Phylogeny in phonology: how Tai sound systems encode their past

Rikker Dockum
Yale University; rikker.dockum@yale.edu

What is phylogenetic signal?

Phylogenetic signal is a measure of statistical dependencies among traits due to phylogenetic relationships (Revell et al. 2008).

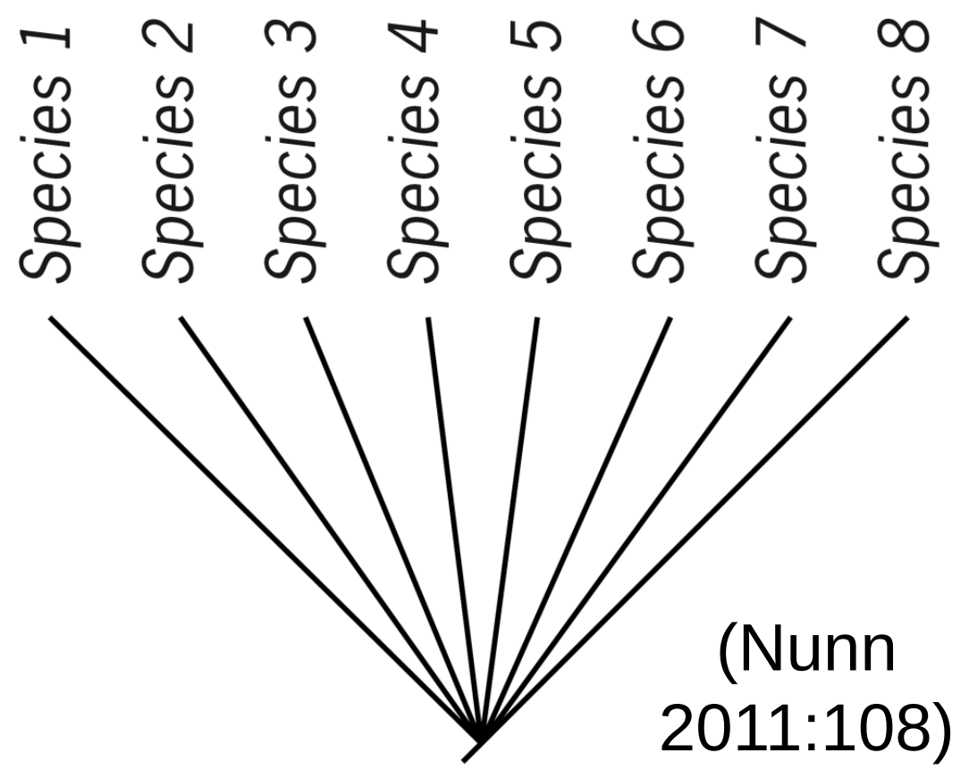
Research questions: (1) *How well do the phoneme inventories and the phonotactics of Tai languages fit a phylogenetic tree?* (2) *Would phonological data be useful for quantitative historical linguistics? (e.g. subgrouping, ancestral state reconstruction)*

What's the intuition? The more closely related two languages are, the more similar their phoneme inventories and phonotactic profiles will be (with usual caveats for coincidence and borrowing).

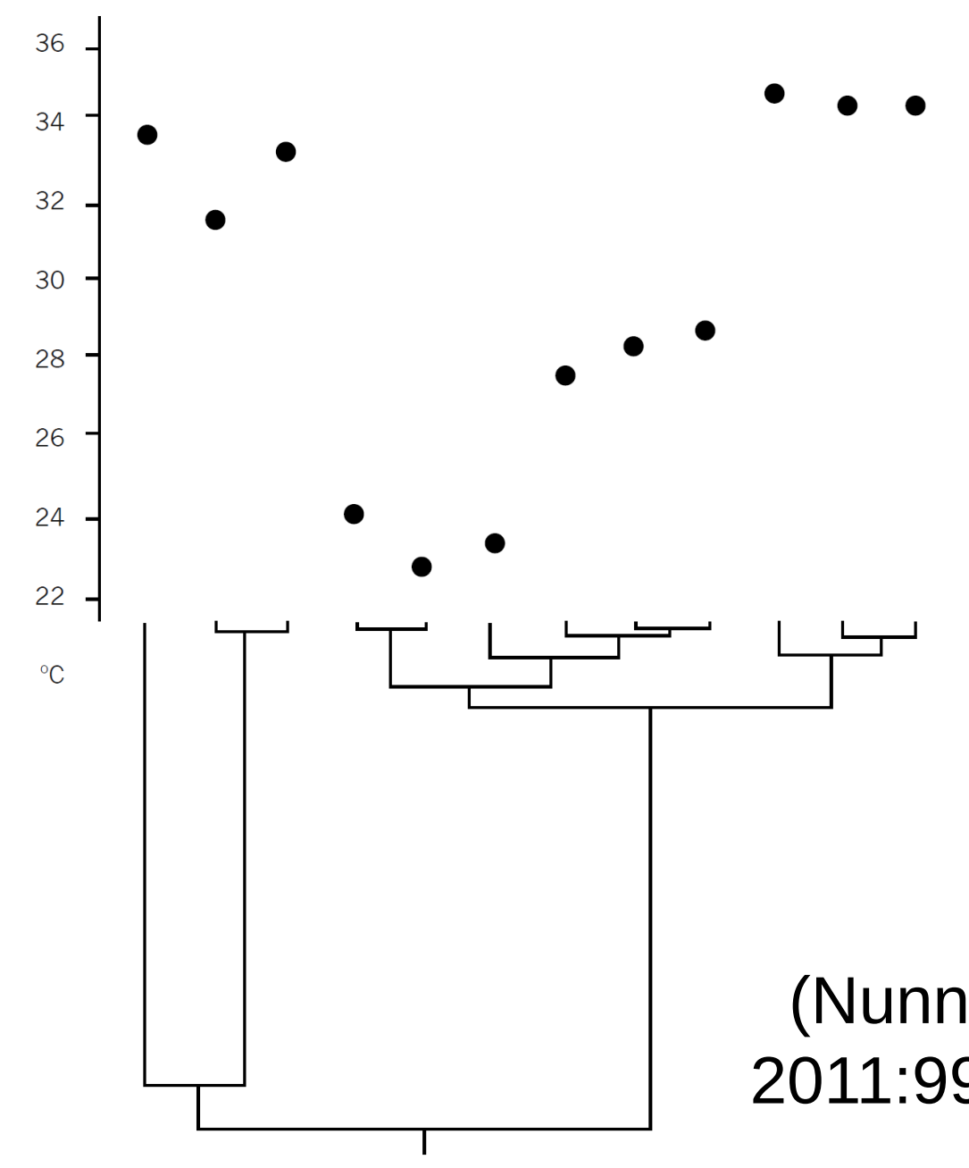
What do we gain by using computational phylogenetics?

- Lets us examine descent in new kinds of traits, and answer questions that may not be tractable with traditional methods
- Replicable and less subjectivity than the comparative method
- To date, lexical cognacy data has most often been used in linguistics (e.g. Gray, Bryant and Greenhill 2010)
- Work on phylogeny in sound systems is quite new (Macklin-Cordes 2015, Macklin-Cordes & Round 2016)

Tree with no phylogenetic signal



(Nunn 2011:108)

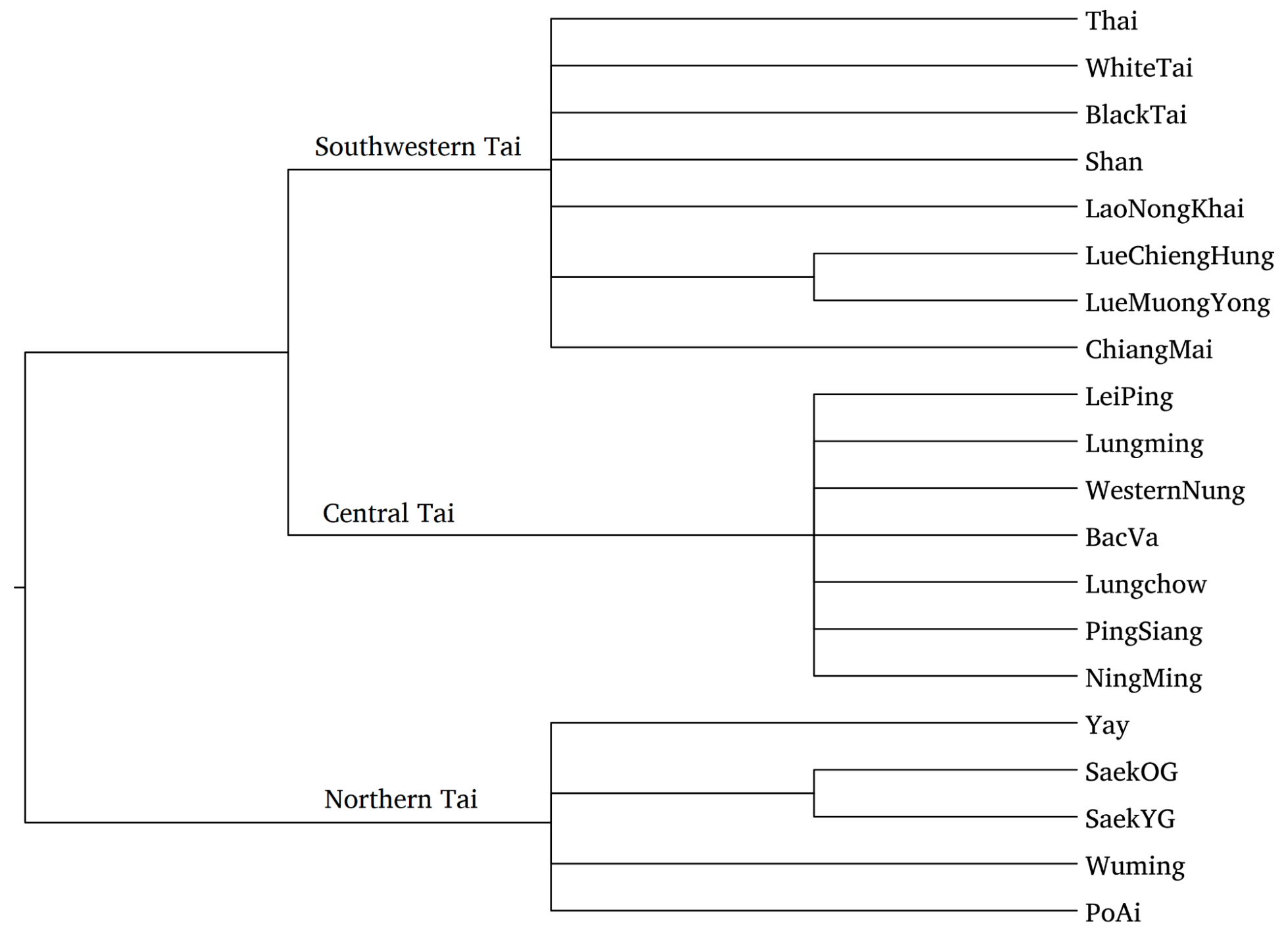


Phylogenetic signal in body temperatures of Australian skinks

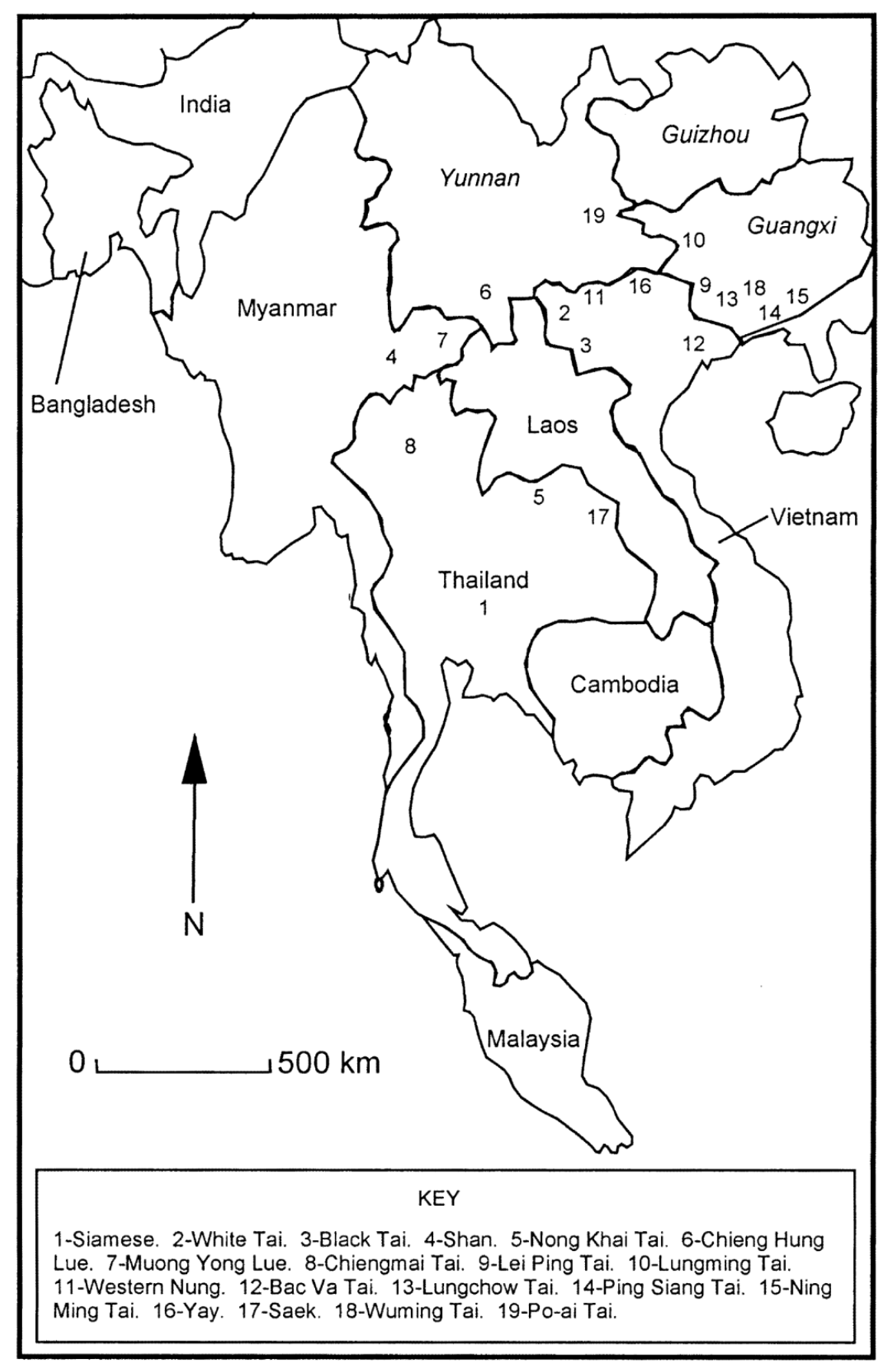
(Nunn 2011:99)

Background on Tai languages

1159 cognate sets from 20 Tai lects used in this study:

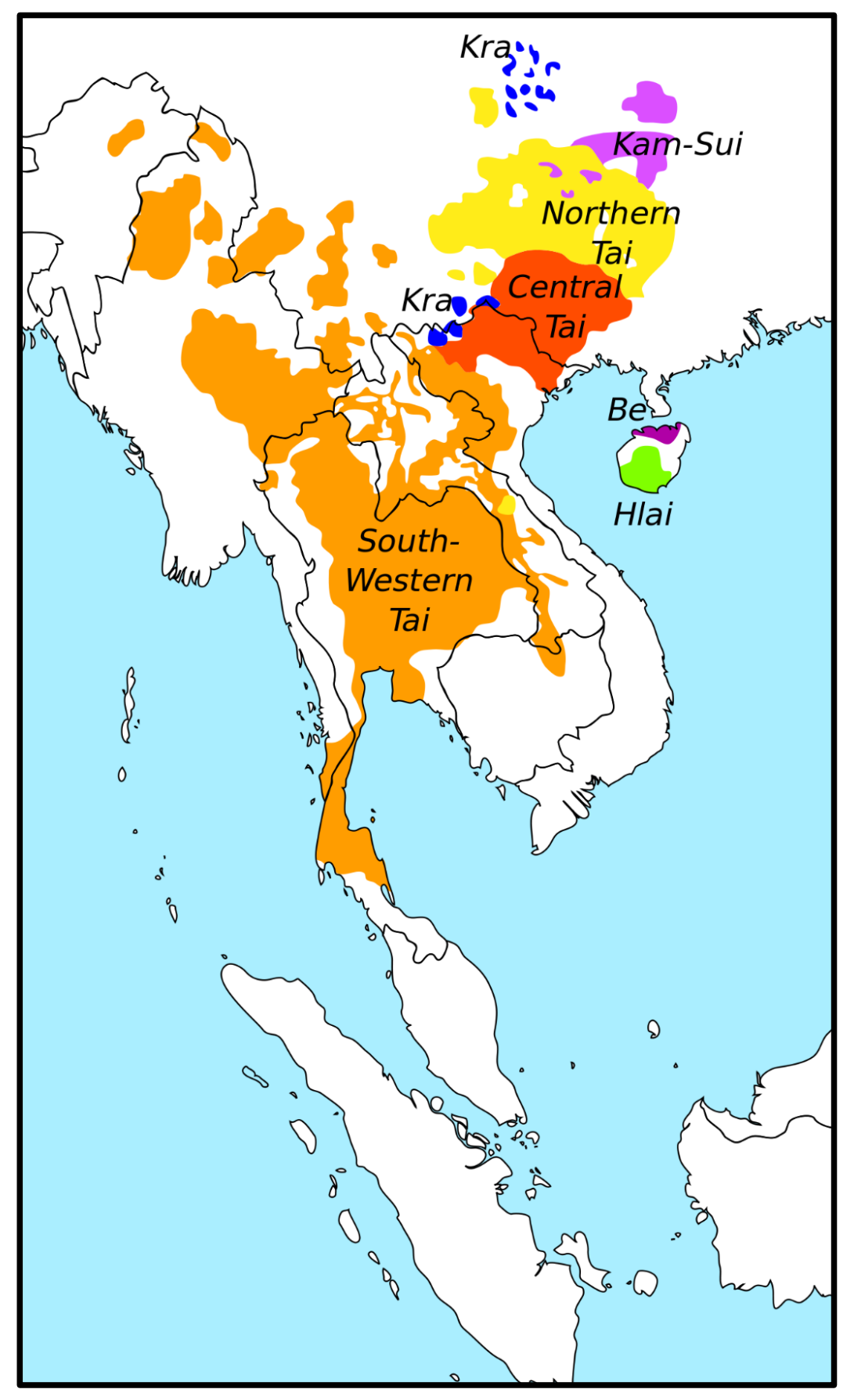


"Classic" comparative method tree of Tai lects used in this study (adapted from Chamberlain 1975)



KEY
1-Siamese. 2-White Tai. 3-Black Tai. 4-Shan. 5-Nong Khai Tai. 6-Chieng Hung Lue. 7-Muong Yong Lue. 8-Chiengmai Tai. 9-Lei Ping Tai. 10-Lungming Tai. 11-Western Nung. 12-Bac Va Tai. 13-Lungchow Tai. 14-Ping Siang Tai. 15-Ning Ming Tai. 16-Yay. 17-Saek. 18-Wuming Tai. 19-Po-ai Tai.

Tai lect locations
(Hudak 2008)



Kra-Dai language distribution
(Wikimedia Commons)

Statistical tests for phylogenetic signal

Method:
1) Two tests for phylogenetic signal (*D* test and Blomberg's *K*)
2) Applied to two types of phonological data (phonemes and biphones)

Different tests exist for different data types

Discrete data (binary and/or multistate)

- *D* test (Fritz & Purvis 2010)
- δ (Holland et al. 2002)
- Mean *Q*-residual (Gray et al. 2004)

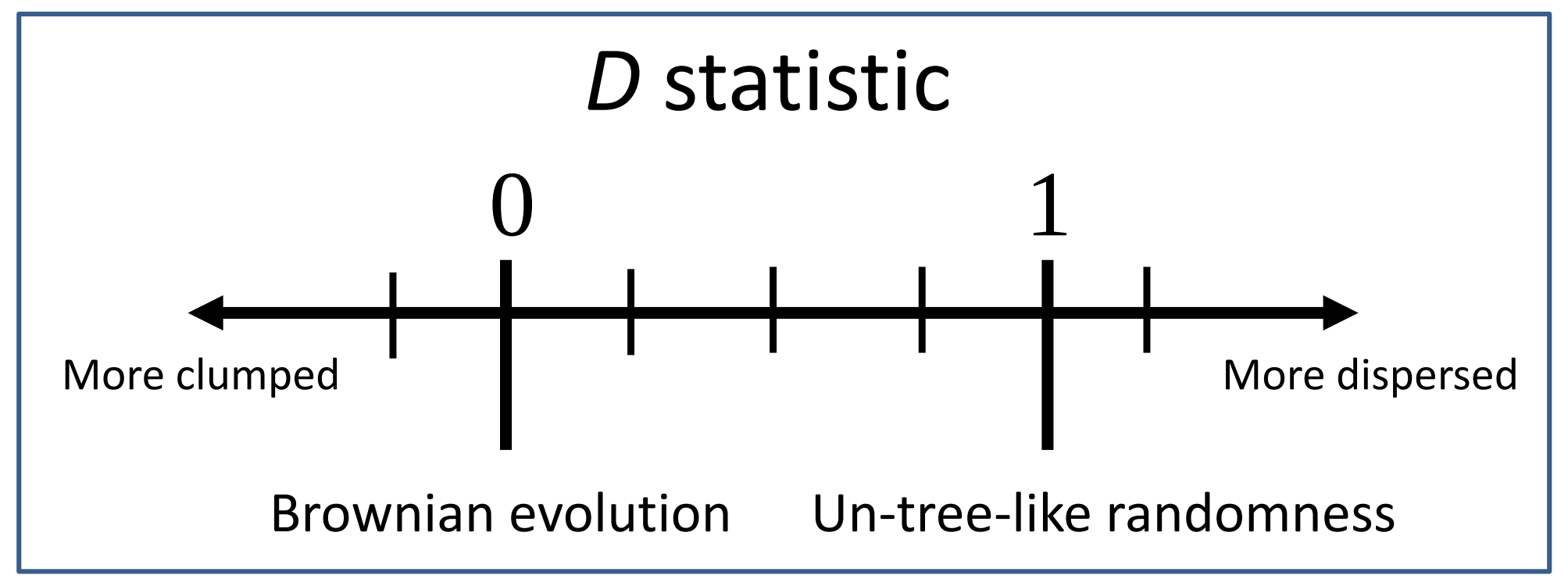
Continuous data

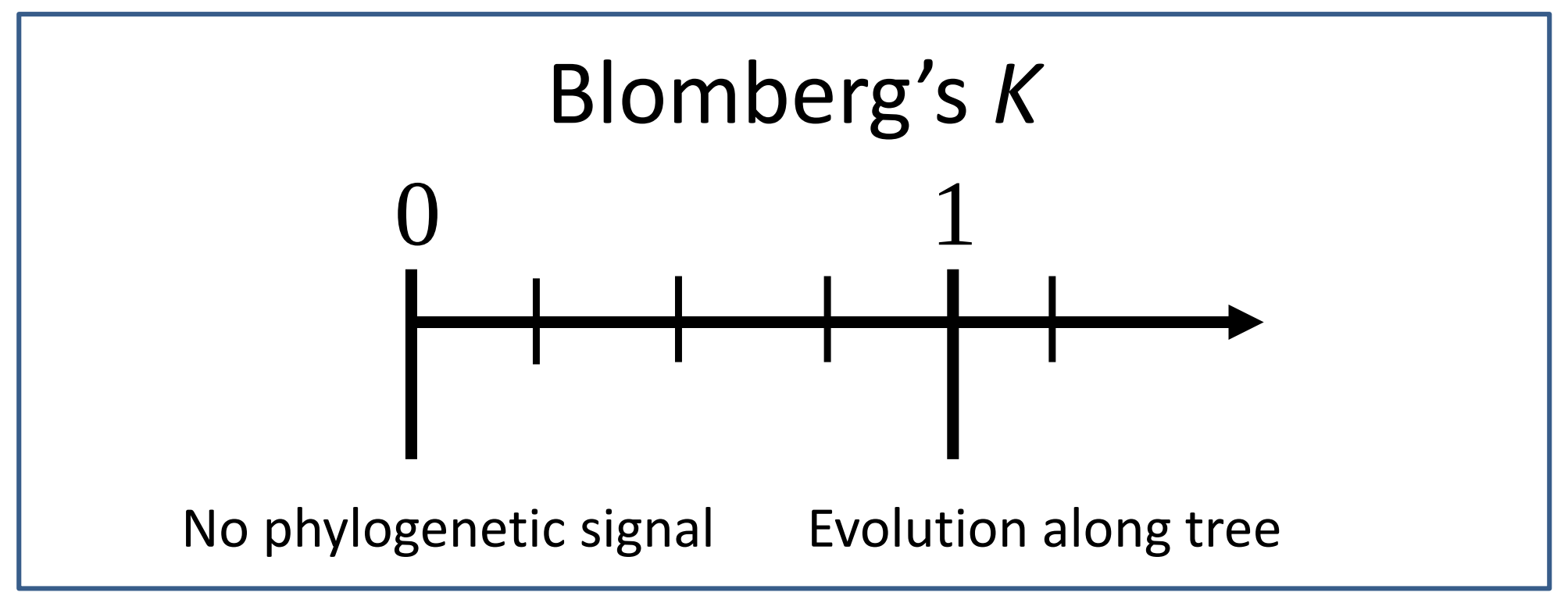
- *K* (Blomberg, Garland & Ives 2003)
- Cmean (Abouheif 1999)

Among others!

Brownian evolution: *model of evolutionary change with randomly fluctuating selection; aka "neutral evolution"*

What do the scores mean?





(Macklin-Cordes and Round 2016)

Data

Two types of data extracted with Python scripts from 1159 cognate sets:

1) Binary data ("coarse-grained" phonological data)
   - Phoneme presence/absence
   - Biphone presence/absence
2) Continuous data ("high-definition" phonological data)
   - Phoneme frequency (calculated as the quotient of total lexical items, as opposed to total phones)
   - Biphone Markov chain transition probability ($P_{ij} = P(x_{n+1} = j \mid x_n = i)$) (Ching & Ng 2006)

- Traits with no variation are pruned; phylogenetically uninformative, and some tests require their removal

Variation in Tai phoneme data

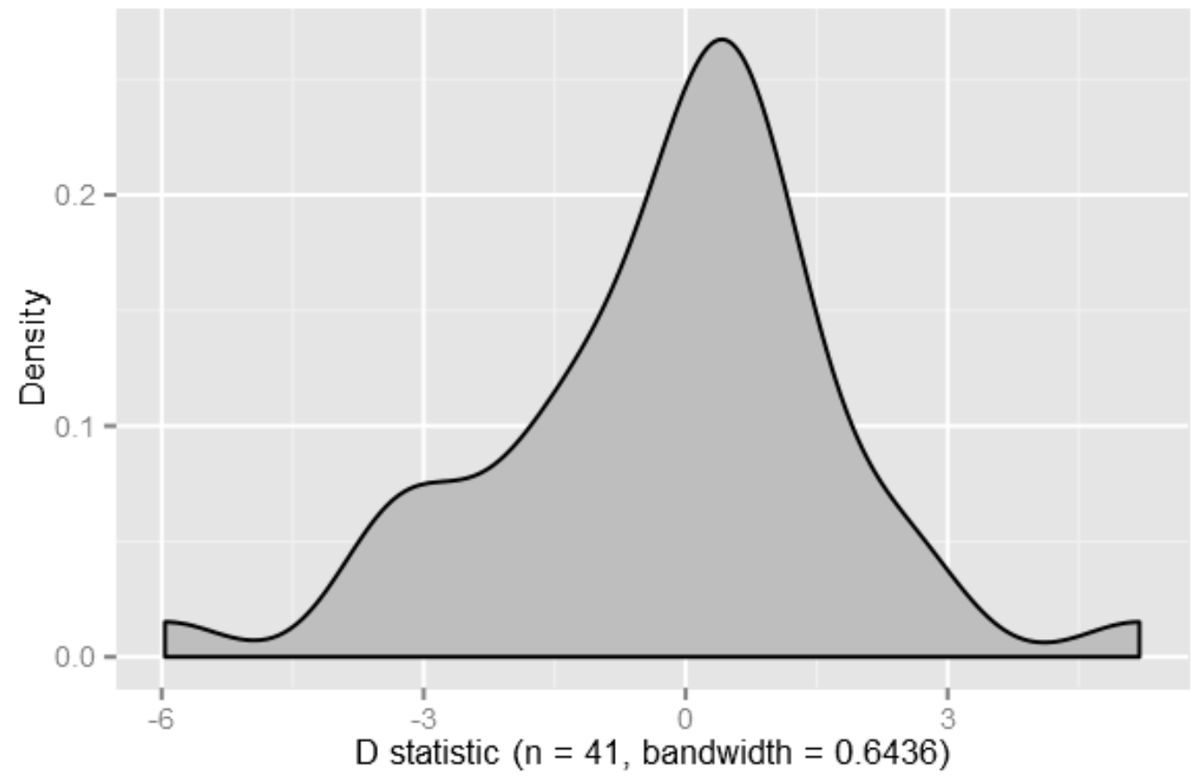
Lects	Phonemes	
	Total	w/Variation
20	54	41

Variation in Tai biphone data

Lects	Biphones	
	Total	w/Variation
20	555	526

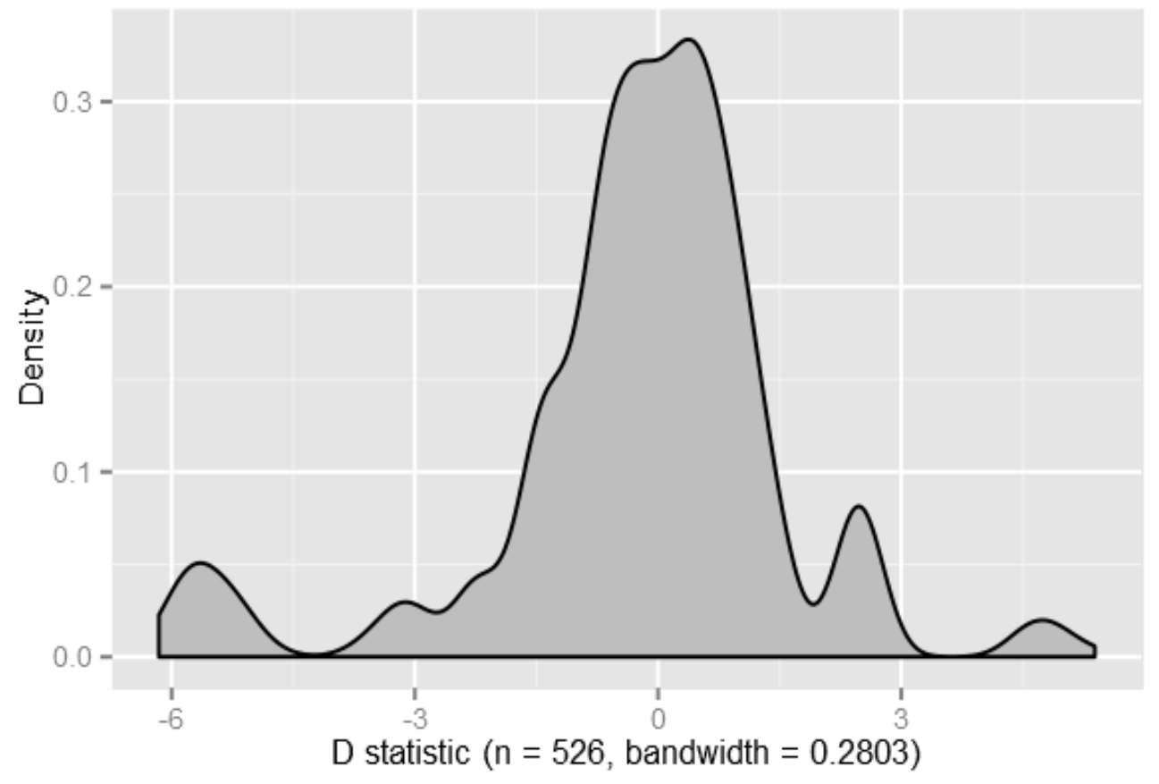
Test 1: *D* statistic (binary data) – R package `caper`

Density plot of *D* for Tai phonemes



	D	*p* $_{(D=0)}$	*p* $_{(D=1)}$
ɣ	-6.02706	0.9496	0
ʔy	-3.36769	0.6626	0.103
ʔd	-3.3493	0.6708	0.0992
ɤ	-3.12975	0.9512	0
ʔb	-3.01933	0.6591	0.097
θ	-2.38469	0.858	0.0104
tʰ	-2.31074	0.8556	0.0103
kʰ	-1.60053	0.9656	0
ă	-1.54586	0.7805	0.0263
r	-1.32693	0.8592	0.0039
ɔ	-1.22613	0.8612	0.0019
pʰ	-1.13276	0.861	0.0059
ɬ	-0.73393	0.8379	0.0022
ɰ	-0.72056	0.8483	0.002
ɯ	-0.35617	0.6508	0.0406
u	-0.3087	0.6433	0.0375
…	…	…	…
oː	0.986454	0.151	0.4283
ɯː	0.997876	0.1492	0.4253
b	1.087499	0.1805	0.4069
e	1.420117	0.2461	0.4974
d	1.495991	0.0934	0.6163
c	1.517884	0.042	0.7487
k	2.460614	0.2287	0.3501
o	2.474653	0.229	0.3529
ð	2.506732	0.2301	0.352
sʰ	4.94437	0	0.7052
Mean D	**-0.11911**		
SD	**1.98**		

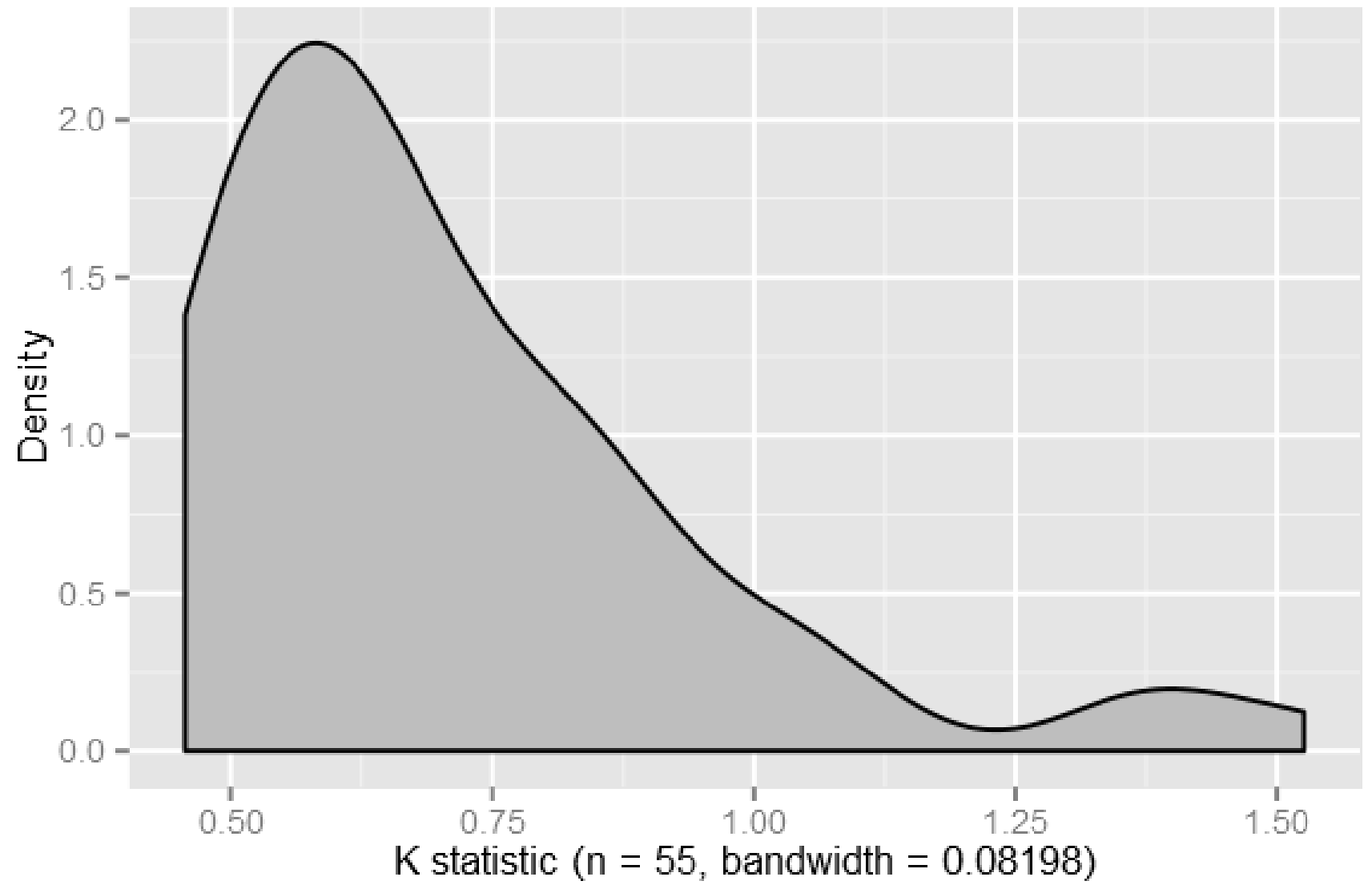
Density plot of *D* for Tai biphones



	D	*p* $_{(D=0)}$	*p* $_{(D=1)}$
rɤː	-6.15479	0.9522	0
hă	-5.97734	0.953	0
ɣaː	-5.95582	0.9483	0
reː	-5.92565	0.9488	0
ɛl	-5.91928	0.8326	0
wɤː	-5.84236	0.9517	0
rɯː	-5.81630	0.9468	0
#ɣ	-5.81201	0.9491	0
uːl	-5.67541	0.8366	0
ă#	-5.67237	0.9479	0
ɣɔː	-5.65405	0.9467	0
ɲiː	-5.63626	0.947	0
xeː	-5.59447	0.9492	0.0068
aːl	-5.57923	0.8313	0
xoː	-5.56562	0.9524	0.0074
xiː	-5.54646	0.9491	0.0062
…	…	…	…
oʔ	4.46611	0	0.696
aʔ	4.55446	0	0.7024
ɤp	4.57207	0	0.6949
eʔ	4.63794	0	0.6944
#sh	4.67346	0	0.7015
sʰɯ	4.81471	0	0.6994
sʰɤ	4.87617	0	0.6996
sʰi	4.89928	0	0.7051
ɤʔ	4.90521	0	0.7
bɤ	4.93359	0	0.7016
Mean D	**-0.23937**		
SD	**1.86**		

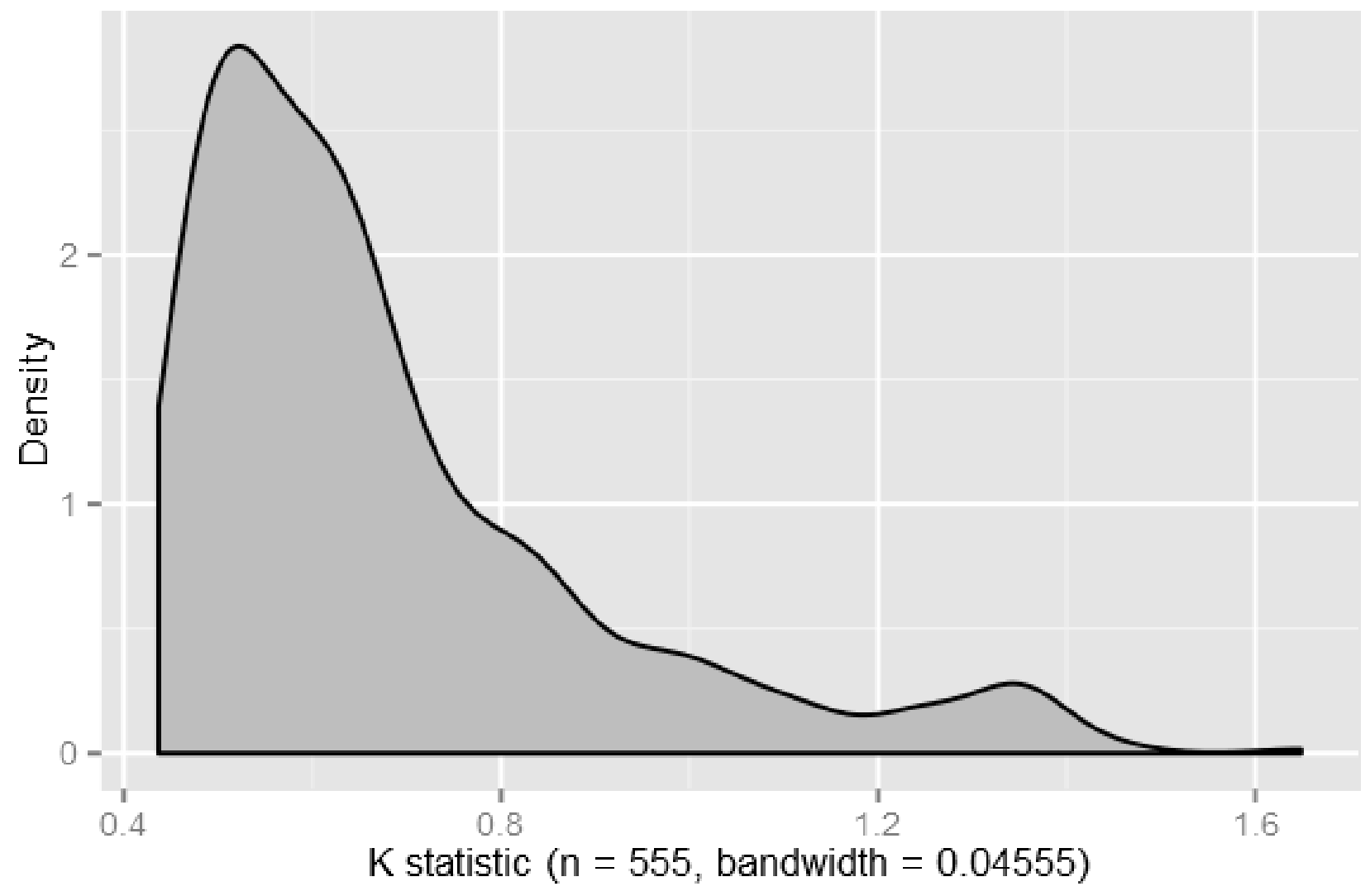
Test 2: Blomberg's *K* (continuous data) – R package `picante`

Density plot of *K* for Tai phonemes



Mean *K* for Tai phonemes: 0.71

Density plot of *K* for Tai biphones



Mean *K* for Tai biphones: 0.68

Conclusions

- This study finds strong phylogenetic signal in "course-grained" binary phonemic and biphone data, contra Macklin-Cordes and Round (2016); this can be attributed to the greater degree of variation in the phoneme systems of Tai languages than in Australian Aboriginal languages.
- This study also confirms the findings of Macklin-Cordes and Round (2016), which observed phylogenetic signal in "high-definition" phonotactic data in Australian aboriginal languages
- Additional tests of δ-score (Holland et al. 2002) and mean *Q*-residual (Gray et al. 2010) also showed signal in the data
- Phonological data of these types shows promise for use in quantitative historical linguistic tasks

Future directions

- A better tree is sorely needed! New lexical phylogenetic tree of the Kra-Dai family in progress
- Testing other types of phonological data for Tai languages
  - e.g. historical tone splits and mergers, as derived from Gedney (1972) tone boxes (in progress)
- Once the new KD tree is ready, perform ancestral state reconstruction on phonological traits