

An introduction to
Culturolinguistic phylogeny and
Phylogeny using continuous data

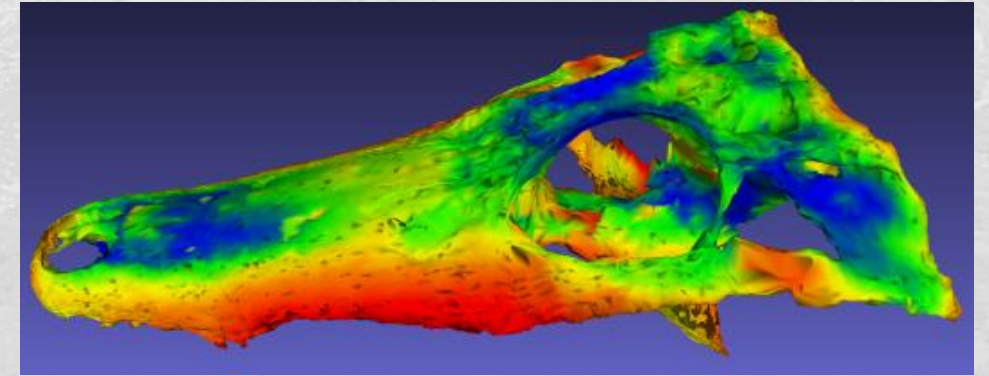
Dr. Roland B. Sookias



sookias.r.b@gmail.com
rsookias.info

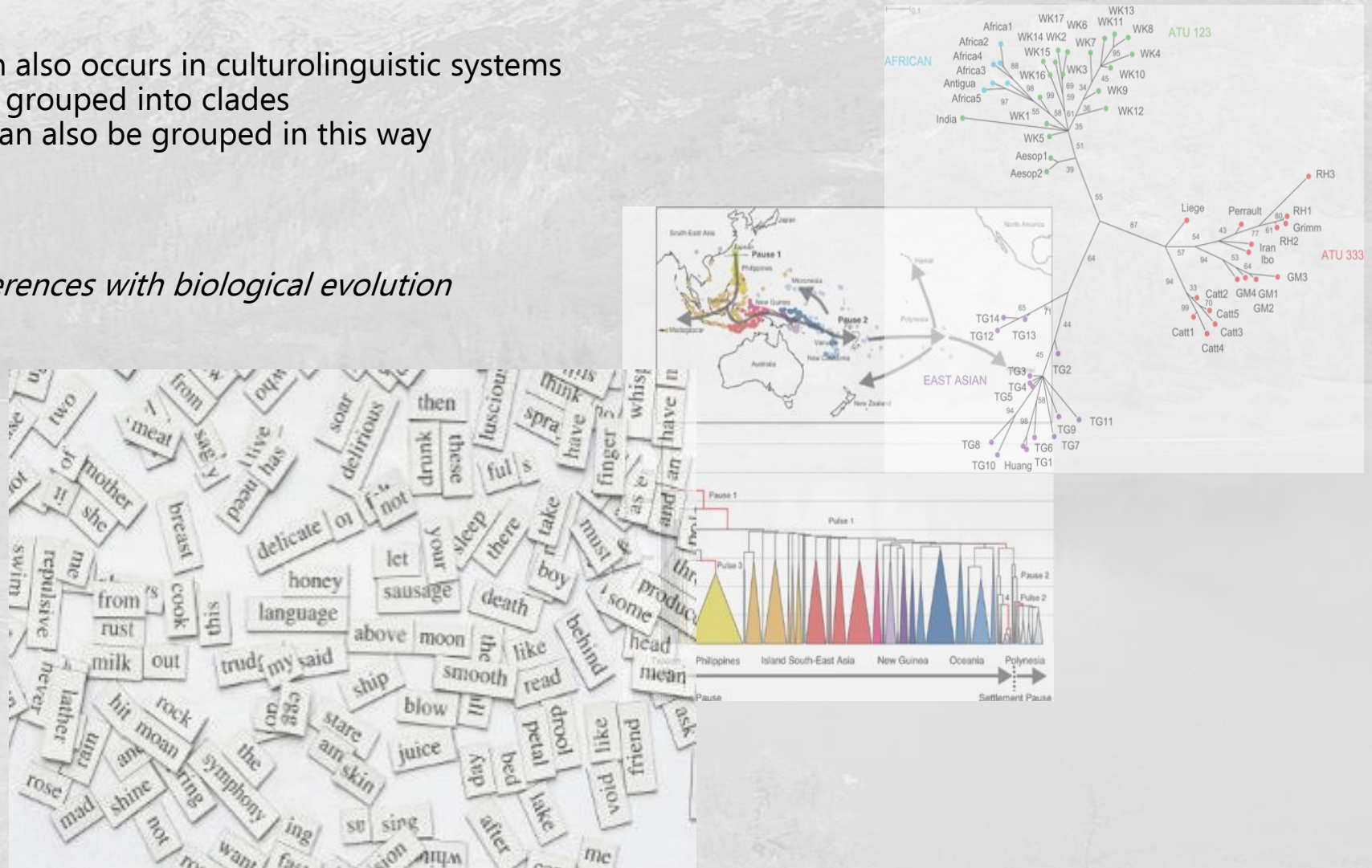
Me

- Background in Triassic archosaurs and their phylogenetics
- Currently working on inferring phylogeny from 3D morphological data
- Previously some work (and interest in!) language phylogeny



Language and culture evolve too!

- Descent with modification also occurs in culturolinguistic systems
- Human languages can be grouped into clades
- Many aspects of culture can also be grouped in this way
- Lecture plan:
 - *Some history*
 - *Methods*
 - *Similarities and differences with biological evolution*
 - *Applications*

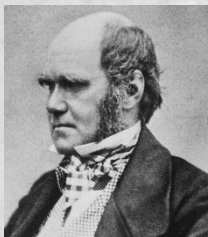


History

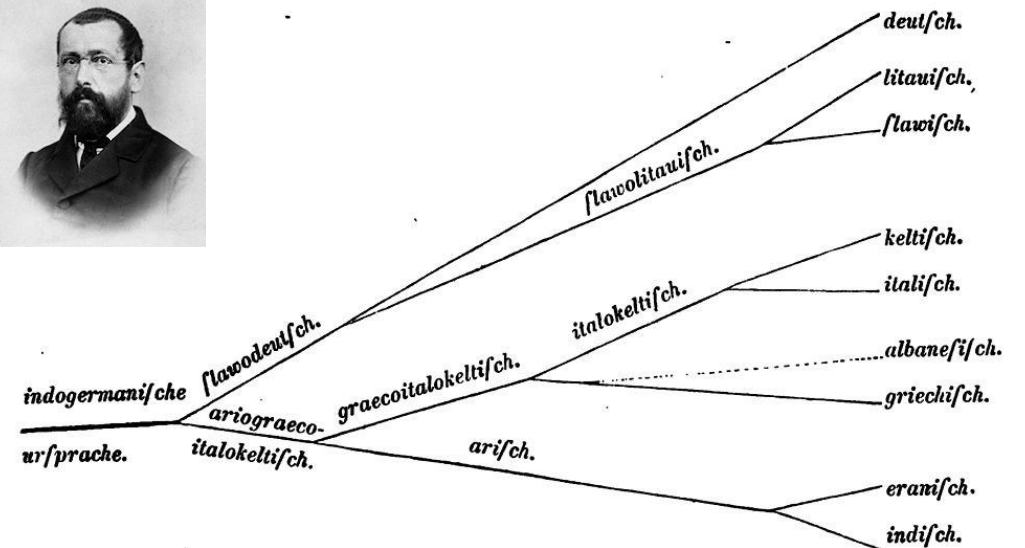
- At least to William Jones (Bengal, 1780)
- Some of the oldest “phylogenetic” trees were linguistic
- Schleicher developed the *Stammbaumtherorie* around the same time as Darwin
- Mentioned by Darwin



“[Sanskrit bears to Latin and Greek] ...a stronger affinity... than could possibly have been produced by accident; so strong indeed, that no philologist could examine them all three, without believing them to **have sprung from some common source**, which, perhaps, no longer exists”



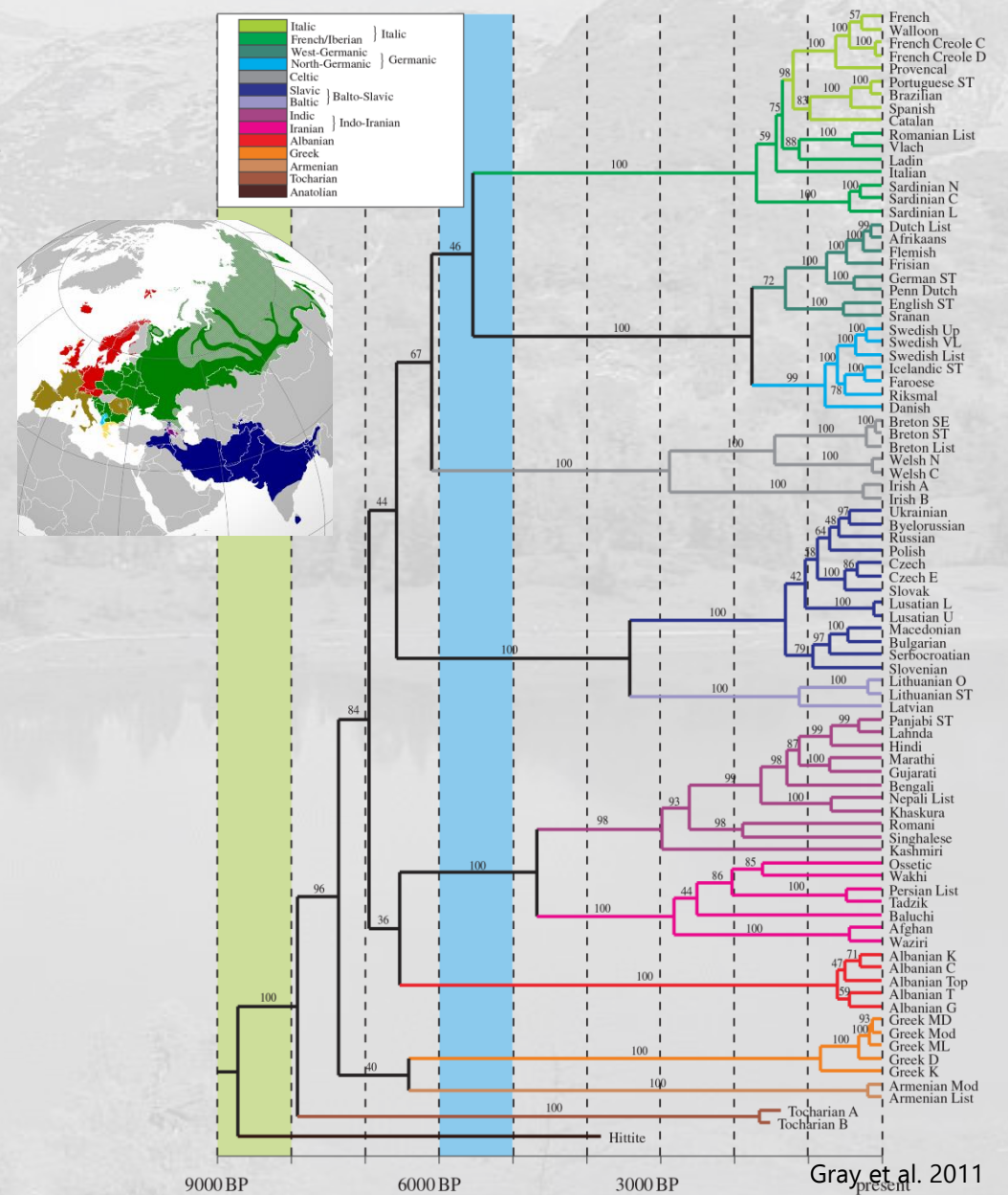
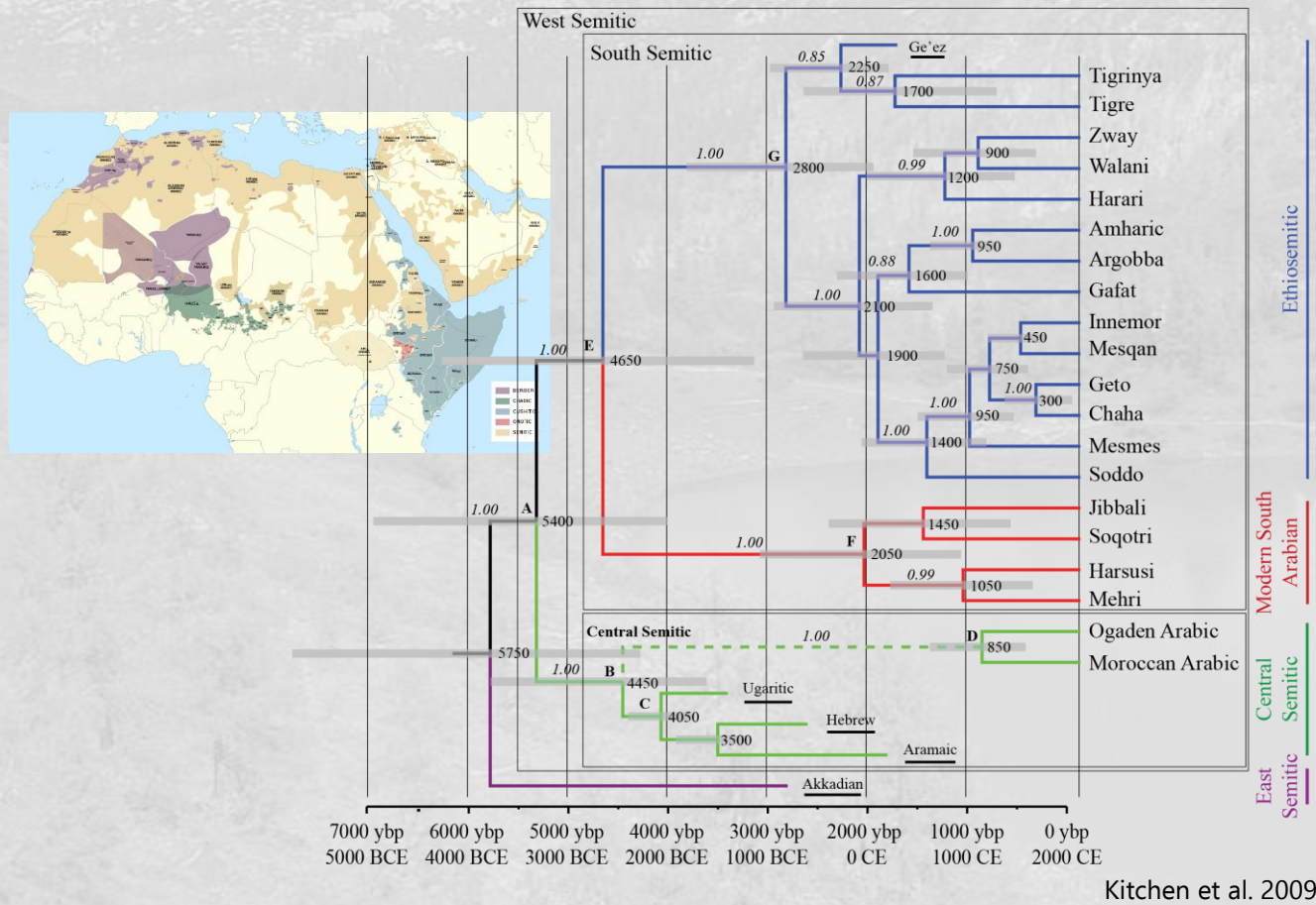
“the proper or even the only possible arrangement would ...be **genealogical**; and this would be strictly natural, as it would **connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin** of each tongue.” (ch. 13: 422)



From Schleicher 1861 *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*


History

- Languages are indeed grouped into clades using trees today, e.g. Indo-European, Semitic (with Afroasiatic)



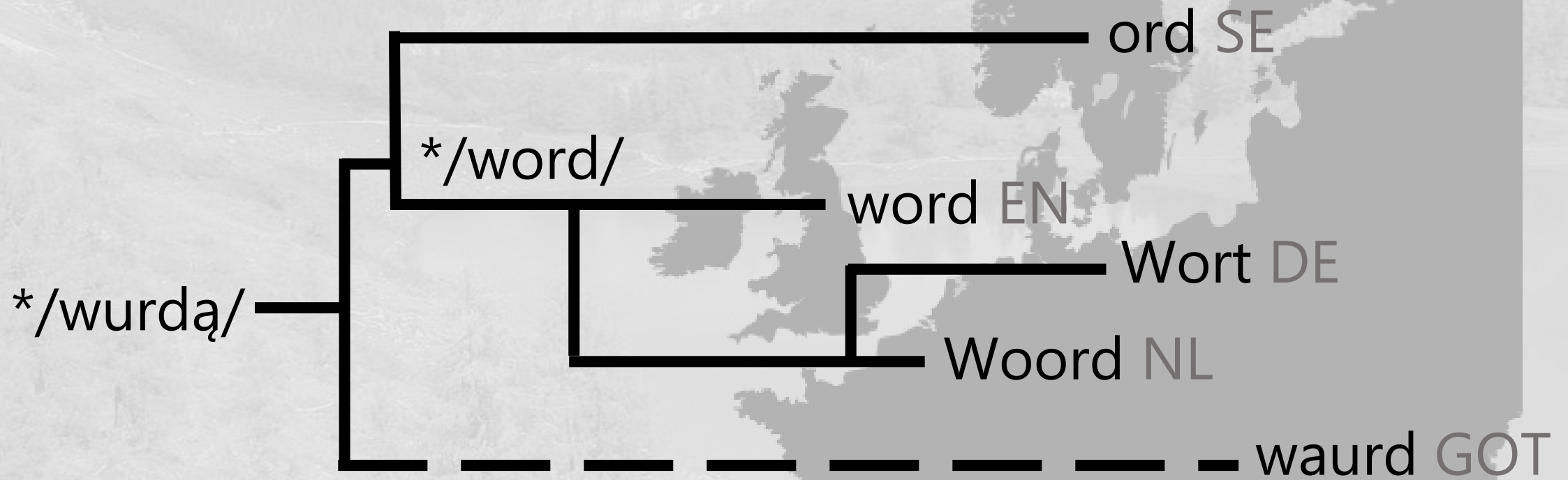
Methods

- Fundamental basis is the idea of **cognates**
- =Homologies in biology, i.e. **features shared due to common origin**
- **Words** can be cognates, but may change their meaning and sound
- **Sounds** can be corresponding phonemes (=homologues) even though they differ
- Shared words due to common origin should all experience the same sound shifts – if not, borrowing (=hybridisation) may have occurred

DE	Wasser	das(s)	Butter		butyrum LAT
EN	water	that	butter		
NL	water	dat	boeter		

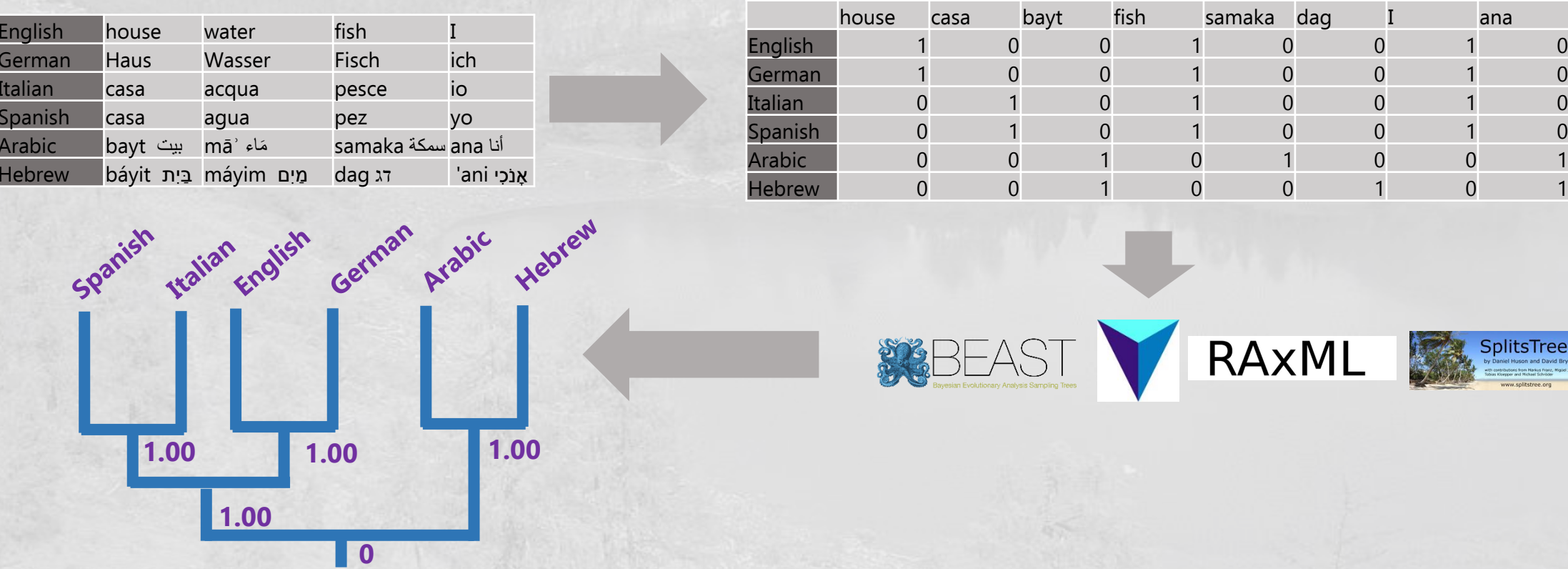
Methods

- Classically “comparative method”
 - Identified **cognate** (=homologue) words (and sounds – **phonemes**)
 - Reconstructed proto language (manually)
 - =ancestral state reconstruction



Methods

- Today similar methods to biological phylogenetics:
 - Common/fundamental words (less subject to borrowing) coded into matrix of “presence” or “absence”
 - Bayesian, ML, parsimony, neighbour joining networks
 - Some specific models (e.g. no regain after loss) can be used



Methods

- Also possible to code using multistate approaches, but rarely done

English	house	water	fish	I
German	Haus	Wasser	Fisch	ich
Italian	casa	acqua	pesce	io
Spanish	casa	agua	pez	yo
Arabic	bayt بيت	mā' ماء	samaka سمكة	ana أنا
Hebrew	báyit בַּיִת	máyim מַיִם	dag דג	'ani אֲנִי

	house	casa	bayt/báyit	fish	samaka	dag	I	ana
English	1	0	0	1	0	0	1	0
German	1	0	0	1	0	0	1	0
Italian	0	1	0	1	0	0	1	0
Spanish	0	1	0	1	0	0	1	0
Arabic	0	0	1	0	1	0	0	1
Hebrew	0	0	1	0	0	1	0	1

Binary

	house	water	fish	I
English	0	0	0	0
German	0	0	0	0
Italian	1	1	0	0
Spanish	1	1	0	0
Arabic	2	2	1	1
Hebrew	2	2	1	1

Multistate

Methods

- Phonemes can also be coded as characters
 - Rarely done, because very labile
- “Morphological” features, i.e. grammatical structure (cases, word order etc.)
 - Difficult to comparably code characters, but may be useful for deep phylogeny



	"th" in "thorn"
English	thorn /θ/
German	Dorn /d/
Dutch	doorn /d/
Swedish	torn /t/
Gothic	thaurus */θ/

English	0
German	1
Dutch	1
Swedish	2
Gothic	0

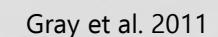
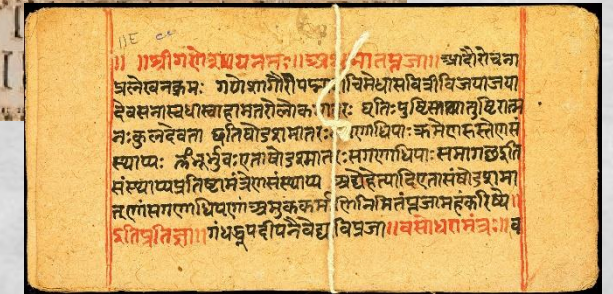
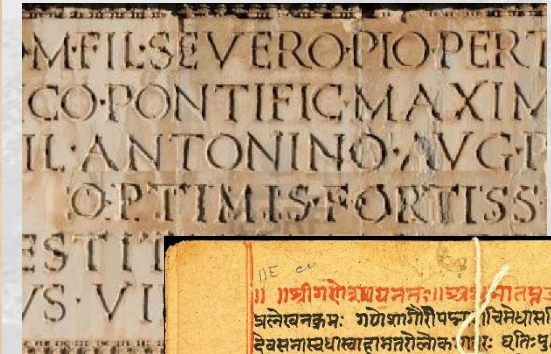
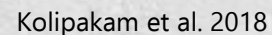
	Voicing	Change to stop
English	0	0
German	1	1
Dutch	1	1
Swedish	0	1
Gothic	0	0

	/θ/	/d/	/t/
English	1	0	0
German	0	1	0
Dutch	0	1	0
Swedish	0	0	1
Gothic	1	0	0



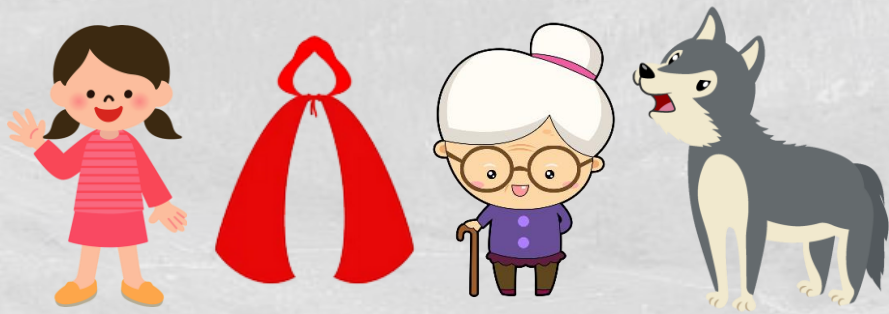
Gender	Masculine
Indefinite	en hest <i>a horse</i>
Definite	hesten <i>horse.DEF</i>
Double definite	den hesten <i>that horse.DEF</i>
Adjective	en fin hest <i>a nice horse</i>
Possessive	min hest/hesten min <i>my horse/horse.DEF</i> <i>my</i>

- Tree dating can be incorporated, as in biological phylogeny
- E.g. minimum divergence dates can be based on manuscript ages
- “Ancestral” languages usually treated as separate tips, as in biology

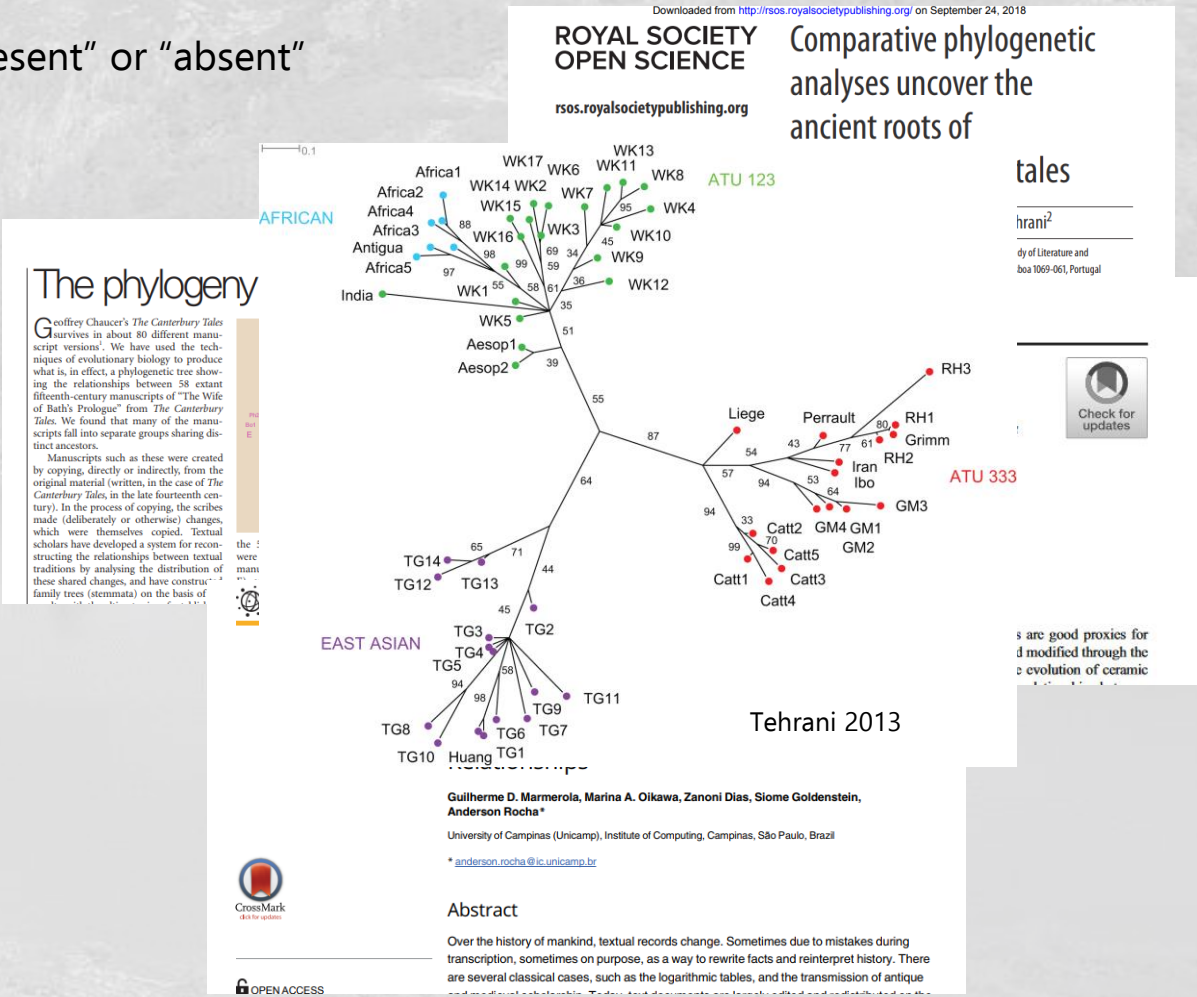


Methods

- Can also code other cultural data
 - For example, folk tales
 - Characters and elements of stories coded as “present” or “absent”
 - Other examples: pots, manuscripts

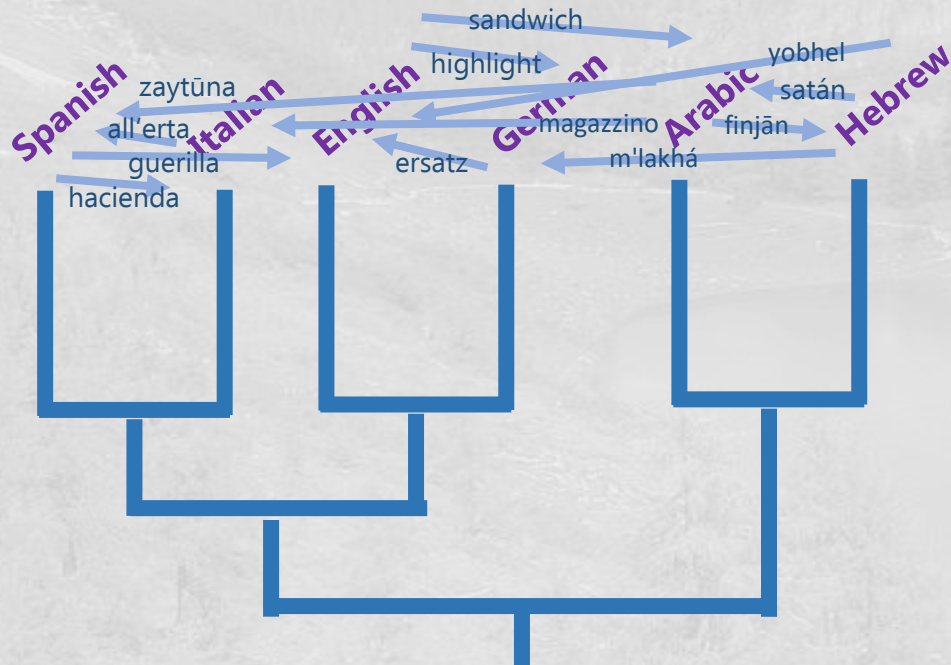


Tale1	1	1	1	1
Tale2	0	1	0	1
Tale3	1	1	0	0



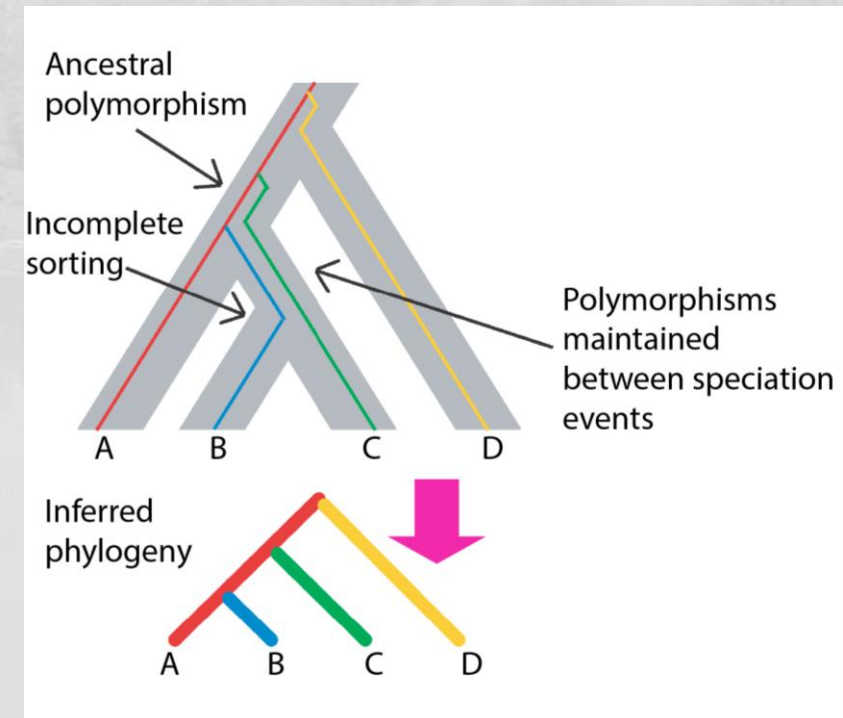
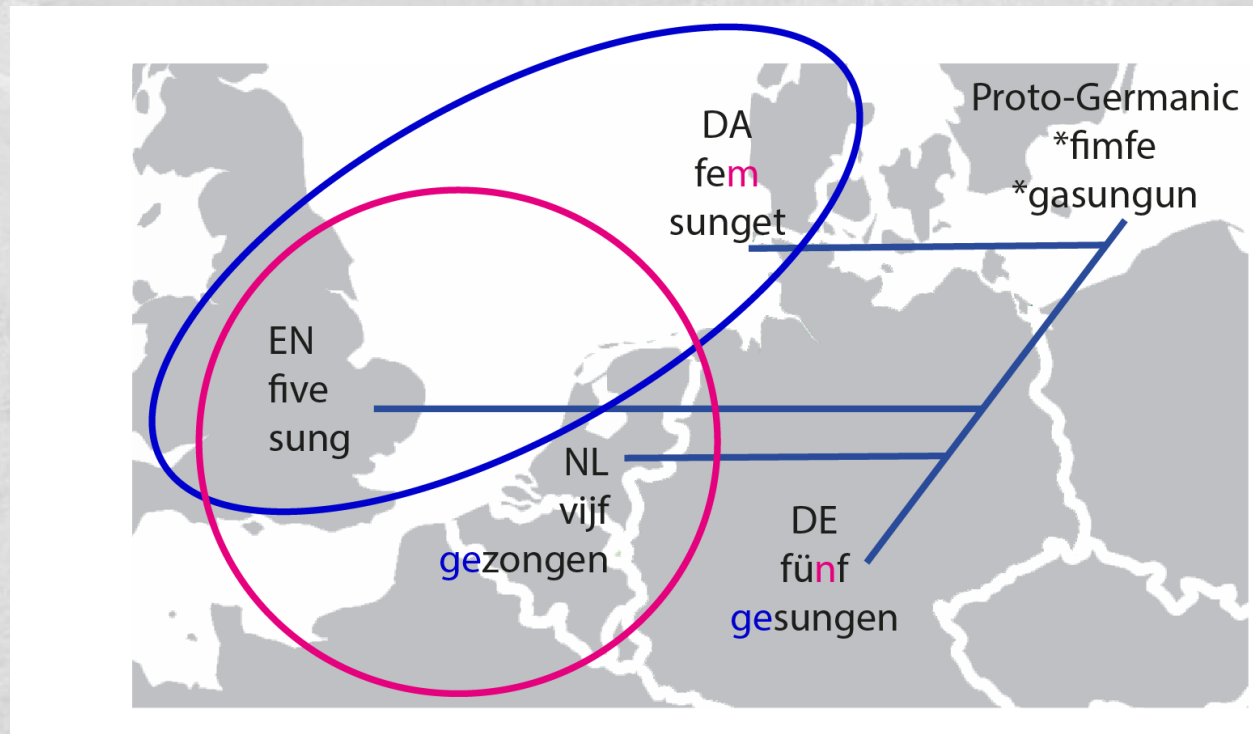
Similarities and differences with biological evolution

- Many aspects of culture, especially core parts of language, are inherited
- There is continued borrowing, even long after initial divergence (*never* complete separation)
- In this way, more similar to plant or bacterial evolution



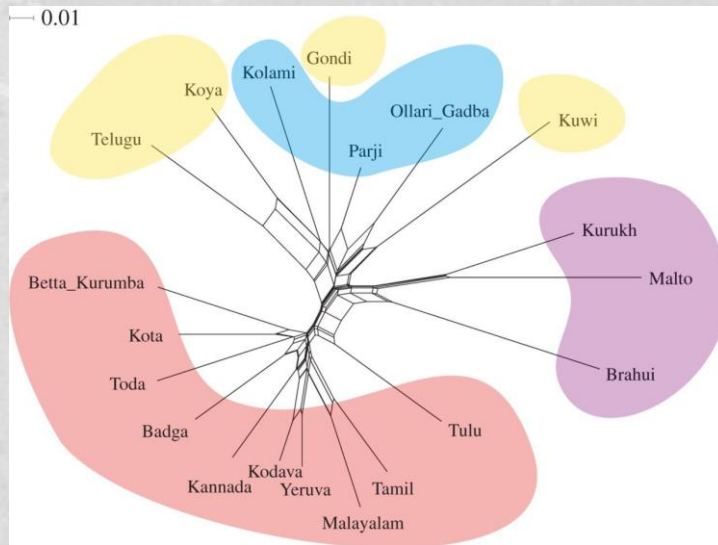
Similarities and differences with biological evolution

- “Wave theory” (*Wellentheorie* – opposing *Stammbaumtherie*) of Schmidt identified waves of radiation of language characteristics – effectively spread of alleles *within* population
- Because new features can arise in particular areas of a diverging language area, they can conflict with main signal and fit geography – not only due to “hybridisation”- effectively incomplete lineage sorting

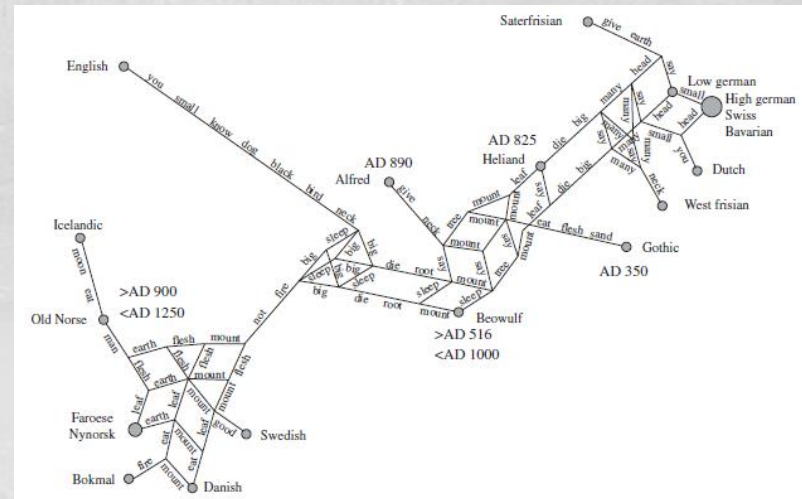


Similarities and differences with biological evolution

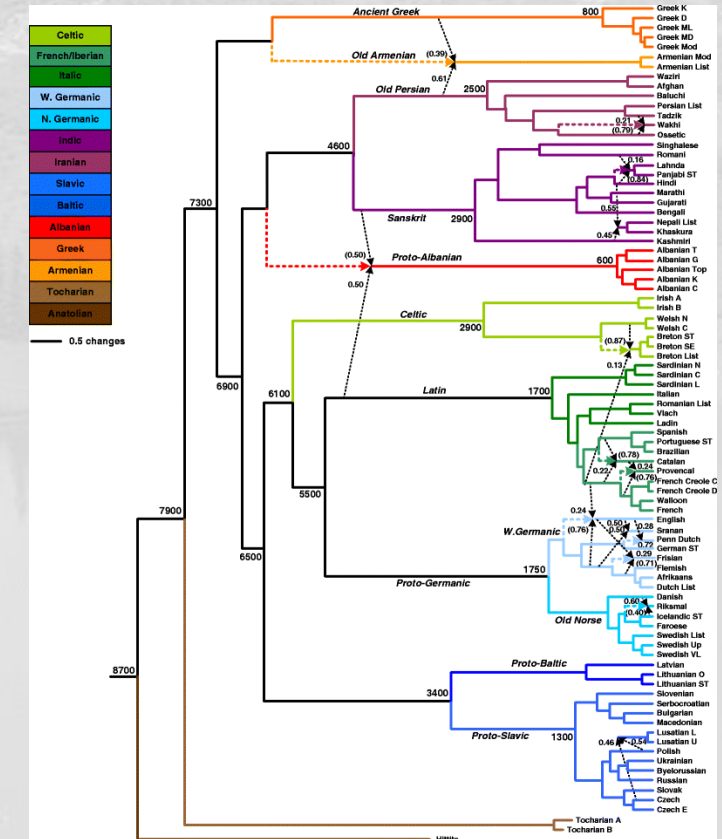
- Network graphs can be more appropriate, especially for very “labile” aspects like phonetics
- Incorporating hybridisation events is becoming possible, and commonly used



Kolipakam et al. 2018



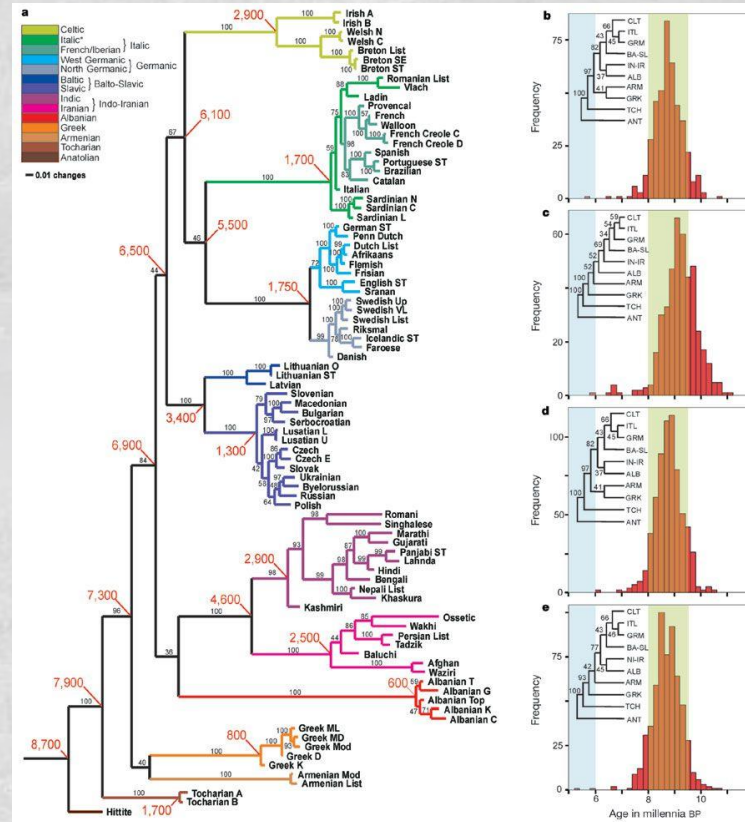
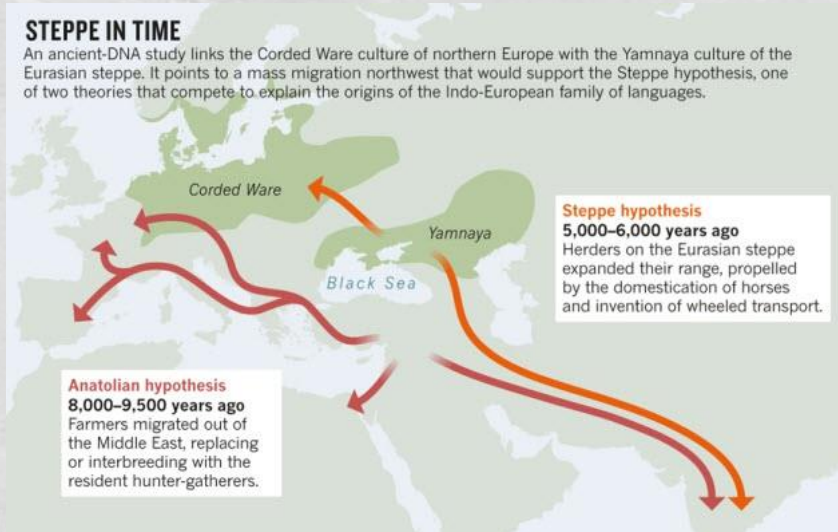
Heggarty et al. 2010



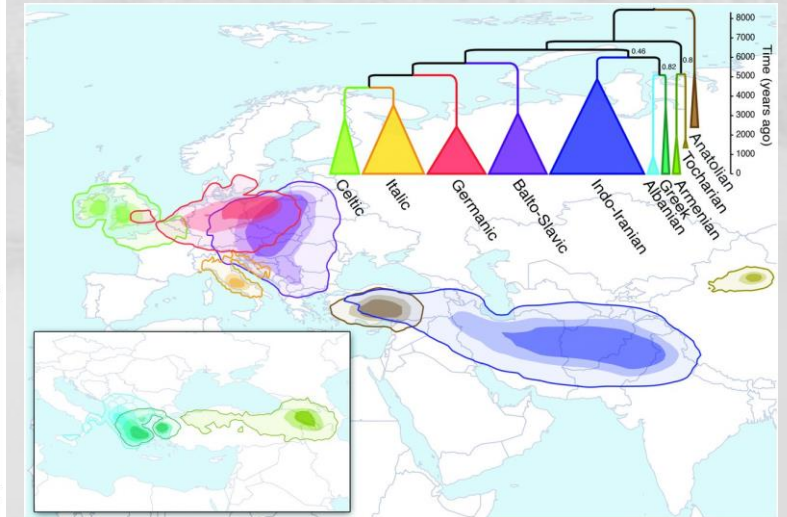
Willems et al. 2016

Applications

- Dating population divergence
- Placing population divergence geographically
- E.g. the Anatolian versus Kurgan hypothesis of agriculture

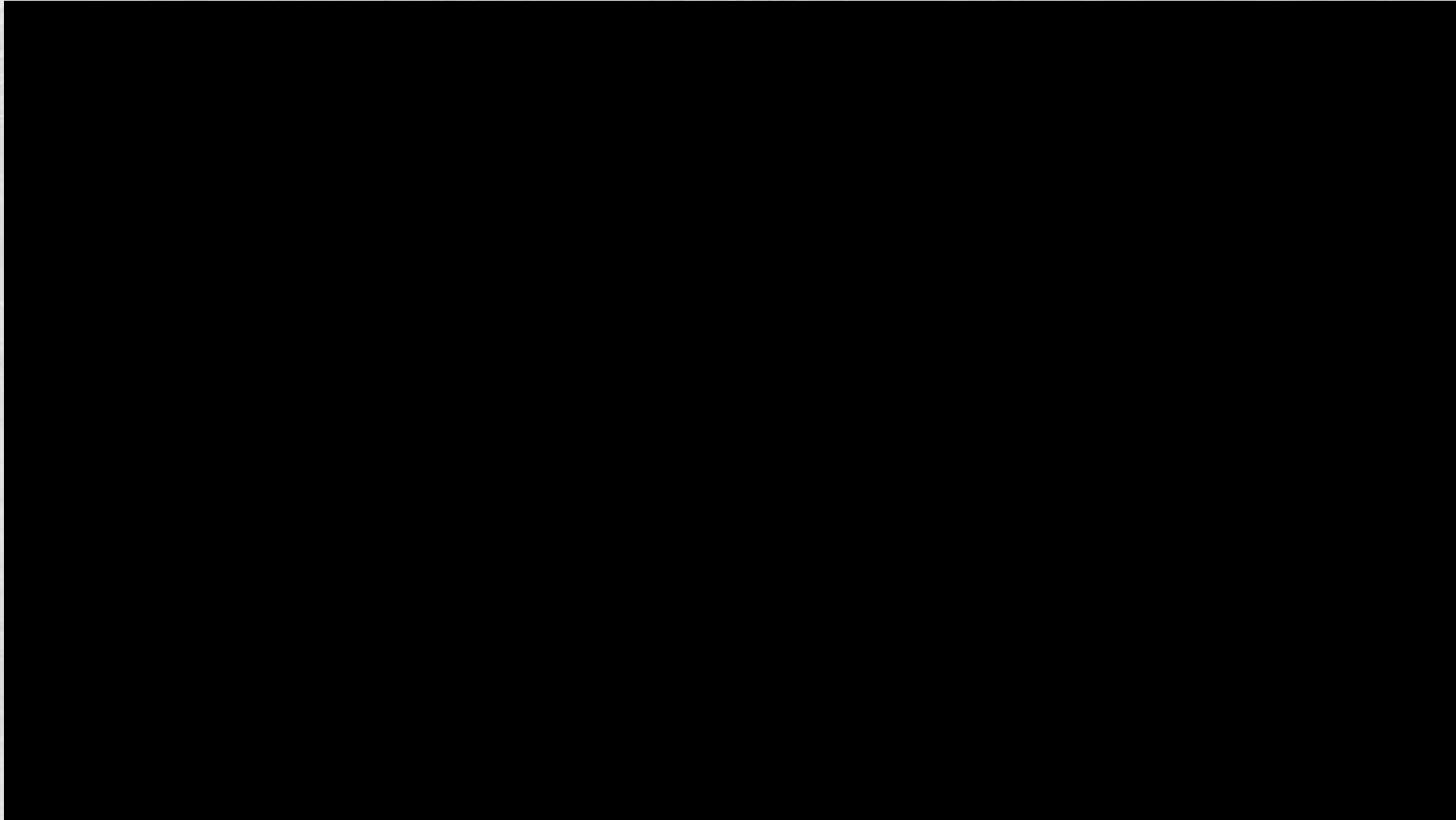


Gray and Atkinson 2003



Bouckaert et al. 2012

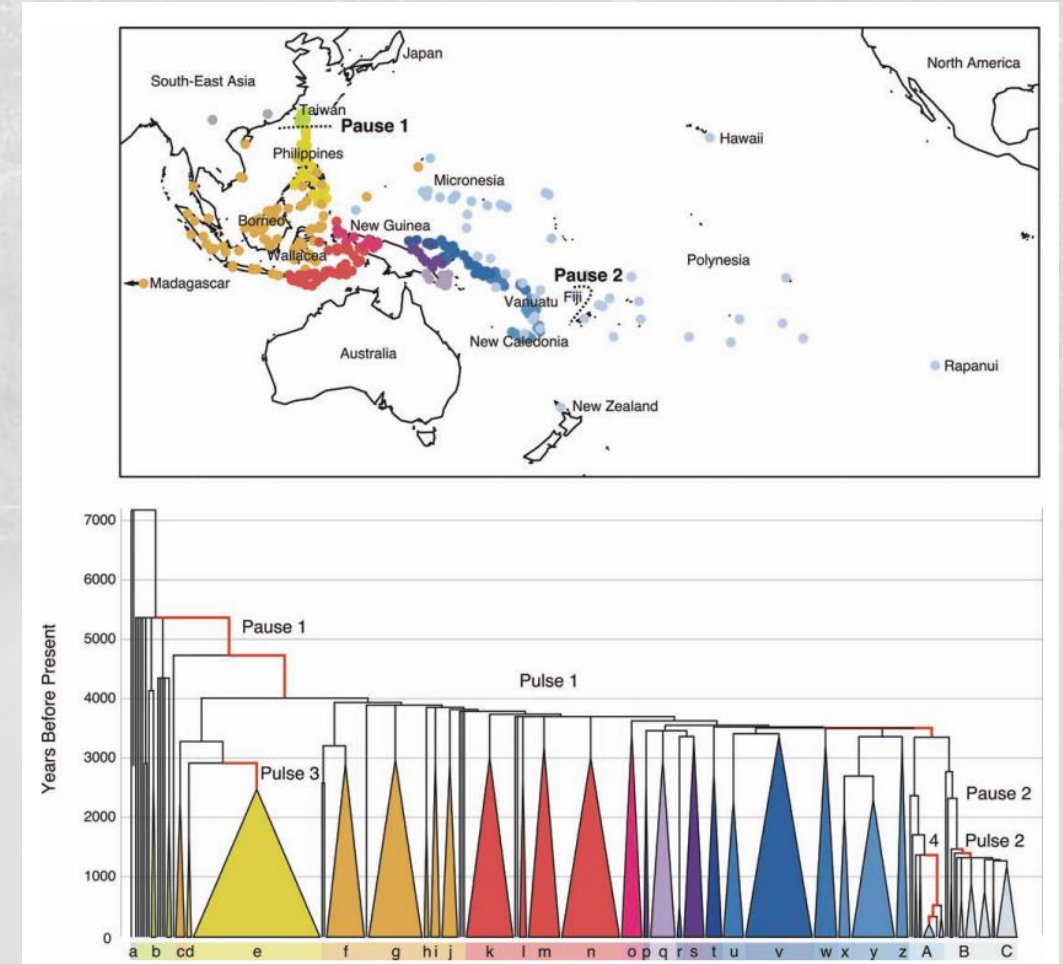
Applications



Bouckaert et al. 2012

Applications

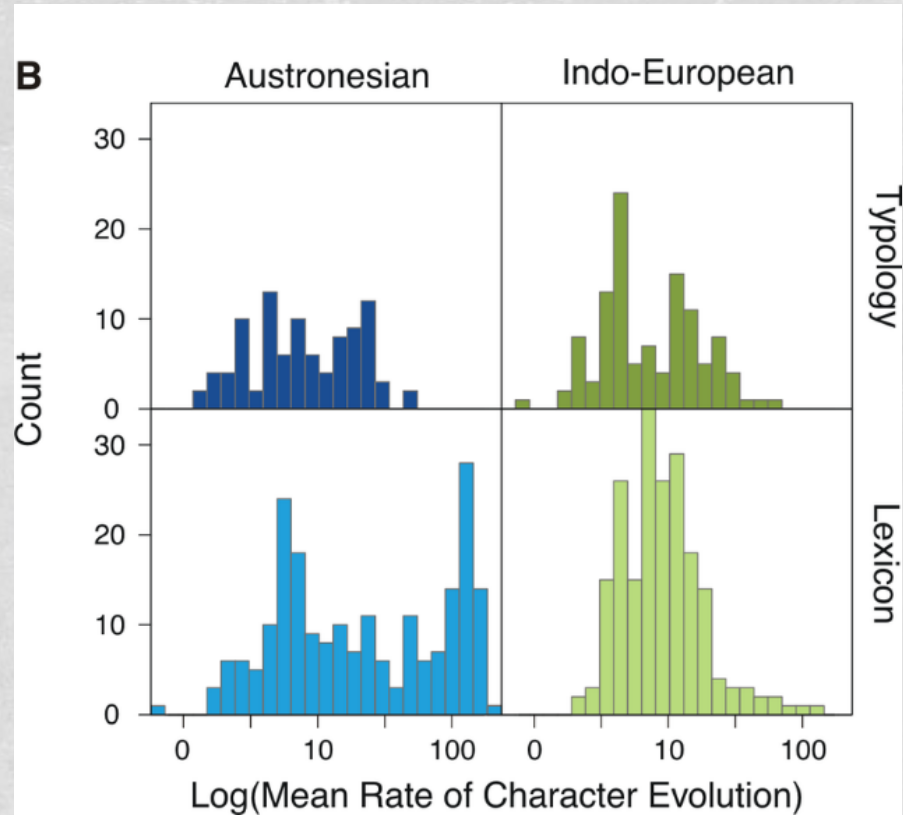
- Another example:
 - Origin and expansion of Austronesian people
 - “Slow boat” from Wallacea v. “pause-pulse” recent from Taiwan



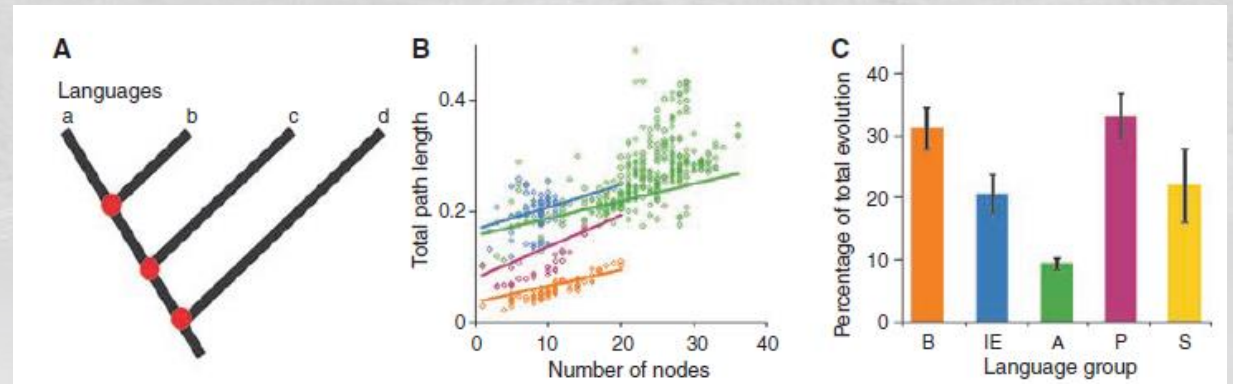
Gray et al. 2009

Applications

- Examining rates of evolution, and how culturolinguistic systems evolve



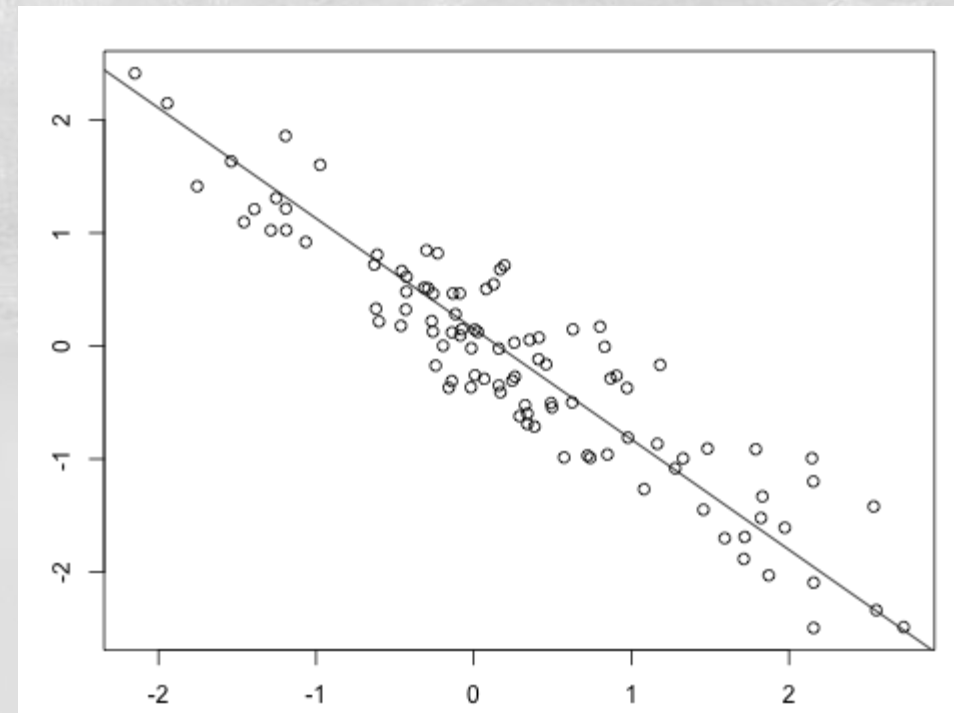
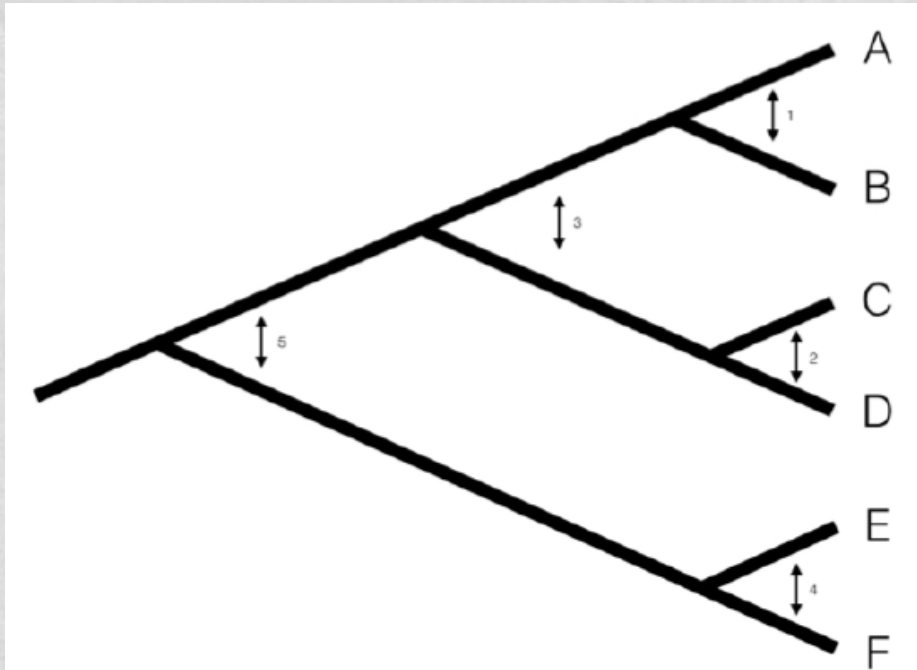
Greenhill et al. 2010



Atkinson et al. 2008

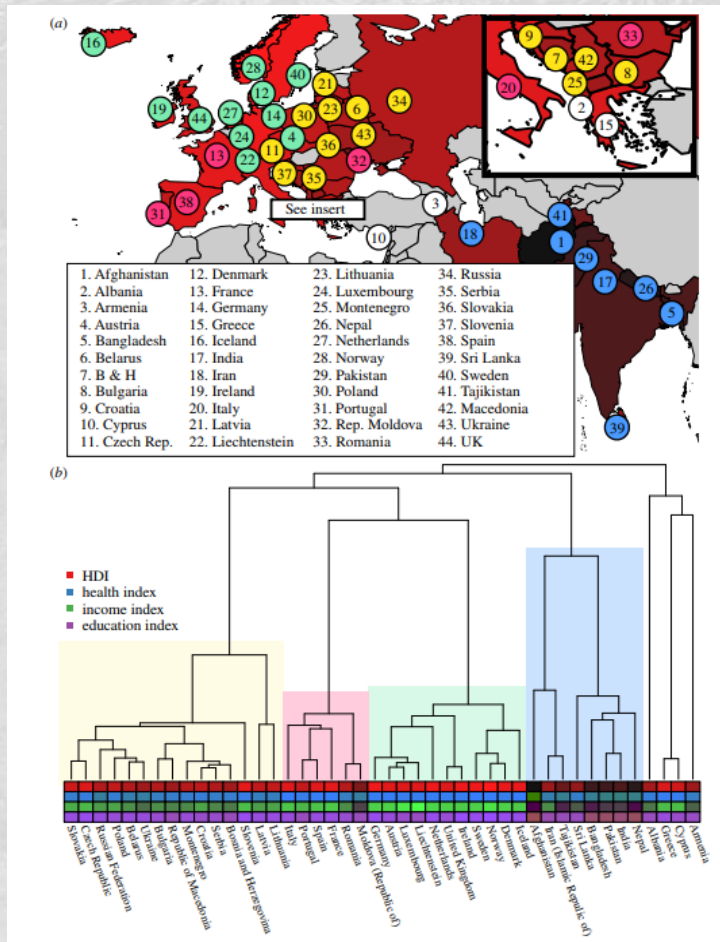
Applications

- Language phylogenies can be used for examining other aspects of cultural evolution – allow **independent contrasts** and **phylogenetically informed regressions**
- Can map any kind of social traits onto phylogeny, and look at how they evolve

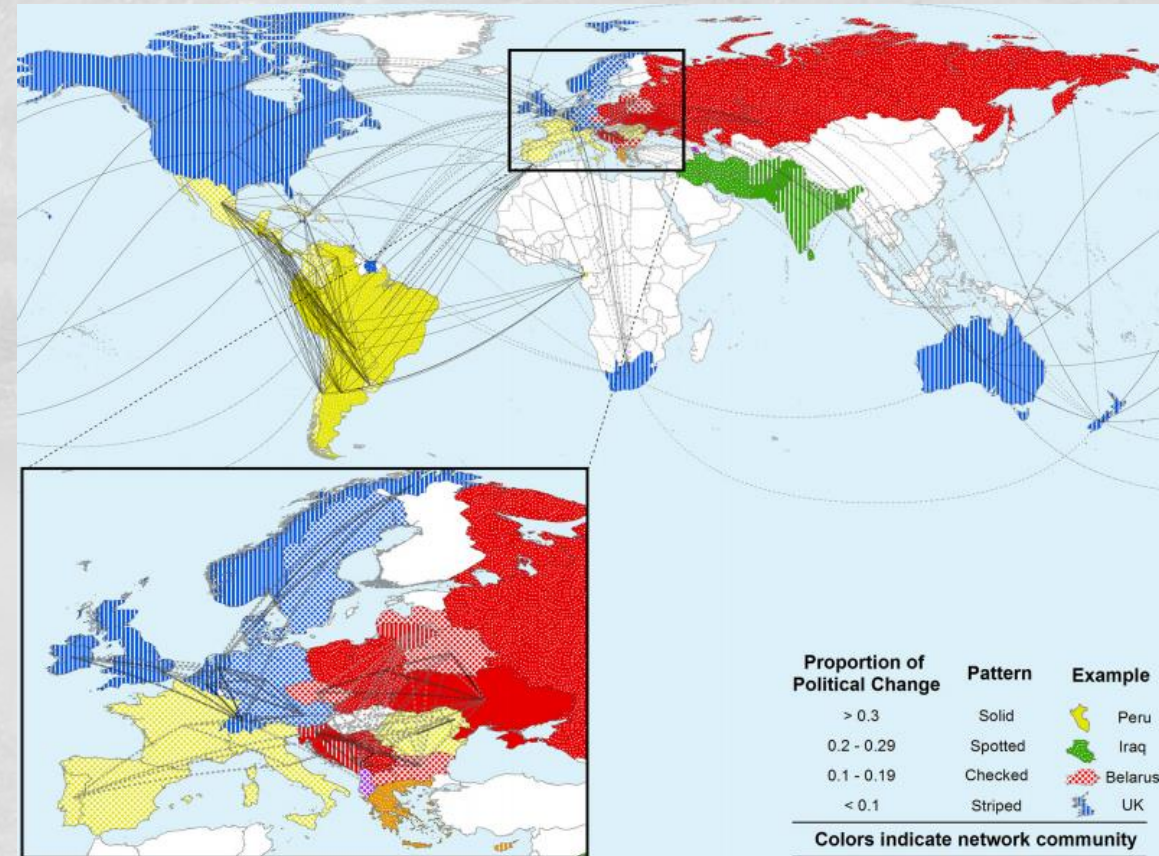


Applications

- Example: economic and social indicators



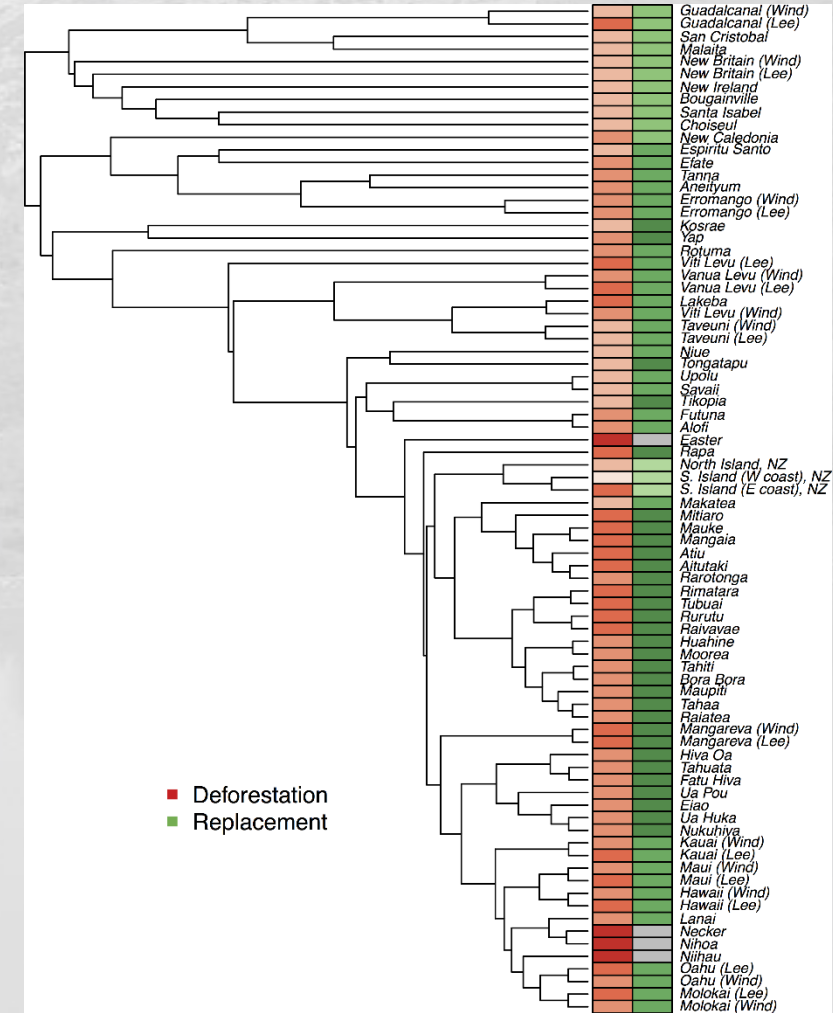
Sookias et al. 2018



Matthews et al. 2016

Applications

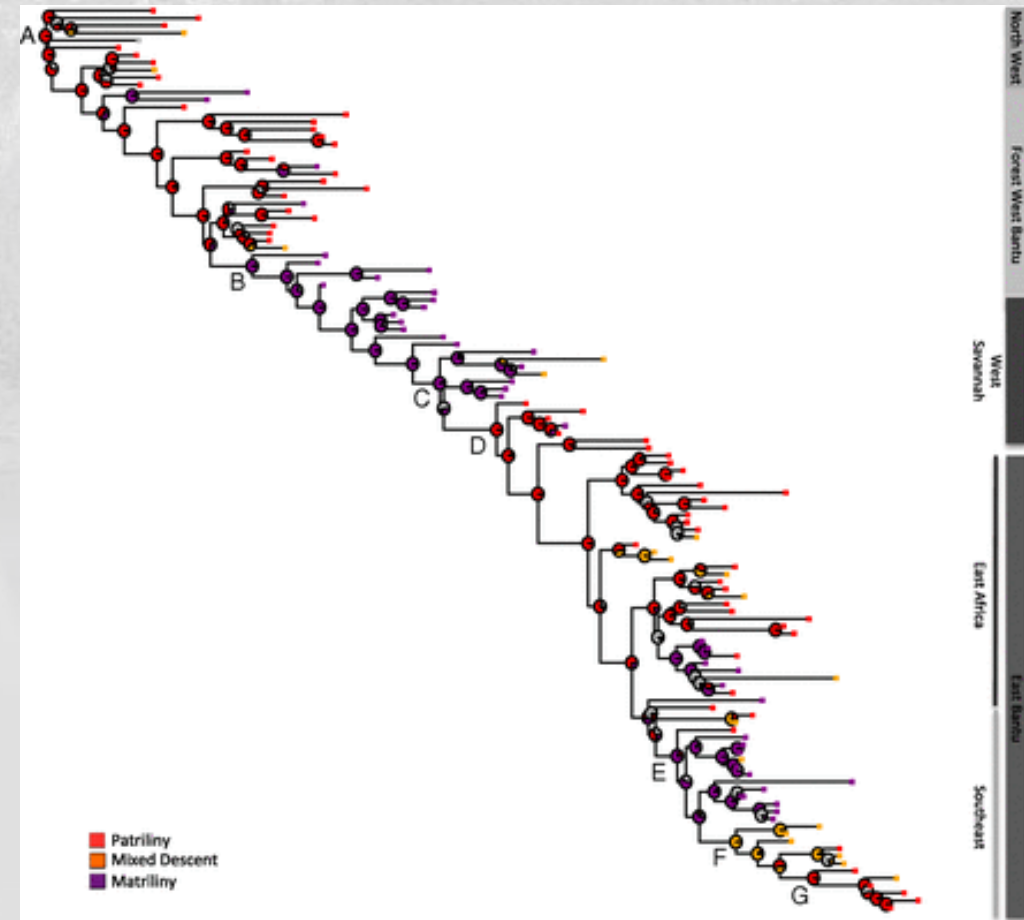
- Further examples: forest use patterns in polynesia



Atkinson et al. 2016

Applications

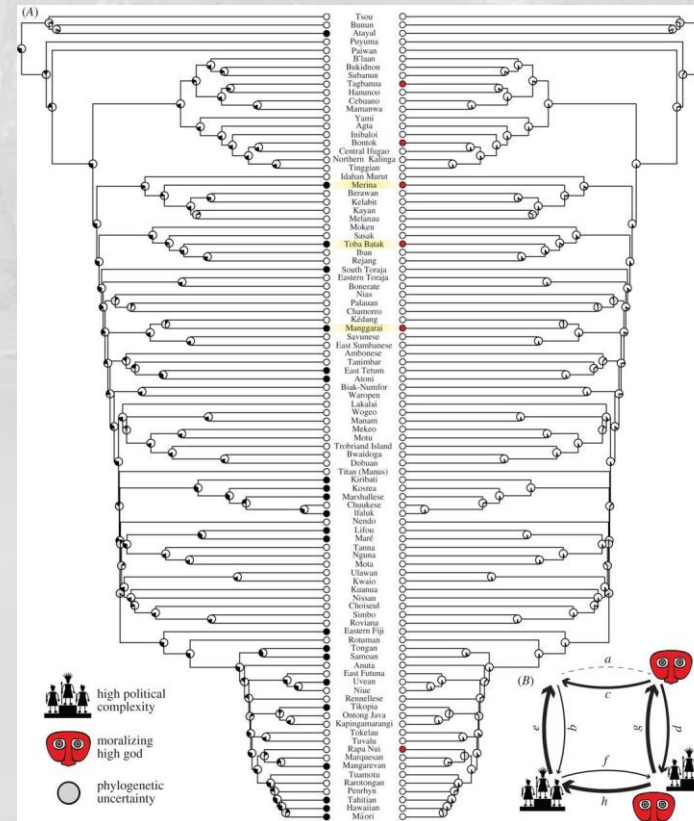
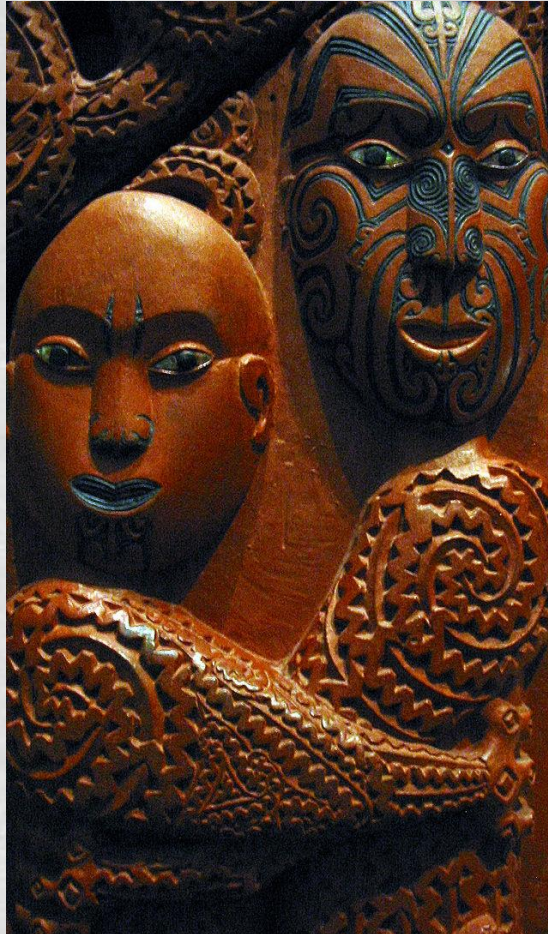
- Further examples: how traditional marriage residence rules (e.g. matri/patrilocal) affect other social structures



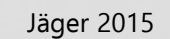
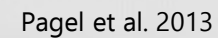
Opie et al. 2014

Applications

- Further examples: do you need high “moralizing high gods” for complex societies? No...

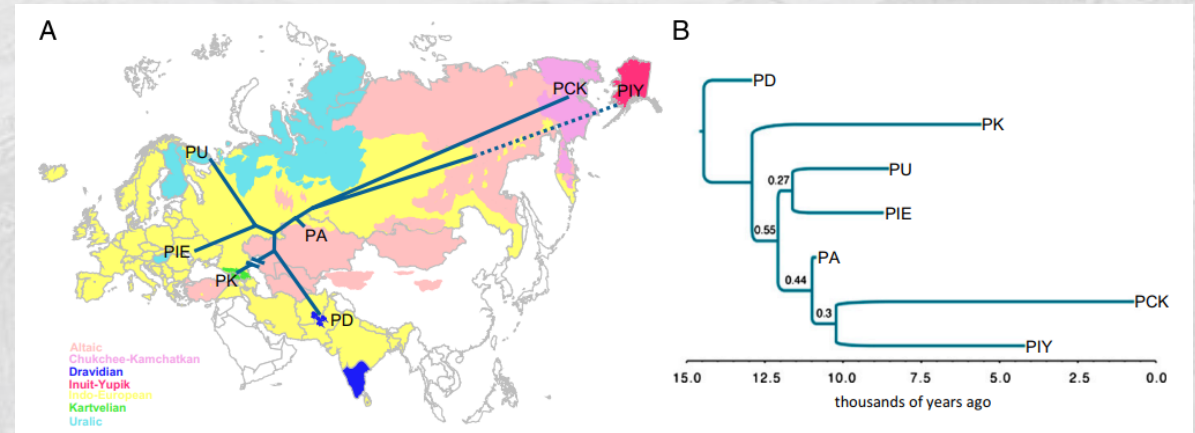


- Deep linguistic phylogeny
- Are certain features more strongly conserved?
- Mass lexical comparison

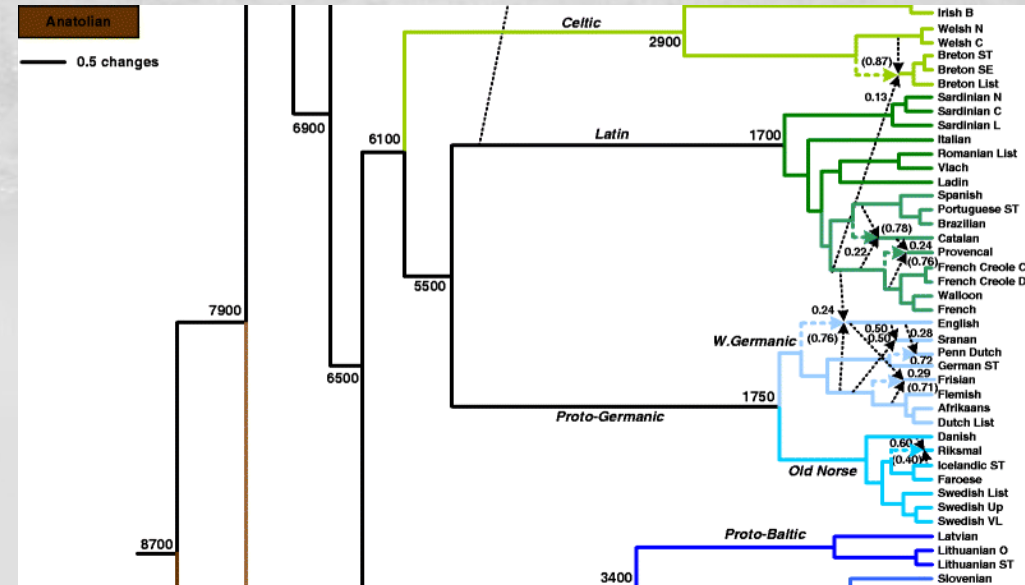


Ways forward

- Expanding knowledge of deep phylogeny
- Automatic cognate judgement – AI?
- Broad incorporation of hybridisation in models
- More and more varied questions!



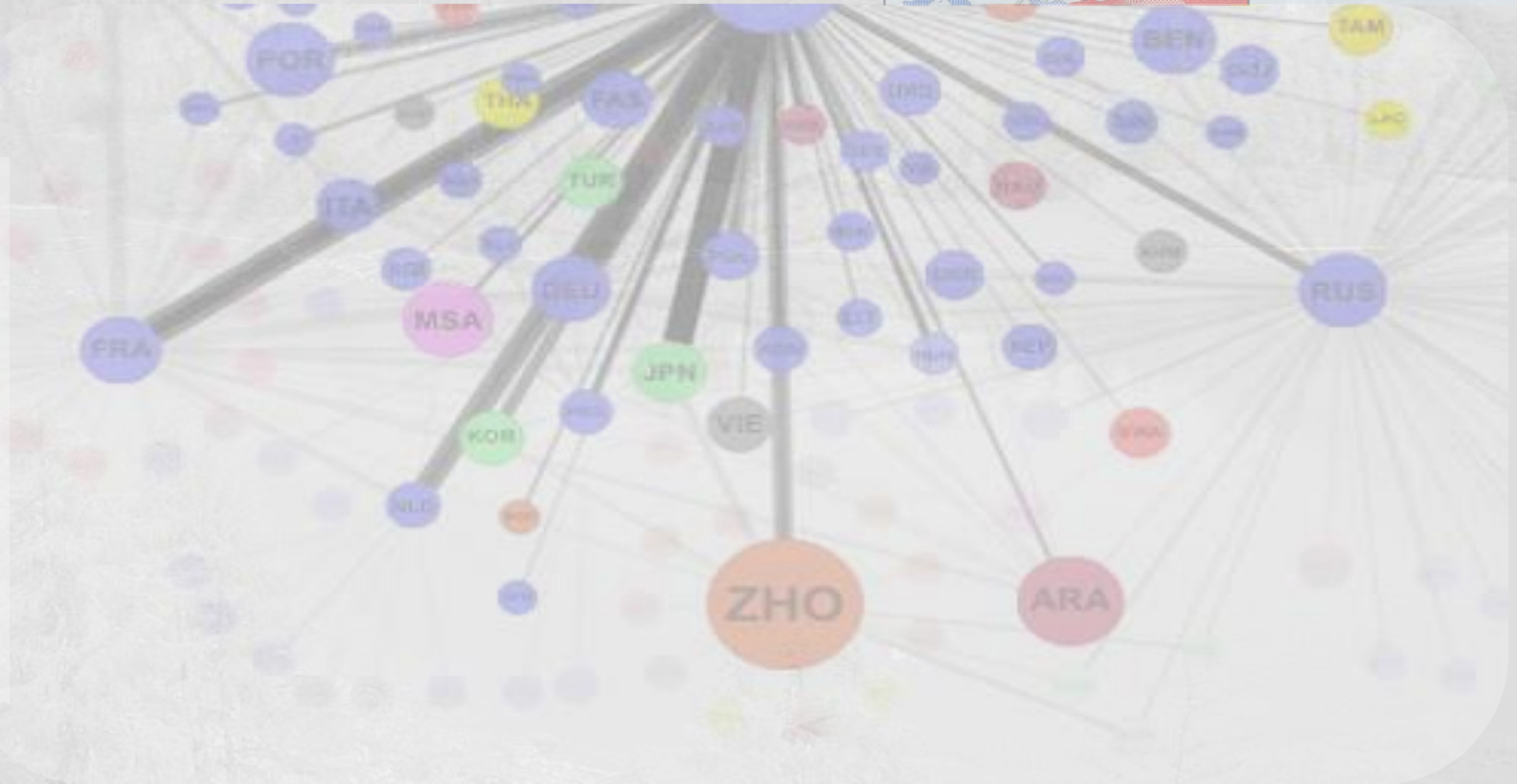
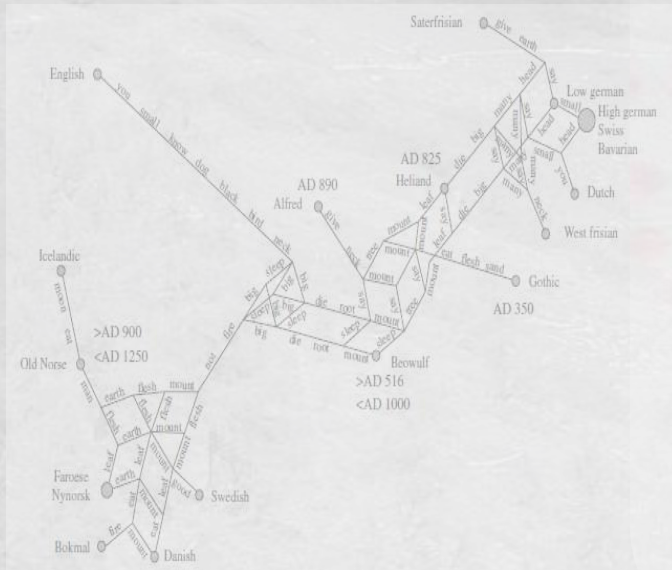
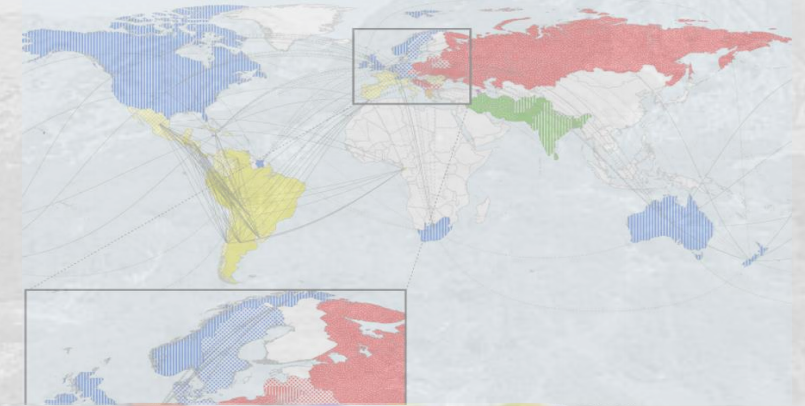
Pagel et al. 2013



Willems et al. 2016

Take-aways

- Language and culture (also non-human) are evolving systems, at least partially inherited
- Very similar methods and concepts to biological evolution
- Huge number of different questions can be addressed
- **Be careful** – can be very sensitive, and know your topic!

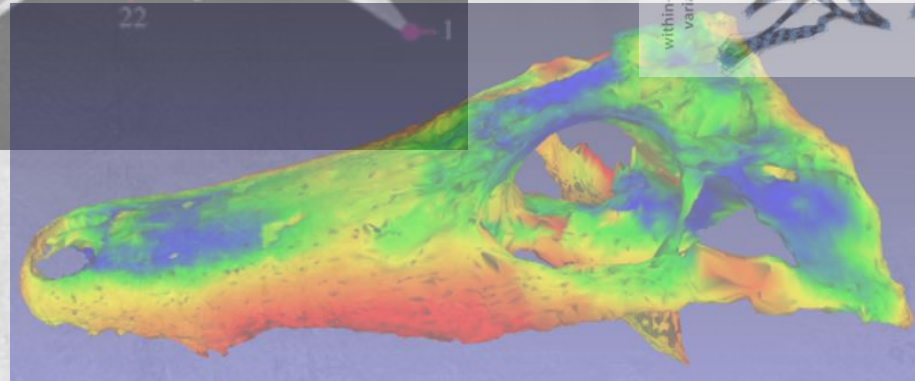
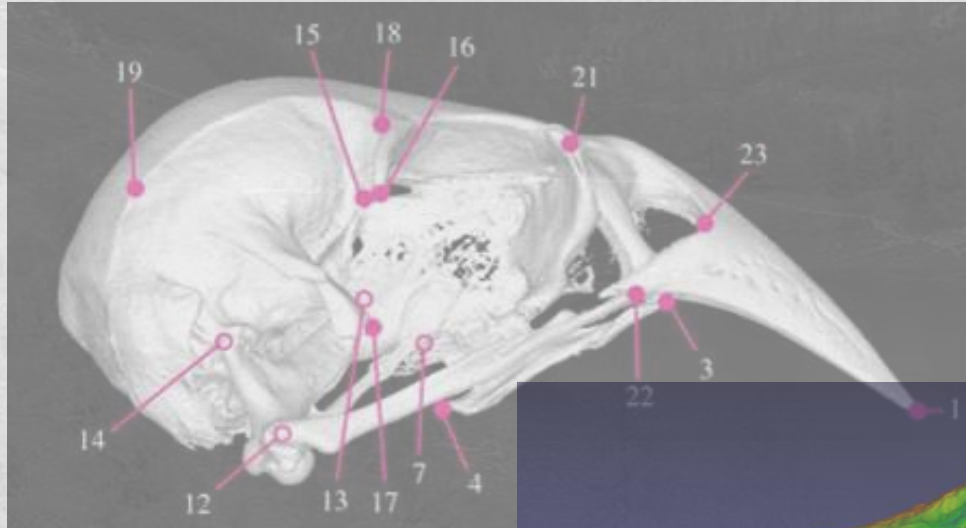


A grayscale photograph of a mountain landscape. In the foreground, a calm lake reflects the surrounding scenery. The middle ground is filled with dense, dark forest covering the lower slopes of the mountains. In the background, rugged, rocky mountain peaks rise against a light sky. The overall tone is somber and naturalistic.

Phylogeny using continuous data

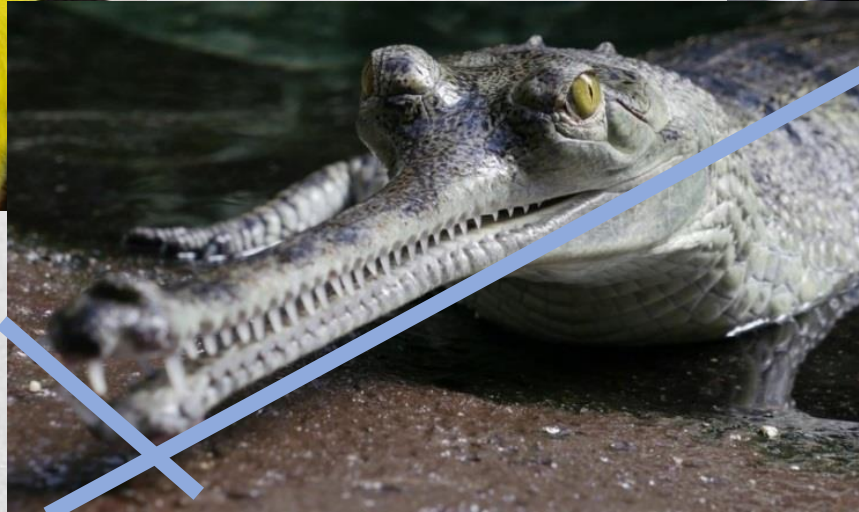
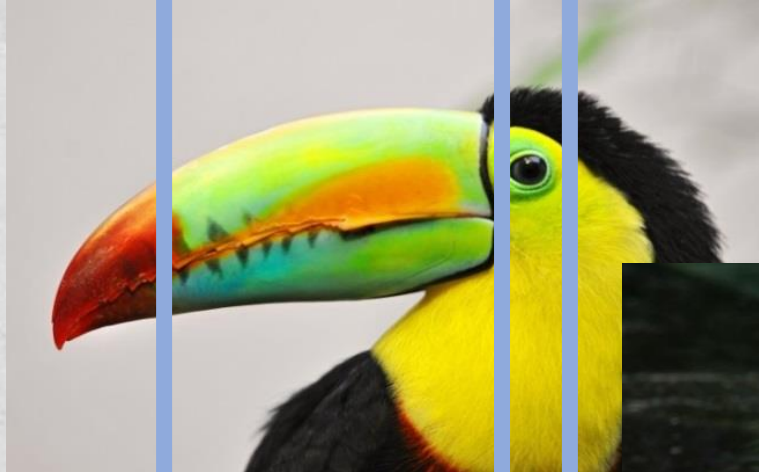
Plan

- Why continuous data?
- What is continuous data and geometric morphometric data?
- Methods of using this for phylogeny
- The effect of ecomorphology



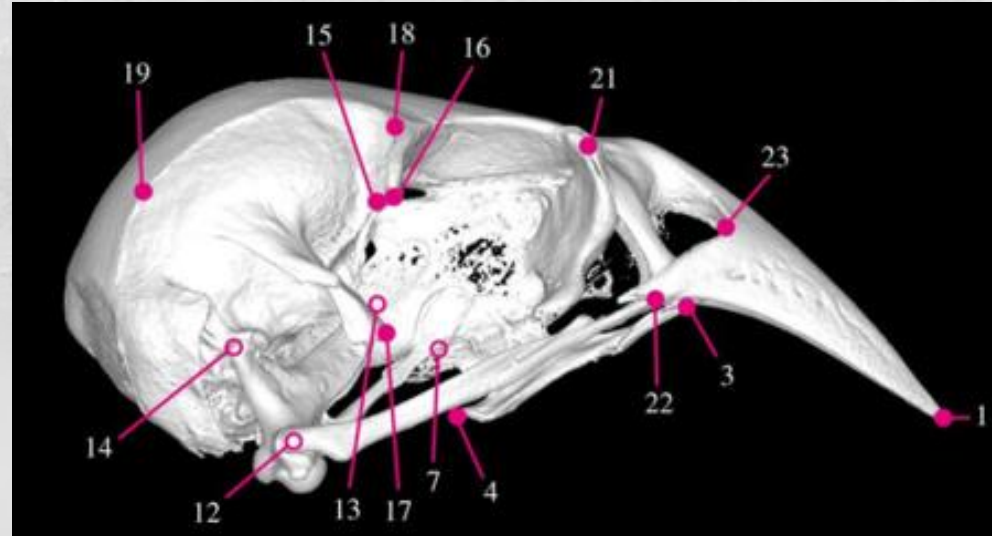
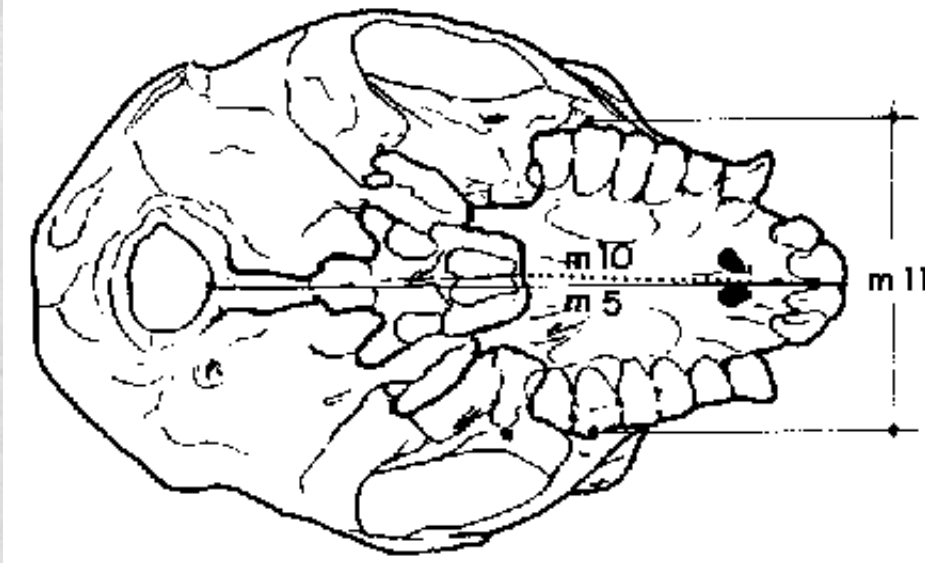
Why use continuous data for phylogeny?

- Life is not discrete!
- Continuous variation, and morphology especially is not easily delimited
- Manual delimitation is often very subjective
- Manual delimitation is very time-consuming



Continuous data: what is it?

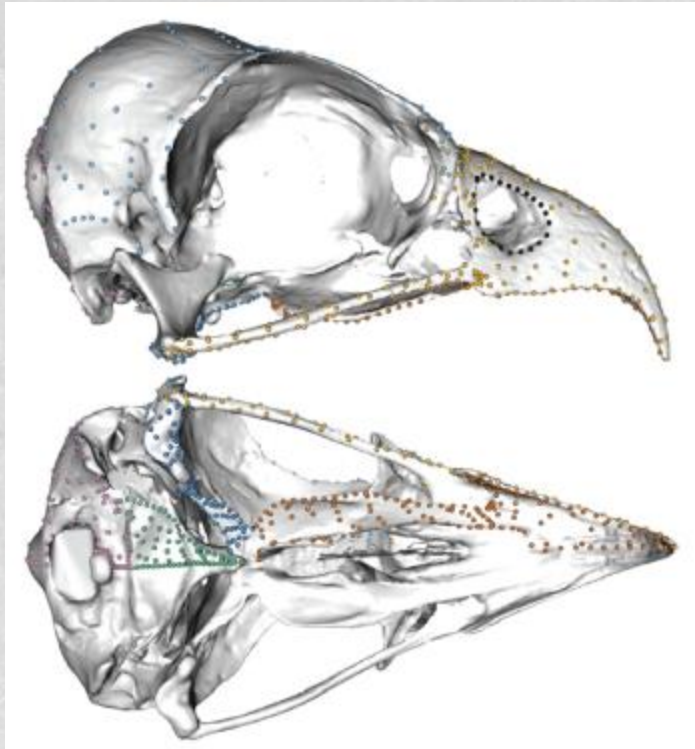
- Any measurement data (=standard morphometrics)
- Also geometric morphometric (GMM) data (=coordinate data)



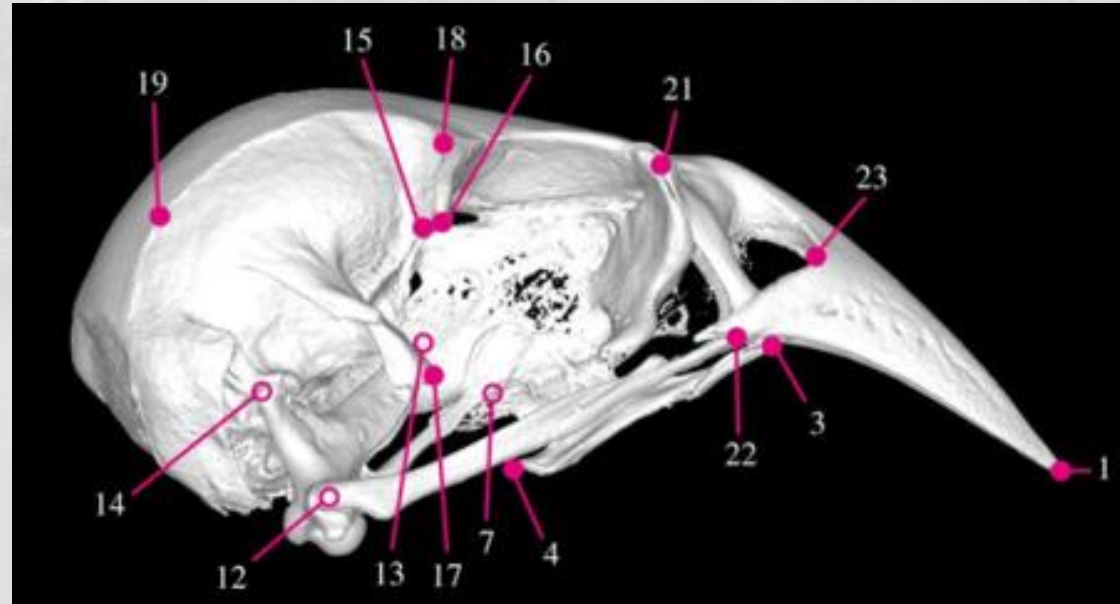
Abzhanov & James 2016

Continuous data: geometric morphometrics

- Typical way of capturing shape variation - a lecture in itself!
- Points put on homologous or functionally homologous points in 2D or 3D
- Semilandmark series along sutures, edges
- Semilandmark patches across surfaces
- Can capture more of overall shape variation than traditional morphometrics, and easier to undertake



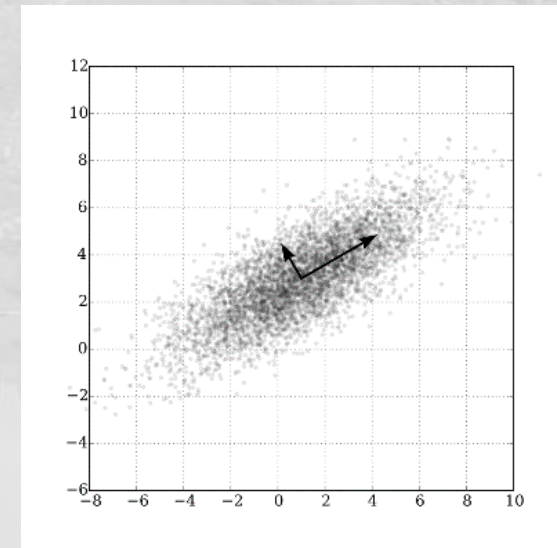
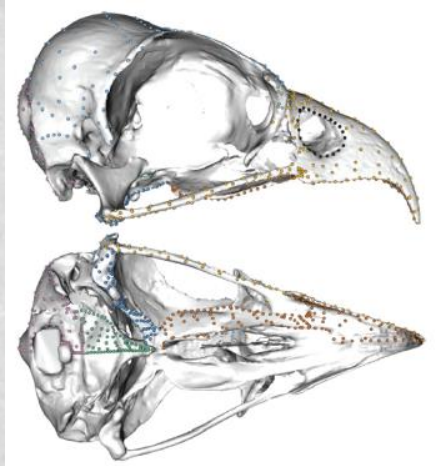
Felice & Goswami 2018



Abzhanov & James 2016

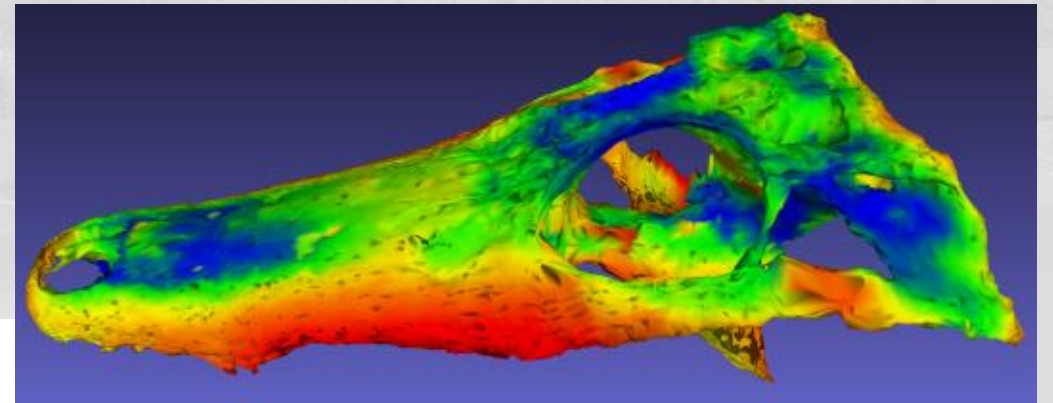
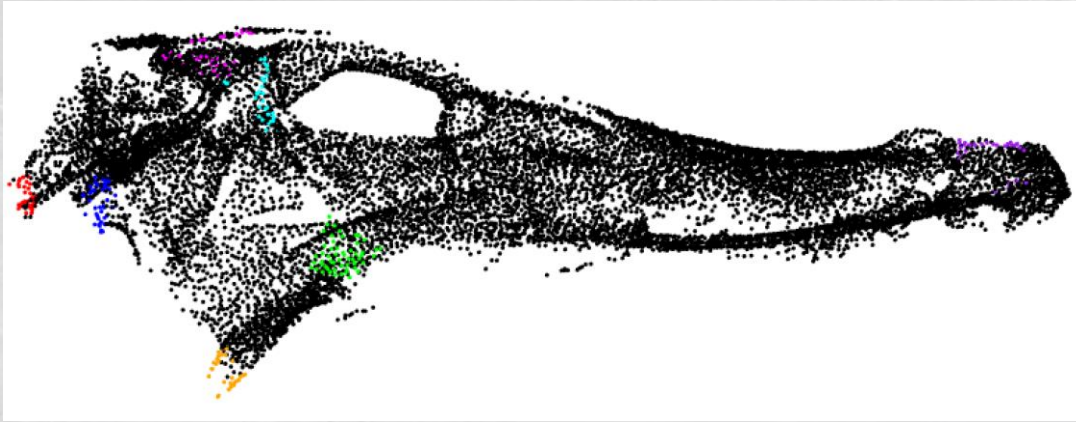
Continuous data: geometric morphometrics

- Points rotated and scaled (Procrustes analysis) to fit mean shape
- Often ordinated (PCA) subsequently, concentrating variation into first few axes



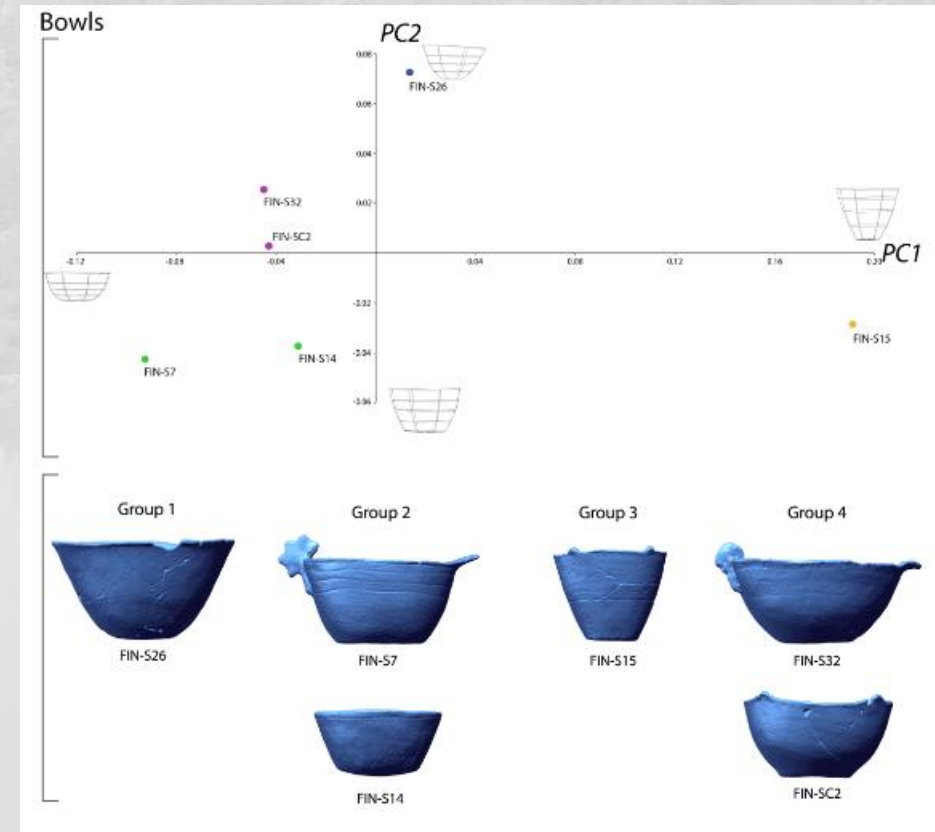
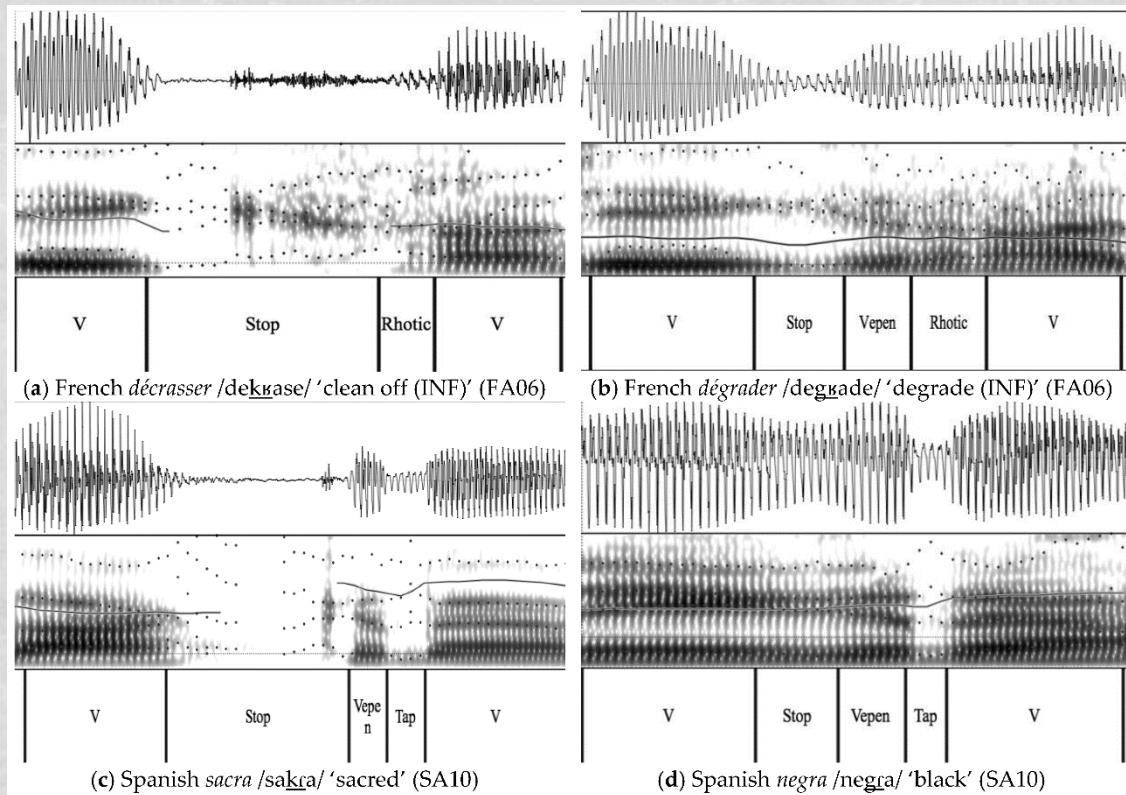
Continuous data: whole surface data

- Subset of GMM data
- Generalized procrustes surface analysis (Pomidor 2016)
- Rotates and scales 3D shapes appropriately, then places landmarks across entire surface automatically
- Data can be used raw or ordinated just like other GMM data



Continuous data: culturolinguistic data

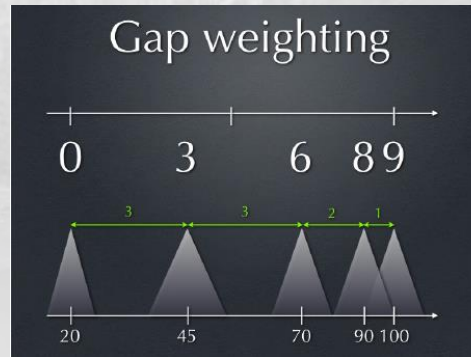
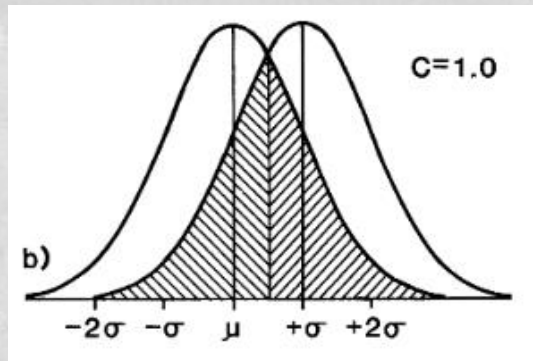
- Phonemes
- Human development indices
- Pot shapes



Selden et al. 2014

Methods: automatic discretization

- Several methods developed based on sample variance
 - Gap coding: new state if separation between means greater than one standard deviation
 - Gap weighting: also weights *size* of gap
 - Not well tested (e.g. against known or molecular phylogeny)
 - Not further discussed here, but would be interesting to test further



- Arbitrary discretization into certain number of states, e.g. maximum number of states in programme



- Continuous characters can be analyzed directly using distance-based, maximum likelihood or Bayesian methods
- Some evidence that this can be successful



■ *Pan paniscus*
 △ *Pan troglodytes*
 ◆ *Gorilla gorilla*
 + *Pongo pygmaeus*
 ◇ *Homo sapiens*

Principal compo

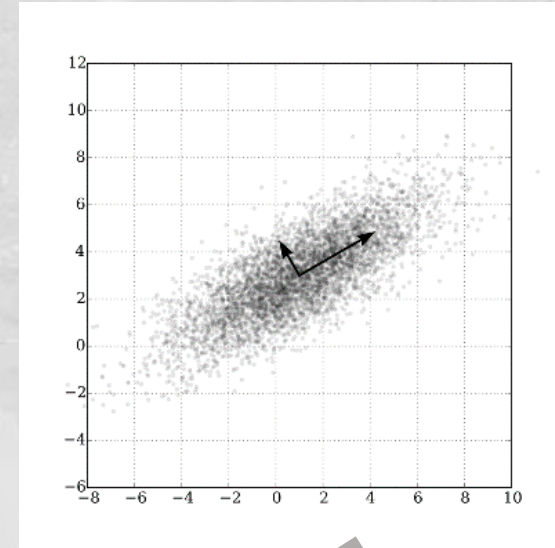
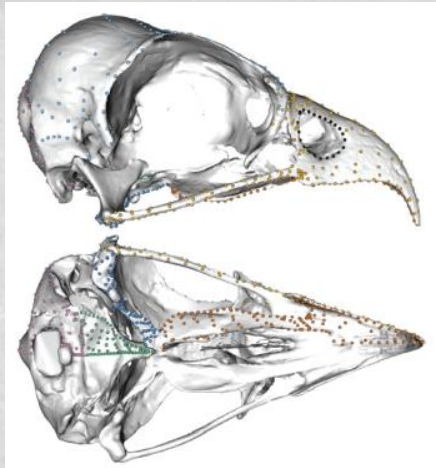
Principal component 1

[illegible]

LIÈGE
université

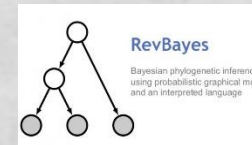
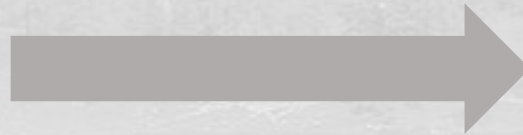
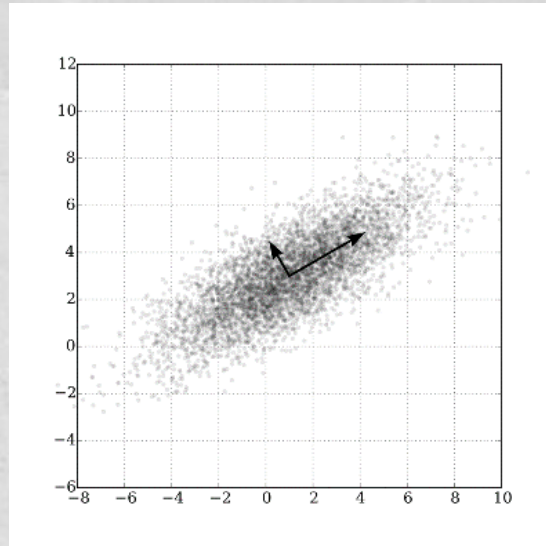
Methods: geometric morphometrics

- Either usable following Procrustes (coordinates still) or ordination (specimen values)



Methods: GMM ordination axis values as characters

- Ordination axis values for taxa can be used as continuous characters
- Should perhaps be weighted by variance on that axis

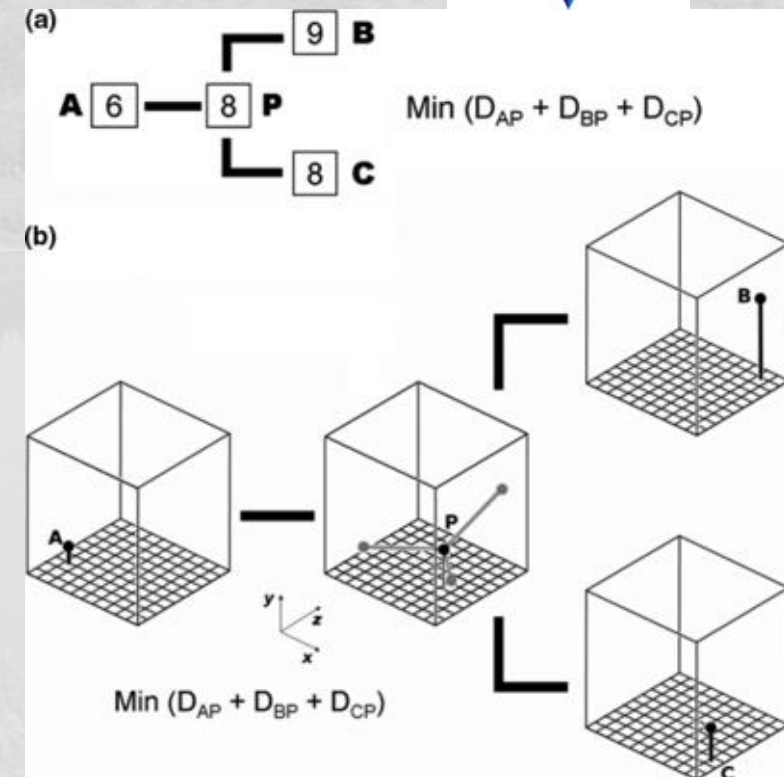
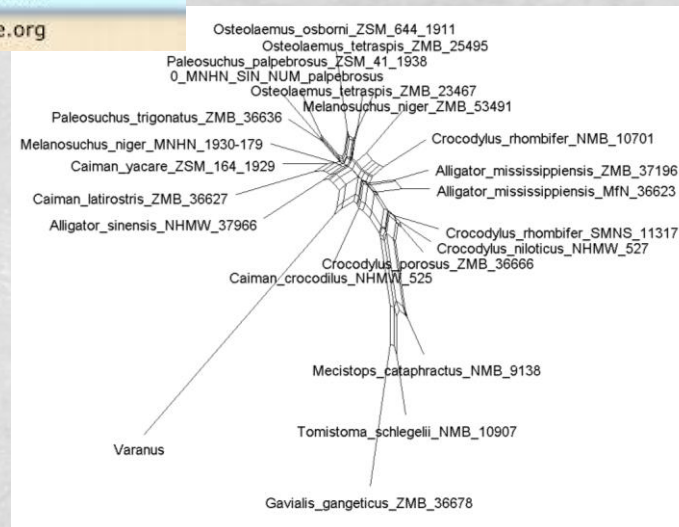
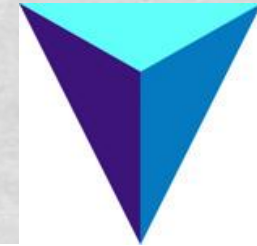


RAxML



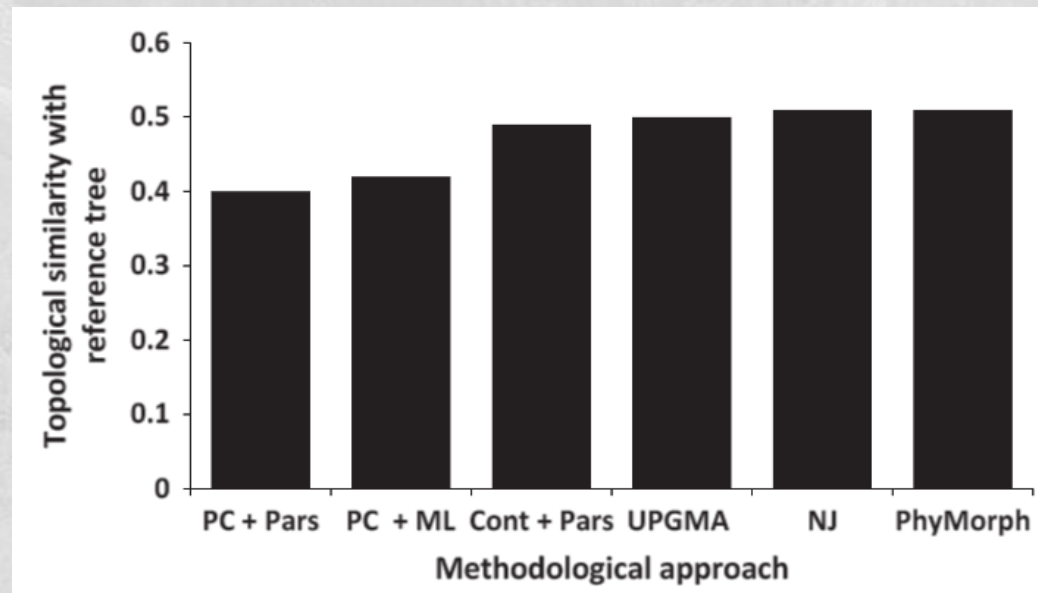
Methods: raw GMM data (after Procrustes)

- Distance-based methods (neighbour joining, UPGMA etc.)
- „Phylogenetic morphometrics“ (Catalano et al. 2010), directly implemented in TNT
 - Tree with minimum distances between ancestor-descendant points
 - Analogous to Farris optimization in parsimony
 - Some controversy methodologically

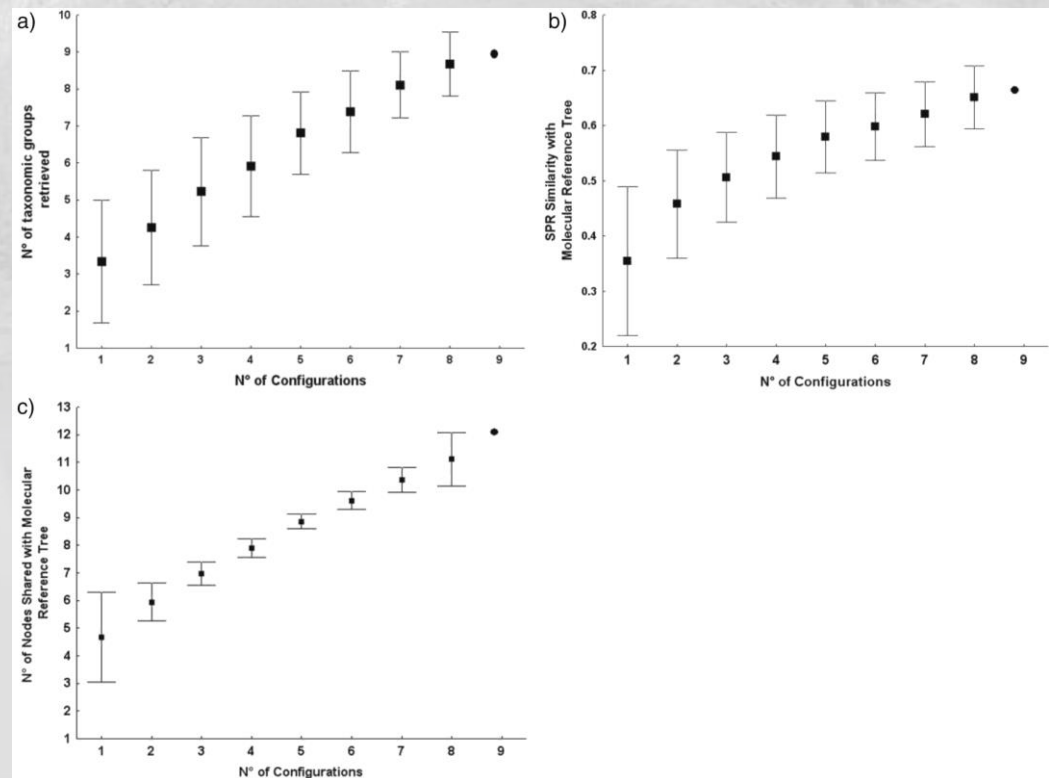


Methods: what's best for GMM data?

- For morphology, raw data seem better than PC values
- Neighbour joining almost as good (and much quicker) than phylogenetic morphometrics
- Accuracy increases with more skeletal elements



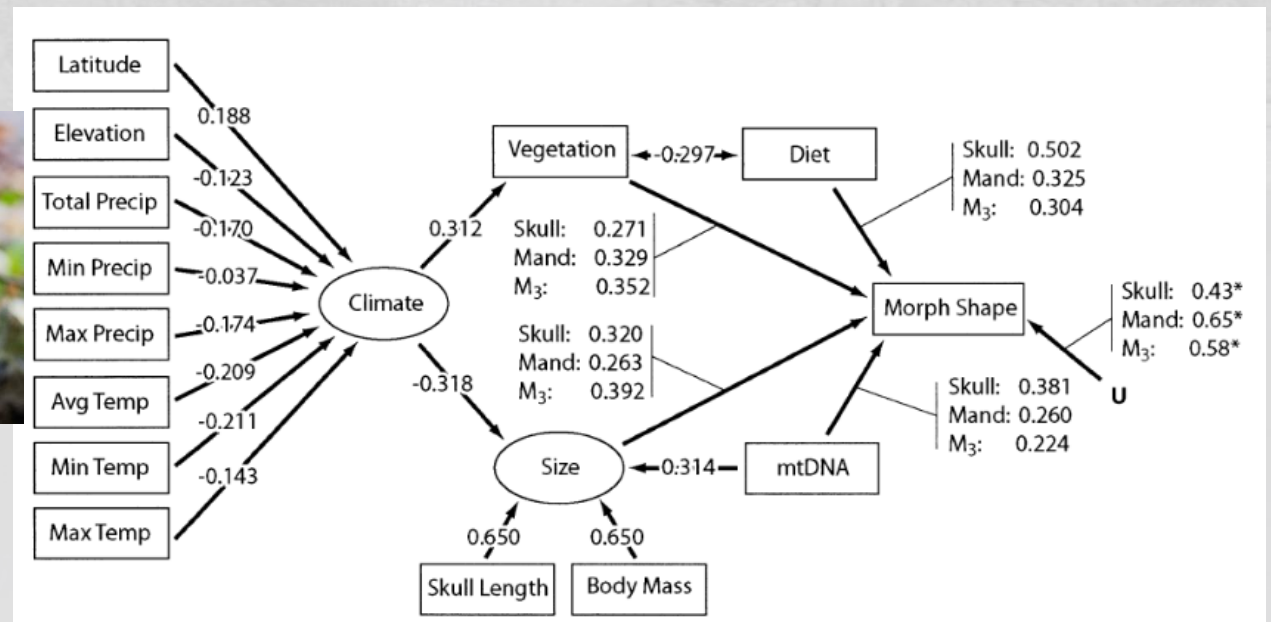
Catalano and Torres 2016



Catalano et al. 2016

Ecomorphological signal

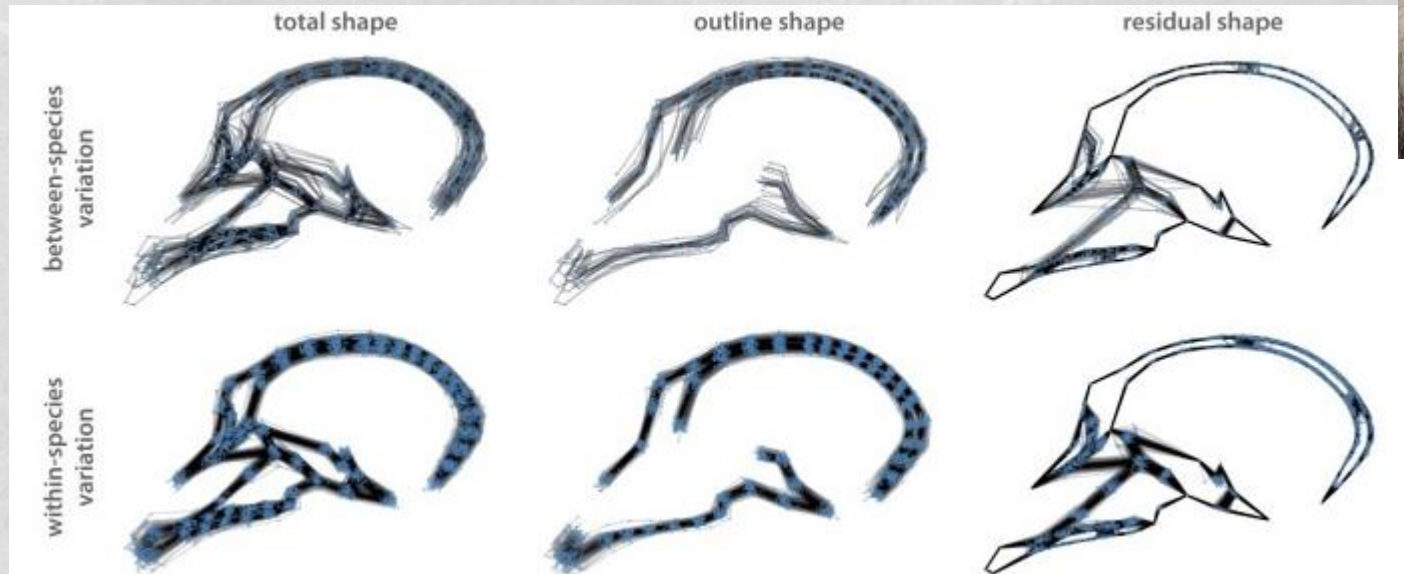
- Is a major part of continuous variation
- Can strongly confound phylogeny



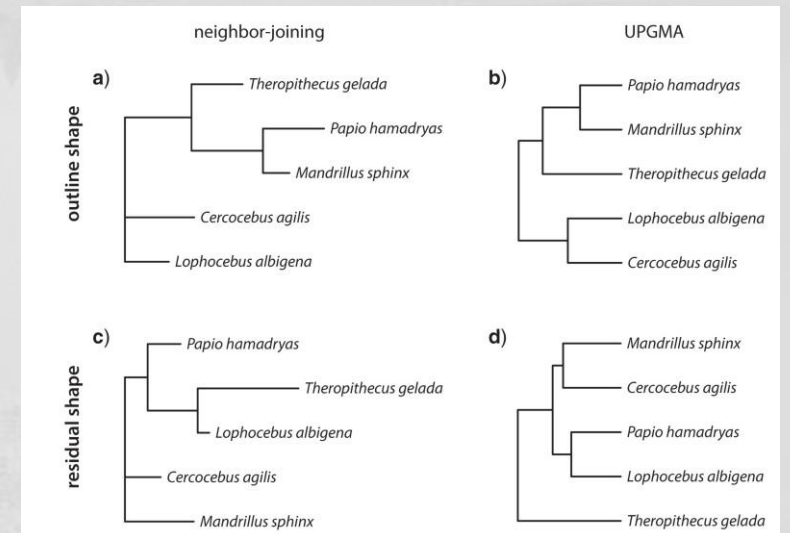
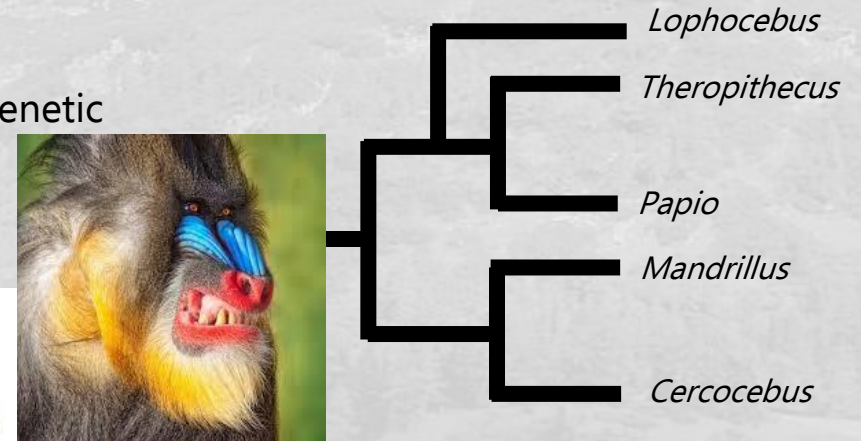
Caumul and Polly 2005

Removing ecomorphological signal?

- Regress out habitat or behaviour?
- Look at specific aspects of shape?
- Grunstra et al. (2021) take out overall (=outline) shape
- Remaining compositional/structural and local shape show stronger phylogenetic signal and yield better tree

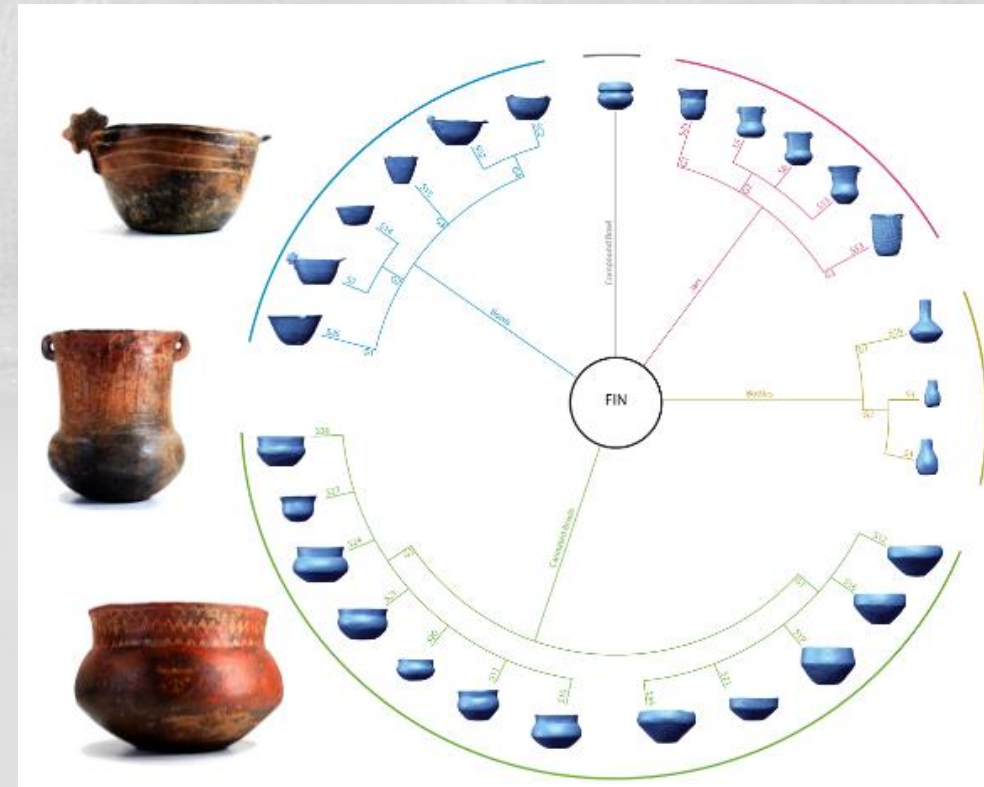
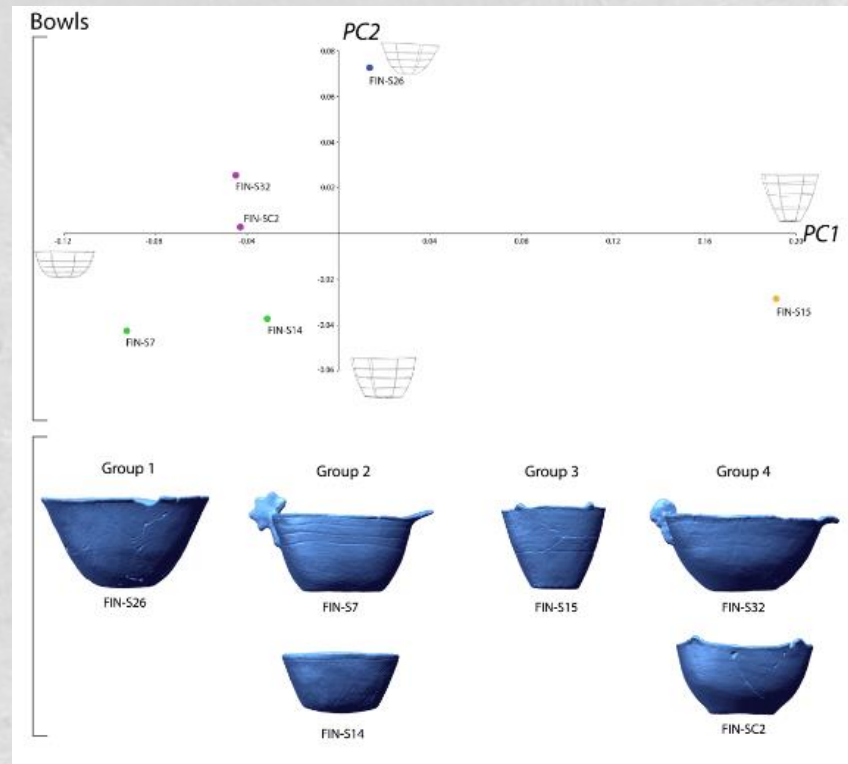


Grunstra et al. 2021



Continuous cultural data

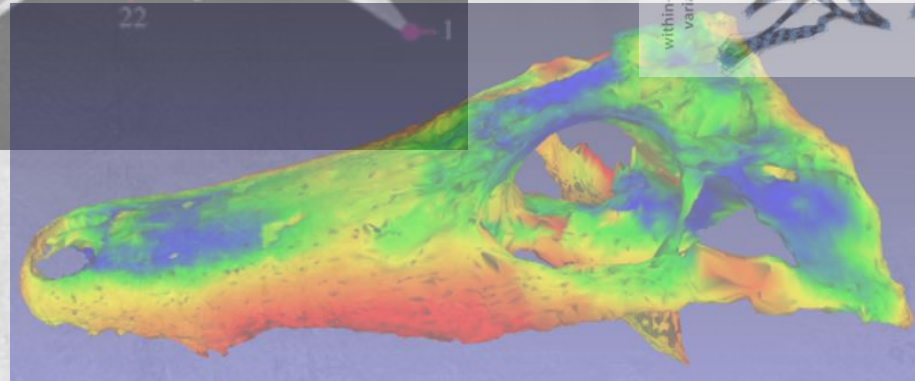
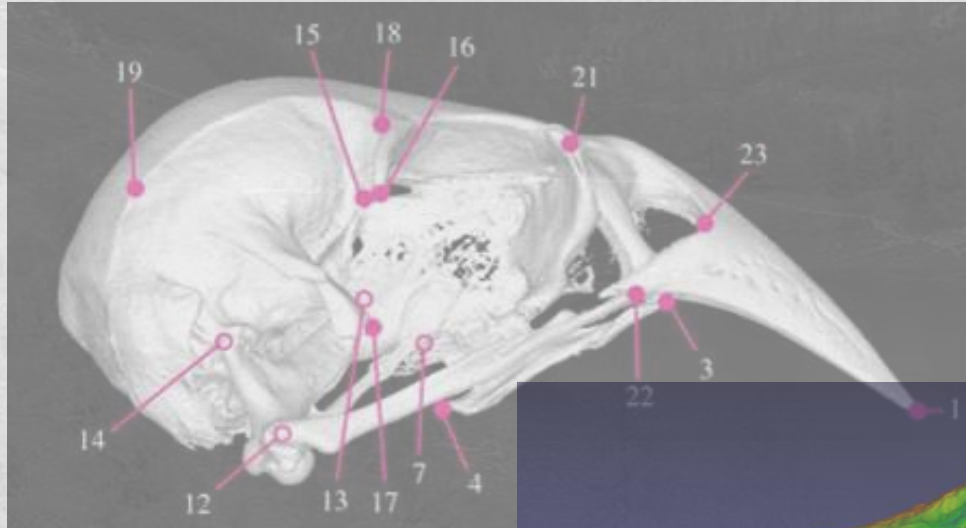
- Little explored, but all same approaches should apply...
- However: very labile and high hybridisation, so may not be as useful if phylogenetic signal is the main interest



Selden et al. 2014

Take-aways

- Continuous data can be used both directly and through discretization
- GMM data can be used directly and after ordination, and has shown some promise
- Ecological signal is very strong, but can possibly be removed by looking at local/structural shape
- **Caution:** the field is still new, and attracted much controversy previously due to theoretical concerns over homology



Any questions?

Also welcome to email later!

sookias.r.b@gmail.com

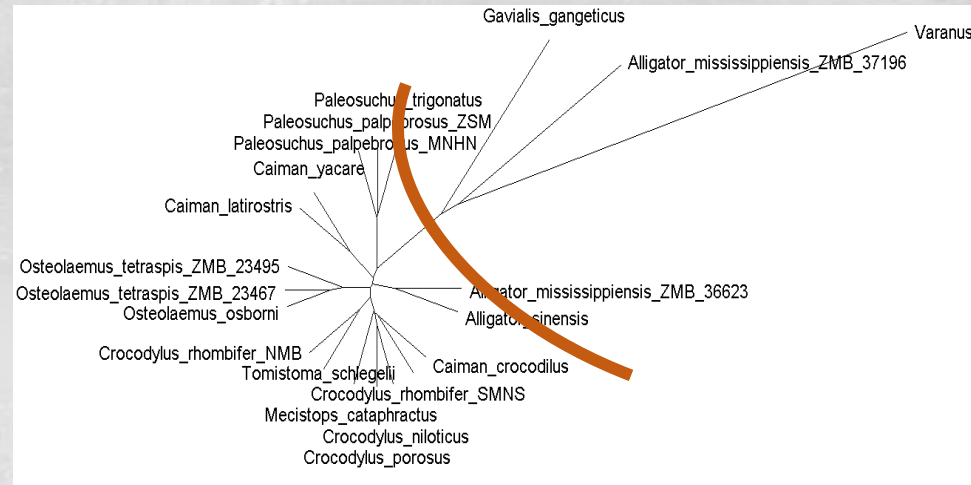
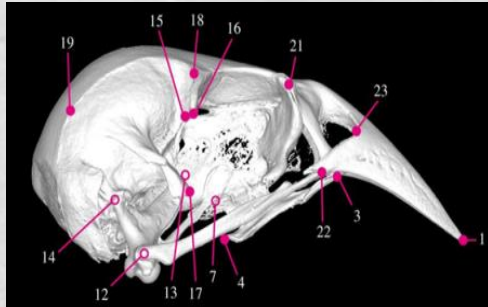
rsookias.info/teaching-resources



EXTRA SLIDES

Discretization from distance trees

- Recently attempted approach (Celik, Phillips)
- Make distance tree from GM data for one structure
- Basically draw line between major branches to score character



0/1