Phylogenetic Signals of Language "External" Factors

Christian Bentz
University of Tübingen

September 14, 2016

Words Bones Genes Tools
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
Language Evolution: The Empirical Turn

Overview

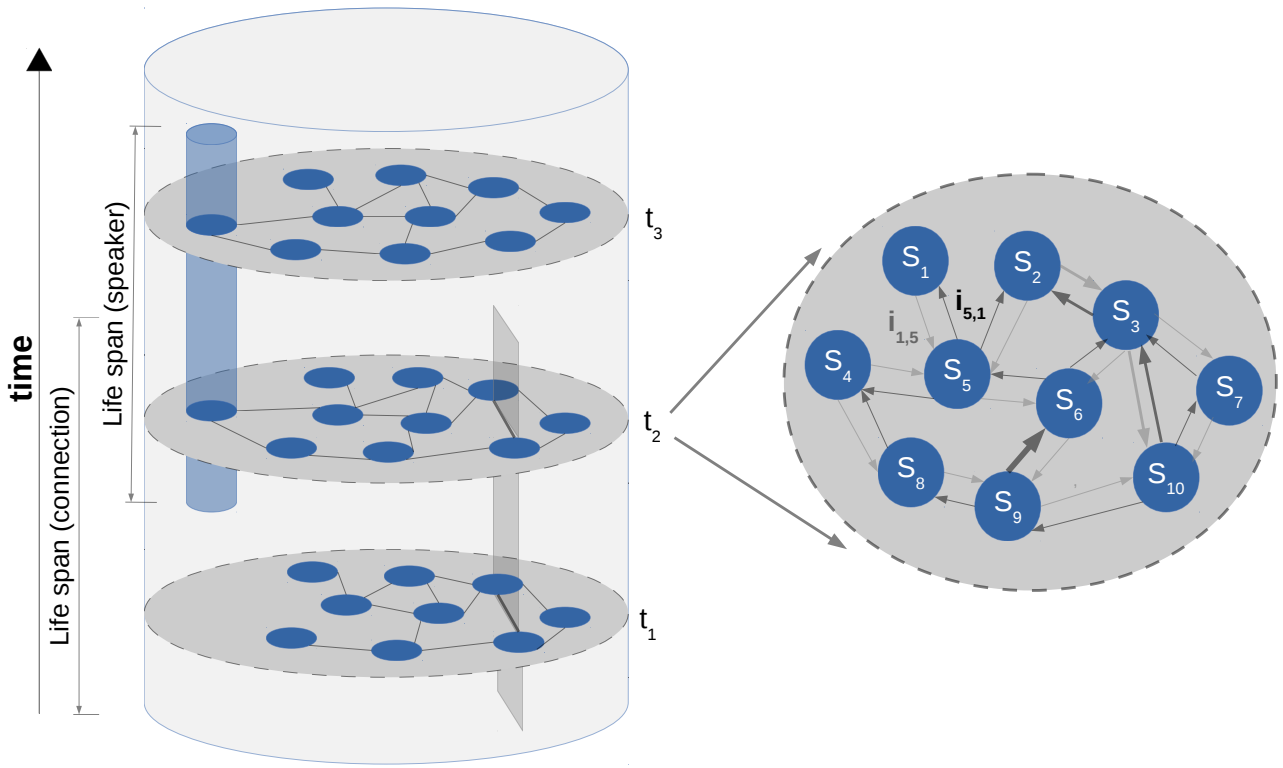
Introduction

Two definitions of language(s)

Languages are accumulations of linguistic interactions between populations of speakers, which are constrained/subdivided by mutual intelligibility (externalized language, E-language)

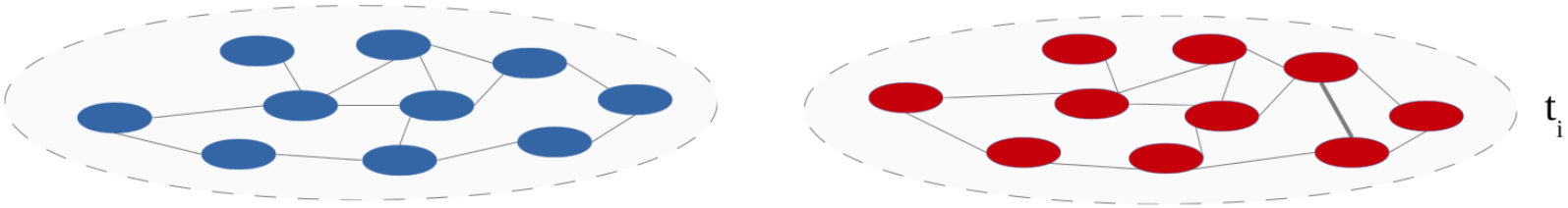
Language is an abstraction over these accumulations, and defined by properties common to all of them (internal language, I-language)

A language from the empirical/quantitative/usage-based perspective



Population: $\mathcal{P} = \{S_1, S_2, ..., S_n\}$
Language: $\mathcal{L} = \{i_{1,2}, i_{2,1}, ..., i_{n-1,n}\}$

Language Comparison

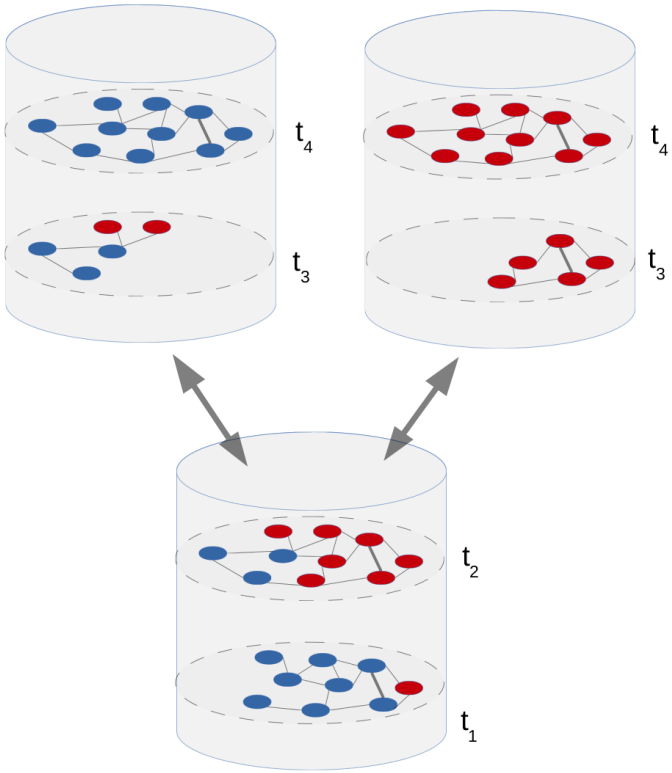


$\mathcal{L}^{A}_{ti} = \{l_{1\to2}, l_{2\to1}, l_{2\to3}, \ldots, l_{n\to n-1}\}$ $\mathcal{L}^{B}_{ti} = \{l_{1\to2}, l_{2\to1}, l_{2\to3}, \ldots, l_{n\to n-1}\}$

How do we approximate languages?

- lexical lists
- structural features (from databases like WALS, AUTOTYP)
- directly by language production, e.g. corpora, recordings

Two Basic Processes: Isolation or Contact, Diversification or Unification

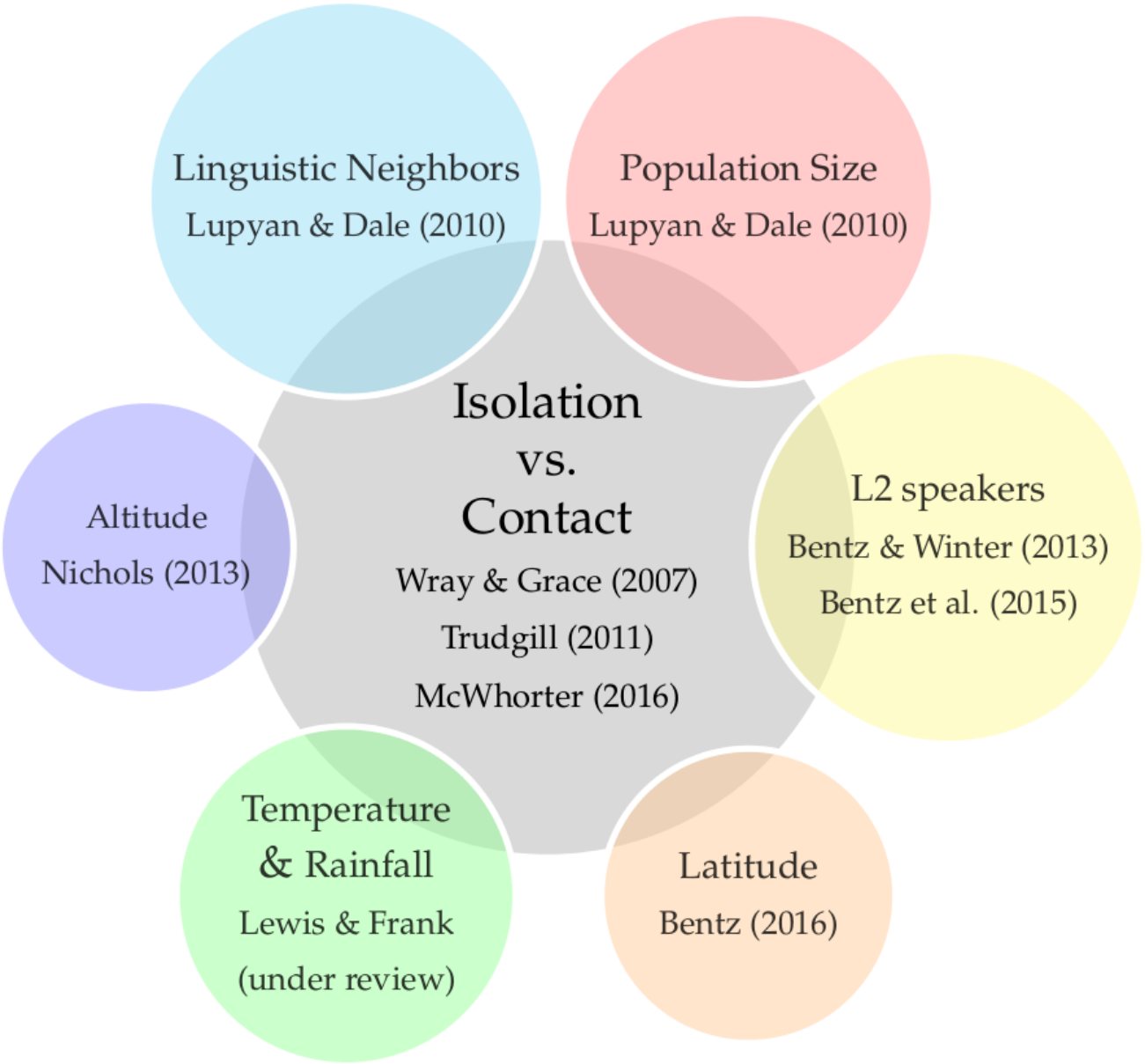


- Geography: Latitude, Longitude, Altitude
- Socioeconomic: Trade, Cultural Exchange
- Language Learning: Native, Non-Native

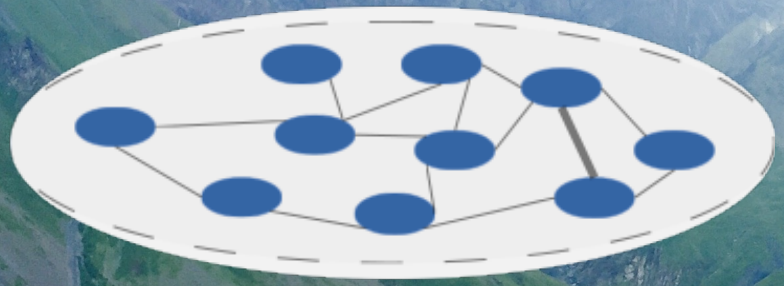
Recent studies on external factors shaping languages

- **Population spread** and lexical change [Bouckaert et al., 2012; Jaeger, 2015]
- **Population size** and morphology/phoneme inventories/lexical change [Lupyan & Dale, 2010; Atkinson 2011, Bromham et al., 2015]
- **Climate** and Tones/Ejectives [Everett & , 2013; Everett, Blasi & Roberts 2015]
- **Altitude** and morphological complexity [Bickel & Nichols, 2003; Bickel & Nichols 2011; Nichols, 2013]
- **Latitude** and word entropy [Bentz, 2016]

"External" Factors

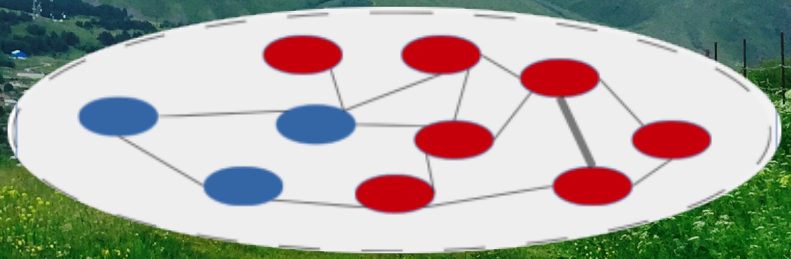


Example: High Altitude Complexity



Asymmetry in language contact, i.e. adult learning, leads to simplification of morphology in lowland languages, but not in highland languages

(Nichols, 2013; Bickel & Nichols 2003, 2011)



Questions

- How much can factors of isolation/contact tell us about genealogical clustering?
- How much do they tell us about past language change, and the time-depth of external factors?

→ **Phylogenetic Signal Analysis**

First step to (potential) solution

Phylogenetic signal analysis

- focus on language **external** features

Data and Methods

Phylogenetic Signal

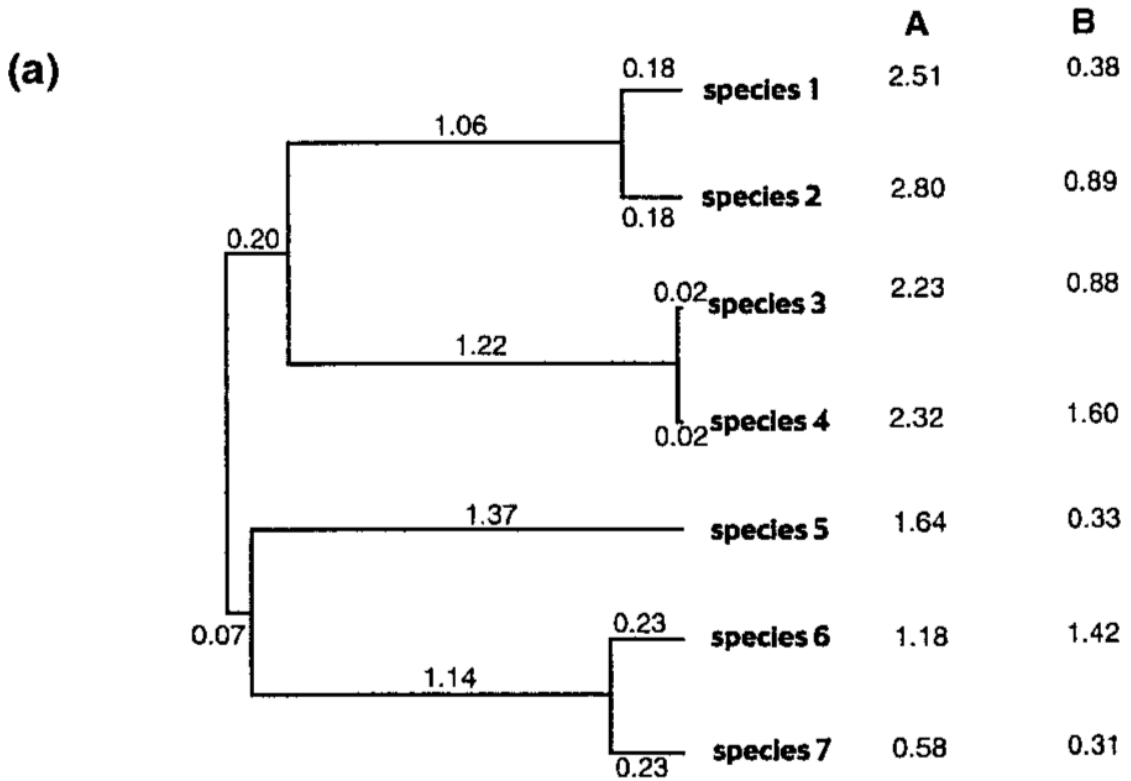
"Phylogenetic signal is the extent to which trait values are statistically related to phylogeny. In other words, phylogenetic signal indicates the extent to which related species [here languages] tend to resemble each other."
[Symonds & Blomberg, 2014]

Data and Methods

Phylogenetic signal indicators

- Pagel's λ: $[0, 1]$ [Pagel, 1999; Freckleton, Harvey & Pagel 2002]
- Blomberg's K: $[0, \infty]$ [Blomberg, Garland & Ives, 2003]
- others such as Moran's I or Abouheif's C, etc.

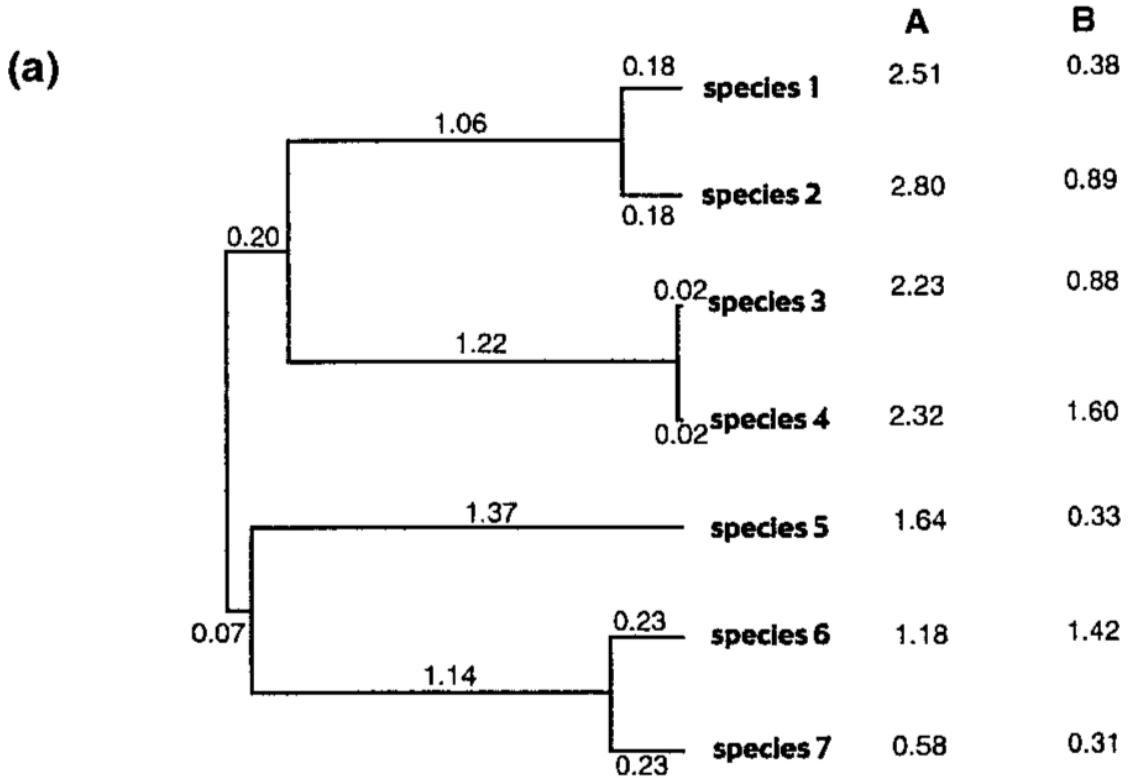
Pagel's λ



[Pagel, 1999; Freckleton, Harvey & Pagel 2002]

Phylogenetic Signal Analysis

$$\lambda^A \to 1$$
$$\lambda^B \to 0$$



Summary

Meta-study has shown that

- only λ and K can be used to compare phylogenetic signals across different trees
- at least 20 species, i.e. languages, are needed to give robust estimates
- λ is more robust under a Brownian motion model (constant rates of change), and to changes in number of species
- K is more robust under a non-BM model (i.e. OU model)

[Münkemüller et al. 2012]

Data

We need

- linguistic trees for different families with > 20 languages
- information on external factors (population size, latitude, longitude, altitude)

Dediu's Forest [Dediu 2015, online at github]

linguistic trees built on topologies from

Ethnologue [Lewis et al., 2014]
WALS [Dryer & Haspelmath, 2013]
AUTOTYP [Nichols, Witzlack-Makarevich & Bickel, 2013]
Glottolog [Hammarström, Forkel, Haspelmath & Nordhoff, 2014]

and branch length information from

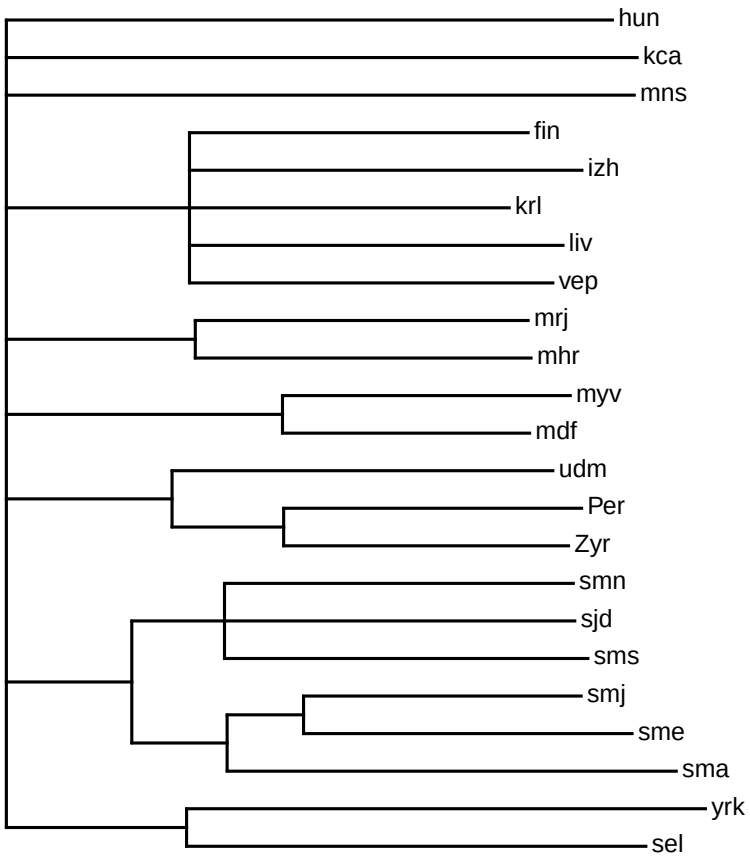
Vocabulary, i.e. Swadesh lists from ASJP16 [Wichmann et al., 2013]
Geography, i.e. great-circle distance
WALS, i.e. distance of languages in terms of structural features
AUTOTYP, also distance in structural features
The *Tree Topology* itself

My Sub-Forest

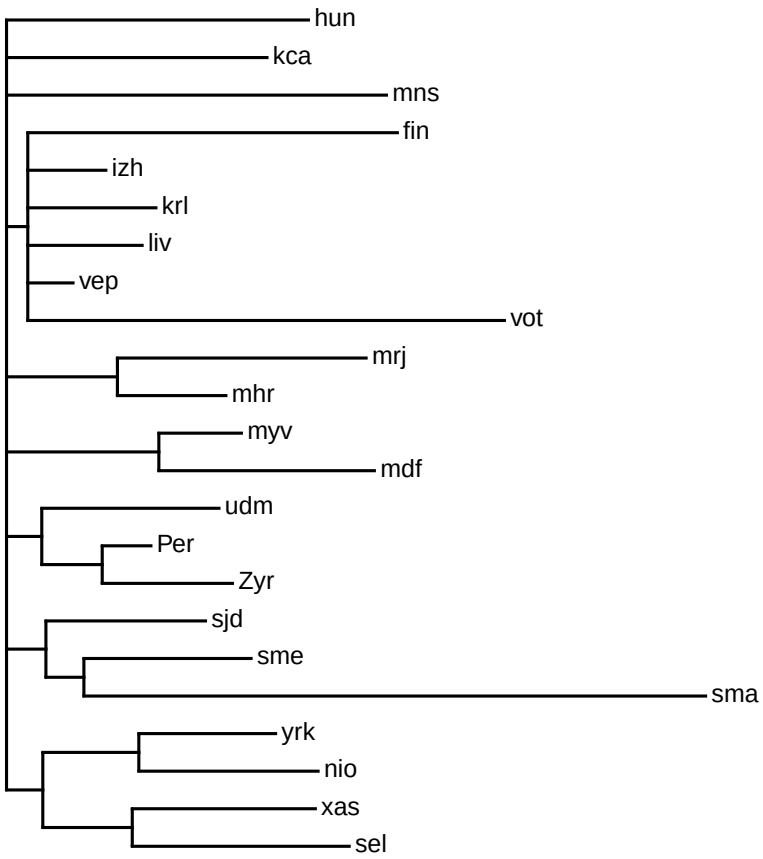
- Topologies: all from ***Ethnologue***, ***WALS***, ***AUTOTYP***, and ***Glottolog***
- Branch length information: ***ASJP***, ***WALS(euclidean)***, ***AUTOTYP***, Geography, Tree Topology
- Branch length method: *ga* (genetic algorithm)

Example trees for Uralic

Ethno, ASJP, ga



Ethno, WALS, ga



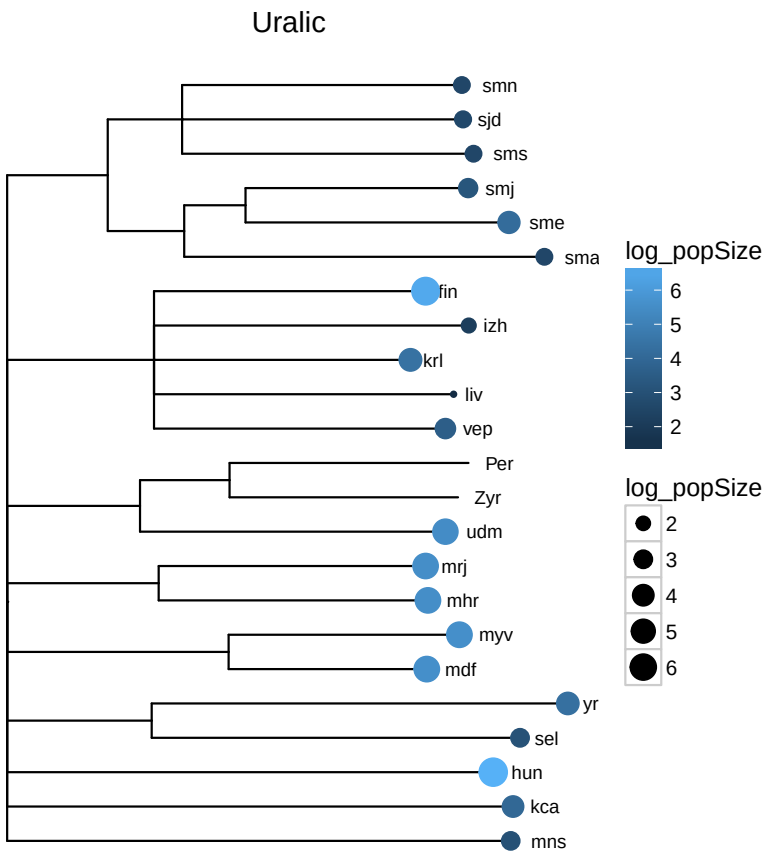
External Factors

6834 languages (unique ISO codes), 376 families

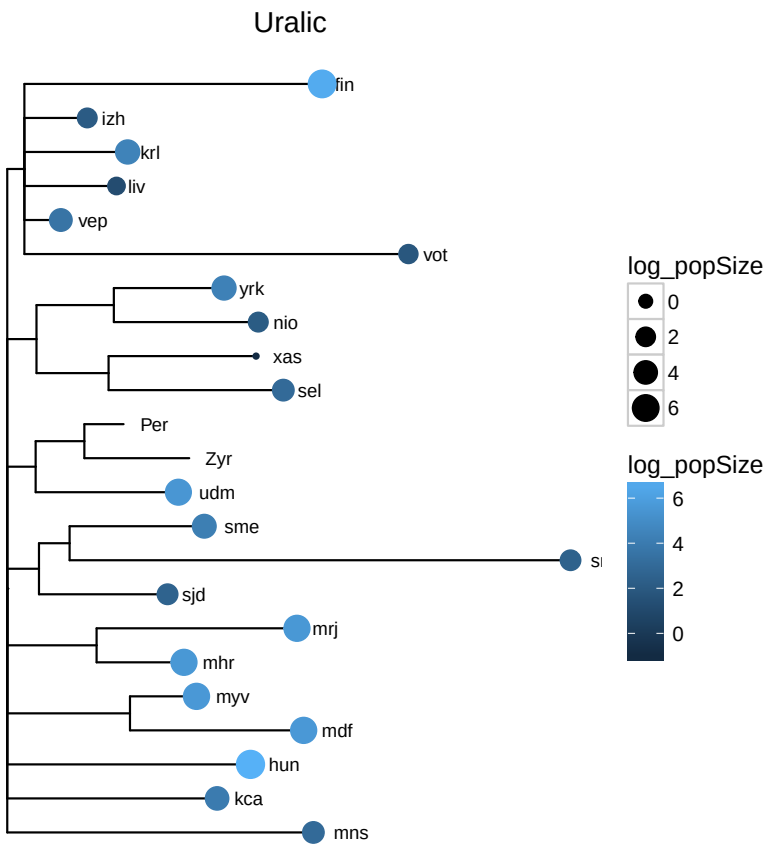
iso	name_glotto	name_ethno	id_glotto	family	latitude	longitude	altitude	log_popSize	popSize
aaa	Ghotuo	Ghotuo	ghot1243	Atlantic-Congo	7.11551	5.95663	248.3353271484	3.9542425094	9000
aab	Alumu-Tesu	Alumu-Tesu	alum1246	Atlantic-Congo	9.01513	8.55956	433.3730163574	3.84509804	7000
aac	Ari	Ari	arii1243	Suki-Gogodala	-7.95526	142.40045	17.7223339081	1.6989700043	50
aad	Amal	Amal	amal1242	Sepik	-4.04517	141.9952	52.0980796814	2.9190780924	830
aae	ArbÃ□Â«reshÃ□Â	Albanian, Arb\"e	arbe1236	Indo-European	38.91104	16.71645	80.7221755981	5	100000
aaf	Aranadan	Aranadan	aran1261	Dravidian	11.35252	75.79538	34.4588470459	2.3010299957	200
aag	Ambrak	Ambrak	ambr1239	Nuclear_Torricelli	-3.51156	142.46152	272.3916931152	2.4623979979	290
aai	Arifama-Miniafia	Arifama-Miniafia	arif1239	Austronesian	-9.15565	149.24638	136.9517974854	3.5403294748	3470
aak	Ankave	Ankave	anka1246	Angan	-7.19444	145.75421	1898.9189453125	3.2041199827	1600
aal	Afade	Afade	afad1236	Afro-Asiatic	12.055125	14.63426	292.415222168	4.4913616938	31000
aan	AnambÃ□Â©	Anamb\'e	anam124	Tupian	-2.71119	-49.30296	37.9944000244	0.7781512504	6
aao	Algerian Saharan	Arabic, Algerian	alge1240	Afro-Asiatic	20.8884	4.80626	564.9817504883	5	100000
aap	ParÃ□Â¡ ArÃ□Â¡r	Ar\'ara, Par\'a	para1310	Cariban	-3.71263	-53.06572	210.4430847168	2.531478917	340
aaq	Eastern Abenaki	Abenaki, Easter	east2544	Algic	45.01121	-68.66167	38.4853134155	0	1
aar	Afar	Afar	afar1241	Afro-Asiatic	12.2281066667	41.8082933333	322.2823791504	6.1072099696	1280000
aas	Aasax	Aas\'ax	aasa1238	Afro-Asiatic	-4.00679	36.86477	1300.5052490234	2.5440680444	350
aat	Arvanitika Alban	Albanian, Arvani	arva1236	Indo-European	38.28299	23.37034	317.3229675293	4.6989700043	50000
aau	Abau	Abau	abau1245	Sepik	-3.97222	141.32359	69.2794952393	3.8615344109	7270
aaw	Solong	Solong	solo1258	Austronesian	-5.86129	148.82509	94.4773254395	3.3424226808	2200
aax	Mandobo Atas	Mandobo Atas	mand144	Nuclear_Trans_N	-5.69357	140.62376	77.6934432983	4	10000
aaz	Amarasi	Amarasi	amar1273	Austronesian	-10.21751	123.96373	188.8751373291	4.84509804	70000
aba	AbÃ□Â©	Ab\'e	abee1242	Atlantic-Congo	5.59682	-4.38497	57.1392402649	5.2304489214	170000
abb	Bankon	Bankon	bank1256	Atlantic-Congo	4.3702	9.6403	85.9485244751	4.079181246	12000
abc	Ambala Ayta	Ayta, Ambala	amba126	Austronesian	14.81558	120.28339	106.2621536255	3.220108088	1660
abd	Camarines Norte	Manide	cama125	Austronesian	14.15712	122.83497	19.7265014648	3.5797835966	3800
abe	Western Abenaki	Abenaki, Wester	west2630	Algic	46.13812	-72.65106	34.6359024048	1	10
abf	Abai Sungai	Abai Sungai	abai1240	Austronesian	5.55394	118.30626	7.863702774	-1	0

Example trees for Uralic

Ethno, ASJP, ga



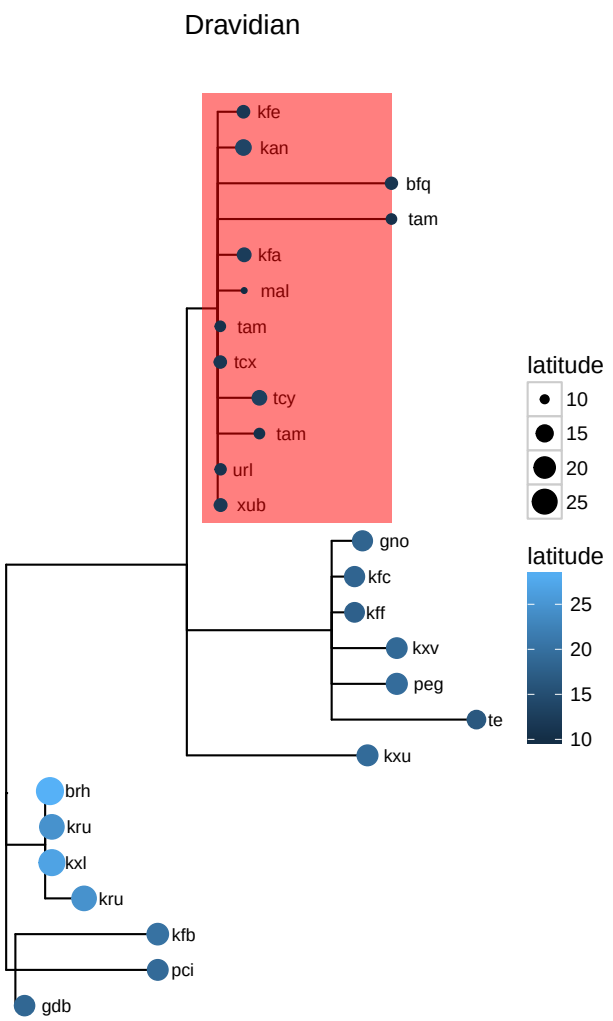
Ethno, WALS, ga



Latitude λ signal

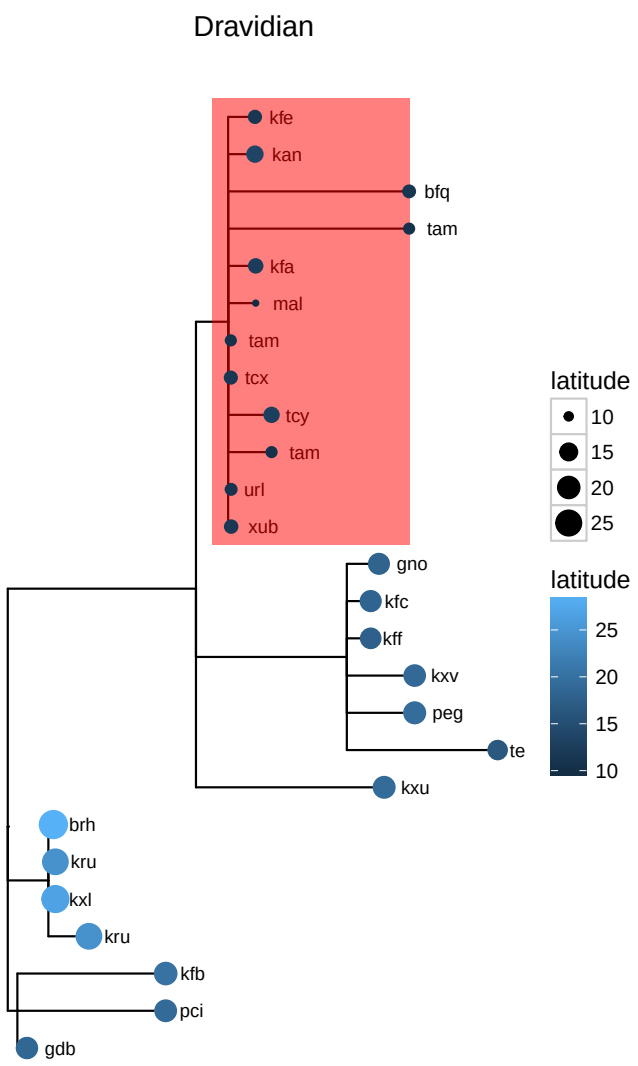
family	lang.	latitude	p-value	sig.	branch source	tree source
Dravidian	23	0.99	< 0.001	λ	autotyp	autotyp
Dravidian	23	0.99	< 0.001	λ	autotyp	ethnologue
Otomanguean	50	0.99	< 0.001	λ	wals(euclidean)	ethnologue
Dravidian	21	0.99	< 0.001	λ	wals(euclidean)	ethnologue
Tupian	37	0.99	0.024	λ	asjp16	glottolog
Dravidian	22	0.99	0.006	λ	wals(euclidean)	wals
Otomanguean	56	0.99	< 0.001	λ	autotyp	ethnologue
				...		
Tupian	21	0	1	λ	wals(euclidean)	wals
Algic	28	0	1	λ	wals(euclidean)	wals
Algic	26	0	1	λ	wals(euclidean)	autotyp
Algic	28	0	1	λ	autotyp	ethnologue
Algic	24	0	1	λ	asjp16	wals
Algic	23	0	1	λ	asjp16	autotyp
Uralic	23	0	1	λ	asjp16	autotyp

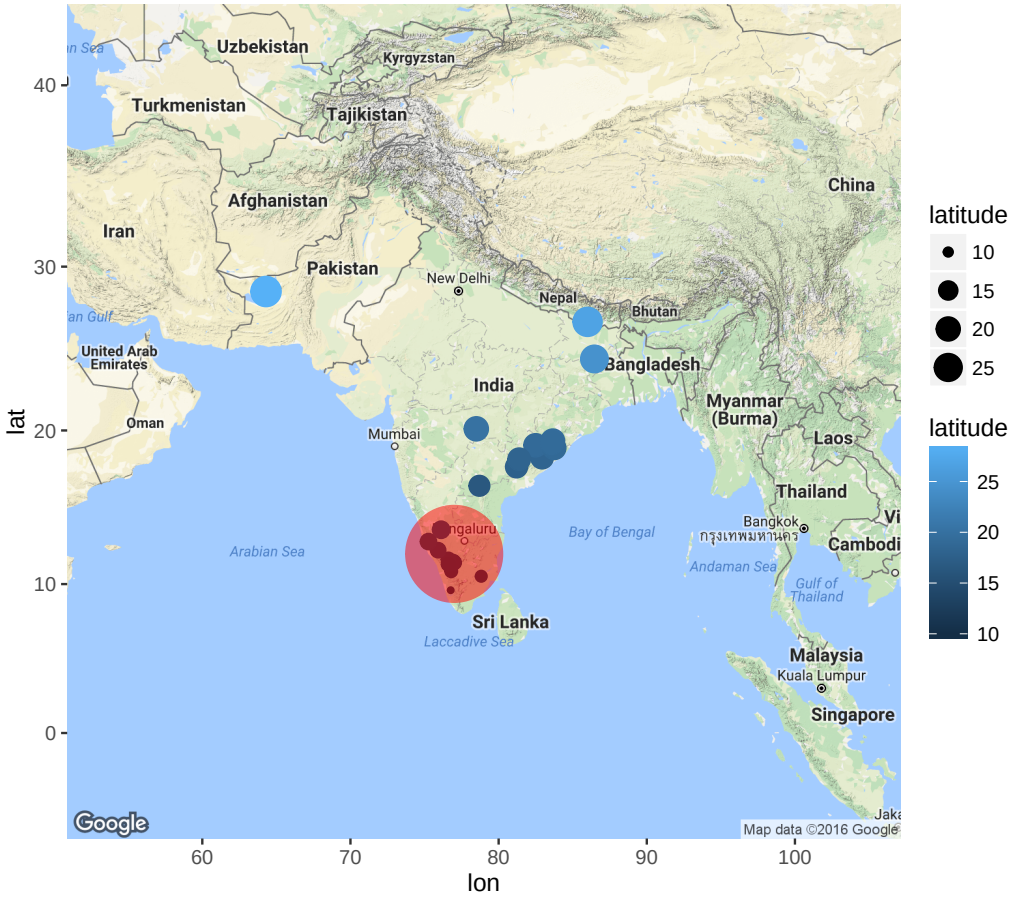
Dravidian, Latitude Signal (branch lengths: Autotyp, tree topology: Autotyp)



Dravidian, Latitude Signal: 0.99

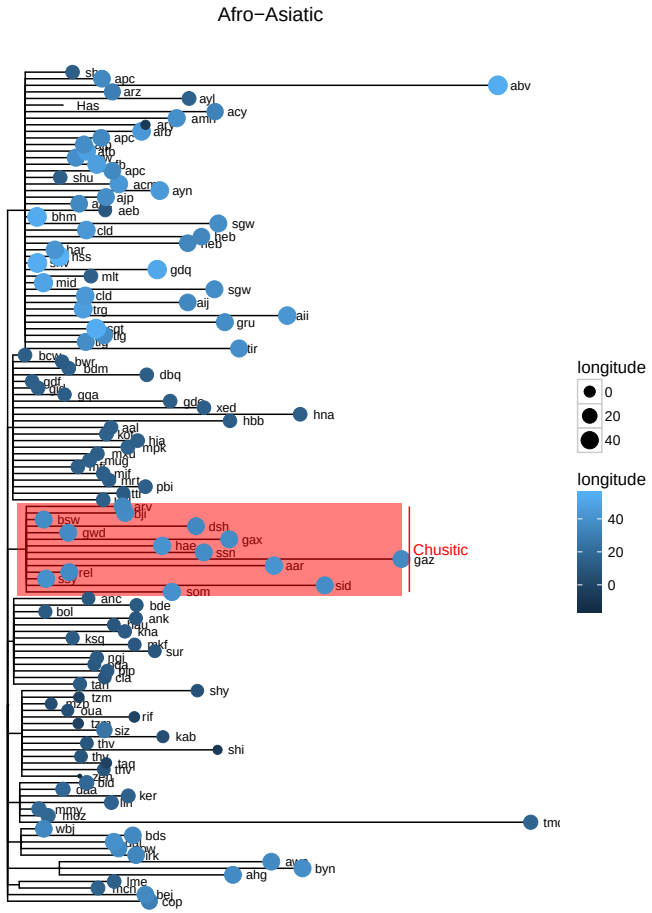
(branch lengths: Autotyp, tree topology: Autotyp)

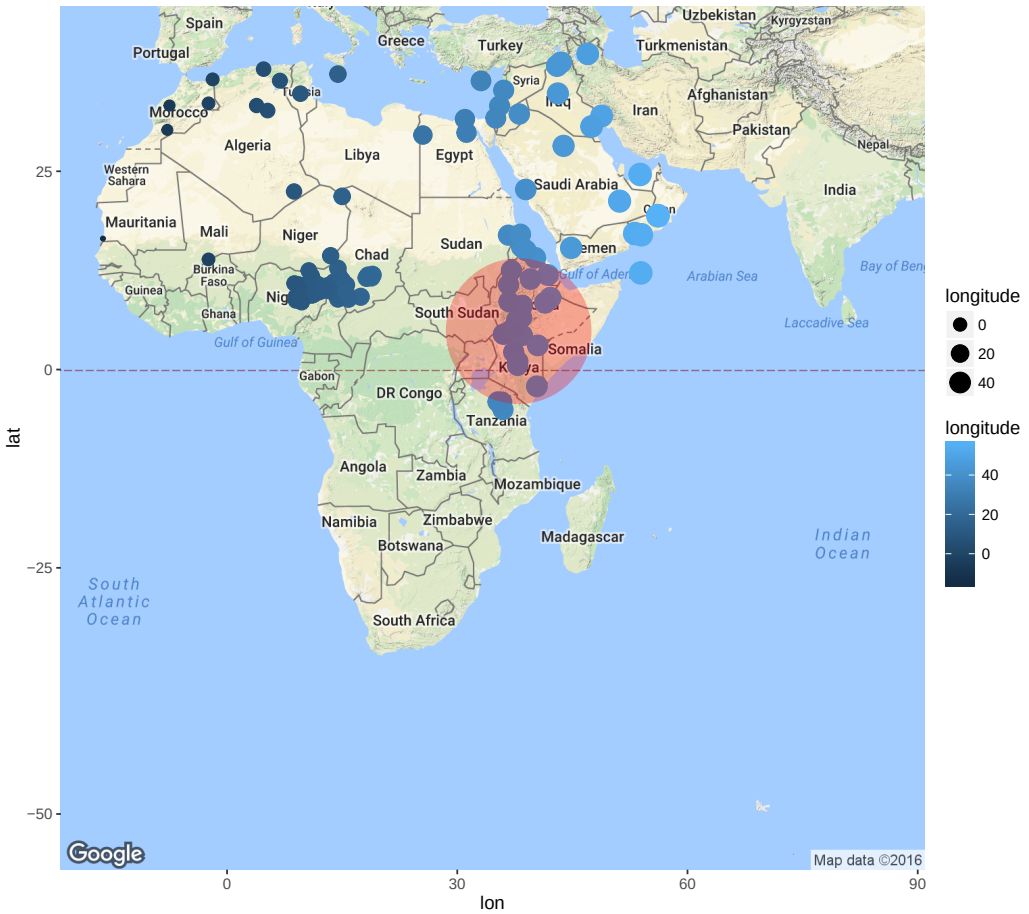




Afro-Asiatic, Longitude Signal: 0.99

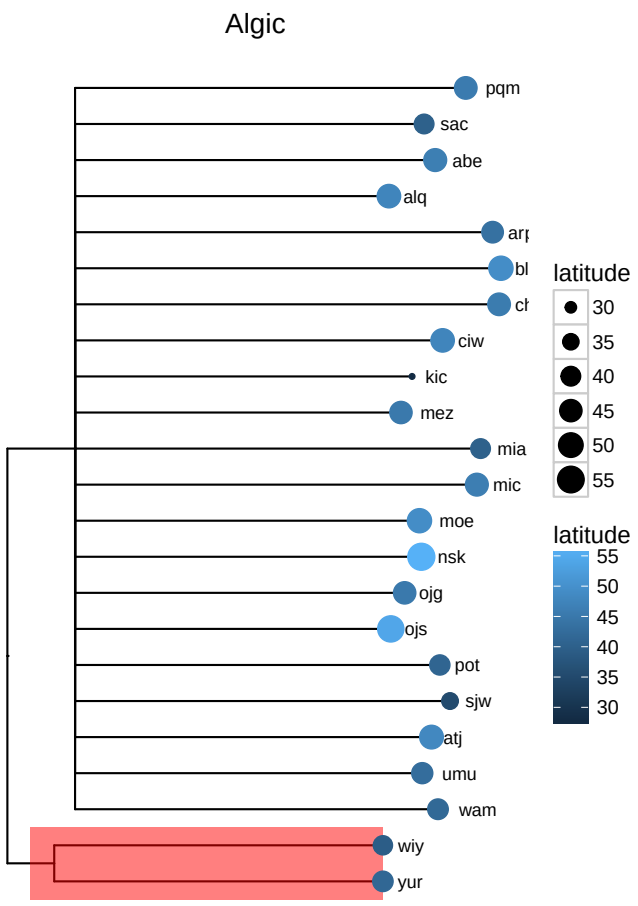
(branch lengths: Autotyp, tree topology: WALS)

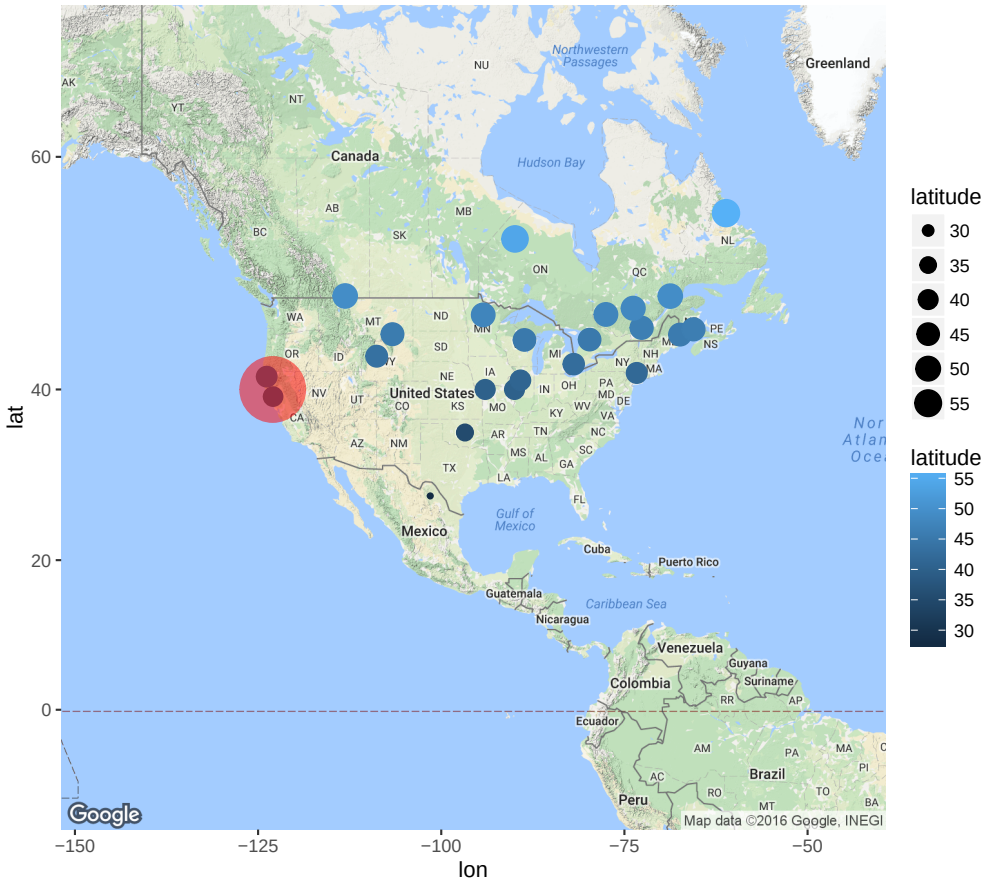




Algic, Latitude Signal: 0

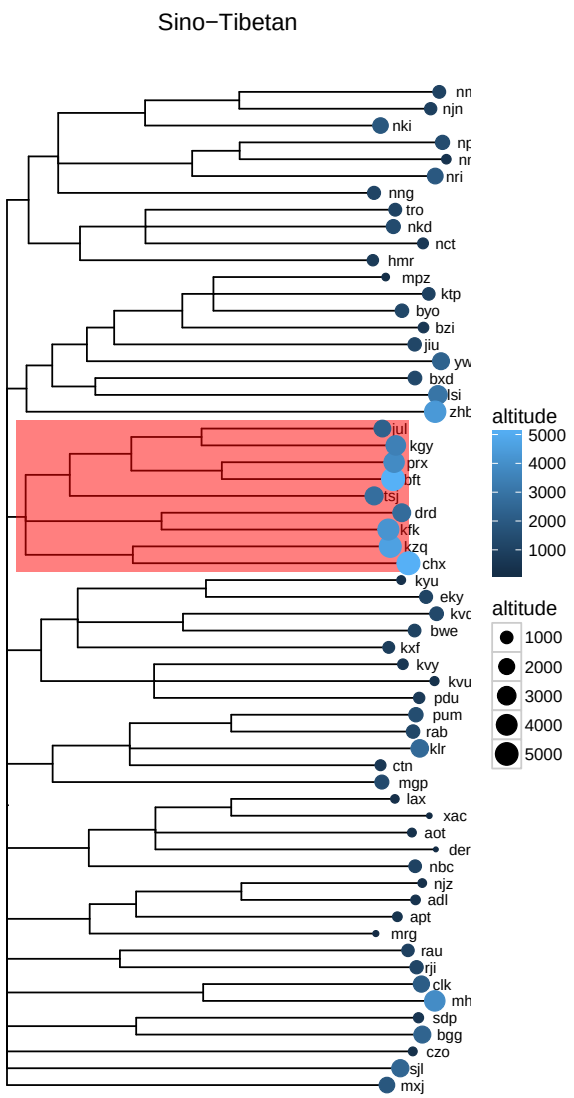
(branch lengths: ASJP, tree topology: Autotyp)

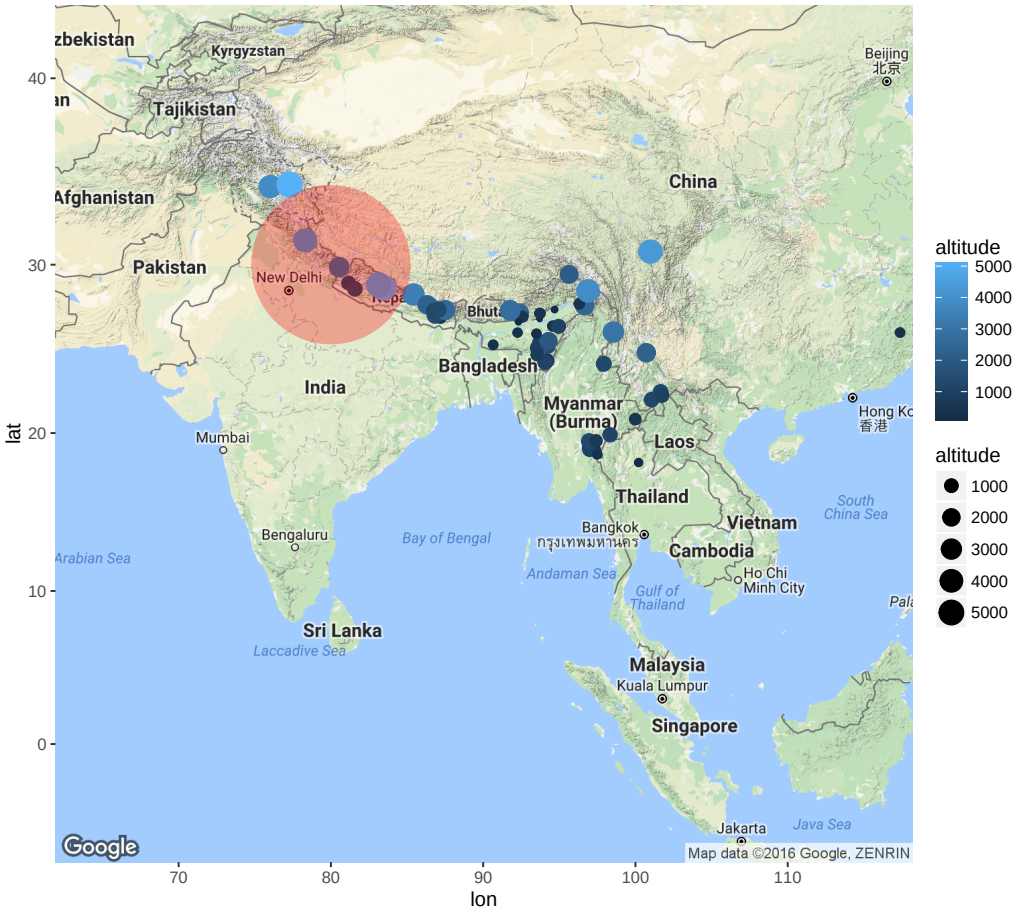




Altitude λ signal

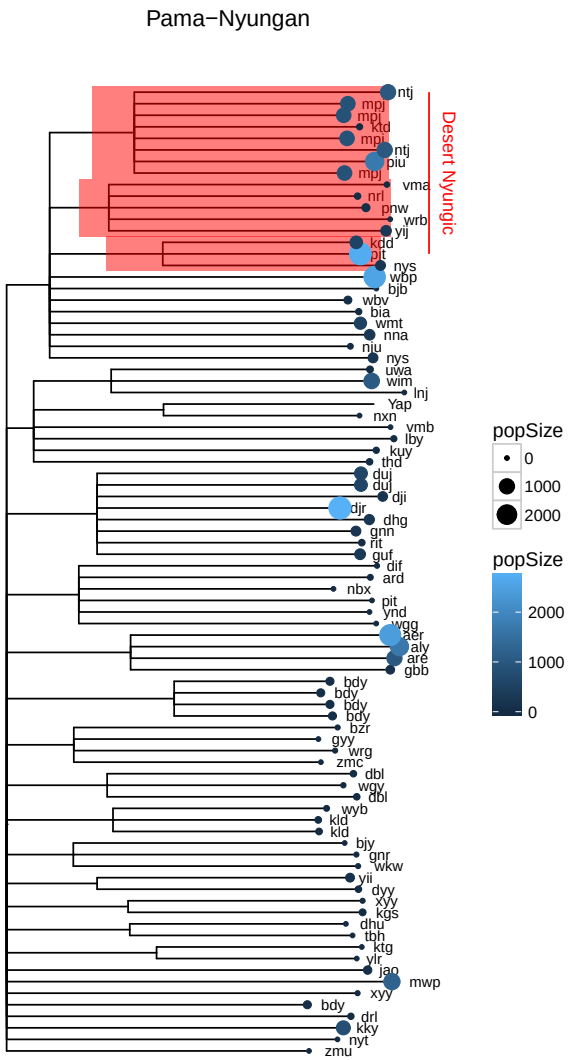
family	lang.	altitude	p-value	sig.	branch source	tree source
Arawakan	26	0.99	0.05	λ	autotyp	autotyp
Arawakan	25	0.99	0.06	λ	wals(euclidean)	autotyp
Sino-Tibetan	60	0.99	< 0.001	λ	asjp16	glottolog
Arawakan	22	0.99	0.06	λ	autotyp	glottolog
Arawakan	22	0.99	0.04	λ	wals(euclidean)	glottolog
				...		
Tupian	22	0	1	λ	wals(euclidean)	ethnologue
Tupian	21	0	1	λ	wals(euclidean)	wals
Dravidian	23	0	1	λ	autotyp	autotyp
Dravidian	22	0	1	λ	wals(euclidean)	wals
Algic	28	0	1	λ	wals(euclidean)	wals
Algic	26	0	1	λ	autotyp	wals

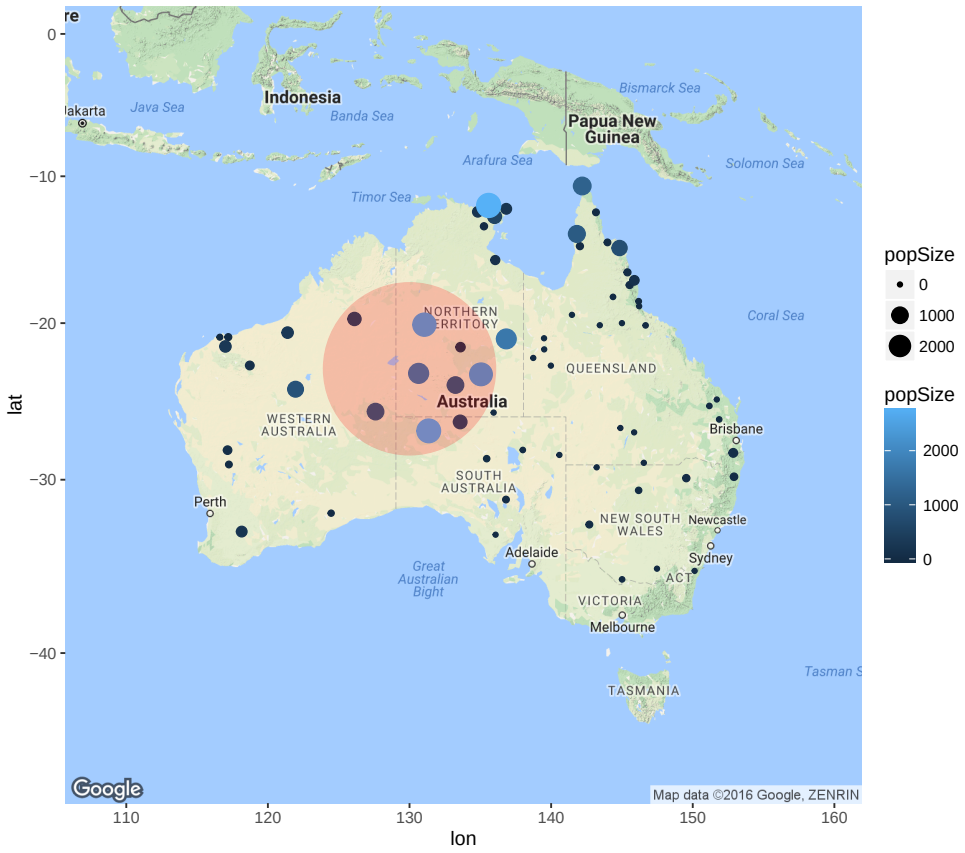

Sino−Tibetan
nn
njn
nki
np
nr
nri
nng
tro
nkd
nct
hmr
mpz
ktp
byo
bzi
jiu
yw
bxd
lsi
zhb
jul
kgy
prx
bft
tsj
drd
kfk
kzq
chx
kyu
eky
kvc
bwe
kxf
kvy
kvu
pdu
pum
rab
klr
ctn
mgp
lax
xac
aot
der
nbc
njz
adl
apt
mrg
rau
rji
clk
mh
sdp
bgg
czo
sjl
mxj
altitude
5000
4000
3000
2000
1000
altitude
1000
2000
3000
4000
5000


Kyrgyzstan
Tajikistan
Afghanistan
Pakistan
New Delhi
China
Beijing
北京
India
Bangladesh
Myanmar
(Burma)
Laos
Thailand
Vietnam
Cambodia
Ho Chi
Minh City
Bangkok
Mumbai
Bengaluru
Sri Lanka
Laccadive Sea
Arabian Sea
Bay of Bengal
Andaman Sea
Gulf of
Thailand
South
China Sea
Malaysia
Kuala Lumpur
Singapore
Jakarta
Java Sea
Hong Kong
香港
Google
Map data ©2016 Google, ZENRIN
lat
40
30
20
10
0
lon
70
80
90
100
110
altitude
5000
4000
3000
2000
1000
altitude
1000
2000
3000
4000
5000

Population size λ signal

family	lang.	popSize	p-value	sig.	branch source	tree source
Uralic	23	0.99	0.29	λ	asjp16	autotyp
Algic	26	0.99	0.02	λ	asjp16	ethnologue
Pama-Nyungan	70	0.99	< 0.001	λ	asjp16	autotyp
Arawakan	25	0.99	0.45	λ	asjp16	autotyp
Sino-Tibetan	107	0.99	< 0.001	λ	asjp16	autotyp
Mayan	27	0.99	0.027	λ	asjp16	autotyp
			...			
Tupian	21	0	1	λ	wals(euclidean)	wals
Dravidian	23	0	1	λ	autotyp	autotyp
Dravidian	22	0	1	λ	wals(euclidean)	wals
Algic	28	0	1	λ	wals(euclidean)	wals
Algic	26	0	1	λ	autotyp	wals
Tupian	22	0	1	λ	wals(euclidean)	ethnologue


Pama−Nyungan
Desert Nyungic
popSize
0
1000
2000
popSize
2000
1000
0


Indonesia
Papua New Guinea
Australia
Java Sea
Banda Sea
Bismarck Sea
Arafura Sea
Solomon Sea
Timor Sea
Coral Sea
NORTHERN TERRITORY
QUEENSLAND
WESTERN AUSTRALIA
SOUTH AUSTRALIA
NEW SOUTH WALES
VICTORIA
TASMANIA
ACT
Perth
Adelaide
Brisbane
Newcastle
Sydney
Melbourne
Great Australian Bight
Google
Map data ©2016 Google, ZENRIN
lat
lon
popSize
0
1000
2000
popSize
2000
1000
0

General Results

lang.	method	fam.	lat.	long.	alt.	popSize
20	λ	21	0.91	0.91	0.62	0.61
50	λ	7	0.99	0.94	0.8	0.7
100	λ	5	0.99	0.92	0.82	0.73
20	K	21	1.68	1.59	0.91	0.85
50	K	7	2.24	2.03	1.15	0.93
100	K	5	2.76	2.26	1.24	0.96

Phylosig cline: *latitude* > *longitude* > *altitude* > *population size*

Summary

- There is a general cline of phylosig:
  latitude >longitude >altitude >population size

Summary

- There is a general cline of phylosig:
  latitude >longitude >altitude >population size
- there is still considerable variation for different factors and specific families

Summary

- There is a general cline of phylosig:
  latitude >longitude >altitude >population size
- there is still considerable variation for different factors and specific families
- according to Symonds & Blomberg (2014) high phylogenetic signal suggests a gradual change over time, whereas low phylogenetic signal suggests either extremely rapid change or extreme stability

Summary

- There is a general cline of phylosig:
  latitude >longitude >altitude >population size
- there is still considerable variation for different factors and specific families
- according to Symonds & Blomberg (2014) high phylogenetic signal suggests a gradual change over time, whereas low phylogenetic signal suggests either extremely rapid change or extreme stability
- for languages this suggests that "external" factors have **gradually co-evolved** with the linguistic features the trees are built on

However

- λ is downwards biased by errors in topologies and errors in tip value estimations
- According to *Revell, Harmon and Collar (2008)* the interpretation of λ as rate of change is problematic, since depending on the evolutionary model

Thanks