# THE MATHEMATICS OF PARAMETRIC COMPARISON

LUCA BORTOLUSSI, ANDREA CEOLIN, GUIDO CORDONI, DIMITAR KAZAKOV,

CRISTINA GUARDIANO, MONICA IRIMIA, GIUSEPPE LONGOBARDI,

NINA RADKEVICH AND ANDREA SGARRO



# THE PROBABILITY OF LANGUAGE RELATEDNESS

- **The probability problem** (Guardiano et al. *in press*)
- **How much similarity is required to demonstrate true language relatedness** when regular sound correspondences are not retrievable?

# A MODEL OF UG AND VARIATION

- Conceptual simplification on the basis of a **Principles and Schemata Model**

  (Longobardi 2005, 2017)

  - 'Initial stage of UG ($S_0$) only contains parameter schemata, and not an extensional list of parameters'

# PARAMETER SCHEMATA (LONGOBARDI 2017)

- Is F, F a feature, **grammaticalized**?
- Does F, F a grammaticalized feature, **Agree** with a category X (i.e. does F probe X)?
- Is F, F a grammaticalized feature, "strong" (i.e. does F **overtly attract** X, probe X with an EPP feature)?
- Is F, F a grammaticalized feature, **spread** on a category X?
- Does a functional category (a set of lexically co-occurring grammaticalized features) X have a **phonological matrix** Φ?
- Does F, F a grammaticalized feature, **probe** the minimal accessible category of type X (or is **pied-piping** possible)?
- Are f1 and f2, the respective values of two grammaticalized features, associated on a category X?
- Are f1 and f2, two feature values associated on X, optionally associated?
- Does a functional feature (set) exist in the vocabulary as a bound/free morpheme?

**Nothing else is a parameter**

# 91 PARAMETERS DEFINING THE SYNTAX OF THE NOMINAL DOMAIN

- The status of **features associated with D**, e.g. person, number, gender and definiteness
- Syntactic properties of **adjectives**, **relative clauses**, **genitival arguments** and **possessives**, **demonstratives**
- Surface position of **N** with respect to its arguments and modifiers

# PARAMETER IMPLICATIONS

- Parameter values: **[+]** or **[-]**
- But parameter values **imply each other** as well: **0** is the state of a parameter which is completely irrelevant owing to the settings of other parameters
- Large number of **parametric interdependencies** (implicational universal principles)
  - Out of 59x91=5369 cells (parameter states), 2393 are null (contain 0), i.e. **44.57%** of the information is redundant

# THE EFFECT OF PARAMETRIC IMPLICATIONS

- To appreciate the typological restrictiveness of a realistic parameter system, we must calculate the **number of possible languages generated**.
- The **first 30 parameters** from TableA (less implicationally constrained than the successive ones) **generate less than $2^{19}$ admissible grammars** (Bortolussi et al. 2011, Ceolin et al. submitted), at least **eleven orders of magnitude less** than the $2^{30}$ expected under total independence (i.e. less than 500k as opposed to more than 1 billion)

# TWO FURTHER STEPS TOWARDS PARSIMONY

i) implied parameters **do not play a role in acquisition** (are never present in the child's mind)

ii) only **[+] valued** parameters are present in the speaker's mind, added from positive evidence in the course of acquisition

iii) to take into account (i) and (ii), we calculated syntactic pairwise distances according to the **Jaccard** formula:

$$\Delta\text{Jaccard}(A,B) = \frac{[\, N_{-+} + N_{+-} \,]}{[\, N_{-+} + N_{+-} + N_{++} \,]}$$

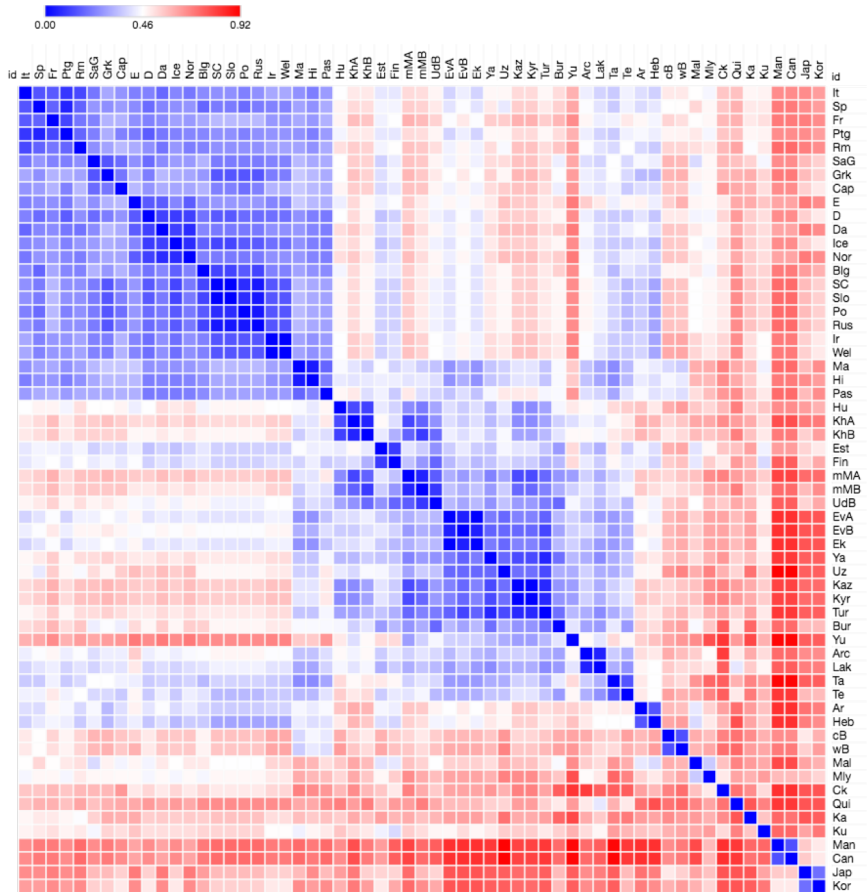
# AN EMPIRICAL IMPLEMENTATION OF THE PCM

- **19** traditionally recognized and irreducible families (scattered across Europe, Asia and the Americas). **59** languages in total.

- **Sino-Tibetan**: Mandarin (Man) and Cantonese (Can)
- **Korean**: Korean (Kor)
- **Japonic**: Japanese (Jap)
- **Muskogean**: Chickasaw (Ck)
- **Guaicuruan**: Kadiweu (Ka)
- **Basque**: Western and Central Basque (wB and cB)
- **Carib**: Kuikuro (Ku)
- **Quechua**: Quichua (Qui)
- **Semitic:** Hebrew (Heb), Arabic (Ar)
- **Dravidian:** Tamil (Ta) Telugu (Te)
- **Austronesian:** Malese (Mls), Malagasy (Mal)
- **North Caucasian :** Archi (Arc), Lak (Lak)
- **Indo-European**:
  - Indo-Iranian: Pashto (Pas), Marathi (Ma), Hindi (Hi);
  - Greek: Salento Greek (SaG), Standard Modern Greek (Grk), Cappadocian Greek (Cap);
  - Romance: Romanian (Rm), French (Fr), Italian (It), Spanish (Sp), Portuguese (Ptg);
  - Germanic: English (E), German (D), Icelandic (Ice), Danish (Da), Norwegian (Nor);
  - Celtic: Irish (Ir), Welsh (Wel);
  - Slavic: Bulgarian (Blg), Russian (Rus), Polish (Po), Serbo-Croat (SC), Slovenian (Slo).
- **Yukaghir**: Yukaghir (Yu).
- **Mongolian**: Buryat (Bur).
- **Turkic**: Turkish (Tur), Yakut (Ya), Uzbek (Uz), Kazakh (Kaz), Kyrgyz (Kyr).
- **Tungusic**: EvenA, EvenB, Evenki (Ek).
- **Uralic**: Balto-Finnic: Finnish (Fin), Estonian (Est); Mari: Meadow Mari (mM); Permic: Udmurt (Ud); Ugric: Hungarian (Hu), KhantyA, KhantyB

# AN EMPIRICAL IMPLEMENTATION OF THE PCM

# MACRO-FAMILIES



# MACRO-FAMILIES

- In order to prove genetic relationship, one needs to rely on **statistical testing** (Kessler and Lehtonen 2006)

- First step: determining a **null distribution**

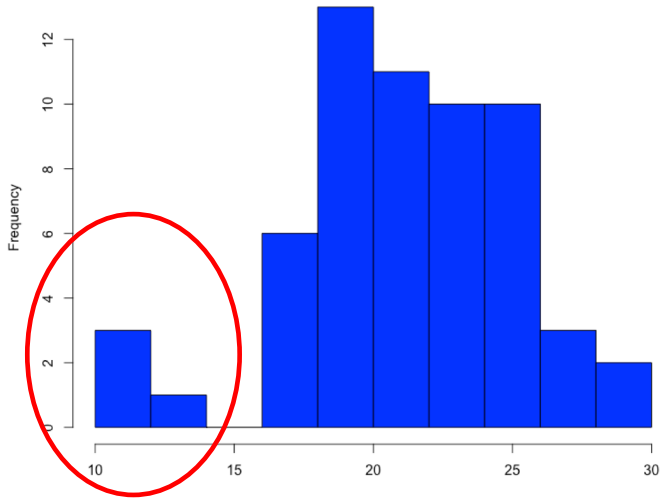
# NULL DISTRIBUTION

- With only **19 families** it's **difficult to determine a null distribution by internal sampling**. Also, with **most** of the languages being from **Eurasia**, it's **difficult to control for historical relations**
- Solutions: we generate artificial languages by recombining the parametric values of the sample. Values are chosen probabilistically, using evidence weighted for families **Implicational constraints are also applied**

# OUTLIERS

- Four languages (Mandarin, Cantonese, Japanese and Korean) have an average of comparable parameters with the other languages = 11 (the average of the entire sample is 21).

We excluded them to avoid a skewed distribution.

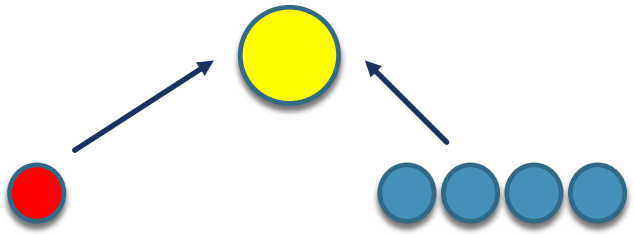


# MEDIAN TESTS

- We **compare distributions** of real and artificial distances by means of Mann-Whitney U tests.
- **First test**:
  - Compare median of distances calculated from a **group of real languages** (e.g. Indo-European) versus median of distances calculated from a **group of artificial languages of the same size**.
  - Repeat 1000 times.
  - Check p-values range.

# MEDIAN TESTS

- **Second test**:
  - Compare median of distances calculated from a **group of real languages of size N** versus median of distances calculated from a **bigger sample** (500 languages).
  - Repeat the same using each of the 1000 artificial samples of the same size N instead.
  - Check p-values range.



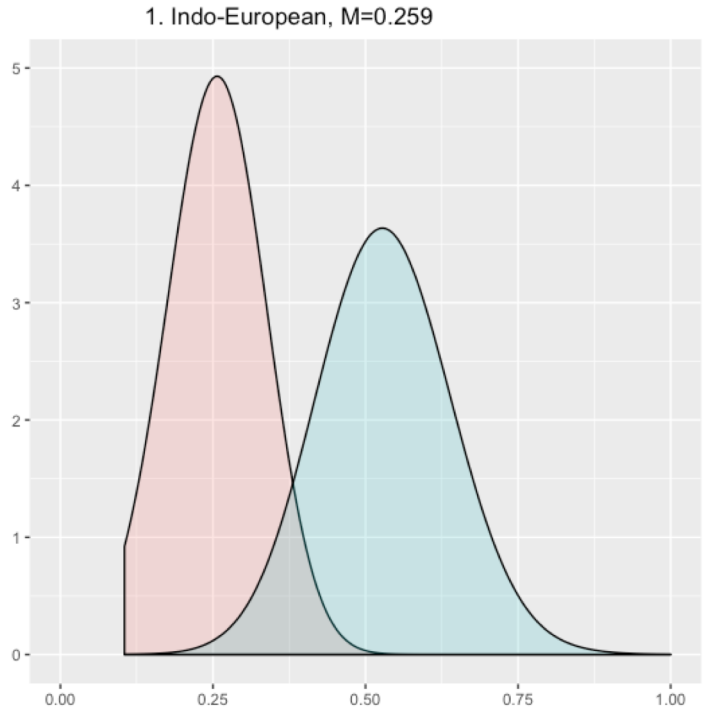
# WHICH MACRO-FAMILIES?

- Potentially, many ($2^N$-(N+1)).
  - We start from **two well-established families**, IE and Uralic.
  - We move to **previously hypothesized macrofamilies**, Altaic, Uralo-Altaic, Indo-Uralic.
  - **Important**: when testing across families, **remove family-internal pairs** (and adjust artificial sample accordingly through sampling)

# A. Well-established families

1. Indo-European

2. Uralic




n_pairs = 216

median = 0.259, sd = 0.051

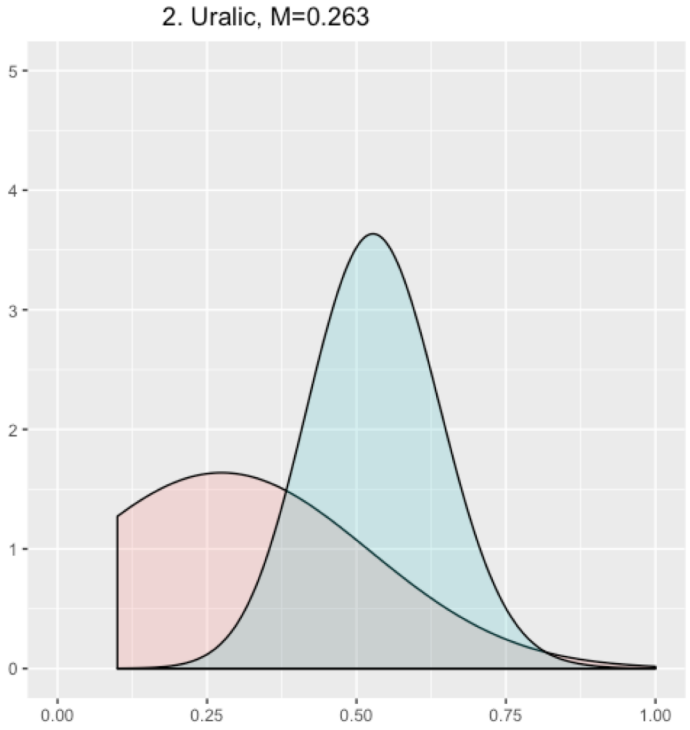
range_mannwhitney = $[2.72*10^{-16}, 4.37*10^{-10}]$

Test1 = <0.001

mannwhitney_bigsample = $3.70*10^{-139}$

range_artificial = $[1.12*10^{-7}, 0.99]$

Test2 = <0.001




n_pairs = 23

median = 0.263, sd = 0.112

range_mannwhitney = $[6.56*10^{-9}, 4.93*10^{-4}]$

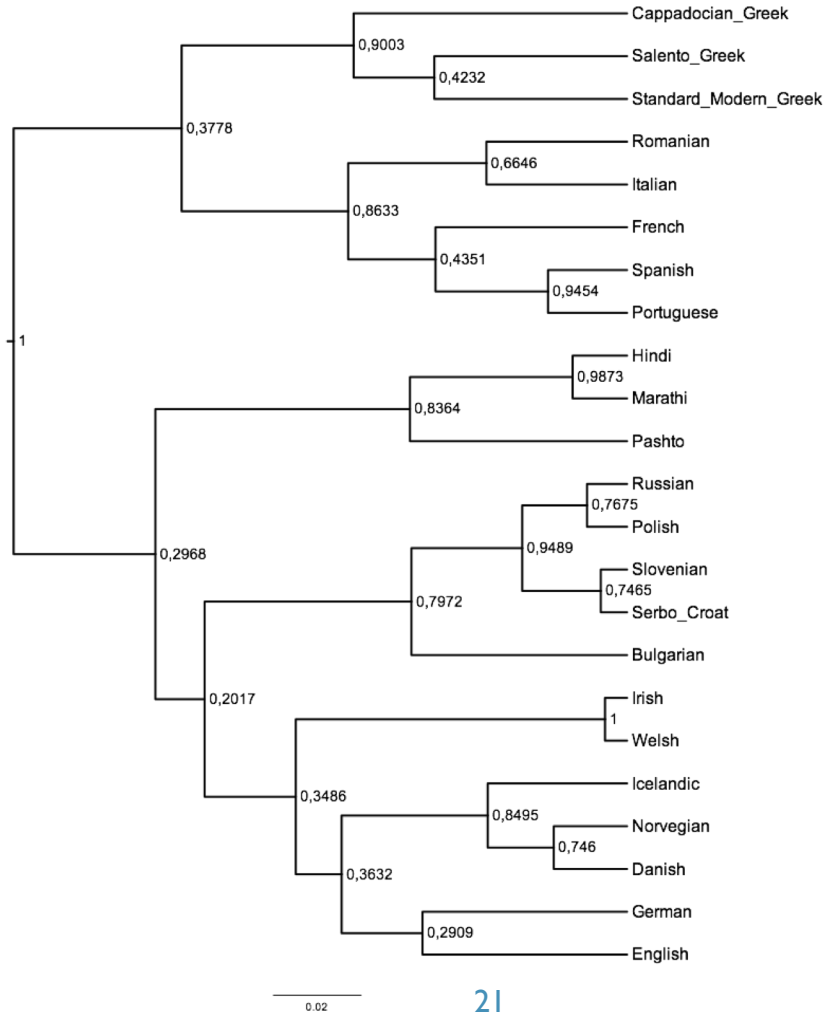
Test1 = <0.001

mannwhitney_bigsample = $3.28*10^{-14}$

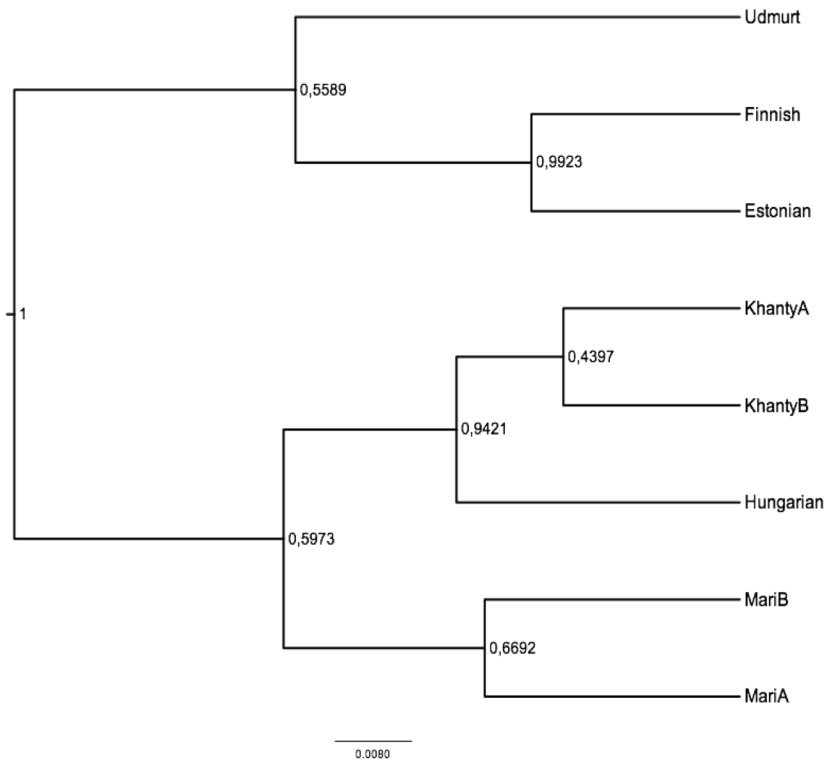
range_artificial = $[3.46*10^{-8}, 0.99]$

Test2 = <0.001

# Phylogenetic Trees - BEAST

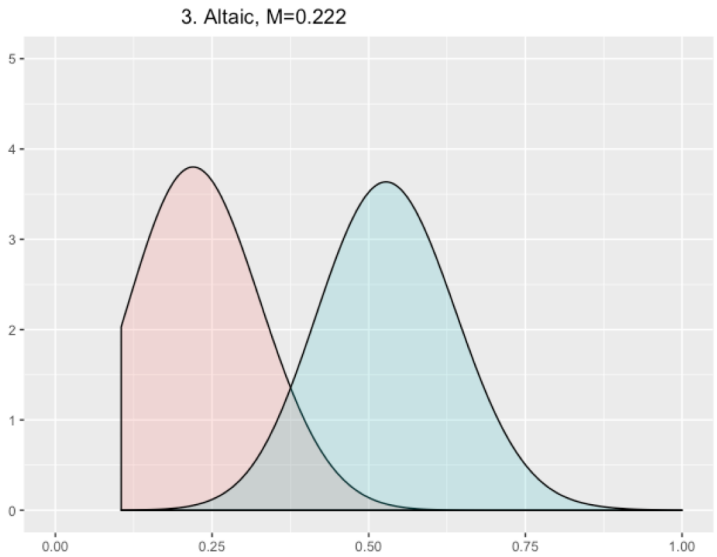


# Phylogenetic Trees - BEAST



# B. **Hypothesized macro-families**

3. Altaic
4. Uralo-Altaic
5. Indo-Uralic




n_pairs = 23

median = 0.222, sd = 0.054

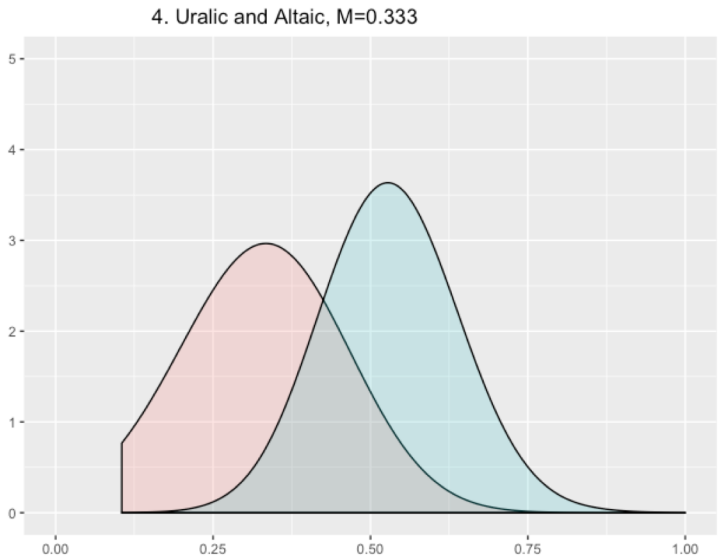
range_mannwhitney = $[5.99*10^{-09}, 2.96*10^{-7}]$

Test1 = <0.001

mannwhitney_bigsample = $1.69*10^{-16}$

range_artificial = $[2.93*10^{-8}, 1.00]$

Test2 = <0.001




n_pairs = 72

median= 0.333, sd = 0.074

range_mannwhitney = $[3.96*10^{-32}, 8.04*10^{-17}]$

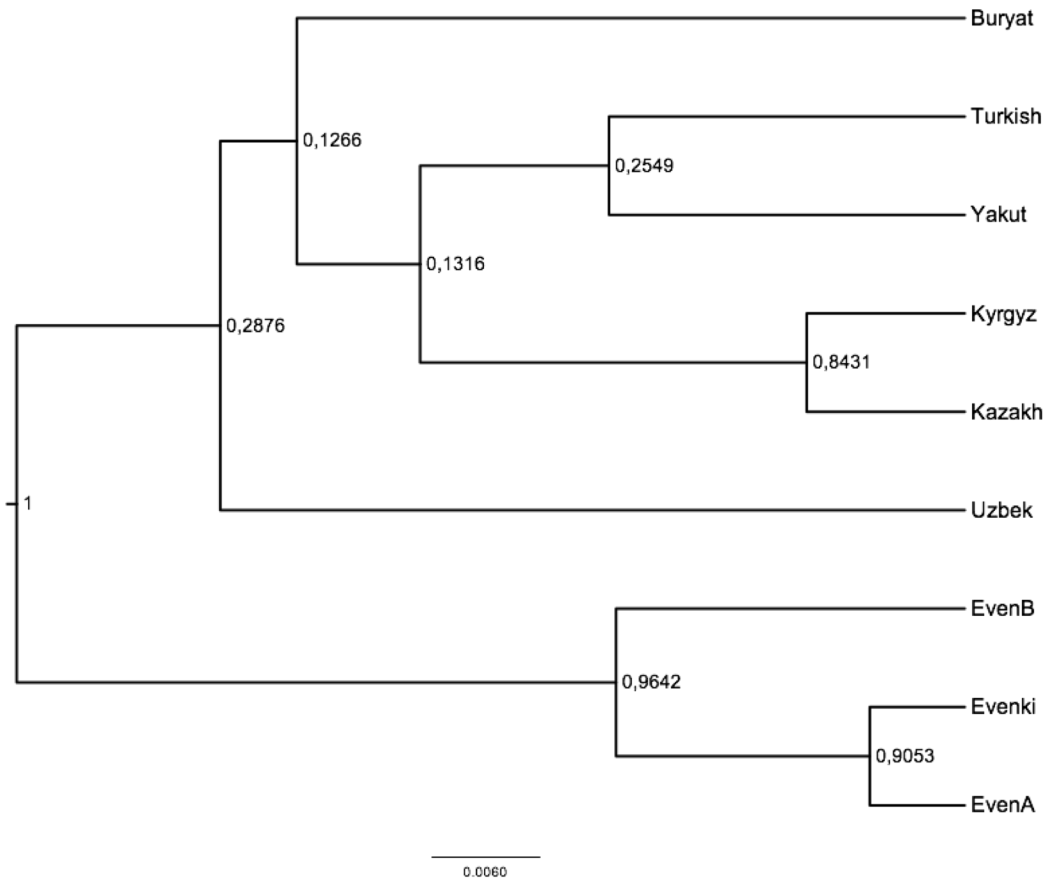
Test1 = <0.001

mannwhitney_bigsample = $3.51*10^{-40}$

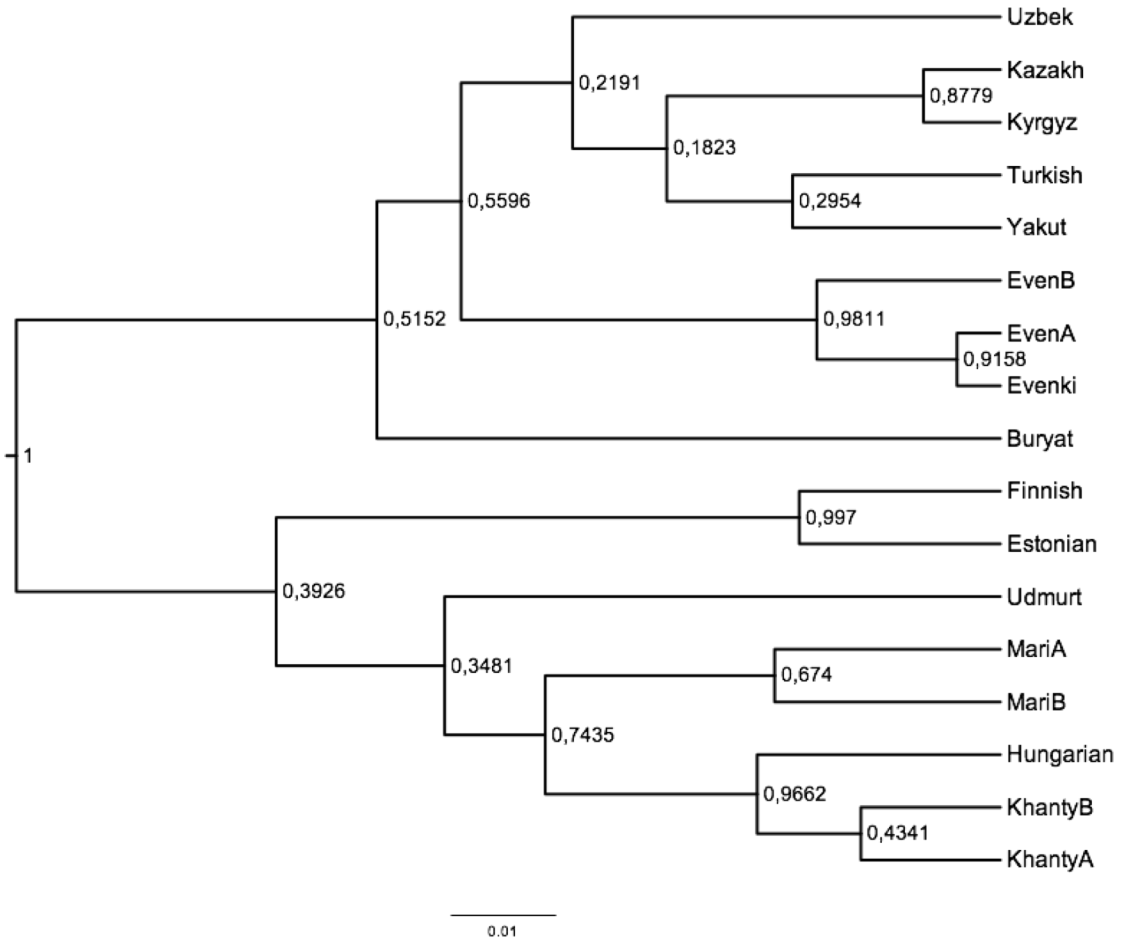
range_artificial = $[2.04*10^{-25}, 0.99]$

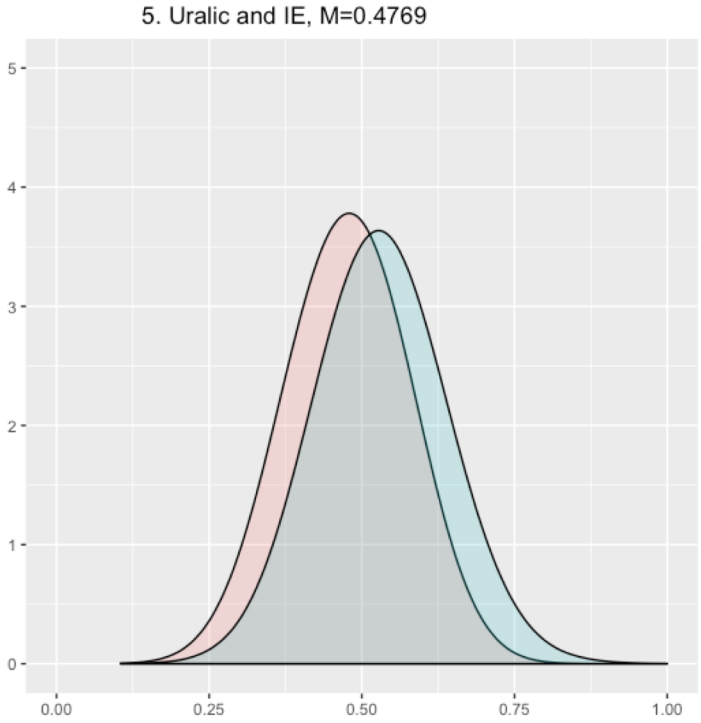
Test1 = <0.001

# Phylogenetic Trees - BEAST



# Phylogenetic Trees - BEAST






n_pairs =184

median = 0.476, sd = 0.063

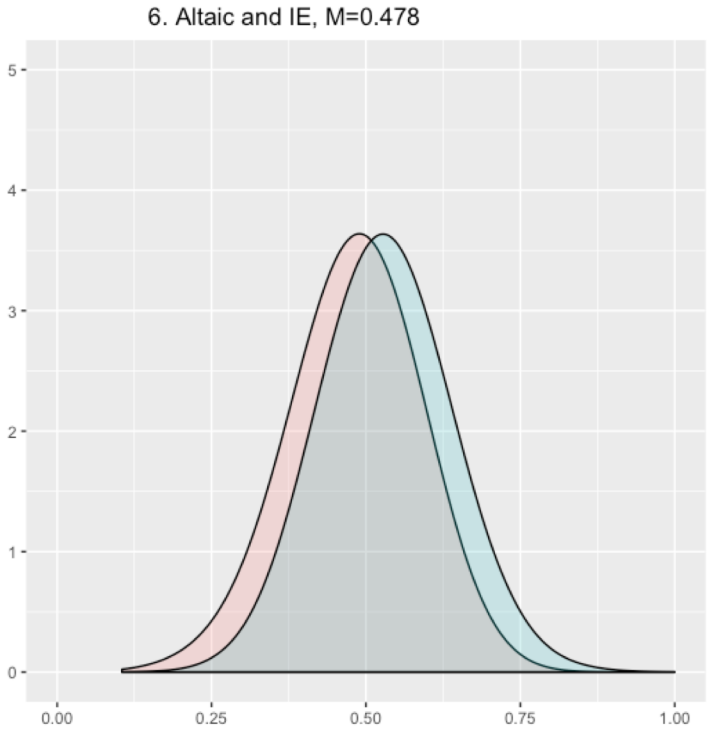
range_mannwhitney = $[3.06*10^{-35}, 0.96]$

Test1 = 0.051

mannwhitney_bigsample = $1.39 *10^{-15}$

range_artificial = $[7.45*10^{-25}, 1.00]$

Test1 = 0.010




n_pairs = 207

median = 0.478, sd = 0.069

range_mannwhitney = [1.67*11$^{-35}$, 1.0]

Test1 = 0.090

mannwhitney_bigsample = 1.58*10$^{-13}$

range_artificial = [1.61*10$^{-23}$, 1.0]

Test2 = 0.015

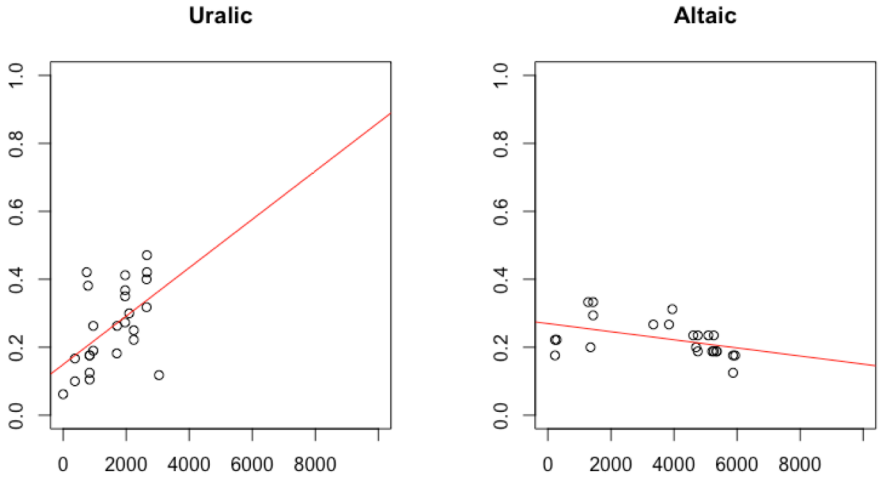
# SUMMARY

- **Indo-European** and **Uralic** are expectedly supported by the test.
- The **Altaic** hypothesis is corroborated.
- Further evidence for a **Uralo-Altaic unit**.
- Larger groups, like Indo-Uralic or Indo-Altaic, are weakly supported by the test.

# GEOGRAPHY

- To what extent are these results influenced by geography?

# GEOGRAPHY

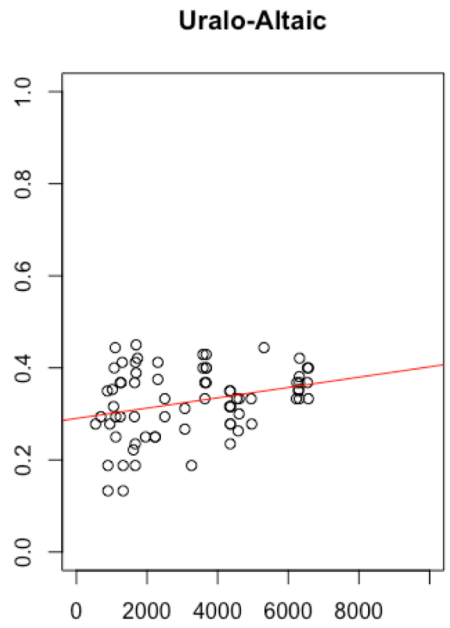






| | Uralic | Altaic |
|---|---|---|
| Kendall tau | 0.498 | -0.466 |
| p-value | 0.001 | 0.003 |
| Pearson r | 0.616 | -0.447 |
| p-value | 0.002 | 0.032 |

# GEOGRAPHY




| | Uralo/Altaic |
|---|---|
| Kendall tau | 0.193 |
| p-value | 0.019 |
| Pearson r | 0.29 |
| p-value | 0.013 |

- Median Balto-Finnic/Altaic: 0.358 (M = 0.333)

Compatible with genealogical relatedness

# CONCLUSIONS

1. Syntax contains a **detectable historical signal**
2. Unlike classical methods, syntax provides **long-range phylogenies encompassing distinct families**
3. Language relatedness can be **tested statistically** thanks to this syntax-based perspective

# THANK YOU!

Thanks to all the members of the LANGELIN project!