ap-i	-il-I-p---	------p---n-ekka-
Nedebang	aaf-i	gel-e-fu--	--ar-ab-u-n------
Reta	aab--	nal-e-bul-	a-ad--b-o-n-a----
Sar-Adiabang	haf--	--p-e-fal-	--ar--buu-n------
Sar-Nule	haf--	nal-e-faj-	
Sawila	-ap-i	gal-impuru	------p-u-n-a-ka-
Teiwa-Madar	xaf--	gel-i-vi--	------buu-n------
Wersing	-ap-i	nej-e-bur-	--ad-ap-u-n-a-k--
Wpantar	hap--	nal-e-bu--	------b-unn-a----

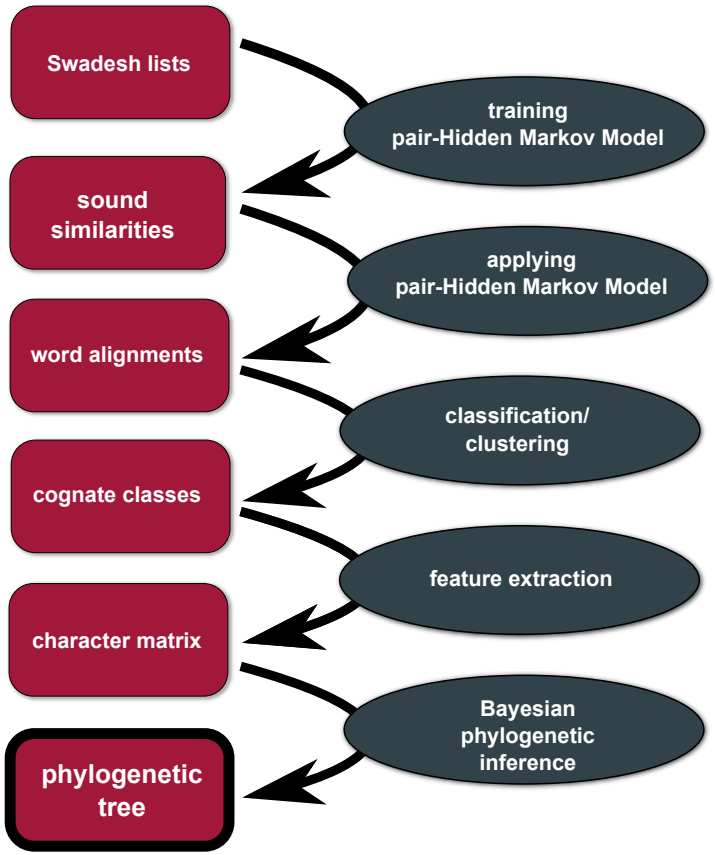


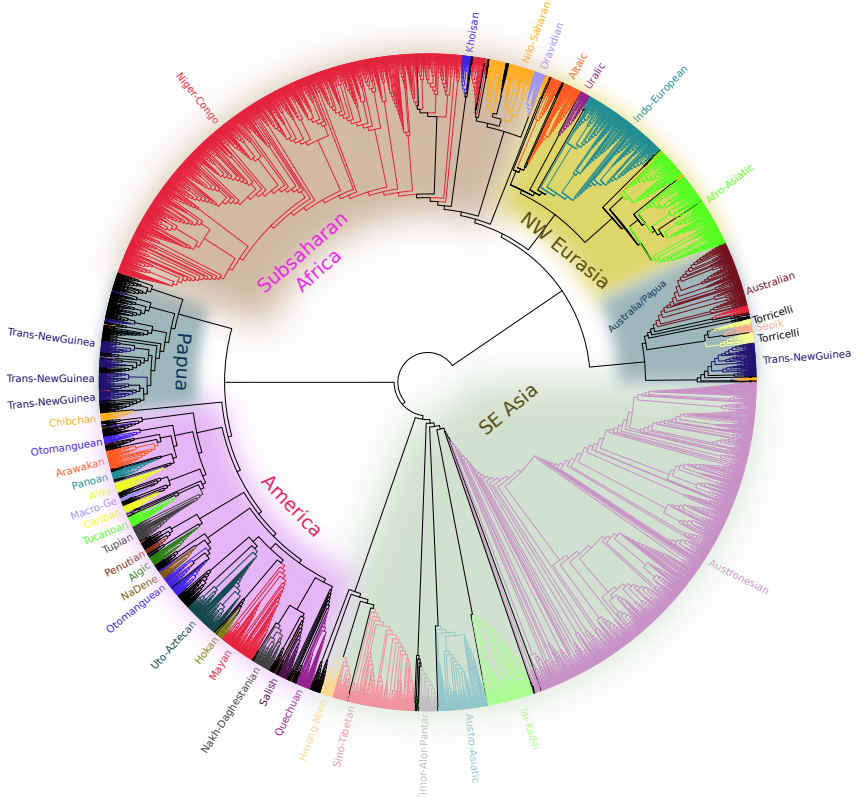


	English	Spanish	Modern Greek	Standard German
I	Ei:A	yo:B	exo:C	iX:D
you	yu:A	ustet:B, tu:C	esi:D	du:E
we	wi:A	nosotros:B	emis:C	vir:A
one	w3n:A	uno:B	enas:C, ena:C	ains:D
two	tu:A	dos:B	8y∼o:C, 8io:D	cvai:E
person	pers3n:A	persona:A	an8∼ropos:B	mEnS:C
fish	fiS:A	peskado:A, pes:A	psari:B	fiS:A
dog	dag:A	pero:B	sTili:C, sTilos:C	hunt:D
come	k3m:A	veni:B	erx∼o:C	kh∼om3n:A
sun	s3n:A	sol:B	ily∼os:C, iLos:C	zon3:A
star	star:A	estreya:A	asteri:A, astro:A	StErn:A
water	wat3r:A	agw∼a:B	nero:C	vas3r:A
stone	ston:A	piedra:B	petra:B	Stain:A
fire	fEir:A	fuego:B	foty∼a:C	foia:D
path	pE8:A	senda:B	8romos:C	pf∼at:A, vek:D
mountain	maunt3n:A	sero:B, monta5a:A	vuno:C, oros:D	bErk:E
full	ful:A	yeno:B	yematos:C, pliris:D	fol:A
new	nu:A	nuevo:A	neos:A, Tenury∼os:B	noi:A
name	nem:A	nombre:A	onoma:A	nam3:A

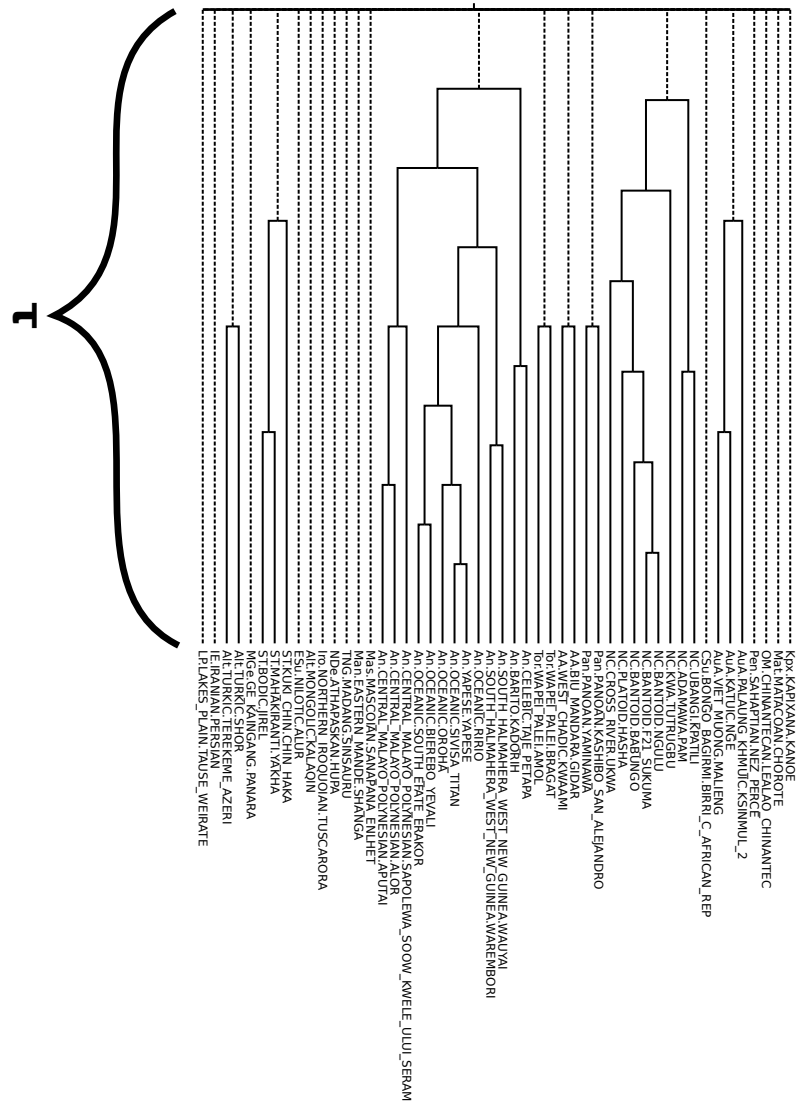



```
TNG.ENGAN.MAIBI                 1000000000000000000000000000000000000000000›
TNG.ENGAN.POLE                  0000000000000000000000000000000000000010000000›
TNG.ENGAN.SAU                   0000000000000000000000000000000000000000000000›
TNG.ENGAN.YARIBA                1000000000000000000000000000000000000000000000›
TNG.FASU.FASU                   0000000000000000000000000000000000000010000000›
TNG.FASU.NAMUMI                 0000000000000000000000000000000000000001000000›
TNG.FINISTERRE-HUON.AWARA       0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.BORONG      0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.BURUM       0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.BURUM_MIND  0000000000000000000000000000000000001010000000›
TNG.FINISTERRE-HUON.DEDUA       0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.HUBE        0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.KATE        0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.KOMBA       0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.KOSORONG    0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.MAPE        0000000000000000000000000000000000001000000000›
TNG.FINISTERRE-HUON.MAPE_2      0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.MIGABAC     0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.MINDIK      0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.MOMOLILI    0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.NABAK       0000000000000000000000000000000000001000000000›
TNG.FINISTERRE-HUON.NANKINA     0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.NEK         0000000000000000000000000000000000000001000000›
TNG.FINISTERRE-HUON.NUKNA       0000000000000000000000000000000000000001000000›
TNG.FINISTERRE-HUON.ONO         0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.SELEPET     0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.TIMBE       0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.TOBO        0000000000000000000000000000000000000010000000›
TNG.FINISTERRE-HUON.WANTOAT     0000000000000000000000000000000000000110000000›
TNG.FINISTERRE-HUON.YOPNO       0000000000000000000000000000000000000001000000›
TNG.GOILALAN.AFOA               0000000000000000000000000000000000000110000000›
TNG.GOILALAN.KUNIMAIPA          0000000000000000000000000000000000000010000000›
TNG.GOILALAN.MAFULU             0000000000000000000000000000000000000010000000›
```

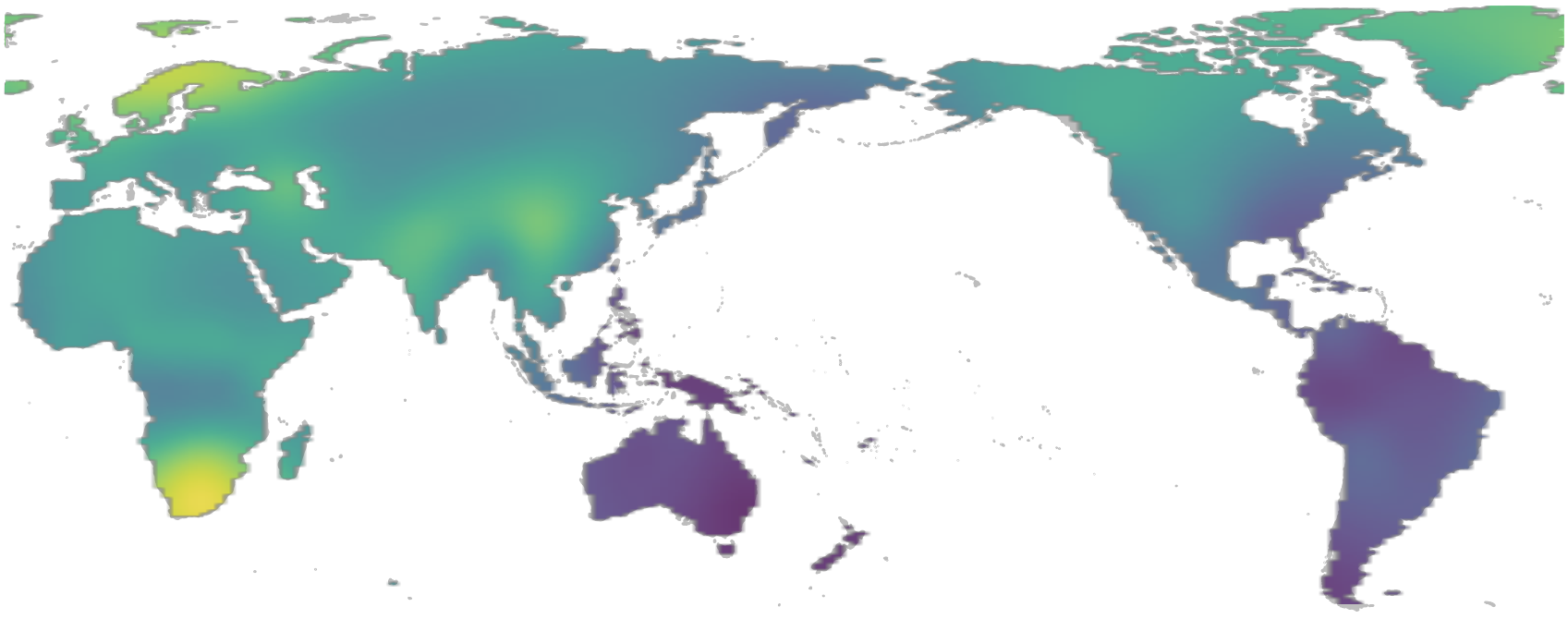

Swadesh lists
sound similarities
word alignments
cognate classes
character matrix
phylogenetic tree
training pair-Hidden Markov Model
applying pair-Hidden Markov Model
classification/ clustering
feature extraction
Bayesian phylogenetic inference

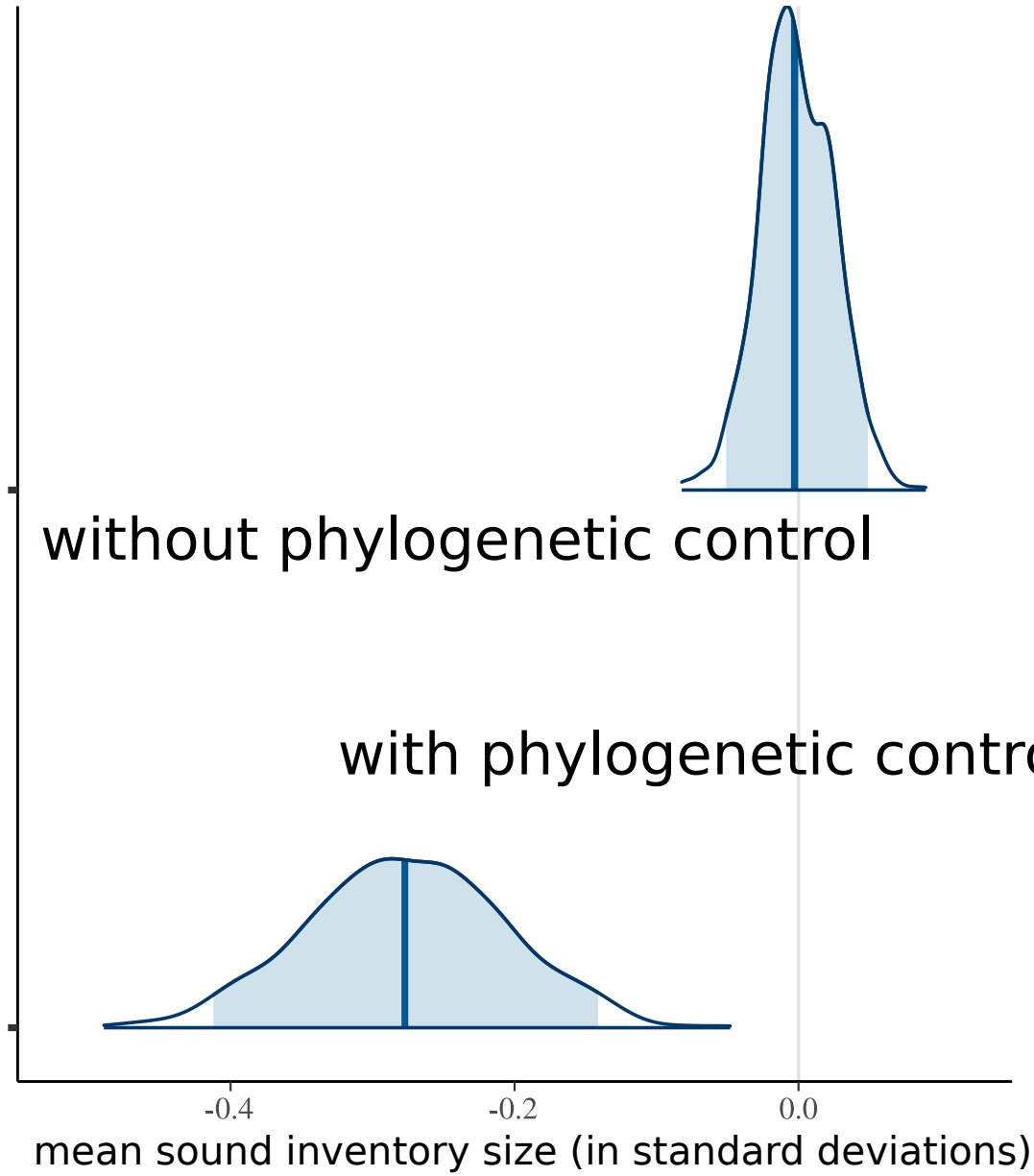

Subsaharan Africa
NW Eurasia
SE Asia
Papua
America
Niger-Congo
Khoisan
Nilo-Saharan
Dravidian
Altaic
Uralic
Indo-European
Afro-Asiatic
Australian
Australia/Papua
Torricelli
Torricelli
Trans-NewGuinea
Austronesian
Tai-Kadai
Austro-Asiatic
Timor-Alor-Pantar
Sino-Tibetan
Hmong-Mien
Quechuan
Salish
Na-Dene
Mayan
Hokan
Uto-Aztecan
Otomanguean
Tupian
Tucanoan
Carib
Macro-Ge
Ainu
Panoan
Arawakan
Otomanguean
Chibchan
Trans-NewGuinea
Trans-NewGuinea
Trans-NewGuinea

- tree estimation above established language families unreliable
  - very high phylogenetic uncertainty
  - branches tend to be much too short
- compromise used here:
  - infer trees for individual language families
  - connect them with a rake-shaped “proto-world” root with unknown depth
  - total tree depth τ is estimated from data
  - isolates are connected directly to the root



Null models for distribution of sound inventory size

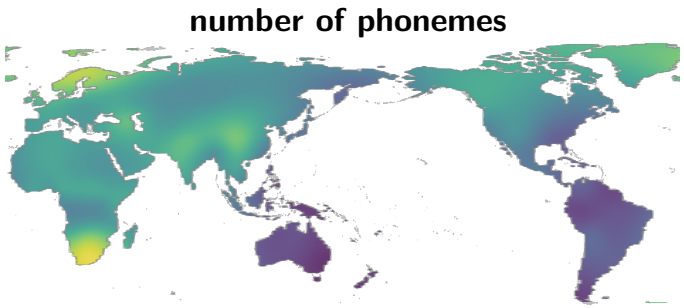


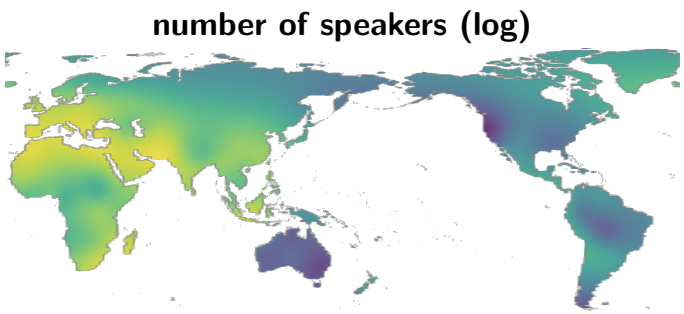


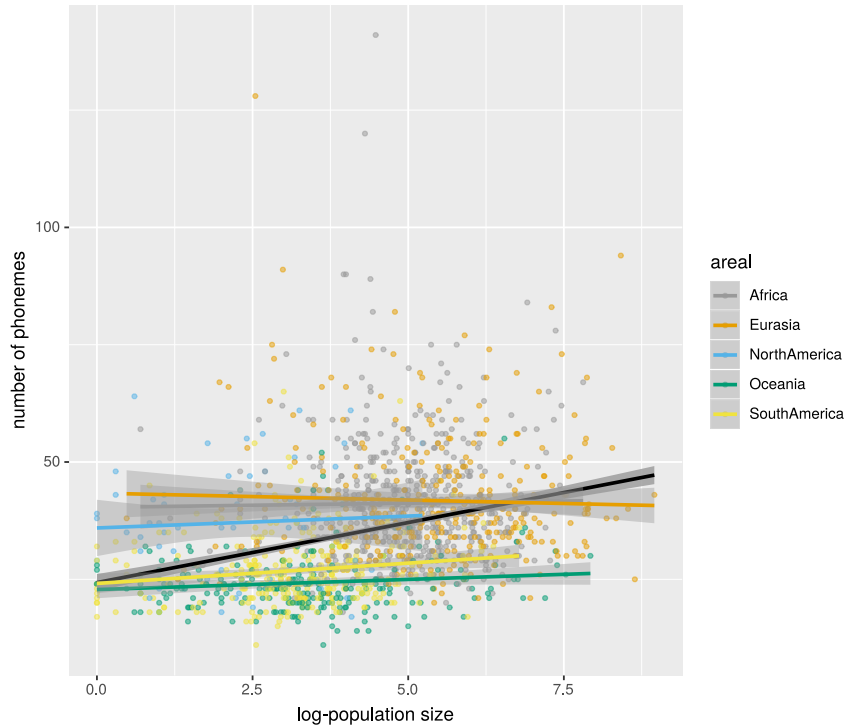
DIC: 4132

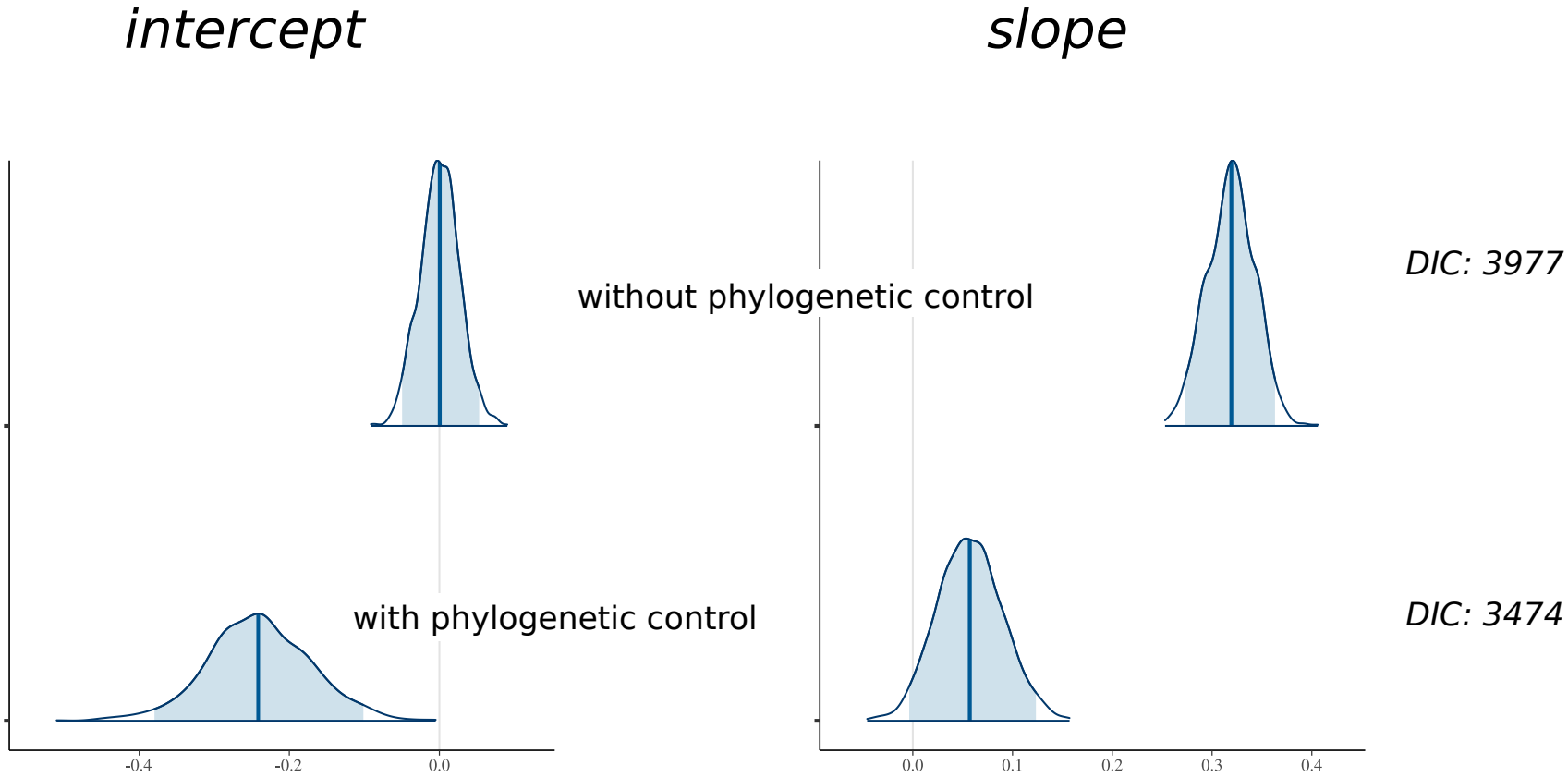
DIC: 3475

Using number of speakers as predictor


number of phonemes

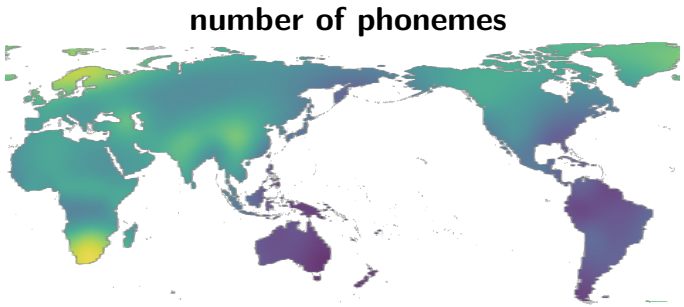

number of speakers (log)


number of phonemes
100
50
0.0
2.5
5.0
7.5
log-population size
areal
Africa
Eurasia
NorthAmerica
Oceania
SouthAmerica

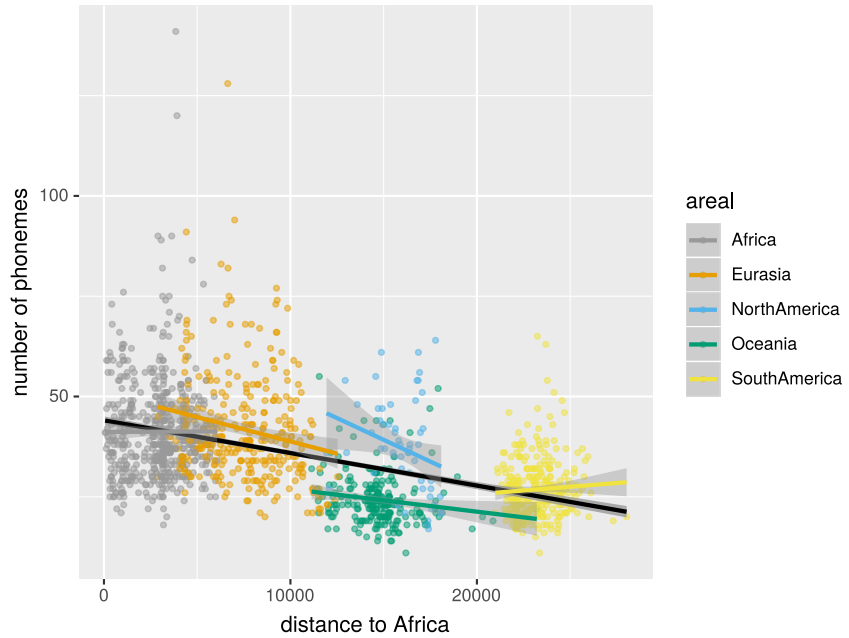


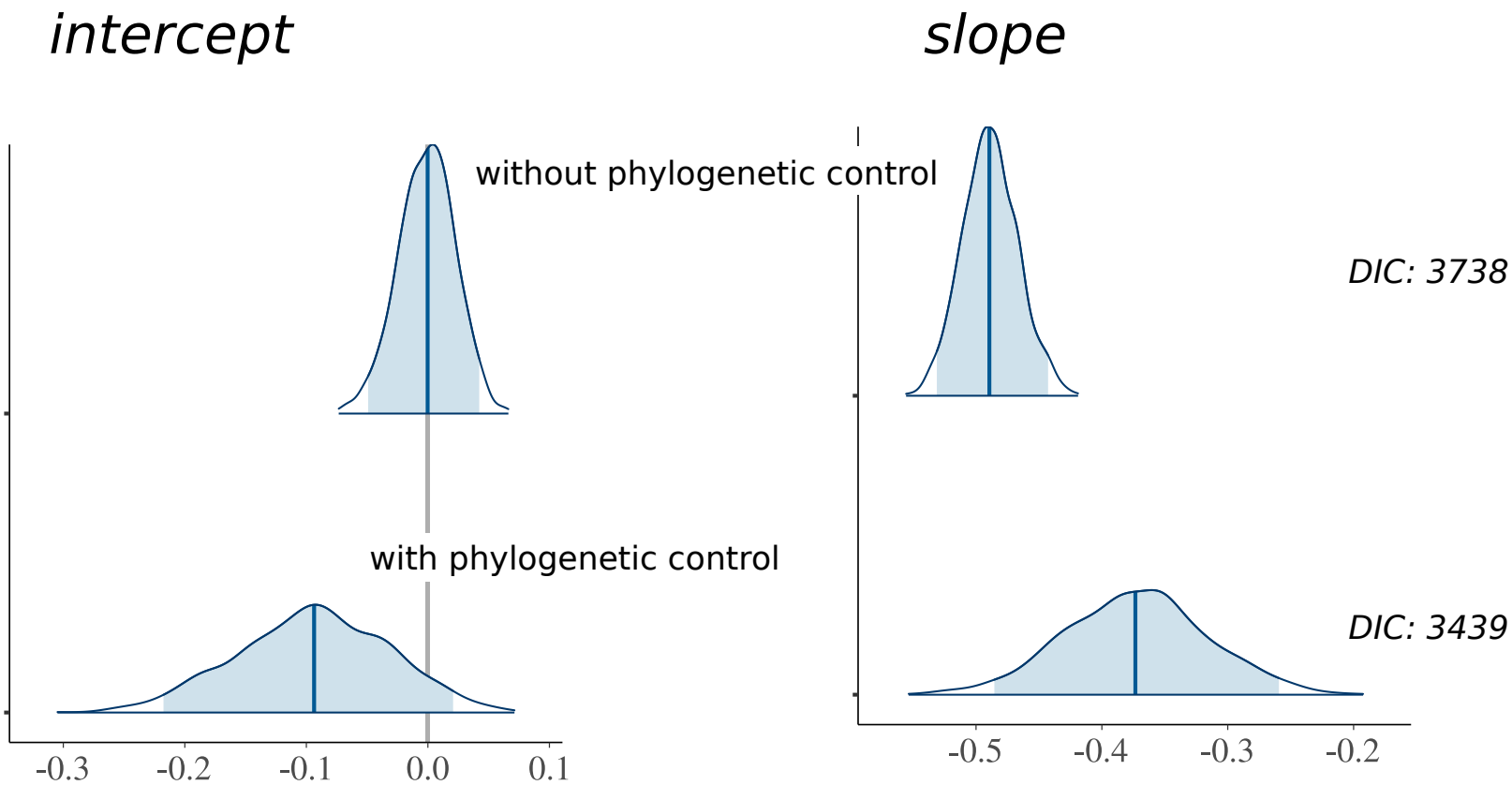
- with phylogenetic control:
  - posterior distribution for intercept includes 0 within 95% HPD (if barely)
  - model comparison shows model with and without predictor to be essentially tied

Using distance from Africa as predictor


number of phonemes

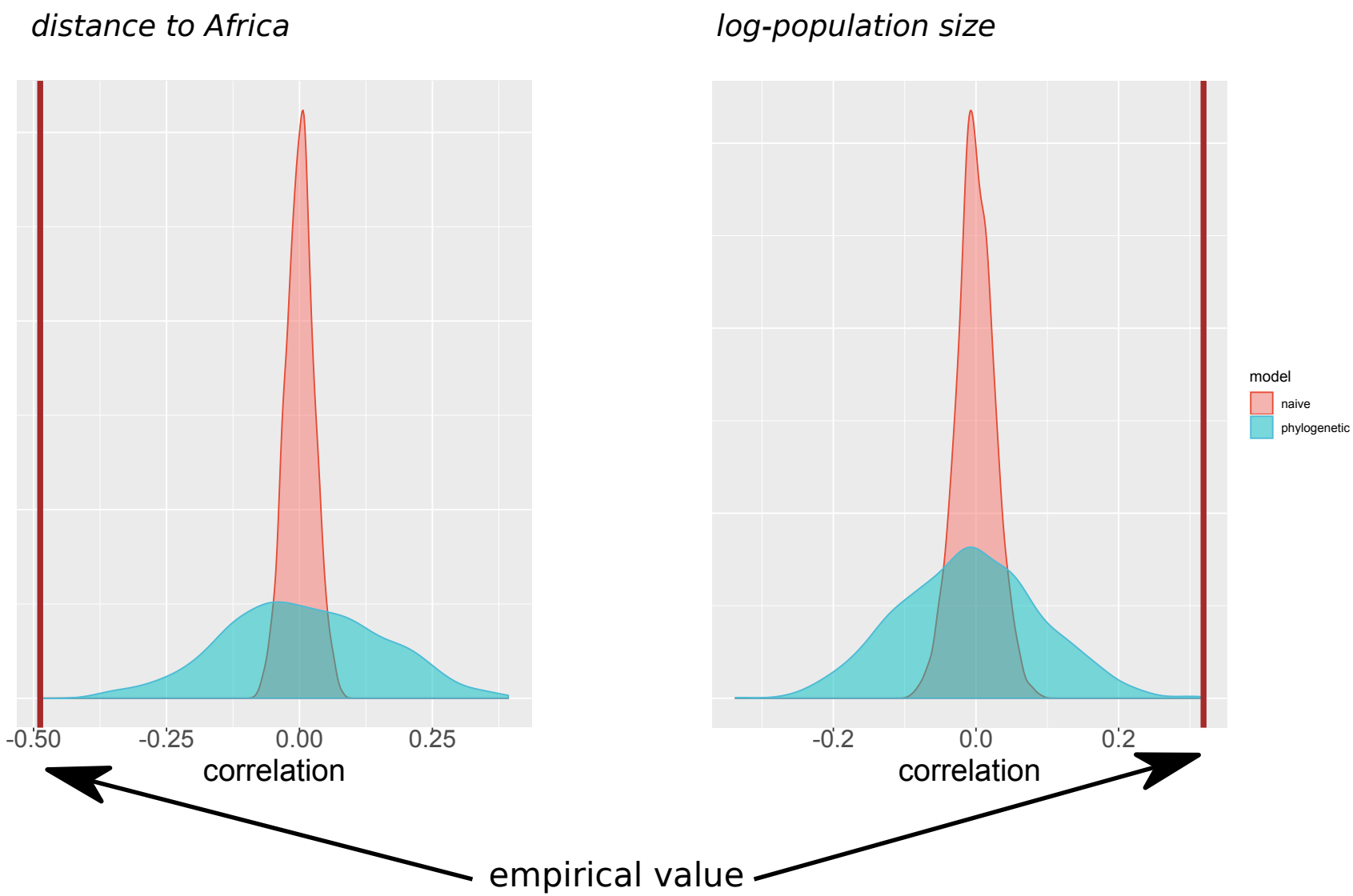

distance to Adis Abeba


number of phonemes
100
50
0
10000
20000
distance to Africa
areal
Africa
Eurasia
NorthAmerica
Oceania
SouthAmerica



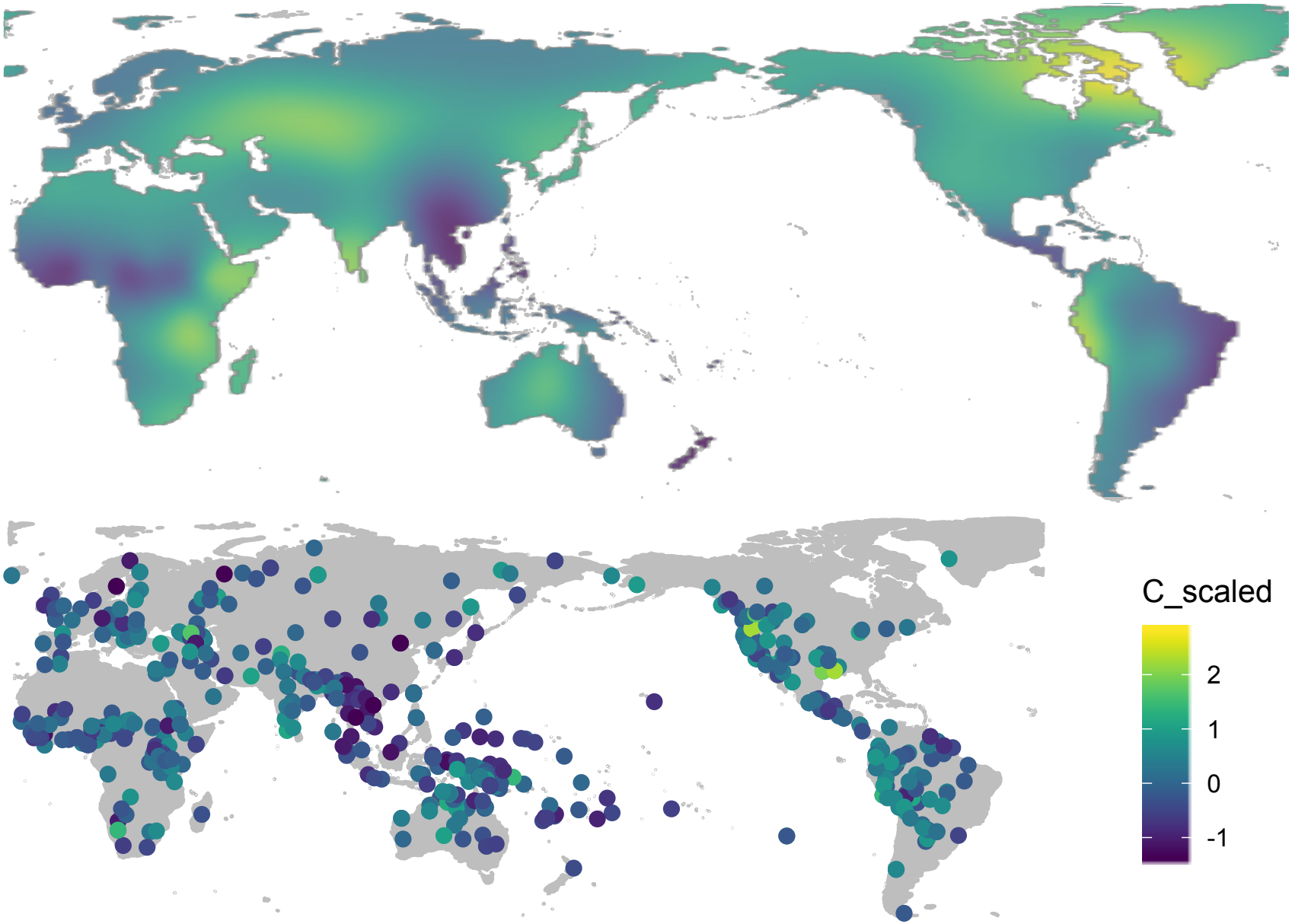
- with phylogenetic control:
  - both posterior distribution and model comparison indicate that distance to Africa is a relevant predictor for sound inventory size
  - with phylogenetic control, the effect becomes weaker, but it does not disappear

Posterior predictive check: correlations under null models

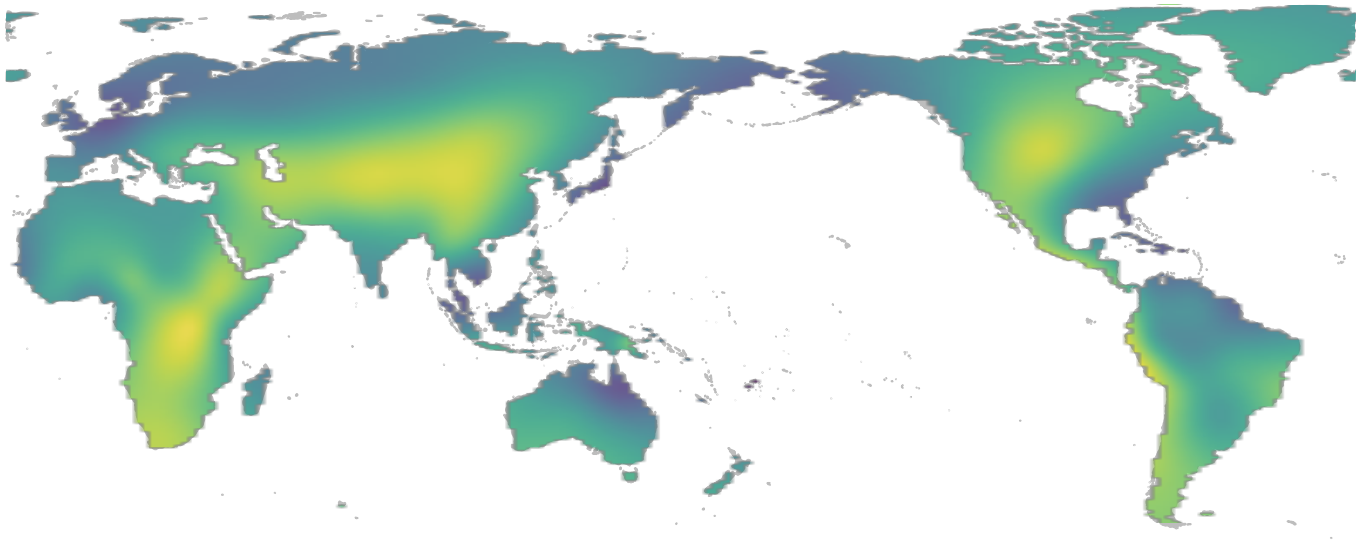


Morphological complexity

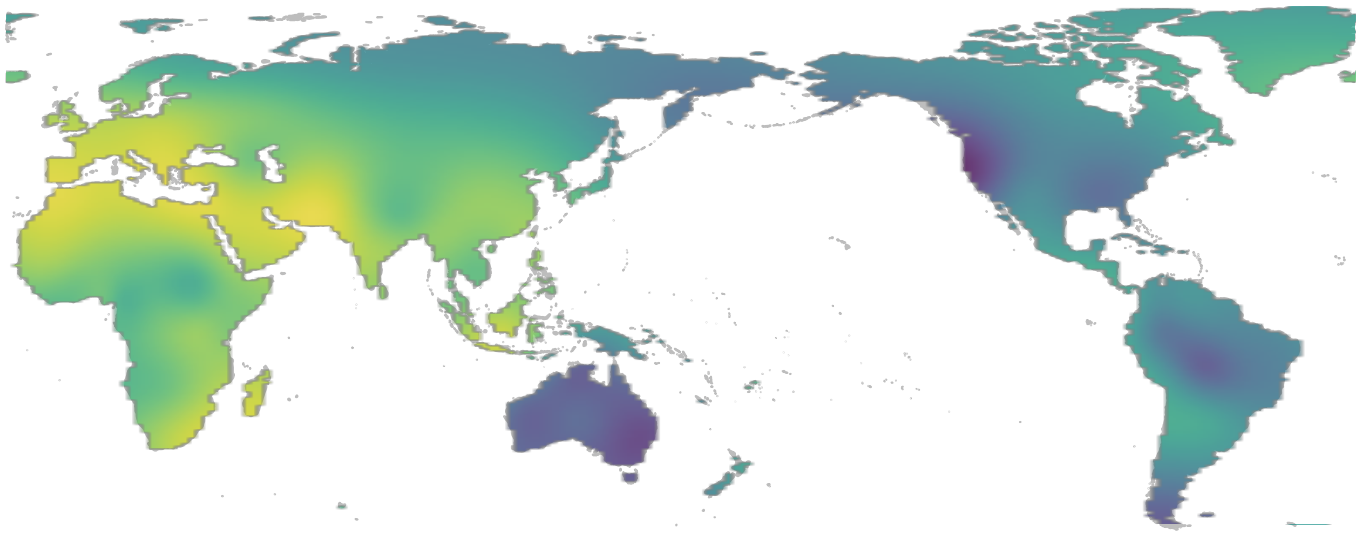
- dataset prepared by Christian Bentz and Johanna Nichols (Nichols and Bentz, 2018)
- numerical measure of morphological complexity
- aggregated from corpus unigram entropy and from number of features according to WALS (Dryer and Haspelmath, 2013) and Autotyp (Bickel et al., 2018)
- 1,060 languages in intersection with ASJP
- various hypotheses about non-linguistic factors influencing complexity, such as
    - distance to equator
    - altitude
    - population size


C_scaled
2
1
0
-1

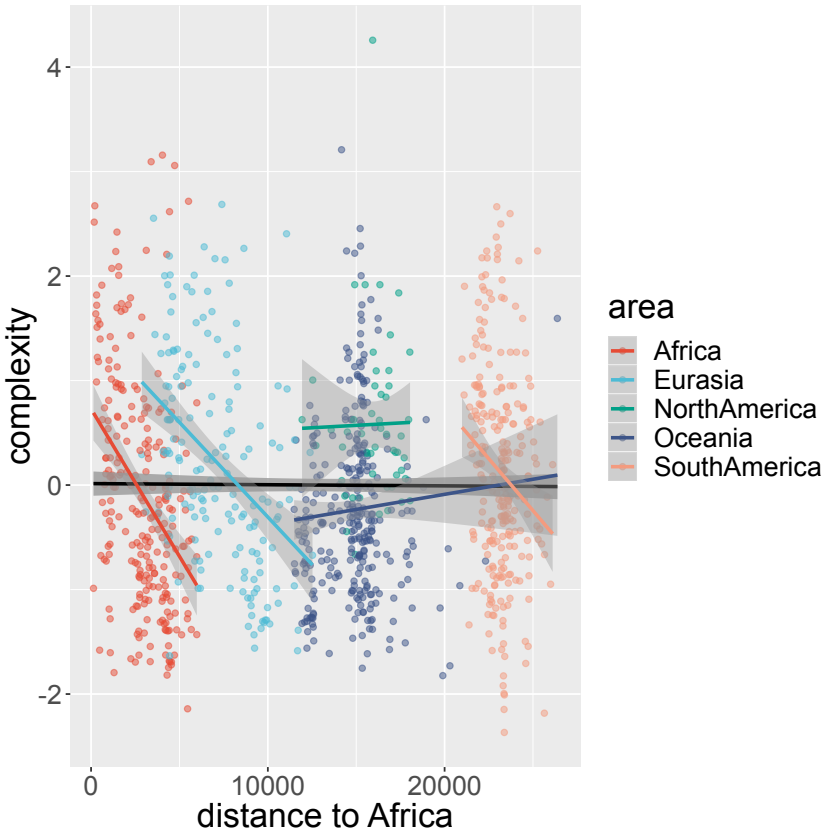
Altitude



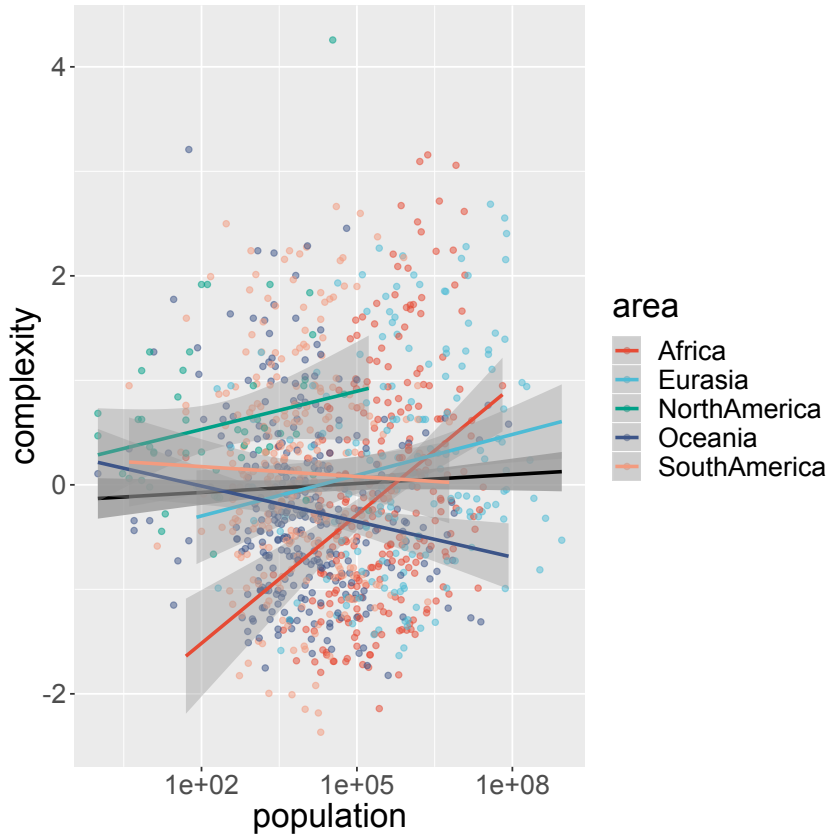
Population size



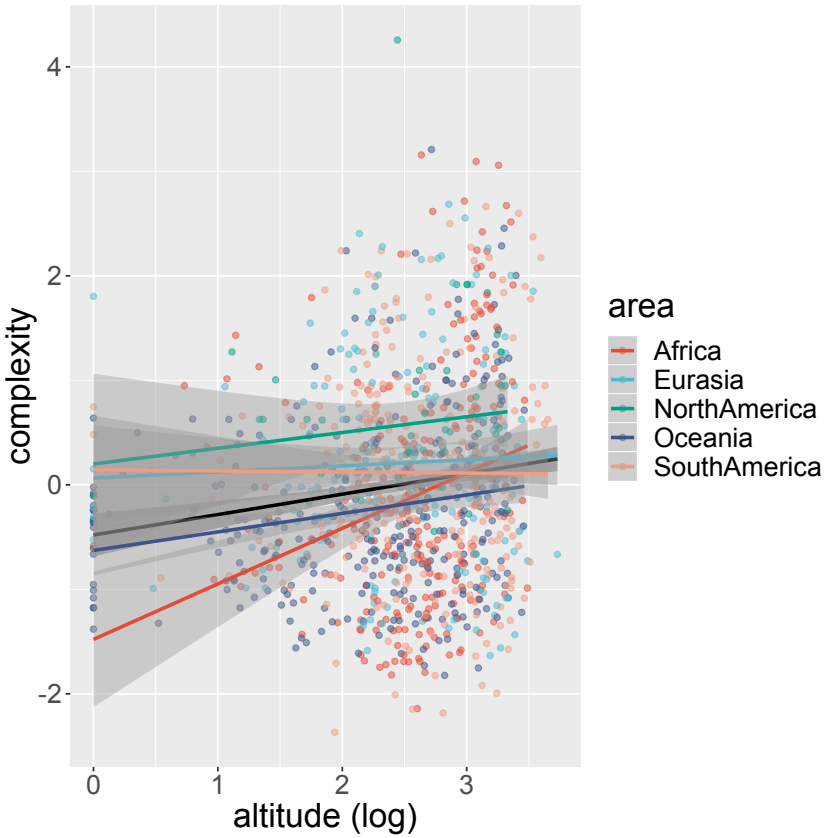
distance to Africa



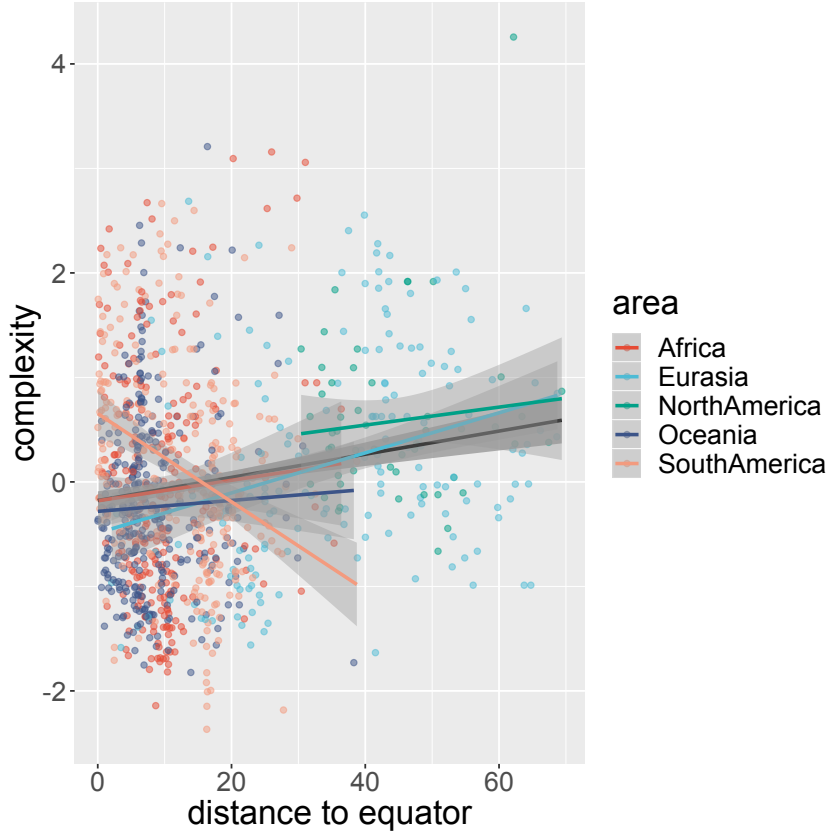
(log-)population size



(log-) altitude



distance to equator



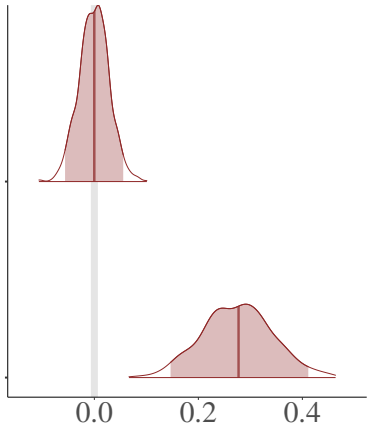
- I fitted two Bayesian regression models

$$\begin{aligned}\text{complexity} \sim \mathcal{N}(&\text{intercept} + \text{distance to Africa} + \\ &\log(\text{population size}) + \\ &\log(\text{altitude}) + \\ &\text{distance from equator}, \Sigma)\end{aligned}$$

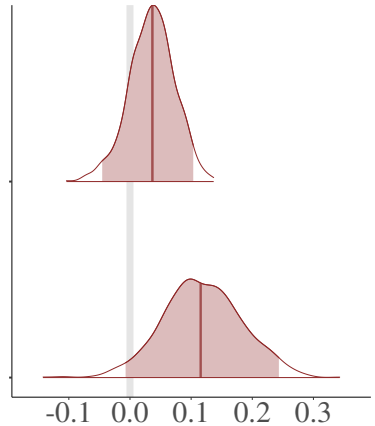
- model 1: Σ is diagonal matrix
- model 2: Σ is derived from phylogenetic trees

posterior estimates of regression parameters

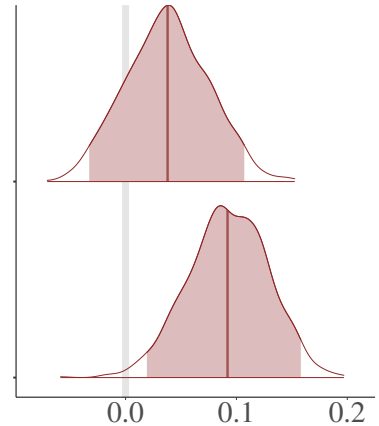
intercept



distance to Africa



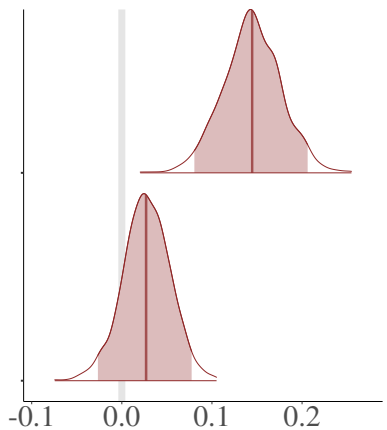
population size



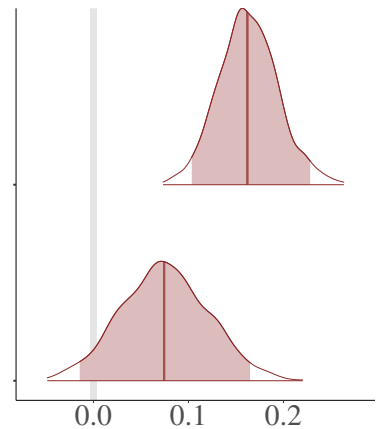
naive model (DIC: 2966)

phylogenetic model (DIC: 2494)

altitude



distance from equator



Conclusion

- controlling for common descent is essential when comparing populations on a global scale
- language trees, as derived via phylogenetic inference, are a suitable starting point for assessing autocorrelation
- resulting model arguably still underestimates autocorrelation since it ignores horizontal transmission
- first experiments with geographic spatial autocorrelation did not alter qualitative picture though

special thanks to Christian Fröhlich

References

Quentin D. Atkinson. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332(6027):346–349, 2011.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. The AUTOTYP database, release 0.1. https://github.com/autotyp/autotyp-data, 2018.

Matthew S. Dryer and Martin Haspelmath, editors. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, 2013. http://wals.info/.

Steven Moran and Daniel McCloy. PHOIBLE 2.0. Available online at http://phoible.org, Accessed on 2019-06-15., 2019.

Steven Moran, Daniel McCloy, and Richard Wright. Revisiting the population vs phoneme-inventory correlation. *Language*, 88(4):877–893, 2012.

Johanna Nichols and Christian Bentz. Morphological complexity of languages reflects the settlement history of the Americas. In Katerina Harvati, Gerhard Jäger, and Hugo Reyes-Centeno, editors, *New Perspectives on the Peopling of the Americas*, pages 13–26. Kerns Verlag, Tübingen, 2018.

Søren Wichmann, Taraka Rama, and Eric W. Holman. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, 15(2):177–197, 2011.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 18). http://asjp.clld.org/, 2018.