Structural features with STRUCTURE

Nataliia Hübler
Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History

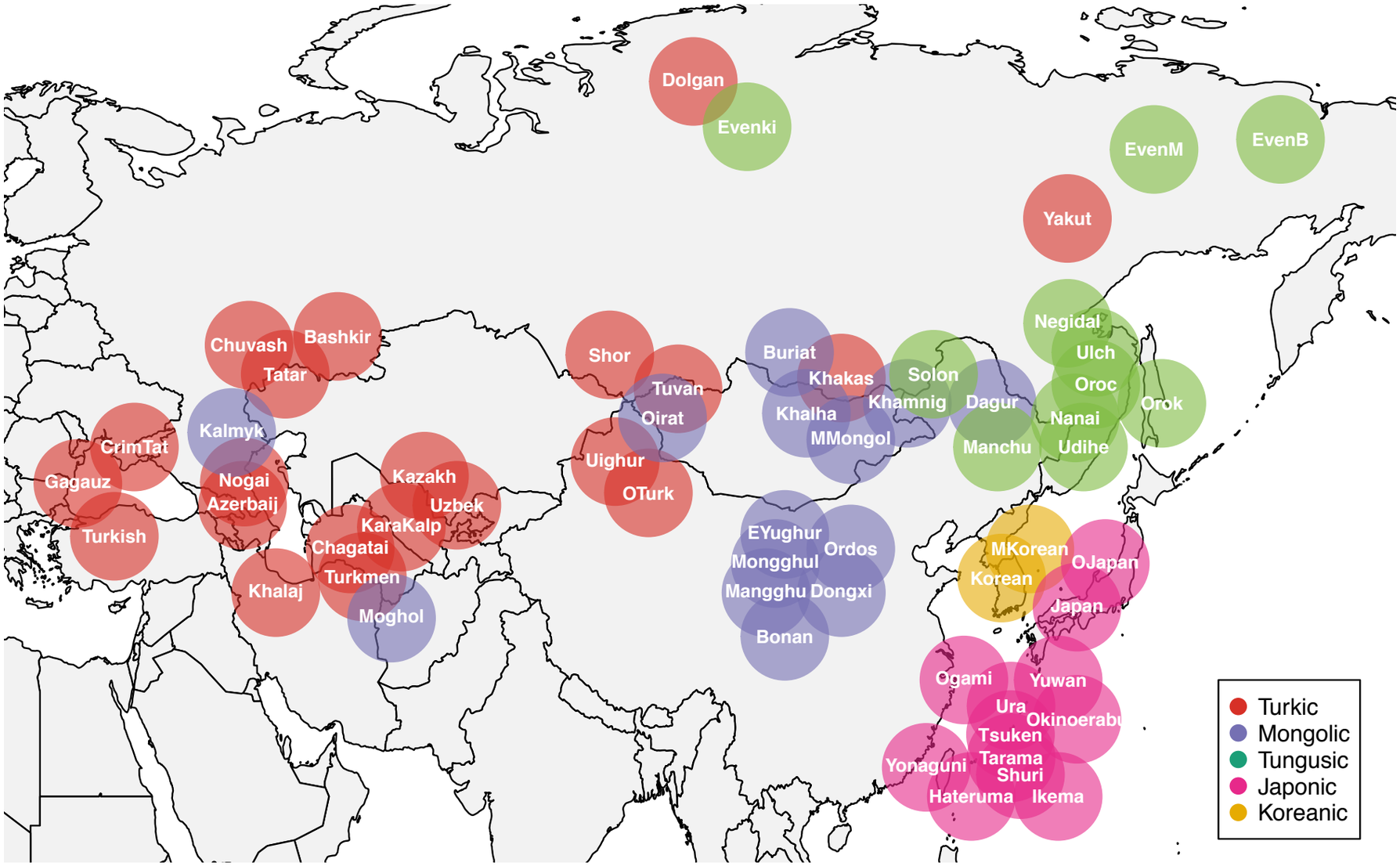
FSU Colloquium 9.07.2021

Idea

- Language structures are a disputed data type in historical linguistics
- It is difficult to disentangle inherited and diffused features.
- There are a few states, namely 1 and 0.
- It is not clear, which methods to use to represent the relationships between languages: trees are inappropriate because of unknown levels of borrowing and low number of character states, neighbour joining provides information on conflicting signal, but doesn't help us with timing and doesn't allow an evolutionary interpretation
- "Transeurasian" is a difficult language grouping

Language sample

- 60 languages spoken all over Eurasia, belonging to 5 language families, known as Altaic, Macro-Altaic, Transeurasian etc.



Idea

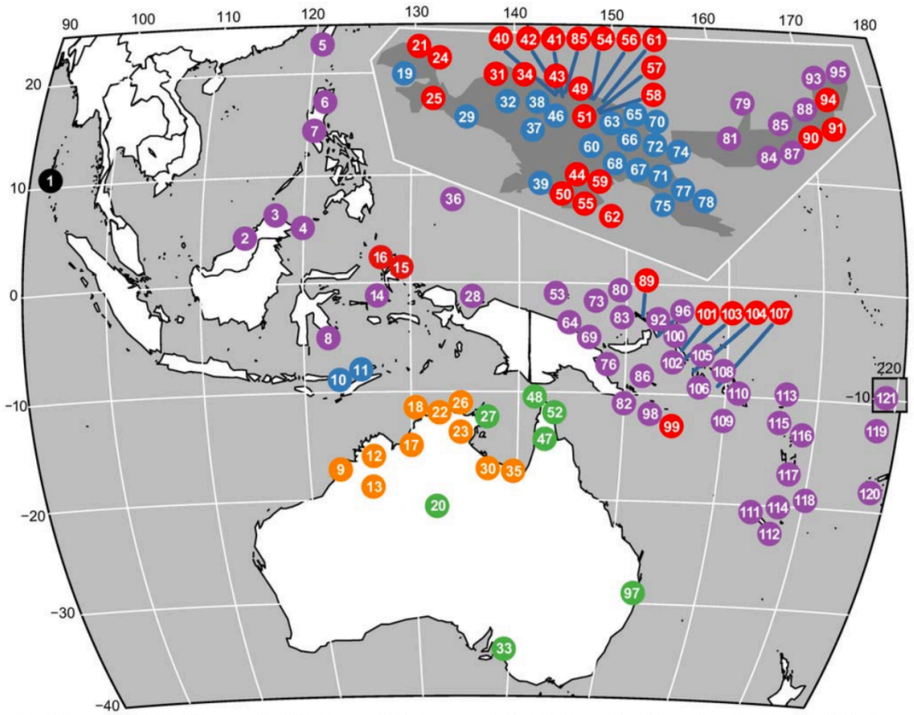
- The languages are known to be similar structurally, but their genealogical relationships are highly disputed
- Continental languages share a long history of contact, despite their vast spread (some Turkic and Tungusic languages co-exist even in the same village, in addition to Russian)
- There are some hypotheses of population movements from Northeast China to the Korean Peninsula and over to the Japanese archipelago -> relatedness of Koreanic and Japonic languages
- There is a long known history of contact between Japonic and Koreanic languages -> horizontal spread of language structures

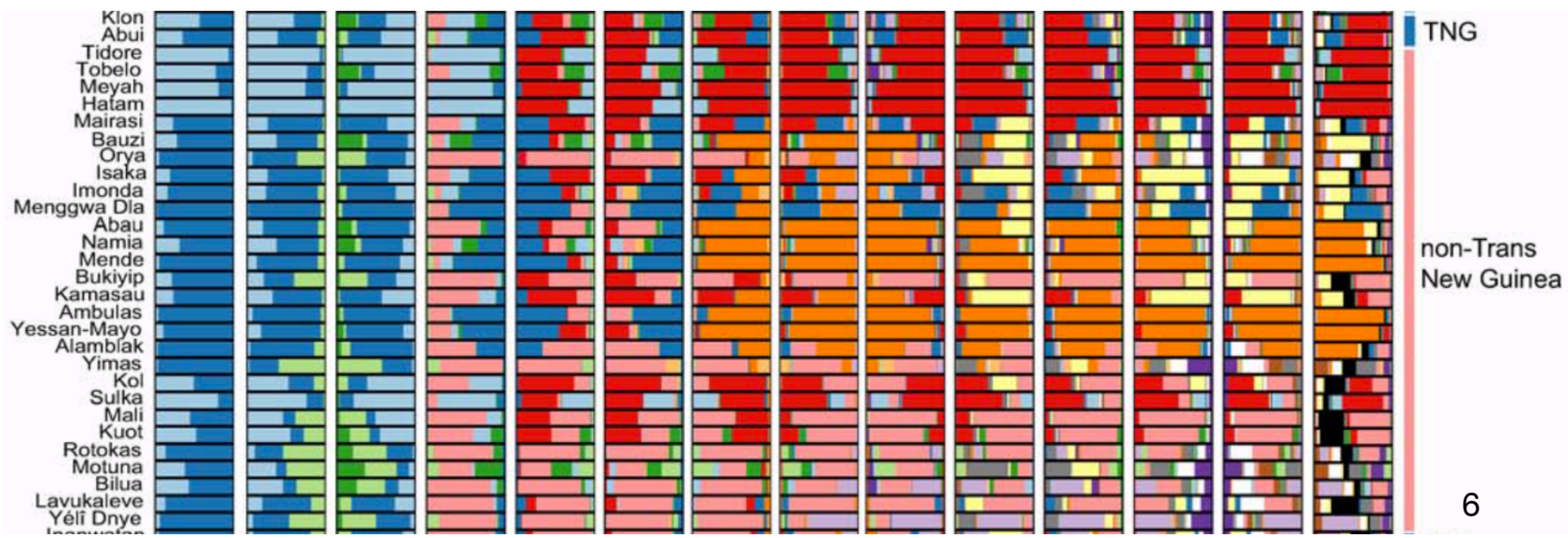
Idea

- Method from population genetics: STRUCTURE
- It allows us to make evolutionary inferences
- Even if it doesn't provide timing, we can see some historical events clearly: languages in heavy contact share half of their structures with their neighbours.

Idea

- In linguistics, STRUCTURE has been previously applied to Sahul languages
- Reesink et al. (2009) use it to determine the contribution of different linguistic lineages to the linguistic diversity of Sahul: they identify 10 populations, some of which can be aligned with known language families and groups and some of which have not been proposed yet.





Data coding: what is a structural feature?

(1) Udihe (Tungusic; Nikolaeva and Tolskaya 2001: 840)

mamasa	ule:-we	olokto-ini
old.woman	meat-ACC	cook-3SG

'The old woman is cooking meat.'

(2) Khalkha (Mongolic; Janhunen 2012: 246)

noxai	mo:r-i:g	barï-eb
dog	cat-ACC	catch-TERM

'The dog caught the cat.'

- Is pragmatically unmarked word order verb-final for transitive clauses? -> yes for both Udihe and Khalkha
- Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause? -> yes for Udihe, 1 -> no for Khalkha, 0

Data set

- The whole dataset divided into 3 sets:
- Phonology: 14 features
- Morphology: 103 features
- Syntax: 103 features

- The division is not ultimate, but based on the codings of the current languages, i.e. some morphological features might belong to the set on syntax for other languages and vice versa.

What is STRUCTURE?

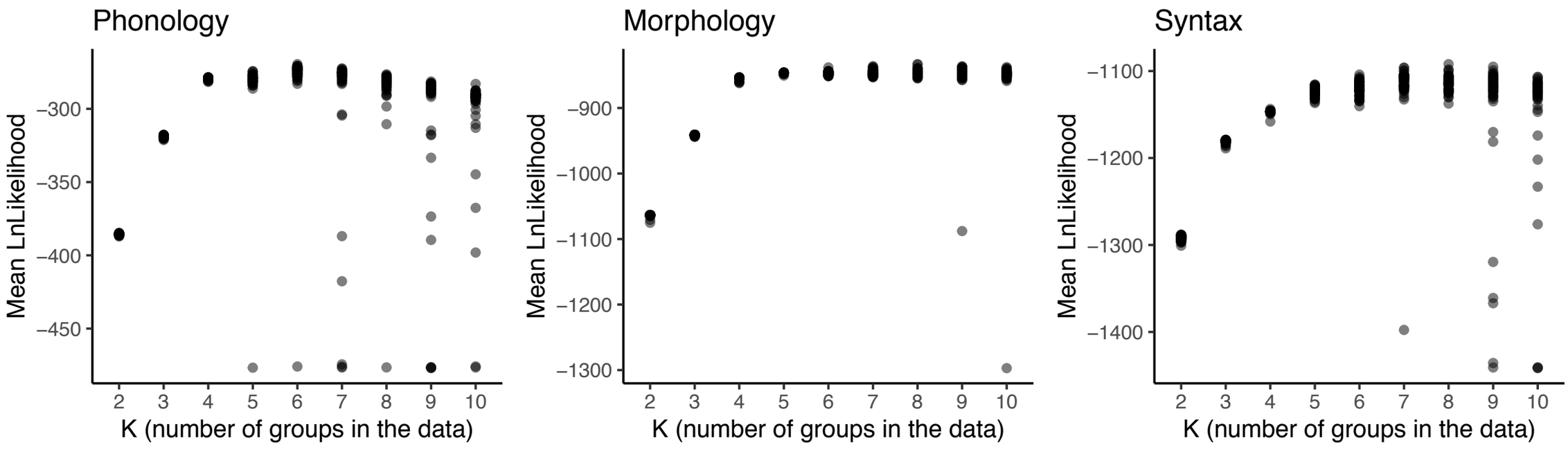
- An admixture model very popular in population genetics
- Allows for a high level of admixture between languages (does not "assign" a language strictly to one language family), therefore ideal for structural features
- Visualises sources of admixture: we can attribute different layers to different language families and determine the features that contribute to each layer
- We can compare the model performance for different assumed number of populations (K) and choose one based on this performance and common sense (the resulting K has to be interpretable in biological - in our case linguistic - sense)

Questions we can answer with STRUCTURE

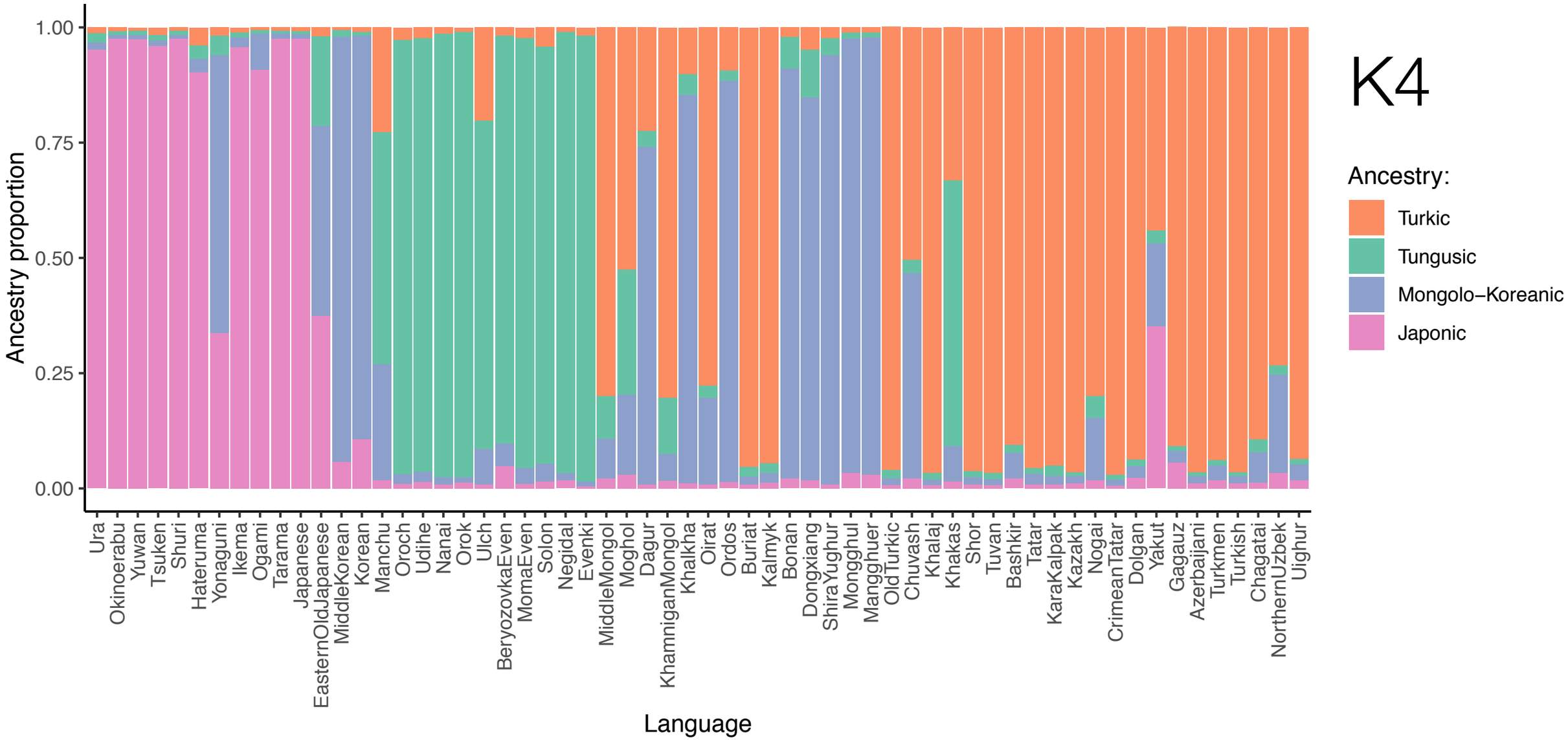
- How many groups can we identify? Do they correspond to the traditional language families?
- Does the division into groups depend on the linguistic level? I.e. do we get different groups if we look at phonology, morphology and syntax separately?
- Are there language levels that are more susceptible to borrowing? (I.e. do we see more admixture in any of the levels?)
- Are there languages with a heavy load of borrowing at any of the levels? What are the sources of this borrowing?

Which K is best?

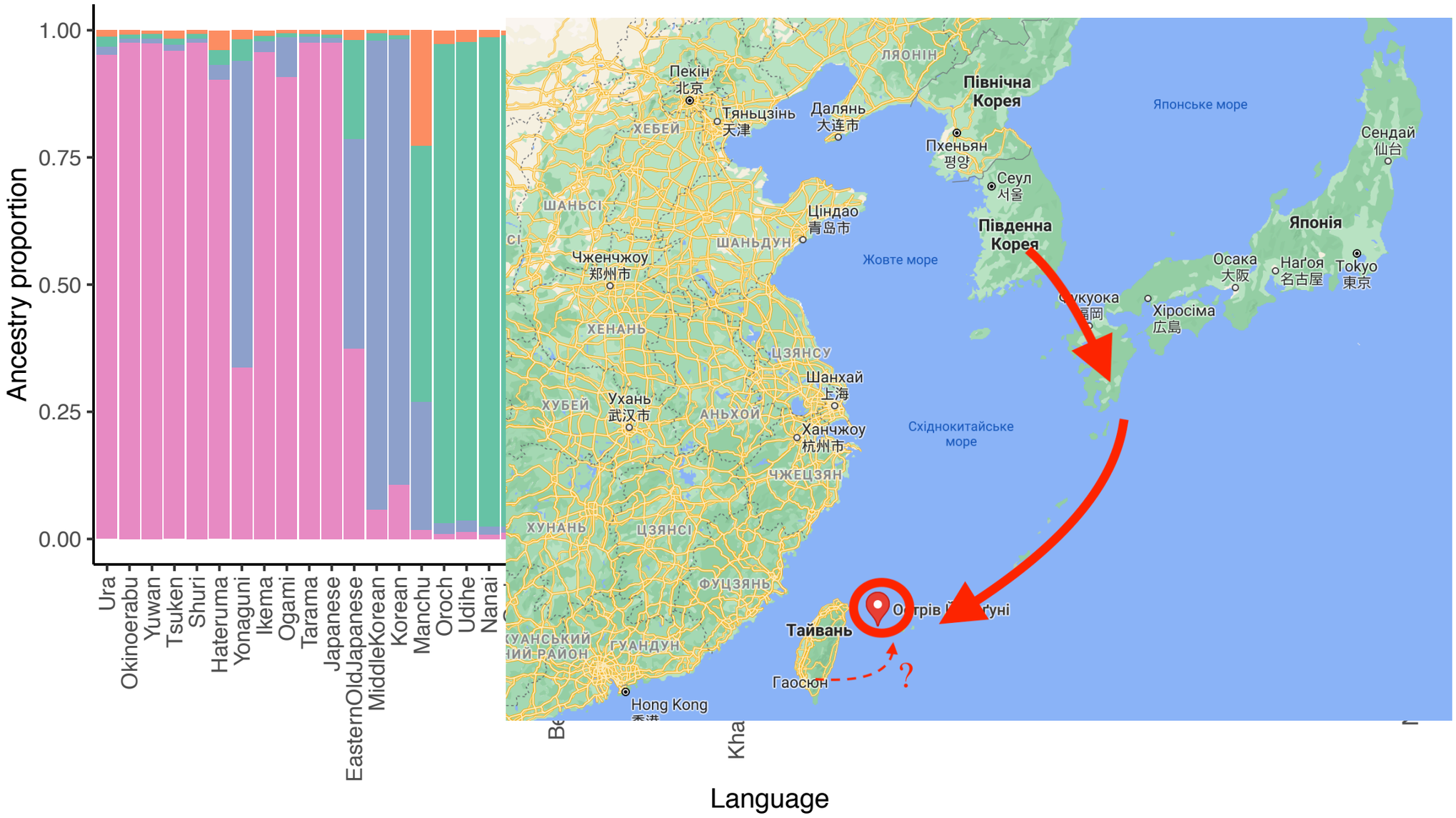
- Algorithm run for each subset (3) and each K (2-10) 50 times, yielding 1350 combinations
- The variation in LnLikelihood between K's has to be higher than between different runs
- A set of rules: an increase in variation between runs after the "true" K, a high jump before the "true" K — the LnLikelihood "plateaus" after the "true" K



Phonology


K4
Ancestry:
Turkic
Tungusic
Mongolo−Koreanic
Japonic
Ancestry proportion
1.00
0.75
0.50
0.25
0.00
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khalaj
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Nogai
CrimeanTatar
Dolgan
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur
Language

Phonology

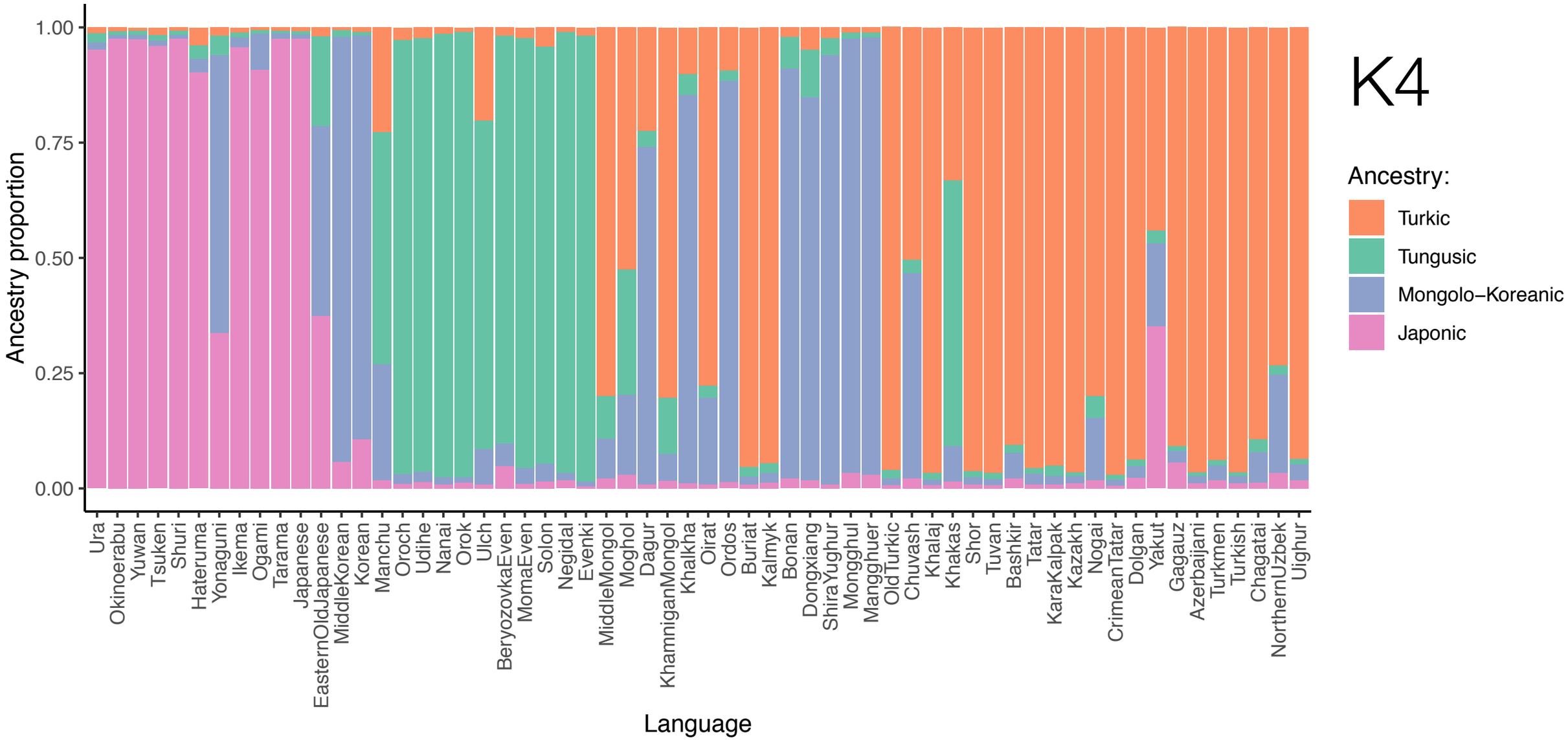


K4

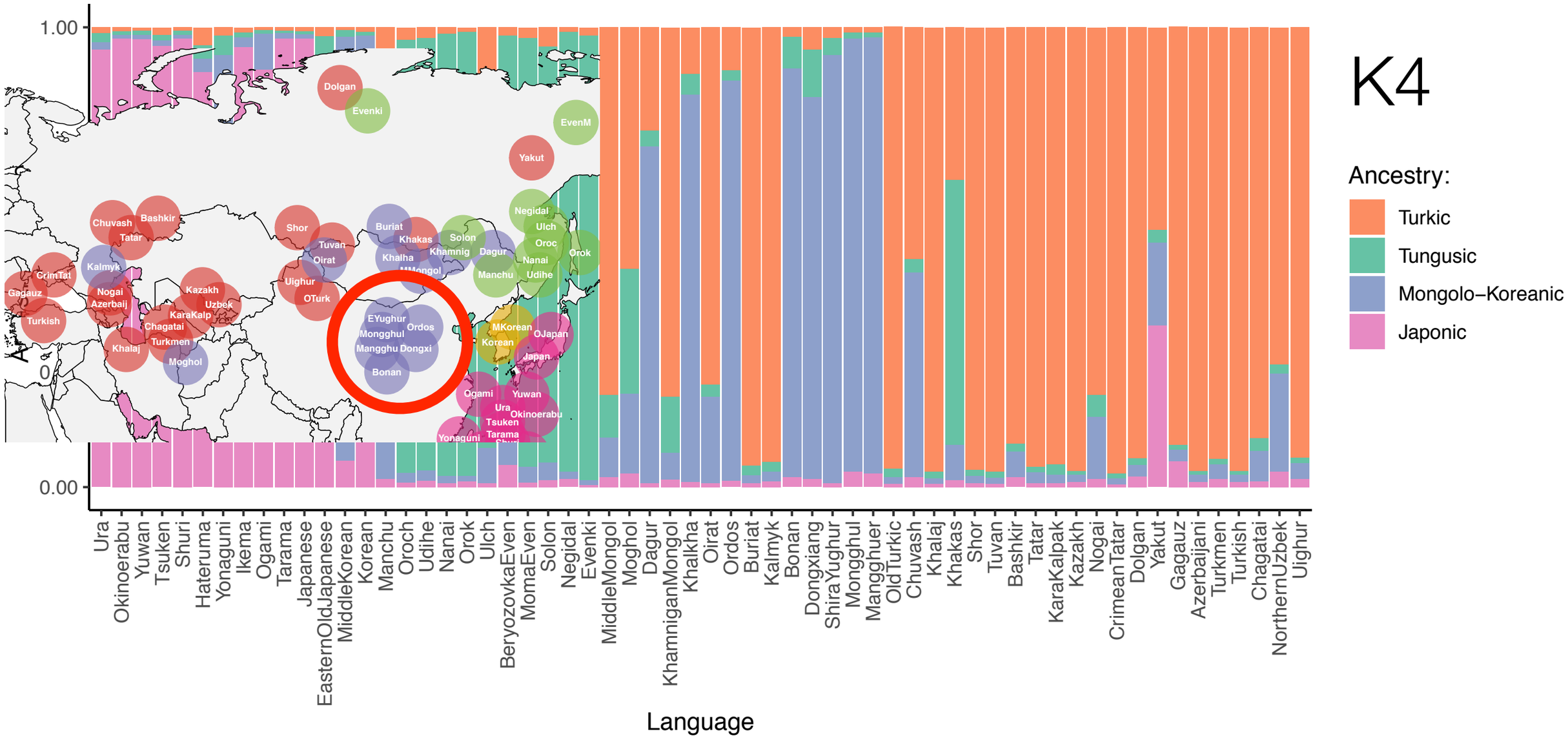
Ancestry:

- Turkic
- Tungusic
- Mongolo–Koreanic
- Japonic

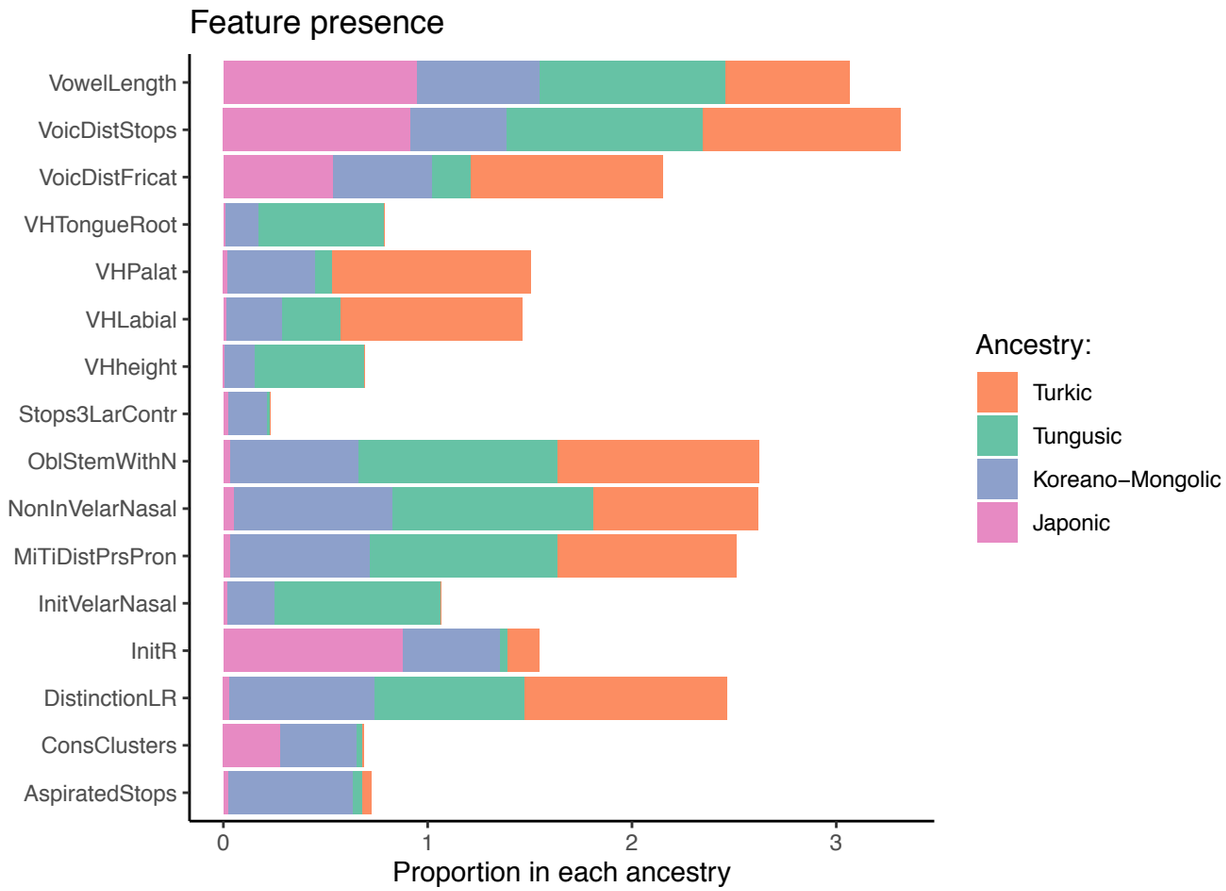
Phonology

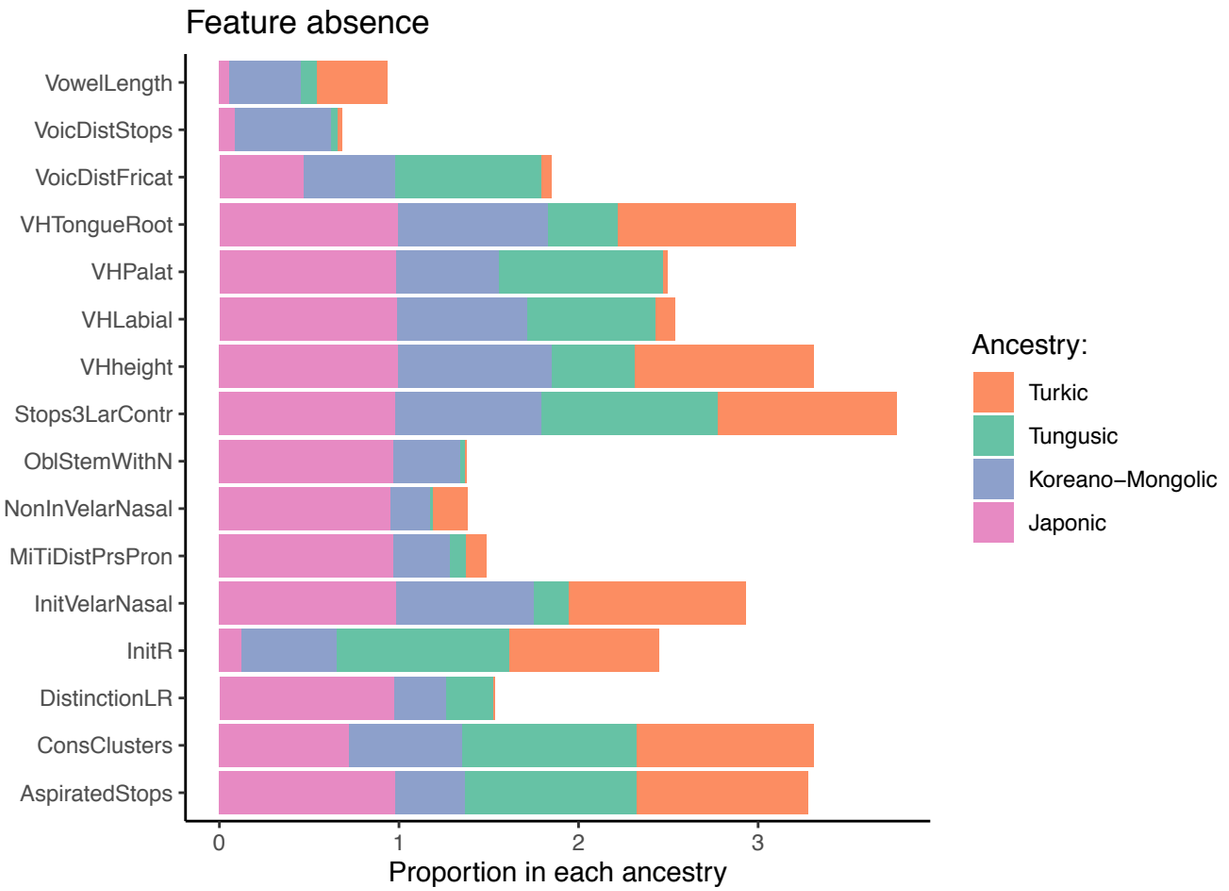

K4
Ancestry:
Turkic
Tungusic
Mongolo−Koreanic
Japonic
Ancestry proportion
1.00
0.75
0.50
0.25
0.00
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khalaj
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Nogai
CrimeanTatar
Dolgan
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur
Language

Phonology

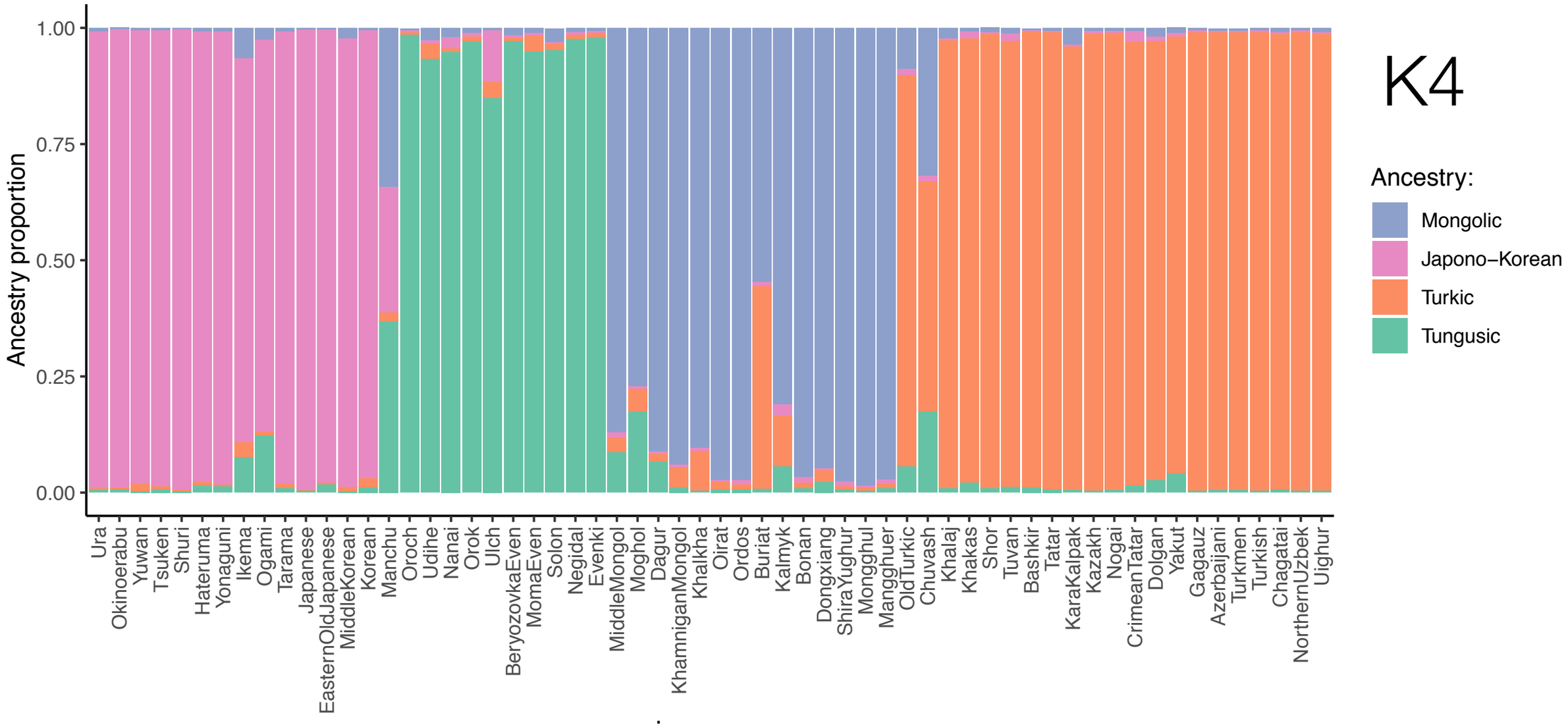

1.00
0
0.00
Ancestry:
Turkic
Tungusic
Mongolo−Koreanic
Japonic
K4
Dolgan
Evenki
EvenM
Yakut
Chuvash
Bashkir
Tatar
Shor
Tuvan
Oirat
Buriat
Khakas
Khamnig
Solon
Dagur
Khalha
MMongol
Manchu
Negidal
Ulch
Oroc
Nanai
Udihe
Orok
Kalmyk
CrimTat
Gagauz
Nogai
Azerbaij
Kazakh
Uzbek
KaraKalp
Uighur
OTurk
Turkish
Chagatai
Turkmen
Khalaj
Moghol
EYughur
Mongghul
Ordos
Mangghu
Dongxi
Bonan
MKorean
Korean
OJapan
Japan
Ogami
Yuwan
Ura
Okinoerabu
Tsuken
Yonaguni
Tarama
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khalaj
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Nogai
CrimeanTatar
Dolgan
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur
Language

Contribution of features to phonological ancestry

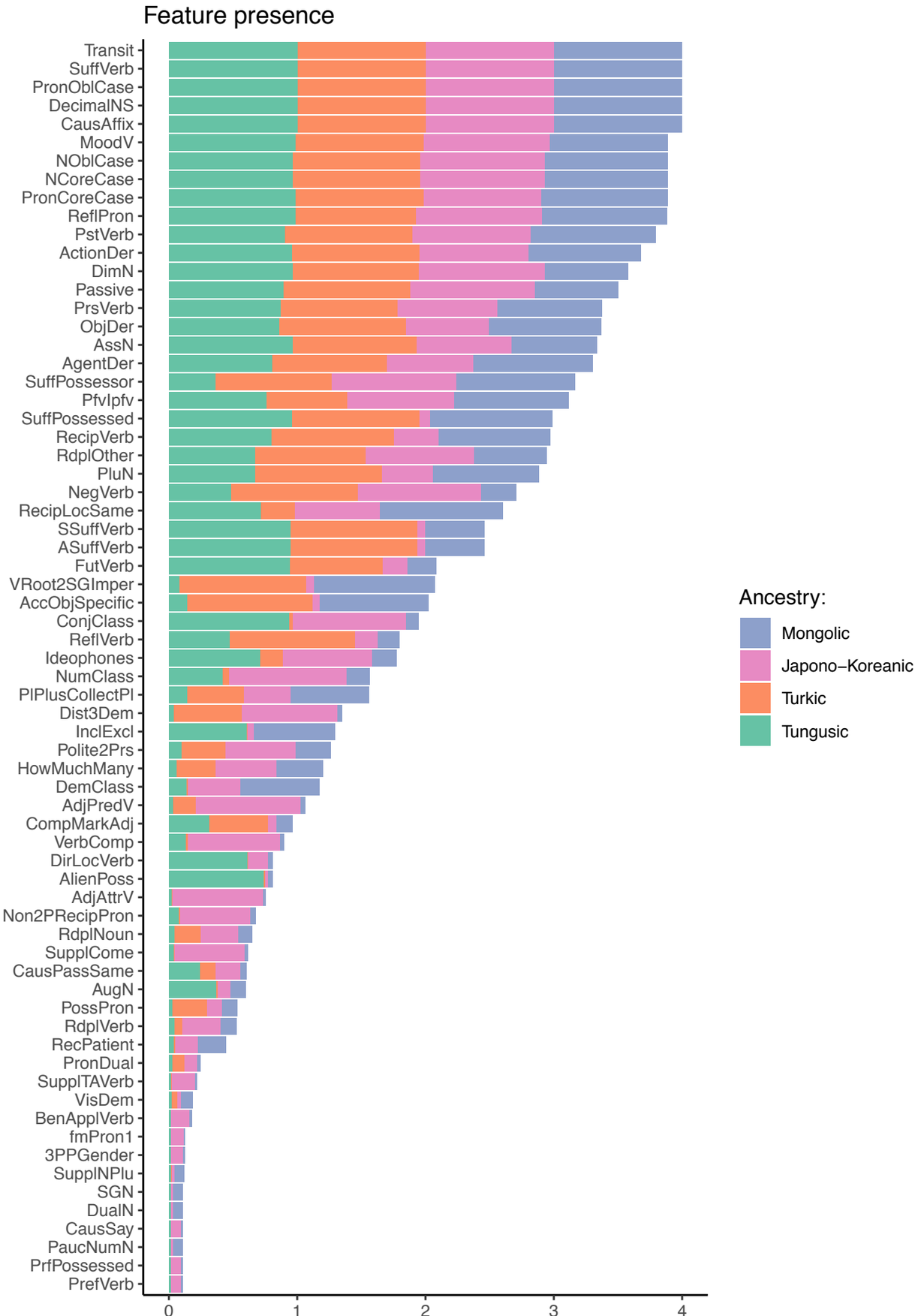


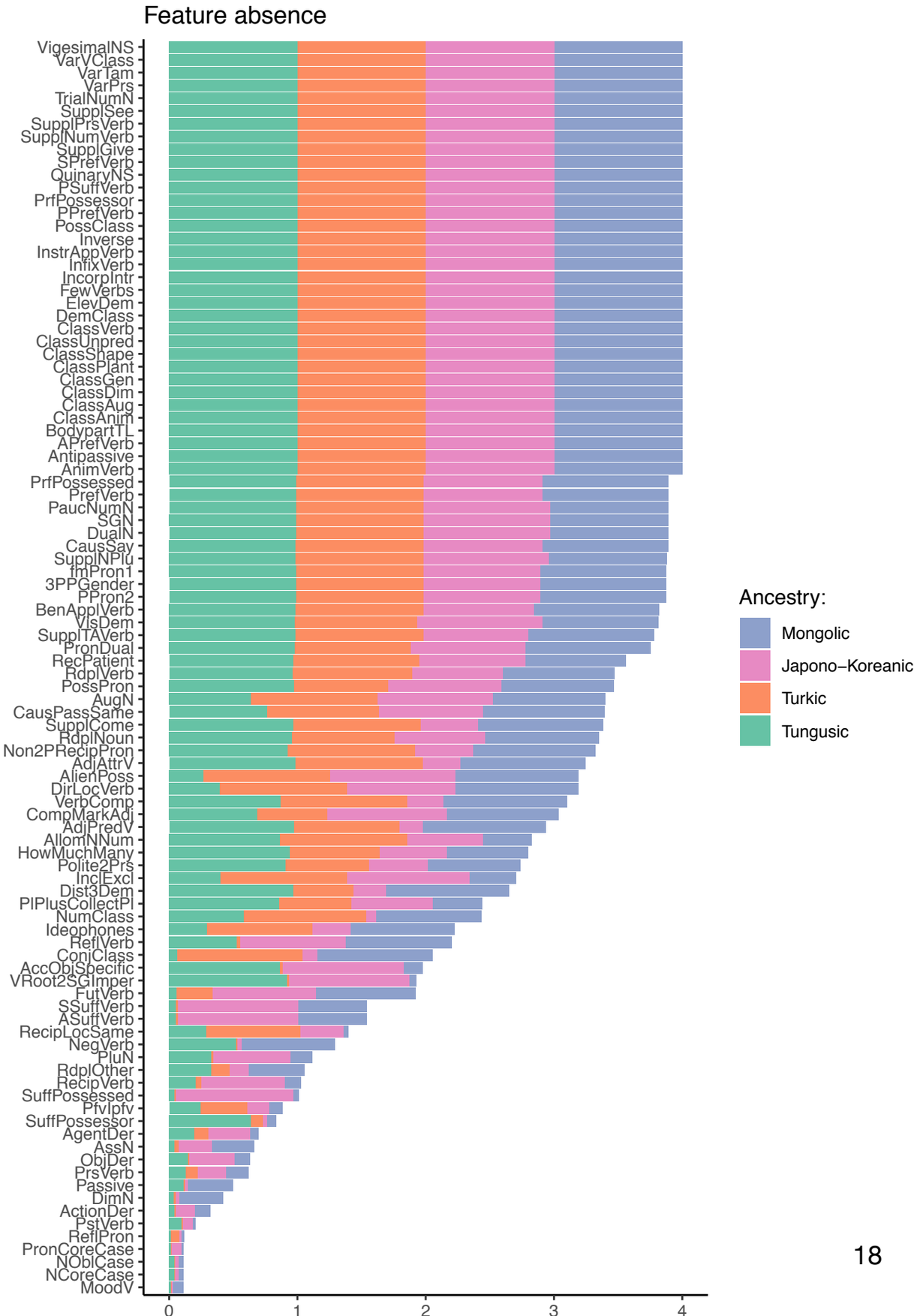


Morphology

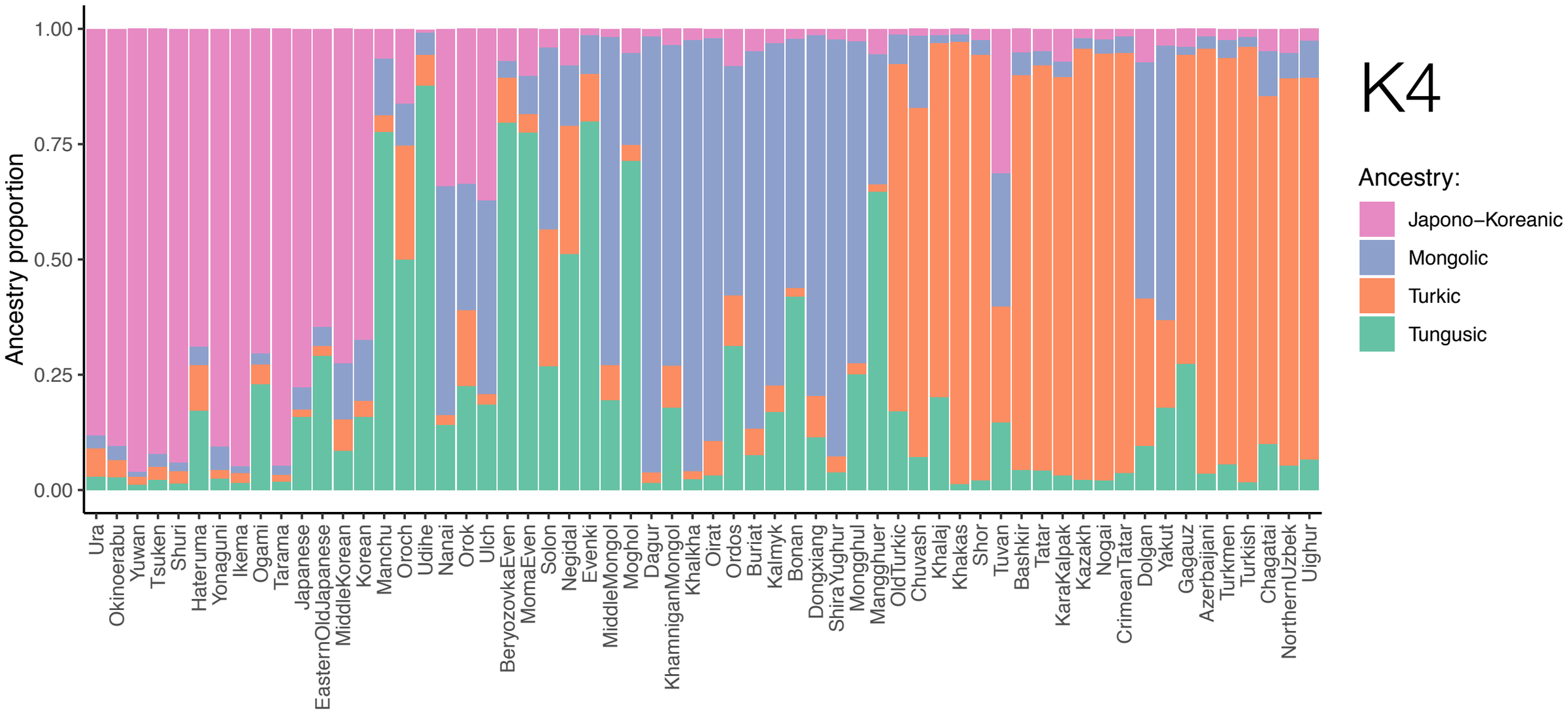

1.00
0.75
0.50
0.25
0.00
Ancestry proportion
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khalaj
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Nogai
CrimeanTatar
Dolgan
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur
K4
Ancestry:
Mongolic
Japono−Korean
Turkic
Tungusic

Contribution of features to morphological ancestry

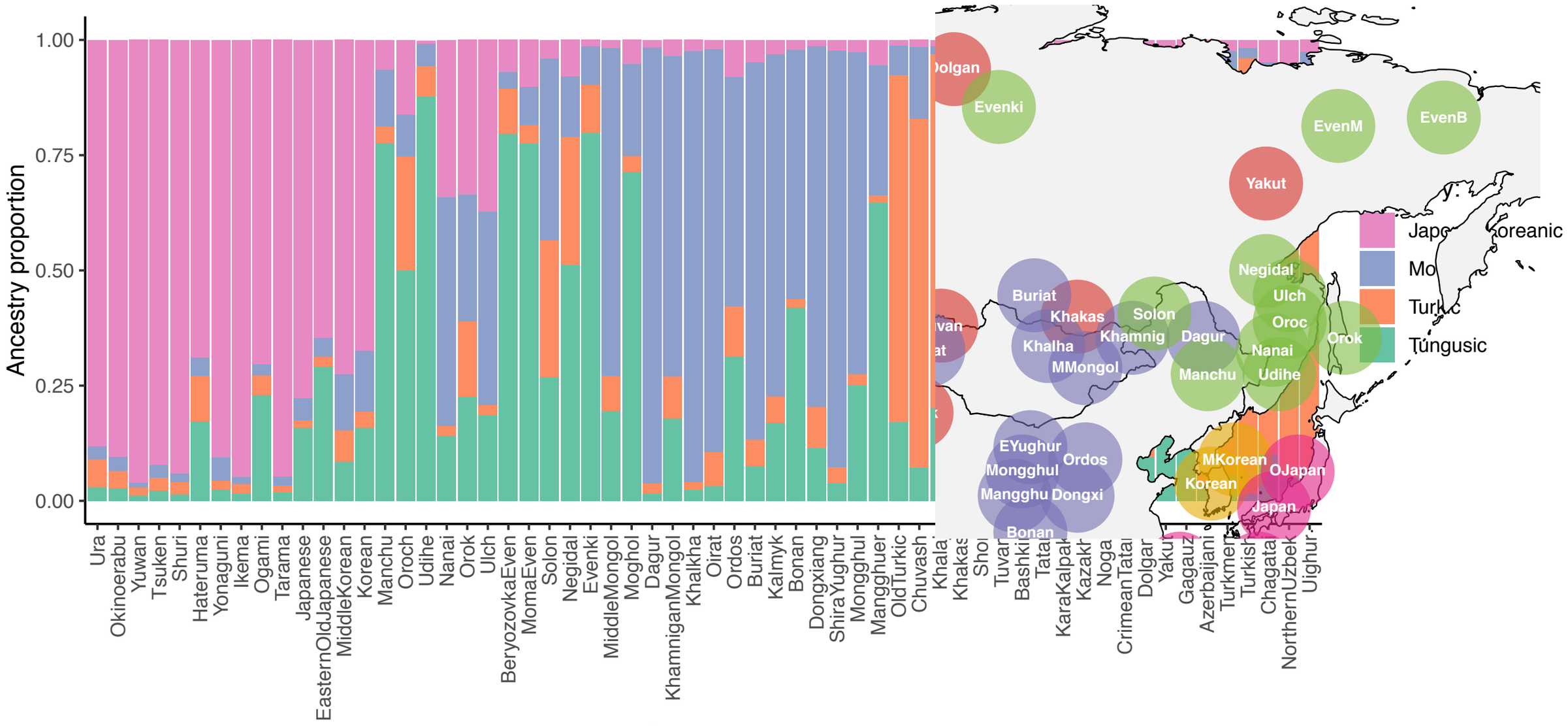




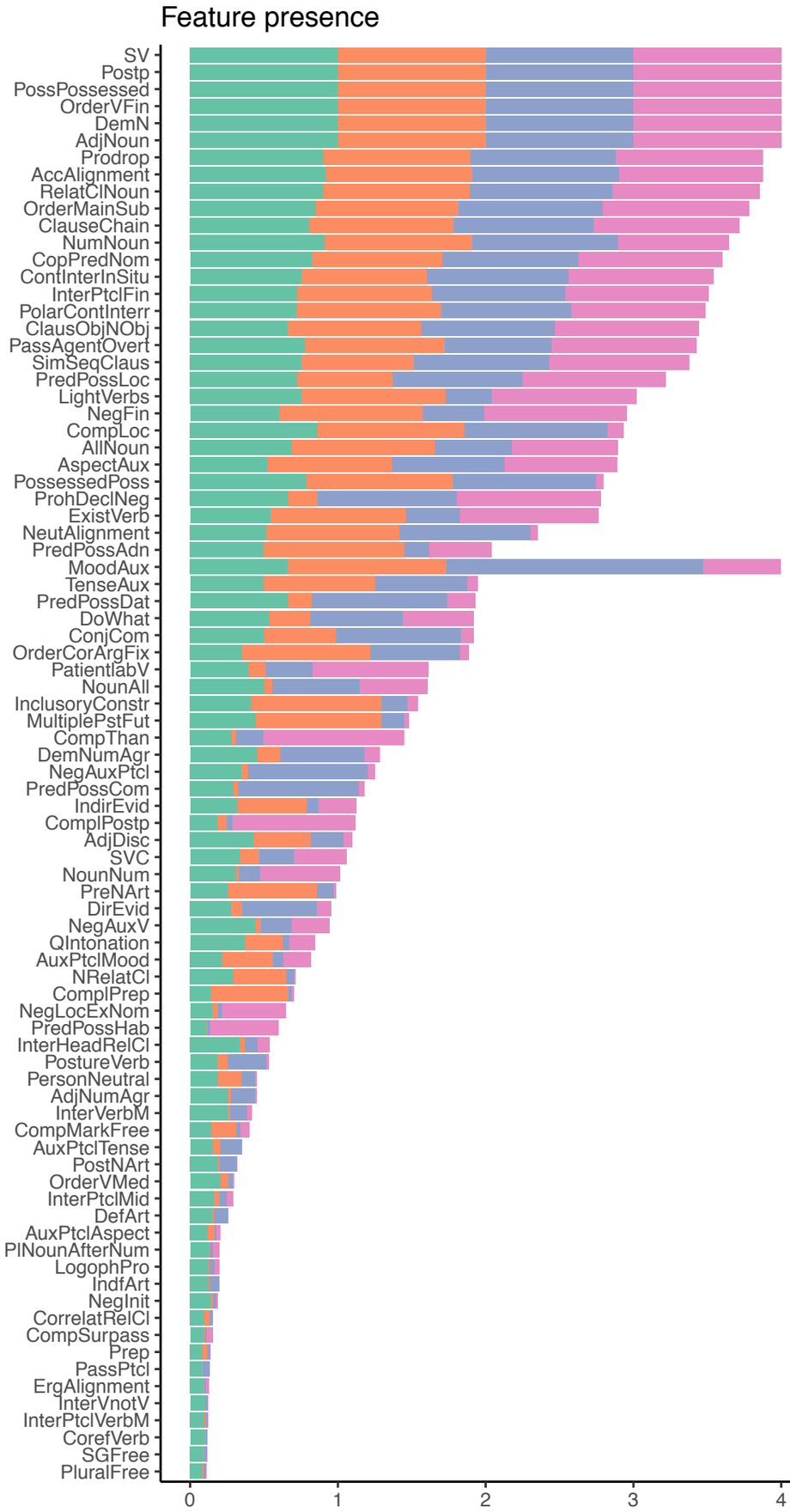
Syntax

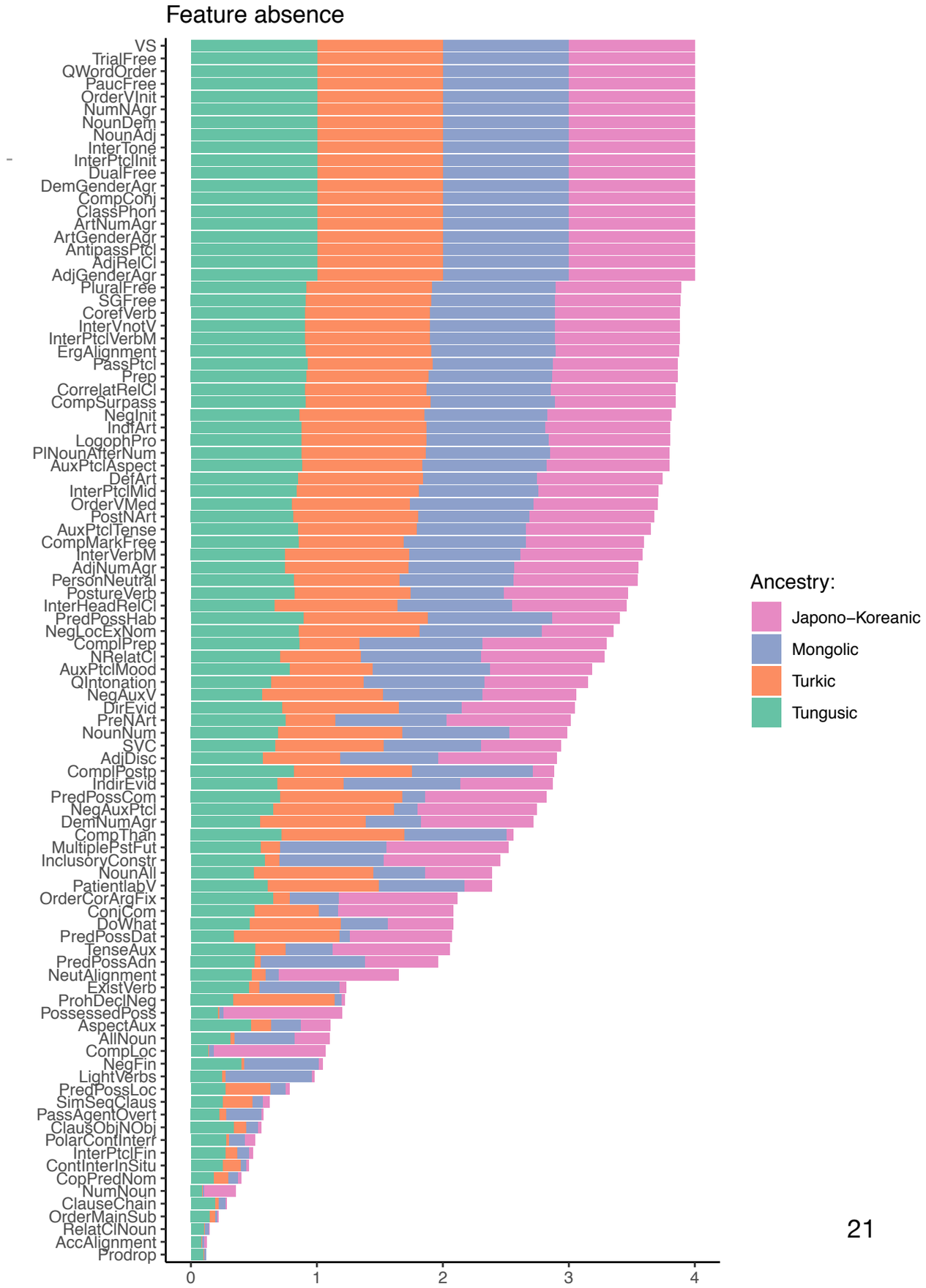

K4
Ancestry:
Japono−Koreanic
Mongolic
Turkic
Tungusic
Ancestry proportion
1.00
0.75
0.50
0.25
0.00
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khalaj
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Nogai
CrimeanTatar
Dolgan
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur

Syntax


1.00
0.75
0.50
0.25
0.00
Ancestry proportion
Ura
Okinoerabu
Yuwan
Tsuken
Shuri
Hateruma
Yonaguni
Ikema
Ogami
Tarama
Japanese
EasternOldJapanese
MiddleKorean
Korean
Manchu
Oroch
Udihe
Nanai
Orok
Ulch
BeryozovkaEven
MomaEven
Solon
Negidal
Evenki
MiddleMongol
Moghol
Dagur
KhamniganMongol
Khalkha
Oirat
Ordos
Buriat
Kalmyk
Bonan
Dongxiang
ShiraYughur
Mongghul
Mangghuer
OldTurkic
Chuvash
Khala
Khakas
Shor
Tuvan
Bashkir
Tatar
KaraKalpak
Kazakh
Noga
CrimeanTatar
Dolgar
Yakut
Gagauz
Azerbaijani
Turkmen
Turkish
Chagatai
NorthernUzbek
Uighur
olgan
Evenki
EvenM
EvenB
Yakut
Negidal
Ulch
Oroc
Orok
Nanai
Udihe
Manchu
Dagur
Solon
Khamnig
Buriat
Khakas
Khalha
MMongol
EYughur
Ordos
Mongghul
Mangghu
Dongxi
Bonan
MKorean
Korean
OJapan
Japan
Japo
oreanic
Mo
Turkic
Tungusic

Contribution of features to syntactical ancestry





Conclusions

- Phonology is determined geographically: very accurately and very systematically.
- Morphology is determined genealogically: we see four clear groupings with only a few exceptions
- Syntax shows the highest levels of admixture and an overall trace of common ancestry with a center in Tungusic
- Interpretation is not 100% reliable because of the sampling: the more languages from the are we take, the more reliable the results: e.g. we don’t see traces of Uralic or other languages of Siberia if we don’t include them in the sample and these traces might be masked as one of the determined ancestries (appear e.g. Mongolic, where it is Uralic in reality)

Conclusions

STRUCTURE can become a useful tool for linguists:

- well applicable for language structures, but also maybe for other kinds of data
- we can verify the existing language families or detect language contact events
- we can also test hypotheses about migrations of speaker groups
- especially nice to detect substratum in case of population replacement and language shift
- a drawback: results heavily dependent on the sample (we can be misguided by excluding particular languages)