H.F.R.I.
Hellenic Foundation for Research & Innovation

ARISTOTLE UNIVERSITY OF THESSALONIKI

University of Crete

Exploring the morphosyntactic characters of WALS for the phylogenetic reconstruction of languages

Dimitris Michelioudakis, Manolis Ladoukakis, Pavlos Pavlidis,
Athanasios Michail Ramadanidis, Maria-Margarita Makri & Elena Anagnostopoulou

ModelGloss

Aristotle University of Thessaloniki

Friday, May 14th 2021

This is a presentation of first results of the H.F.R.I.-funded project Modeling Glossogeny

Objectives:

- To investigate language change with tools and methods inspired by evolutionary biology.
- To employ similarities and differences in the syntax and morphology of languages in order to probe genetic relatedness and effects of geographical proximity between languages.
- To combine qualitative and quantitative methods in order to uncover what is specific to the evolution of languages as opposed to the evolution of species.
- To assess the historical signal contained in linguistic data from different levels.

Research Team:

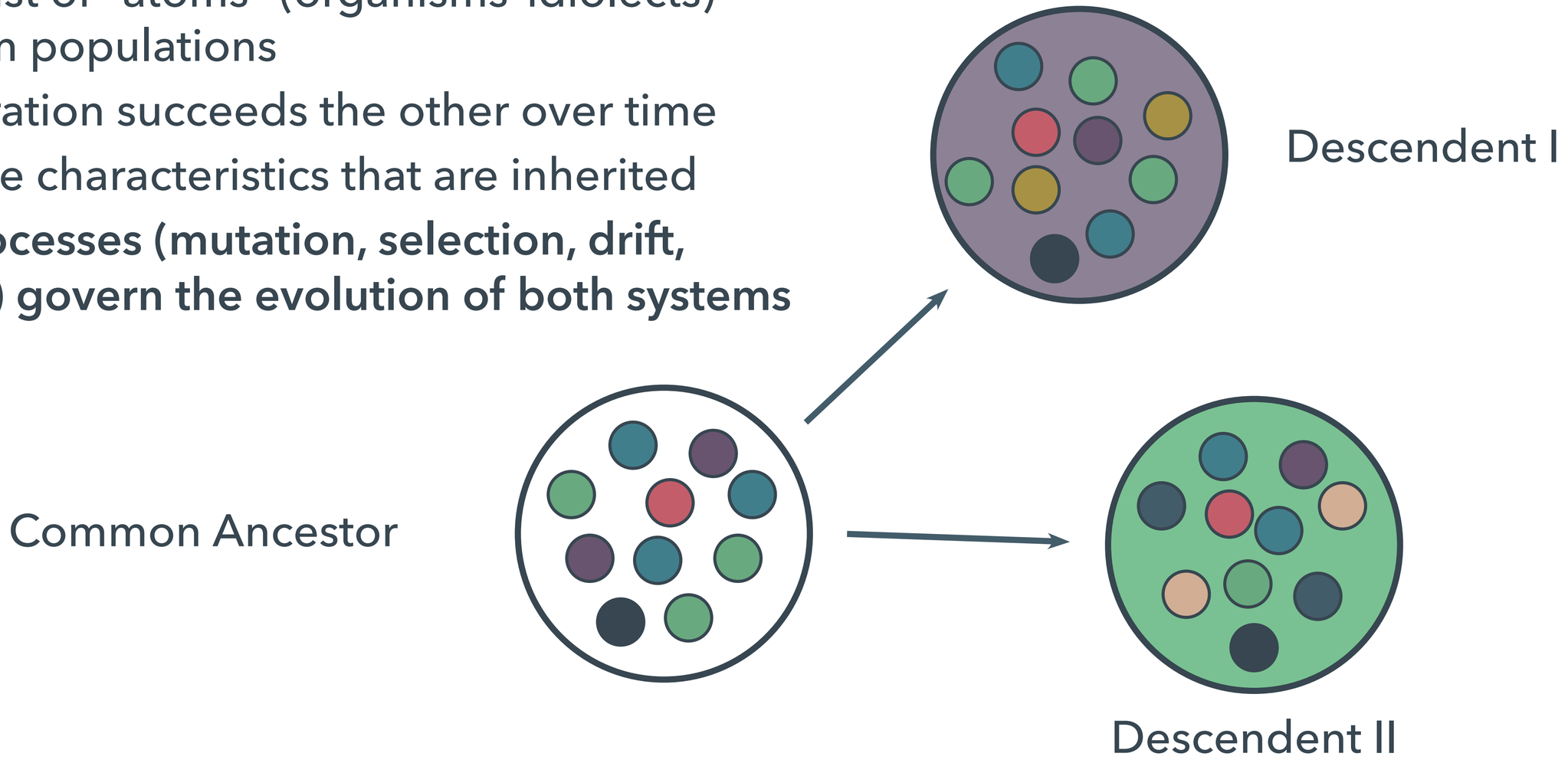
- Elena Anagnostopoulou (University of Crete)
- Manolis Ladoukakis (University of Crete)
- Dimitris Michelioudakis (Aristotle University of Thessaloniki)
- Maria-Margarita Makri (University of the Aegean, University of Crete)
- Pavlos Pavlidis (Foundation for Research and Technology—Hellas, FORTH)
- Athanasios Michail Ramadanidis (University of Crete)

Project Website:

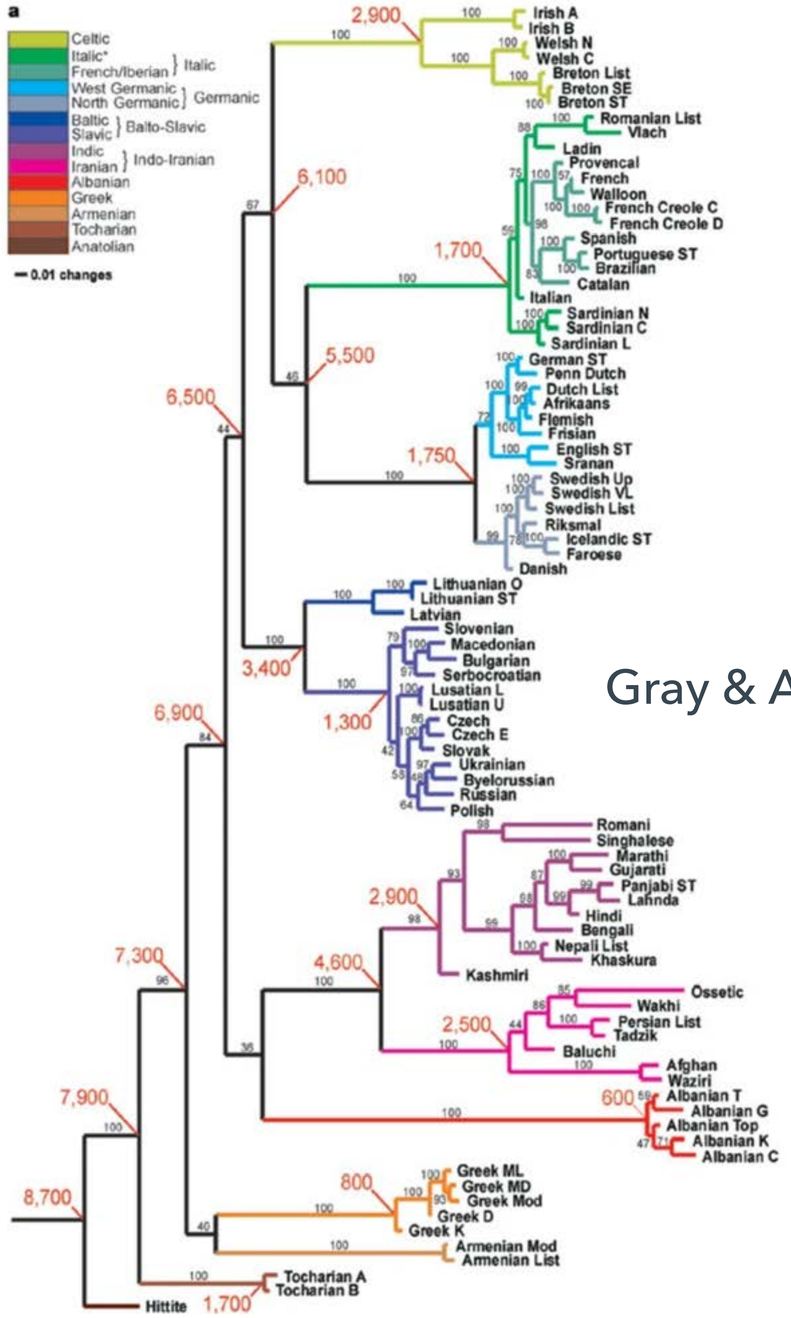
http://modelgloss.philology.uoc.gr/

Similarities between languages and species

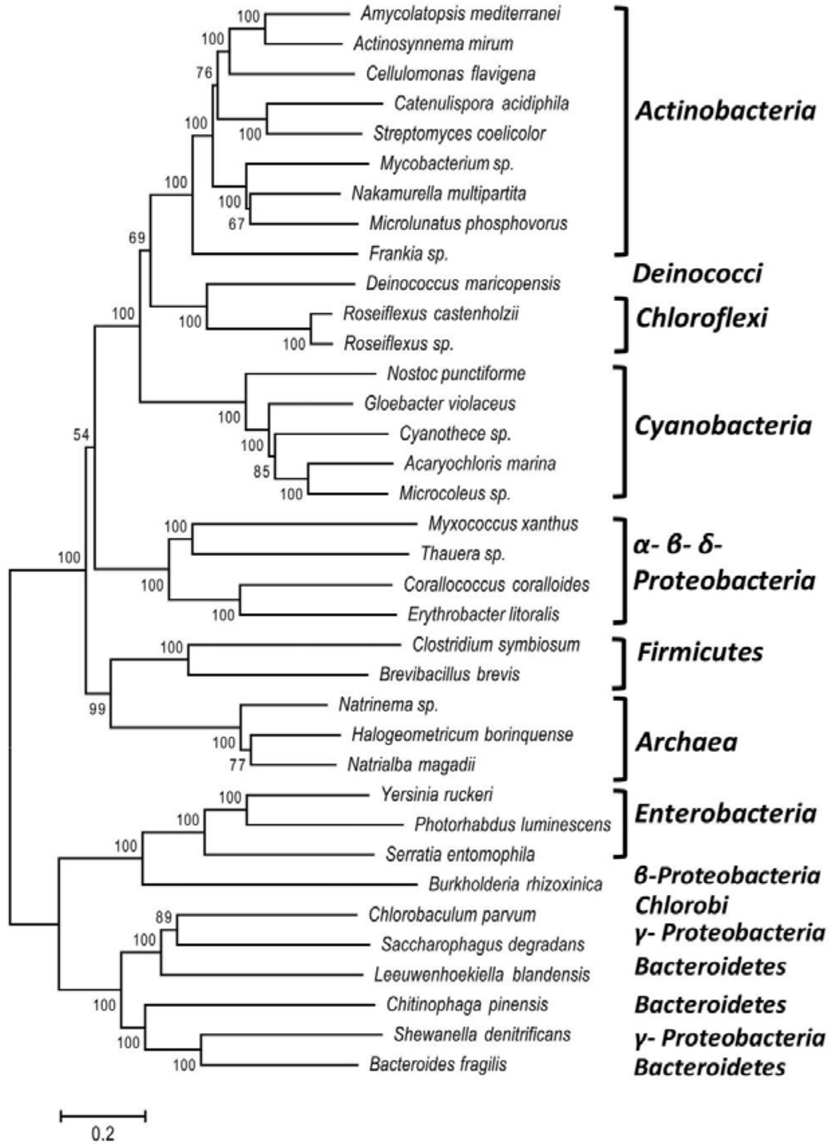
- both consist of “atoms” (organisms–idiolects) which form populations
- one generation succeeds the other over time
- atoms have characteristics that are inherited
- **similar processes (mutation, selection, drift, migration) govern the evolution of both systems**



Reconstruction of genealogical relations with phylogenetic methods

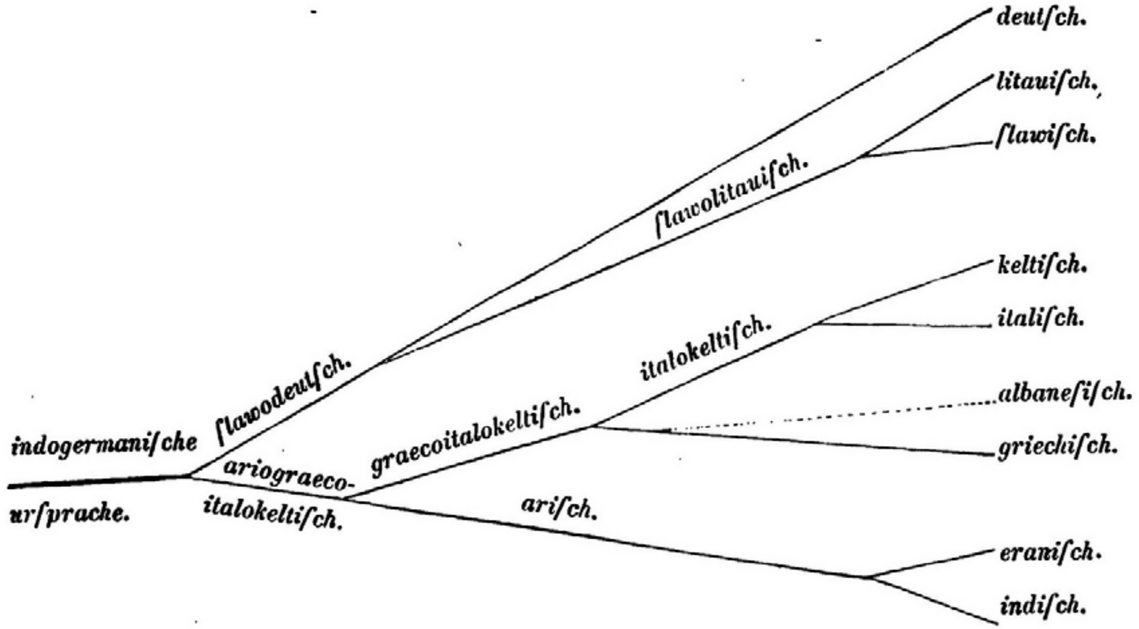


Gray & Atkinson, 2003

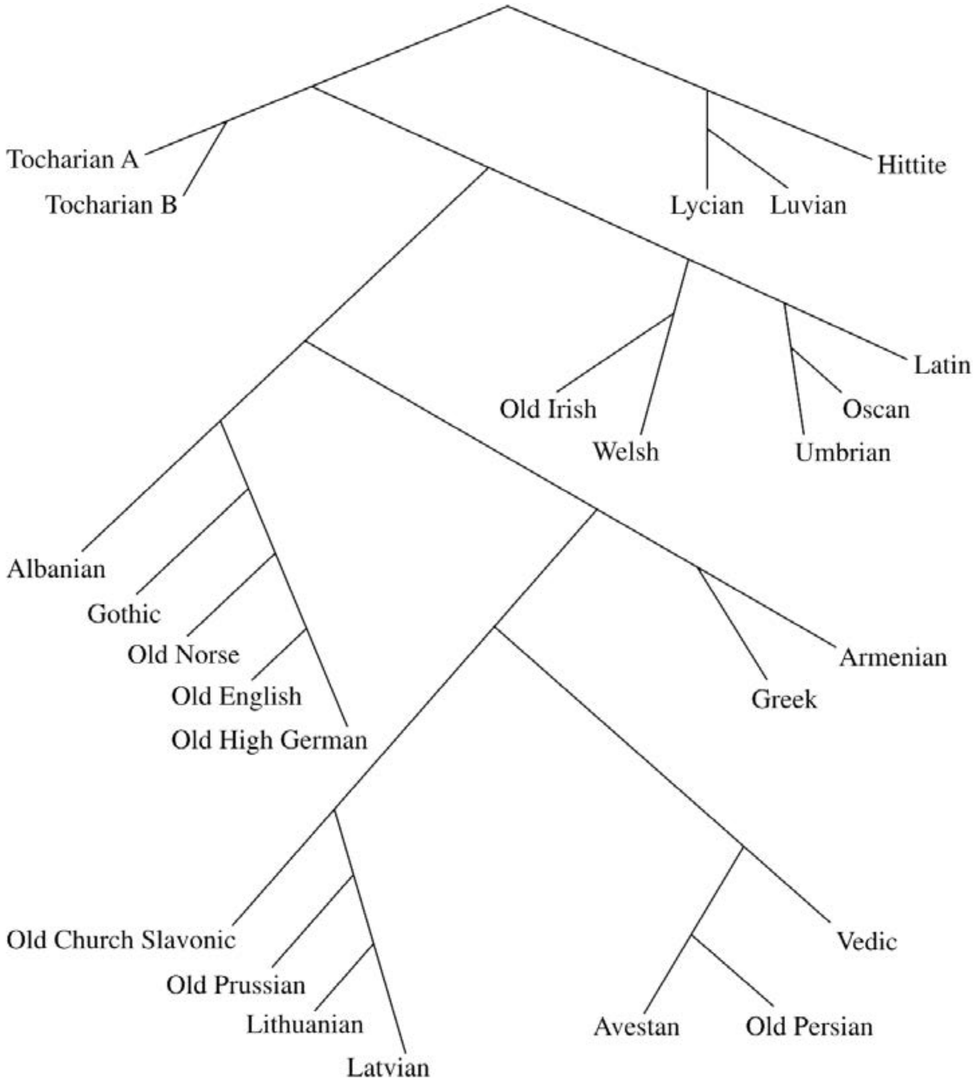


Sarris, Ladoukakis et al., 2014

Traditional *vs* computational phylogenetics



Schleicher, 1863



Ringe, Warnow, & Taylor, 2002

Cognates *vs* morphosyntactic data I

- In modern historical research based on phylogenetic methods, the primary data employed are **cognate words drawn from the conservative vocabulary** of languages, which has been shown to resist borrowing (Swadesh, 1955).
- According to the orthodox view in historical linguistics, this method saturates at a **historical depth of 6000–10000 years** (though see Gray & Atkinson, 2003; Jäger, 2015).
- Nichols (1992) proposes that **grammatical features can help trace deeper** history, at least 15000 up to potentially 50000 years.

Cognates *vs* morphosyntactic data II

- On the other hand, Greenhill et al. (2017) argue that grammatical features are less reliable for two, possibly related, reasons:
    a. many **grammatical features change faster** than the conservative part of the lexicon;
    b. structural data show a much higher level of **conflicting signal** due to parallel evolution and areal diffusion.
- Greenhill et al. observe that the more **slowly evolving grammatical features** seem to be more **abstract** and less available to speaker reflection.
- Partly in the same spirit, Ringe, Warnow and Taylor (2002) **integrate grammatical characters** to improve the phylogenetic reconstruction of Indo-European.
- Longobardi and Guardiano (2009), Longobardi et al. (2016) and related work build phylogenies on the basis of **exclusively morphosyntactic data** from the **nominal domain**, presented as a network of interdependent parameters.

The ModelGloss approach

- **ModelGloss investigates the potential and the challenges of morphosyntactic data.**
- We aim to investigate and disentangle the overall signal of morphosyntax, teasing apart:
    1. features that preserve the **historical signal** from
    2. features that are prone to **contact-induced transfer**, and
    3. **homoplastic** features.
- Our guiding hypothesis is that:
    - › Type-1 features relate to **macro-/meso-**parameters
    - › Type-2 features relate to **micro-/nano-**parameters (cf. Biberauer & Roberts, 2016; Roberts, 2019)
    - › Type-3 features reveal **universal** tendencies

Exploring WALS and other typological databases

Our starting point is the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013), which has the following advantages:

1. It is the **largest and most complete database** covering a wide array of grammatical phenomena ('features' in WALS) as manifested in a great number of languages across the world.
2. It provides fairly complete information on a **geographically and historically balanced** sample of 100 languages.
3. The combination of a big number of features coming from all major areas of morphology and syntax **minimizes the effects of homoplasy**.
4. It allows for a higher order of magnitude in the **number of available taxonomic characters**, which is imperative for phylogenetic research.
5. Data of this type are the only data that can help us reconstruct **relationships across families** (deep phylogenies).
6. It **can be combined** with other typological databases such as SSWL leading to further enrichment of the data.

Limitations of the 100-language sample of WALS

1. The 100-language sample **may only be used for large-scale phylogenetic research** as it is a representative sample of language families around the globe.
2. It contains **very few languages per family**. Ideally it could be used to reconstruct deep genealogical relationships between families. In practice, it could turn out that it only reconstructs shallow relationships between the languages that belong to the same family.
3. Practical considerations: many missing values, inaccuracies, inconsistencies, and mistakes which are difficult to control for and cross-check, given that the languag-es included have been researched to a lesser degree.

A WALS-based database of 60 IE languages I

WALS features are multi-valued and are classified into the following categories:

- phonology (20), morphology (12), nominal categories (29), nominal syntax (8), verbal categories (17), word order (56), simple clauses (26), complex sentences (7), lexicon (13).

This is a good starting point, because we were able to select characters that cover all available domains of morphology and syntax.

In order to overcome the limitations of the 100-language sample, we developed **a novel table with 60 IE languages,** taking into account WALS features from the following domains:

1. morphology (12), nominal categories (28), nominal syntax (8), verbal categories (17), word order (56), simple clauses (24), complex sentences (7), lexicon (4). At this stage, a small number of features was excluded as non-informative.

A WALS-based database of 60 IE languages II

2. Of these features, ≥75% completeness has been attained for the following domains:
    - morphology (8),
    - nominal categories (23),
    - nominal syntax (7),
    - verbal categories (13),
    - word order (20),
    - simple clauses (18),
    - complex sentences (1),
    - lexicon (2).

Most of the features excluded were non-applicable in IE (e.g. a big number of word order features could only be defined under conditions absent in all IE languages) or features for which it was difficult to find precise data, due to the fuzziness of the definition (e.g. 22A “Inflectional Synthesis of the Verb”).

A WALS-based database of 60 IE languages III

3. Out of the 60x92 = 5520 resulting data points, roughly 30% could be consistently filled on the basis of values contained in the WALS database.

   In order to get maximally reliable data:

   i. we cross-checked the accuracy of the existing information, and

   ii. we provided values for the missing cells,

   based on existing grammatical descriptions, the formal linguistic literature and native speaker intuitions.

A WALS-based database of 60 IE languages IV

4. We then **turned the originally multi-valued features into binary characters**, essentially turning each of the 5345 set values (of the features defined in 3 above) into a binary question asking whether the state described by that value is true of the language under investigation or not. This resulted into 16380 binary characters.

5. We also **annotated our characters in terms of the scale** developed in Wichmann and Holman (2009), which divides the WALS features and their states into four types according to their diachronic stability: **very stable, stable, unstable, very unstable**.

A WALS-based database of 60 IE languages V: language selection

A. Contemporary Languages

- European Languages (44)
  i. Albanian
  ii. Armenian
  iii. Standard Greek, Cypriot Greek, 3 Italiot & 5 Asia Minor varieties
  iv. Romance (7)
  v. Celtic (4)
  vi. Balto-Slavic (12)
  vii. Germanic (9)

- Indo-Iranian Languages (9)
  i. Indic (7)
  ii. Iranian (2)

B. Older systems

i. Classical Greek
ii. New Testament Greek
iii. Latin

iv. Old Church Slavonic
v. Gothic
vi. Old English

vii. Sanskrit

ModelGloss data

First ModelGloss data include:

- Table of language characters and character values from The World Atlas of Language Structures (https://wals.info)
- Data from the 100-language sample of WALS

		Characters					
		SC98	SC99	SC100	SC101	SC102	SC103
Languages	hix	1	0	1	0	0	0
	hmo	1	0	1	0	1	1
	imo	0	0	0	0	1	0
	ind	1	0	1	0	0	1
	jak	1	0	1	0	0	0
	jpn	0	1	0	1	1	0
	knd	0	1	0	1	0	0
	krk	1	0	1	0	0	0
	kay	0	1	0	1	1	0
	kew	0	0	0	0	0	0

ModelGloss data: enriching the database

- Table of 156 WALS features with values that were specifically collected for 60 Indo-European languages

		Characters					
		37A	38A	39A	40A	41A	42A
Languages	Italian	1	2	3	3	2	1
	Spanish	1	2	3	3	3	1
	Catalan	1	2	3	3	3	1
	French	1	2	3	3	1	2
	EU Portug.	1	2	3	3	3	1
	Br. Portug.	1	2	3	3	3	1
	Romanian	3	2	3	3	2	1
	Latin	5	5	3	3	4	1
	Cl. Greek	1	4	3	3	3	1
	New Test.	1	4	3	3		1

ModelGloss data: converting to binary

- Out of the 156 WALS features, >75% coverage was attained for 92 of them.
- We turned the values of the 92 features into binary questions asking whether the state described by that value is true of the language under investigation or not.
- This resulted into 425 binary characters.

		Characters				
		37A-1	37A-2	37A-3	37A-4	37A-5
Languages	Italian	1	0	0	0	0
	Spanish	1	0	0	0	0
	Catalan	1	0	0	0	0
	French	1	0	0	0	0
	EU Portug.	1	0	0	0	0
	Br. Portug.	1	0	0	0	0
	Romanian	0	0	1	0	0
	Latin	0	0	0	0	1
	Cl. Greek	1	0	0	0	0
	New Test.	1	0	0	0	0

Computational methods

- Distance-based, specifically neighbour-joining, phylogenetic algorithms
- Maximum Parsimony (character based) algorithms
- Bayesian phylogenetic software (including strict and relaxed clock models) following Greenhill et al. (2017)

WALS-based IE trees: main results

Our preliminary results include:

- The topology best matching received wisdom
- The effects of varying certain variables, while keeping others constant:
  - › Clock models, older languages and the Balkan Sprachbund
  - › The effects of stable and very stable characters
  - › Effects of homoplastic features

WALS-based IE trees: main results

1. Compared to the other tree-constructing methods, the **bayesian trees** had stronger statistical support and better reflected known historical and geographical relationships.
2. Within the bayesian framework, the following variables were manipulated:
   a. strict *vs* relaxed clock models;
   b. inclusion/exclusion of non-contemporary languages;
   c. different subsets of characters, according to their stability.

WALS-based IE trees: main results

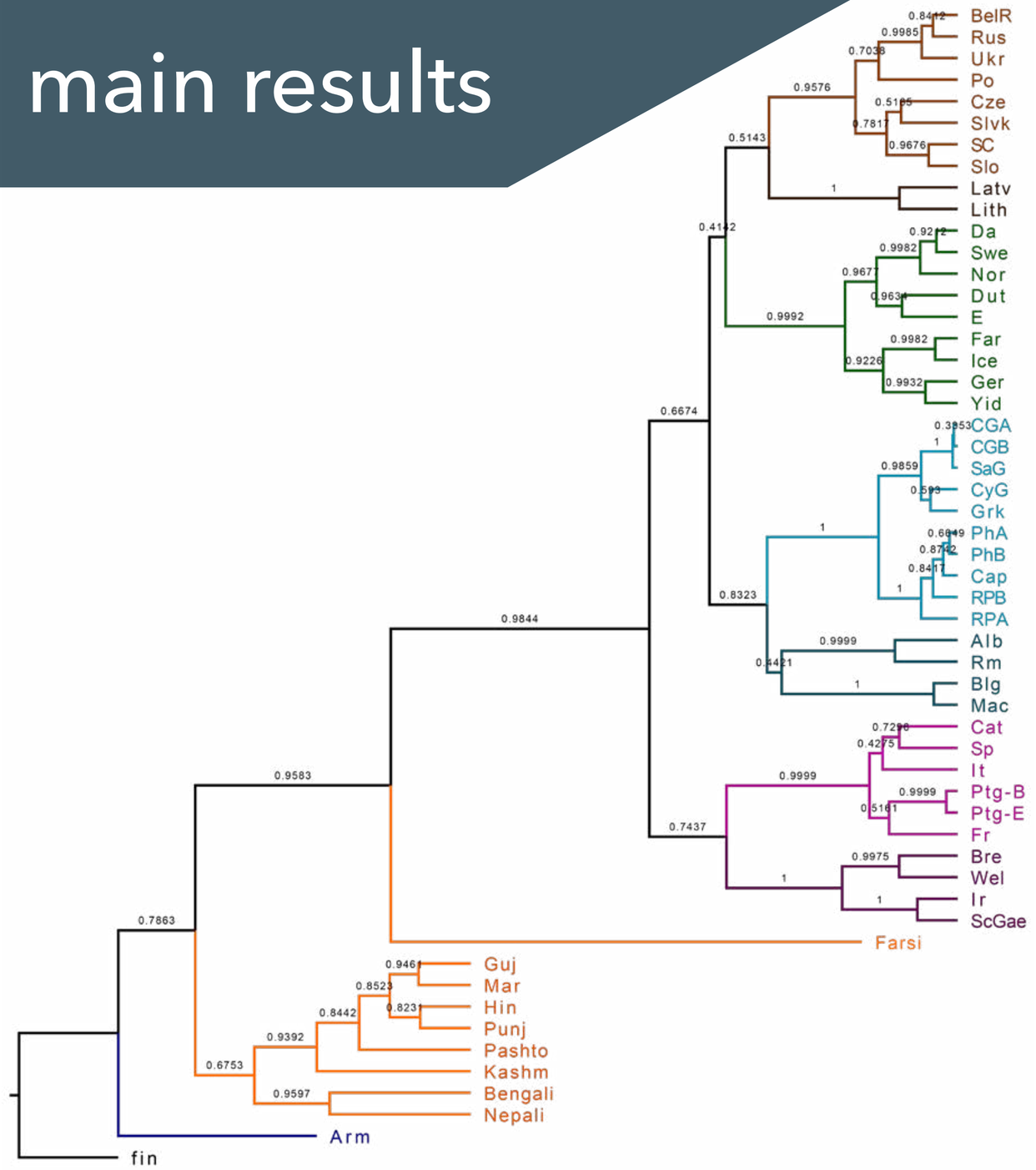
3. Most experiments reconstruct genera such as Romance, Celtic, Greek (with its dialects), Slavic, Germanic and Indo-Iranian, with the exception of:

   i. a clade that groups together the languages of the, so-called, Balkan Sprachbund;

   ii. a clade including some (or sometimes all) of the non-contemporary languages, when taken into account.

WALS-based IE trees: main results

4. The tree that is closest to received wisdom results from a bayesian analysis with:
   i. a sample including contemporary languages only;
   ii. a strict clock model;
   iii. very stable, stable and unstable characters (i.e. excluding very unstable ones).



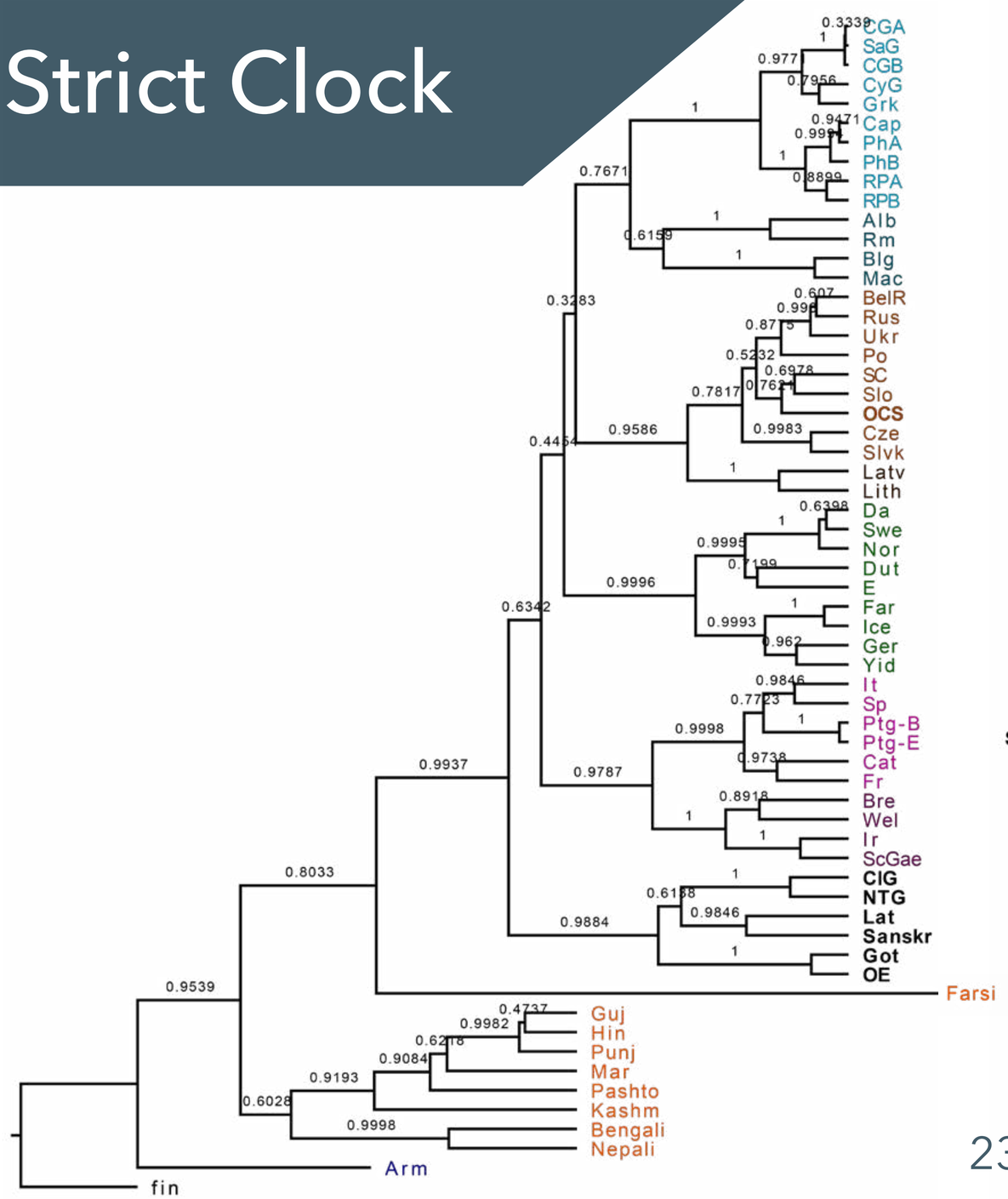
Clock models, older languages and the Balkan Sprachbund

The choice of clock model matters:

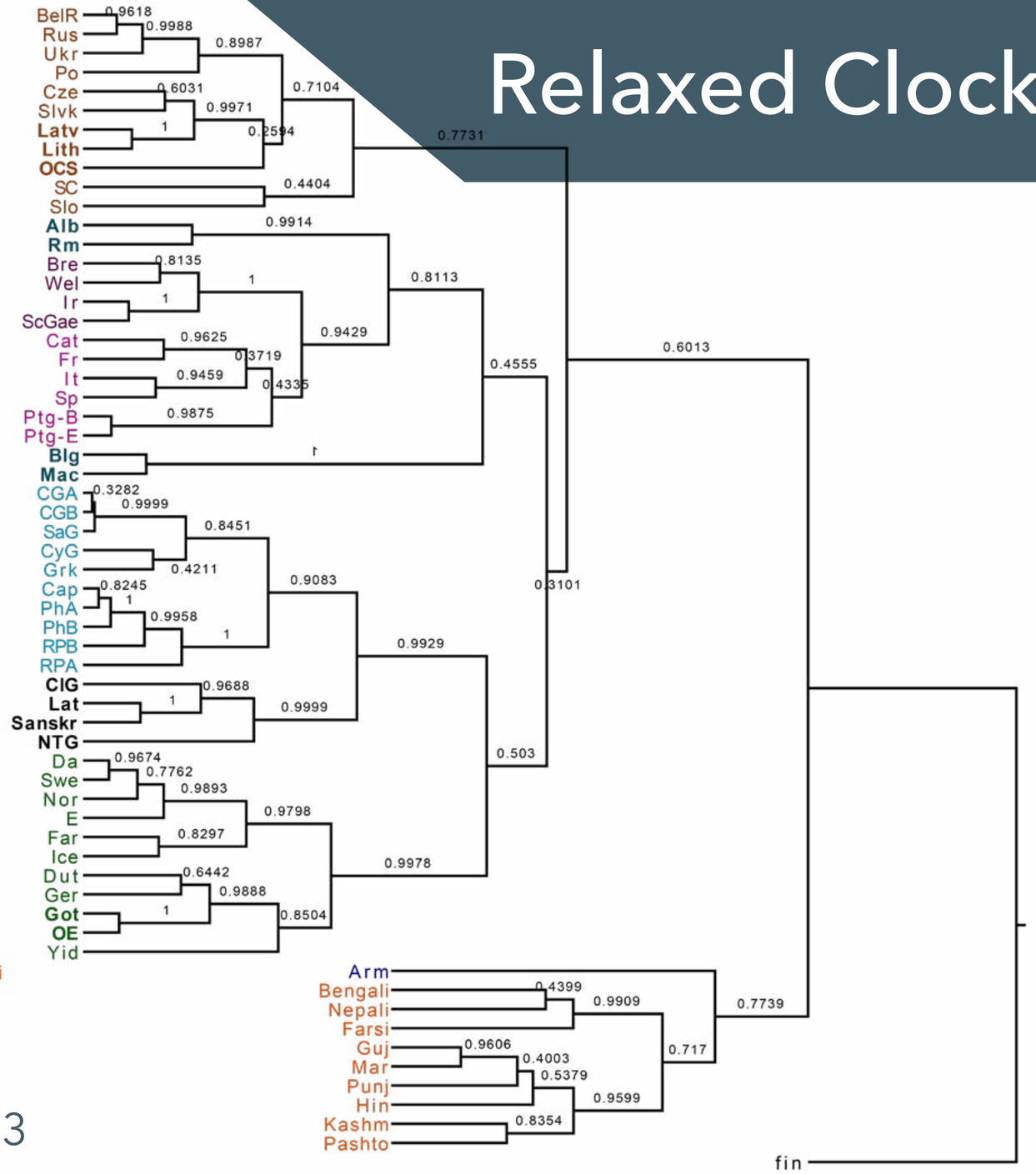
- When all characters are taken into account, only a relaxed clock model can link the medieval languages (though not the more ancient ones) to their respective genera.
- On the other hand, connections due to long-term contact, such as the Balkan Sprachbund, are lost with a relaxed clock model and preserved with a strict clock model.

Historical classification in the case of the Balkan Sprachbund has not been attained with the models so far applied.

Strict Clock



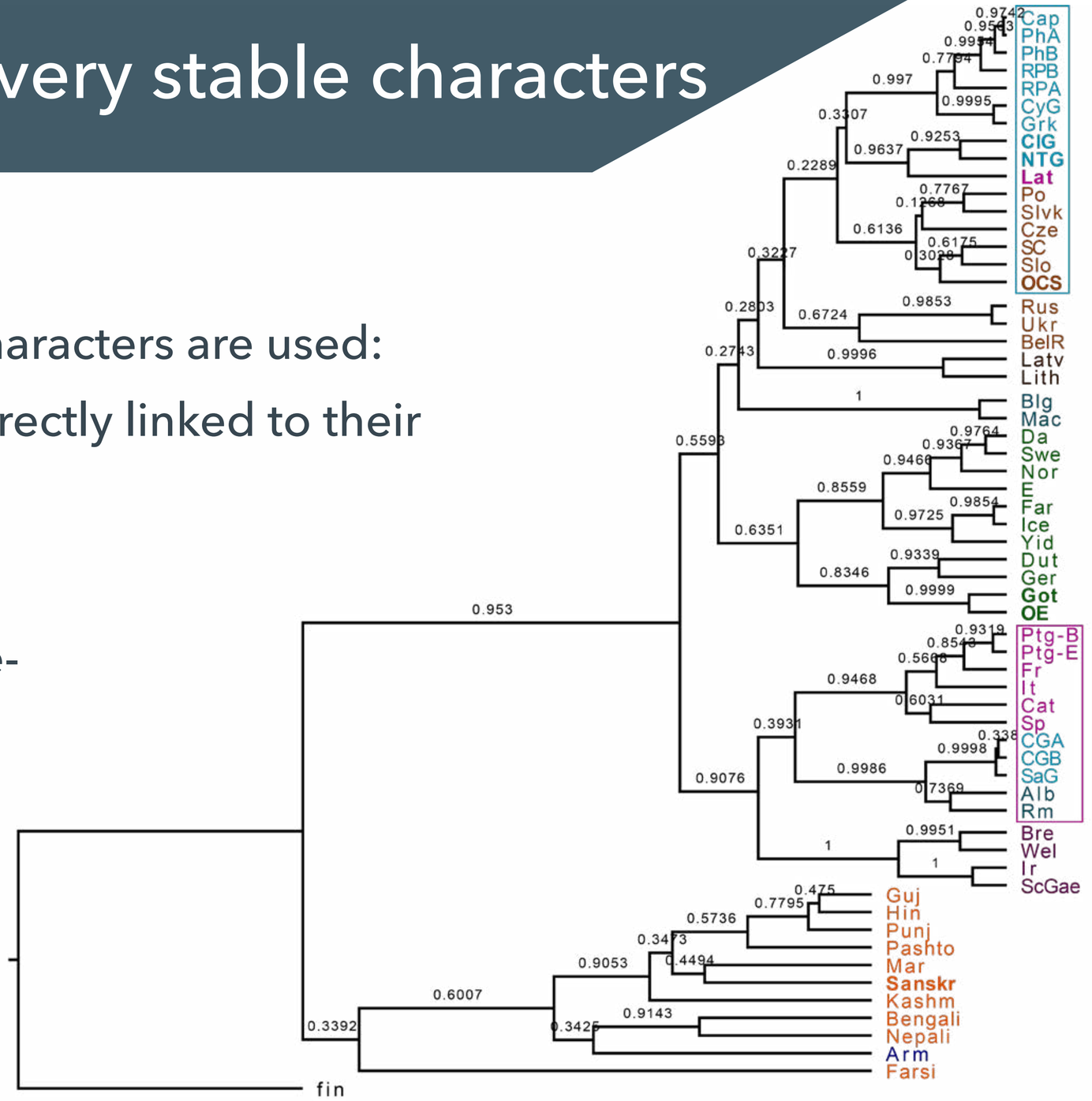
Relaxed Clock



The effects of stable and very stable characters

When only stable and very stable characters are used:

- the older languages are (more) correctly linked to their respective genera.
- However, the unity of families such as Greek, Balto-Slavic or Romance, and of the Balkan Sprachbund is seriously disturbed.

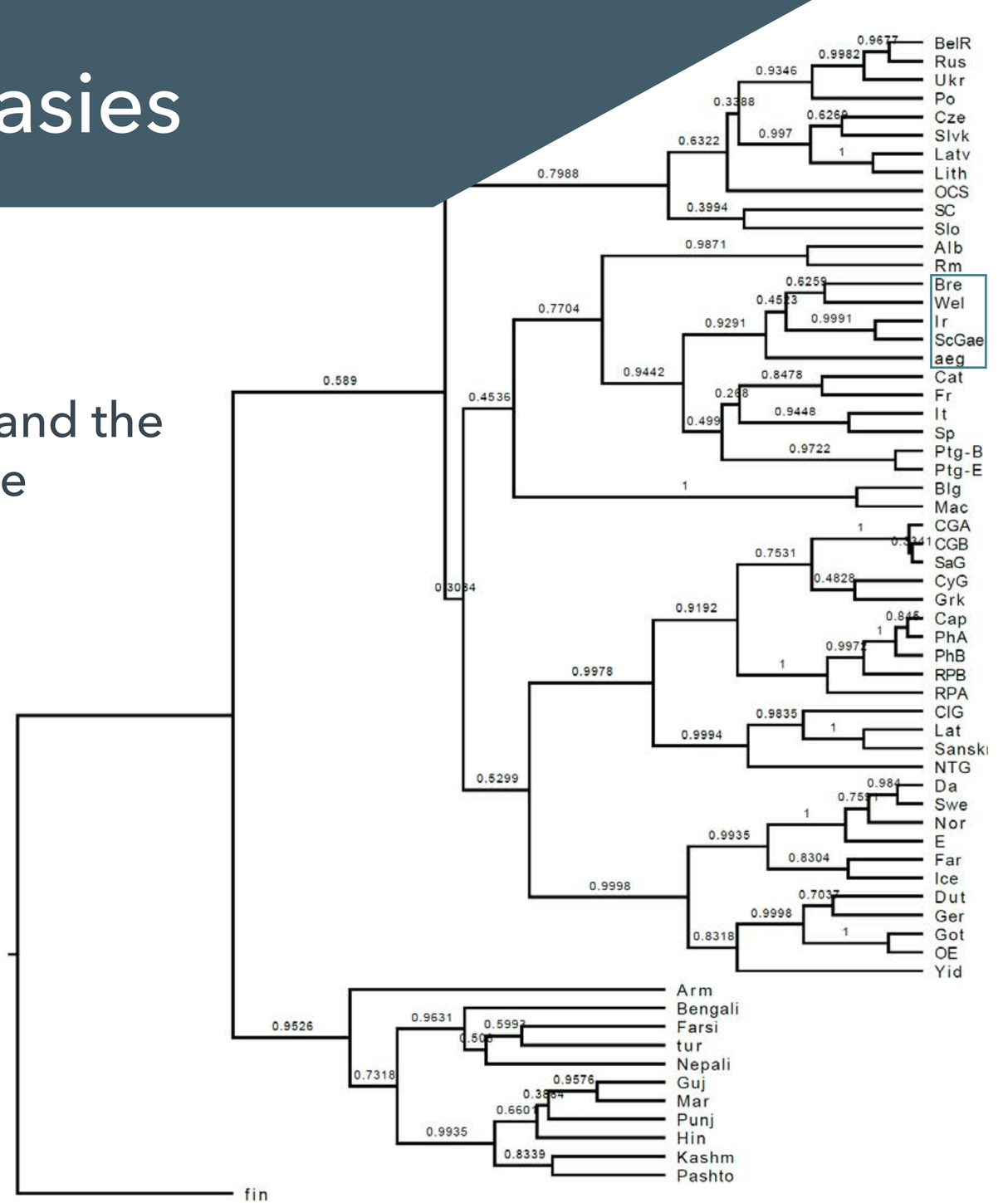


The contribution of very stable characters

- Very stable characters are significant for very deep historical connections, e.g. they secure the placement of Farsi within IE.
- In all trees above Farsi is always within IE and connected to the rest of Indo-Iranian.
- In trees not including very stable characters Farsi is classified as a non-IE language.

Effects of homoplasies

- Languages with strong similarities in the IP and the DP domain are clustered together in the tree (Arabic is linked to Celtic).
- The Celtic-Semitic connection is remedied when very stable characters are excluded.
- This suggests that very stable characters include a considerable amount of homo-plastic morphosyntactic features.



Conclusions

1. Despite some limitations (e.g. conflation of surface patterns with potentially multiple structural sources and, thus, inevitable homoplasies), WALS features do provide useful historical information.
2. Wichmann and Holman's (2009) classification of characters according to stability has significant implications for historical reconstruction.
3. No combination of variables among those tested reverses the effects of long-term contact, e.g. languages of the Balkan Sprachbund never cease to be connected to some extent. Languages such as Bulgarian or Romanian cannot be made to attach to their respective genera. This could be linked to the nature of WALS characters.
4. The models we tested do not simultaneously capture ancient languages, micro-variation and contact effects.

ModelGloss

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: HFRI-FM17-44).

References

Biberauer, T., & Roberts, I. (2016). Parameter typology from a diachronic perspective. *Theoretical approaches to linguistic variation*, *234*, 259–291.

Dryer, M. S., & Haspelmath, M. (2013). The world atlas of language structures online.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439.

Greenhill, S. J., Wu, C. H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, *114*(42), E8822-E8829.

Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, *112*(41), 12752–12757.

Longobardi, G., & Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua*, *119*(11), 1679–1706.

Longobardi, G., Ceolin, A., Ecay, A., Ghirotto, S., Guardiano, C., Irimia, M. A, Michelioudakis, D., Radkevich, N., Pettener, D., Luiselli, D. & Barbujani, G. (2016). Formal linguistics as a cue to demographic history.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press.

Ringe, D., Warnow, T., & Taylor, A. (2002). Indo-European and computational cladistics. *Transactions of the philological society*, 100(1), 59–129.

Roberts, I. (2019). *Parameter hierarchie s and universal grammar*. Oxford University Press, USA.

Sarris, P. F., Ladoukakis, E. D., Panopoulos, N. J., & Scoulica, E. V. (2014). A phage tail-derived element with wide distribution among both prokaryotic domains: a comparative genomic and phylogenetic study. *Genome biology and evolution*, *6*(7), 1739–1747.

Schleicher, A. (1863). Darwinism tested by the science of language, trans. AV M. Bikkers. Böhlau. Reprinted in:(1983) Linguistics and evolutionary theory: Three essays, ed. K. Koerner.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, *21*(2), 121–137.

Wichmann, S., & Holman, E. W. (2009). *Assessing temporal stability for linguistic typological features*. München: LINCOM Europa.