Linguistic Areas and Prehistoric Migrations

Funding Body

This research was funded by the Centre for Language Studies at Radboud University, Nijmegen, the Netherlands.

International Max Planck Research School (IMPRS) for Language Sciences

The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics

Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-55-4
Cover design and lay-out by Jeremy Collins
Printed and bound by Ipskamp Drukkers, Enschede

The research reported in this thesis was conducted at the Center for Language Studies at Radboud University and the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands.

Linguistic Areas and Prehistoric Migrations

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 19 februari 2024
om 12.30 uur precies

door

Jeremy Charles Collins
geboren op 24 december 1987
te Parijs, Frankrijk

Promotoren:

Prof. dr. P.C. Muysken (†)

Prof. dr. S.C. Levinson

Prof. dr. D. Dediu (Universitat de Barcelona, Spanje)

Manuscriptcommissie

Prof. dr. A.M.C. van Kemenade (voorzitter)

Prof. dr. M.J. Dunn (Uppsala Universitet, Zweden)

Prof. dr. J.-M. List (Universität Passau, Duitsland)

Prof. dr. R.W.N.M. van Hout

Dr. M. Dingemanse

Acknowledgements

I would like to thank my supervisors Pieter Muysken, Stephen Levinson and Dan Dediu for all of their support and patience during my project. Pieter lent his expertise on language contact to my project as well as providing extensive feedback and encouragement on many drafts of my thesis. Steve provided a lot of useful critical feedback and support as well, especially in the last rounds of submission, and his work inspired many of the basic conceptual questions such as the use of phylogenetics to analyse structural features. Dan provided a lot of help in improving the conceptual clarity of many sections, the statistical analyses and discussions of the the relationship between phonological features and genetics, for which his work (e.g. Dediu and Ladd 2007) was the primary inspiration.

I would like to thank Nick Enfield too, who took me under his wing as his student when I applied to Radboud and during my first two years there, and invited me to join the Language & Cognition group in the MPI. His work on semantic typology, linguistic areas and Southeast Asia was my main reason for applying for Radboud and the inspiration for my own fieldwork on Tai-Kadai and Austroasiatic languages, as well as for my interest in formalising the concept of linguistic areas.

I had a wonderful four years in the MPI, with the L&C meetings on Wednesdays being a highlight of the week for me, as were conversations that I had with different members of the department such as Seán Roberts, Mark Dingemanse, Simeon Floyd, Harald Hammarström, Francisco Torreira, Elisabeth Norcliffe, Asifa Majid, Kobin Kendrick, Judith Holler, Gunter Senft, and my co-PhD students Emma Valtersson, Giovanni Rossi, Tayo Neumann, Ewelina Wnuk, Matthias Barthel, Gabriela Garrido, Sebastian Sauppe, Kang-Suk Byun, and others. Above all I thank Stephen Levinson for creating such a stimulating and friendly research environment and for inviting me into the group.

Beyond the L&C department, I spent time with many great people in the MPI and Radboud whose company over beers made living in Nijmegen more fun, among them Tomas Lehečka, Gwilym Lockwood, Elliot Sollis, Johanne Tromp, Sarah Graham, Amie Fairs, Joshua Birchall, and Tulya Kavaklıoğlu. Seán Roberts deserves special mention for his hilarious movie nights and Space Ghost Train Halloween parties.

Thank you to Michael Dunn too for inspiring me to pursue phylogenetics (especially structural phylogenetics) and for good conversations over coffee in the Midi Planck building and at conferences in various parts of the world. I have fond memories of exchanging ideas late into the night in the Midi Planck with Damián Blasi during that period as well, as well as of all of our adventures in Auckland, Hong Kong, Leipzig, Vienna and other places.

Thank you to the members of the Typology Clubhouse, the shared office that housed Luis Miguel Rojas Berscia and me originally and later Hedvig Skirgård, Jakob Lesage, Jesse Peacock and Hilário de Sousa, for making working in the office so fun, for happy memories such as the Santa Lucia parties, and for your continued friendship.

The workshops of the Glottobank group organised by Russell Gray and Quentin Atkinson were an experience that I will always cherish. I worked on the Grambank database led by Hedvig Skirgård, Hannah Haynie, Harald Hammarström, and Stephen Levinson, which has since become one my main academic foci. I worked on analyses with Quentin Atkinson and Luke Maurits (who contributed to the analysis of phonotactics data in Chapter 3), and thanks to Remco Bouckaert too for guidance on issues with using BEAST and the spherical phylogeography package. I enjoyed conversations and spending time with other participants of these workshops too: Simon Greenhill, Martin Haspelmath, Michael Dunn, Damián Blasi, Nicholas Evans, Mattis List, Robert Forkel, Thiago Chacon, Nataliia Hübler, and Kyla Quinn. Thank you to Jay Latarche for help in co-managing the SOAS coding group and largely taking over my patronage of Grambank features.

I collaborated briefly with Mark Donohue on an analysis of the World Phonotactics Database in section 3 of Chapter 2 on the spread of tonal languages; although the analysis is my own, he helped to rewrite parts to make it more presentable for publication, and contributed the map of tonal languages in Figure 1.1 in Chapter 1. I greatly appreciated his quick response times and willingness to help, as well as his enthusiasm for the subject, frank and humorous communication style, and expertise on matters of phonotactics and language contact.

I would like to thank the Manuscript Committee of this thesis: Mark Dingemanse, Mattis List, Ans van Kemenade, Roeland van Hout, and Michael Dunn, and Luis Miguel Rojas Berscia as well for being part of the Doctoral Examination Committee. The manuscript also benefited from Brigitte Pakendorf's feedback on the first submitted version.

I am very grateful to my paranymphs Gabriela Garrido and Tayo Neumann, particularly to Gaby for help with organisation, and to Tayo for making the time to come from Berlin; Kevin Lam for help and encouragement with all issues related to the IMPRS, and him and Ludy Cilissen for help with the design and printing of the dissertation; and Monique Wakkerman at Radboud for very patient handling of administrative matters and encouragement during the latter stages of the PhD process, and similarly to Edith Sjoerdsma in the MPI. Thanks too to Pim Levelt, whose piano in the basement of the MPI provided me with many hours of joy.

Thank you to my informants in Xishuangbanna, Ximeng, and the other locations in Chapter 4, particularly to Yuni in Mangjing for her patience in answering questions for a few hours and for introducing me to her relatives in other villages.

I would like to mention some good friends here, such as Tayo Neumann, who I had memorable times with in Shanghai, Hong Kong and Berlin, exchanging daily jokes on WhatsApp and having enjoyable arguments about geopolitics; Peter and Morgan Jackson who hosted me a few times in Shanghai, and Ed and Sarah Jackson; Aleem Meghji, Alexander & Ally Goldsmith, and Sing Sing Ngai who regularly badgered me about when I would finish; and people from my time in Nijmegen such as Damián Blasi, Hedvig Skirgård, Jakob Lesage, Hilário de Sousa and Luis Miguel Rojas Berscia.

Finally, love to all my family and thank you for your support through this time: Mum, Dad, Paul, Grandma, Grandpa, Granny, Grandad, Cherry, Naomi, Nick, Megan, and Michaela.

Linguistic Areas and Prehistoric Migrations

Jeremy Collins

jeremyccollins@gmail.com

Contents

Chapter 1: Introduction

1. Summary

One of the greatest achievements of linguistics has been the discovery of language families. Over five hundred languages stretching from Western Europe to Nepal descend from a common ancestor, whose existence was first demonstrated by Sir William Jones (Jones 1786) and which was given the name 'Proto-Indo-European' by Thomas Young (Robinson 2007). Another example is the Austronesian languages of the Pacific and the Indian Ocean, from Madagascar to Hawaii, which have been shown to be the descendants of a proto-language probably spoken in Taiwan (Schmidt 1899; Blust 2013). The primary evidence for language families is shared innovations in their core vocabulary, as demonstrated for instance by the word for 'father' across languages of Europe and India (German *Vater*, Latin *pater*, Spanish *padre*, Ancient Greek πατήρ, Tocharian *pacer*, Persian *pedar*, Vedic Sanskrit *pitā́*), and by systematic patterns of sound change such as Grimm's law (Campbell 2004:49).

This thesis is about the history of languages, but from a different perspective: the history of language structures. These are properties of languages such as grammatical features, and phonological features, as documented by linguistic databases such as the World Atlas of Language Structures (Dryer and Haspelmath 2013) and the World Phonotactics Database (Donohue et al. 2013): properties such as basic word order, the presence or absence of particular phonemes, and so on.

Like words, structural properties of languages may be informative about language history, and about the history of human migration. For example, some features of language have striking geographical distributions, such as in Figure 1.1 which shows where tonal languages are found. Tonal languages such as Mandarin make distinctions between morphemes using pitch, for example *mā* (high tone) 'mother', *má* (rising tone) 'numb', *mǎ* (low dipping tone) 'horse', and *mà* (falling tone) 'scold'. About a third of languages are tonal, but they are not evenly distributed, instead appearing in large clusters around the equator.

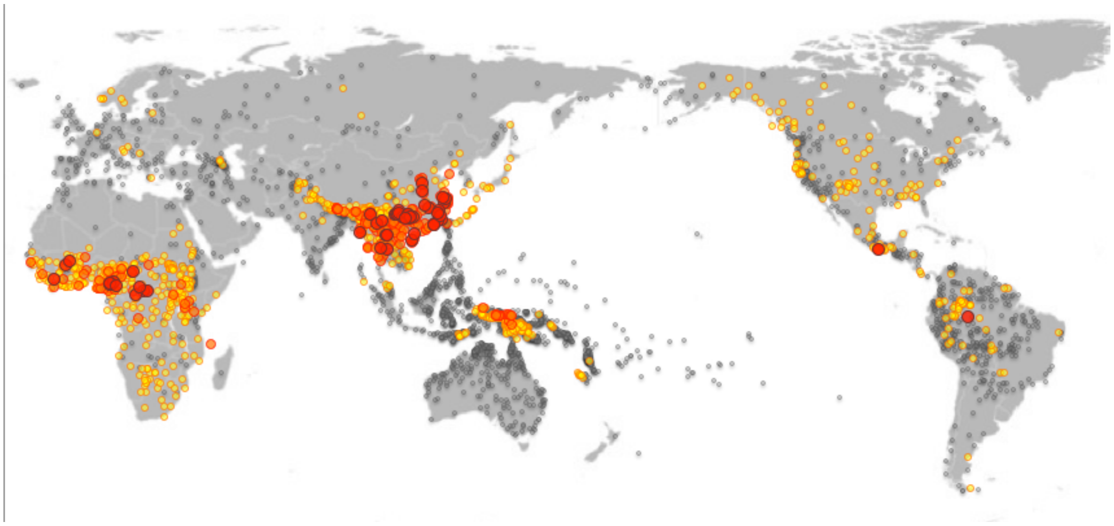


Figure 1.1: A map of tonal languages produced by Mark Donohue using data from the World Phonotactics Database (Donohue et al. 2013), with number of tones indicated by size and darkness of the colour of the circles (red = four or more, orange = 3, yellow = 2, grey = non-tonal)

Why does this linguistic feature cluster so strikingly? One of the main ideas pursued in this thesis is that features such as tone show the history of migration. The clustering in the image partly reflects the history of large tonal language families which dominate in Africa and Southeast Asia, such as Niger-Congo and Sino-Tibetan. But it also shows that there has been contact between different language families, with tone spreading between them; for instance, Chinese has influenced Vietnamese and other languages in Southeast Asia, primarily through loan words (Haudricort 1954).

This suggests that features such as tone can be informative about how languages have been influencing each other, an idea explored further in Chapter 2. The way that tone clusters has also drawn other hypotheses, such as the idea that tone may be influenced by climate. Since tonal languages are found primarily around the equator in humid climates, it has been suggested that humid environments are more conducive to the precise control of phonation that tonal languages require (Everett, Blasi and Roberts 2015). The experimental evidence that they cite shows that dry air makes precise movements of the vocal cords more difficult, making it less likely that tonal languages would be found in arid regions. Chapter 2 discusses this hypothesis further and suggests that, contrary to their argument, the clustering of tone in these areas may simply be due to language contact, and the fact that humid environments have more languages than non-humid environments; simulations on how tone could have spread randomly show that the correlation between tone and humidity may not be causal, but could be an artefact.

If individual features such as tone can be informative about language history, then the question arises about how much we can learn about language history if we take evidence from many different features. This question is explored in Chapter 3, which takes 184 features from the World Phonotactics Database, and uses a phylogenetic method to analyse how languages form clusters. The reason for using a phylogenetic method is explained in sections 2.1 and 2.2, since it has been used successfully for analysing vocabulary data and for providing a statistical basis for language families (e.g. Gray and Jordan 2000, Bouckaert et al. 2012, Jäger 2015); and also applied to structural data, with intriguing results (e.g. Dunn et al. 2005, Dediu and Levinson 2012, Jäger and Wichmann 2016). The questions examined in this chapter are: what would happen if data from the World Phonotactics Database were analysed using this type of algorithm, what kind of clusters emerge, and how much these clusters can tell us about language history, either of language families or of contact. The chapter concludes that many interesting clusters emerge, such as South China; India/Southeast Asia; East India; North China; the Caucasus/Middle East; and Indo-European/Turkic. Zones such as these are areas where unrelated languages have phonological properties in common, suggesting that they have been interacting with each other.

In contrast with this continent-scale study, Chapter 4 is a fieldwork study on East Palaungic languages in Southwest China to research how linguistic features spread through closely related neighbouring languages. The main finding that there is surprising diversity in these languages, and that this diversity reflects the diversity of Southeast Asia as a whole; they vary in their basic word orders, in their use of numeral classifiers, and in their semantics. There is evidence that this diversity is partly created by language contact: several features seem to be spreading into Palaungic languages from nearby, unrelated Tai-Kadai languages. I present evidence for this from census data, showing that the distribution of typically Tai features is correlated with the proportion of Tai speakers nearby.

This leads onto a hypothesis pursued in Chapter 5, that the distribution of linguistic features is related to migration: for example, if tone has spread from one language into an unrelated language, then it may indicate that people have been moving from one community to the other. This hypothesis is explored by using genetic data to reconstruct migrations of people in Eurasia over the last 100,000 years, specifically of mitochondrial DNA lineages. The reconstructed migration data is then compared with measures of linguistic similarity developed in Chapter 3. The result is that linguistic relatedness correlates to some extent with how likely people are to have migrated between particular locations, even after controlling for spatial auto-correlation. Similarity in linguistic structures is also a predictor of how likely people are to have migrated between locations: the more similar languages are in their phonology, the more likely there is to have been migrations of people between those languages, even after controlling for geographical distance. Although the correlations are small, this suggests that there is some truth to the notion that language contact comes about partly because of movement of people between communities; and that linguistics can help us to interpret patterns in genetic data.

In summary, this thesis presents evidence for the following main findings:

1. Features such as tone show the history of language contact, through the way that they cluster geographically and change in language families.

2. Languages in Eurasia show evidence of forming clusters based on phonological features, which cross-cut language families. These clusters can to some extent be revealed by phylogenetic methods.

3. Language contact is more likely the more speakers of another language there are in that location. This may sound trivial, but does not have to be true. Features can also travel by cultural diffusion between languages: Japanese has many English loan words (Miller 1997), despite there not being mass migration of English speakers to Japan. By contrast with loan words, some structural features are good indicators of the nearby presence of speakers of some other language, as the fieldwork study in Chapter 4 shows.

4. On a continental scale, this is also true to some extent: the more similar languages are in their phonology, then the more likely there is to have been mitochondrial DNA migrations between those two languages, after controlling for geographical distance.

The remainder of this chapter elaborates further on the background to the study of language contact, and of the methods employed in this thesis. Section 2.1 defines linguistic areas and the controversies over inferring large linguistic areas. Section 2.2 describes Bayesian phylogenetic methods, and how they can be applied to the study of linguistic areas. Section 3 describes the ‘demic hypothesis’ of how linguistic areas are formed at least some of the time by migration of people across linguistic communities, rather than simply because languages borrow features from each other; and how this can be supported from genetics in particular. Section 4 suggests some avenues for future research.

2.1 Linguistic Areas

Areas where languages from different families are structurally similar have been labelled 'linguistic areas' (Trubetzkoy 1928, Thomason and Kaufman 1992:95-97). Southeast Asia is a particularly striking example, as is shown by Enfield (2005), who lists phonological properties such as tone, large vowel inventories, and constrained syllable codas, as well as syntactic properties such as Subject-Verb-Object word order, numeral classifiers, and lack of inflectional morphology. Linguistic areas have been described in the Balkans (Trubetzkoy 1928, Joseph 1983), Meso-America (Campbell et al. 1986), Northern California (Sherzer 1973), India (Emeneau 1956, Masica 1976), Australia (Dixon 2001), Africa (Heine and Nurse 2008), South America (Aikhenvald 2001, Crevels and van der Voort 2008) and various other parts of the world (see van Gijn and Muysken 2016 for a comprehensive review).

A main focus of this thesis is Eurasia, and so it is useful to list areas which have been proposed on that continent: the Balkans as mentioned, but also Europe as whole (Haspelmath 2001); the Middle East, especially between Iranian languages and Turkic (Watkins 2001, Johanson and Bulut 2006); South India (Indo-European and Dravidian contact, Emeneau 1956) and India as a whole (Masica 1976); Mainland Southeast Asia (Enfield 2005); Central Asia between Turkic and Mongolic (Schönig 2003); and Siberia (Anderson 2006).

These are all areas where languages show structural similarity across genealogical boundaries, suggesting that they have been in contact with each other. Language contact can take different forms, from borrowing of loan words to more extensive grammatical change because of processes such as substrate influence (Thomason and Kaufman 1988:50). An example of the latter is English as spoken in Singapore, which has grammatical features more akin to Chinese or Malay than they are to English such as deleting the copula ('My handwriting not clear'), using sentence final particles ('Hurry up la'), and using pitch in a level tone on words, perhaps an incipient form of lexical tone (Goh 1998). When there is widespread or elite multilingualism, neighbouring languages can influence each other through borrowing, substrate influence, or extensive bilingualism, and help form linguistic areas (Matras 2011, van Gijn and Muysken 2016).

While there are many examples of linguistic areas around the world, their existence is far less strongly substantiated than that of language families, and their history is often poorly understood. Thomason (2000) writes in conclusion to a survey of linguistic areas in which language families have often influenced each other in a multi-directional way: "As we have seen, it is often possible to establish a source language or language family for a particular areal structural feature in a Sprachbund, but very often no source can be established or, in many cases, even guessed at. For these features, the short answer to the question 'where do the features come from?', therefore, is a large question mark: we don't know." The term 'linguistic area' itself is often criticised, for reasons such as its definition being unclear or problematic, and there are various methodological pitfalls to showing their existence in particular cases (Masica 2001, Thomason 2001, Stolz 2006, Muysken 2008, Campbell 2013).

This thesis therefore aims to provide an updated definition of 'linguistic area' and to show that linguistic areas can be supported statistically. Like language families, linguistic areas show that particular events took place such as migrations of people, which can be usefully compared to findings from genetics and archeology.

A linguistic area can be defined as a group of languages whose structural properties show evidence for having a common origin, to the extent that structural properties can be assumed to travel together through time and space as a group. For example, a model can be tested which assumes that certain phonological features are transmitted together as a package, and that the phonological features of the languages in question seem to be derived from a common ancestor. If this model is well supported statistically, then one can call that group of languages a linguistic area.

This definition differs from the use of 'linguistic area' in most previous work, which requires that there be transmission of structural features primarily through language contact; this is true both in practice when the term is used (which is primarily about areas spanning multiple language families, as reviewed above), and in theory when definitions are attempted. Campbell (2006) cites definitions of linguistic areas such as Sherzer (1973:760), 'A linguistic area is defined here as an area in which several linguistic traits are shared by the languages of the area and furthermore, there is evidence (linguistic and non-linguistic) that contact between the speakers of the languages contributed to the spread and/or retention of these traits and thereby to a certain degree of linguistic uniformity within the area; or Katz (1975:16), 'One can speak of a Sprachbund if: at a given time, in a continuous geographical region, that is intersected by at least one language boundary, is encompassed by at least one isogloss.'

The definition employed in this thesis, by contrast, does not insist on language contact happening at all; in fact, a language family can also be a type of linguistic area. This is because there is not necessarily a coherent definition of linguistic area which can encompass the cases that we are trying to explain (e.g. languages with shared structural properties in Southeast Asia) but exclude other cases (e.g. languages with similar structures because they are in the same language family). Defining 'linguistic area' to be specifically about cases with contact may be like attempting to define a word such as 'ape' so that it artificially excludes humans: a scientific definition of 'ape' (a clade in the mammal family tree) inevitably includes humans, and an attempt to define the term to mean 'non-human ape' is misguided (at least from a cladistic point of view). 'Linguistic area' is similar, in that a useful definition of the term (e.g. showing that languages across multiple families cluster together in their structural properties) is inevitably also applicable to other groups of languages (e.g. related languages) which also cluster together in their structural properties. In retrospect, the phenomenon that Trubetskoy (1928) and work since then identified was not 'linguistic areas' in the sense of structural similarity across unrelated languages, but rather the fact that structural features can be transmitted in bundles, and that the transmission history of these bundles overlaps with the history of language families but can also be different from it.

There is an additional pragmatic reason for not insisting on language contact as part of the core definition of linguistic area, which is that the transmission history of linguistic structures can be elucidated, but it may be impossible to say historically whether this transmission history has overlapped with the transmission history of vocabulary ('inheritance') or not ('contact). The best examples of this include contact between closely related languages (which may be quite prevalent since closely related languages are also often neighbours), and ancient contact between language families (where it may be ultimately unknowable whether their similarity is due to contact or relatedness); and this point also applies to areas of the world such as South America, where languages are often presumably related in the last ten thousand years, but where the relatedness has not been demonstrated (Muysken et al. 2014). A group of languages can have structural properties in common either due to contact or relatedness, and even when it is unclear what exactly the cause is, it may nonetheless be useful to note this clustering of languages, for making historical inferences or for other types of statistical application.

Dryer (1989) is a good example of this latter type of study that employs a definition of 'linguistic area' which resembles that used in this thesis: 'By linguistic area I mean something rather different from what is often intended by the term. Generally, the term is used to refer to an area in which a large number of typological characteristics have diffused among languages which are genetically unrelated or at best remotely related. However, by linguistic area, I intend an area in which at least one linguistic property is shared more often than elsewhere in the world to an extent which is unlikely to be due to chance, but which is probably due either to contact or remote genetic relationships. In other words, the number of typological characteristics shared may not be enough to satisfy the normal notion of linguistic area; all I demand is that at least one property be shared to an extent that is likely to be due to areal or genetic factors. Furthermore, I remain uncommitted to what extent the existence of properties in a large area is likely to be due to diffusion as opposed to genetic relationship; hence by linguistic area I do not preclude the possibility that the underlying cause is partly or largely genetic.' The areas that he identifies (including Eurasia as a whole) are then employed to assess the frequency of particular structural properties, and for generating independent samples of languages for studying word order correlations. It is this sense of linguistic area that will be used in this thesis, a group of languages which show structural similarities that can be either due to contact or inheritance.

There are a few other questions of usage of the term in this thesis that need to be answered. For example, Campbell (2006) asks the following questions about the term: (i) How many languages does there need to be in a linguistic area? (ii) How many different language families does there need to be? (iii) How many features must be shared between these languages? (iv) How does one define the boundaries of a linguistic area? (v) Are there different kinds of linguistic area, given the various different ways that they can be formed (trade, substrate influence, repeated migration)? (vi) Is historical evidence for the formation of the linguistic area needed, or just circumstantial evidence for similarity? (vii) Does a linguistic area have to be 'a geographically delimited area' (Aikhenvald and Dixon, 2001:11), and how is that defined? (viii) How distinctive do the areal traits have to be? Can common traits be used? (ix) What is the relative weight of different traits? His paper finally poses a different question, why one should use the term 'linguistic area' at all, if it is easier to talk about more fundamental concepts such as inheritance and contact in particular languages.

These questions can all be answered by recalling the definition of linguistic area that this thesis uses: a group of languages whose structural properties can be shown statistically to be likely to have been transmitted from a common origin (to the extent that these structural properties can be assumed to be travelling in a package between languages). This definition results in the following answers to these questions:

(i) **How many languages do there need to be in a linguistic area?** There can be any number, as long as it is more than one. The question is like asking 'How many languages does there need to be in a language family?': the key point is that you can demonstrate common ancestry of structural features, which can be done between just two languages, or as many as several thousand. This leads naturally on to a related question, which is whether there can be linguistic areas within linguistic areas: to which the answer is yes, just as one can have genealogical units within genealogical units. A language can belong to Germanic, and also belong to Indo-European; similarly, a language can belong to (for example) an East Indian linguistic area, and also belong to an Indian linguistic area, and even larger areas such as Eurasia.

(ii) **How many different language families do there need to be?** As discussed above, a linguistic area as defined in this thesis can be a group of related languages, so the answer here is just 'one'. The question is in fact badly formulated, as it is possible that many language families are related in macro-families. Even if Sino-Tibetan and various other language families in Southeast Asia were shown to be related in one large family, this would not invalidate prior work showing that languages in different genealogical units have been in contact with each other, and hence describing Southeast Asia as a linguistic area.

(iii) **How many features must be shared between these languages?** As discussed in section 2.2, the key point is that structural similarity between languages can be statistically supported, not that there be a particular absolute number, which would be arbitrary. In a Bayesian analysis for example, the relevant metric is the posterior probability of a group of languages being a cluster, as opposed to other alternative clusters, which is calculated as the proportion of trees in the posterior sample which contain that particular clade (e.g. O'Reilly and Donoghue 2017). This depends on the strength of the signal in the data, but in practice, as Chapter 3 shows, the number of features used for such a demonstration may be between 100 and 200 (the size of a database such as the World Phonotactics Database (Donohue et al. 2013).

(iv) **How does one define the boundaries of a linguistic area?** In the phylogenetic method explained in section 2.2 and in Chapter 3, this is not an issue: languages are either found to be in a group or not. In a more sophisticated method using multiple trees, then languages can be in a group according to one set of structural features and not in that group according to another.

(v) **Are there different kinds of linguistic area, given the various different ways that they can be formed (trade, substrate influence, repeated migration)?** 'Linguistic area' is defined here purely in terms of statistically supported structural similarity between languages, and so the reasons for the linguistic area's existence are not relevant. It is again worth substituting ‘linguistic area’ with ‘language family’ by analogy: language families may have historically different reasons for existing (trade, migration etc.) but still be language families.

(vi) **Is historical evidence for the formation of the linguistic area needed, or just circumstantial evidence for similarity?** In this thesis, the definition used just requires ‘circumstantial evidence’, namely evidence that a certain cluster is statistically supported.

(vii) **Does a linguistic area have to be ‘a geographically delimited area’?** In fact, a linguistic area does not need to be geographically coherent, from the definition used in this thesis. The key point is the statistical demonstration of structural similarity between languages. In practice, geographical coherence should be used to validate the statistical method used, as Chapter 3 attempts to: while most of the linguistic areas discovered by the phylogenetic method are indeed in ‘geographically delimited areas’, there are a few strongly supported groupings of languages which are more scattered, which indicate that the method may be flawed in some instances and needs refinement (e.g. by using multiple trees rather than just one). This is also in line with literature such as Muysken and Smith (1995), which defines linguistic areas based on networks of people (such as the link between the Caribbean and West Africa) rather than geographical contiguity.

(viii) **How distinctive do the areal traits have to be?** The approach taken in this thesis is to use a database of structural properties, the World Phonotactics Database (Donohue et al. 2013). There can be statistical support for a particular grouping of languages being a linguistic area from just a handful of unusual features, or from many common features. The question is once again not the number or the nature of the features, but the strength of statistical support.

(ix) **What is the relative weight of different traits?** The weighting of different traits is in fact the key strength of the phylogenetic method used in this thesis, which estimates how stable and unstable different features are, at the same time as searching through possible groupings of languages (trees). This is again discussed in more detail in Chapter 3.

(x) **Why use the term ‘linguistic area’ at all, if it is contentious?** The term is a simplification, since languages influence each other through language contact in different ways; and different structural properties are likely to have different histories (such as phonological and grammatical properties), meaning that languages will

> cluster in different ways depending on the choice of features. The simplification is useful, however, when defined for a particular set of features. If a particular structural profile is found in a group of languages, then it suggests that some event has taken place that spread that structural profile, whether it is by migration, trade, or the influence of a particular language. Ultimately, this thesis is not wedded to the term 'linguistic area', but is rather attempting to model language contact, initially with a clustering method which lends itself naturally to producing groupings that can be described using that term. A more sophisticated model will identify multiple trees (perhaps even individual trees for each feature), and when this is possible this will obviously supersede the simplified method of identifying linguistic areas.

One goal of studying linguistic areas is to identify particularly large areas and to explain why they exist. For example, languages across Siberia, India and as far east as Japan share similar patterns of head-final word orders, as described in databases such as the World Atlas of Language Structures (Dryer and Haspelmath 2013). It is possible that some of these languages are ultimately related in an 'Altaic' or 'Transeurasian' family, as work such as Robbeets (2005) has argued on the basis of putative cognates in vocabulary. But because languages interact with each other, one does not have to postulate a language family to explain these similarities, if languages across Siberia have converged on similar structural properties by contact. There are suggestions of such areas, as in Bickel and Nichols (2009)'s study on case-marking which concludes that 'Linguistically speaking, the similarity of Europe to Asia supports – as far as case marking is concerned – the decision to regard Eurasia as a single macro-area.' Like macro-language-families, these macro-areas are controversial, but if supported would raise significant historical questions about how they came into existence.

The remainder of this section provides an overview of the challenges in quantifying similarity between languages, which is necessary for identifying such areas. The first challenge is deciding which properties to compare, and how to measure them objectively. There then needs to be a statistical procedure for determining which languages are most similar to each other, and assessing whether that similarity is due to chance. Some papers statistically test differences between geographical regions, which Muysken et al. (2014) calls a 'top-down' approach because these regions are defined in advance. An example is Birchall (2014), who tests for whether particular features are more common inside a geographical area within South America than outside of it. Another paper, Daumé III (2009), presents a Bayesian method distinguishing between these processes by modelling linguistic features as both inherited genealogically and subject to areal influence by languages within a particular radius. The method identifies plausible linguistic areas such as the Balkans, parts of India, Meso-America and so on, as well as a ranking of features by how likely they are to be borrowed.

The problem with these 'top-down' approaches is that they often do not take into account the fact that within a proposed area, there is also a certain amount of statistical non-independence. First, languages are expected to be similar if they are closely related. Second, languages may be in contact with each other on a more local scale: a language in Southeast Asia may borrow some properties from another language, without this meaning that we should treat the entire region as a linguistic area. The effects of contact need to span many languages, and should be detectable even once we have controlled for more banal forms of local contact.

For example, Birchall (2014)'s attempt to demonstrate the existence of linguistic areas in South America does not take into account that there may simply be more local instances of contact (or relatedness) causing the higher frequency of these variables in certain regions, without this meaning that there is a linguistic area covering the entirety of the proposed region. The problem is present in quantitative work, but also more implicitly in non-quantitative work. A variant of this problem is deciding where the boundaries of a linguistic area are. The Southeast Asian area tends to be defined as the mainland (e.g. Enfield 2005), but can also be argued on the basis of certain linguistic features to extend into Myanmar (Jenny 2015), or even into Indonesia and West Papua (Gil 2015).

This problem is particularly acute for a controversial study by Bickel and Nichols (2006) on a putative macro-area called the Pacific Rim, an area spanning New Guinea, Southeast Asia, Northeast Siberia, and the whole of the Americas. Their claim is based on the fact that many linguistic variables are more common within this area than outside of it, such as numeral classifiers, possessive classes, tone and about forty other variables. However, their sample includes many related languages, as they sample one language per 'genus' (a term denoting a group of closely related languages; see Dryer (1989)), and hence this could simply reflect the higher frequency of these variables in certain families. Even discounting that possibility, it may simply demonstrate that there are linguistic areas which are already known about and are rather smaller than their proposed area, such as a Meso-American area, or a Southeast Asian area.[1]

[1] This criticism applies less to more recent work by Bickel, which uses techniques for controlling for non-independence more akin to phylogenetics, such as the Family Bias method (Bickel 2013).

A different way of stating this problem is that linguistic areas may have a hierarchical structure. There could be linguistic areas within linguistic areas, just as there are sub-families within language families (Germanic within Indo-European, for example): languages in southern India have been argued to form a linguistic area (Emeneau 1956), while at the same time being part of a larger linguistic area in India as a whole (Masica 1976). Within a cluster such as southern India, related languages will be more structurally similar to each other than non-related languages on average, and neighbouring languages will be more similar than non-neighbouring languages. Ideally, a quantitative method for demonstrating the existence of a linguistic area should also demonstrate how areas are nested, showing their hierarchical structure.

A further challenge for traditional approaches to linguistic areas such as the 'top-down' approach, which specify geographical areas in advance and then examine their linguistic properties, is that linguistic areas could have non-obvious shapes. We will see this in Chapter 3, where the existence of some very large linguistic areas is proposed (e.g. the entirety of Eurasia except for Southeast Asia), or which have a non-obvious geographical shape, such as a crescent running from the Caucasus through the Middle East to North Africa. In principle, the best way of identifying linguistic areas is to examine structural similarity between languages, regardless of their geographical location, and see how the linguistic areas emerge (the 'bottom-up' approach). Geographical information could be added into this analysis by somehow weighting the probability of particular groupings by how close they are to each other, but ideally these groupings would emerge from linguistic facts alone, just as language families do.

With these challenges in mind, we can turn to a quantitative way of studying linguistic areas pursued in this thesis. Problems of multiple testing, how to delineate linguistic areas given that there are areas within areas, and how to quantify similarity of languages, will be dealt with by a particular method from phylogenetics, the branch of biology that studies evolutionary relationships among species.

2.2 Overview of Phylogenetic Methods

In biology, species are related to each other in a family tree, reflecting the fact that they have diverged from common ancestors; humans are closely related to chimpanzees, more distantly to gorillas, more distantly still to orangutans, and so on (e.g. Dawkins 2004). Languages also form family trees which are richly structured, such that we can be confident of the existence of a language that was the ancestor of the Germanic languages (Proto-Germanic), and further back in time, the ancestor of all Indo-European languages (Proto-Indo-European).

For this reason, phylogenetic methods have been successfully applied to reconstructing family trees of languages, by treating vocabulary as analogous to biological traits. Languages can be demonstrated to be related by using cognacy judgements (typically using short lists, such as Swadesh lists (Swadesh 1952)); the presence or absence of particular cognates can then be turned into a sequence of 0's and 1's that can be read into phylogenetic software such as BEAST 2 (Bouckaert et al. 2014) in the same way that a DNA sequence can be. A fuller explanation of this process, with reference to works that pioneered this approach such as Gray and Jordan (2000), is in Chapter 3, section 2.

Phylogenetic methods have been instrumental both in demonstrating quantitatively that families actually exist and have a well-defined structure; and also in investigating their origins, including their ages and their homelands (e.g. Bouckaert et al. 2012, Chang et al. 2015). There has even been a statistical move towards finding macro-language-families, by comparing vocabulary lists and quantifying similarity between them (Jäger 2015).

The use of phylogenetic methods to study languages raises an immediate problem however, namely that languages borrow from each other in ways that species in biology typically do not (with exceptions even there, such as inter-breeding between closely related species, or horizontal gene transfer in bacteria; Page and Charleston (1998)). The use of a model where vocabulary is transmitted in a single tree inevitably is a simplification, and potentially leads to significant errors. This should be taken seriously, and new models need to be developed to overcome this obstacle; in this vein I discuss using a mixture model at the end of Chapter 2, which allows for two or more trees. But it is also worth noting that this simplification is no worse than the implicit simplification made by historical linguists when they look for the structure of families such as Indo-European and Austronesian (defining sub-groups) and trying to exclude borrowing of vocabulary when doing so.

The methods above that have been applied to vocabulary can also be applied to structural data. Dunn et al. (2005) pioneered this approach by collecting grammatical data on languages in Island Melanesia, and applied a phylogenetic method to show that languages also seem to form family trees based on their structures, in the absence of using vocabulary data. The aim of that paper was to argue that structural data can be used to infer genealogical relationships. Some work since then has argued against this, most recently Greenhill et al. (2017), which shows that structural features evolve quickly in known language families such as Austronesian.

Rather than using phylogenetics to study language families, this thesis instead uses this method to study linguistic areas. The approach is identical to the analysis of vocabulary in languages, but simply taking linguistic structures as the data rather than vocabulary. A language is summarised as a set of features, such as how many consonants it has, what word orders it has, whether it has tone, and so on.

One does not normally think of a family tree of structures, but it is not much stranger than assuming that vocabulary forms a family tree. If instead one also imagines the language as a set of grammatical and phonological structures, the transmission of the language can continue by substrate influence, such as Cantonese into the variety of English spoken in Hong Kong (e.g. Hung 2000). Hong Kong English in this view is a descendant of Cantonese - although of course it will also have many of the structures of English. The resulting family tree of languages according to their structures will overlap with the tree that is based on core vocabulary (varieties of Chinese are structurally similar as well as related), but not entirely, because there is also horizontal transmission of structures.

Chapter 3 applies this method to a phonological data from languages in Eurasia, from the World Phonotactics Database (Donohue et al. 2013), a database with data on over 3000 languages about the presence of particular phonemes (clicks, tones, fricatives, types of liquids, etc.); and the number of consonants that are allowed in various places in the syllable, and other constraints on syllable structure. The main question pursued in this chapter is whether phonology and phonotactics can inform us about language history, in the way that vocabulary can. There are plenty of examples of unusual phonological properties being found in particular regions, such as clicks in sub-Saharan Africa (Güldemann and Stoneking 2008) or complex tonal systems in Southeast Asia (Matisoff 1973). These properties often cross-cut language families, suggesting that they may be especially informative about language contact.

A Bayesian phylogenetic analysis is used (see Chapter 3 section 2 for a full description), sampling ways of arranging languages in a family tree based on their phonological properties, according to their posterior probability. This sample of trees is then summarised in a consensus tree with the most probable clades. The resulting tree of languages should be interpreted as an analysis of how languages are similar to each other in their sound systems. Much of this similarity will be due to recent contact between languages, and so it may seem odd to summarise this similarity in a tree form. It is justified nonetheless by the observation in the preceding section that linguistic areas have a hierarchical structure, namely that there are always areas within areas.

A tree is a useful way of representing this multi-level clustering of languages. But one can go further and say that a phylogenetic analysis shows the history of sound systems, treating the phonology of a language as something that can replicate and have a history of its own. The sound systems of two languages can be related by common descent, in the same way that the vocabulary of two languages can be related. It is just that in the case of sound systems, the common descent process can also be by what we think of as language contact (relative to the vocabulary system). The idea that language is made up of parts which have independent transmission histories is elaborated on in Enfield (2014), *Natural Causes of Language*.

As with vocabulary, the tree model is ultimately a simplification, because different aspects of sounds systems might each have their own history: some properties such as syllable structure may have a different transmission history than other features such as fricatives. We can acknowledge this problem once more (as we do for vocabulary) and reiterate the need for multiple tree models, as Chapter 3 outlines in brief, and still see a single tree model as a useful first approximation of the history of sound systems.

With these points in mind, there is evidence for groupings of languages which are geographically coherent. Several groups of languages in Southeast Asia emerge, for example, which cross-cut language families and show how they have been interacting. Other interesting results include a linguistic area extending from the Caucasus through the Middle East to northwest India; and a large linguistic area covering Indo-European, Turkic and other families in Siberia. Many of these linguistic areas are nested, with intricate structure showing how pairs of languages have influenced each other, as well as possible areas within areas. Perhaps the most impressive result is that a phylogenetic analysis can find these groupings without using any geographical or genealogical information at all, showing that phonological systems can be distinctive enough to be able to show where languages are spoken and what languages they are most closely related to.

One can also add genealogical information to the phylogenetic analysis, by constraining the lower clades of the tree to be language families. When this is done, many of the above results are simplified, but the overall picture is the same, of a clade in Southeast Asia, another clade covering most families in western and northern Eurasia, and another clade comprising the Middle East and the Caucasus. This result, combining known language families with phonological similarity, turns out to be especially useful for analysing migration patterns, as the following section shows, and lends quantitative support for the notion of these three regions being distinct linguistic areas.

3. The Demic Hypothesis of the Spread of Language Structures

The analysis in Chapter 3 demonstrates that there are groupings of languages which are similar in their structures, specifically in what sounds they use, and their phonotactic rules.

Why do languages borrow sound patterns from each other? The hypothesis pursued in Chapters 4 and 5 is that linguistic structures travel between languages because of people migrating between communities. An especially important role is played by second-language learners, in this scenario, who may introduce linguistic structures from their first language into the second language; and this may be especially true of phonological structures (e.g. Hansen Edwards and Zampini 2008). Migration can cause language contact in various forms, such as substrate influence, as in the case of native Chinese (Mandarin/Cantonese/Hokkien) and Malay speakers in Singapore speaking English (e.g. Goh 1998). Another way is through superstrate influence, where a prestige language can be a source of loan words and grammatical structures, such as the influence of Chinese on Vietnamese (Haudricourt 1954).

The hypothesis that migration is the primary reason for the diffusion of linguistic structures is in contrast with an alternative, that languages borrow properties from each other simply by virtue of having some moderate bilingualism. In this view, two linguistic communities could have relatively little migration between them and still influence each other because they have some familiarity with each others' languages.

We could call these hypotheses the demic diffusion and cultural diffusion hypotheses respectively, after the debate over how agriculture spread from the Middle East to Europe, in which it was similarly proposed that agriculture was brought by a population expansion or 'demic diffusion' (Ammerman and Cavalli-Sforza 1984; Chikhi et al. 2002), and a counter-proposal that farming could have been borrowed between neighbouring communities ('cultural diffusion') without there necessarily being much migration of people (Zvelibel 2000). No doubt both processes may be involved in the spread of linguistic features, but this thesis presents evidence that the distribution of structural features may be a window into migration in particular.

Chapter 4 elaborates on the demic diffusion hypothesis with a case study on a local scale, by studying the way that a group of closely related languages have been changing by language contact, and specifically because of the migration of people speaking other languages. The languages are in the Palaungic branch of the Austroasiatic family and are spoken in Yunnan province in Southwest China. I elicited data through fieldwork on syntactic and semantic structures of these languages, such as word order, the use of numeral classifiers, and the semantics of verbs such as ‘eat’ and ‘drink’ (which vary in what type of objects they can take in different languages). The variation in these languages is presented, and discussed in comparison with data on demographics.

The main result is that these languages have been under the influence of language contact with Tai languages, and especially Tai Lü. Various linguistic features in the Palaungic languages show evidence of having been influenced by Tai, such as subject-verb order (in contrast with verb-subject order used by other Palaungic languages), the phrase ‘eye of the day’ to mean ‘sun’, the greater use of numeral classifiers, and a simplified system of ‘eat/drink’ verbs. This is confirmed for at least some features by a significant positive correlation with the ratio of Tai speakers to Palaungic speakers in nearby towns. There are also features such as the use of Tai numerals and the replacement of Wa cultural imagery (such as bull skulls) by Buddhism which also show a strong correlation with a higher proportion of Tai speakers, providing further evidence that these linguistic and cultural features travelled together demically. This shows that linguistic features such as syntactic and semantic structures can be a window into migration, as confirmed by census data and the presence of Tai numerals. It therefore provides justification for using structural features to investigate migration further back in time.

Chapter 5 provides a larger test of this hypothesis, by examining the relationship between structural features and genetic evidence for migrations, using mitochondrial DNA data. Mitochondrial DNA is transmitted from the mother, making it amenable to an analysis using Bayesian phylogeography that has previously been applied to the spread of viruses (Lemey et al. 2009) and languages (Bouckaert et al. 2012). A set of mitochondrial DNA genomes from around the world, but Eurasia and Africa in particular, was downloaded from GenBank (Benson et al. 2013) and analysed using the phylogeographic software ‘spherical geography’ in BEAST 2 (Bouckaert 2016), in order to reconstruct a family tree of individuals according to their mtDNA.

The analysis is presented as a proof of concept of applying Bayesian phylogeographic modeling of migration to human mtDNA markers, and then trying to explain the resulting patterns in terms of linguistics. MtDNA only shows part of the story of human migrations, of course, namely the migration of women. The reason why mtDNA was chosen was because phylogenetic analysis can be applied to uniparental markers and this allows a model of migration to be inferred based on reconstructing the locations of ancestral nodes in the tree; and because there was more data available for mtDNA than for Y chromosomes in Genbank, as well as more detailed location data. The analysis could also be done for Y chromosome data, which may be fruitful given the fact that languages often seem to be spread by male migrations, such as military conquests and movement of traders (Forster and Renfrew 2011). Finally, work that analyses single nucleotide polymorphisms, whole genomes or ancient DNA and compares the findings with linguistics is obviously desirable as well (e.g. Dediu 2007; Malaspinas 2016; Haak et al. 2015; Pakendorf 2014); some such studies are reviewed in Chapter 5.

The locations of people's ancestors can be reconstructed by treating longitude and latitude as continuous variables that evolve along the tree (see the summary of the phylogeographic method from Bouckaert (2016) in Chapter 5 section 3.1). To take a hypothetical example of a person who lives in England, but whose immediate relatives live in Germany, it is likely that there was a migration from Germany to England at some point. In other cases the ancestral location might be more uncertain, but can still be narrowed down to a likely range: if some relatives live in England and some in Germany, the common ancestor of these people might have lived in England or Germany, or geographically intermediate locations such as Holland.

The result of the phylogeographic analysis of mtDNA genomes is therefore a reconstruction of how people are likely to have been migrating in the past. Since mtDNA is transmitted specifically from the mother, these are all migrations that women have made, which may be systematically different from male migrations. The mtDNA genomes that survive today also represent a non-random sample of the genetic diversity of past populations, since they are from women in the past who have had many descendants on the maternal line.

Nevertheless, the migrations that women made in the past may offer some insights into the way that populations have been interacting. For example, the analysis in Chapter 5, section 3.2 shows that there are systematic differences in the directions of migration in different regions, and in particular whether people have tended to move along the lines of latitude (east-west) or longitude (north-south): in much of Eurasia, people have tended to migrate along lines of latitude, suggesting ways that large language families such as Indo-European or Turkic may have ended up covering such broad areas (recalling Diamond (1997)'s explanation for how Eurasia became especially prosperous due to the exchange of agricultural plants and other innovations across the continent). By contrast, migrations in China and the rest of Southeast Asia have tended to be from north to south, again overlapping some large genealogical units, such as as the Sinitic branch of Sino-Tibetan, or Tai-Kadai, which seems to have originated in southern China and spread to Southeast Asia (Jenks and Pittayaporn 2017); and Austronesian, which spread from Taiwan southwards to the Philippines, Indonesia and the rest of the Pacific (Gray and Jordan

2000; Gray, Drummond and Greenhill 2009). People have also tended to spread from north to south in Africa, in line with the southward expansion of Bantu languages (Grollemund et al. 2015). Concerted directions of migrations may in theory also help to create linguistic areas, since the tendency to move along the north-south axis in East Asia may reflect movements that helped to spread structures such as tone and movements of people across linguistic communities.

The impression that mtDNA migrations overlap with language families or linguistic areas can be quantified. A way of doing this to see how well the similarity of languages (be it their relatedness or their structural similarity) is predicted by the presence of mtDNA migrations between these languages. Languages are expected to be more similar to each other if there are more migrations between locations where they are spoken. This can be tested by taking a measure of linguistic similarity based on the phylogenetic analysis in Chapter 3, which builds a family tree of languages based on their phonological properties (as section 2 summarised).

The number of migrations between languages can indeed be used to predict how similar languages are, as Chapter 5 shows. However, this might be a trivial consequence of the fact that both variables - the number of migrations between languages, and how similar two languages are - are both correlated with geographical distance. One must control for geographical distance, in order to test whether mtDNA migrations themselves are useful in predicting how similar languages are. The approach taken to do this is to model the relationship between geographical distance and linguistic similarity, and then to see whether an additional term predicted from the number of mtDNA migrations improves the accuracy of the model. An additional method is to use partial Mantel tests to test the correlation between the number of migrations between languages and their similarity, controlling for geographical distance. In both approaches, the number of mtDNA migrations is shown to have a small but significant correlation with linguistic similarity, after controlling for geographical distance, in line with previous work on language and genetics such as Dediu (2007).

This is because there are many examples of pairs of languages which are more similar than expected given their geographical distance, a fact which is explained by the existence of mtDNA migrations between them. In some cases, these pairs of languages are within the same family, such as Afro-Asiatic, Uralic, Indo-European and Turkic. However, in some cases they are unrelated languages, for instance in different families in northern Eurasia, which are also more similar phonologically than expected by their distance, and this is again predicted by the presence of migrations between these languages. This is also true for languages in the Caucasus and the Middle East, where movements of people across language families also predict the phonological similarity of languages in that region. Surprisingly, this is generally not true of language families in Southeast Asia (with the exception of Austroasiatic and Austronesian), although this may be more because of the smaller number of mtDNA sequences from there in the sample.

Much of individual human migration appears to follow a pattern of random Brownian motion, as people move between communities and speak different languages, with linguistic barriers rarely being a barrier to gene flow (Pakendorf 2014), explaining why this correlation is small. However, it is also possible for people to move in a more concerted way, such as during the Arab conquests of the Levant and North Africa (Nebel et al. 2002), the expansion of the Roman empire (Séguy 2019), and of Turkic groups such as the Göktürks in the Eurasian steppe (Yanusbayev et al. 2015). Known linguistic groupings (Arabic, Romance, Turkic) reflect these historical migrations, and these provide an explanatory framework for increased probability of mtDNA migrations within these regions. Larger language families show a relationship with mtDNA migrations, pushing the evidence for the co-spread of languages and people even further back in time; and finally, unrelated languages in some areas such as Northern Eurasia or the Caucasus have phonological similarity that correlates with mtDNA migrations, showing the role that structural features can play in reconstructing human history when the languages in question are not related.

4. Conclusion

Chapter 6 summarises the main points of the thesis, namely that there is a way of finding linguistic areas which controls for local contact and relatedness, using phylogenetics (Chapter 3); and that there is evidence that these areas show the history of migration in particular, on a local scale from modern demographic data and ongoing cases of language contact in action (Chapter 4), and on a continental scale from the point of view of migrations reconstructed from mtDNA evidence (Chapter 5). Particular linguistic areas emerge from this analysis, such as Northern Eurasia, the Caucasus/Middle East, and Southeast Asia, which have been given new quantitative support here, as well as an explanation for why they exist, by reference to mtDNA migrations that they correlate with.

There are various future directions that these studies open up. Chapter 3 discusses the possibility of using a mixture model to more accurately capture the multiple histories of structural features. Phylogeography can then be used to reconstruct the way that individual sets of features have spread. Functional dependencies between features are also not investigated, which are both interesting in themselves, and necessary to discover in order to improve the phylogenetic analysis. This chapter does show that phylogenetics of language structures is worth pursuing on a global scale, as they provide support for historically meaningful clusters, both by themselves and in conjunction with genealogical information.

The fieldwork study in Chapter 4 would benefit from a larger sample of languages and features to support the findings regarding language contact. One additional contribution of the chapter is in showing the value of semantic features which are fine-grained (such as eliciting types of classifiers, types of ingestion verbs with different objects, and semantic calques), beyond what is traditionally captured by morphosyntactic databases. This shows ways of enriching databases in the future, and perhaps ultimately by crowd-sourced efforts to record and analyse language (such as the website Language Landscape[2]).

Chapter 5 reconstructs migrations through using mtDNA, but a much better method is to use whole genomes. Recent advances in computational techniques for analysing whole genomes (Kelleher et al. 2019, Wonhs et al. 2022) make it possible to reconstruct full family trees rather than simply along a maternal or paternal line; and other work on ancient DNA in particular can be used to provide time calibrations for large-scale migration events (e.g. Haak et al. 2015). Fuller models of migration based on whole genome data and especially from ancient DNA will therefore be useful to compare with linguistic similarity. The chapter provides a method for comparing the two, and shows that information from migrations does indeed improve the accuracy of a model for predicting language similarity.

The work in this thesis is my own, but has had input from colleagues which I acknowledge here. In Chapter 3, Luke Maurits and Quentin Atkinson allowed me to use a methodological innovation from a forthcoming paper, using Glottolog phylogenies as clade constraints in a phylogenetic analysis; Luke Maurits also allowed me to use a substitution model that he devised for that paper, and helped with pre-processing the data using Beastling. In Chapter 2, Seán Roberts allowed me to use data on humidity used in Everett, Blasi and Roberts (2015), and Dan Dediu allowed me to use phylogenies based linguistic distances (now in Dediu 2015). The discussion in Chapter 2, section 3 had input from Mark Donohue in the writing when we collaborated on submitting it as a paper (as yet not published), and he also contributed the map of tonal languages in Figure 1.1 in this chapter. The phylogeographic analysis in Chapter 5 benefited from discussion with Quentin Atkinson and Remco Bouckaert.

[2] www.languagelandscape.org

Data and code used for this thesis is referenced in the Supplementary Information at the end, which describes a repository on GitHub (https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials) aside from data taken from the World Phonotactics Database, which was publicly available at the time of writing but has since been taken offline. This data can be obtained by contacting the author requesting data.csv for Chapter 3 (or see the Github repository for the latest status); and the data used for the phylogenetic analysis is in any case available on the Github repository in files such as 1.xml. These are also references to data and code in each chapter in the relevant methods sections.

I hope in this thesis to provide an explanation for why linguistic areas exist, by showing that one can predict in a quantitative way how languages can be similar over large distances, if there is genetic evidence that there have been migrations between them. Conversely, this method provides a formal way to explain patterns in genetic data, such the movements of mtDNA lineages, by reference to language.

Chapter 2: A Brief History of Tone

1. Summary

This thesis begins with a case study of a structural feature, tone, and what it reveals about the history of language contact and prehistoric migration in different regions.

Tonal languages such as Mandarin make distinctions between morphemes using pitch, for example mā (high tone) 'mother', má (rising tone) 'numb', mă (low dipping tone) 'horse', and mà (falling tone) 'scold'. About a third of languages are tonal, but they are not evenly distributed, instead appearing in large clusters around the equator.

One paper proposed that tonal languages are more likely to occur in warm, humid environments, because arid environments are detrimental to the precise control of vocal cord movement that tonal languages require (Everett, Blasi and Roberts 2015). This chapter argues that this striking geographical pattern is instead due to the history of language contact in Africa and Southeast Asia, and in particular because of demographic expansions in these regions associated with agriculture. Because of the history of language contact in these regions, the correlation between tone and climate that Everett et al. identify may be an artefact. Section 2 shows in a series of simulations that language contact can create clusterings of tonal languages in regions of high humidity, where there is the greatest density of languages; and that the correlation between tone and humidity is found in these simulations, even when using the same tests and statistical controls for language family and geographical region that Everett et al. use.

Section 3 elaborates on how tone has spread in these regions by reconstructing the movement of two large language families, Niger-Congo and Sino-Tibetan, and how their tones have changed over time. It is also shown that tonal systems tend to decrease in complexity as tonal languages spread, most likely due to simplification by second-language learners who shifted to speaking tonal languages from their originally non-tonal native languages. Section 4 shows that the cline in complexity of tonal languages overlaps suggestively with the way that agriculture spread in these regions, giving a first example of the hypothesis pursued in this thesis, that structural features such as tone can show the history of migration. Section 5 provides a brief discussion of other non-linguistic factors that may affect tone, such as population size (Atkinson 2011) or genetics (Dediu and Ladd 2007). Sections 6 and section 7 summarise the main conclusions of the chapter.

2. Tone and Climate

Everett, Blasi and Roberts (2015) find a correlation between humidity and complex tone using the World Phonotactics Database (Donohue et al. 2013)[3], a correlation that holds up within different families and parts of the world. They suggest that dry air is known to affect the larynx and make precise phonation more difficult, precisely the kind of thing that really could (in principle) affect the way that people use a tonal language.

The experimental evidence that they cite shows that dry air makes precise movements of the vocal cords more difficult, raising questions that are worth exploring in the context of tonal languages. Do speakers of complex tonal languages such as Cantonese alter their use of tone in dryer conditions, for example? This may be a realistic expectation, if the effect of desiccated air on the larynx is as strong as it is reported in experiments. China is a natural testing ground for work of this kind, given that varieties of Chinese vary in their number of tones and in their climactic conditions.

In addition to the experimental evidence that humidity can affect phonation, humidity correlates with the number of tones that languages use within five different large regions (Africa, Eurasia, South America, North America, and the Pacific), and within four different language families (Sino-Tibetan, Austro-Asiatic, Afro-Asiatic, Niger-Congo). This is better statistical support than for other linguistic correlations such as word order universals, which despite having some support when sampling from different macro-areas (Dryer 1992) do not seem to hold consistently within large language families (Dunn et al. 2011), suggesting that the correlation between tone and humidity is indeed very strong.

But despite the apparent statistical and experimental support for their causal claim, the correlation may be an artefact of the history of language families and language contact. To illustrate this, I show in a series of simulations that random selection of languages followed by language contact can create a positive global correlation between tone and humidity with as much as a 83% probability, and a 47% probability of holding

[3] The results that they get are also replicated on another database of tone, the chapter 'Tone' in the World Atlas of Language Structures (Maddieson 2013). The rest of this chapter uses the World Phonotactics Database for further analysis of tone, because this database contains information on the number of tones that languages use, unlike Maddieson (2013) which only classifies languages as having no tone, simple tone (2 tones) or complex tone (more than two tones). The World Phonotactics Database was publicly available when this analysis was performed, but has since been taken off-line: the data is currently only available by contacting the author, although see https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials for the current status. In private correspondence with the authors of the database, citations for particular datapoints are available and are broadly in line with how others would code a feature such as tone (Mark Donohue p.c.). The geographic distribution of tone is also in line with Maddieson's database, suggesting that this dataset is unlikely to be unreliable with regards to the coding of tone.

within at least two different macro-areas. Language contact is additionally responsible for these correlations holding up when controlling for language relatedness, as I show that when using the random independent samples test employed by Everett et al., their result is still expected by chance as much as 60-80% of the time.

These simulations rely on certain assumptions, such as tone being randomly innovated in a small number of languages, and then spreading to other languages. The particular model used assumes a 100% probability of contagion over a fixed distance or number of languages, creating clusters of languages which are all tonal. This assumption is obviously simplistic, but is a way of simulating the way that tonal languages form clusters, either because of language families or because of language contact.

The reason why a structural feature such as tone can correlate with humidity is that languages are not randomly distributed around the world: there tend to be more languages in more humid areas such as West Africa and New Guinea, for example; humidity correlates with language density (number of languages per 30,000 km2, Pearson's r=0.31, p<0.001). This is illustrated by Figure 2.1, which shows density of languages for every language location in the World Phonotactics Database (Donohue et al. 2013). The size of the stars is proportional at each point to the number of languages within a 100 km radius of that point (typically this is between 1 and 5 in Europe, and as high as 134 in parts of New Guinea).

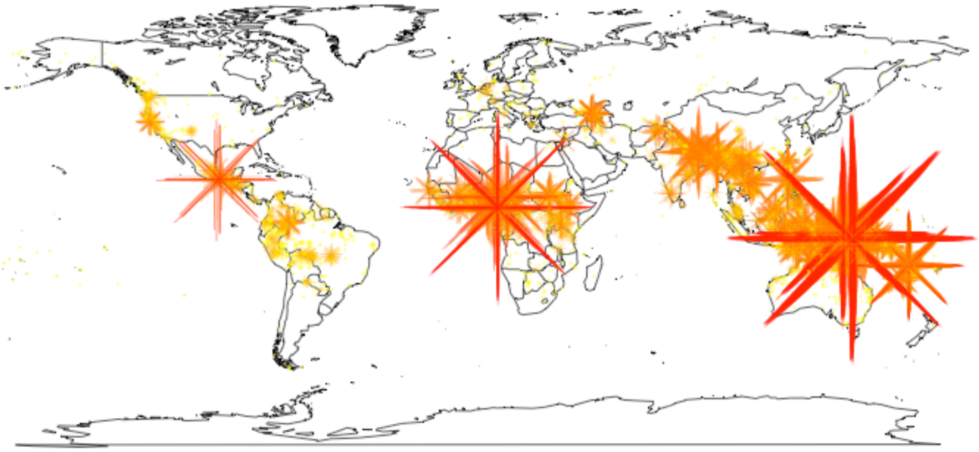


Figure 2.1: A map of language density, with each star proportional to the number of languages in a 100 km radius, using data from World Phonotactics Database (Donohue et al. 2013).

Does the distribution of languages make certain linguistic features more likely to occur in humid environments by default, simply because there tend to be more languages there? That is the first concern. The other problem is non-independence of languages, both by language families and language contact.

When testing a correlation such as between humidity and number of tones, one has to be careful to avoid counting several related languages as independent data points. A trivial example of this would be if a hundred dialects of English were counted; they would be expected to be non-tonal and all spoken in a relatively non-humid area in England.

Everett et al. take language relatedness into account when testing their correlation; this they do by sampling one language per known family when comparing the humidity values of tonal and non-tonal languages. However, they do not control for a different kind of non-independence, which arises when languages borrow from one another.

Everett et al. use a random independent samples test rather than a logistic regression. They sample one language per family in order to control for relatedness, and then compare the humidities of complex tonal languages with non-complex-tonal languages; they are predicting that in especially dry environments, complex tonal languages are unlikely to occur. Their main way of testing this is to look at the 15th percentile of humidities in the two groups and compare them, done over 5,000 samples. I replicated this with 100 samples, in which the 15th percentile of the humidities of the complex tonal languages was always higher than the 15th percentile of the humidities of the non-complex-tonal languages, in line with their prediction.

But how likely is this particular result to occur by chance, perhaps helped by the clustering of tonal languages due to language contact? I conducted simulations to model the way that this association with humidity could arise simply because of the way that tone clusters, using the following method. The R script and data are provided in Github repository referenced in the Supplementary Materials (section 1.2).

1. A small number of languages *N* can be chosen at random, out of those available in the World Phonotactics Database, and assigned complex tone. For each of those languages, the nearest *L* languages are then assigned complex tone. This is a basic way of creating clusters of languages which are tonal, where full contagion is assumed; another way not tried here is to give languages in the vicinity a particular probability of acquiring tone by contagion.

2. The random independent samples test that Everett et al. employ was done on this simulated data. The procedure in (1) and the random samples test were repeated 100 times to see how often simulated data has an association between tone and humidity (i.e. that the fifteenth percentile of the humidities of complex tonal languages is higher than that of non-complex-tonal languages).

3. Different values of the parameter *N* between 4 and 6 were tried, and the number *L* varied between 100, 200, and 300.

4. A different method of random sampling was then tried, in which languages were selected in a phylogenetically weighted way, by randomly selecting a language family, and then choosing a random path from the root of the family to one of the tips. This was tried with *N* = 6 languages chosen, and *L* varying between 100, 150 and 200.

The result was that the number of neighbouring languages *L* was the main determining factor of how likely an association with humidity was. The probability of their result occurring in these random simulations is 33%, 45%, and 49% respectively for values of *L* equal to 100, 200 or 300. The more languages that are affected by language contact from the initial 4-6 languages, the more likely there is to be a positive association. This result seems to be due to the high concentration of languages in humid areas, making random selections of geographically contiguous languages likely to be found in humid regions.

However, many of these languages in these regions are closely related (for example, there are many Niger-Congo languages in West Africa), making this test perhaps unfair. If instead of choosing languages purely at random they are chosen in a phylogenetically weighted way (see point 4 in the method above), with *L* being 100, 150 and 200, the probability of their result in the random independent samples test is 38%, 33% and 42% respectively.

In all of these cases, the probability of their result in the random independent samples test is much higher than normally accepted by conventional significance; in fact, under some models of language contact given above, it is highly likely that tone will correlate with humidity, including after controlling for language family.

The global correlation between tone and humidity should therefore be considered a possible by-product of a more basic fact: that tone seems to be rarely innovated, but spreads widely in the regions where it is found. Niger-Congo and Sino-Tibetan are two examples of families that have spread across a large region, probably causing other languages to become tonal, as explored in section 3.

The more impressive part of Everett et al.'s results is that tone is associated with humidity within several different macro-areas, namely Africa, Eurasia, the Pacific, North America and South America. They test the correlation between number of tones and humidity within areas; I employ instead the measure adopted so far of complex tone (1) versus non-complex-tone (0), in which complex tone is positively correlated with humidity in a logistic regression ($p < 0.05$) in three regions, Africa, Eurasia, and North America. How likely is it that these correlations will hold up within these regions purely by random evolution of tones and language contact?

Continuing with simulations where the number of languages N=6 (without phylogenetic weighting) and language contact spreads over L=100 languages, there is a 47% chance of holding within at least two macro-areas and a 15% chance of holding within at least three. With a phylogenetic weighting, there is a 39% chance of holding within at least two areas and an 11% chance of holding within 3.

In short, under these different models of language contact, it is quite unlikely that tone and humidity will correlate in three different macro-regions purely by chance, with that probability ranging, for example, between 11% and 15% depending on what assumptions are made in the simulations; but in general these probabilities are unexpectedly high, and moreover above the conventionally accepted significance level of 5%.

Another finding of Everett et al.'s paper is that number of tones correlates with humidity within large language families, such as Sino-Tibetan (Pearson's r=0.16, p<0.01) and Niger-Congo (Pearson's r=0.3, p<0.001). However, the major confound here is once again language contact. As section 3 shows, Sino-Tibetan languages also have fewer tones when they are near to generally non-tonal Indo-European languages, and have more tones when near highly tonal Hmong-Mien languages; Niger-Congo languages similarly lose tones near Afro-Asiatic languages, which are often non-tonal or have few tones[4]. This matters because speakers of non-tonal languages may be affecting the tonal systems of Niger-Congo and Sino-Tibetan languages; it is likely that there has been a lot of simplification of complex tonal systems over time due to second-language learning. This explanation will be expanded in section 3, which uses a phylogeographic approach to reconstruct the expansion of two tonal language families, and evidence that they have undergone simplification.

[4] The mean number of tones in Afro-Asiatic is 1.17, compared with 2.7 in Niger-Congo, using the World Phonotactics Database.

The possible effect of aridity on phonation is worth testing in naturalistic contexts, such as in conversations in different Chinese varieties, and cannot be ruled out as a factor influencing the distribution of tonal languages. However, language contact should be considered a serious confound in the way that it can create a positive global correlation between humidity and tone, including after controlling for language family, and even within specific macro-areas. More generally, a moral of these simulations is that correlational studies should take into account the geographical distribution of languages, as this provides an unexpected confound to claims about how languages culturally evolve.

3. Simplification of tonal systems due to language contact

Section 2 argued that tone is likely to have spread over long distances by language contact. This section reconstructs how this is likely to have happened, using evidence from the way that tonal complexity decreases from particular locations, and by reconstructing the spread of tonal language families.

A hypothesis explored in this section is that the complexity of tonal languages should decrease the further away they are from where tone was originally innovated in that region. This is because an expanding society speaking tonal languages would encounter populations speaking both tonal and non-tonal languages; when this happens, the number of contrasting tones typically decreases, because tones are difficult for second language learners to acquire if their native language is non- tonal (Gottfried and Suiter 1997), and they tend to simplify the tonal system. In a hypothetical scenario, Chinese imposed on a population of Europeans would be likely to lose at least some tonal distinctions due to simplification of the tone system by second- language learners. This has been argued to have happened with northern Chinese varieties, which have been learnt partially by speakers of (originally non-tonal) Turkic, Mongolic and Tungusic languages (Hashimoto 1986, LaPolla 2010).

A converse of this is that some languages acquire tones by contact, but again in a typically simplified form. Some English Creoles use tone because of influence from tonal West African substrate languages, but with typically fewer tonal contrasts than the substrate languages themselves (see examples such as Krio, Ghanaian Pidgin English and Nigerian Pidgin in Maurer et al. 2013). Vietnamese dialects in turn became tonal because of influence from Chinese (Haudricourt 1954), although they again often have one tone fewer than the six tones in southern Chinese varieties. This process of tonogenesis could in theory happen in a few different ways, such as by language shift, when speakers of a tonal language introduce tone into a language they are speaking, as in the case of

Nigerian or Hong Kong English (Maurer et al. 2013); or because speakers adopt an incoming tonal language, and subsequently introduce tone into their now less dominant local languages.

The pressure for languages to lose or gain complex tone due to language contact can be demonstrated quantitatively by examining large, tonal language families such as Niger-Congo and Sino-Tibetan. When a language moves closer to a set of non-tonal languages, for example, it may lose tones; conversely it may gain tones if it moves towards languages with more tones. This hypothesis can be tested by reconstructing the locations of past languages such as Proto-Chinese, and also by reconstructing the number of tones that those languages had.

This method can be referred to as a phylogeographic method, in that it reconstructs the likely locations of nodes in a phylogeny. Chapter 5 employs a more sophisticated Bayesian implementation in the software package BEAST 2 (Bouckaert et al. 2014; Bouckaert 2016). This approach was pioneered in the case of reconstructing language history by Bouckaert et al. (2012) in their study on the homeland of Indo-European; they controversially found support for the homeland of the family being in Anatolia, consistent with the fact that most of the primary branches of Indo-European are in that region (Anatolian, Tocharian, Armenian, Greek, and Indo-Iranian being the first groups to split off in the phylogeny).

For the purposes of this study, a simpler method of maximum likelihood reconstruction of locations and number of tones is used, using the 'ace' function in the 'ape' phylogenetic package in R (Paradis et al. 2015), and family trees of Niger-Congo and Sino-Tibetan from Glottolog (Hammarström et al. 2014).

The logic of phylogeographic reconstruction can be summarised with an example. If there is a person who lives in England, but whose immediate relatives live in Germany, then one can infer that the person in England has recently migrated from Germany. In other cases the ancestral location might be more uncertain, but can still be narrowed down to a likely range: if some relatives live in England and some in Germany, the common ancestor of these people is quite likely to have lived in England, Germany, or geographically intermediate locations such as Holland; of course it is also possible, but less likely, that they were in a location further away such as China. This assumption can be formalised by giving the latitude and longitude of where people are currently, and inferring the way that they may have changed over time assuming that they evolve by Brownian motion. Two related people who are at different longitudes (e.g. 0 in London and 13 in Berlin) may have had a common ancestor anywhere between these two longitudes, with varying probabilities, and these probabilities will change depending on where more distant relatives are located (e.g. if they are all also in Berlin, then the location of the common ancestor of these two individuals is more likely to be in or near Berlin).

The maximum likelihood approach in this section treats the phylogeny as known (the Glottolog phylogenies) and seeks to infer values for the latitudes and longitudes of the inner nodes that maximise the probability of the latitudes and longitudes of the tips of the phylogeny. This contrasts with a Bayesian approach, which can search through different phylogenies and also sample reconstructed values for the longitudes and latitudes. The Bayesian method can sample from possible migration scenarios more fully, as well as incorporating uncertainty in the phylogenies.

Another key difference is that the Bayesian approach can be geographically more accurate. Bouckaert (2016)'s implementation of the package 'spherical phylogeography' models migration as a process of diffusion on a sphere, which produces more accurate results than just modelling migration as a change in latitude and longitude, which suffers from distortions due to the Mercator projection. In addition, one can take into account the shape of landmasses by assigning zero probability of a node being over water; or take into account the fact that movement over water can be different from movement over land (e.g. people may be less likely to move over water, but move over it faster by boat). These different extensions are all discussed in Bouckaert et al. (2012). The main disadvantage of the Bayesian method is how time consuming it is to run, especially on large families such as Niger-Congo, as well as the complexity of implementing these methods. The maximum likelihood methodology in this section, while crude, is simpler and faster to implement, and still gives reasonable results when reconstructing the locations of many language families (e.g. Austronesian is reconstructed to an origin in Taiwan, and Bantu to an origin in Cameroon).

To recapitulate, the method employed here to test this hypothesis is to use maximum likelihood reconstruction of locations and number of tones, using the 'ace' function in the 'ape' phylogenetic package in R (Paradis et al. 2015), and family trees of Niger-Congo and Sino-Tibetan from Glottolog (Hammarström et al. 2014). The R script and data are provided in Github repository referenced in the Supplementary Materials (section 1.2). A full description of the method is below:

1. Phylogenies from Glottolog were used and assigned branch-lengths of length 1; and then this was replicated using phylogenies based on neighbour-joining of languages using similarity between ASJP word lists (Dediu 2015).

2. The longitude and latitude of modern languages were used to reconstruct the locations of each node in the tree, using the 'ace' function in the 'ape' phylogenetic package (Paradis et al. 2015). This gave maximum likelihood reconstructions, treating longitude and latitude as continuous variables that evolve by Brownian motion.

3. The number of tones was reconstructed at each node of the tree by using the 'ace' function. The model assumed that the number of tones is a discrete variable which can change with different rates between different numbers, using an 'All Rates Different' (ARD) substitution matrix; the probability that a language that has 2 tones can change to having 3 tones can be estimated during the reconstruction.

4. The number of tones that languages transition to at each node can then be compared with the proximity of the node's location with other language families, as a measure of how much they are being influenced by language contact. In the case of Niger-Congo, Afro-Asiatic was the family used. In the case of Sino-Tibetan, the highly tonal language family Hmong-Mien and the generally non-tonal language family Indo-European were used.

An example of a reconstruction of a particular lineage is shown in Figure 2.2, which shows the path that two Niger-Congo languages (Zulu and Swahili) have historically taken through Africa and the way that their number of tones has changed while they were moving. Proto-Niger-Congo is reconstructed in this case to West Africa, and with five tones, then moving into Cameroon, then southward and eastwards. As this language moved through Africa, the number of tones began to decrease, going through a period of having no tones and then regaining tones as it moved into southern Africa. This is probably not wholly accurate (for example, Narrow Bantu did not necessarily have zero tones, as is reconstructed here), but it represents what is most parsimoniously inferred from the modern distribution of numbers of tones in Niger-Congo.

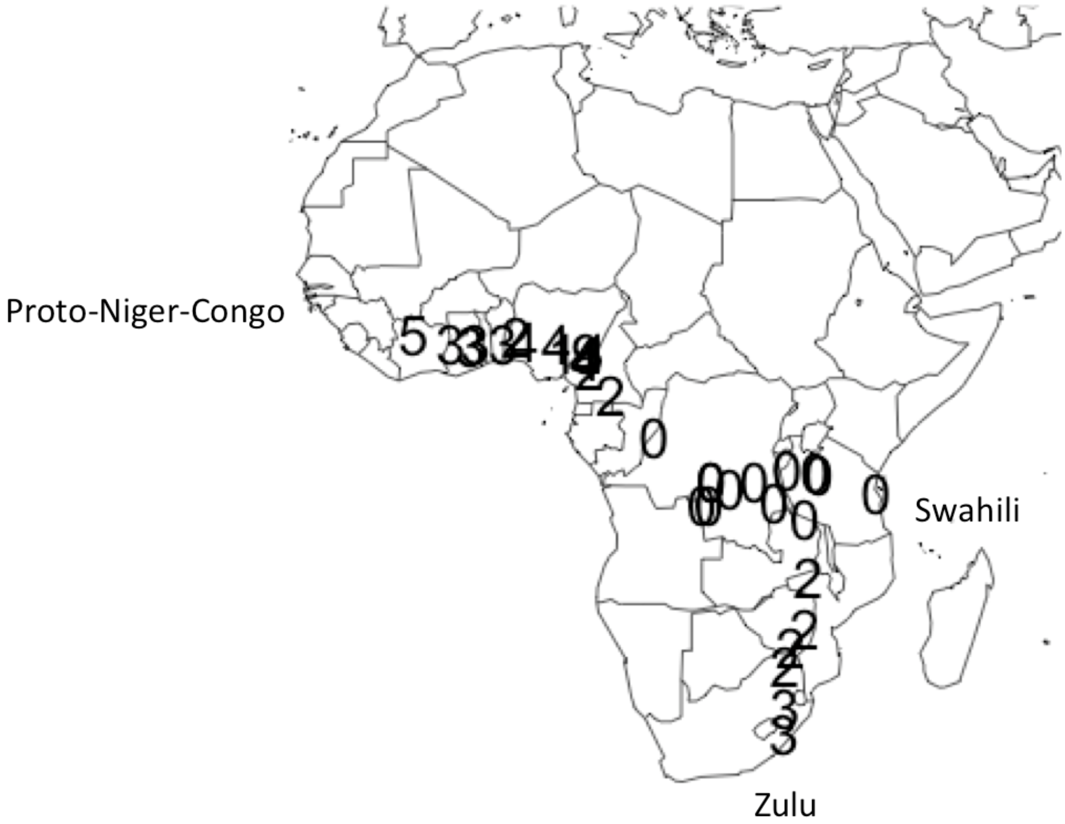


Figure 2.2: The paths taken by two Niger-Congo languages, Swahili and Zulu, in a phylogeographic analysis using R.

When taking transition events in number of tones, the number of tones that Niger-Congo languages transition to correlates negatively with proximity to the nearest Afro-Asiatic language (Pearson's r=-0.23, p<0.001). A reasonable explanation for this could be that speakers of languages that are non-tonal or with relatively few tones are simplifying the tonal systems of Niger-Congo languages. This is expected if there are interactions between these language families (whether by language shift or other kinds of bilingualism: see Güldemann 2010 for proposals of language contact zones, such as the Macro-Sudan belt that covers both Niger-Congo and Chadic languages).

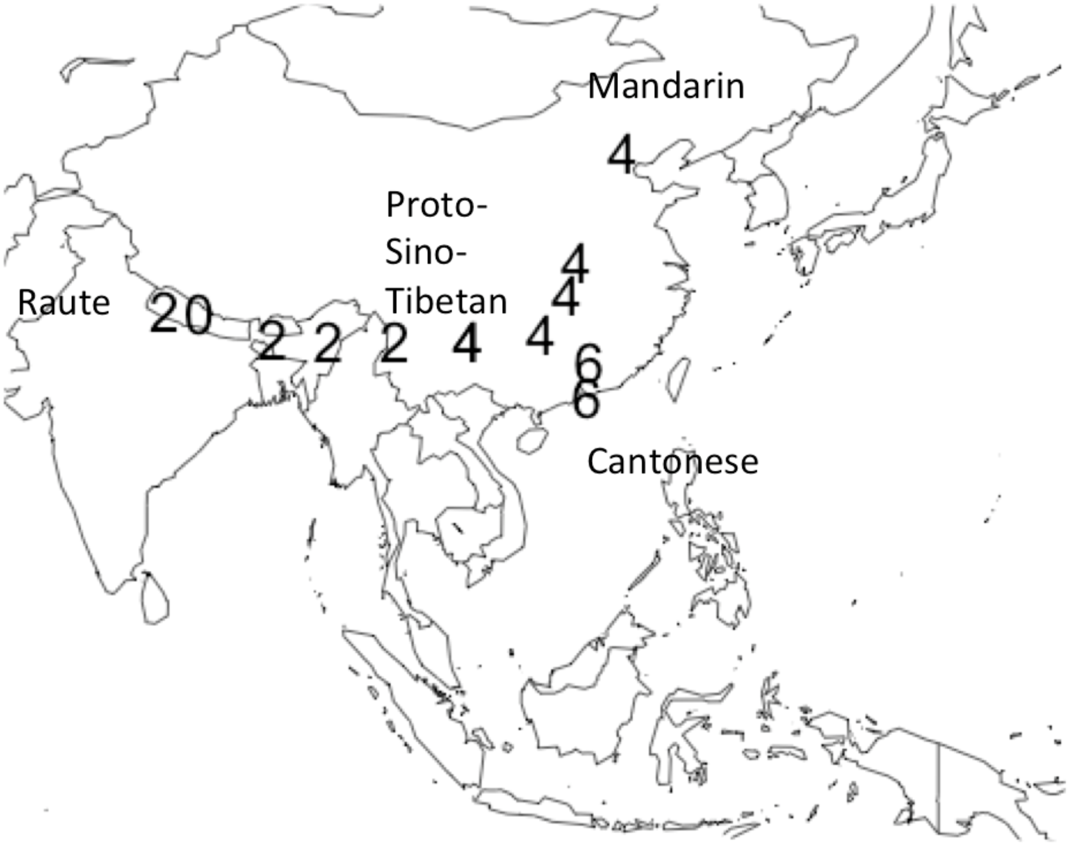


Figure 2.3: The paths taken by three Sino-Tibetan languages, Raute, Mandarin and Cantonese, in a phylogeographic analysis using R.

Furthermore, in Sino-Tibetan, the number of tones that languages transition to correlates positively with proximity to the nearest Hmong-Mien language (Pearson's r=0.33, p<0.001) and negatively with proximity to the nearest Indo-European language (Pearson's r=-0.47, p<0.001). Hmong-Mien languages have the highest number of tones in Asia, suggesting that proximity to these languages caused some Sino-Tibetan languages to gain tones due to contact. Similarly, Indo-European languages generally have no tones or very few, suggesting that Sino-Tibetan languages have been losing tones near India due to simplification by contact with Indo-European. The path that three Sino-Tibetan languages (Raute, Mandarin and Cantonese) took from a reconstructed location for Proto-Sino-Tibetan in western China is shown in Figure 2.3.

4. Clines in tonal complexity and the overlap with agricultural expansions

To recapitulate the finding of the previous section, languages tend to lose tones when they are near non-tonal languages. They can also gain tones when they are near tonally complex languages, as in the case of Sino-Tibetan languages having an especially high number of tones near Tai-Kadai and Hmong-Mien languages; however, the number of tones that these Sino-Tibetan languages have is typically less than the number of tones that the donor languages have.

If tone spread through Asia by language contact, then in general we expect tonal complexity to be highest near where it originated, and to decline in languages which are further away from that origin. This pattern is indeed found in both Africa and Asia, and also perhaps other tonal areas of the world, where there is a decreasing number of tones from a particular point, an epicentre that tone seems to have radiated out from.

In Africa, languages with the most tones are found in Nigeria and Cameroon. Successively fewer tones are found in languages further away, first in West Africa, and then in a large area reaching down into South Africa. The number of tonemes that languages use, as reported in the World Phonotactics Database (Donohue et al. 2013), decreases gradually, as is visualized in the heat map in Figure 2.4. To produce this map, a curve was drawn around languages with the most tones (in this case twelve) and coloured, and then around languages which have at least the second most tones (11), and so on down to the lowest number of tones (2). Each curve is coloured with heat colours corresponding to numbers of tones.

Why does tone decrease in complexity from Cameroon in particular? It is striking that the distribution of tonal complexity matches the archeological and linguistic evidence for the expansion of Niger-Congo languages, and particularly the Bantu expansion that began in Cameroon and spread eastwards to Lake Victoria and then south. This is shown in more detail in Figure 2.5 from a phylogeographic analysis of the Bantu languages by Grollemund et al. (2015). These languages are likely to have spread demically (de Filippo et al. 2012) by pottery-making horticulturalists and farmers (Russell et al. 2014). If there is one demographic event that has the power to explain why tone has spread not just in Niger-Congo languages but also in surrounding language families such as Afro-Asiatic, this is likely to be it.

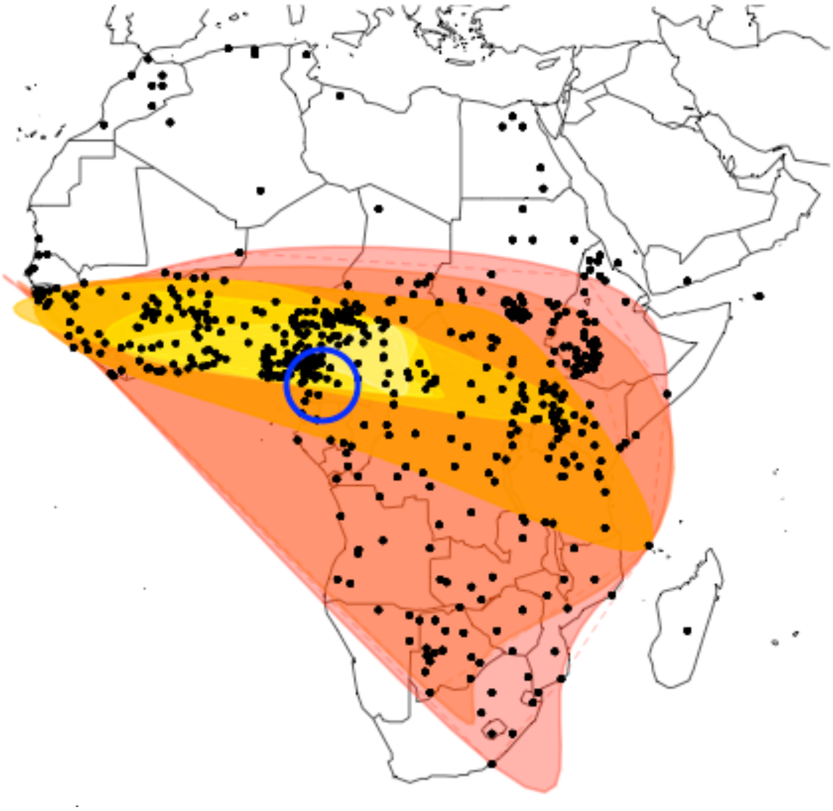


Figure 2.4: A heat map of number of tones in different languages in Africa, from 2 (light red) to 12 (yellow, in Cameroon). The blue circle shows the putative origin of the Bantu expansion in Cameroon (Grollemund et al. 2015; see Figure 2.5)).

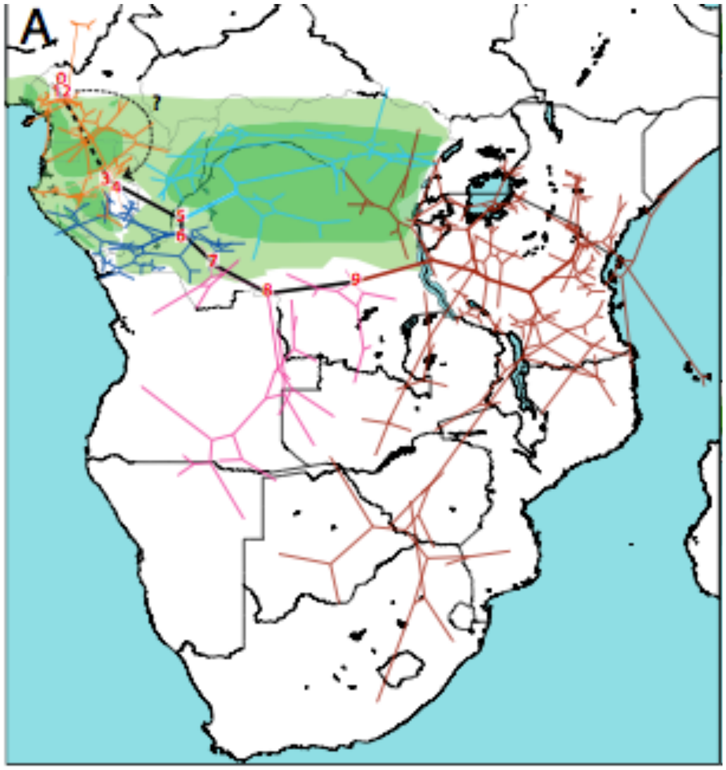


Figure 2.5: A map of the Bantu expansion from Cameroon according to linguistic evidence from Grollemund et al. (2015), with different coloured lines representing the paths of particular lineages, and a green area showing the rainforest at 2500 years BP (dark green) and 5000 years BP (light green).

A similar pattern is found in Asia, where again the number of tones that languages use decreases, forming concentric circles that radiate out of southern China, as shown in Figure 2.6. The most complex tone systems are found in the languages of the Hmong-Mien family around Guangxi. Successively fewer tones are found in Tai-Kadai languages of the area, then in southern varieties of Chinese, then more distantly related Sino-Tibetan languages, and finally in Austroasiatic languages in Southeast Asia, languages of the Himalayas, Japanese and some Korean dialects, and some Malayo-Polynesian languages.

This shows an overlap again with the spread of agriculture in this region, in this case the way that rice-farming spread in Asia, on the basis of archeological evidence from a recent paper by Silva et al. (2015), as shown in Figure 2.7. Rice domestication is first attested in southern China 9000 years BP spreading out of southern China 4-5000 years into Southeast Asia and India (Fuller et al. 2009) building on an earlier tradition of cereal agriculture (Lu et al. 2009).

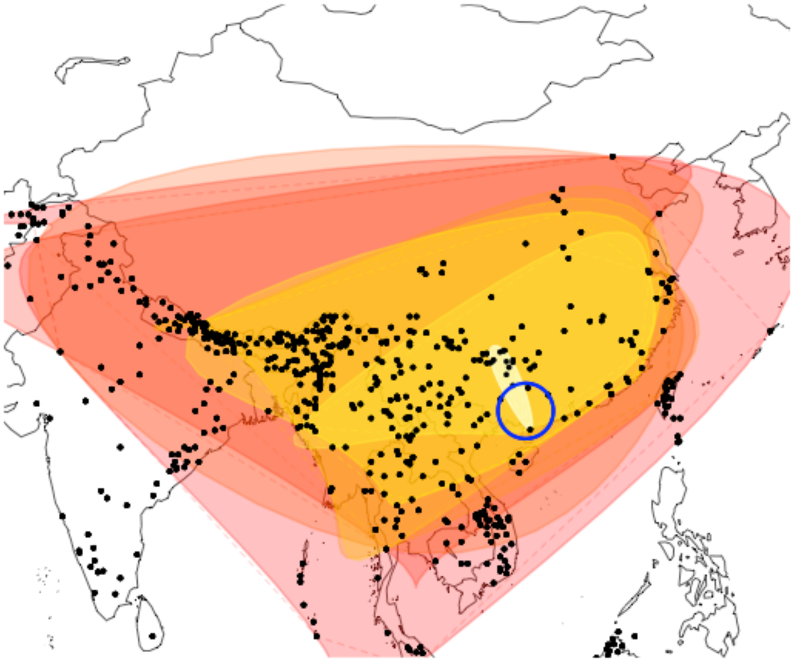


Figure 2.6: A heat map of number of tones in Asia (white/yellow = most number of tones (9), red = fewest (2), with the blue circle marking the origin of domesticated rice according to genetic and archeological evidence (Huang et al. 2013, Molina et al. 2011, Silva et al. 2015; see Figure 2.7).

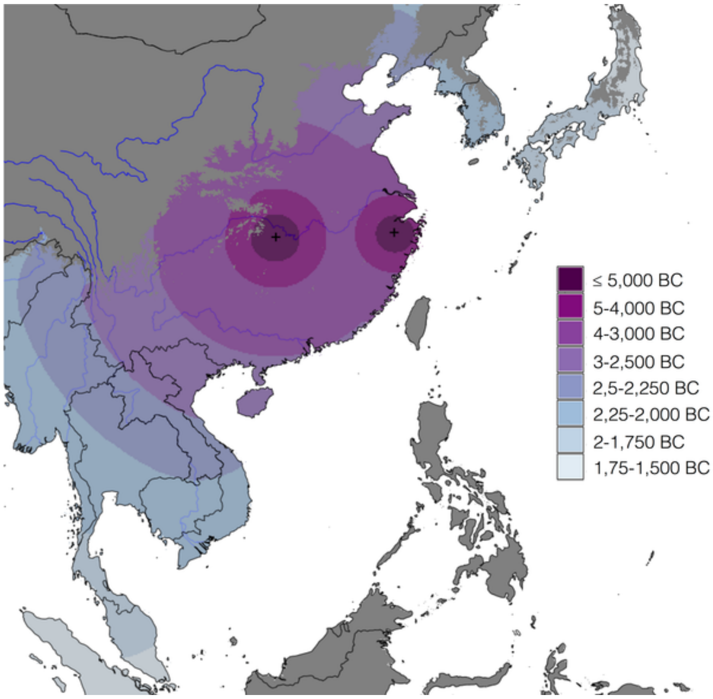


Figure 2.7: Silva et al. (2015)'s map of the spread of Japonica rice farming according to archeological evidence.

In both regions, there is a striking similarity between the number of tones that languages use, and distance from a point of agricultural expansion. This association can be demonstrated more formally by using a linear regression, and linear mixed effects models, using the following method:

1. Languages were chosen from the World Phonotactics Database that are found in Africa or Asia. These two regions were tested because they both have tonal languages, which are assumed to have developed tone independently. The region 'Africa' is defined by the macro-region label 'Africa' in the World Phonotactics Database, while 'Asia' is defined by the two Autotyp labels 'South Asia' and 'Southeast Asia' in the same database.

2. In these two regions, a linear regression was used to predict the number of tones using distance from a point where agriculture arose in Africa, chosen as the coordinate (11.52, 3.87) based on a map of the Bantu expansion in Bellwood (2013). In Asia, the dependent variable was distance from sites in southern China where rice was domesticated according to genetic evidence (Huang et al. 2013) or where the oldest rice remains are (Silva et al. 2015), represented by the coordinates (110, 22.5) and (120.7, 28) respectively. The regression was also repeated using just languages with tone.

3. A linear mixed effects model was then used to control for the non-independence of languages within the same family, using top level family (e.g. Sino-Tibetan) as a random effect. This was done using the packages 'lme4' (Bates et al. 2014) in R. The R script and data are provided in Github repository referenced in the Supplementary Materials (section 1.2).

The number of tones declines with distance from the Pearl River valley area, where genetic evidence suggests that domesticated rice originated (Huang et al. 2013, Molina et al. 2011) (n=674, Pearson's r=0.5, p<0.001), and the correlation is as strong from the location of the oldest rice-farming sites further north in Zhejiang (Bellwood 2013). The number of tones correlates inversely with distance from this region within major rice-farming families such as Sino-Tibetan (Pearson's r=0.41, p<0.001) and Austro-Asiatic (r=0.35, p<0.001).

In both regions, there is a correlation between the number of tones that languages use and distance from a point of agricultural expansion. These correlations are not simply due to the non-independence of tonal languages within the same families. In order to control for phylogenetic relatedness, the correlations reported above were retested using mixed effects models, with language family as a random effect. These models were compared with a model with these as random effects which also include distance from the location of the proposed demographic expansion (such as Cameroon), and in Africa and Asia the models that included distance from the origin were found to be significantly better (p<0.001). In addition, the correlations reported hold up within the large families in these regions, Niger-Congo, Afro-Asiatic, Sino-Tibetan and Austro-Asiatic.

The claim is of course not that agriculture predisposes languages to develop tone, but that the spread of tone in both regions could be explained by demographic expansions. It is unknown exactly how old the main tonal language families Niger-Congo, Hmong-Mien, and Sino-Tibetan are but their ages have been estimated using automated dating techniques of vocabulary at 6227 years old, 4243 years old, and 5261 years old respectively (Holman et al. 2011), congruent with or predating the timing of agricultural spreads in these regions.

This gives a possible explanation for why tone spread so far in each of these regions across multiple language families: it is not that tone is especially fashionable, for instance, but that it was brought by the expansion of these language families, and by people migrating. on the heel of population expansions associated with the advent of agriculture. In the process of migration, tonal languages that they brought will have caused other non-tonal languages to gain tones (and at the same time, causing the expanding tonal languages themselves to lose tones).

5. Other factors influencing tone

The possible role of the environment in influencing tone was discussed in section 2. Other papers have also speculated on non-linguistic factors that may have influenced the distribution of tonal languages. For example, population size has a systematic relationship with various aspects of language structure, such as morphological complexity (Lupyan and Dale 2010). It is also possible that population size has a relationship with the number of phonemes that a language has, as Atkinson (2011) argues; he further claims that the number of phonemes that languages have decreases with distance from Africa, as if there is a founder effect of phoneme diversity as people migrated, and that this holds separately for consonants, vowels, and numbers of tones. It is still unclear what similar mechanism would do this for phoneme inventories, although he cited independent evidence that population size is also a predictor of phoneme inventory size, suggesting that some process of simplification of the phoneme inventory may come about if a part of linguistic community breaks away. This is, of course, part of the argument in this chapter with regard to clines in tonal complexity, that the number of tones used can decrease as languages spread, although the mechanism proposed is simplification due to language contact, rather than a founder effect.

Another paper by Dediu and Ladd (2007) argues that genetics may also influence the distribution of tonal languages. Speakers with ancestral alleles of the genes *ASPM* and *Microcephalin* are found in regions with tonal languages. The correlation between these variants and tone is strong (stronger than 97.3% of all gene-language correlations that they tested), and it remains significant in a partial Mantel test controlling for language relatedness and distance between languages. Since these two genes are expressed in the brain, Dediu and Ladd argue that the reason for this may be that these two genes have an effect on speakers' processing of tone, causing some languages to be less likely to develop tone than others.

The two genes *ASPM* and *Microcephalin* had no particular reason to be tested, as their effect on cognition was unclear at the time (Dediu and Ladd 2007, Mekel-Bobrov et al. 2007). However, since that paper, there have been two experiments which seem to vindicate this hypothesis.

The first study by Wong, Chandrasekaran and Zheng (2011) that found that people with the derived allele of ASPM were better than those with the ancestral allele at a tone perception task. There are two reasons, however, why Wong et al.'s study cannot be taken as strong support for Dediu and Ladd's claim. The first is that the result of their experiment went in the opposite direction to that predicted by Dediu and Ladd; the ancestral allele is the one that is found in regions with tonal languages, not the derived allele. The second reason is that the sample in this experiment only contained thirty-two participants, making it quite possible that their result is a false positive.

The second study, by Wong et al. (2020), is an experiment on native Cantonese speakers which seems to support Dediu and Ladd's hypothesis more directly. In this study, they showed that speakers with the derived allele of *ASPM* perform worse on a tone perception task than speakers with the ancestral allele, which is exactly what Dediu and Ladd's original paper predicts.

The experimental evidence in Wong et al. (2020) now makes the existence of a causal relationship between *ASPM* and tone quite likely, but not certain. It is possible, for example, that there were some speakers in the experiment who were not entirely native in their phonology, or whose knowledge of Cantonese tones had interference from Mandarin, which could happen if some mainland Chinese people were in the sample. Since the frequency of the derived allele of *ASPM* is higher in Northern and Western China, it is possible that the results were affected by a few participants who were worse at tone perception who also happened to have the derived allele of *ASPM*. Another, less likely, possibility is that *ASPM* is not directly related to tone, but rather to intelligence generally, as measured for instance by IQ. This is suggested by the fact that IQ predicts performance on the tone perception task in Wong et al.'s study (r=0.224, p<0.001; however, past studies have investigated whether *ASPM* is related to IQ, and have not found any positive association (Mekel-Bobrov et al. 2007).

Another question is why the Cantonese speakers in the experiment were apparently native speakers, and presumably competent at producing tones in Cantonese, while simultaneously in some cases performing badly (almost at chance level) on a tone perception task. However, this has also been found in studies on people who are tone-deaf musically (e.g. inability to detect when someone sings out of tune, or wrong notes in a familiar melody): Yun et al. (2010) find that native Mandarin speakers with congenital tone-deafness (amusia) tend to do worse on discriminating lexical tones, despite their tone production being normal. Amusia is also known to be hereditary: for instance, in Peretz and Vuvan (2017), amusia is found in 1.5% of the population, but is found in 46% of first-degree relatives of someone with amusia. This makes it quite plausible that there are indeed single genes that can have a large effect on ability to perceive tones accurately, while not affecting tone production, or indeed cognition more generally (Peretz and Vuvan 2017).

This last point, that *ASPM* could be affecting linguistic tone perception via amusia, raises two further questions. One is whether the Cantonese participants of Wong et al. (2020)'s experiment with the derived allele of *ASPM* were also more likely to be tone-deaf than the participants with the ancestral allele, which would strengthen the case for *ASPM* affecting linguistic tone. Wong et al. test for this, asking participants to judge whether two melodies are identical or slightly different (by one note), and do not find an effect of *ASPM* on this ability. The second question is whether, if *ASPM* is linked with amusia, this condition will turn out to be more common in Northern and Western Eurasia than elsewhere, given the higher frequencies of the derived allele in this region. Given the negative result in Wong et al.'s experiment, it is reasonable to assume that this is not the case.

Another question is why Dediu and Ladd's hypothesis may turn out to be correct, despite a seemingly valid criticism of it from a statistical standpoint. I have previously summarised my criticism of this line of reasoning in Collins (2017) as follows: 'Picking two genes to focus on because they occur in the same regions as tone, itself a very spatially clustered linguistic feature, automatically makes this correlation better than most randomly selected correlations between genes and linguistic features. This makes the fact that this correlation is in the top 97.3% of gene-feature correlations unimpressive, as a correlation which is visually striking could well be stronger than 97.3% of randomly chosen gene-feature correlations.' They discovered the visual resemblance between the distribution of the ancestral allele of *ASPM* (and *Microcephalin*) and tonal languages, and this then led them to test the correlation: as Dediu said in his doctoral thesis (Dediu 2007:192), 'This hypothesis was based on apparently congruent geographical patterns and on a putative decomposition of the various linguistic strategies into sequential and parallel components (D. R. Ladd, pc), supported by data from linguistics and neurosciences. Unfortunately, for the moment, we do not have a coherent theory concerning the parallel and sequential mechanisms in language, and, subsequently, there is no clear mechanism linking ASPM-D and MCPH-D to tone.' They are therefore testing the correlation not because there was a clear causal theory (at the time at least), but mostly because it was visually striking.

Another way of saying this is that one should not *test* a hypothesis on the same data that *suggests* a hypothesis. As Casella and Berger (2001: 7121) put it, 'For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and hence … is likely to be accepted unless the bias is corrected for'. But why would this line of criticism be wrong, and Dediu and Ladd's approach turn out to be valid after all? One reason may be that a gene that affects ability to perceive tone may have quite a strong effect; if a community with many effectively tone-deaf speakers is unlikely to develop tone languages, or adopt them by language contact, then the existence of such a gene may well be visible from the geographical distribution of tonal languages - namely, that there may be an unusual cluster of *non*-tonal languages, as there is in Northern and Western Eurasia. This fact might justify using this geographical distribution to discover candidate genes that can be causally linked with tone, after appropriate controls for language family and geography. The criticism raised by Collins (2017) is therefore overstated, in so much as it underestimated the probability of hypotheses such as Dediu and Ladd (2007)'s turning out to be true (if the experiment by Wong et al. (2020) is valid and is replicated).

Finally, how does Dediu and Ladd's hypothesis affect the other historical hypotheses explored in this chapter? In some ways, it does not affect the historical account of how tone may have spread out of particular regions such as Southern China; if anything, it makes these historical accounts more likely to be accurate, if by contrast populations in Northern China have a higher frequency of the derived allele of *ASPM*, and hence are presumably less likely to innovate tone. In other ways, it may affect the validity of using tone as a feature for investigating the history of languages, since it does not evolve neutrally but is more likely to be innovated in certain regions. It also further complicates the hypothesis about humidity affecting tone, since Everett, Blasi and Roberts do not control for the frequency of alleles of *ASPM* in their analysis. Further work is needed to investigate how these different factors may interact to produce the modern distribution of tonal languages, and it would be especially useful to investigate regions such as Northern China to disentangle these effects on Chinese varieties spoken there.

6. The stability and borrowability of tone

Much of this chapter argues that the striking clustering of tone in particular regions reflects its historical stability, but also the ability for tone to be borrowed across neighbouring languages. This may sound like a paradox, as the stability of a feature within a language family and its ability to be borrowed are usually thought of as opposites; this is especially true in historical linguistics when studying vocabulary, where there is often a distinction made between basic vocabulary (word lists such as Swadesh (1952)), and the rest of the lexicon which tends to be easier to borrow, and is therefore less reliable for reconstruction.

Models of language history have often conflated stability and borrowability in a single metric which they refer to as stability. For instance, Dediu and Levinson (2012) calculates the stability of different features in the World Atlas of Language Structures; but this is defined (broadly, and measured with different methods) as the probability for a particular value of a feature to remain the same in a language family over the time. This definition ignores the difference between features changing in a family simply because they are prone to change, and changing in a family because values for that feature are borrowed from other languages.

A full model of language history would need two different metrics for a particular feature, its stability and borrowability. A feature can be both highly stable and highly borrowable, and in fact this is what is largely assumed in this chapter for tone. Whole language families such as Sino-Tibetan and Niger-Congo can be tonal (and conversely, most language families in mainland Eurasia have no trace of tone), suggesting that the basic property of having tone or not having tone is very stable. This is indeed found by Dediu and Levinson (2012), where tone is ranked fifteenth out of 68 features in stability. The number of tones that a language has is, of course, a less stable property, since languages transition between different numbers of tones within families such as Niger-Congo.

At the same time, tone is very borrowable. Section 3 demonstrates this by showing that changes in number of tones is predicted by what unrelated languages there are nearby (e.g. Sino-Tibetan languages tend to lose tone when they move towards Indo-European languages). This can be seen more generally by the fact that tonal language families cluster together, such as Hmong-Mien, Sino-Tibetan, and Tai-Kadai in Southeast Asia. Assuming that they have not inherited tone from a common ancestor, the clustering of unrelated tonal families in Southeast Asia, as in other parts of the world, suggests that tone has spread between these families. A quantification of stability and borrowability in comparison with other features is beyond the scope of this chapter, but it does seem that there is evidence for tone being both stable and borrowable, the conditions needed to explain the clustering pattern found, particularly as argued in the simulations in section 2.

7. Conclusion

This chapter began with the observation that tonal languages are clustered in particular regions, and then attempted to rebut an explanation by Everett et al. (2015) that tonal languages have adapted to humid environments. Tone is likely to have spread long distances by language contact, which section 2 showed in simulations is the set of conditions especially likely to create a spurious correlation between tone and humidity, by virtue of the fact that humid regions are where the greatest number of languages are found.

Tone has spread across several language families in Asia, such as Sino-Tibetan, Tai-Kadai, Hmong-Mien, Austronesian, Japanese, and Austroasiatic. In Africa, tone has spread across numerous families such as Nilo-Saharan, Khoe, Mande, Afro-Asiatic, Tuu, as well as the main large tonal family Niger-Congo. Similarly in Mexico, tone is found in Otomanguean, Mayan, Uto-Aztecan, Chibchan, Totozoquean-Chitimacha, Cuitlatec, and several others.

Why is tone so contagious? This chapter explored the hypothesis that tone was brought across these regions by migration of people, as the cline in complexity of tonal systems in each region overlaps with the way that agriculture is known to have spread, and also aligns with the history of the main large tonal families. There may, however, be other explanations for the pattern of clines in tonal complexity, such as the encroachment of non-tonal languages on areas of tonal languages; and there are other factors influencing the history of tone, such as language-internal reasons to do with syllable complexity, the use of particular consonants, word length, and other aspects of phonation; and of course, it is possible that the environment or other non-linguistic factors could affect the history of tone as well, as other papers have suggested (Dediu and Ladd 2007 on the possible role of genes; and Atkinson (2011) on the effect of population size).

The hypothesis outlined here is instead an example of the more general idea pursued in the remainder of this thesis, that linguistic structures can be used to show the history of languages interacting with each other, and that this in turn is likely to show the way that people have been migrating. The following chapter investigates this hypothesis with a larger set of phonological features, using the World Phonotactics Database (from which data on tone was used) to find linguistic areas using phylogenetics; while Chapters 4 and 5 continue to explore the relationship between these linguistic areas and data on migration.

Chapter 3: The Family Tree of Sound Systems

1. Summary

The history of languages can be illuminated not just by comparison of basic vocabulary, but also by comparison of syntactic and phonological properties. This is hinted at in the World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2013), which shows that some structural properties tend to predominate in certain areas of the world, such as tone and numeral classifiers in Mainland Southeast Asia (e.g. Enfield 2005).

A particularly useful source of data for investigating this hypothesis is the World Phonotactics Database, a database of phonological and phonotactic structures (Donohue et al. 2013), which has data on over 3700 languages for about 160 features. The justification for using phonological and phonotactic properties of languages is that they may be useful for reconstructing the history of languages, given that whole language families such as Sino-Tibetan can be tonal (and conversely whole language families non-tonal), or have quite distinctive properties such as clicks. These properties may also reveal the history of language contact, as tone and clicks also cross-cut language families in certain regions (Güldemann and Stoneking 2008; Matisoff 1999).

This chapter uses phylogenetics to investigate how similar Eurasian languages are to each other in their phonology. Phylogenetic methods have been used to study the history of language families and, more speculatively, to discover clusters of languages using their grammatical structures, such as by Dunn et al. (2005) on on a set of Melanesian and Papuan languages, who found that closely related Austronesian languages in Melanesia cluster hierarchically in a tree similar to the tree found by using vocabulary. They also found using the same method that unrelated languages in Papua New Guinea shared structural properties, suggesting either ancient relatedness or contact between these languages. Other papers applied these methods to the WALS data, finding some evidence for large-scale clustering that might again indicate macro-families or large-scale contact (Dediu and Levinson 2012, Dediu and Cysouw 2013). There have been other Bayesian analyses of databases of linguistic structures with the aim of elucidating language contact in particular. Reesink, Singer and Dunn (2009) apply a method for inferring founding populations to data on grammatical structures gathered for languages in the Pacific and Australia, using the software STRUCTURE (Pritchard, Stephens and Donnelly 2000). This model reconstructs a set of ancestral languages which have influenced different languages to different degrees. Daumé III (2009) uses a Bayesian model to discover linguistic areas based on geographical proximity. Other quantitative modelling of language contact includes Chang and Michael (2014) on phonological inventories in South America.

When phylogenetics is applied to structural data, as in this chapter, the clusters discovered overlap to some extent with known language families, but also cross-cut them, showing how languages have been interacting through language contact. This means that phylogenetics can be used to discover linguistic areas, in the sense of groups of languages which share structural properties due to contact, an example being languages in Southeast Asia (Enfield 2005) which share properties such as numeral classifiers, tone, and so on. Dahl (2008) analysed languages in the World Atlas of Language Structures, measuring typological distance between languages, and found that languages of Southeast Asia were often about as typologically similar to each other as closely related languages in Europe, despite being unrelated (e.g. the distance between Polish and Russian is 12.4, while it is 11.4 between Thai and Vietnamese). A disadvantage of the method that Dahl used is that it treats all features equally, whereas phylogenetics allows the stability of features to be estimated at the same time as discovering how languages cluster.

It may seem paradoxical to use phylogenetics to study language contact, given that these methods are usually used when features are assumed to be inherited from only one parent. This chapter argues that, at a minimum, phylogenetics is a useful method for discovering clusters of languages which are similar in their phonological and phonotactic properties. These clusters are hierarchical, showing how individual neighbouring languages have been interacting, as well as how languages cluster into larger areas (such as India) and then into more weakly supported macro-areas (western and northern Eurasia). Phylogenetics has some advantages over other clustering methods, because it is character-based rather than using a single metric of similarity as neighbour-joining does. In particular, the analysis employed here allows features to have different rates of change, meaning that features which are slower changing will contribute more information to the deep divisions in the tree than the noisier, fast-changing features will.

Phylogenetics is more than a glorified clustering method, however, as it is also a model of history: a tree represents the way that a package of linguistic features has been transmitted through time. Ultimately, the best model of the history of language would have multiple trees, ideally for each individual feature or bundle of co-transmitted features. This is usually unfeasible, since a single feature does not contain enough information to distinguish between convergence and relatedness (two languages that have tone could have shared history or have innovated it independently). A method which comes closer to this is a phylogenetic mixture model that allows two or more trees, as implemented in Bayes Phylogenies (Pagel and Meade 2004) and discussed further in section 5.1.

With these points in mind, this chapter presents evidence for groupings of languages based on phonology which are geographically coherent. Several groups of languages in Southeast Asia emerge, for example, which cross-cut language families and show how they have been interacting. Other interesting results include a linguistic area extending from the Caucasus through the Middle East to northwest India; and a large linguistic area covering Indo-European, Turkic, and other families in Siberia. Many of these linguistic areas are nested, with intricate structure showing how pairs of languages have influenced each other, as well as possible areas within areas. Perhaps the most impressive result is that a phylogenetic analysis can find these groupings without using any geographical or genealogical information at all, showing that phonological systems can be distinctive enough to be able to show where languages are spoken and what languages they are most closely related to.

2. Methods: Bayesian Analysis of the World Phonotactics Database

As described above, the World Phonotactics Database has information for over 3700 languages around the world, coded for 161 features to do with phonology and phonotactics (as well as some additional information on population sizes and word order). The World Phonotactics Database was publicly available when the data was downloaded in 2014, but has since been taken off-line: the raw data is therefore currently only available by contacting the author or Mark Donohue, although see https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials for the current status.

The questions range from asking how many consonants can maximally appear in the onset or coda of syllables, what kind of consonants can appear in certain positions (e.g. ‘Is the second consonant preferentially a glide?’), question about the number of consonants in the language, and about more specific types of consonants (‘Does the language have rhotics?’, ‘How many coronal nasal places are there?’). The questions therefore are divided between binary questions which have either ‘yes’ or ‘no’ as an answer, and questions which have a number such as the number of consonants in a language or the number of consonants in a coda.

Each language has data on features such as the number of consonants, number of plosives and so on, as well as phonotactic features such as the number of consonants maximally allowed in the onset and coda. There are dependencies between features, both logically (in the sense that some features are more specific versions of other features, such as ‘Are there clicks?’ and ‘Are there dental clicks?’) and functionally (there may be trade-offs in languages in the complexity of different domains, such as tonal languages being less likely to allow complex consonant clusters). One drawback of the database is that it did not provide references in the publicly downloadable version at the time (and the data itself has since been taken offline). It is also unknown how much agreement there would be between different people in coding a feature for a particular language (a drawback of most other linguistic database such as WALS as well). However, a check of a small number of languages for a few features such as syllable complexity, where I had queries, showed that data collection did systematically record references and that the coding did follow procedures that the coders could defend when asked (Mark Donohue p.c). The main advantage of using this database compared with WALS is the large number of features and languages, as well as the intrinsic interest of the question of how informative phonology and phonotactics are about language history.

This database has not been analysed using Bayesian phylogenetics before, and so the main focus of this chapter will be on converting this data into a format that can be analysed with these methods, and seeing how much historical signal they contain. The reason for using a Bayesian phylogenetic method is that it is a way of quantifying similarity between languages in their sound systems, whether this similarity is due to common ancestry or language contact. Phylogenetic methods model the way that traits have evolved along family trees, and hence are controlling for clustering at lower levels of the hierarchy; this is in contrast with simpler methods such as those employed by Bickel and Nichols (2006), which attempt to quantify commonalities shared by languages in a region such as the Pacific Rim but do not control for local areality such as in Meso-America. If there are such local clusters, a phylogenetic method would find them (making Meso-America a clade for example), and they would not affect the likelihood of putative higher-order clades such as the Pacific Rim.

A Bayesian approach is a direct way of finding the posterior probability of trees and hence in quantifying uncertainty about how languages are related. One approach taken since Gray and Jordan (2000) is to code for the presence or absence of cognate vocabulary. For example, *main* in French and *mano* in Italian are similar enough that they are likely to be related (or borrowed); hence one can code French and Italian as 1 (meaning they have a word which is part of this cognate class), and English as 0 because it has *hand*, which is not in this cognate class. English then gets a 1 for the next cognate class (cognates of *hand*, which exist in Germanic languages), while French and Italian get 0.

This string of 0's and 1's can be read into phylogenetic software in the same way that a DNA sequence can be, and can be analysed using a Bayesian method to work out what family trees are particularly probable. The approach is to ask what the likelihood of a family tree is, namely with what probability the data would be the way that it is if the hypothesis (the family tree being proposed) were true. This can be calculated by summing over all possible scenarios of how an individual trait may have evolved: *main* may have been in the ancestor of all Indo-European languages and then lost in the Germanic branch for instance, or it may have been absent in Proto-Indo-European but innovated by the Romance languages; or it may have been present in Proto-Germanic as well but then lost by all of the modern Germanic languages; and so on. The probability of all of these possible scenarios can be calculated by using a model of trait evolution which has the probability of a trait being innovated, in this case how likely a language is to have *main* or a cognate of it as a word for 'hand'; and the probability of the language losing that trait, in this case a language having *main* as the word for 'hand' but then innovating a different word for 'hand' and using that instead. Some words will be slow-changing, such as the word for 'two' which has been conserved in all branches of Indo-European (recall the modern words for 'two' such as French *deux*, Greek *duo*, Russian *dva*, and Hindi *do*), while most words change more quickly.

The probability of a particular scenario of how a word evolved can be calculated by using these substitution rates and the proposed family tree. The family tree not only specifies how languages are related, but also how much time has elapsed since they diverged. The amount of time is taken into account when calculating the probability of words being lost or gained, as the more distantly that languages are related, the more probable it is that changes in vocabulary will occur.

Finally, these scenarios are summed over (this is calculated using Felsenstein's tree pruning algorithm, Felsenstein (1981)) to produce the likelihood of the tree for that particular word. This is calculated for each word, and the product of all of these likelihoods is the overall likelihood of the tree. This number can then be multiplied by the prior probability of the tree (a number that reflects how probable a particular tree topology is, given that certain tree shapes may be quite unusual), and this gives the posterior probability of the tree. The details of this calculation are described in Huelsenbeck et al. (2001).

There are many possible topologies for a rooted tree (e.g. more than 10^{96} when there are only 60 taxa), and it is impossible to calculate the probability of each one of them. Instead, an algorithm is used that searches through trees and samples them according to their probability, the Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al. 1995).

The logic of the MCMC algorithm is summarised in Dunn (2009): ‘The Monte Carlo Markov chain searches the parameter space following a conceptually simple algorithm in which the likelihood is compared between the current position in the space and another randomly chosen (hence Monte Carlo) position nearby. If the new position in the parameter space has higher likelihood than the current one, it becomes the current position for the next iteration of the search (these repeated searches form a Markov chain, a procedure which retains no memory of its previous states)…if the newly sampled likelihood is lower in an iteration of the Markov chain, a random choice is made between keeping the priors from the current round or using the posteriors, with a probability of choosing the latter equal to the ratio of the new to old likelihood scores… Each iteration moves stochastically through the parameter space, tending towards areas of greater likelihood, until it reaches the equilibrium zone, an area of the parameter space with consistently higher likelihood. The goal of the search is a sample of trees/parameters from the equilibrium zone. Before this state is reached the likelihoods fluctuate wildly with each iteration (likelihood values over a search are plotted in Fig. 4), and the trees produced do not mean anything much—this is called the burn-in period, and these trees are discarded from the analysis.’

The result of running a MCMC chain for typically many millions of iterations will be a set of trees, most of which will fit the data well. This set of trees will reflect the different possible solutions to the question of what the real tree is. It can be summarised by a single consensus tree, which shows groupings (clades) that occur most often in the set of trees found by the algorithm; for example, it may turn out that in 97% of trees in the sample, English is the closest relative of Dutch, and so a consensus tree would put them together and give that clade a 97% probability. Other clades will have much lower probabilities, especially those further back in time.

The same methodology can be applied to structural features, which are often binary (presence or absence of a particular structure, but can also be multi-state. In particular, in modelling the evolution of inventory sizes, a substitution model developed by Maurits et al. (forthcoming) is used, which assumes that phoneme inventories evolve in a step-wise way, being more likely to transition to a number of phonemes that is close by: a transition from 6 tones to 7 or 8 tones should be more likely than transitioning to zero tones.

This analysis will focus on language families in Eurasia in particular, an area that is the general focus of this thesis. The languages under analysis are given a label for what continent they are on, so this was done simply by retaining languages labelled ‘Eurasia’, comprising thirty-six language families according to the (conservative) Glottolog classification. This includes languages in some language families such as Afro-Asiatic and Austronesian, most of which are found outside of Eurasia (in North Africa and the Pacific respectively).

Dependencies are a problem for any Bayesian phylogenetic analysis, because the likelihood of a tree is calculated for each feature and then multiplied to give a total likelihood for the tree. This can only be done if each feature is statistically independent. In practice, it is not clear how much dependencies affect the result of analyses. In addition, there is no easy way of getting rid of them in this data, as both general features ('Are there clicks?') and more specific features ('Are there dental clicks?') may both carry historical information. Functional dependency is also difficult to get rid of, both because phonological systems may have a large number of complex dependencies between features, making them difficult to get rid of entirely, and also because there is no easy statistical way of distinguishing features that correlate in languages because of functional dependency and those that correlate because of shared inheritance and contact. Two possible ways of dealing with these problems are discussed in the conclusion of this chapter, but the main analysis does not attempt to address these issues. It should be noted that it is unlikely that dependencies have a large effect on the results of the analysis, given that the clades in the resulting phylogeny are mostly geographically coherent, suggesting that languages cluster by features mainly because of inheritance and contact, rather than because of dependencies between features.

An additional analysis was run using a methodological innovation suggested in a forthcoming paper by Maurits et al. (forthcoming), which is to use known language phylogenies as monophyletic clade priors. The purpose of doing this, as Maurits et al. write, is to be able to find out which features change quickly and which change slowly, following on from other work that has used Bayesian phylogenetic methods to estimate the stability of typological features (e.g. Dediu 2011). To implement this, the family tree classification from Glottolog was used (Hammarström et al. 2014). Each grouping and sub-grouping in the Glottolog family tree was used as a monophyletic clade constraint (e.g. Indo-European, Germanic, North Germanic, etc.).

The Markov Chain Monte Carlo analysis was run for 10 million iterations in each analysis, with a burn-in of 1 million iterations. The effective sample size (ESS) of all of these analyses for the tree topology is only 4, meaning that the analysis had not been run for long enough to explore the space of possible trees (an analysis is likely to have converged if the ESS is above one hundred according to the BEAST documentation[5]; see section 5.3 for discussion of this point). Reasons for the slowness of the analysis may

[5] https://beast.community/ess_tutorial

include the large number of languages considered (785), but also the time taken in computing the likelihood for the multi-state features, some of which have over one hundred states (see step 3 below). Some ways of speeding up the calculation may include having a randomised sample of Eurasian languages instead of every language (at the expense of losing some information about how features evolve in language families, which requires dense sampling); using threading for the likelihood computations on a cluster; or running several parallel MCMC chains on a cluster to get a fuller sample. The analysis in this chapter is therefore exploratory, and can be improved by implementing these changes to the methodology as well as changes suggested in other places in the chapter (such as dealing with dependencies or using multiple trees).

The main result of this analysis does nevertheless have value because it shows that there is historical signal in phonological/phonotactic features, such that geographically coherent clusters of languages emerge; and that these clusters demonstrate patterns of language contact, because they cross-cut language families. For comparison, a hierarchical cluster analysis of the data was also done using a distance-based method UPGMA (Sokal and Michener 1958), implemented in the Python package scipy and described in section 4.2.

The method is summarised below:

1. I took data from the World Phonotactics Database, choosing specific features (excluding word order). This file can be obtained by contacting the author requesting the file data.csv in the Supplementary Materials (Chapter 3). The languages used were all in Eurasia, defined as having the WALS classification ‘Eurasia’ (which Beastling uses). The features that were excluded were 'Language', 'Latitude', 'Longitude', 'ID', 'Language family', 'Language family1', 'Language family2', 'Language family3', 'Language family4', 'Language family5', 'Is this language an isolate', 'World macro-region', 'WALS area', 'Autotyp', 'Country', 'Ancient language', 'Papuan', 'Altitude', 'Eastness', 'SOV', 'SVO', 'VSO', 'VOS', 'OVS', 'OSV', and 'Maximal CV contrasts'. All other features were used, making a total of 141 features (102 binary and 39 non-binary, as described in points 2 and 3 below).

2. 102 binary features, such as whether a language has clicks (‘yes/no’), were modelled using a covarion model. A description of the covarion model in Beastling (Maurits et al. 2017) is provided in the documentation[6]: ‘The binary Covarion model is defined for binary datasets, i.e. sets where every datapoint is either a 0 or a 1. This

[6] https://beastling.readthedocs.io/en/latest/substitution.html.

model introduces a latent "fast" or "slow" state, which controls the rate of transitions between 0 and 1 (transitions in either direction are always equally probable). This model is typically used for cognate data, but can be used for binary structural data also…This model estimates two parameters, a switching factor which governs how frequently the latent state switches between "fast" and "slow", and a parameter denoted "alpha" which controls the difference in speed between the two states. By default, these parameters are shared across all features in the dataset, i.e. BEAST will estimate 2 parameters for n features.'

3. 39 multi-state features, such as the number of tones that a language has, were modelled using an ordinal model developed by Luke Maurits (Maurits et al. forthcoming), which is described in the following way, using the example of WALS consonant inventory sizes: '[The model] permits transitions only between adjacent values on the scale…with all such transitions being equally probable. We call this the Ordinal model. Note that when we speak of permitting or not permitting certain transitions, we are considering an infinitesimal interval of time. A language cannot increase its consonant inventory from 'small' to 'large' in a single such interval, however, it can make four permitted transitions in a series of consecutive intervals which occur over an arbitrarily short period of time, so that if the data truly demands a tree topology in which consonant inventory changes rapidly in a short period of time, the model can accommodate this. These rapid changes are therefore simply strongly dispreferred over more gradual explanations.' The same model is used here, applied to numbers such as the number of tones that languages have. The way that both binary features and multi-state features can be accommodated is to have two separate 'partitions' in BEAST, one for the binary features and one for the multi-state features; this means that both feature sets share the same tree topology, but have separate substitution models.

4. The data was converted to a xml file for BEAST using the Python package Beastling (Maurits et al. 2017). This file is available on Github in the Supplementary Materials (1.xml).

5. I ran the xml file in BEAST for 10 million iterations, resulting in trees and log files (1.trees, 1.log). The trees were summarised using the TreeAnnotator package in BEAST in a consensus tree and plotted in Figtree.

The procedure was the same for the clade-constrained analysis described in section 4.1, except for the following additional step:

6. When converting the data to a xml file for BEAST using Beastling, an additional line monophyletic = True was included in the Beastling file, which uses Glottolog trees to constrain the topology of the trees being searched for in Beast (e.g. the trees being searched for must include Indo-European as a monophyletic clade).

3. Results

The main result is a tree of Eurasian languages, comprising clades with different degrees of support. The full consensus tree is presented in the supplementary materials and is plotted in Appendix 2 (Figs. S1.1-11), but a summarised version is given in Figure 3.1.

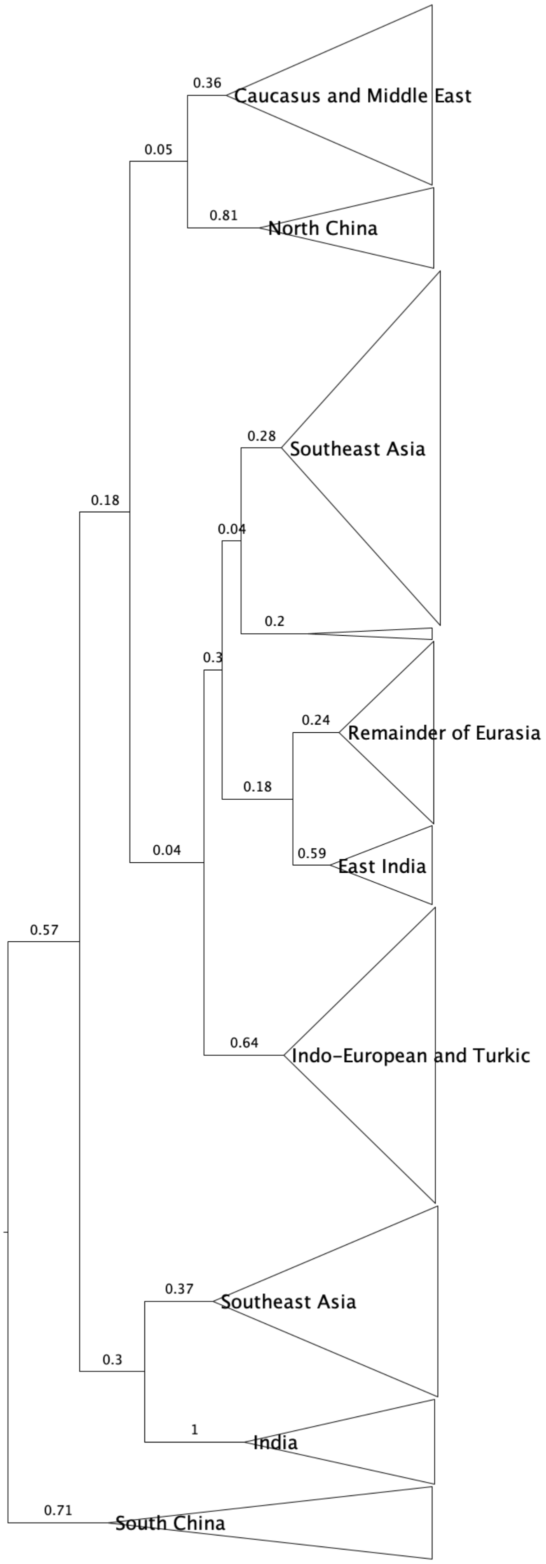


Figure 3.1: A consensus tree of languages in Eurasia according to the phylogenetic analysis of the World Phonotactics Database. The main clades are collapsed and summarised with a geographical label for readability. Numbers on each node reflect the posterior probability of the clade, between 0 and 1.

Because the tree is too large to describe in full, the large clades in Figure 3.1 will be summarised below, along with some of the more strongly supported sub-clades. The posterior probability is given in brackets, with brief descriptions of the location and main language families. The general result is that several geographically coherent clusters of languages emerge from the analysis, which is surprising given that only information about phonology and phonotactics is being used. A certain amount of parallel development of features is to be expected, meaning that some of these clusters also have a few outliers in other regions.

1. South China (71%), mostly comprising Sino-Tibetan, Hmong-Mien, and a few Tai-Kadai languages. The languages are shown in Figure 3.2, and are coloured to show their language family: Sino-Tibetan in dark blue, Hmong-Mien in red, Tai-Kadai in yellow and Austroasiatic in light blue. This area may be thought of as a south China area, mostly comprising Sino-Tibetan and Hmong-Mien languages, with a few Tai-Kadai languages and one Austroasiatic language in southern Laos. In the consensus tree that emerges from the phylogenetic analysis, this clade is in fact on the outside of the tree, with all other languages in Eurasia forming a clade with 57% posterior probability. This means that other Eurasian languages have more in common with each other than they do with languages in the south China area. A more contentious way of putting it would be that there is evidence for all Eurasian languages outside of south China together forming a linguistic area, that is, having some properties in common. Perhaps this is just an artefact of the fact that only Eurasian languages are used in this analysis; if languages from other continents were included (an analysis not yet performed), it may turn out that this clade is no longer strongly supported, or that it is supported and is due to most languages in Eurasia having a common historical signal (either due to inheritance or contact). This large area is therefore hard to interpret, but is worth noting as a theoretical strength of the method - that it can in principle detect large areas that linguists would not even have thought of looking for (an entire continent minus a linguistic area could in principle be itself a linguistic area).

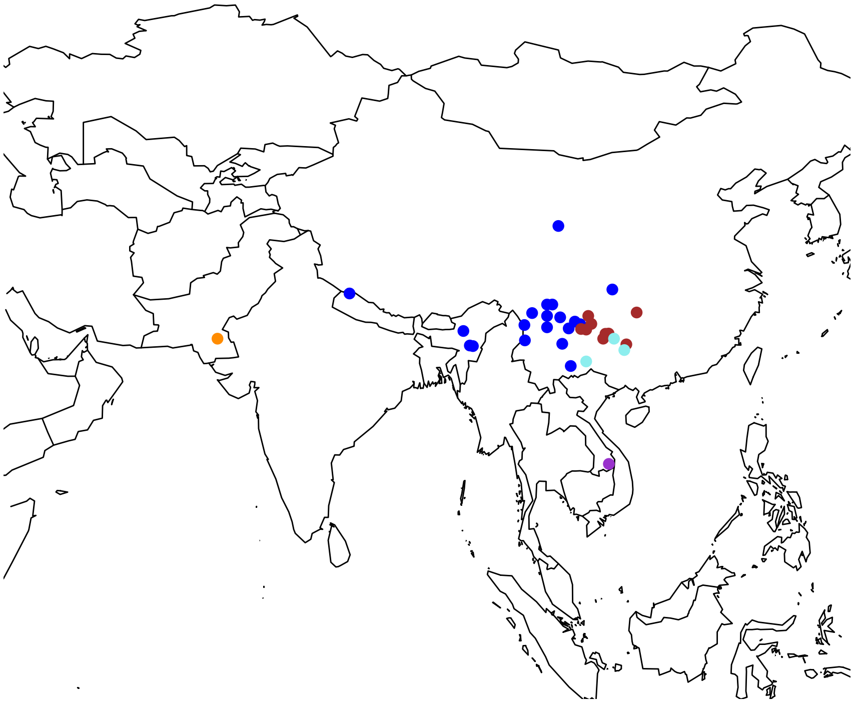


Figure 3.2: A clade with 72% posterior probability comprising mostly languages in southwest China, coloured by language family (blue: Sino-Tibetan, brown: Hmong-Mien, orange: Indo-European, purple: Austroasiatic, light blue: Tai-Kadai).

2. India and western Southeast Asia (30%), shown in Figure 3.3. This group splits into two sub-groups, one in India (100%) and one in Southeast Asia (37%). The group in India is mainly made up of Indo-European and Tibeto-Burman languages, shown in Figure 3.4 (as well as Scots Gaelic as an outlier). The group in Southeast Asia is mainly made up of Tibeto-Burman languages in the Himalayas, with a few Austroasiatic languages.

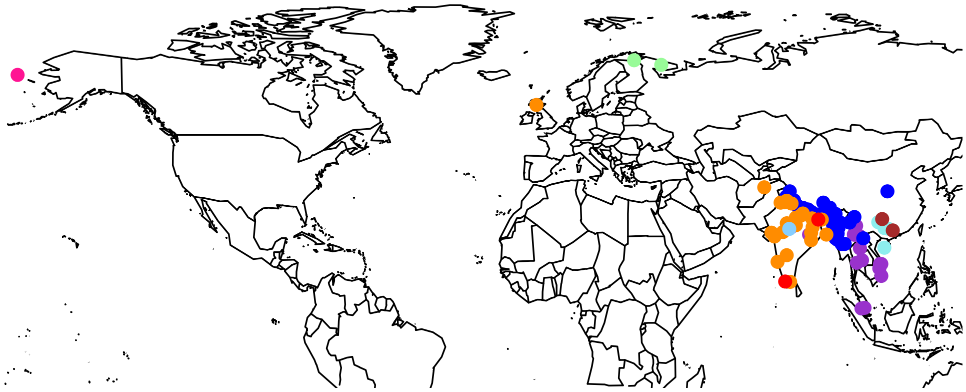


Figure 3.3: A clade with 30% posterior probability mostly comprising languages in Southeast Asia and India, coloured by language family (purple: Austroasiatic, blue: Sino-Tibetan, light blue: Tai-Kadai, orange: Indo-European, red: Dravidian, sky blue: Nahali, green: Uralic, pink: Eskimo-Aleut).

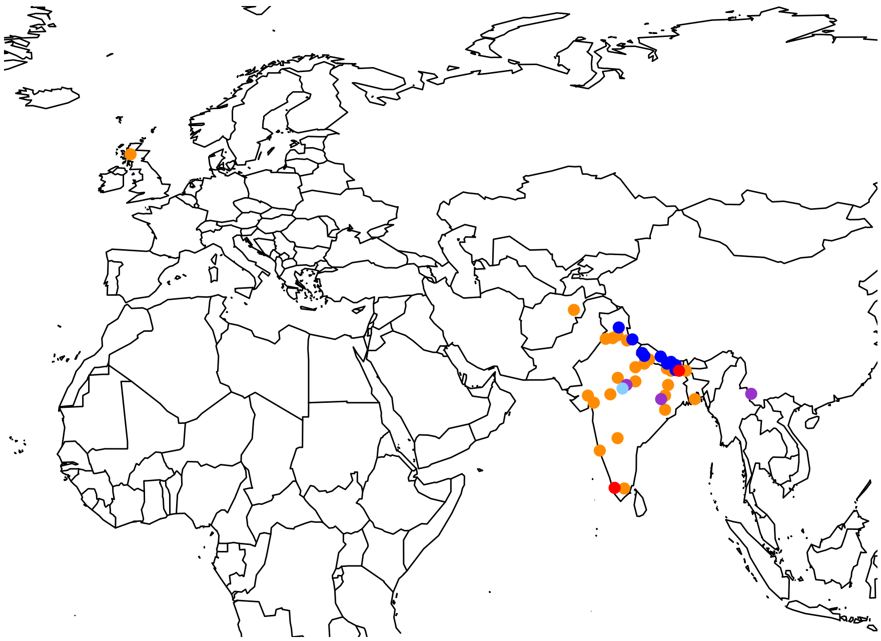


Figure 3.4: A sub-clade of the clade in Figure 3.3, with 100% posterior probability, which mostly comprises languages in India, coloured by language family (orange: Indo-European, red: Dravidian, purple: Austroasiatic, sky blue: Nahali).

3. Indo-European and Turkic languages (64%), shown in Figure 3.5. This clade splits into two overlapping groups, one primarily Indo-European in Europe, comprising some Slavic and Romance languages (100%). The other group splits in two, one group containing the remainder of the Indo-European languages (45%), and a second group (17%) which contains a Turkic sub-clade (82%) and two strongly supported but strange groups, described in (4) and (5). There are also outliers in this cluster such as two Uralic languages and an Eskimo-Aleut language, reminding us that independent convergence in phonological features does occur.

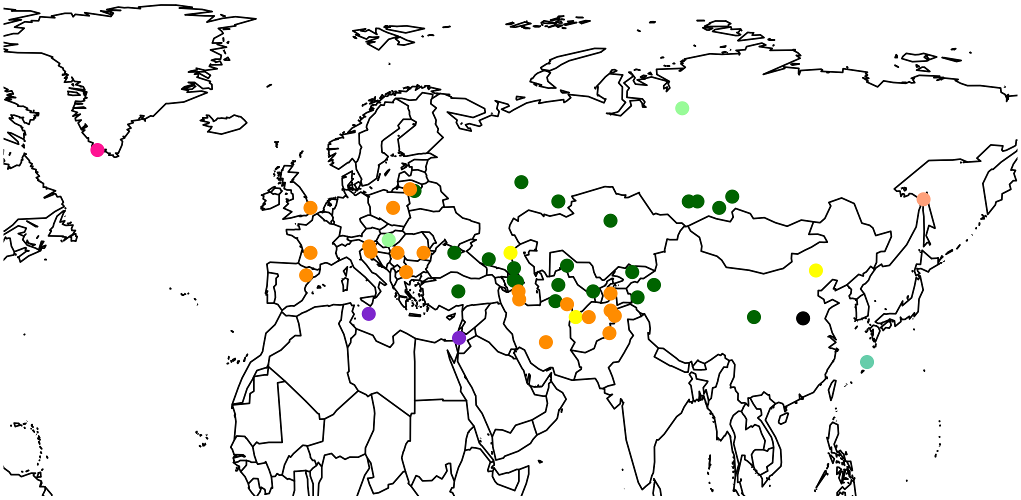


Figure 3.5: A clade comprising languages in western and northern Eurasia, mostly Indo-European and Turkic (orange: Indo-European, dark green: Turkic, black: Tungusic, yellow: Mongolic, purple: Afro-Asiatic, aquamarine: Japonic, light salmon: Chukotko-Kamchatkan, pink: Eskimo-Aleut, light green: Uralic).

4. A disparate group (99%) containing Yenets, Maltese, Greenlandic, Tajik, Romany and Moghol.

5. Another disparate group (99%) containing Amami Ryukyuan, Nivkh, and two Indo-European languages, Avestan and Waigali.

6. Eastern and southern India, containing mostly Austro-Asiatic and Dravidian (58%), shown in Figure 3.6. There are also a few Indo-European languages such as Kumauni, suggesting cases where they have been in contact with Dravidian. Nynorsk

(Norwegian) is also included in this clade, again an outlier in an otherwise geographically coherent cluster.

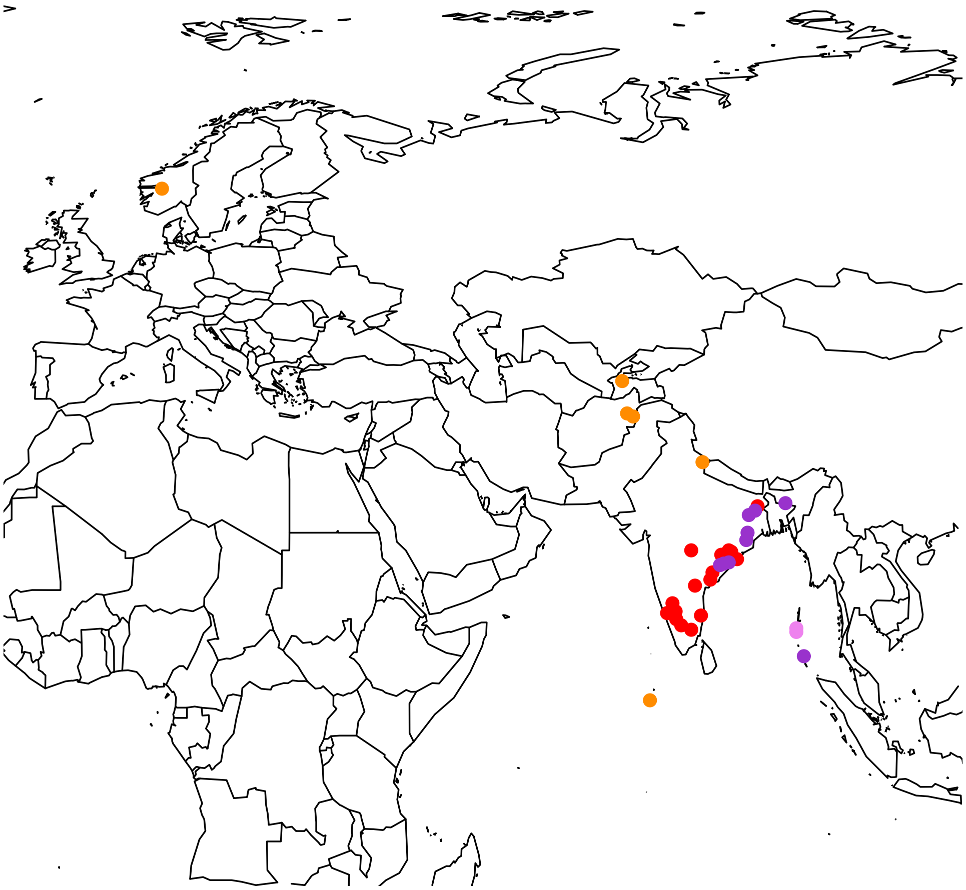


Figure 3.6: A clade comprising languages in India with 59% posterior probability, coloured by language family (violet: Andaman, purple: Austroasiatic, red: Dravidian, orange: Indo-European).

7. Southeast Asia (28% posterior probability), shown in Figure 3.7, mostly comprising Sino-Tibetan, Austro-Asiatic and Tai-Kadai, but also Korean. This clade has a relatively low posterior probability, in contrast to the expectation that Southeast Asian languages would form a clear linguistic area.

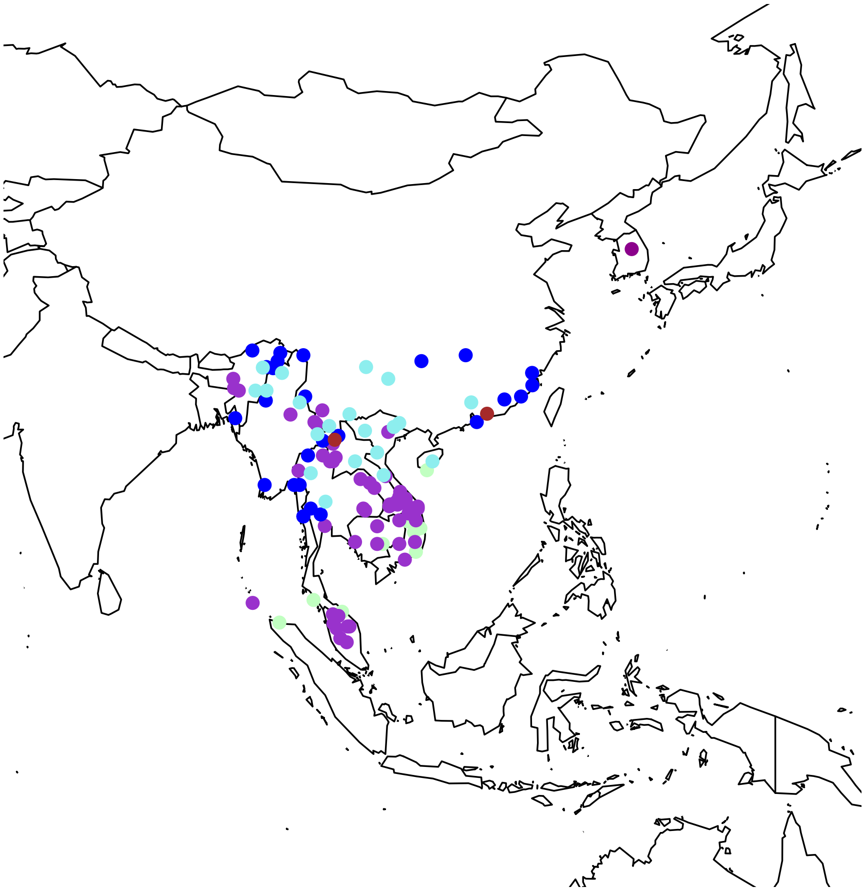


Figure 3.7: Southeast Asia, coloured by language family (purple: Austroasiatic, light blue: Tai-Kadai, dark blue: Sino-Tibetan, light green: Austronesian, magenta: Korean).

8. An arc of languages running from the Caucasus through the Arabian Peninsula and into Afghanistan and Pakistan (36%), shown in Figure 3.8. This group is richly structured, comprising smaller groups such as just the Caucasus and Middle East (29%); just Nakh-Daghestanian and Kartvelian (47%); the Northwest Caucasus (83%); and Arabic varieties along with Kurdish and Brahui (64%).

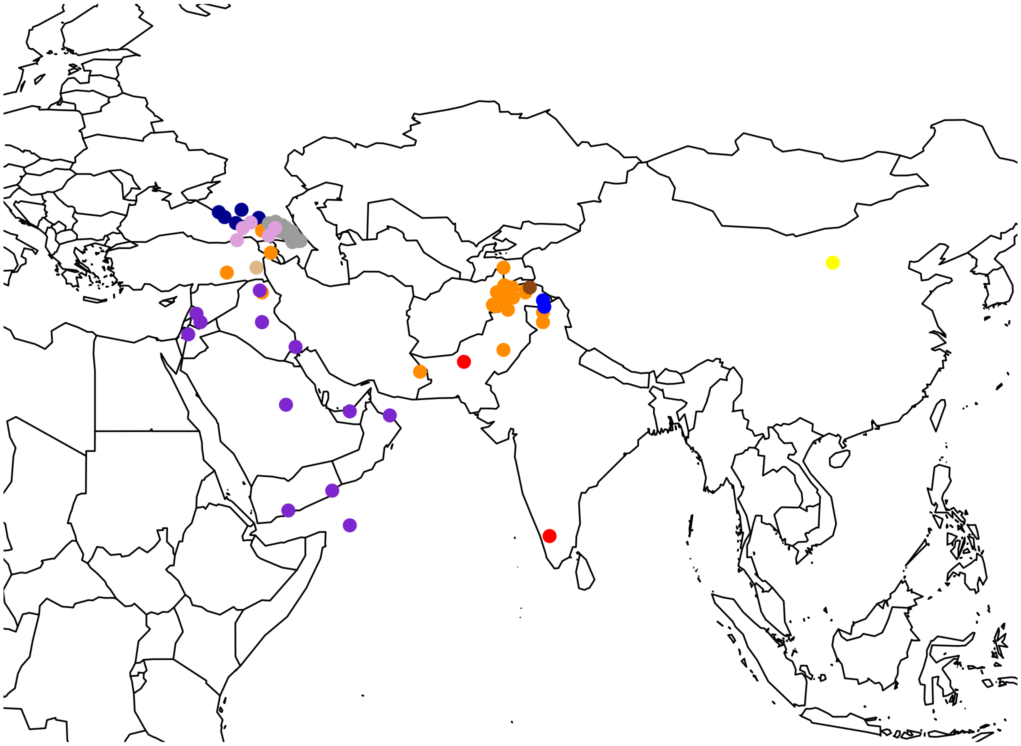


Figure 3.8: A clade with 36% posterior probability mainly comprising languages in the Caucasus, Arabian Peninsula, and northwest India, coloured by language family (midnight blue: Northwest Caucasian, orange: Indo-European, red: Dravidian, plum: Kartvelian, blue: Sino-Tibetan, gray: Nakh-Daghestanian, brown: Burushaski, light brown: Hurro-Urartian, purple: Afro-Asiatic, yellow: Mongolic).

9. Western and northern China (81%), shown in Figure 3.9. This group contains a sub-clade of Mongolic, Turkic and northern Sinitic languages (81%), suggesting that languages such as Mandarin have influenced or been influenced by non-Sinitic languages in north China.

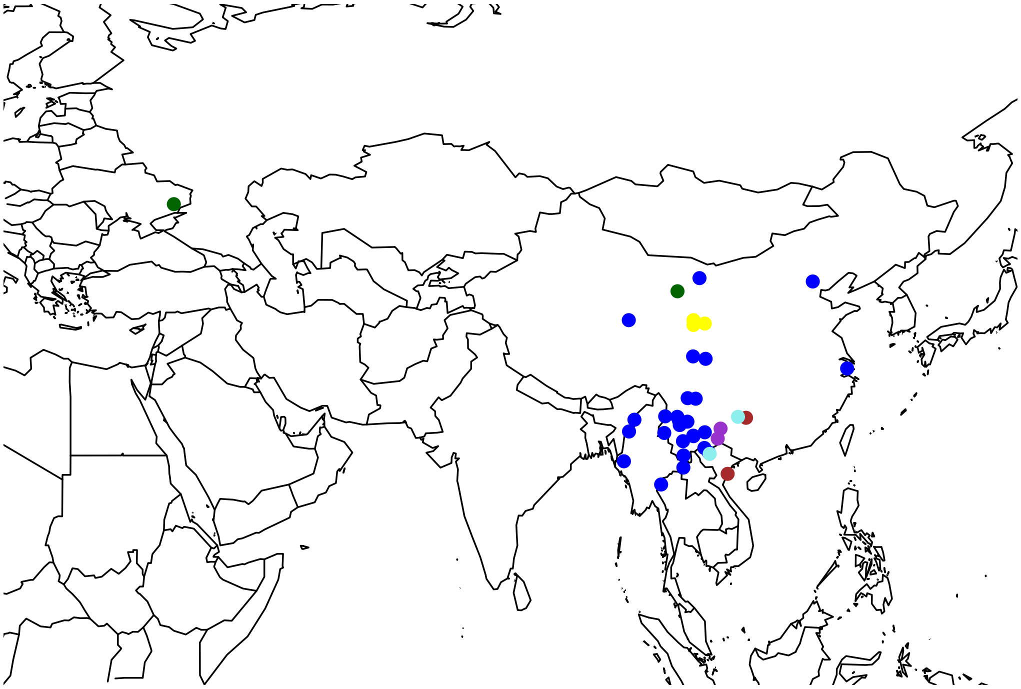


Figure 3.9: A clade comprising languages in mostly in western and northern China (81%), coloured by language family (blue: Sino-Tibetan, yellow: Mongolic, green: Turkic, purple: Austroasiatic, light blue: Tai-Kadai, brown: Hmong-Mien).

These groups are plausible candidates for being linguistic areas, in the sense that they are groups of languages with relatively high posterior probabilities, and are geographically coherent (with the exception of 4 and 5). To some extent this may be due to the fact that linguistic sub-families tend to be geographically contiguous; but they are also often restricted to a particular region, such as Southeast Asia or India (perhaps with occasional languages that are mistakenly included, such as Nynorsk mentioned above).

There is a lot more to explore in the results, such as the interactions between particular languages and how plausible these interactions are historically. An example is the structure of the outer clade in Southeast Asia (clade 1 in the list above). The structure of this clade is shown in Figure 3.10.

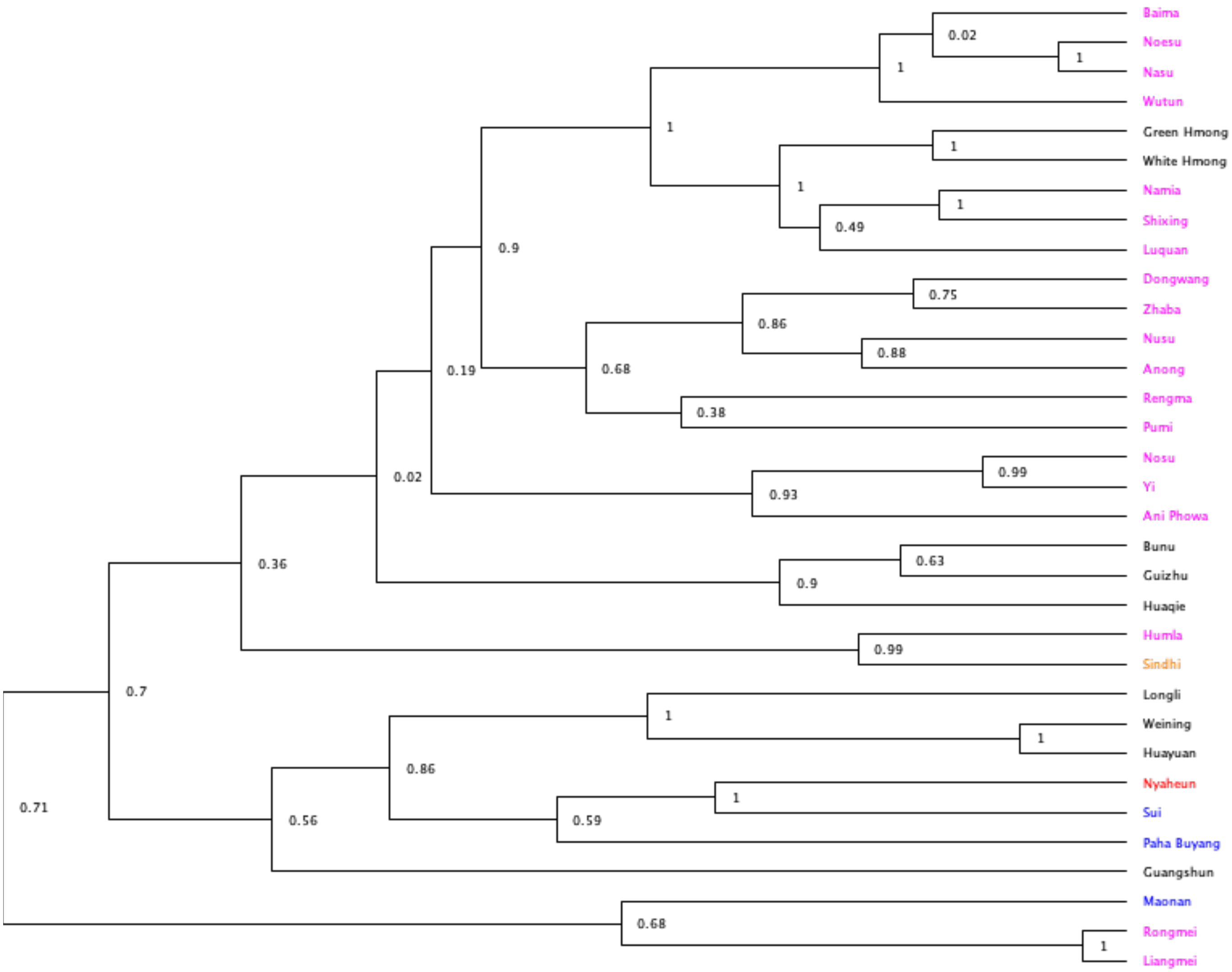


Figure 3.10: The structure of clade 1, comprising languages mostly in South China (see Figure 3.2). The taxa names are coloured by language family; Sino-Tibetan in purple, Hmong-Mien in black, Tai-Kadai in blue, Austro-Asiatic in red, Indo-European (Sindhi) in orange.

The tree shows that two Hmong-Mien languages, Green Hmong and White Hmong, are phonologically similar to a group of Sino-Tibetan languages, placed in a group with 100% posterior probability, suggesting convergence in their phonology due to contact. This clade is in a larger clade with a larger group of mostly Hmong-Mien and a few Tai-

Kadai languages with 70% probability. Finally, on the outside of the clade are two Sino-Tibetan languages of northeast India, grouped with a Tai-Kadai language Maonan with 67% probability.

A different clade in Southeast Asia is clade 7 in the list above. A particularly interesting sub-clade is shown in Figure 3.11, showing interactions between Korean and Sino-Tibetan languages. Korean is placed in a clade with 96% probability with two southern Chinese Sinitic languages (Fuzhou and Changsha), a Hmong-Mien language She in Guangdong, and the Tibeto-Burman language Akha in Laos. This is most likely because of the phonological influence of southern Chinese languages on Korean, for example through borrowing of vocabulary (e.g. de Roulet 2018). This sub-clade is part of a larger clade of mostly Sino-Tibetan languages again, with 96% probability.

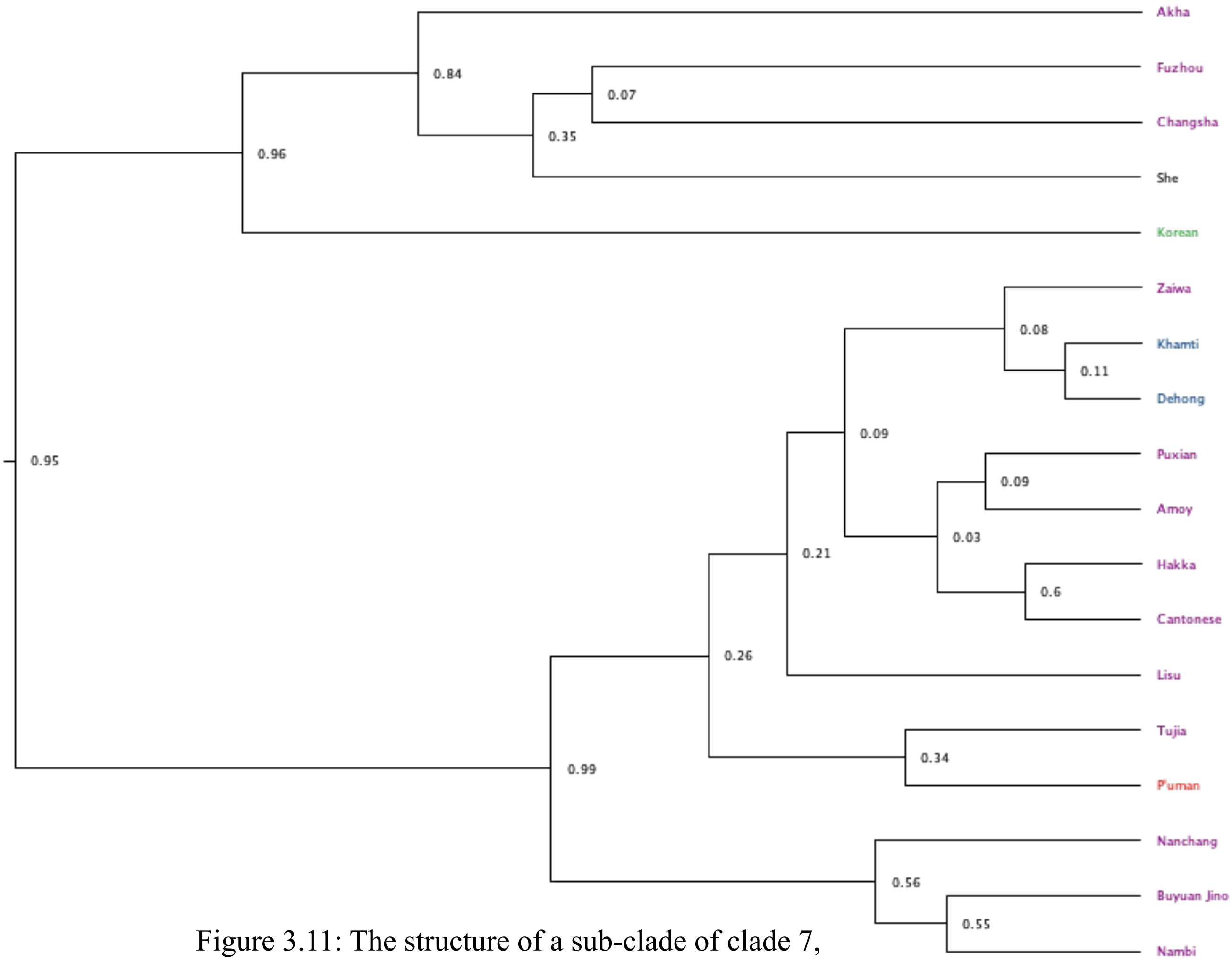


Figure 3.11: The structure of a sub-clade of clade 7, comprising Korean and Sino-Tibetan languages that may have influenced it.

In many cases the analysis recovers groups of related languages, as expected because phonotactic properties are inherited from ancestral languages. But the method is unreliable for reconstructing common ancestry, and does particularly badly in reconstructing families such as Indo-European. Indo-European languages are not placed in a single clade for example, being placed with nearby languages such as in Southeast Asia, the Caucasus, or Uralic languages. Some relationships within Indo-European are suggested, such as among some Slavic languages (40%), but many languages in these groups are otherwise scattered around the tree. Some sub-clades show clear language contact, such as one that contains Basque and Spanish dialects (83%), but many are likely to be simply due to independent convergence, such as a clade containing Greek, Sicilian, Manx and the Chukotko-Kamchatkan language Itelmen (43%).

4. Additional Analyses

4.1 Results of the Clade Constrained Analysis

This section describes the results of the analysis where the clades are constrained to include known language families. This is to test a hypothesis due to Maurits et al. (forthcoming) that this method can potentially recover macro-families; and in doing so, some clearer results about linguistic areas are also obtained.

As with the analysis in section 3, the main result is a tree of Eurasian languages, this time with known language families as clades. The interpretation of these results is unclear, since phonological properties do not simply travel in language families; but several large macro-areas emerge from this analysis, which are likely to reflect more ancient relatedness or language contact. The main results are:

1. A clade covering most of Eurasia, in a way that resembles the Eurasian linguistic area of Bickel and Nichols (2009) (85%). This clade covers all language families in the sample except for families of Southeast Asia.

2. A clade covering language families in Southeast Asia (Austronesian, Tai-Kadai, Sino-Tibetan, Hmong-Mien, Austroasiatic, Shompen) (82%). This is less surprising as it corresponds to a well-known Southeast Asian linguistic area (e.g. Enfield 2005), although it is gratifying to see this supported in a quantitative way based entirely on phonological features.

3. Within the main Eurasia clade, there is a clade comprising the Caucasus and Afro-Asiatic (94%).

4. This clade is also part of a larger clade (48%) containing various language families in Eurasia, such as Turkic, Uralic, Nivkh, Ket, Chukotko-Kamchatkan, and Eskimo-Aleut.

5. A Turkic-Uralic clade (96%), suggesting recent interaction between these two northern Eurasian families.

6. Within Southeast Asia, there is a clade comprising Tai-Kadai, Austroasiatic, Austronesian and Shompen (81%).

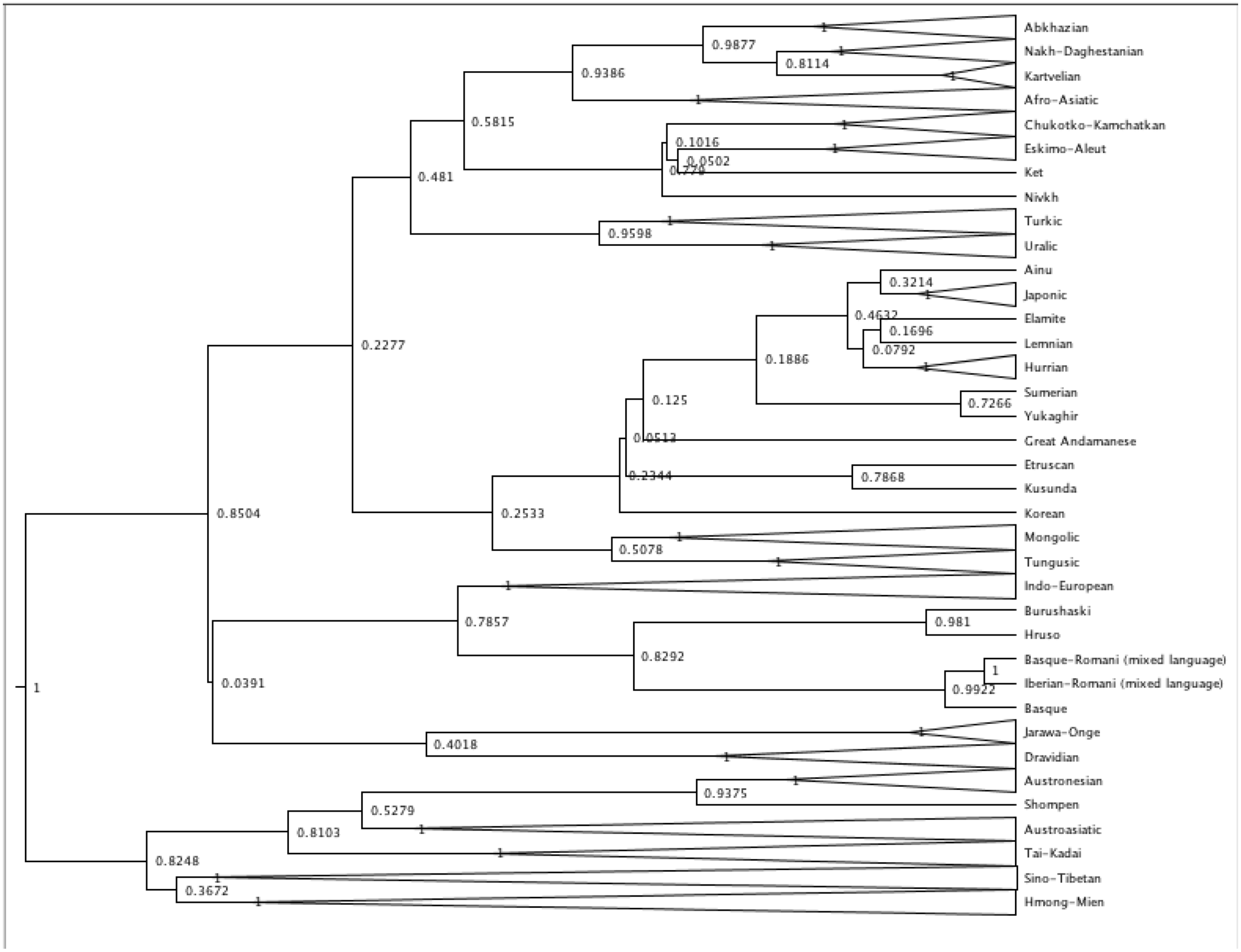


Figure 3.12: The family tree of language families in Eurasia based on phonotactic properties, with constrained clades for language families. Numbers on clades are posterior probabilities.

4.2 Distance-based Cluster Analysis

For comparison, a distance-based cluster analysis using the UPGMA algorithm (Unweighted Pair Group Method with Arithmetic Mean; Sokal and Michener 1958) was also performed on the same data. The implementation was from the Python package Scipy using the 'cluster hierarchy' module. The method was as follows:

1. An array of shape 785 x 141 (785 languages, 141 features) was made; the array was normalised by dividing the value for a feature by the square root of the sum of the values squared for that feature across all languages (in Python, x / numpy.sqrt(numpy.nansum(numpy.square(x), axis=0)))

2. A distance matrix was made for the languages by taking the mean absolute difference in the arrays, ignoring missing values (numpy.nanmean(abs(x - y))).

3. The UPGMA algorithm implemented in the function 'linkage' in Scipy was used, which iteratively merges clusters based on which are closest together (defined by taking the mean of the distances between members of the two clusters). It constructs a rooted tree, which was then plotted in Figtree (see Appendix 2, Figure S2).

The purpose of the comparison is to answer some possible objections to the use of the Bayesian phylogenetic method in section 3, such as: why use a phylogenetic approach, rather than just another standard hierarchical clustering method? Why use a Bayesian method that is computationally more intensive (and which in this case resulted in a low effective sample size)? As discussed in Chapter 1 and section 1 of this chapter, the Bayesian approach models the way that the features are evolving along the tree, whereas the distance-based method does not, instead using a cruder heuristic of similarity between languages. The Bayesian method also searches through the possible rates of change of features, which is useful because features evolve at different rates; by contrast, the UPGMA algorithm gives equal weighting to the features, at least in the version implemented here. It is possible that a version could be tried which iteratively updates the weighting of the features in some way, at the same time as finding the clusters. Finally, the UPGMA algorithm is not attempting to search through the space of trees and make a sample based on likelihood, instead returning only a single tree based on a simple bottom-up heuristic. By contrast, the Bayesian approach searches through the space of trees more and returns trees sampled in proportion to their posterior probability, and so even with a small effective sample size is arguably a more extensive search than a method that returns a single tree. For these reasons, the Bayesian approach is expected to yield a result that is closer to the reality of how languages have inherited phonotactic features (either in language families or by contact).

The results of the distance-based analysis bear this out. The dendrogram is plotted in Appendix 2 in S2.1-11. There are some similarities with the Bayesian phylogeny, plotted in S1.1-11, as one would expect. For instance, in S2.8 there is a clade of languages in the Caucasus belonging to different families such as Kartvelian, Nakh-Daghestanian, and Indo-European; it continues into S2.9 with other families such as Northwest Caucasian and some Semitic languages. However, some Afro-Asiatic and Northwest Caucasian languages appear in a different part of the tree, in various clades in S2.11. By contrast, in the Bayesian phylogeny in S1.1 there is not only a single coherent clade comprising all of those languages, but also some further structure to its sub-clades, such as one of mostly Afro-Asiatic languages (64% confidence), one of mostly Nakh-Daghestanian (47%), one of Northwest Caucasus (83%) and one of mostly Indo-European (89%) (although each of these clades also has ‘mistakes’, such as the Indo-European language Yazgulami placed with the Northwest Caucasus languages). The Bayesian analysis in this case seems to uncover more meaningful clusters than the distance-based analysis.

There are other similarities between the two trees, such as the clade comprising Korean, She (Hmong-Mien), Fuzhou (Sinitic), Puxian (Sinitic) and various other Sino-Tibetan languages (S1.4 in the Bayesian tree and S2.5-6 in the distance-based tree). Clades which appear in the results of both analyses may be more likely to be reflecting genuine similarity, rather than a spurious result in the Bayesian analysis due to lack of convergence. Another interesting example is the placement of Nynorsk with languages in India (S2.1) which replicates the result from the Bayesian analysis, perhaps due to the convergent evolution of features such as tone and retroflex consonants (which Nynorsk is described in the database as having). But there are also many dissimilarities: there are no shared clades of languages in India for example (e.g. the placing of Malayalam with different languages in S1.10 and S2.1). In some cases, a ‘mistake’ in the Bayesian phylogeny may be due to lack of convergence, for example the placing of Scots Gaelic with languages in India in the Bayesian phylogeny (S1.10); in the distance-based analysis, Scots Gaelic is placed on the outside of a clade of languages mostly in Western Europe (S2.2).

Impressionistically, the distance-based tree has many small geographically coherent clades. However, it is difficult to find much geographical coherence to larger clades, unlike in the Bayesian phylogeny, where several large clades comprise languages in a single region such as East India or South China. This may be due to the three properties of the Bayesian analysis described above that the distance-based analysis lacks: modelling the history of the features, modelling the different rates of change of the features, and more extensive search through the space of possible trees. While a distance-based analysis provided here provides a baseline clustering for the languages in the database, the Bayesian analysis in section 3 seems to provide more meaningful results despite shortcomings such as the lack of convergence.

5. Discussion

5.1. Phylogenetic mixture models

One disadvantage of the phylogenetic method used (as well as UPGMA) is clearly that it forces languages to belong to a particular clade, whereas in reality languages may have more than one parent. For example, Tibeto-Burman languages in the Himalayas may have structural properties from both Indo-European languages in India and Sino-Tibetan or other families in Southeast Asia (the 'Indosphere' and 'Sinosphere': see Matisoff 2015). What is needed is a method that can allow languages to have some properties inherited from one tree, and some features inherited from another. The mathematical framework for this method exists, having been described in Pagel and Meade (2004). Gray, Greenhill and Ross (2010) allude to this framework, adding that this method has not been applied to cultural data: 'We suggest that Bayesian phylogenetic mixture models (Pagel and Meade 2004) could be used to investigate complex histories at the level of specific characters. Instead of forcing all the characters onto a single tree, these mixture models allow different models of evolution to be applied to each character in the data. Essentially this allows the characters to "choose" between alternative trees. A multiple topology mixture model is currently implemented in Bayes Phylogenies (Pagel and Meade 2004) but, to the best of our knowledge, has not yet been used in studies of cultural evolution.' This remains true at the time of writing, and the application of a multiple topology model is particularly appropriate for structural data.

A mixture model can select between two trees, as illustrated in Figure 3.13. For instance, while much of the basic vocabulary of English is Germanic (such as *hand*), some of its vocabulary is from French (such as *bottle*). Tree A, which places English in a clade with Germanic languages, may receive a high likelihood for some traits (such as the word for 'hand'), but a low likelihood for other traits (such as the word for 'bottle'). Tree B places English with Romance languages, and would receive a high likelihood for traits such as 'bottle', but a low likelihood for traits such as 'hand'.

Neither tree is a good fit for the entire data, but a model which assumes that some traits follow tree A and other traits follow tree B can do better than a model which assumes a single tree. The probability of a trait is calculated for each tree, and the maximum value (or the mean, or a weighted average) of the two likelihoods is then taken. This becomes the likelihood of that pair of trees for that trait; the likelihoods are then calculated for each trait in this way, and the product of all of the likelihoods for all of those traits is the likelihood of the model. This model can be sampled by MCMC in exactly the same way as a model which has just one tree, but sampling pairs of trees instead.

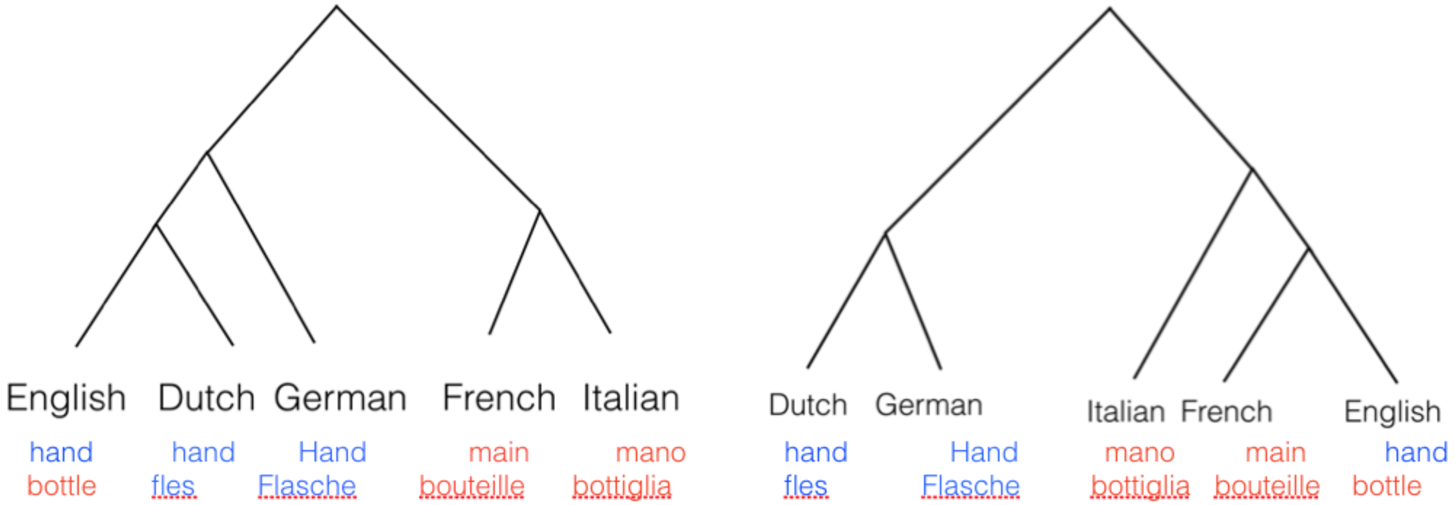


Figure 3.13: An illustration of the phylogenetic mixture model, in which two trees are assumed to generate the data. Each trait, such as ‘hand’ or ‘bottle’, can be chosen to have been generated from one of the two trees.

A further possibility is constraining trees in a mixture model to be correlated with each other, for instance by forcing trees to be within a certain ‘edit distance’ of each other. This would allow the inference of two trees, but also allow them to use information from each other. To use the example of the words ‘hand’ and ‘bottle’ above, it is unlikely that words have entirely unrelated histories from each other - since German and Dutch share cognates for both ‘hand’ and ‘bottle’, and Italian and French share cognates for both of them too, making the histories of those two words highly correlated. It would make sense to use information from both of these words in constructing the two trees, but to also allow the trees to have differences; for instance, in making English belong to a clade with Germanic languages for the word ‘hand’, and with Romance languages for the word ‘bottle’.

Whether this constraint is added or not, the benefits of this model will be that it produces a pair of trees, showing how one set of structural features may have been transmitted in a different way from another set. This will then show how some languages belong to two different areas, such as both an Indian area and a Southeast Asian area; or in particular, how languages can have certain properties similar to languages they are closely related to, and certain other properties that are similar to their neighbours because of language contact.

5.2. Logical and functional dependencies

Another problem in the model implemented in this chapter is that structural features are treated as statistically independent of each other. In reality, as mentioned, structural features are dependent on each other for two reasons. Certain features are logically dependent on others, such that if one knows the answer to one question, an answer to another question is entailed automatically; for instance, if an answer to the question ‘Are there clicks?’ is no, then the answer to ‘Are there dental clicks?’ must also be no, given that dental clicks are a type of click. The other reason is that certain features are statistically more likely to occur together for functional or historical reasons. For instance, if a language has one type of click such as dental clicks, it is quite likely to have another type of click such as alveolar clicks, the probability being 21/27 in the World Phonotactics Database. A more subtle example is that a language with complex consonant clusters in the coda of syllables may be unlikely to have complex tone (i.e. three or more tonal contrasts), the probability being 25/489. Other examples of functional dependencies include Greenberg’s ‘universals’ of grammar (Greenberg 1966), such as the observation that "With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun”; Dryer (1992) and Dunn et al. (2011) are two studies that study word order dependencies (the former using sampling from different families and areas, the latter testing the strength of these dependencies within four language families).

These two types of dependencies can be referred to as ‘logical’ and ‘functional’ dependencies respectively. Logical dependencies can in fact be handled in a phylogenetic analysis, by using a method also employed in a phylogenetic analysis of the story ‘Little Red Riding Hood’ (Tehrani 2013). In that paper, variants of the tale were coded for the absence or presence of particular story traits, which have logical dependencies between them (certain events can only happen in a story variant if an earlier event has happened). If event B is logically impossible because an event A that it depends on has not occurred, then the question about whether event A has occurred is answered with a ‘0’, and the question about whether event B has occurred is answered with a ‘?’, a symbol standing for missing data. Software such as BEAST can ignore values in the data which are missing, and so two stories which differ for their value for event A will not be counted on events which depend on event A. Stories which do have event A can then be compared on whether they have event B. This can be continued indefinitely for chains of logical dependencies; for instance, another event C can be dependent on event B, and can similarly receive a ‘?’ if event B has received either a ‘0’ or a ‘?’.

This approach can be used to handle logical dependencies in a linguistic database, namely using ‘?’ for features which are dependent on another question. A question such as ‘Are there clicks?’ may be answered with ‘0’, which means that ‘Are there dental clicks?’ should be answered with ‘?’. The question about the number of clicks, ‘How many clicks are there?’ should also be answered ‘?’ rather than ‘0’. This can easily be implemented in the analysis described in this chapter, and should be if it is replicated or tried on another dataset which has logical dependencies.

Functional dependencies are harder to handle in a phylogenetic analysis. In an extreme case, a large number of correlated traits could cause an analysis to form erroneous clusters, if each dependent trait supports one particular clustering. A humorous example by Dawkins (2004) is in classifying millipedes: if there are red millipedes and blue millipedes, one could attempt to classify them using morphological traits, but it would be a mistake to count the colour of each leg as one thousand independent features (the colour of leg 1, of leg 2, etc.), rather than as a single general feature (the colour of the legs), since that would be counting a single embryological feature (leg colouring) one thousand times. One may be making an analogous mistake by treating different types of click sounds (alveolar clicks, dental clicks etc.) as functionally independent, since they are the result of a single property of the language, that it uses clicks at all[7].

One possible way of dealing with functional dependencies is to run the phylogenetic analysis here, and then use the resulting phylogenies to test for the correlation between different features. The correlation test would be the same as that used to test word order correlations by Dunn et al. (2011), which takes known family trees and then tests models for how features co-evolve (namely whether some combinations are favoured over others). This test could in principle be used on the tree(s) discovered in this chapter, testing whether particular features are likely to have co-evolved even after taking into account their shared history. For instance, no matter how you arrange languages in a family tree, it may be likely for certain properties such as tone to be linked with syllable structure, and this would show up in the way that languages in different parts of the tree have independently evolved certain combinations of properties (such as simple syllable structure and complex tone). The phylogenetic analysis could then be rerun by either

[7] This is not to claim that there is a deep analogy to be made between biology and language: the processes which causes dependencies between features are superficially similar but distinct, such as natural selection and embryological constraints in biology, and constraints of use, acquisition and historical development in linguistics. The analogy is invoked here merely to point out that the same quantitative problem can arise in both disciplines when using a phylogenetic method that assumes independence of traits.

removing some functionally dependent traits, or collapsing them into a single multi-state value; for instance, assuming that a language is more likely to have alveolar clicks if it has dental clicks, these two features can be collapsed into a multi-state feature 'Does the language have dental and alveolar clicks?', with four possible answers: (1) Neither, (2) Just dental clicks, (3) Just alveolar clicks, and (4) Both. The transition rates between the four states can then be estimated, with the transition between (2) and (4) for example being more likely than a transition between (1) and (4). This is again left for future implementations of this model.

5.3 Convergence and Effective Sample Size

An important remaining issue with the analysis in this chapter is the small effective sample size (ESS) of the samples in the MCMC. Drummond et al. (2006) describe effective sample size as 'the number of independent samples that would be the equivalent to the autocorrelated samples produced by the MCMC. This provides a measure of whether the chain has been run for an adequate length (for example, if the effective sample sizes of all continuous parameters are greater than 200).' It is worth noting that the threshold of 200 is not given any justification in their paper. The BEAST 2.0 documentation[8] (which also suggests a lower threshold of 100) states that a low ESS will result in a poor estimate of a parameter, but does not explain why the threshold should be in the hundreds of samples, or what 'poor estimate' would mean: in other words, how accurate on average an estimate of a parameter would be with a small number of samples.

Lanfear et al. (2016) similarly describe the threshold of 200 as an 'arbitrary but pragmatic rule of thumb'. They say this in the context of making the point that auto-correlation among tree topologies is in fact even higher than the ESS of tree likelihood would suggest: 'Surprisingly, we detected autocorrelation [of tree topologies] in all of the empirical studies we analysed here (fig. 3*A*). This is despite these MCMCs having been run for 20,000,000 generations, with samples collected at large intervals of 10,000 generations—parameters that would usually be considered more than adequate for a Bayesian phylogenetic analysis. This is more notable because the number of taxa in these analyses was relatively small (from 9 to 61), and the trees were inferred with simple models of molecular evolution (a single GTR + G model applied to each data set).'

One might conclude from Lanfear et al.'s study that the aim of having a representative sample of trees from the posterior distribution is potentially quite unrealistic, especially for analyses with large numbers of taxa and complex substitution

[8] https://beast.community/ess_tutorial

models, and adopt a more modest aim of having a sample that gives some indication of the posterior distribution to within some degree of accuracy. A way of testing this in the case of the analysis conducted here would be to run simulations to see how accurate the consensus tree is compared with the real tree that was used to generate the simulated data. An informal way is simply to compare the result against some known information about the taxa, in this case, which language families the languages are known to belong to, and which regions they are found in. Given that the Bayesian tree in this chapter recovers some clusters of languages which are in the same geographical regions, and in some cases related languages, this suggests that the results are not inaccurate, let alone spurious, even if they are a 'poor' estimate of the true posterior distribution by the standards of a fully Bayesian analysis (as is possible with a small number of taxa).

The less stringent aim of getting to within some accuracy of the true tree (if we assume that there is a tree-like process generating the data) is shared by other clustering methods such as UPGMA. One could ask why a Bayesian approach is pursued in this chapter, rather than simply using one of these clustering methods, as is common for analysing large numbers of taxa (e.g. Simonsen et al. 2010, Jäger 2015, Jäger 2018). Section 4.2 provides the latter analysis as a baseline clustering of the languages in the World Phonotactics Database, in response to this potential objection, at the same time arguing that the UPGMA algorithm used is both theoretically inferior to using a Bayesian method, and also seems to reconstruct less meaningful clusters, so is therefore likely to be less accurate overall. There is the potential problem of spurious clades in the Bayesian analysis, perhaps because of the random initialisation (although removing samples from the burn-in should have mitigated this), or perhaps because of the MCMC chain falling into local optima. There may be a few cases of this, such as the placement of Scots Gaelic with languages in India in the Bayesian analysis but not the UPGMA analysis; but in other cases (such as a similar problem in the placement of Nynorsk), apparently spurious clades turn out to be potentially due to the genuine convergent evolution of phonotactic features.

Finally, if one criticises the Bayesian approach in this chapter on the grounds of the low effective sample size, this criticism is even more applicable to a simple clustering approach such as UPGMA, which does not explore the space of trees at all but instead heuristically finds a single tree (a non-randomly drawn effective sample size of one). Using MCMC to search through the space of trees with a more sophisticated substitution model is offered in this chapter at least as an improvement over this baseline clustering method.

6. Conclusion

The methods used in this chapter can be seen as a type of cluster analysis: languages which are similar to each other in their phonology are grouped together, whether the similarities are due to inheritance or contact. Several geographically coherent groups of languages emerge, which can be interpreted as linguistic areas because they cross-cut language families: a South China area; an Indian/Himalayan area; an eastern and southern Indian area; a Caucasus/Middle East area; an Indo-Turkic area; and so on. When clade constraints are added, there is strong support for linguistic areas such as Southeast Asia, and even a large macro-Eurasian area, in terms of the posterior probabilities assigned to the clades on the tree. This should be taken as a preliminary finding, however, given the low effective sample size in the analysis.

The main advantage of using a phylogenetic method is that it is hierarchical, finding clades within clades. It can be thought of as an automatic way of grouping languages in the way that linguists already do when they propose linguistic areas, but in a way that allows for sophisticated parameters such as different rates of change of features, and controlling for local instances of language contact and relatedness.

Some of the areas that emerge are reminiscent of linguistic areas identified by analyses such as Chirikba (2008) on the Caucasus, or Emeneau (1956) and Masica (1976) on India. In both of these cases, there is some indication that widespread multilingualism is the cause of structural convergence. For example, Chirikba (2008:8) describes the Caucasus as a region where different forms of bilingualism have been common, without a single dominating language before the introduction of Russian; he gives evidence for a linguistic area in this region using different types of linguistic data, including phonological properties such as ‘rich consonantism’, ‘ternary contrast of stops and affricates’, ‘glottalization’ and so on. This analysis contrasts with some previous analyses quoted by Chirikba such as Klimov (1978), who says that ‘it is very doubtful that the observed common Caucasian parallels are due to any sort of areal interaction’. The cluster analysis in this chapter contributes to this debate by showing both that there is areal clustering of phonological features (that is, that the commonalities are not due to independent development), and also that the clusters tend to be evidence of contact rather than shared inheritance (for instance, some Indo-European and Afro-Asiatic languages are also found in this cluster).

India is similarly described in Thomason (2000) as having clearly had different forms of language contact. A case study of the village Kupwar by Gumperz and Wilson (1971) shows one type of extreme grammatical convergence between Urdu, Marathi and

Kannada because of multilingualism, which Thomason speculates ‘might be a miniature reflection of the Sprachbund as a whole’ (p.10). A different type of contact also mentioned by Thomason is unidirectional substrate influence from Dravidian to Indic, suggesting language shift. Although it does not help to disentangle these different causes, the cluster analysis in this chapter does vindicate the notion (perhaps obvious, but often not supported in a quantitative way) that different language families in India share phonological properties due to contact, and that Indic languages primarily cluster with these other families as opposed to with their other Indo-European relatives.

Finally, some clades pick out commonalities between Southeast Asian languages, such as the area shown in Figure 3.7 comprising Tai-Kadai and Austroasiatic languages (among other groups). This is in agreement with surveys such as Enfield (2005:184), who describes a typical example of convergence between the unrelated languages Khmer (Austroasiatic, in the Mon-Khmer clade), Lao (Tai-Kadai) and Cham (Austronesian) in their vowel systems: these three languages are present in the clade in Figure 3.7. As Enfield writes, Tai-Kadai migrations influenced many populations in Southeast Asia, primarily by causing those populations to adopt Tai languages, and the expected outcome of this is for the languages of these groups (often Mon-Khmer or other Austroasiatic) to have influenced Tai-Kadai by substrate effects, so it is possible that the commonalities between languages in this clade are due to Austroasiatic substrate effects. Enfield argues this in particular for the case of vowel systems: 'That this close similarity is contact induced is clear when we consider Cham in the context of other Austronesian languages, whose vowel systems are normally much simpler than this, with ~4 vowels (Himmelmann & Adelaar 2005). Cham has undergone radical change under pressure from Mon-Khmer (Thurgood 1999). Similarly, Tai languages further north of Lao have simpler vowel systems.’ The reverse process of Austroasiatic languages adopting Tai-Kadai features such as tone is discussed in the next chapter.

Besides the interpretation of the phylogeny in this chapter as clusters corresponding to linguistic areas, it can also be read in a more literal way, as a history of the sounds of languages and the way that they have been transmitted. If two languages are placed together in a clade, then it means that they have ‘inherited’ their phonological properties from a common source. In some cases this could be because one language has influenced another by language contact, meaning that the common source in this case is an older version of one of the languages; e.g. if Basque was influenced by Spanish, then Basque and Spanish might be placed in a clade together, with the common ancestor in this case being an older version of Spanish.

One can even use this information to reconstruct the spread of features geographically, using a method such as phylogeography outlined in Chapter 2. A greater problem is inferring how old these areas are, and hence evaluating hypotheses linking the formation of linguistic areas with agricultural expansions. Some calibrations could be used such as the age of particular families or sub-families, helping to estimate the age of linguistic areas from the amount of structural diversity that accrues within them. Both projects would benefit from using a multiple topology model, rather than assuming a single topology, as described in section 5.1.

Despite the current limitations, the statistical support for the existence of these linguistic areas helps to make predictions about the movement of people and cultures which can be usefully compared with other demographic data. The following chapter is a case study on how that can be done, comparing the spread of linguistic variants by language contact with the movement of people across linguistic communities.

Chapter 4: Language Contact in Austroasiatic Languages of Southwest China

1. Summary

Mainland Southeast Asia is an example of a linguistic area, in that language families in that region share a set of distinctive properties, often due to language contact. Enfield (2005) reviews several examples of features which seem to have spread in this region by language contact, and Dahl (2008) shows that Southeast Asian languages are often typologically very similar, often as similar as closely related languages, despite being from different families. In Chapter 3 I also demonstrated this by showing that Southeast Asian languages often cluster together (although forming several different clusters rather than just one) according to their phonological properties.

Southeast Asian languages have other structural properties in common besides phonological properties, such as similar word orders, complex systems of numeral classifiers, and calques such as ‘eye of the day’ to mean ‘sun’. This chapter explores the way that these features have been spreading in Southeast Asia, and in particular asks whether the languages involved came to share these properties when people migrated across linguistic communities.

The approach taken in this chapter is to survey a small group of languages in southwest China to show how the convergence may have happened, in the East Palaungic branch of the Austroasiatic family. Despite being closely related and situated within a few hundred kilometres of each other, they vary in complex ways in their syntax and semantics.

Some of this variation may be due to contact with Tai-Kadai languages, making them a useful case study of language contact in action. Two hypotheses are put forward: (a) that Austroasiatic languages in general have been changing under influence of Tai-Kadai languages, of which the Palaungic group is an example; and (b) that this change is due to the presence of Tai-Kadai speakers in areas where Palaungic languages are spoken, with bilingual Palaungic speakers introducing Tai-Kadai features into their own languages. These points are argued for in section 3, by presenting each feature individually and summarising how it varies in the Palaungic languages, Austroasiatic, and the rest of Southeast Asia.

The amount of Tai-Kadai influence on each Palaungic language seems to be linked with the degree of Tai religious influence in the use of Buddhist temples and the erosion of Wa animistic culture, and above all to the presence of Tai speakers. This is tested by comparing the linguistic features with the proportion of Tai-Kadai speakers to Palaungic speakers in each region. The result is that subject-verb order and the presence of the calque ‘eye of the day’ in particular seem to be linked with greater ratios of Tai speakers in nearby towns.

2. The present study

I conducted fieldwork on thirteen closely related languages in the mountainous area in Yunnan province along the border between China and Myanmar, all in the East Palaungic branch of Austroasiatic. These thirteen varieties are dialects of approximately five languages in Glottolog (Hammarström et al. 2018): Zhenkang Wa [wbm], Ximeng-Menglian Awa [vwa], Blang [blr], Hu [huo] and U [uuu]. Nuclear Wa, Awa and Blang are in the Waic branch, while Hu and U are in the Angkuic branch. The language affiliation of each variety is based on my elicitation of basic vocabulary and comparison with Paul Sidwell’s database of Austroasiatic vocabulary (Sidwell 2015). In the remainder of this chapter, Zhenkang Wa is sometimes abbreviated to ‘Wa’, and Ximeng-Menglian Awa to ‘Awa’.

These languages were chosen because they are on the boundary of the southeast Asian area, as surveys of the geographical distribution of each linguistic feature in this chapter will show; on the one hand they are part of the Austroasiatic family, which has unusual values for these features especially further west near India; and on the other hand they are spoken near Tai-Kadai and Sino-Tibetan languages, making them a good case study in how languages can change by language contact. These languages turn out to vary in these features in a way that relates to their geographical location, the proportion of Tai-Kadai speakers in the region, and the presence of another Tai cultural feature, Mahayana Buddhism. This suggests that these features have been changing in these languages due to contact with Tai speakers.

The East Palaungic languages are a branch of Khasi-Palaungic, itself a primary branch of Austroasiatic (other primary branches including Aslian, Nicobaric, Mundaic, Vietic, Khmeric, Khmuic, etc.). The languages surveyed in this chapter are related in the following family tree:

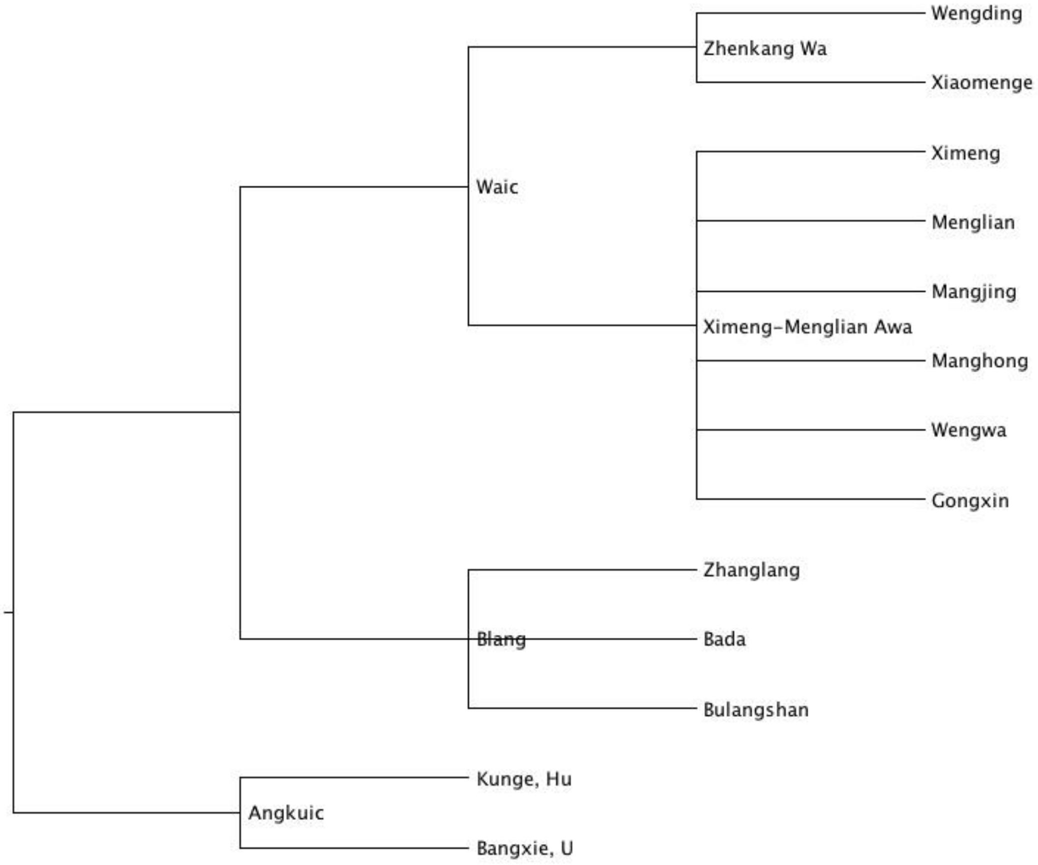


Figure 4.1: A family tree of the varieties surveyed in this chapter, using the tree from Glottolog. All branch lengths are set to 1 (in the absence of data on how long ago these languages diverged). Blang, Zhengkang Wa, Ximeng-Menglian Wa, Hu and U are names of languages, while Angkuic and Waic are names of sub-families. The remainder are place names.

Figure 4.2 plots these varieties, colouring them by branch. This map uses Google Earth to show their location in Southeast Asia, while subsequent maps in this chapter plot locations using the maps library in R (Brownrigg 2018). Figure 4.3 shows the locations and names of the varieties.

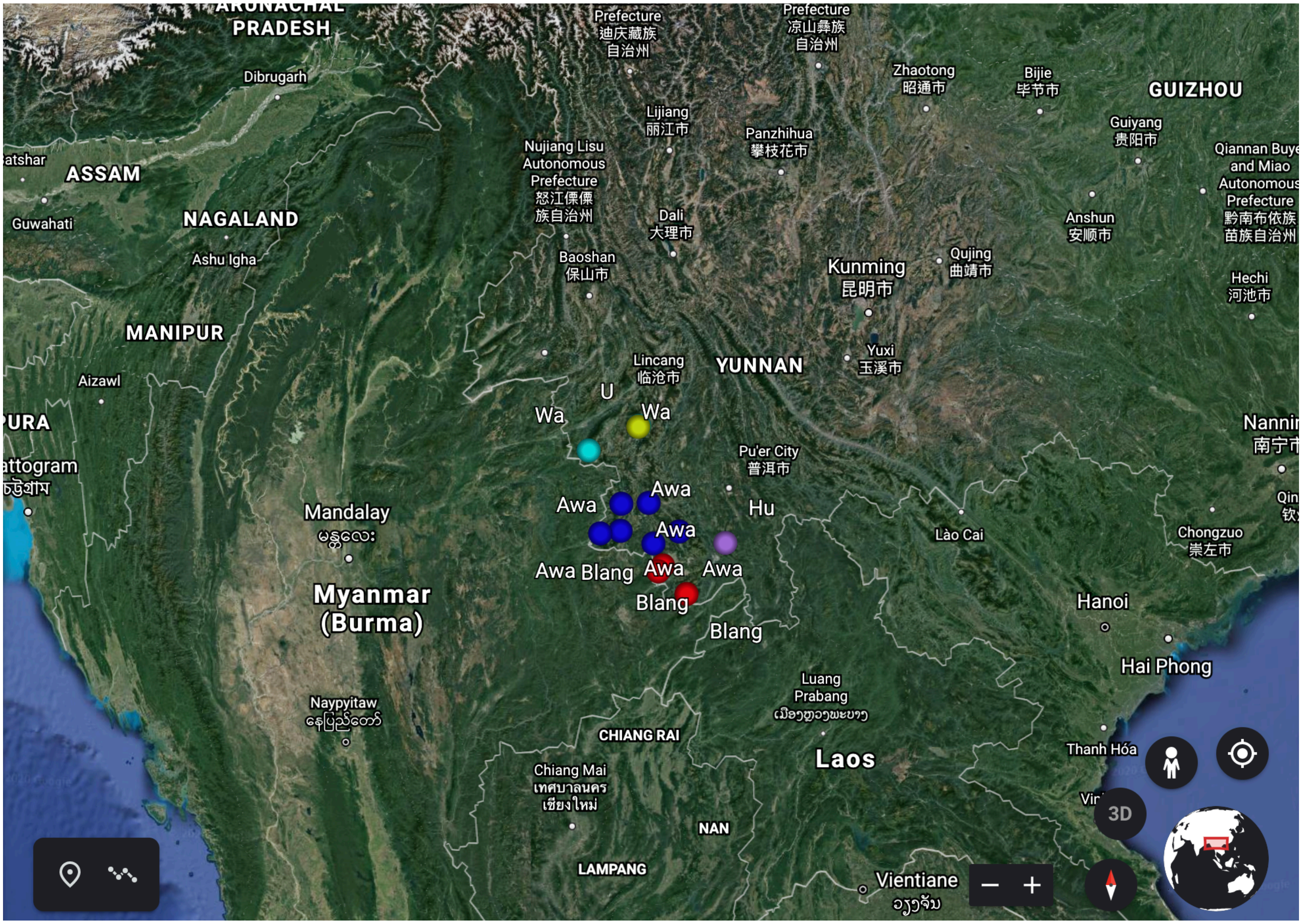


Figure 4.2: The varieties studied in the chapter, labelled by Glottolog language name (background map produced using Google Earth, http://www.earth.google.com, 2018): U (yellow), Wa (light blue), Awa (dark blue), Blang (red), and Hu (purple)

There are various other languages spoken in this region, including varieties of the Tai-Kadai language Tai Lü (henceforth referred to as Tai) and the Tibeto-Burman language Hani. Census data is available for major towns in the region, and shows that in some places in the east and far north Tai speakers are in a majority (Jinghong, Menghai, Dehong), while in the west near Myanmar it is speakers of Palaungic languages that are in the majority (primarily Awa), such as Ximeng and Wengding. While the locations where the varieties are spoken are in some cases some distance away from the towns (0-58km, with a mean of 33 km), the census data for the towns is taken as a proxy for the proportions of Tai and Palaungic speakers in the locations themselves: the distances are given in Table S1 in Appendix 1.

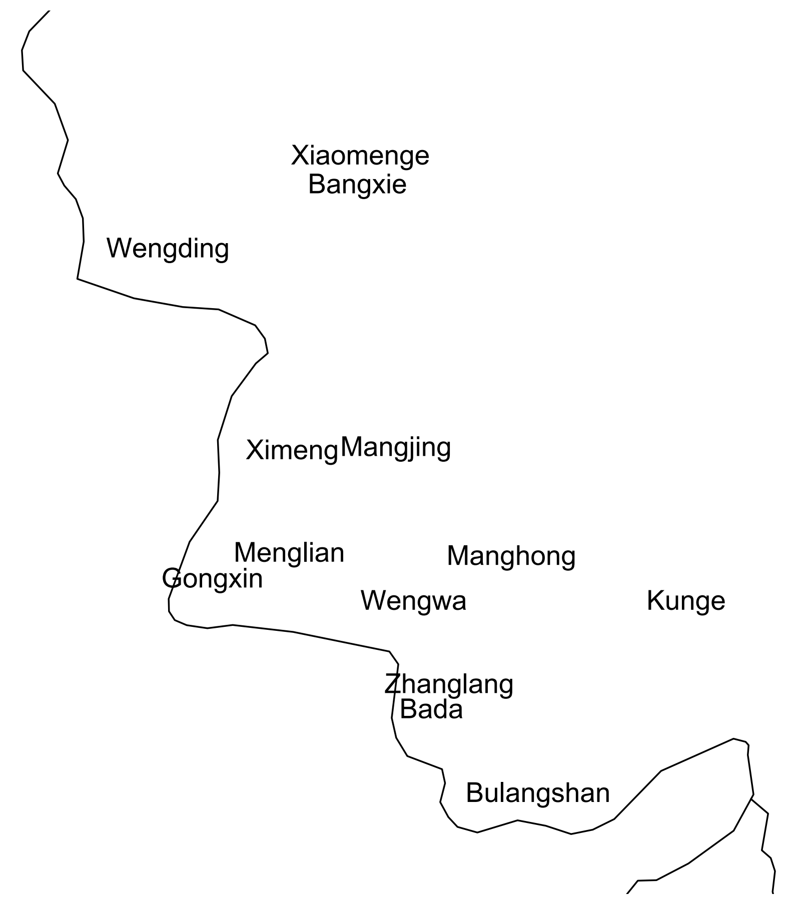


Figure 4.3: A map of the varieties surveyed in this chapter, with the border between China and Myanmar also shown.

The proportion of Tai speakers to speakers of the relevant East Palaungic language is shown on the map below as ratios. Red locations are where Tai speakers are in a majority (with ratios indicated), and blue locations are where they are in a minority.

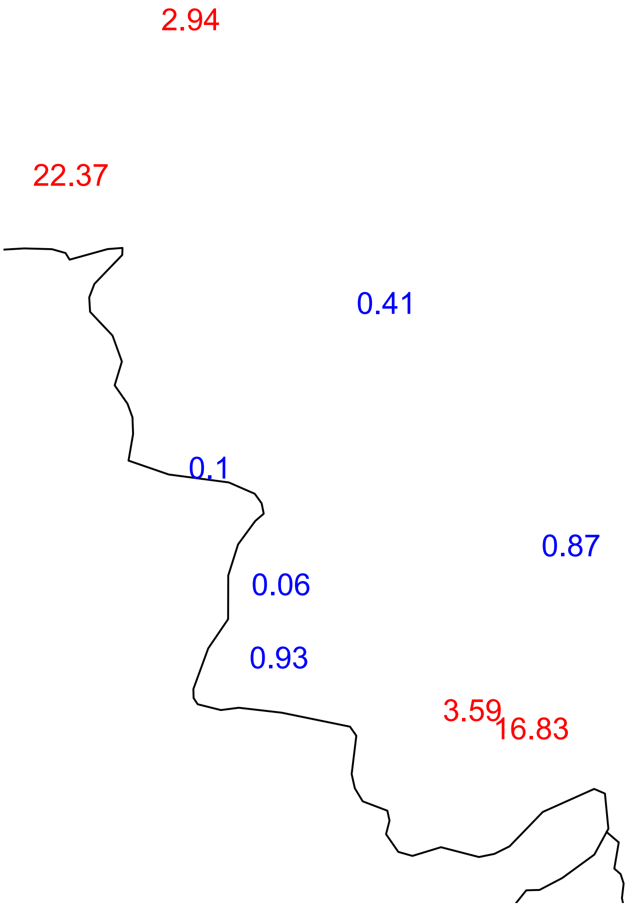


Figure 4.4: Ratio of Tai speakers to speakers of the relevant East Palaungic language in major towns, e.g. 22.37 means that there are 22.37 times as many Tai speakers as Palaungic speakers (see supplementary information, section 1.6, for census data). Red points are locations with a Tai majority (i.e. where the ratio is above 1), while blue points are locations with a Palaungic (Wa/Bulang) majority.

In line with the differing numbers of Tai speakers in each region, there are also striking cultural differences between these places, with Buddhist temples being mainly found further east in Tai-majority areas such as in Bada, Zhanglang and Mangjing; while traditional Wa symbols such as the bull skull is used on clothes and buildings further west, such as Ximeng, Wengding and Gongxin (from own observation).

The proportion of Tai speakers in each location is negatively associated with the use of bull skull imagery. This can be shown by assigning each location a demographic ratio of Tai speakers to speakers of Palaungic languages in the town closest to the location. The mean ratio is 0.5 in locations which use bull skull imagery (Ximeng, Wengding, Gongxin), and is 3.7 in locations without. A permutation test, permuting the values associated with each location, shows that this mean ratio would only be this low in 6% of random samples (taking 10,000 samples), giving a significance of p = 0.06.

This analysis is not comprehensive (there was no exhaustive search for bull skull imagery everywhere, and is just including particularly conspicuous examples), but it illustrates a way of quantifying the association between the presence of Tai speakers and different cultural or linguistic features that will be used for the remainder of this chapter. The sample size is small, and there is no control for dependencies by using a phylogeny; the quantification is simply to put a measure of confidence on how much these features are associated with Tai contact. A fuller quantitative test of the association between migration and linguistic features on a global scale will be outlined in Chapter 5, using mtDNA and the World Phonotactics Database.

The following sections show the results of eliciting linguistic features in the Palaungic languages, and tests for the relationship with the presence of Tai speakers, covering basic word order (3.1), the semantic calque ‘eye of the day’ (3.2), verbal semantics (3.3), the order of modifiers (3.4), and numeral classifiers (3.5). Each section discusses the variation in these features in southeast Asia, and then their variation within the East Palaungic languages. The method was as follows:

1. I visited the following towns: Kunge, Bulangshan, Zhanglang, Bada, Wengwa, Manghong, Mangjing, Menglian, Gongxin, Ximeng, Wengding, Bangxie, and Xiaomenge. Their locations are shown in Figure 4.3.

2. In each place I interviewed between 1 and three informants and recorded them with their consent; they gave consent to use elicited sentences in my research. The numbers of informants varied in each location (Kunge: 1, Bulangshan: 2, Zhanglang: 3, Bada: 3, Wengwa: 1, Manghong: 1, Mangjing: 1, Menglian: 2, Gongxin: 1, Ximeng: 2, Wengding: 2, Bangxie: 1, Xiaomenge: 1).

3. I asked for translations from Mandarin for the following phrases: a) ‘I eat rice’; b) ‘I eat/drink’ with other nouns (beef, fruit, banana, vegetables, congee, soup, alcohol, tea, water); c) ‘one book’, and ‘one’ + other nouns (e.g. dog, tree, flower, chopstick, etc.: the complete list is in the Github repository under classifier_list.txt); d) ‘this/that man’, to elicit demonstrative word order; a selection of basic vocabulary, in order to help classify the language using Glottolog and Sidwell (2015); and the words ‘sun’ and ‘rainbow’ in particular, to investigate semantic calques.

4. The results were then compared with demographic data, gathered from online sources given in Appendix 1. Each variety was assigned an approximate ratio of Tai speakers to Palaungic speakers based on the data available for the nearest town, which was on average 33 km away from the location in question (with a range of 0-58 km; distances shown in Table S1 in Appendix 1). A typical comparison is of the ratio of Tai speakers to Palaungic speakers, with particular linguistic features; for instance, whether this is ratio higher for languages which have the expression 'eye of the day' for 'sun'. In this case the ratio of Tai speakers is high (5.59:1) in languages with 'eye of the day' and low (0.88:1) in languages without.

3. Features surveyed in East Palaungic languages

3.1. Order of Verb, Subject and Object

One of the defining features of the Southeast Asian linguistic area is the use of Subject-Verb-Object (SVO) word order (Enfield 2005, Gil 2013). This is shown in red on the map below from Dryer (2014). Five language families in mainland Southeast Asia and parts of Indonesia use this word order but are surrounded by languages of the rest of Eurasia which tend to use Subject-Object-Verb (SOV) order shown in blue. To the east are the Austronesian languages which tend to use Verb-Subject-Object (VSO) or Verb-Object-Subject (VOS) order, shown in yellow. These word orders cross language family boundaries, suggesting that they have been spreading partially by language contact. For example, Sino-Tibetan and Austroasiatic languages near India use SOV word order, while in China and Southeast Asia they are more likely to use SVO word order. The fact that several neighbouring language families in Southeast Asia use SVO word order also suggests that there has been a spread of this word order through these families, although it may be independent development or inheritance from a common ancestor further back in time as well.

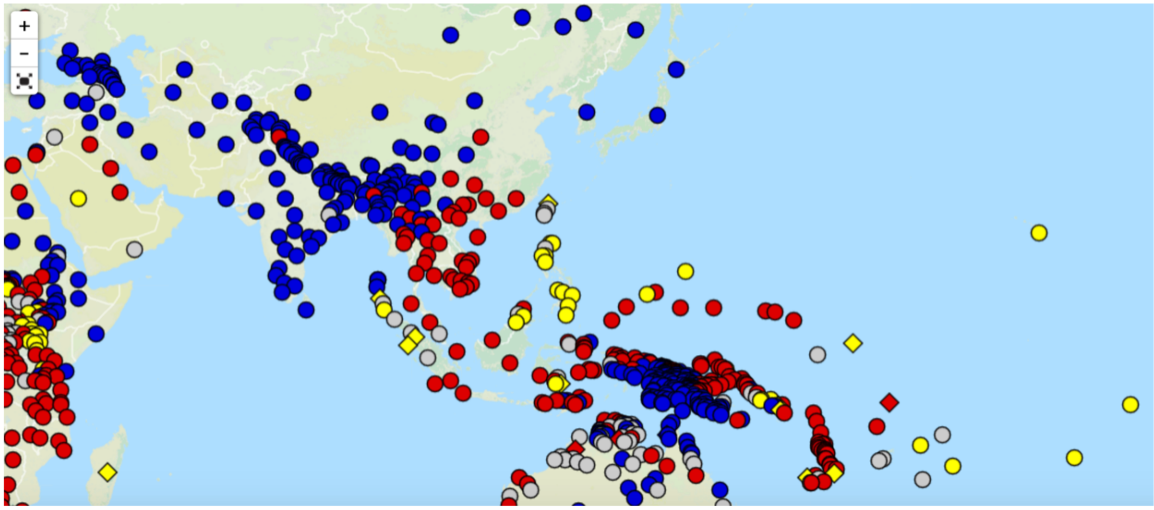


Figure 4.5: Map from the World Atlas of Language Structures (Dryer 2013) showing the order of subject, verb and object in Southeast Asia (blue = subject-object-verb, red = subject-verb-object, yellow circle = verb-subject-object, yellow diamond = verb-object-subject, grey = mixed word orders).

East Palaungic languages are unusual (even unique, judging from the map of languages from WALS) in mainland Southeast Asia and eastern Eurasia as a whole in using VSO word order. However, there are some interesting indications in the rest of Austroasiatic that this word order may once have been more widespread. Nicobarese, an Austroasiatic language of the Nicobar Islands, uses VSO order (Dryer 2013), and some languages in the Aslian of the same family have been described as having some verb-initiality (Kruspe 2004). Most relevant is Khasi, an Austroasiatic language in northeastern India that is most closely related to the Palaungic languages, according to Glottolog (Hammarström et al. 2018) and Sidwell (2011): this language also displays VSO order in some clause types (Rabel 1961).

Austroasiatic languages to the west near India therefore show some signs of verb-subject order, while further east they have subject-verb order. One hypothesis is that subject-verb order has been spreading in Austroasiatic due to the influence of Tai, Sinitic and other mainland Southeast Asian languages, while VSO order may have been much more widespread in the past in Austroasiatic languages.

Palaungic languages in particular are an ideal case-study in the spread of SVO word order, as they are split in their use of VSO order and SVO word order. I elicited the intransitive sentence ‘I eat rice’ (‘eat rice’ is a single morpheme, e.g. *sɔm*) to illustrate variation in the order of verb and subject. Languages which use SV order are shown in red on the map below. Languages which use VS order are shown in blue, and one language which uses both order simultaneously, SVS, is shown in purple. A transcription of ‘I eat.rice’ is shown for each variety as an example, although the same word order also applies to transitive sentences elicited (such as ‘I eat beef’).

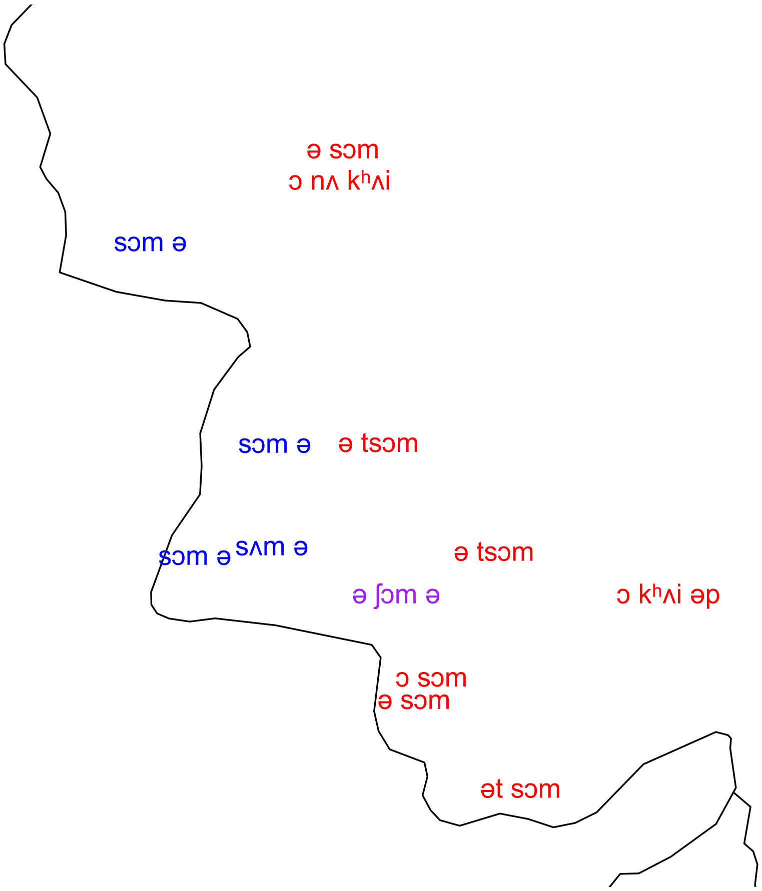


Figure 4.6: Verb-subject order in Palaungic languages (red = verb-subject, blue = subject-verb, purple = subject-verb-subject).

Palaungic languages are divided between those which use verb-subject order, shown in blue on the map, and those that use subject-verb order shown in red. One language, Wengwa shown in purple, uses both at the same time (ə ʃɔm ə ‘I eat I’), as if there is a person agreement system.

The fact that there is a neat geographical split in the use of these word orders, and furthermore that the variety that uses both simultaneously is found in between them, suggests that subject-verb order has been spreading in these varieties by contact with Tai languages.

This can be shown by comparing the distribution of these word orders with the proportion of Tai speakers in each region. Languages which use subject-verb order have a higher mean proportion of Tai speakers, 4.01:1, than the other languages, which have a mean ratio of 0.5:1 ($p = 0.04$ in a permutation test). If Wengwa is included, which uses subject-verb-subject order, then the ratios are 4.13:1 and 0.51:1 ($p = 0.01$). This provides support for the hypothesis that the innovation of subject-verb order is associated with the presence of Tai speakers.

3.2 Eye of the day

Another example of a linguistic feature which is reliably associated with the presence of Tai speakers is the expression ‘eye of the day’ used to mean ‘sun’. This is common in Southeast Asia, such as the Thai *ta wan* and the Indonesian *mata-hari* (both ‘eye day’). Urban (2010) showed that this semantic pattern is cross-linguistically rare, not being found outside of Southeast Asia and the Pacific, where Austronesian languages have brought it. Given that it is widespread in Malayo-Polynesian languages, it may have been present in Malaysia about 3000 years ago (Gray, Drummond and Greenhill 2009), and Urban suggests that it was transmitted into Malayo-Polynesian from mainland Southeast Asian languages.

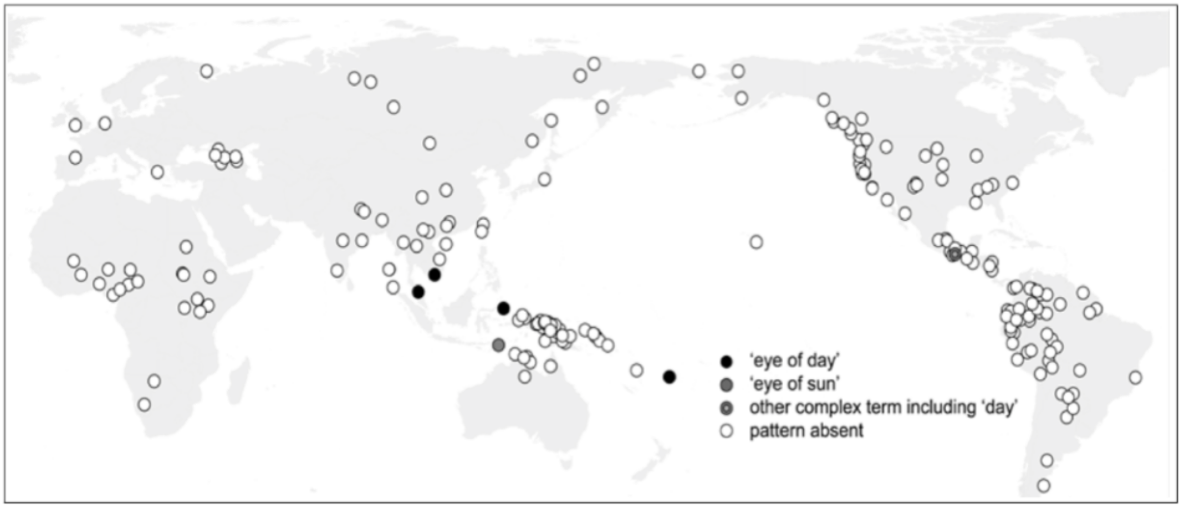


Figure 4.7: The global sample for Matthias Urban's study on 'eye of the day' (Urban 2010): key given in the diagram.

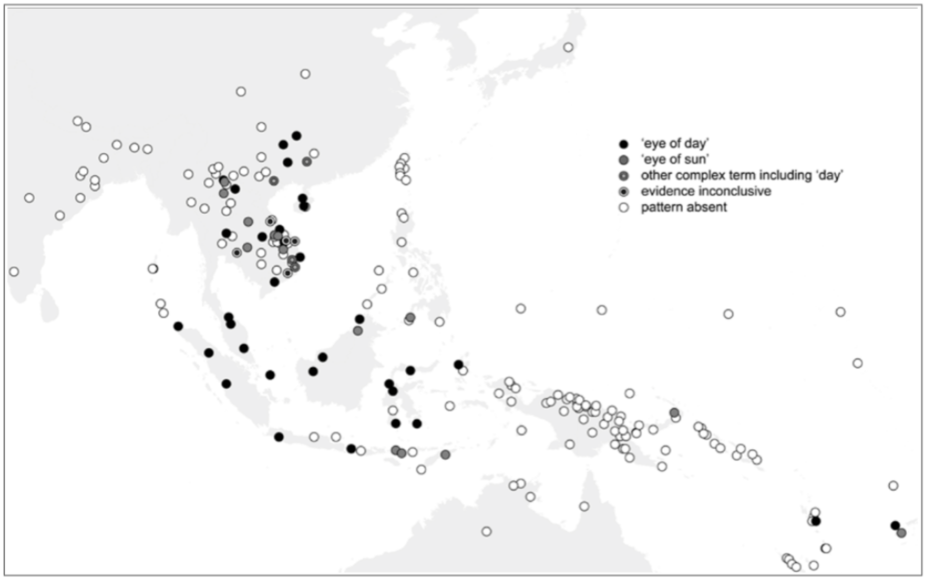


Figure 4.8: Urban's (2010) map of 'eye of the day' in southeast Asia and Oceania: key given in the diagram.

However, it is also very prone to being borrowed, for example appearing in several Papuan languages that have had contact with Austronesian. It is a calque found in Tai-Kadai languages in particular, while Austro-Asiatic languages are more divided in whether they have it.

Wa varieties again turn out to be a good illustration of the way that the calque can spread by language contact. There is a split in whether they have this feature, and this feature is principally found in the southern varieties close to Tai in Jinghong. Tai also has this expression, *ta wan* 'eye [of the] day'.

The Wa varieties are split between those which have sŋi or a cognate such as *hŋi*, *ŋi* or *ŋʌi*, meaning 'sun' or 'day', and those which have *ŋʌi sŋi* 'eye of the day'. One complication is that *ŋʌi* 'eye' itself may be etymologically related to 'sun' *sŋi*, perhaps by some derivational process of adding *s-* as a prefix. Languages with the calque 'eye of the day' are shown in red, while languages with a monomorphemic word for 'sun' are shown in blue.

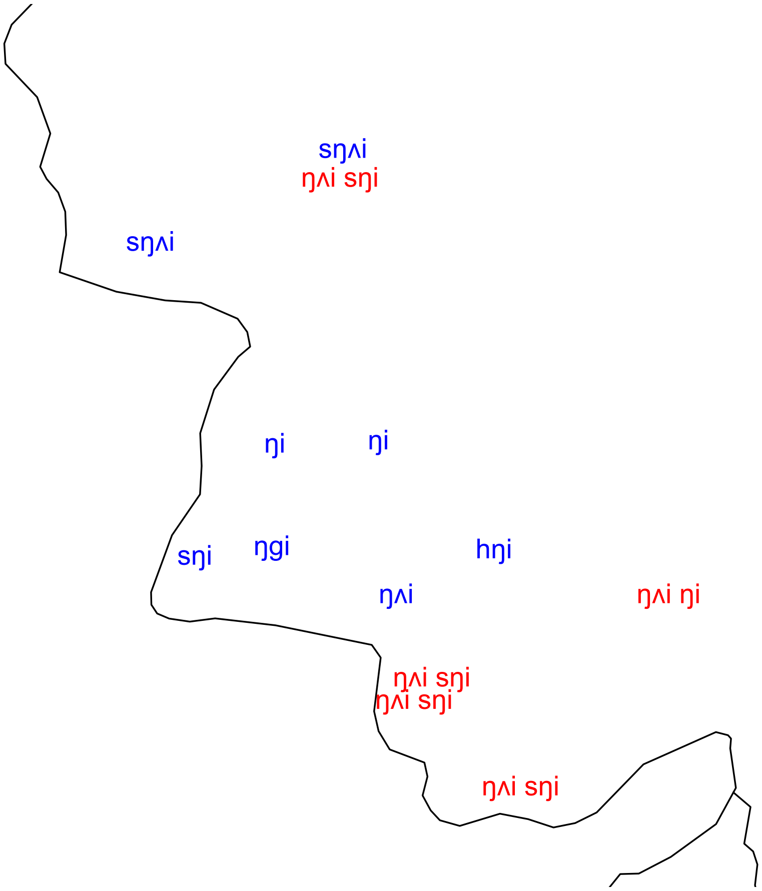


Figure 4.9: 'Eye of the day' in the Palaungic languages (red = present, blue = absent).

In agreement with the hypothesis that the calque is borrowed from Tai, the mean ratio of Tai speakers to Palaungic speakers is 5.59:1, whereas in languages without the ratio is 0.88:1 (p = 0.01 in a permutation test). This shows that a calque can be an especially good marker of language contact, as others have noted for areas such as Meso-America (Campbell et al. 1986), and the relationship with the demography of Tai speakers also suggests that it is a marker of demically induced contact in particular.

Perhaps another example is the word for 'rainbow', which in some Palaungic languages is 'drink water', *niə rɔm*, and was explained by informants to refer to a dragon in the sky drinking water. It resembles the expression for 'rainbow' in Thai, รุ้งกินน้ำ *roong gin naam* 'rainbow drinking water'; and also the word for 'rainbow' in the Tibeto-Burman language Naxi *mɯ ɕi dʑi t'ɯ* ‘sky tongue drink water’ (Zhao 1995:114). This was not plotted on a map because informants in most locations did not know the word for ‘rainbow’, but illustrates how calques may be especially informative about language contact. The following section explores a different type of semantic pattern, the use of basic verbs such as ‘eat’ and ‘drink’.

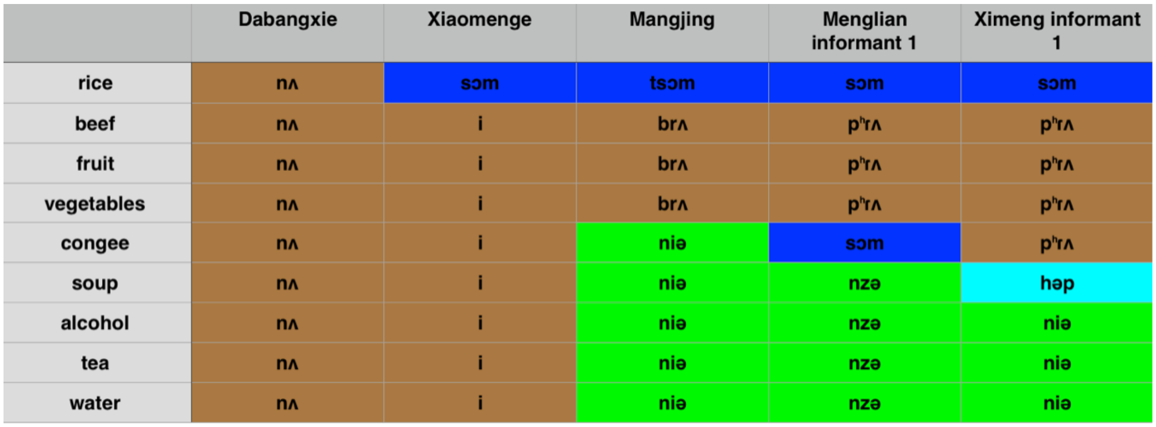
3.3 Semantics of ingestion verbs

Semantic typology is relatively understudied compared with syntactic typology, perhaps because there is a large number of grammatical descriptions of languages that tend to include information such as word order, but not detailed descriptions on how for example verbs are used. Work by Majid, Boster and Bowerman (2008) pioneered the use of standardised elicitation tasks to find out the semantic ranges of ‘cut’ and ‘break’, and showed that languages can partition the semantic space of even quite concrete verbs very differently. Verbs of eating and drinking are one example, as Newman (2009) explores with examples from different languages.

A common system in Southeast Asia is to allow a single verb to cover both eating and drinking, such as in Thai and other Tai-Kadai languages (Rzymski and Tresoldi 2019), some Sinitic languages in Zhejiang and Fujian (ibid.), and some Austroasiatic languages such as Minor Mlabri (Rischel 1995), suggesting that it has been spreading by language contact. Other Austroasiatic languages by contrast often have verbs denoting specific types of eating, such as the Aslian languages which make quite specific and obligatory distinctions such as between eating vegetables and eating meat (Kruspe 2004).

The variation is especially complicated in Wa varieties, for which I elicited ingestion verbs used with nine nouns: rice, beef, fruit, vegetables, congee, soup, alcohol, tea and water. These verbs were elicited by asking in Mandarin how ‘eat/drink’ was used with these different direct objects.

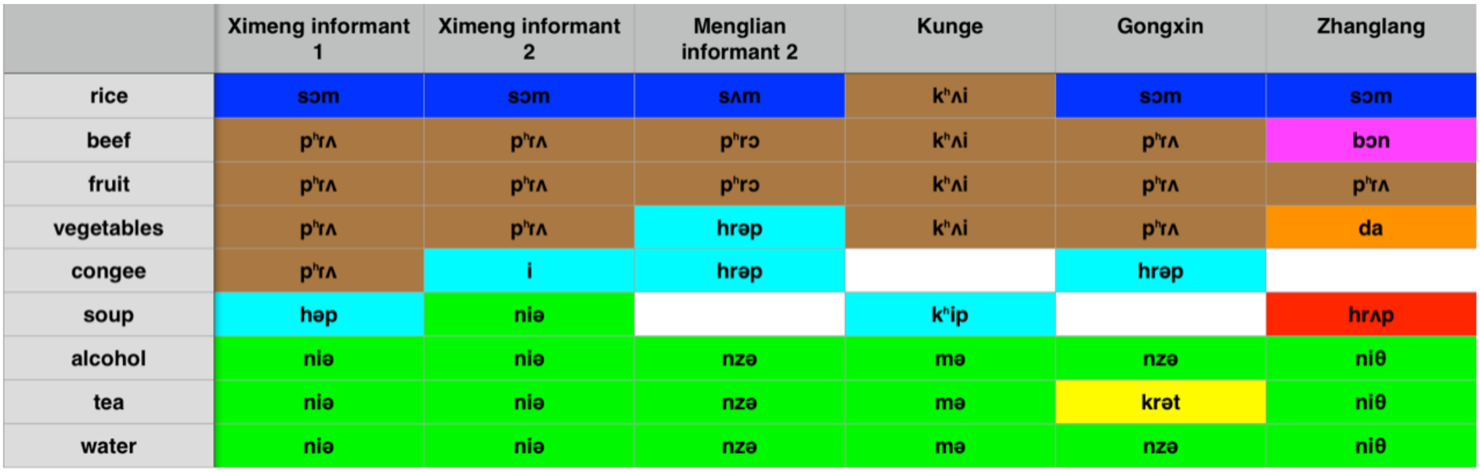
The simplest system is if there is a single verb that is used with all of these, as in the language U in Dabangxie. This can be represented schematically in Table 4.1, which colours all of the nouns in the first column brown, showing that they take the same verb. In another variety in Ximeng according to one informant, there was a distinction between ‘eat’ (for solids such as beef, fruit, vegetables and congee), ‘drink’ (for liquids such as alcohol, tea and water) and a separate verb for ‘eat rice’ and another verb specifically meaning ‘drink soup’. This is represented in the second column by colouring these nouns with four separate colours according to the verb that they take. The third column shows a different informant from Ximeng, who used a specific verb for ‘eat congee’ and a different boundary for ‘eat’ and ‘drink’ from that of the first informant.



	Dabangxie	Xiaomenge	Mangjing	Menglian informant 1	Ximeng informant 1
rice	nʌ	sɔm	tsɔm	sɔm	sɔm
beef	nʌ	i	brʌ	pʰrʌ	pʰrʌ
fruit	nʌ	i	brʌ	pʰrʌ	pʰrʌ
vegetables	nʌ	i	brʌ	pʰrʌ	pʰrʌ
congee	nʌ	i	niə	sɔm	pʰrʌ
soup	nʌ	i	niə	nzə	həp
alcohol	nʌ	i	niə	nzə	niə
tea	nʌ	i	niə	nzə	niə
water	nʌ	i	niə	nzə	niə

Table 4.1: Simple systems of partitioning ingestion verbs in Palaungic languages. The colours are to help visualise the partitions between different verbs.

There are many further systems of dividing up the semantic space of these verbs. Some examples of these are given in Table 4.2 and are plotted on the map in Figure 4.10. White cells are for missing data, in many cases because the informants did not know how to translate the names of particular types of food such as congee.



	Ximeng informant 1	Ximeng informant 2	Menglian informant 2	Kunge	Gongxin	Zhanglang
rice	sɔm	sɔm	sʌm	kʰʌi	sɔm	sɔm
beef	pʰrʌ	pʰrʌ	pʰrɔ	kʰʌi	pʰrʌ	bɔn
fruit	pʰrʌ	pʰrʌ	pʰrɔ	kʰʌi	pʰrʌ	pʰrʌ
vegetables	pʰrʌ	pʰrʌ	hrəp	kʰʌi	pʰrʌ	da
congee	pʰrʌ	i	hrəp		hrəp	
soup	həp	niə		kʰip		hrʌp
alcohol	niə	niə	nzə	mə	nzə	niθ
tea	niə	niə	nzə	mə	krət	niθ
water	niə	niə	nzə	mə	nzə	niθ

Table 4.2: More complex partitions of ingestion verbs in Palaungic languages, with specific verbs for ‘drink tea’ (yellow, in Gongxin) or ‘eat meat’ (pink, in Zhanglang).

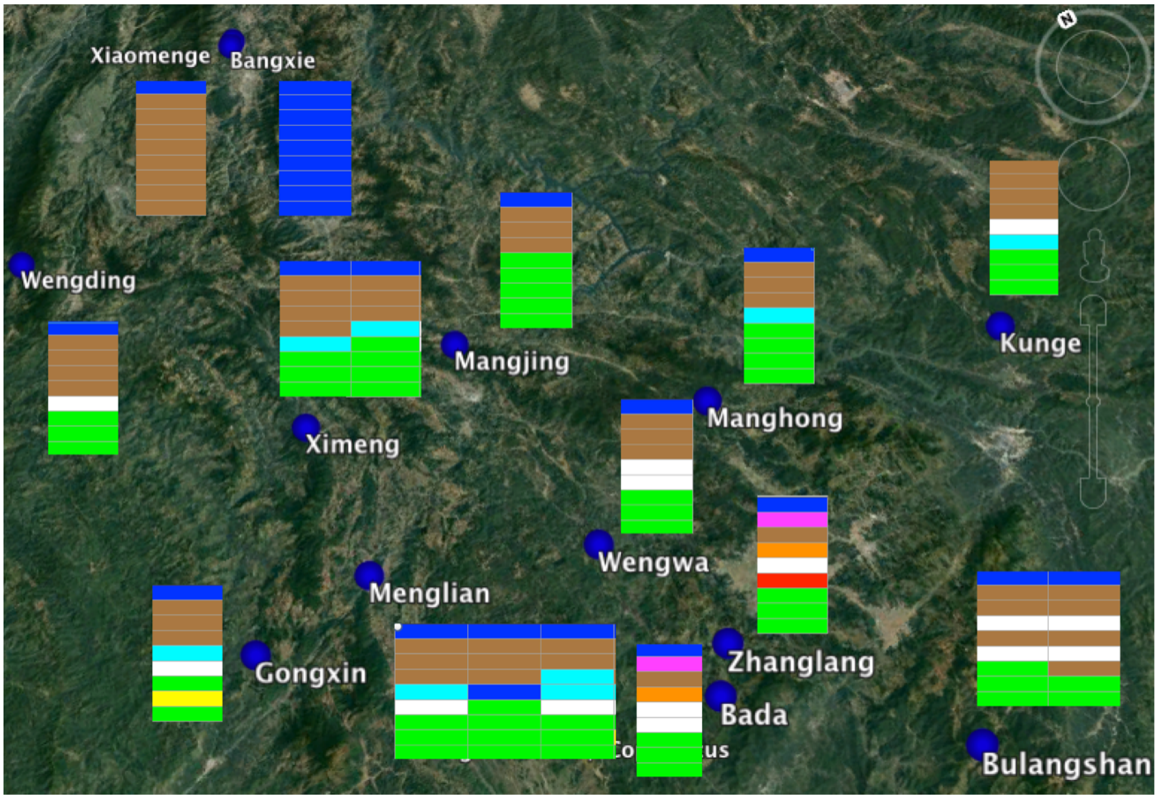
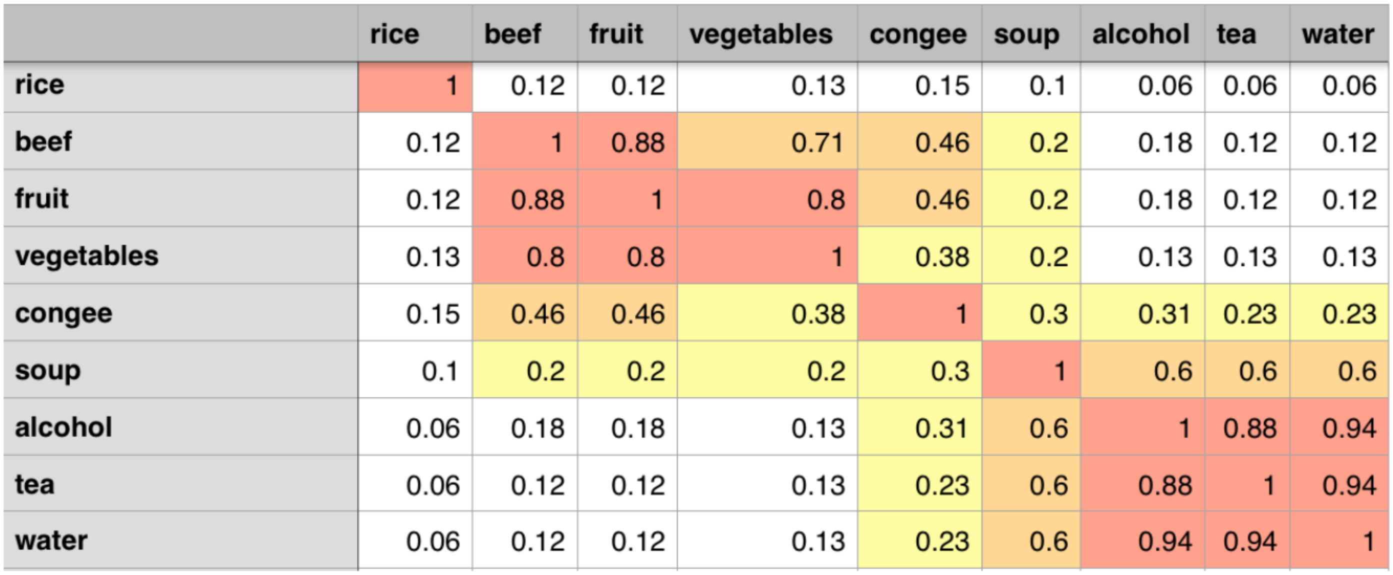


Figure 4.10: The geographical distribution of different ingestion verb systems (background map produced using Google Earth, (http://www.earth.google.com, 2018).

Despite the diversity of these systems, there are clear recurring tendencies such as solids being grouped together, but also for some food stuffs such as rice, congee and vegetables to be treated as on the borderline between solid and liquid. Table 4.3 summarises the probability of two nouns taking the same ingestion verb across the different systems elicited.



	rice	beef	fruit	vegetables	congee	soup	alcohol	tea	water
rice	1	0.12	0.12	0.13	0.15	0.1	0.06	0.06	0.06
beef	0.12	1	0.88	0.71	0.46	0.2	0.18	0.12	0.12
fruit	0.12	0.88	1	0.8	0.46	0.2	0.18	0.12	0.12
vegetables	0.13	0.8	0.8	1	0.38	0.2	0.13	0.13	0.13
congee	0.15	0.46	0.46	0.38	1	0.3	0.31	0.23	0.23
soup	0.1	0.2	0.2	0.2	0.3	1	0.6	0.6	0.6
alcohol	0.06	0.18	0.18	0.13	0.31	0.6	1	0.88	0.94
tea	0.06	0.12	0.12	0.13	0.23	0.6	0.88	1	0.94
water	0.06	0.12	0.12	0.13	0.23	0.6	0.94	0.94	1

Table 4.3: The probability of a verb for one noun being used with another noun (e.g. ‘eat fruit’ has a probability of 0.46 of being used with ‘eat congee’). Red squares show higher values, while lighter (e.g. yellow and white) show lower values.

The word for 'eat' *k*^{h}*ʌi* in Hu is the same as the word for 'rice' in the other Angkuic language U. This may reflect a tendency in Austroasiatic languages to derive verbs from nouns, such as the verb ‘drink’ from the noun ‘water’ in the Khmuic language Mlabri (Rischel 1995). We also see this in Zhanglang, where ‘I eat meat’ is expressed as ‘I meat meat’, using the noun *bɔn* as both a noun and a verb:

(1) *ə*   *bɔn*   *bɔn.mui*

1sg meat. meat.cow

‘I eat beef’.

The use of verbs for specific kinds of eating is found in other parts of Austroasiatic such as the Aslian language Jehai (Burenhult 2005), and Mlabri (Rischel 1995), suggesting that it might be quite an old feature of the family. The simplification of these systems in some languages, such as the Angkuic languages Hu and U, may have been due to contact with Tai (which like Thai has a single verb covering 'eat' and ‘drink’).

However, contrary to this hypothesis, these two languages are associated with locations with Tai minorities, and so there is a positive relationship between the number of ingestion verbs used and the proportion of Tai speakers. Locations with a Tai majority in fact use 4.2 verbs on average, and locations with a Palaungic majority use 3.1 verbs on average. The specificity of ingestion verbs is therefore not reliably associated with Tai contact, or at least using this particular sample and method.

3.4 Modifier word order

This section returns to a different type of word order variation, the order of modifiers such as adjectives, numerals and demonstratives. In languages in Southeast Asia, there is a general tendency for modifiers to be placed after the noun, especially in Tai-Kadai languages, while Austroasiatic languages are again more mixed. An example of this geographical distribution is shown in the map from WALS below in Figure 4.11 (Dryer 2013), showing the order of demonstrative and noun.

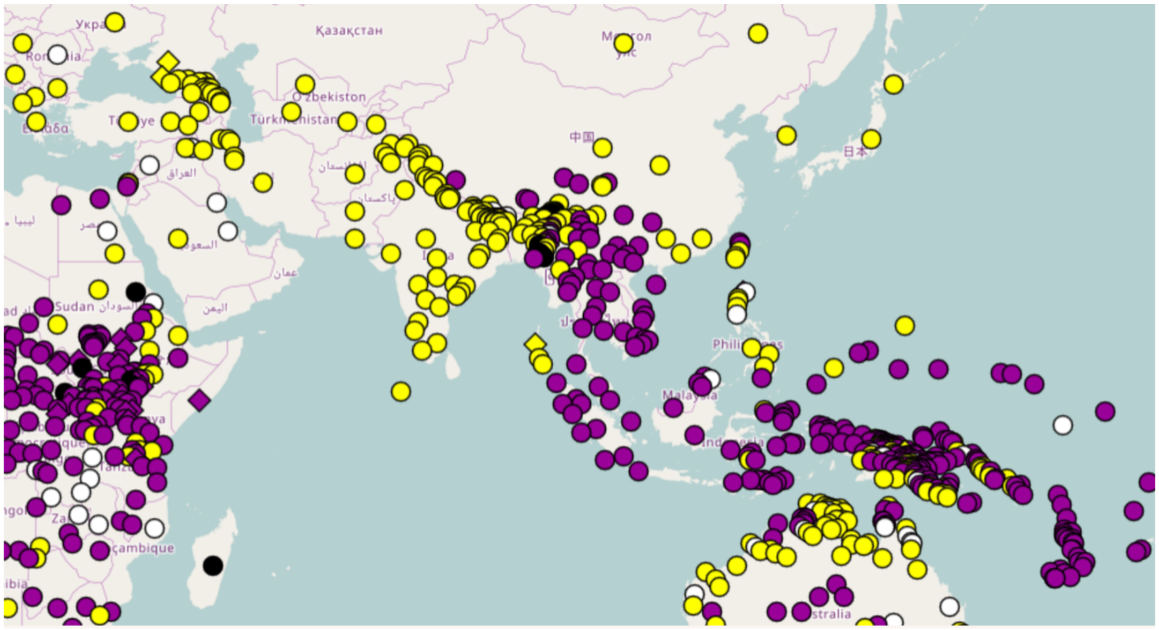


Figure 4.11: The ordering of demonstrative and noun in languages of southeast Asia and surrounding areas (purple = noun-demonstrative, yellow = demonstrative-noun, black = demonstrative simultaneously before and after noun, white = no dominant order).

The East Palaungic languages all have noun-adjective order, and generally also use noun-numeral order (except when a numeral classifier is not used, as discussed in section 3.5). The main variation is in demonstrative-noun ordering, where varieties that use noun-demonstrative order are again found in regions with a higher ratio of Tai to Palaungic speakers, suggesting that they have been influenced by noun-demonstrative order in Tai. Languages with noun-demonstrative order have a mean ratio of 3.77:1, while languages with demonstrative-noun order have a mean ratio of 1.53:1; this difference does not reach conventional significance in a permutation test (p = 0.19), however.

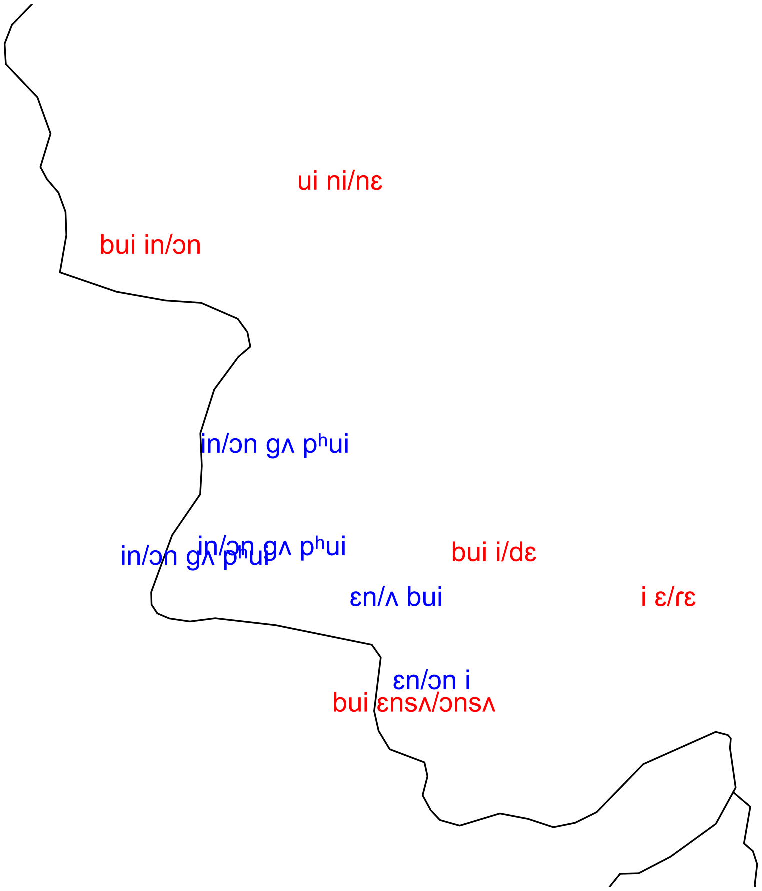


Figure 4.12: The ordering of demonstrative and noun in East Palaungic languages, using the sentence ‘this/that man’ as an example (blue = demonstrative noun, red = noun-demonstrative).

3.5 Numeral classifiers

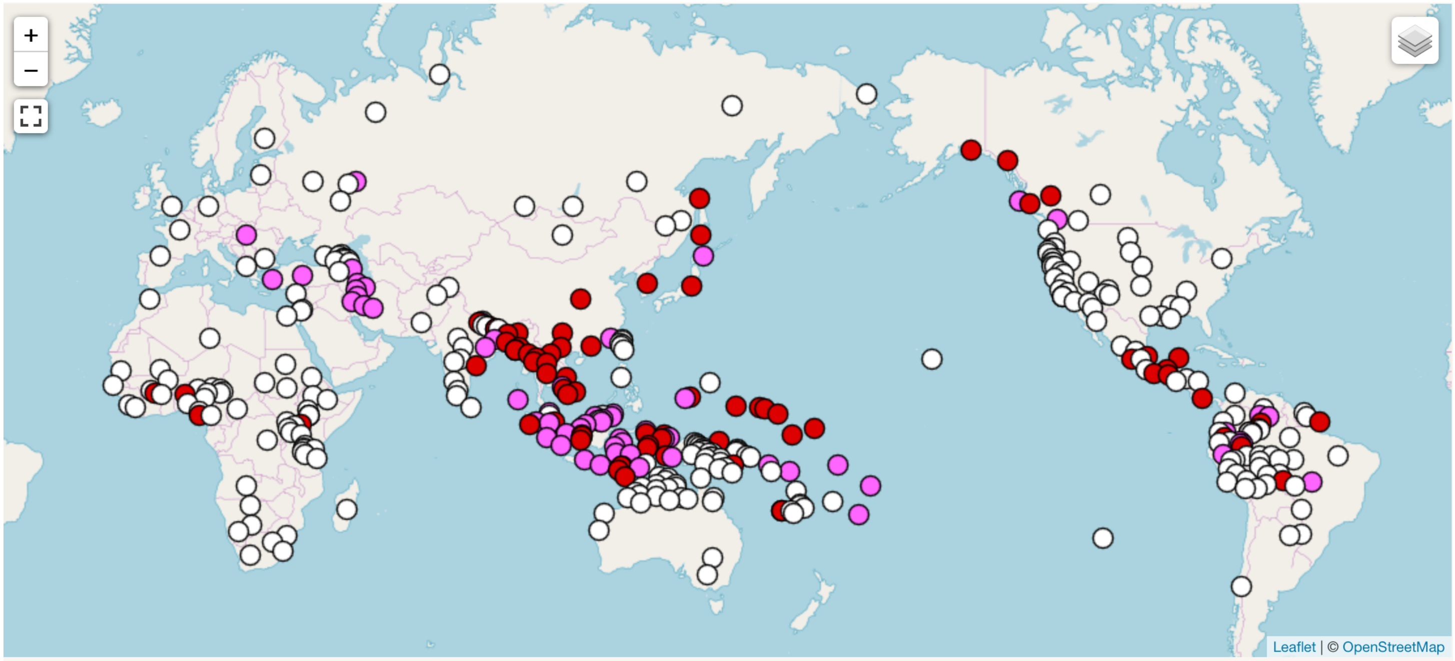


Figure 4.13: A map of languages which have numeral classifiers (Gil 2013) (red = obligatory classifiers, pink = optional classifiers, white = absent)

Numeral classifiers are another prototypical feature of southeast Asian languages, which extends into Austronesian languages and even into languages in west Papua, as Figure 4.13 from Gil (2013) shows.

East Palaungic languages tend to use numeral classifiers, but vary in the complexity of the classifier systems. A set of about 73 nouns were used to elicit phrases of the form ‘one person’, ‘one dog’ and so on (the full list and data are provided in Appendix 1), designed to find classifiers for specific categories such as animals, vehicles, long or flat objects, plants, and the like.

The number of classifiers used in this elicitation ranged from 10 to 24, plotted on the map below in Figure 4.13 (with the three languages in red being the languages with the most classifiers). There is only one language in this sample spoken in a place with a Tai majority (3.59:1), Zhanglang, which also happens to be the language with the most classifiers, 24. The remaining non-Tai-majority languages have a mean of 13.7 classifiers. The sample is too small to conclude very much (the significance in a permutation test is p = 0.14), but suggests that the number of classifiers may be associated with the presence of Tai speakers. Tai itself as spoken in Jinghong uses 27 classifiers for the same task.

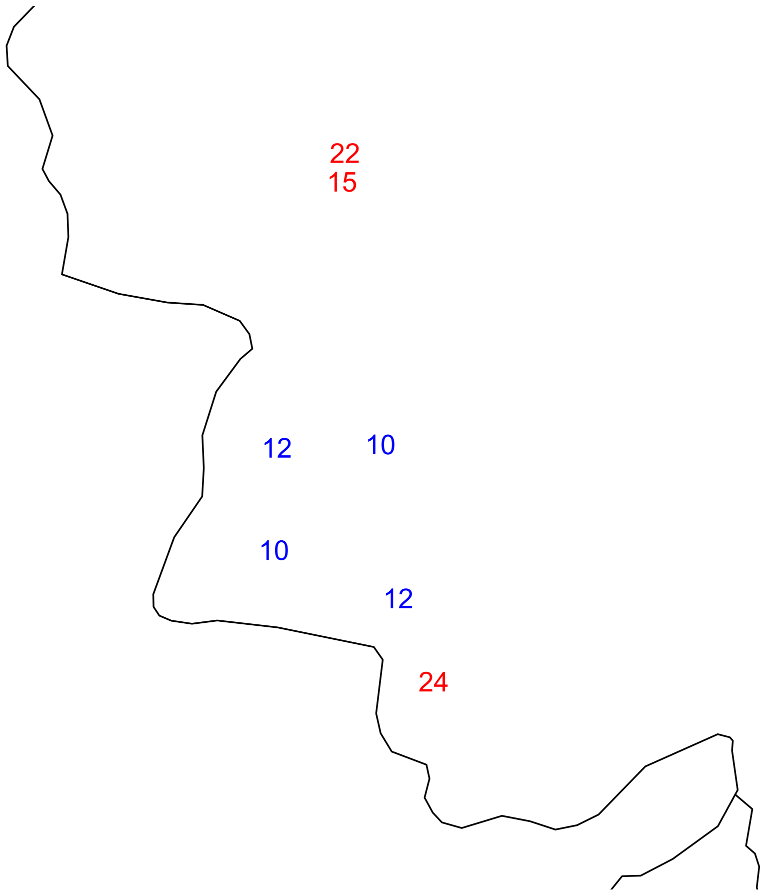


Figure 4.14: The number of numeral classifiers in each of the seven sampled languages (with the three highest numbers shown in red).

An interesting additional grammaticalization pathway was observed in the Angkuic language U in Bangxie. The numeral ‘one’ in this language (and its relative Hu) seems to be derived from the general numeral classifier mu in other languages, while the other languages in the sample use *dɛ*. This affects the ordering of the numeral and the classifier in U: if a classifier is used which is not *mu*, such as *do* used for classifying animals, then the order is noun-numeral-classifier:

(1) *so mu do*

dog one classifier

‘one dog’

If the classifier used is mu itself, then the construction is the noun followed by ʌmu:

(2) *mɔk ʌmu*

hat one.classifer

‘one hat’

The grammaticalization pathway of numeral classifier > ‘one’ is unusual but is found in some Sinitic languages of southeast China, such as in Shaowu (Ngai 2015), and so may be a more widespread tendency in languages of Southern China more generally. Like ‘eye of the day’, it is a possible example of a semantic calque that indicates contact between languages, pending further data on its geographical distribution.

3.6 Numerals

The clearest example of Tai contact is the borrowing of Tai numerals. Several languages in this survey use Waic numerals, at least below ten, although they may also use Tai numerals to express numerals to express larger quantities (include ‘thirty’).

However, several varieties use Tai numerals for at least some numbers below ten. The two Angkuic languages Hu and U are examples, Hu expressing all numbers with Tai numerals, and U expressing numbers above three using Tai numerals. The Blang variety in Bada also uses entirely Tai numerals. The Blang variety in Zhanglang uses Tai numerals above five (despite the system being decimal, as they all are), according to one informant; another informant used entirely Blang numerals; while the other Blang variety in Bulangshan does not use Tai numerals below ten. The Wa and Awa varieties that were sampled do not use Tai numerals below ten (Ximeng, Wengwa, Wengding, Menglian, and Gongxin).

This is clearly a case of Tai contact, and accordingly there is a relationship with the ratio of Tai speakers to Palaungic: varieties which use Tai numerals below ten have an average ratio of 6.1:1, while varieties which do not have an average ratio of 1.1:1 (p = 0.032).

4. Conclusion

This chapter reviewed syntactic and semantic variation in East Palaungic languages, and hypothesised that this variation is due to contact with Tai-Kadai languages.

Hu, U, and Blang dialects seem to be especially influenced by Tai-Kadai languages since they use Tai numerals, and have subject-verb order and the calque ‘eye of the day’. Wa and Awa varieties by contrast all lack Tai numerals and ‘eye of the day’, and are more likely to use verb-subject word order (1/2 and 3/6 dialects). There is some statistical support for the association of these features with proportion of Tai speakers, although the sample is generally too small, and does not attempt to control for non-independence of data points.

To demonstrate these points statistically would require a large quantitative test, perhaps on Austroasiatic languages more generally, and controlling for phylogenetic dependencies. There is also some need for correcting for multiple testing, as several features were chosen and tested in this chapter, and specifically because of their suggestive variation in Southeast Asia. An illustration of this comes from the fact that several features thought to be likely to be influenced by Tai turn out have only non-significant relationship with the presence of Tai speakers (number of classifiers, and demonstrative-noun order), or no relationship at all (number of ingestion verbs). These relationships are even weaker when aggregated by languages, such as number of classifiers (U 15, Wa 22, Awa 11, Blang 24), number of eat verbs (Hu 3 and U 1, Wa 2.5, Awa 4, Blang 4.7), noun-demonstrative order (found in Hu, U and Wa, and mixed in Awa and Blang), none of which show the pattern of Tai influence in the previous paragraph. A larger set of syntactic and semantic features is therefore needed to test the claim of contact with Tai more systematically.

This is therefore a case study which illustrates the demic hypothesis examined in this thesis, that linguistic structures often spread through languages by migration of speakers to other linguistic communities. The Palaungic languages are a microcosm of changes that happened in Southeast Asia, such as the spread of subject-verb-object word order, numeral classifiers, and semantic calques, which might similarly be explained on a larger scale as spreading with people, perhaps speakers of Tai-Kadai languages in particular, which are thought to have originated in Southern China and spread from there to the rest of Southeast Asia (Jenks and Pittayaporn 2017).

The speakers of the languages surveyed in this chapter are not usually bilingual between Palaungic and Tai-Kadai languages, and are also not aware of the properties that languages such as Tai have. The Tai-like properties are therefore not necessarily due to current bilingualism in these villages, but suggest a situation in the past where speakers of some Palaungic languages were using Tai languages as well, for example in Myanmar in Tawngpeng state where intermarriage between Palaung and Tai people is common (Simms 2017:177). This contrasts with an example of Austroasiatic contact with Tai-Kadai given in the previous chapter, in which there was convergence in vowel systems (Enfield 2005:184) and arguably in other phonological properties (given that these languages were in a Southeast Asian clade together in the phylogenetic analysis in the previous chapter) due to Austroasiatic substrate influence.

It is possible that factors such as the spread of Buddhism played a role in bringing the Tai language and Tai linguistic features to Waic communities (Patterson Giersch 2006:22), which may not have needed very much migration of speakers; however, in this particular case, the presence of Buddhism is also well predicted by the proportion of Tai speakers. The following chapter explores the demic hypothesis further in a study of mtDNA data, by comparing evidence for movement between particular locations and similarity in linguistic structures.

The study in this chapter also shows the intrinsic interest of features which are elicited in fieldwork rather than through grammatical descriptions, such as the semantics of ingestion verbs, the number of numeral classifiers, and semantic calques. Elicitation of semantic features, as well as phonological and structural features, for language varieties on a local scale is a promising way in future of providing a more detailed picture of language history.

Chapter 5: Linguistic Areas and Mitochondrial DNA Lineages

1. Summary

Chapter 3 identified geographical clusters in which languages tend to share phonological properties, and Chapter 4 suggested that one way that linguistic structures spread is by movement of people. It is possible that these linguistic areas are formed partly because of migration of people, for example of speakers out of southern China to the rest of Southeast Asia (Jenks and Pittayaporn 2017, Enfield 2005). These migrations of people could bring the descendants of a language (such as the spread of Tai-Kadai languages), or could cause situations of language contact where Tai-Kadai structures are introduced into other languages, as was suggested in the previous chapter and has been documented elsewhere (see Chapter 1 for the review of the literature on linguistic areas). This relationship between the movement of people and the movement of languages or linguistic structures can be addressed by using genetic data to reconstruct migration patterns in the past, and comparing it with linguistic data.

This chapter uses mitochondrial DNA data from GenBank (Benson et al. 2013) and Bayesian phylogeography to reconstruct common migration routes. A set of 2000 mtDNA sequences from 423 locations around the world was analysed using a spatial model implemented in BEAST by Bouckaert (2016). This chapter may be the first model-based reconstruction of mtDNA migration routes, despite the large literature on analysing the geographical distribution of mtDNA haplogroups (e.g. Harcourt 2016), and despite the use of Bayesian phylogeography for other domains (such as the transmission of viruses, Lemey et al. 2009, Magee et al. 2017; and the spread of languages, Bouckaert et al. 2012).

The main result is a dataset of the movements of mitochondrial DNA lineages in various parts of the world, including Africa, Eurasia, Oceania, and the Americas. Particular routes taken by these lineages, sometimes over large distances, may help to explain the shape of certain large linguistic areas. There are also significant differences in this data in the overall direction that people have tended to move in different regions, such as whether people have tended to move horizontally (east/west) such as in most of central Eurasia, or vertically (north/south) such as in East Asia, Africa and Western Europe.

The relationship between migration and languages is then examined, by modelling how they are correlated with each other and with geographical distance. The relationship between geographical distance and linguistic distance is first modelled, by using the results of the phylogenetic analysis of Eurasian languages in Chapter 3. The number of mtDNA migrations between linguistic communities is then included in the model, to find out whether they increase its predictive accuracy. It is found that there is a large increase in accuracy once migrations are included, in the case of predicting linguistic similarity from the clade-constrained phylogenetic analysis; but that there is no increase in accuracy in predicting linguistic similarity from the non-constrained phylogenetic analysis, which turns out to be very well correlated with geographical distance. The main conclusion of this is that, from one measure of linguistic similarity (constrained distances), linguistic areas have been shaped in part by migrations and not just by sharing of features by linguistic communities in close proximity. The analysis therefore offers modest support for the demic hypothesis explored in this thesis, although it is based on a selective line of evidence (mtDNA data only, and a measure of linguistic similarity using only phonology/ phonotactics) and should be extended to other types of linguistic and genetic data.

2. The present study: overview and relationship with previous studies

2.1 The 'demic' hypothesis of the spread of language families and linguistic structures

Chapter 3 described how languages share common phonological properties across language families. The clade-constrained analysis of the World Phonotactics Database showed that languages in Eurasia seem to fall into two main clusters. The first cluster is Southeast Asia, which has been suggested to be a linguistic area by linguists (e.g. Enfield 2005). The second cluster is the rest of Eurasia, covering Europe, most of Siberia, India, the Middle East and North Africa. This cluster will be referred to as 'Eurasia' for short for the rest of the chapter. In the non-clade-constrained analysis, Southeast Asian languages fall into several different clusters, with perhaps the clade in Figure 3.7 in Chapter 3 being the closest to what has been described in Enfield (2005).

Why did these areas form? One reason may be because of concerted migration of people across language boundaries, and in particular if people have expanded out of a certain area. Southeast Asia may be an example of this, as certain language families such as Tai-Kadai spread out of southern China to the rest of Southeast Asia (Jenks and Pittayaporn 2017), and likely influenced the phonology of languages that they came into contact with while doing so.

This may be viewed as a hypothesis about the 'demic' spread of phonological features, namely that these linguistic areas were formed by movement of people. 'Movement' here specifically means settlement, in order for it to leave a genetic trace. Of course, language contact happens regularly by other means, such as the movement of traders, and modern communications allow language contact to happen without people moving at all. It is therefore also possible that phonological features can spread between neighbouring languages without there necessarily being concerted movement of people. This is often argued by analogy to have happened with agriculture, which can be borrowed between people without there being a genetic signal of agriculturalists moving (Zvelebil 2000), although in practice there is also genetic support for the demic model in the case of agriculture, such as in Chikhi et al. (2002).

De Filippo et al. (2012) is an example of a linguistic study testing similar hypotheses for the expansion of Bantu languages in Africa, namely whether these languages spread by demic or cultural diffusion. They find that populations speaking Bantu languages are more genetically similar to each other than they are to populations speaking non-Bantu languages, after controlling for geographical distance using partial Mantel tests. They conclude that a demic diffusion model is supported in this case: ‘Our comparison of the genetic distances among Bantu populations with those of Bantu versus all other linguistic and ethnic groups…indicates that even geographically distant Bantu-speaking populations are closely related to each other, as expected with demic diffusion, and argues against a major role for language shift in the Bantu expansion.’

The aim of this chapter is similarly to ask whether linguistic areas are formed by demic or cultural diffusion, although the implementation is quite different from de Filippo et al. (2012), for instance in using ‘migrations’ (see section 3.1 for details) inferred from a mtDNA phylogeny as a measure of genetic distance, rather than F_{ST}.

2.2 Studies comparing linguistic structures and genetic data

Mitochondrial DNA (mtDNA) is inherited from the mother only, and so mtDNA sequences form a family tree showing the way people are related to each other along the maternal line; put another way, mtDNA sequences allow us to construct a global family tree of women.

There is a large literature on investigating the history of mitochondrial DNA and how long ago particular migration events may have happened (Hasegawa et al. 1985, Kivisild 2015). There are also surveys of the geographical distributions of mitochondrial DNA clades (haplogroups) and proposals of the way that these can be used to reconstruct history in different regions (e.g. Kayser et al. 2008 on mtDNA diversity in Melanesia). There is not work using model-based phylogeographic reconstruction of mtDNA haplogroups of the type implemented in software such as BEAST, developed by Lemey et al. (2009) for viruses and since applied to the spread of Indo-European languages by Bouckaert et al. (2012).

One aim of this chapter is therefore to conduct such a phylogeographic analysis, to provide a visualisation of mtDNA lineages that would allow some comparison with the spread of languages. Although the results in this analysis are preliminary, the kind of migrations that can be discovered by analysing mtDNA in fact do seem to carry meaning: they are a proxy for population movements which shaped how language families moved, and how languages in different families have come to share phonological properties.

This is a quantitative version of what has been said more informally in much of the literature on mtDNA, which pinpoints particular haplogroups and suggests that their movement may be associated with particular language families, such as Austronesian (Kayser et al. 2008) and Austroasiatic (Zhang et al. 2015). Bellwood (2013) suggests many such links between genetic variants and language families. In non-academic discussion, the links between language families and particular haplogroups can become quite sweeping, such as on the website Eupedia, which has titles for posts such as 'Haplogroups of Bronze-Age Proto-Indo-Europeans'[9].

Other papers link the movements of language families to other types of genetic evidence, using ancient DNA (e.g. Haak et al. 2015 on the potential spread of some Indo-European languages from an origin in the steppe; and Malaspinas et al. 2016 on the genomic history of aboriginal Australians). Y chromosome DNA is also particularly interesting because of the suggestion by Forster and Renfrew (2011) that language families may sometimes be carried into new regions by a small number of males, rather than large-scale migration (for instance, they suggest 'It may be that during colonization episodes by emigrating agriculturalists, men generally outnumbered women in the pioneer colonizing groups and took wives from the local community.') Other localised studies of the relationship between different types of genetic data and situations of language contact and the spread of language families are described by Pakendorf (2014), one example being de Filippo et al. (2012) on the spread of Bantu languages mentioned in section 2.1.

Despite the assumption that there is a relationship between the distribution of genetic variants and the movement of language families, these observations are often non-quantitative, and a solid demonstration of this relationship has been surprisingly elusive.

[9] https://www.eupedia.com/genetics/haplogroups_of_bronze_age_proto-indo-europeans.shtml

Perhaps the first quantitative attempt was by Cavalli-Sforza, Minch and Mountain (1992), which studied 38 global populations. This study compared a tree of these populations using genetic data (based on Cavalli-Sforza et al. 1988), and a tree of these populations based on putative language families, including some controversial members such as Altaic and even macro-families such as Eurasiatic, and finds a highly significant correlation between the two. There was no control for geographic distance, however, which makes the relationship between the two trees somewhat trivial: languages which are closer together tend to more closely related, and human populations which are closer together tend to be more closely related, but these two facts by themselves are not enough to conclude that populations and languages have a shared history beyond that fact.

Dediu (2007) is a more sophisticated study of the relationship between genes and languages, employing a spatial statistic approach. He uses a partial Mantel test to test the correlation between genetic distance between a set of populations, geographic distance, linguistic genealogical distance, and typological linguistic distance using features from the World Atlas of Language Structures. The Mantel test is a way to compute the correlation coefficient between two matrices, by finding the correlation coefficient of the elements of the two matrices; a partial Mantel test computes the correlation between two matrices while controlling for the effects of a third (Dediu 2007:239, Mantel 1967).

Interestingly, once geographic distance is controlled for, he finds that the correlation between genetic distance and linguistic genealogical distance is small ($r = 0.1041$, $p = 0.0308$; Dediu 2007:275), and the correlation between genetic distance and typological distance is non-significant (p.244). He notes that this is in agreement with previous studies which had tested the relationship between genetic distance and linguistic distance using partial Mantel tests, several of which find that the relationship disappears once geographic distance is controlled for (p.183, citing among others Sokal et al. 1992 and Rosser et al. 2000).

Another paper which investigated typological distance and genetic distance is Creanza et al. (2015), which uses phoneme inventories. They again find that the relationship between genetic distance and linguistic distance is non-significant after controlling for geographic distance. A more local study is on indigenous populations in Taiwan by Brown et al. (2014), who find a relationship between linguistic distance and genetic distance which again becomes non-significant ($p=0.07$) once they controlled for geographic distance.

By contrast, Longobardi et al. (2015) is one study that finds a high correlation between genes and linguistic features, even once geographic distance is controlled for, but it should be noted that the sample is restricted to just fifteen languages, all in Europe. They find that the correlation between genetic distance and linguistic distance is high (0.49-0.51), and higher than the correlation between genetic distance and geographic distance (0.38). This result may be due to the fact that they include Finnish and Basque people, who are genetically quite differentiated from the other European populations in their sample, and this matches well with the fact that Finnish and Basque are linguistically distinct from the other populations which speak Indo-European languages. This is therefore a confirmation of the idea that linguistic and genetics can be mutually informative, but is probably specifically about the differentiation of Basque and Finnish from Indo-European languages.

An additional recent quantitative study on the correlation between genes and linguistic features, as well as music, is Matsumae et al. (2021) on populations of northeast Asia. They study 14 populations encompassing 11 different families or isolates, for which genetic, linguistic and musical data were available. They compared genetic distance between populations with different types of linguistic distance: lexical (based on data from the Automated Similarity Judgement Project (ASJP) database, Wichmann et al. 2016), grammatical (from WALS (Dryer et al. 2013) and AUTOTYP (Bickel et al. 2017)), and phonological (the World Phonotactics Database (Donohue et al. 2013) and PHOIBLE (Moran et al. 2014)). To calculate genetic distance they used 37,093 SNPs from 245 individuals (including their newly genotyped 15 Nivkh individuals), taking pairwise F_{ST} between populations, the proportion of the total genetic variance due to between-population differences. To compare these distance measures they use Redundancy Analysis (RDA), an alternative to the Mantel test which performs a regression of multiple response variables on multiple predictor variables. According to their summary, RDA yields an adjusted coefficient of determination (adjusted R^2), which captures the variation in the response that can be explained by the predictors, that is then compared with adjusted R^2 values under random permutations. They find correlations between genetic distance and grammatical distance in particular (R^2=0.54), which is significant after adding controls for geographical distance and genealogical distances in a partial RDA. Other correlations such as between lexicon and genetic distance (or with music) are not significant after adding these controls.

Matsumae et al. suggest that the correlation may be due to relationships between languages dating to before recent contact and inheritance within known families. They base this conclusion on the fact that the correlation survives controls for geographical and genealogical distance. There is an alternative possibility, which is that genetic distance is a better proxy for language contact than geographical distance is. If two populations have had migration between them, then it is also more likely that some type of language contact may have taken place, presumably raising the probability above what you would expect from knowing the geographical distance between the two languages (since there may often be neighbouring languages which do not systematically have contact). To the extent that this explanation may be true, this is indirect support for the 'demic hypothesis' pursued in this chapter, that contact-induced structural change is mainly caused by migration of people.

This is borne out by inspecting the clusters that emerge in their visualisations of genetic and grammatical distance using neighbournet networks produced with SplitsTree (Huson and Bryant 2006). Some clusters in common include Selkup and Nhanasan (both Uralic languages), Chukchi and Koryak (both Chukotko-Kamchatkan), and Even and Evenki (both Tungusic). The correlation between genetics and grammar seems to be mostly due to these known families, but also due to the clustering of Japanese, Korean, Nivkh, and Ainu in both trees. While it is possible that these languages are related, it is also possible that there has been simultaneously a lot of interbreeding between these populations, as well as language contact.

2.3 Phylogeography using whole genome data

The quantitative studies above raise some interesting points of comparison with the analysis employed in this chapter. First, the analysis in this chapter uses reconstructed mtDNA migrations, rather than genetic distance between populations defined by SNPs. This is clearly a deficiency of the current analysis, as mtDNA offers only partial information compared with variants across the whole genome. Furthermore, only a small number of mtDNA lineages survive, meaning that the picture of migrations that surviving mtDNA diversity offers is quite limited; for example, over 40% of Europeans have mtDNA haplogroup H, making them the descendants of a woman on the maternal line 22,000 years ago (Brotherton et al. 2013). An additional problem with the analysis in this chapter is that the mtDNA samples from each population are not randomly selected, introducing another type of bias. Perhaps the one advantage of using mtDNA is that it is transmitted from a single parent, and hence it is easy to use to reconstruct a phylogeny of people, and to use that to reconstruct a complex sequence of migrations. It is also an opportunity to quantify the type of statements that are made in the literature on mtDNA, by using a model-based approach to reconstructing these migrations, and comparing them with linguistic distances; but one can and should use whole genome data to model the history of populations (e.g. Gravel et al. 2013 on the history of Native American migrations).

One study in 2022 demonstrates how whole genomes can be used to build phylogenies which can then be used for reconstructing migrations (Wohns et al. 2022). They use their previously published package *tsinfer* (Kelleher et al. 2019) which differs from Bayesian phylogenetics in many ways, employing simpler heuristics that arguably make it somewhat less accurate, but also allowing inference of multiple trees that can be scaled to analyse millions of genomes.

The *tsinfer* method aims to produce a 'succinct tree sequence' from genome data, which they describe as a sequence of marginal trees, each encoding the genealogy for a particular segment of DNA. Adjacent trees moving along a chromosome tend to be highly correlated, so shared edges are stored only once. The way that they infer these correlated trees is to assume that each site has mutated only once, and that it is possible to infer which state is ancestral and which is derived. The algorithm starts at a particular site and moves both leftwards and rightwards from it, estimating the tree topology and plausible values for the ancestral haplotype. They use various heuristics to decide when to stop reconstructing the tree for a particular haplotype and to begin a new tree.

Once they have trees for particular haplotypes, they use a simple method of taking midpoints of modern locations as a way to reconstruct ancestral locations. They mention the locations of the root of the different ancestral haplotypes in Africa (and also surprisingly in Papua New Guinea ~140,000 years ago, potentially consistent with the time-depth of Denisovan lineages found in Papuans). However, their paper does not explore in detail patterns of migration in the more recent past, so is not immediately comparable in its findings with the study in this chapter. Their dataset of reconstructed migrations from autosomal data may be useful in future for comparison with languages, building on the type of comparison developed in this chapter between mtDNA migrations and structural data.

The remainder of this chapter is structured as follows. Section 3 presents a summary of the phylogeographic method, and broad patterns found when this is applied to studying mtDNA sequences. Section 4 presents an analysis of the relationship between migrations and measures of linguistic similarity.

3. Phylogeographic analysis of 2,000 MtDNA sequences

3.1 Method

The main analysis uses 2,000 sequences, a sample constructed by downloading 31,845 complete mtDNA sequences from a public database, GenBank (Benson et al. 2013). 10,302 sequences had information on location, allowing them to be assigned a

GPS coordinate using the Python package geocoder, giving 621 locations around the world. The locations are unevenly sampled, with some countries such as Japan and the USA having hundreds of sequences, and some areas such as the Solomon Islands having extremely rich geographic information (due to papers such as Duggan et al. 2014).

The analysis was restricted to complete mitochondrial DNA genomes. Non-indigenous sequences in the Americas were removed to simplify the analysis by using an online program (Haplofind: https://haplofind.unibo.it/new/, Vianello et al. 2013) to classify mtDNA sequences by haplogroup, and retaining only A, B, C, D and X, as these are the five haplogroups which have been described in the literature on Native American mtDNA (e.g. Kumar et al. 2011). Because sequence alignment software such as CLUSTALW (Thompson et al. 1994) can currently align at most ~6000 sequences, the number of sequences used was capped to 6,028. This was arrived at by sampling sequences randomly by location with a maximum of 97 sequences per location, and the resulting sequences were then aligned using CLUSTALW. 2,000 sequences were then sampled, weighted by location, yielding a total of 423 locations, in order to yield a manageable number of sequences for the Bayesian phylogenetic analysis. An overview of the distribution of the sequences by country is given in Table 5.1.

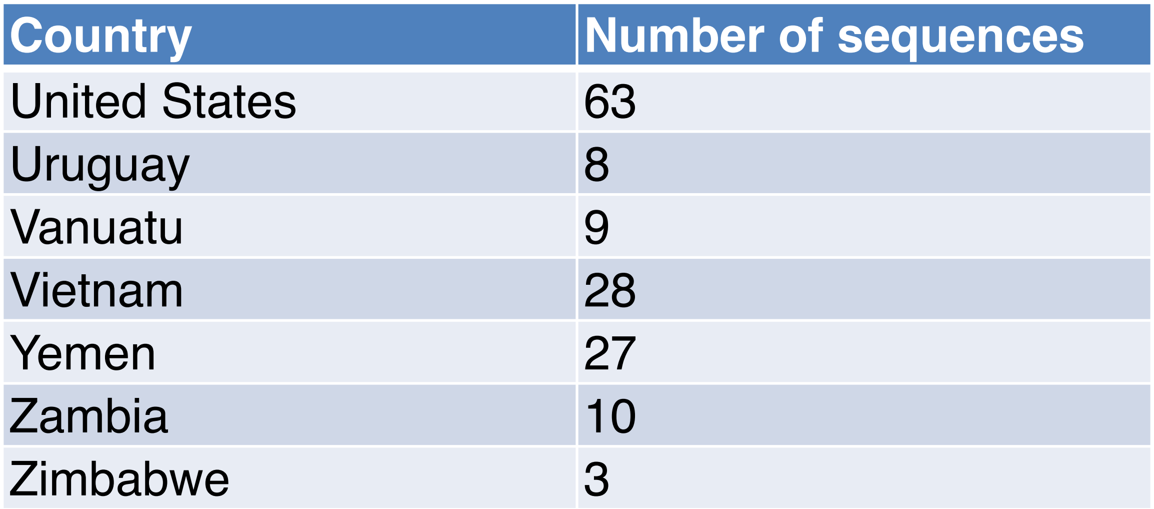
Country	Number of sequences
Albania	4
Algeria	10
Angola	10
Argentina	34
Armenia	3
Australia	9
Austria	1
Azerbaijan	9
Belarus	10
Bolivia	16
Bosnia and Herzegovina	3

Country	Number of sequences
Botswana	10
Brazil	11
Brunei Darussalam	8
Bulgaria	8
Burkina Faso	1
Cambodia	10
Cameroon	6
Canada	14
Chad	11
Chile	18
China	55
Colombia	28
Costa Rica	4
Croatia	1
Cyprus	3
Czech Republic	15
Denmark	10
Dominican Republic	4
Ecuador	11
Egypt	12
El Salvador	1
Equatorial Guinea	3
Estonia	6
Ethiopia	25
Fiji	10
Finland	10

Country	Number of sequences
France	13
Gambia	1
Germany	10
Ghana	3
Greece	20
Greenland	2
Guatemala	2
India	23
Indonesia	57
Iran	23
Iraq	10
Israel	10
Italy	142
Japan	10
Jordan	10
Kazakhstan	2
Kenya	5
Kiribati	1
Kuwait	10
Laos	10
Lebanon	13
Libya	10
Lithuania	2
Madagascar	7
Malaysia	55
Marshall Islands	2

Country	Number of sequences
Mauritania	9
Mauritius	9
Mexico	34
Micronesia, Fed. Sts.	5
Moldova	2
Mongolia	3
Morocco	30
Mozambique	10
Myanmar	3
Namibia	10
Nepal	20
New Zealand	2
Nicaragua	3
Niger	1
Nigeria	13
Oman	4
Pakistan	8
Panama	22
Papua New Guinea	113
Paraguay	3
Peru	43
Philippines	32
Poland	12
Portugal	10
Romania	15
Russia	213

Country	Number of sequences
Sao Tome and Principe	5
Saudi Arabia	10
Senegal	1
Serbia	10
Slovakia	10
Solomon Islands	161
Somalia	10
South Africa	47
South Korea	10
Spain	96
Sri Lanka	1
Sudan	10
Sweden	19
Switzerland	10
Syria	5
Taiwan	16
Tanzania	10
Thailand	4
Timor-Leste	10
Tunisia	10
Turkey	10
Tuvalu	10
Uganda	3
Ukraine	10
United Arab Emirates	20
United Kingdom	16



Country	Number of sequences
United States	63
Uruguay	8
Vanuatu	9
Vietnam	28
Yemen	27
Zambia	10
Zimbabwe	3

Table 5.1: Number of sequences included in the analysis by country/territory.

In addition to the above sequences, two Neanderthal sequences were included (Green et al. 2008, Sawyer 2013 respectively) as well as a Denisovan sequence (Krause et al. 2010).

The main analysis was run using BEAST, inferring a phylogeny of mtDNA sequences using the same methodology as described in Chapter 3: one can propose a particular family tree of how they are related to each other and how much change they have undergone, and calculate the probability of that tree generating those mtDNA sequences. The program then searches through possible trees and generates a sample of phylogenies sampled according to their posterior probability, using the Metropolis-Hastings algorithm (Metropolis et al. 1953). This calculation is implemented in the software package BEAST 2 (Bouckaert et al. 2014). The analysis was run for 10,000,000 iterations, with a burn-in of 50%.

The package ‘spherical phylogeography’ (Bouckaert 2016) was used to reconstruct ancestral locations of mtDNA lineages, and the results plotted in Google Earth. In the words of Bouckaert et al. (2012), the method uses a relaxed random walk model of continuous spatial diffusion along the branches of a phylogeny, treating location as a continuous vector (longitude and latitude) that evolves through time, and seeking to infer ancestral locations at internal nodes on the tree while simultaneously accounting for uncertainty in the tree. The resulting reconstruction therefore takes phylogenetic information into account; in the case of mtDNA migrations, for example, it predicts that humans originated in Africa, since the primary branches of the mtDNA family tree are known to be there (Ingman et al. 2000).

The reconstruction provides most likely longitudes and latitudes for the location of each ancestral node in the tree, as well as a 95% confidence interval. This allows migration routes to be reconstructed.

The method is summarised below:

1. I downloaded 31,845 mtDNA sequences from GenBank, which are contained in the text file 31845 documents.gb in the Github repository. 10,302 complete sequences had information on location, allowing them to be assigned a GPS coordinate using the Python package geocoder, giving 621 locations around the world.

2. 6,028 of these sequences were randomly selected, weighted by location, and aligned them using CLUSTALW, which are in the file all-aligned.txt.

3. In addition, some sequences were removed if they were in the Americas but they belonged to a haplogroup which was not Native American (A, B, C, D and X), using Haplofind.

4. A xml file in BEAST was prepared for the analysis, which contained 2000 taxa, weighted by location. The substitution model used was HKY (Hasegawa et al. 1985) with four gamma categories, and a Yule prior for the tree topology. There was a separate clock model used for location, meaning that people can be assumed to migrate at rates which are independent of the molecular clock (i.e. people can migrate at different speeds)[10]. The tree is not time-calibrated. The resulting file is in the repository, 1.xml.

5. I ran the analysis in BEAST for 10,000,000 iterations, with a burn-in of 50%, with the resulting trees in 1.trees.

6. The trees can be summarised in a consensus tree. Each node in this tree has a reconstructed location, which is the average of the locations of that node in the sample. From this, one can reconstruct migration routes that were taken along each branch of the tree. One migration is defined as the path from the reconstructed location of a node to the location of its immediate daughter.

[10] One possible objection to using BEAST for a phylogenetic analysis of the mtDNA samples is if any of the samples are ancestors of any others (e.g. mtDNA from a mother and child), since the analysis assumes that the taxa are all tips. However, there are very few (if any) examples of this in the data, and if there were then the result would just be that the taxa would be treated as very close siblings; this is not expected to significantly affect the topology of the rest of the tree.

7. In order to compare the consensus phylogeny with a standard mtDNA phylogeny such as Phylotree (van Oven and Kayser 2009), the sequences were run through the mtDNA haplogroup classification tool Mthap (Lick 2018). The haplogroups for each sequence are appended to the taxon names in the consensus tree, to visualise how well the consensus tree corresponds to Phylotree.

A 'migration' is defined above as **a movement from the location of an ancestral node to the location of its immediate child node**, yielding 4,318 distinct migrations in the current dataset. Migrations are defined this way because they are independent cases of someone in a location having a descendant (as close as possible in time to them) in a different location. The time scale of these migrations can be very different; in an extreme case, for instance, the earliest migrations in the reconstruction may span tens of thousands of years (migrations from Africa to Southeast Asia), while more contemporary migrations may be reconstructed on the scale of centuries.

This allows some quantitative study of tendencies in migrations in different regions. It is imperfect, because these are not random samples of people (they are sampled from what is available from genetic studies), and because of the small sample sizes and uneven sampling from different locations. There are patterns in these migrations however, such as particular routes which occur frequently, or differences in the probability of moving in particular directions.

'Migration' is also being used here to refer to the movement of individual lineages, and not to concerted movement of many people from a population. It would be fallacious to equate movements of lineages with concerted migrations, for two reasons: first, populations contain many different maternal lineages, and the ones which survive in modern mtDNA data are only a fraction of those. Second, the presence of an individual with a particular mtDNA haplogroup in a location does not necessarily imply a large-scale migration, as it could simply be due to the movement of one individual. An example would be if an ethnically Chinese individual in England were included in the analysis; the 'migration' from China to England that would be reconstructed in the phylogeny refers to the migration of one individual (the individual in the sample, or one of their ancestors), not to a concerted migration of people between these two regions.

The results of the analysis are therefore also very dependent on the choice of samples. In practice, it is assumed here that the individual samples reflect what is typical in the indigenous populations that they are drawn from. This could be improved by incorporating population frequencies of the haplogroups of the samples.

Despite the inability to infer anything about concerted migration of people from a small sample, there is still value in reconstructing migrations of individual lineages. This is because literature on mtDNA and Y chromosome haplogroups has in some cases reconstructed the movement of these lineages more informally, and compared these movements with the distribution of language families. To the extent that this exercise has value, then the analysis is in this chapter is a way of formalising the reconstruction and the comparison.

To give an example, Kayser et al. (2008) states: 'Linguistic (Blust 1999) and archaeological (Bellwood 2004; Bellwood and Dizon 2005) evidences strongly suggest a Taiwanese origin for the Austronesian expansion. This hypothesis is also supported by mtDNA evidence, as a genetic trail for the origin of the mtDNA "PM" [haplogroup B4a1a] (via its immediate precursor haplogroup B4a1) can be traced back to Taiwan (Redd et al. 1995; Trejaut et al. 2005)[…].' This statement makes five assumptions:

1. That haplogroup B4a1a does in fact originate in Taiwan;

2. That the Austronesian language family originates in Taiwan;

3. That there is a statistically supported correlation between the distribution of Austronesian languages and that of B4a1a;

4. That this correlation is meaningful, as opposed to the correlations with many other genetic markers that could have been used and could have had a spurious correlation with the distribution of Austronesian languages due to multiple testing;

5. And that this correlation is not simply because of spatial autocorrelation (neighbouring regions will tend to have both similar genetic markers and languages), but does in fact imply that Austronesian languages and the lineage B4a1a migrated together.

The analysis in this chapter aims to formalise assumption (1) by reconstructing the movement of mtDNA lineages using a model. Trejaut et al. (2005) are referenced, for example, for the Taiwanese origin of the lineage B4a1a, which is based on the fact that they find five primary branches of B4a1a in Taiwan; this is a type of intuitive phylogeographic reconstruction, which can be formalised by having a model reconstruct probable locations for ancestral nodes in a phylogeny.

Assumption (2) is not discussed in this chapter, since it involves phylogeographic reconstruction of language locations. However, assumption (3) is explicitly tested in section 4, which compares the paths that mtDNA lineages have taken with similarity between languages using phonotactics data from Chapter 3. Assumption (4) is tested by not cherry-picking particular haplogroups and language families, but by testing the general hypothesis that mtDNA lineages routes correlate with linguistic similarity. Finally, assumption (5) is tested by including a control for geographical distance (either in the regression models or in the Mantel test).

The following section presents the main results, and examples of patterns that emerge from the analysis of migration routes.

3.2 Results

The result of the phylogenetic analysis is a sample of trees, which can be summarised in a consensus tree (the maximum clade credibility tree, using BEAST's TreeAnnotator).

The consensus tree is plotted in Appendix 2 in Figures S3.1-52, which is plotted over several pages given the size of the tree. The effective sample size of the tree likelihood was low (4), due to the slow running time for the analysis when using 2000 sequences. As argued in section 5.3 of Chapter 3, this does not necessarily make the phylogeny completely invalid (especially compared with simpler heuristic methods such as UPGMA, which do not explore the tree space at all), but it is worth comparing the phylogeny in this case against a standard mtDNA phylogeny such as Phylotree (van Oven and Kayser 2009).

The phylogeny shows the Denisovan as the first primary branch to split off, followed by the two Neanderthals in a clade together as the next primary branch, followed by lineages in Africa: L1 coming off first as a primary branch (100% confidence), followed by L0 clades (100%), then L2a1 (99%), and then L4 and L6 together (99%). There are some differences with Phylotree, which has L0 coming off first before L1, and which does not have L4 and L6 in a clade together; and one taxon classified by Mthap as L2e1 which comes off first before L1.

The remainder of the tree is more rake-like, with low posterior probabilities on many clades. This perhaps reflects the lack of convergence in the analysis, as also indicated by the low ESS. However, the way that taxa are grouped together in the phylogeny still resembles Phylotree, whose primary branches are shown for comparison in Appendix 2 in Figure S4. For instance, taxa in haplogroup R (including R0, JT, HV etc.) are mostly placed in the same part of the tree (Figures S3.34-3.50), although not in a monophyletic clade. This applies similarly to other large haplogroups such as haplogroup M (Figures S3.3-3.19), but also to some smaller ones such as B4a1a1 (Figures S3.29-32 and S3.42-44). This is generally true of other clades in the tree, indicating that the analysis has found groupings that largely agree with Phylotree, but with low posterior probabilities and some inaccuracies due to lack of convergence.

Using this consensus tree, migration routes were plotted using the definition above (the path from the reconstructed location of a node to the location of its immediate daughter). A sample of these routes were then plotted in Google Earth. Since there are over 4000 routes reconstructed here, it is difficult to visualise them all simultaneously and interpret patterns. One way of summarising some more common patterns is to show regions where there has been especially frequent movement.

As an example, a region can be defined as a circle, namely all points within a certain distance (e.g. 400 km) of a particular location. An example of how some migrations can be summarised is shown in Figure 5.1, which shows migrations between one circle of radius 400 km in eastern New Guinea, and another circle of radius 400 km in the Solomon Islands. Each migration is represented by a blue dot for its starting point, and a red dot for its end point.

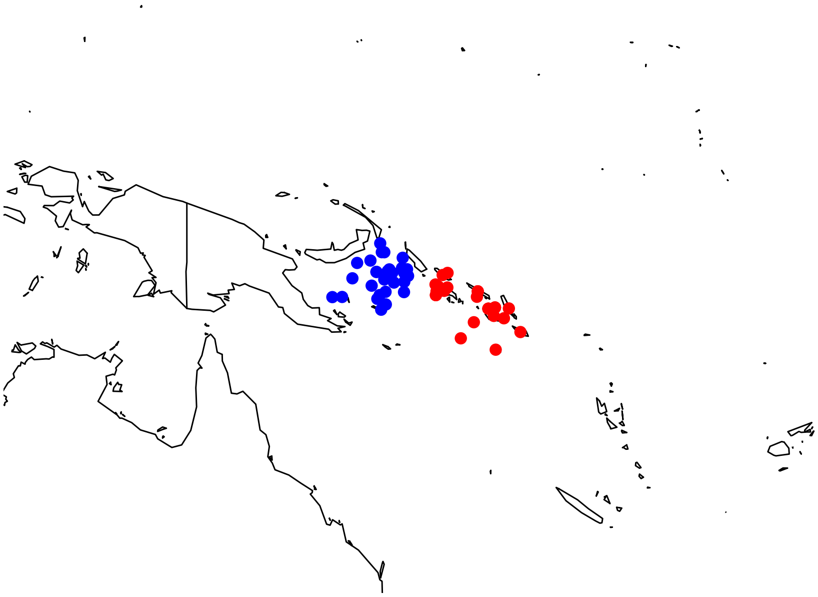


Figure 5.1: Movements from eastern New Guinea to the Solomon Islands. Blue dots mark the beginning of the migrations, and red dots mark the end point.

It is not particularly surprising that there is movement of people between local neighbouring regions, but some long-distance movement can be surprising and potentially illuminate the history of linguistic areas, or language families. An example is movement from Southeast Asia to New Guinea, which illustrates migration that possibly includes those that brought Austronesian languages, which originated in Taiwan and subsequently arrived in Indonesia, New Guinea, and islands in the Pacific (Greenhill, Drummond and Gray 2009).

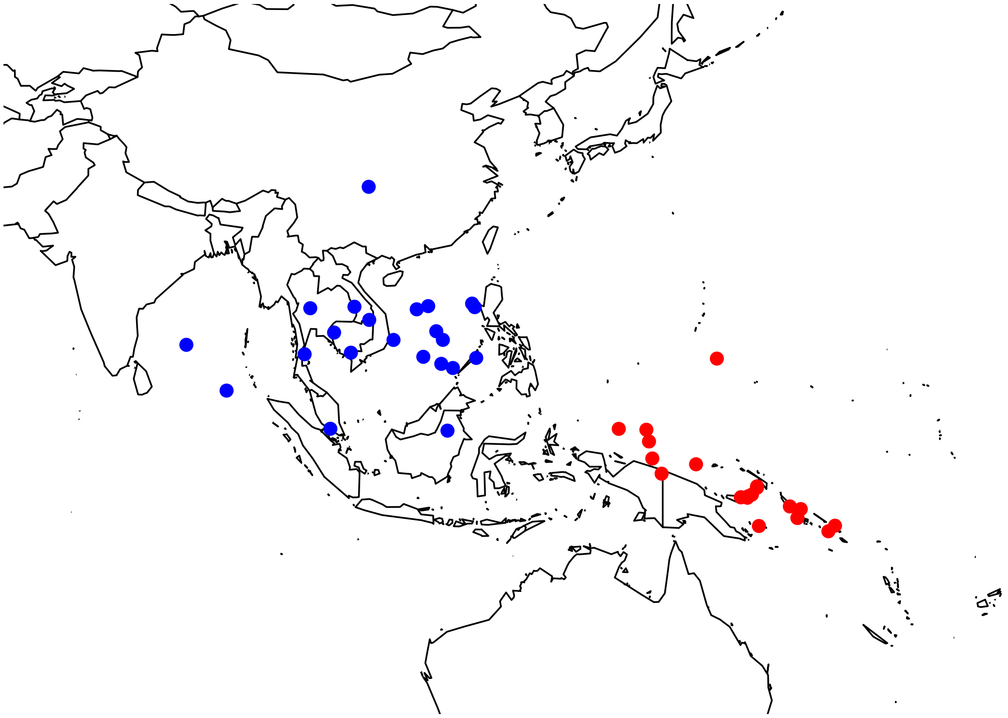


Figure 5.2: Movements from Southeast Asia to New Guinea. Blue dots mark the beginning of the migrations, and red dots mark the end point.

Other plausible examples include movement from the Steppe into northeastern Eurasia which may be associated with the movement of Uralic languages (Figure 5.3); or movement between India and Southeast Asia, associated with languages families spanning both regions such as Sino-Tibetan or Austroasiatic (Figures 5.4 and 5.5). Sino-Tibetan is likely to have originated in Northern China according to recent phylogenetic analyses (Zhang et al. 2019, Sagart et al. 2019) and spread westwards to India. How Austroasiatic spread is less clear, although there is some evidence that it moved outwards from South China (Sidwell 2015b). Regardless of the direction that these language families moved in, there is likely to have been migration between these two regions associated with the movement of languages, and accordingly there are migrations in both directions between India and Southeast Asia.

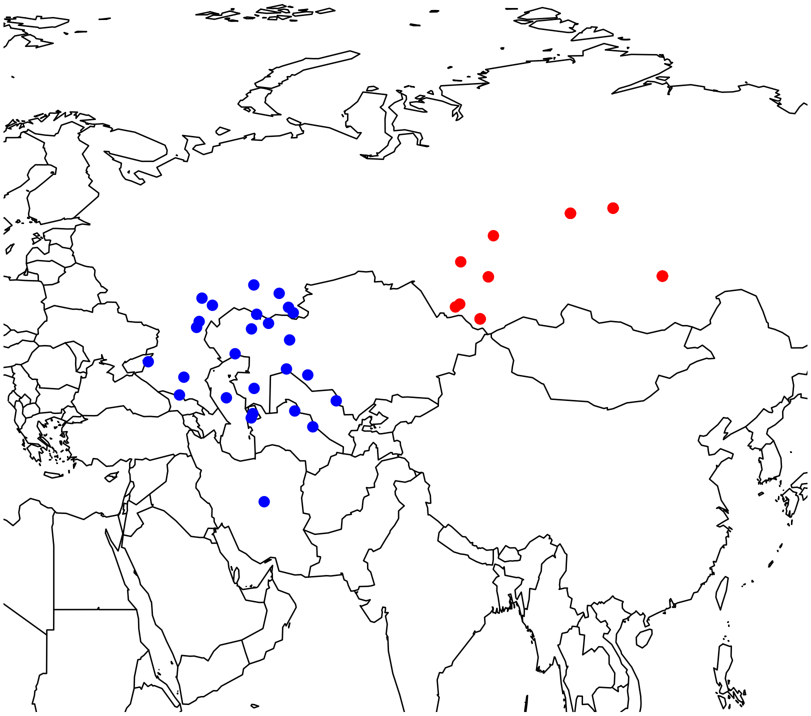


Figure 5.3: Movements across northern Eurasia. Blue dots mark the beginning of the migrations, and red dots mark the end point.

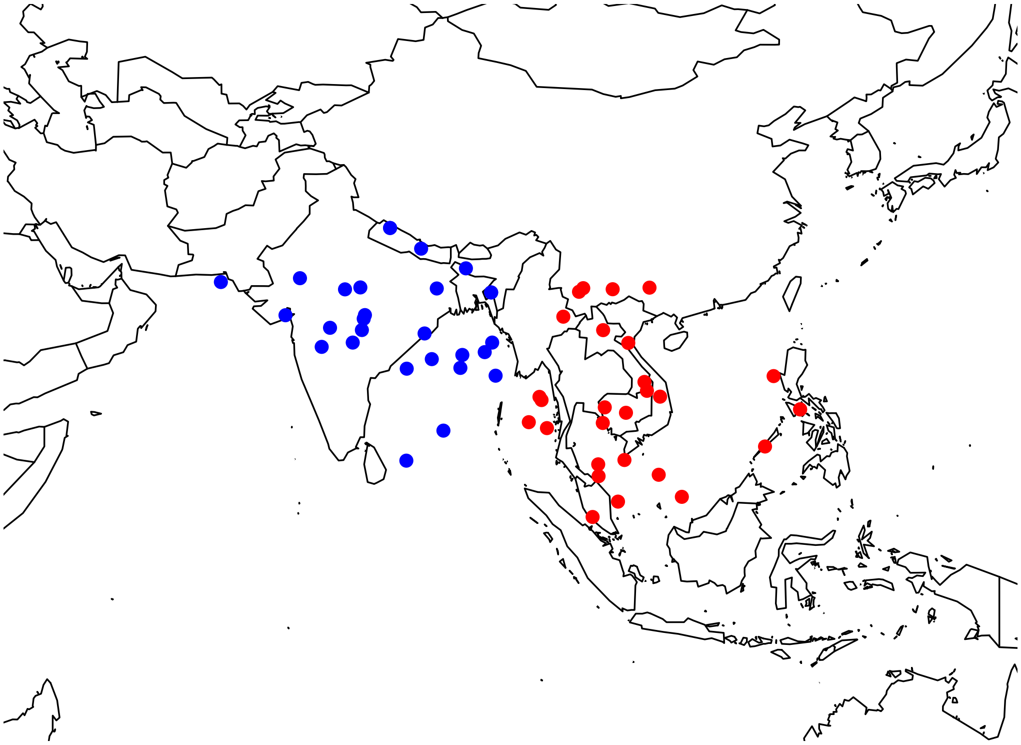


Figure 5.4: Movements from India to Southeast Asia. Blue dots mark the beginning of the migrations, and red dots mark the end point. Note that some points are on water, since the phylogeographic model used from Bouckaert (2016) reconstructs latitude and longitude without awareness of landscape.

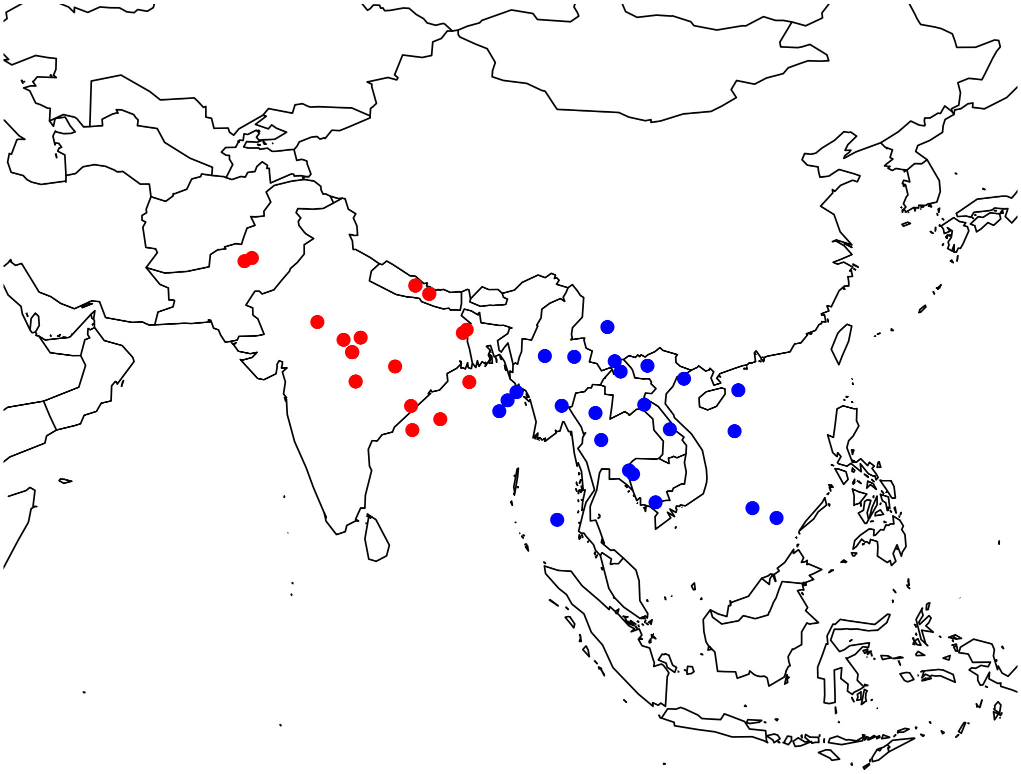


Figure 5.5: Movements from Southeast Asia to India. Blue dots mark the beginning of the migrations, and red dots mark the end point.

Other migrations may make more sense not from the point of language families, but from the point of view of linguistic areas. Figure 5.6 shows movement between northern India and the Caucasus and Middle East, which resembles the linguistic area found in the previous chapter that links these regions. Finally, some examples of migrations between regions do not seem to have any linguistic correlate; for instance, there is plenty of movement between North Africa and southern Europe, as Figure 5.7 shows, but there is no language family spanning both areas.

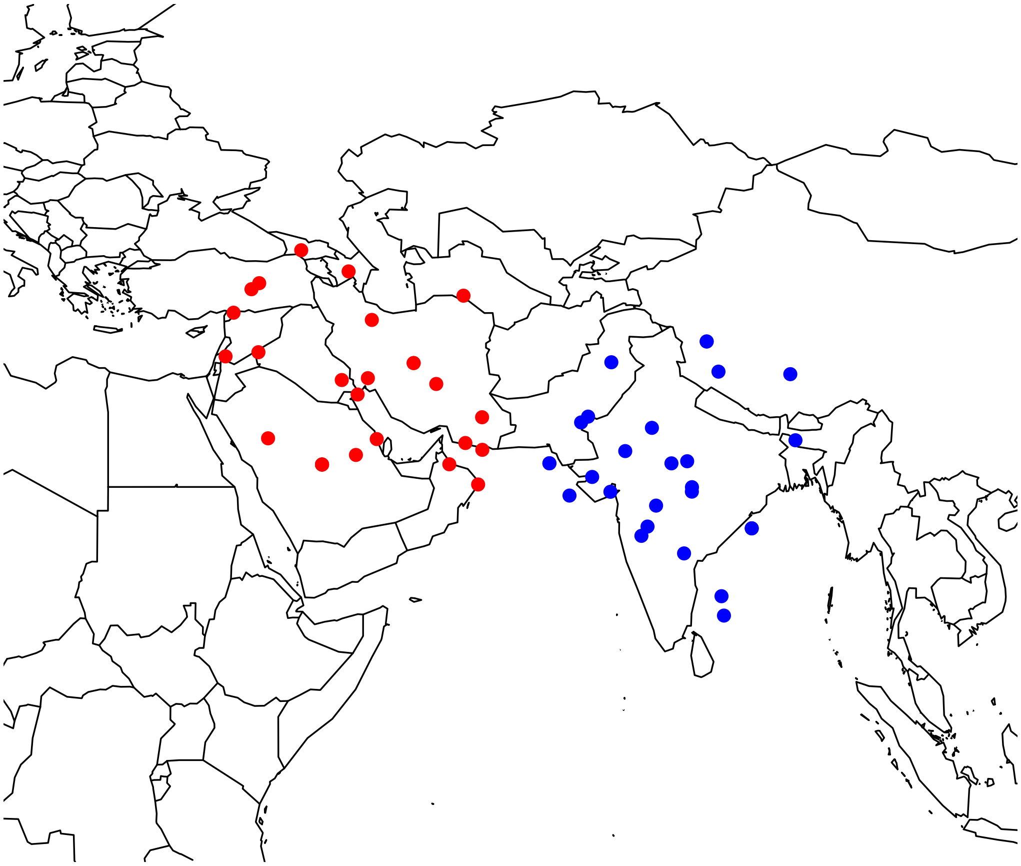


Figure 5.6: Movements from the Indian subcontinent to the Caucasus and the Middle East. Blue dots mark the beginning of the migrations, and red dots mark the end point.

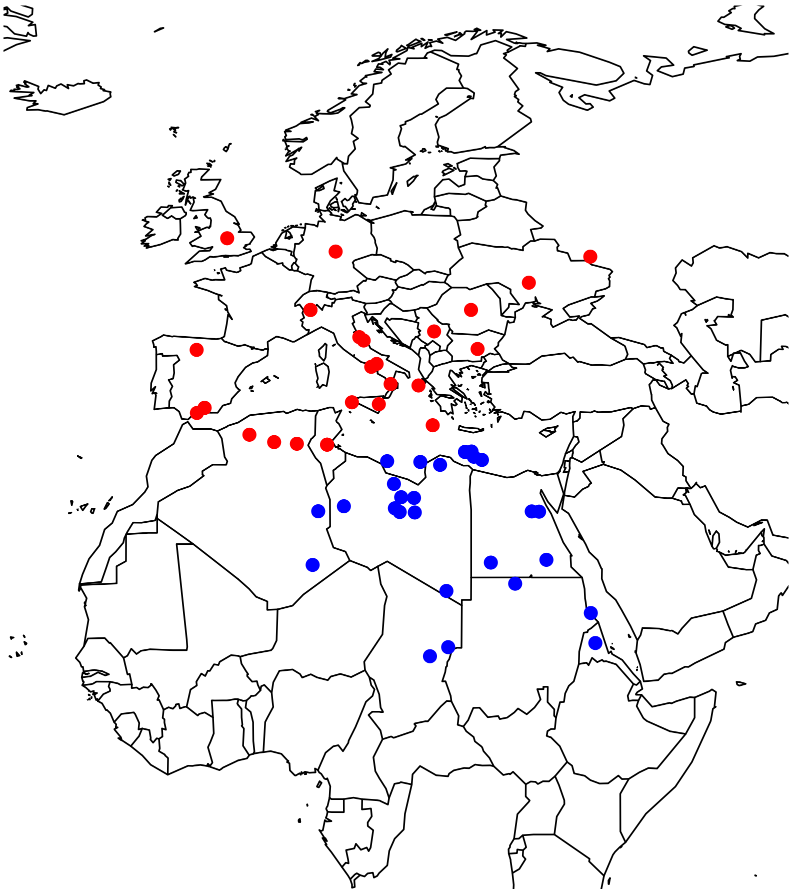


Figure 5.7: Movements from North Africa to Europe. Blue dots mark the beginning of the migrations, and red dots mark the end point.

Another pattern in the migrations is that there are different directional tendencies in different parts of the world; in particular, whether people are more likely to move in an east-west direction, or whether they move north-south. The angle that a migration is at relative to the equator can be calculated, and this angle differs on average in different places. To visualise this point, squares are plotted where this angle has tended to be less than 25 degrees, and squares where the angle has tended to be greater than 25 degrees. Each square contains thirty migrations, and the cut-off point for statistical significance is

67% (the proportion above which there is a less than five percent probability of there being an equal probability of movements above and below 25 degrees).

To focus again on Africa and Eurasia in particular, regions where migrations have tended to be at more vertical, with a greater than 25 degree angle from the equator, are shown in Figure 5.8, while regions where migrations have tended to be more horizontal (less than a 25 degree angle from equator) are shown in Figure 5.9.

There are some suggestive directional tendencies which may be behind the formation of linguistic areas found in Chapter 2. There is a tendency for horizontal movement across the Himalayas, for instance, which may be behind the linguistic areas crossing from India to Southeast Asia. Another area contained Indo-European and Turkic, an area which covers a large distance from east to west, in line with the tendency for there to be concerted horizontal movement in the region between Europe and northern China. Finally, the crescent of languages from the Middle East and the Caucasus to northwest India is also seen here, since there is a tendency for horizontal movement in this region. Eurasia generally has had a lot of east-west movement, as Jared Diamond's *Guns, Germs and Steel* famously noted as an explanation for how Eurasia was especially prosperous due to the exchange of agricultural plants and other innovations across the continent.

There are also some striking areas where vertical movement (at a greater than 25 degree angle) has predominated. This is most true in Africa, where the angles of migrations have in fact tended to be above 45 degrees, perhaps mostly due to the southern expansion of Bantu languages. It is also true in island Southeast Asia, probably due to the Austronesian expansion; and in China and Southeast Asia, where several vertical clusters of languages were discovered in Chapter 2 based on similarity of phonology. This is plausibly driven by the expansion of particular families, such as the southern movement of Tai-Kadai speakers out of China to the rest of Southeast Asia; as Enfield (2005:184) writes, 'MSEA geography is dominated by major river systems running north to south…In their Southern reaches, these rivers empty into wide plains now dominated by dense populations of paddy farmers speaking varieties of Vietnamese, Khmer, Siamese, and Lao. Highland areas north of these plains and running south along the Annamite Cordillera are home to ethnic minorities who practice mostly shifting agriculture (i.e., slash-and-burn). This pattern has resulted from major migrations over the past two millennia, mostly southward from China toward the lowlands. The most significant historical migrations have been the southwest fanning spread of Tai speakers from southwest China (Enfield 2003b, Wyatt 1984).'

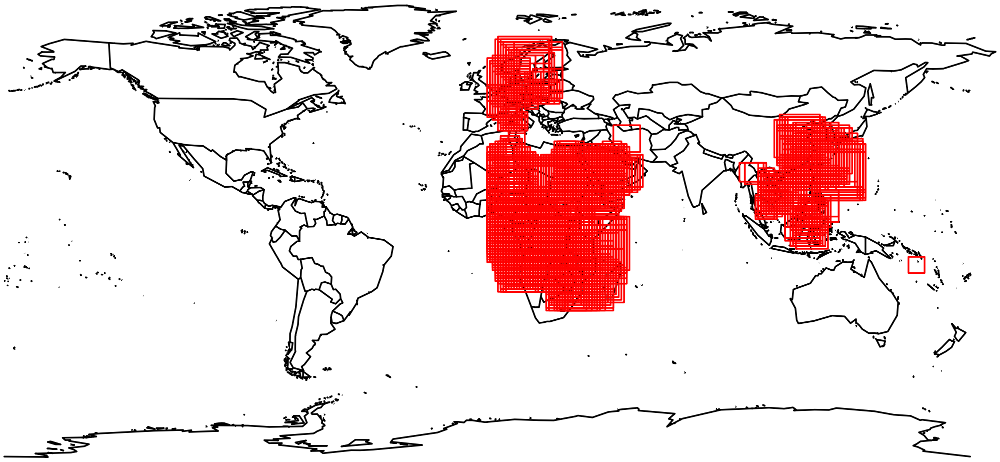


Figure 5.8: Regions where there is a significant tendency for migrations to be move north-south.

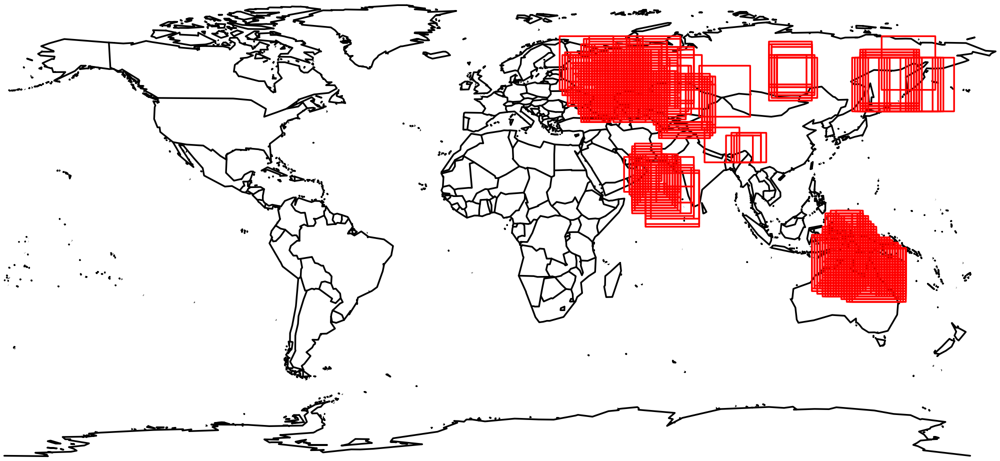


Figure 5.9: Regions where there is a significant tendency for migrations to be move east-west.

The regions shown in these three figures are suggestive, since a tendency for concerted movement in particular directions (east-west, or north-south) may be what drives the linguistic homogenisation of some regions. A mass migration southwards in southeast Asia, for instance, is detectable both from genetic data here as a statistical tendency to move in a particular direction (southwards, at a greater than 25 degree angle), and also from linguistics, from the similarity of languages in this region in their structural properties.

4. Measuring the effect of migrations on language similarity

4.1 Method

This section reports on a more explicit test of the hypothesis that migrations have shaped the distribution of languages. The preceding sections have hinted at this, suggesting for example that concerted migrations may cause some regions to be especially similar linguistically.

Linguistic similarity can be calculated by using the results from Chapter 3 of this thesis, which described how one can use the World Phonotactics Database to analyse according to their phonological and phonotactic properties. This was done by using a phylogenetic method, producing a family tree of languages that reflected both inheritance in known language families and language contact. In a second analysis in that chapter, the method was used again but with known language families as constraints.

These two measures of linguistic distance will be called ‘unconstrained’ and ‘constrained’ linguistic distance, respectively. The two family trees are given in Figures 3.1 and 3.12 in Chapter 3, and are reproduced here:

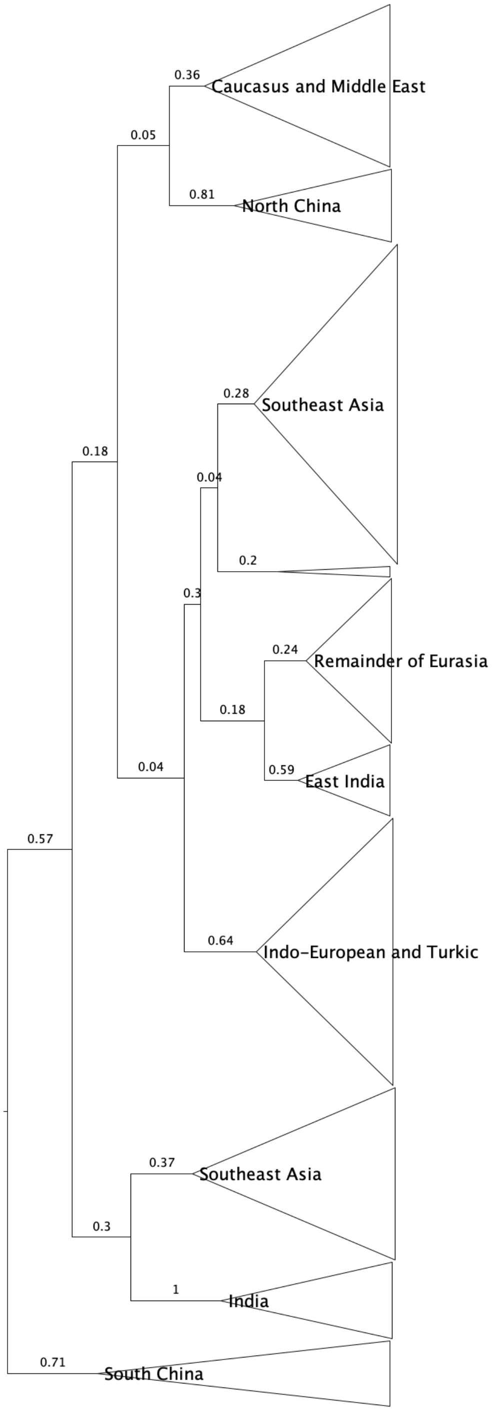


Figure 5.10: A consensus tree of languages in Eurasia according to the phylogenetic analysis of the World Phonotactics Database (Donohue et al. 2013), reproduced from Chapter 3, Figure 3.1. The main clades are collapsed and summarised with a geographical label for readability.

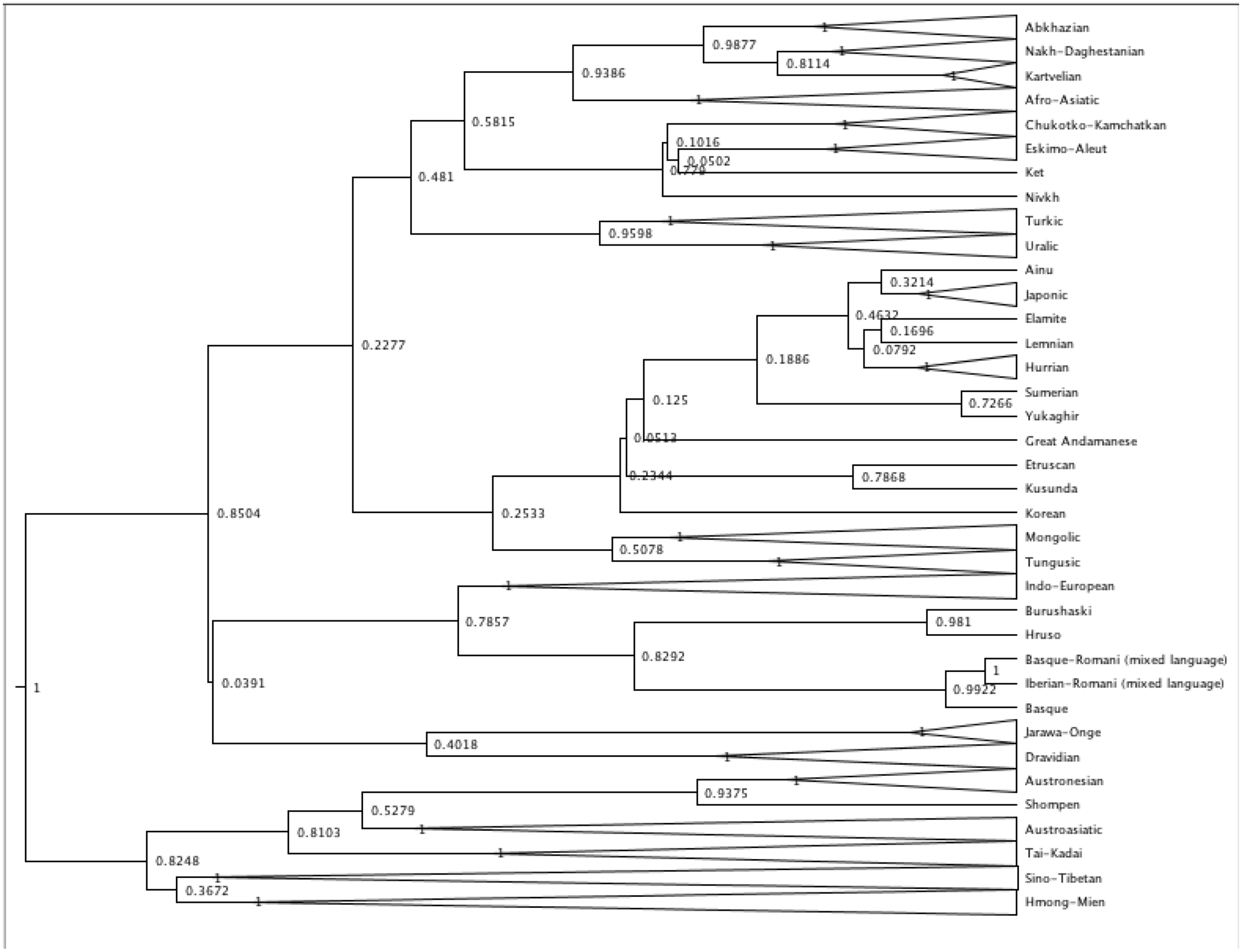


Figure 5.11: The family tree of language families in Eurasia based on phonotactic properties, with constrained clades for language families, reproduced from Chapter 3, Figure 3.12. Numbers on clades are posterior probabilities. The lower primary clade is referred to as the 'Southeast Asian' clade in parts of this chapter, and the upper clade is referred to as the 'Eurasian' clade.

Constrained linguistic distance was defined in Chapter 3 by finding a phylogeny of Eurasian languages based on their phonological properties, but with known language families as constraints. This means much of the information that constrained linguistic distance provides is equivalent to saying how closely related they are. However, beyond the level of language family, it shows what linguistic areas languages belong to; for instance, language families in Southeast Asia are grouped together in the tree, meaning that languages are close to each other partly due to what linguistic area they belong to.

These family trees can be used to provide a measure of linguistic distance between pairs of languages A and B, calculated by taking the patristic distance between the two languages (the sum of the branch lengths of those two languages back to their common ancestor; e.g. Stuessy and König 2008). The linguistic distance between two languages can be predicted in part from their geographical distance: the closer two languages are, the more likely they are to share phonological properties in common. This can be demonstrated by plotting geographical distance to linguistic distance (here referring to constrained linguistic distance).

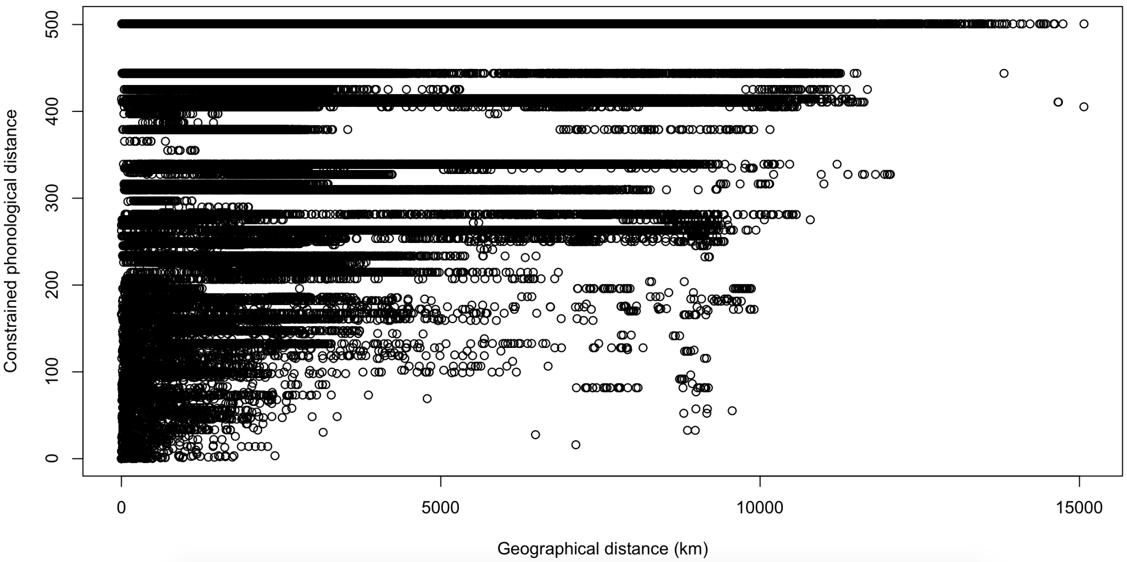


Figure 5.12: A plot of geographic distance (km) and constrained linguistic distance. The graph seems to suggest that languages are more similar to each other on average the closer they are to each other. The units on the y-axis refer to the patristic distance between pairs of languages, calculated as the sum of branch lengths of the branches connecting the languages to their common ancestor in the tree: these units are arbitrary as the tree is not time-calibrated, and so should be understood only as providing a relative measure of phonological similarity between languages.

An additional factor shaping how languages are similar may be the number of migrations taking place between them. For instance, two languages far apart from each other may be similar because there were migrations between those two places: modern examples include British and American English, spoken on different continents but both a type of English, because of migration from England to America. Examples in the case of older language families may include the migration of speakers of Polynesian languages, who will be genetically quite similar to each other as well as speaking similar languages, due to recent divergence between 800 and 1000 years BP (Ioannidis et al. 2021; Gray, Drummond and Greenhill 2009), despite the large distances that they covered across the Pacific.

In these cases, there are languages which are more similar to each other than one would expect, which can be explained by long-distance migrations. Data from genetics can therefore have predictive power, not just in these reasonably clear and recent cases, but in cases of migrations far in the past before historical records; conversely, linguistics can help to explain patterns in genetics.

Using the dataset of migrations inferred from mtDNA in this chapter, a model can be made which predicts linguistic distance not just from geographical distance, but also from the number of migrations that have taken place between particular pairs of languages.

Since the linguistic analysis in Chapter 3 uses only languages from Eurasia, the analysis in this section is limited to migrations in Eurasia. The start and end points of each migration are taken and assigned languages which they are closest to. We can then say for each pair of languages in our sample how many migrations occur in the dataset. This methodology is of course crude, in that a person making a migration perhaps several thousand years ago is assigned to a language in the present day; in many cases, an ancient migration has nothing to do with the two languages which are at the beginning and end points of the migration pathway. However, there are also examples of migrations within the last few thousand years which will be relevant to modern day languages. The first type would be migrations which happened in recent history, say within the last few hundred years, where the migration is in fact between the two linguistic communities that we are assigning it to. A second type would be migrations further back in the past, before the languages they are assigned to even existed, but which were between languages related (or ancestral) to the ones that they have been assigned to. One example may be the movement of Indo-European speakers, say from Eastern to Western Europe (see the reconstructed migrations of Indo-European languages in Bouckaert et al. 2012): this type of migration made the modern day languages that the migration is assigned to (say, Polish and French) more similar, by virtue of Polish and French being descendants of Indo-

European when it spread into Europe. Even though the migration was not literally from a Polish-speaking community into a French-speaking community, the migration is therefore still relevant to understanding the linguistic similarity between Polish and French. Another example is how language families may have influenced each other, or may even be related, and how the similarities between these families may correspond to migrations between them; for example, migrations within the last few thousand years between Turkic, Tungusic, Mongolic and Japonic languages may reflect the relatedness of these families (Robbeets 2005) or ancient contact between them.

It is therefore expected that the number of migrations between each pair of languages should also show a systematic relationship with linguistic distance, and indeed it does; the more migrations there are between two languages, then the more similar they are. However, this could be a trivial by-product of geographical distance: the closer two languages are, the more likely they are to be similar, and the more likely that people will move between them. One needs to control for geographical distance in some way to find out what the relationship between genetics and linguistic similarity is.

The approach taken in this section is to examine two competing models, one where linguistic distance is a function entirely of geographic distance; and another where linguistic distance is partially a function of both geographic distance and number of migrations. These two models are then evaluated by their Bayesian Information Criterion (BIC; Schwarz 1978), which includes a penalty term for the number of parameters in model. The difference between the BIC of the two competing models is taken, with differences of more than 10 being strong evidence in favour of the model with the lower BIC (Kass and Raftery 1995).

A summary of the method is provided below:

1. Migration pathways inferred from the phylogeny in section 3 were used. These are given in the data file movements2000.csv.

2. Patristic distances between each pair of languages in Eurasia were calculated from the phylogenies discovered in chapter 3. These distances will be referred to as ‘linguistic distance’. This was done using two different phylogenies, the ‘constrained’ phylogeny which has clade constraints, and the ‘unconstrained’ phylogeny which does not. Steps 3 and 4 were each done for constrained and unconstrained linguistic distance.

3. A model was made predicting the linguistic distance between pairs of languages from geographical distance, and a second model which also incorporates the number of

mtDNA migrations between the pair. This is calculated by finding the nearest language to the points on the migration pathway. The functions lm and poly in R were used in each case (R Core Team 2021).

4. The Bayesian Information Criterion of two different models was taken. The first model predicts linguistic distance just from geographical distance, fitting it as a polynomial function of the form $f(x) = ax^3 + bx^2 + cx + d$. The number of parameters for the function (3 in this example) is tested up to 25, and the model with the lowest BIC found, which fits the data best while at the same time using as few parameters as possible. The second model takes this function but also includes a polynomial function of linguistic distance, again going up to 25 parameters; the lowest BIC of this second model is found and compared with the BIC of the first model, to assess the strength of evidence that mtDNA migrations can explain linguistic distance. The function BIC from the R package flexmix (Gruen and Fleisch 2008) was used for calculating the BIC.

5. The second model in step 4 uses absolute number of mtDNA migrations, but the model may be improved by using a *proportion* of migrations to a particular destination from a particular starting point; for example, instead of using an absolute value such as 5 migrations from French to Italian, one can take the probability of a migration from French being to Italian (as opposed to German). The use of language labels here refers to end points of migrations and their assignment to a language that is nearest to that point, as explained in step 3. Finally, the model can be improved further by taking the probability of a migration being between French and Italian (in either direction) given that it the migration involves at least one of the two languages.

4.2 Results

There are separate results for constrained and unconstrained linguistic distance. The results for constrained linguistic distance are summarised first in the following section.

4.2.1 Constrained linguistic distance

The model that includes migrations as well as geographic distance is strongly preferred to a model which just uses geographic distance (Bayes factor = 1740). The preferred model has fourteen parameters for geographical distance, and five parameters for number of migrations.

The model which uses *proportion* of migrations to a particular destination given a particular starting point was also tested, and this was found to improve the model over the model which uses an absolute number (Bayes factor = 55). A symmetrical model was then tested, which takes the number of migrations between two languages and divides it by the total number of migrations involving at least one of the two languages. This model was found to improve the model even further (Bayes factor = 1208) over the unsymmetrical model. This final model uses fourteen parameters for geographic distance and nine parameters for migrations.

To investigate why mtDNA migrations improve the model, one can look at the individual pairs of languages which are better predicted by migrations than by geographical distance. The squared error of each prediction of the geography-only model and the model with migrations can be taken, and the pairs of languages with the largest differences between these squared errors can be found. A sample of these languages sorted by largest to smallest difference is shown in Table 5.2.

Most of these pairs of languages are within the same language family (as this measure of linguistic distance is constrained by language family), and often spread over large distances. Some especially interesting examples include Maltese, an Afro-Asiatic language spoken in Southern Europe, whose similarity to other Semitic languages is predicted by migrations from the Middle East. Other examples include Chinese varieties, such as Cantonese and Mandarin, whose relatedness is again predicted by migrations between points where they are spoken, contrary to the expectation from their geographic distance that they should be very different from one another. Other examples include languages in Indo-European branches, often within the same branch (Spanish and Portuguese, Bulgarian and Czech, German and Danish); and similarly, Turkic languages and Japanese varieties are often similar to each other despite being spread over large distances.

However, it is not only languages within the same family that are better predicted by migrations than by geographic distance. In many cases these pairs of languages are from different families but from the same broader linguistic area. A list is given in Table 5.3 of these families, which are part of linguistic areas and also have migrations between them. Particularly interesting examples include movements in northeastern Eurasia, such as between Japanese and Korean, and with other families such as Eskimo-Aleut, Chukotko-Kamchatkan, Tungusic, and Mongolic. There are also interactions between Afro-Asiatic languages and languages in the Caucasus. A complete list of cross-family interactions which contribute to making languages more similar is in Table 5.3.

Language 1	Language 2
Kazakh (Turkic)	Uzbek (Turkic)
Karakalpak (Turkic)	Uzbek (Turkic)
Romanian (Indo-European)	Ladino (Indo-European)
Neapolitan (Indo-European)	Galician (Indo-European)
Spanish (Indo-European)	Neapolitan (Indo-European)
Arabic Omani (Afro-Asiatic)	Arabic Syrian (Afro-Asiatic)
Ukrainian (Indo-European)	Slovene (Indo-European)
Arabic Syrian (Afro-Asiatic)	Arabic Yemeni (Afro-Asiatic)
Neapolitan (Indo-European)	Aragonese (Indo-European)
Kirghiz (Turkic)	Uzbek (Turkic)
Sicilian (Indo-European)	Galician (Indo-European)
Neapolitan (Indo-European)	Ladino (Indo-European)
Ukrainian (Indo-European)	Dacian (Indo-European)
Mordvin (Uralic)	Cheremis (Mongolic)
Spanish (Indo-European)	Asturian (Indo-European)
Catalan (Indo-European)	Asturian (Indo-European)
Italian (Indo-European)	Sicilian (Indo-European)
Japanese (Japonic)	Irabu Ryukyu (Japonic)
Spanish (Indo-European)	Aragonese (Indo-European)
Turkish Crimean (Turkic)	Uzbek (Turkic)
Bashkir (Turkic)	Urum (Turkic)
Slovak (Indo-European)	Czech (Indo-European)
Arabic Omani (Afro-Asiatic)	Arabic Saudi (Afro-Asiatic)
Corsican (Indo-European)	Galician (Indo-European)
Arabic Saudi (Afro-Asiatic)	Arabic Palestinian (Afro-Asiatic)
Arabic Saudi (Afro-Asiatic)	Arabic Syrian (Afro-Asiatic)
Bulgarian (Indo-European)	Russian (Indo-European)
Neapolitan (Indo-European)	Catalan (Indo-European)
Bulgarian (Indo-European)	Ukrainian (Indo-European)
Ukrainian (Indo-European)	Slovak (Indo-European)
Sicilian (Indo-European)	Catalan (Indo-European)
Portuguese (Indo-European)	Spanish (Indo-European)
Romanian (Indo-European)	Italian (Indo-European)
Ukrainian (Indo-European)	Polish (Indo-European)
Dimili (Indo-European)	Yazdi (Indo-European)
Turkish Crimean (Turkic)	Urum (Turkic)
Italian (Indo-European)	Asturian (Indo-European)
Sardinian (Indo-European)	Aragonese (Indo-European)
Sardinian (Indo-European)	Asturian (Indo-European)
Sinhalese (Indo-European)	Maldivian (Indo-European)
Catalan (Indo-European)	Sicilian (Indo-European)
Khorasani Turkic (Turkic)	Turkish (Turkic)
Shodon (Japonic)	Japanese (Japonic)
Slovak (Indo-European)	Ukrainian (Indo-European)
Romansh (Indo-European)	Galician (Indo-European)
Koryak (Chukotko-Kamchatkan)	Kerek (Chukotko-Kamchatkan)
Finnish (Uralic)	Kildin Sami (Uralic)
Catalan (Indo-European)	Aragonese (Indo-European)
Sicilian (Indo-European)	Neapolitan (Indo-European)
Tsat (Austronesian)	Chru (Austronesian)

Language 1	Language 2
Sardinian (Indo-European)	Galician (Indo-European)
Catalan (Indo-European)	Galician (Indo-European)
Urum (Turkic)	Turkish Crimean (Turkic)
Turkish (Turkic)	Uighur (Turkic)
Urum (Turkic)	Uzbek (Turkic)
Lombard (Indo-European)	Asturian (Indo-European)
Arabic Syrian (Afro-Asiatic)	Arabic Gulf (Afro-Asiatic)
Russian (Indo-European)	Belarussian (Indo-European)
Arabic Saudi (Afro-Asiatic)	Arabic Gulf (Afro-Asiatic)
Japanese (Japonic)	Old Japanese (Japonic)
Albanian Tosk (Indo-European)	Albanian Gheg (Indo-European)
Uighur (Turkic)	Yugur (Turkic)
Catalan (Indo-European)	Spanish (Indo-European)
Irabu Ryukyu (Japonic)	Amami Ryukyu (Japonic)
Aromanian (Indo-European)	Romansh (Indo-European)
Sardinian (Indo-European)	Corsican (Indo-European)
Ukrainian (Indo-European)	Czech (Indo-European)
Ukrainian (Indo-European)	Russian (Indo-European)
Yazdi (Indo-European)	Dimili (Indo-European)
Russian (Indo-European)	Ukrainian (Indo-European)
Irabu Ryukyu (Japonic)	Ainu (Ainu)
Aromanian (Indo-European)	Italian (Indo-European)
Lombard (Indo-European)	Occitan (Indo-European)
Itelmen (Chukotko-Kamchatkan)	Kerek (Chukotko-Kamchatkan)
Arabic Omani (Afro-Asiatic)	Hebrew (Afro-Asiatic)
Slovak (Indo-European)	Polish (Indo-European)
Catalan (Indo-European)	Neapolitan (Indo-European)
Italian (Indo-European)	Walloon (Indo-European)
Khalkha (Mongolic)	Buryat (Mongolic)
Belarussian (Indo-European)	Slovene (Indo-European)
Aromanian (Indo-European)	Dalmatian (Indo-European)
Azerbaijani (Turkic)	Turkmen (Turkic)
Corsican (Indo-European)	Italian (Indo-European)
Mordvin (Uralic)	Karelian (Uralic)
Maltese (Afro-Asiatic)	Hebrew (Afro-Asiatic)

Table 5.2: The first 86 pairs of languages that which are better predicted by migrations between them than by their geographic distance.

Language family 1	Language family 2
Afro-Asiatic	Kartvelian
Afro-Asiatic	Nakh-Daghestanian
Afro-Asiatic	North-west Caucasus
Afro-Asiatic	Turkic
Ainu	Mongolic
Ainu	Tungusic
Ainu	Eskimo-Aleut

Language family 1	Language family 2
Ainu	Chukotko-Kamchatkan
Andaman	Dravidian
Austronesian	Austroasiatic
Chukotko-Kamchatkan	Afro-Asiatic
Chukotko-Kamchatkan	Eskimo-Aleut
Chukotko-Kamchatkan	Turkic
Chukotko-Kamchatkan	Mongolic
Dene-Yeniseic	Turkic
Dravidian	Tibeto-Burman
Eskimo-Aleut	Chukotko-Kamchatkan
Eskimo-Aleut	Turkic
Eskimo-Aleut	Uralic
Eskimo-Aleut	Korean
Eskimo-Aleut	Japonic
Eskimo-Aleut	Ainu
Indo-European	Tibeto-Burman
Indo-European	Tungusic
Indo-European	Basque
Indo-European	Hurro-Urartian
Japonic	Ainu
Japonic	Uralic
Japonic	Korean
Japonic	Tungusic
Japonic	Mongolic
Japonic	Chukotko-Kamchatkan
Japonic	Eskimo-Aleut
Japonic	Turkic
Kartvelian	North-west Caucasus
Kartvelian	Nakh-Daghestanian
Kartvelian	Turkic
Korean	Japonic
Korean	Mongolic
Korean	Eskimo-Aleut
Korean	Turkic
Mongolic	Tungusic
Mongolic	Uralic
Mongolic	Sumerian
Mongolic	Turkic
Mongolic	Kusunda
Mongolic	Nakh-Daghestanian
Mongolic	Eskimo-Aleut
Mongolic	Afro-Asiatic
Mongolic	Chukotko-Kamchatkan
Mongolic	Tibeto-Burman
Nahali	Indo-European
Nakh-Daghestanian	Afro-Asiatic
North-west Caucasus	Afro-Asiatic
North-west Caucasus	Turkic
North-west Caucasus	Uralic
Sumerian	Indo-European
Tai-Kadai	Tibeto-Burman
Tibeto-Burman	Indo-European

Language family 1	Language family 2
Tibeto-Burman	Mongolic
Tibeto-Burman	Tai-Kadai
Tungusic	Mongolic
Tungusic	Korean
Tungusic	Sumerian
Tungusic	Uralic
Tungusic	Nakh-Daghestanian
Tungusic	Turkic
Tungusic	Eskimo-Aleut
Tungusic	Chukotko-Kamchatkan
Turkic	Uralic
Turkic	Mongolic
Turkic	Eskimo-Aleut
Turkic	Kartvelian
Turkic	Afro-Asiatic
Turkic	Tungusic
Turkic	Japonic
Uralic	Mongolic
Uralic	Turkic
Uralic	Chukotko-Kamchatkan
Uralic	Yukaghir

Table 5.3: A list of language families which have migrations between them that better predict linguistic distance than geographic distance, sorted alphabetically by language family 1.

In some cases it is difficult to know how to interpret these migrations, especially cases such as migrations between Japonic and Eskimo-Aleut. The particular languages in this list are not necessarily accurate, as there is uncertainty around the locations of the migrations themselves (they are reconstructed from mtDNA and approximate modern day positions of people); and because the method of selecting the languages the migration is associated with (by selecting nearest languages to the start and end) is also approximate. Furthermore, in some cases, the relevant reconstructed migration may have taken place long before the languages which those points are being assigned to existed.

What can be learnt from this analysis, if the individual pairs of languages and migrations here are not precise? It demonstrates a general point that linguistic areas such as Northern Eurasia are created by people moving across linguistic communities; and also that language families are likely to have spread at least in part because of people migrating, rather than because languages were adopted without people moving. Although it is possible for linguistic features to travel between languages without people moving (loanwords being the archetypal example), in practice a model predicting linguistic similarity is helped by including migration data. Conversely, it shows a way that linguistic information can be usefully employed in genetics research, a topic that will be returned to in the discussion in section 4.2.3.

4.2.2 Unconstrained linguistic distance

By contrast with constrained linguistic distance, unconstrained linguistic distance is only slightly better predicted by adding migrations than by a model which just uses geography (Bayes factor = 3). The best fitting model has 22 parameters for geographic distance and one parameter for migrations. Again by contrast with constrained linguistic distance, using proportions rather than absolute numbers of migrations makes the model worse (Bayes factor = 16).

This suggests a qualitative difference between constrained and unconstrained linguistic distance: the former is very well predicted by adding migrations to the model, while the latter is mostly predicted by geography. This is perhaps in line language families being good indicators of migration (e.g. Japanese speakers to the Ryukyuan islands, or the spread of Arabic), whereas phonological similarity that is not constrained by relatedness may be more likely to reflect random language contact that often does not show such a systematic relationship with migration.

Put more starkly, this analysis therefore fails to support one key claim of this thesis, namely that linguistic areas discovered in the analysis of the World Phonotactics Database are formed by migration of people, at least so far as the analysis without clade constraints is concerned. However, these results are exploratory in any case, based on mtDNA data and similarity according to phonological/phonotactics data, and so should not be taken as refuting the general hypothesis that migration influences typological similarity; and as discussed, when clade constraints are used, then there is some support for some of the resulting linguistic areas (e.g. Northern Eurasia, the Caucasus) having been shaped by migration of people.

4.2.3 Mantel tests

One difference between previous quantitative studies and the analysis in the preceding two sections is that these studies use a Mantel test for the correlation between genetic distance (e.g. based on single nucleotide polymorphisms) and linguistic distance. A Mantel test computes the correlation coefficient between two matrices, by finding the correlation coefficient of the elements of the two matrices, and then using a permutation test where rows and columns are randomly shuffled; this randomisation procedure is used because the elements of a matrix are not independent, making the p-values of standard correlation coefficients inadequate (Dediu 2007:239, Mantel 1967). An intuitive way of thinking about this is that if one compares the linguistic and geographical distances between *N* languages, moving one of the languages would cause *N-1* distances to change. A correlation between the elements of the two matrices could therefore accidentally emerge due to the non-independence of these elements.

As a matter of fact, the measure of mtDNA genetic distance used in this chapter does not have this property: if there are two migrations from German to Dutch in the dataset, this is independent of the number of migrations from Dutch to German, or from German to any other language. It is therefore possible to treat the number of migrations between pairs of languages as independent, although the linguistic distances themselves are still non-independent.

One can therefore use a Mantel test to make the findings of this chapter more directly comparable to other quantitative work. A partial Mantel test can be used to test the correlation between linguistic distance and number of migrations, controlling for geographic distance.

For constrained linguistic distance, there is a strong correlation with geographic distance as one would expect (r=0.36, p<0.02), but also with number of migrations after controlling for geographic distance (r=-0.05, p<0.01: negative because the more migrations there are, the less the linguistic distance between two languages). When the proportion of migrations is made symmetrical (number of migrations between two languages divided by the total number of migrations involving either language), this correlation increases to r=-0.07. For unconstrained linguistic distance, there is a much higher correlation with geographic distance (r=0.64, p<0.02), and a small correlation again with migrations after controlling for geographic distance (r=-0.03, p<0.01).[11]

These results are similar to those obtained by Dediu (2007:275), which finds that the correlation between genetic distance and linguistic genealogical distance is small (r = 0.1041, p = 0.0308, and the correlation between genetic distance and typological distance

[11] The fact that unconstrained linguistic distance correlates more strongly with geography than constrained linguistic distance is slightly unexpected, given the impression that some of the clades discovered in the unconstrained phylogenetic analysis are not very geographically contiguous, such as a large clade covering almost the whole of Eurasia; or the inclusion of some Scandinavian languages such as Nynorsk in a clade of languages mostly in India. From another perspective, it is perhaps to be expected that if the unconstrained analysis is picking up on language contact, then this will correlate well with geographical distance (although it is not clear why this correlation would be stronger than that for language families).

is non-significant (p.244). One could conclude from this test, as Dediu does, that ‘the correlation between linguistic (family) distribution and genetics is mostly explained by geography, confirming previous conclusions’ (2007:275). But this does not mean that genetics can be ignored when explaining linguistic diversity. The main analysis in this chapter showed that a model which uses information from migrations has a higher likelihood than a model which does not, and helps to explain linguistic similarity both among languages which are known to be related (Maltese and Semitic languages, or varieties of Chinese), and also languages in northern Eurasia which are more similar phonologically than expected because there has been widespread migration between these places. A further difference between the main analysis and the Mantel test is that the main analysis uses a non-linear function to map geographic distance onto linguistic distance (a polynomial function which levels off at around 5000 km); even with a model that fits the function mapping from geographic distance to linguistic distance as closely as possible, genetic distance still turns out to make the predictive model more accurate.

One might ask whether improving the accuracy of a predictive model is the same as explanation (see for example Schmueli 2010). From some points of view, it is worth making a distinction between a correlation and a causal explanation, and there are statistical techniques such as causal graphs that can help disentangle the interactions of variables (Pearl 1995). This study cannot be said to have demonstrated a causal relationship between migration and linguistic similarity, in the sense that it has not included every other conceivable confounding variable and used a technique such as causal graphs; but most studies do not reach this high bar for a demonstration of causality, instead usually relying on testing an association between two variables, controlling for plausible confounds, and having a causal theory for why one variable would affect the other, such as independent evidence for a causal relationship between the two. In this case, it is perfectly plausible that the migration between two places can *cause* similarity between the languages of those two places; trivial examples include the migration of English to North America, or the migration of Japonic languages to the Ryukyuan islands, while more speculative but not implausible cases would include the movement of people bringing related languages, or the movement of people across communities resulting in structural convergence of languages. The association between migration and linguistic similarity was tested, controlling for the most obvious confound, that both variables are affected by geographical distance; in this sense, it is justified to say that the number of migrations helps to explain (not just predict) linguistic similarity between languages.

Conversely, it means that linguistics can help explain data from genetics which might otherwise seem to be without much pattern. The migration pathways reconstructed from mtDNA data in this chapter, for example, show a lot of Brownian motion, as one might expect. But there are also statistically supported, non-trivial patterns which emerge once this data is compared to languages. This includes movements between locations known to have related languages, such as mainland Japan and the Ryukyu islands, the Middle East and Malta, or Spain and Portugal; but also includes migrations across Northern Eurasia, which are predicted to have occurred despite the large distances involved, purely because of the phonological similarity between languages there. Although many papers have commented on patterns such as this (e.g. comparing the spread of mtDNA haplogroups in the Pacific to the spread of Austronesian languages, Kayser et al. 2008), this chapter provides a way of demonstrating these associations statistically, and factoring out trivial associations which arise purely because of the correlation of both genes and languages with geographical distance.

5. Conclusion

This chapter reconstructed mtDNA migrations using Bayesian phylogeography, and presented patterns of concerted migrations. Section 4 showed how these migrations contributed to making languages more similar, beyond what is expected from their geographical distance.

Future work on the relationship between migration and language should use whole genome data rather than just mtDNA data, particularly because of recent advances in computational methods for inferring genealogies from whole genomes (e.g. Kelleher et al. 2019, Wohns et al. 2022); and also include ancient DNA samples, which gives a fuller picture of genetic diversity that has not survived in modern populations (e.g. Haak et al. 2015).

The correlation between linguistic distance and migrations is not very strong, in line with previous studies, and as expected since language is rarely a barrier to genetic admixture (Pakendorf 2014). Many migrations will be therefore between languages without contributing to making these languages more similar. However, information from migrations does significantly improve models predicting the similarity of languages (when using linguistic distance with clade constraints), lending support to the demic hypothesis discussed elsewhere in this thesis (e.g. Chapter 1 section 3) that the distribution of structural linguistic features sheds light on patterns of concerted migration.

The following chapter summarises this idea in the context of the rest of the thesis: that linguistic areas can be identified (Southeast Asia, Northern Eurasia, and the Caucasus/Middle-East) from the phylogenetics of language structures, and because of the relationship between linguistic similarity and migration demonstrated in this chapter as well as in the fieldwork case study in Chapter 4, these linguistic areas are likely to have been partially formed by movement of people across language communities.

Chapter 6: Future Directions

1. Summary

This thesis focused on inferring linguistic areas and explaining why they exist by reference to migration patterns. Chapter 2 explores the history of a particular areal feature, tone, and attempts to explain why it clusters areally, arguing against a rival hypothesis that tone is influenced by climate (Everett et al. 2015). Chapter 3 uses a phylogenetic method to infer clusters of languages that share common history due to relatedness or language contact, using phonological data from the World Phonotactics Database (Donohue et al. 2013). Chapter 4 is a case study in language contact, showing how migrations of Tai speakers into areas of Palaungic languages may cause syntactic and semantic changes in these languages. Chapter 5 reconstructs migrations according to mitochondrial DNA data using Bayesian phylogeography, and provides some preliminary suggestions for why these are relevant to understanding the shape of linguistic areas by comparing these findings with the phylogenies inferred in Chapter 3. This chapter summarises these findings in more detail, and provides some areas for further research in conclusion.

2. Main points of the thesis

Chapter 2 discusses an example of how a single structural feature can be revealing about language history. Tone clusters in particular areas, showing that it is likely to have spread through families by language contact. In particular, it shows significant decreases in complexity from particular points, suggesting that it was simplified as it expanded. This is supported by a phylogenetic analysis of tonal language families, which lose tone as they move towards non-tonal language families. This was contrasted with an alternative line of explanation from Everett et al. (2015), that tone is influenced by humidity. Language contact is a confound for their hypothesis, as is demonstrated by simulations of the spread of tone from random locations. These simulations ends up producing correlations with humidity, using the same statistical tests and controls for genealogy, as much as 60-80% of the time. The correlation holds up in addition within two macro-areas 47% of the time, and in three macro-areas as much as 21% of the time, suggesting that the association between humidity and tone may be an artefact of language contact.

Chapter 3 shows that languages form clusters according to their phonological and phonotactic properties, and that these clusters can be detected by phylogenetics. These clusters include languages broadly covering southeast Asia; a smaller group of languages in south China; India and western southeast Asia; western and northern China; Europe and a large part of Siberia; eastern and southern India, mostly comprising Dravidian and Austroasiatic languages; and an area comprising the Caucasus, the Middle East, and northwest India.

If clade constraints are added, a simpler picture emerges of three main areas: southeast Asia, a Caucasus/Middle East area, and a large area covering the remainder of Eurasia. In either case, the general picture is that sounds convey information about history, in particular the relative isolation of southeast Asia from the rest of the continent (although with a lot of interaction with India); the massive amounts of movement within the rest of Eurasia; and the relative isolation of the Caucasus and Middle East.

The identification of linguistic areas may provide an explanatory framework for other disciplines such as genetics and archeology. This is especially true to the extent that the movement of linguistic features reflects the movement of people. This type of language contact is referred to in this thesis as demic diffusion, and the hypothesis that some types of language contact reliably indicate migration of people is referred to in some sections as the demic hypothesis (see Chapter 1 section 3).

A case study on how this demic process of language contact might work is in Chapter 4. The phonology, syntax and semantics of Palaungic languages in southwest China vary, and appears to have been influenced by Tai languages to differing degrees. Some languages show evidence of Tai contact, such as using Tai numerals, having the calque 'eye of the day', using subject-verb order rather than verb-subject order, and perhaps other features such as having more numeral classifiers and fewer ingestion verbs.

In keeping with demic hypothesis, there is a relationship between the distribution of Tai features in Palaungic languages, and the proportion of Tai speakers in each location where they are spoken. The calque 'eye of the day' shows the strongest relationship, with the average ratio of Tai speakers to Palaungic speakers being 5.59:1 in languages which have that calque, and 0.87:1 in languages without. A permutation test shows that the average ratio of Tai to Palaungic speakers in languages with that calque would be that high in only 1.4% of random samples, giving some quantitative support for the hypothesis that they are correlated. Similarly, Tai numerals are associated with high ratios of Tai speakers (p=0.037) and the use of subject-verb order (p=0.01).

This seems to vindicate the demic hypothesis on a local level, namely that some features can be a reliable predictor of the presence of speakers of a different language (and where they are in fact in a majority), and that these features often do not diffuse further into regions beyond that. The speakers of these languages interviewed are not bilingual between Palaungic and Tai-Kadai languages, and are also not aware of the properties that languages such as Tai have. The Tai-like properties are therefore instead tentatively suggested in the conclusion of that chapter to be due to past bilingualism, in which speakers who were more dominant in the Tai language that they adopted than in the language of their Palaungic community. A more detailed analysis of the social interactions between these communities is one direction for further research, as would be an analysis of the genetics of these populations, with the prediction being that there should have been more gene flow from Tai populations into Palaungic populations which speak languages that are more structurally similar to Tai.

Chapter 5 explored the demic hypothesis on a continental scale, constructing a dataset of mitochondrial DNA sequences and using a phylogeographic method to reconstruct ancestral movements of maternal lineages. It was found that there are statistically significant differences in the directions that people have tended to move in different regions, which may relate to the shape of linguistic areas. For example, there is a tendency for horizontal movement (i.e. moving along lines of latitude) across the Himalayas, for instance, which may be behind the linguistic areas crossing from India to Southeast Asia. Another linguistic area contained Indo-European and Turkic languages, an area which covers a large distance from east to west, in line with the tendency for there to be concerted horizontal movement in the region between Europe and northern China. Finally, the crescent of languages from the Middle East and the Caucasus to northwest India is also seen here, since there is a tendency for horizontal movement in this region.

There are also some striking areas where vertical movement (at a greater than 25 degree angle) has predominated. This is most true in Africa, where the angles of migrations have tended to be above 45 degrees, mirroring the southern expansion of Bantu languages. It is also true in island southeast Asia, probably due to the Austronesian expansion; and in China and southeast Asia, where several vertical clusters of languages were discovered in Chapter 3 based on similarity of phonology. This is plausibly associated with the expansion of particular language families, such as the southern movement of Tai-Kadai speakers out of China to the rest of southeast Asia (Enfield 2005).

A tendency for concerted movement in particular directions (east-west, or north-south) may be what drives the linguistic homogenisation of some regions. A mass migration southwards in Southeast Asia, for instance, may be detectable both from genetics as a statistical tendency to move in a particular direction (southwards, at a greater than 25 degree angle), and also from linguistics, from the similarity of languages in this region in their structural properties.

The association between migration of people and linguistic similarity is then tested in a quantitative way, by using measures of phonological similarity developed in Chapter 3; these measures of similarity are referred to as 'constrained' and 'unconstrained', depending on whether known language families were included as constraints. The number of migrations between locations helps to predict the similarity between two languages, even after taking their geographical distance into account. This is shown first by a Bayesian approach, comparing the likelihood of two models, one predicting linguistic similarity from geographic distance, and the other including a term predicted from the number of migrations between the two languages. The correlations are also tested using partial Mantel tests, with the result that the number of migrations shows a small but significant correlation with linguistic similarity after controlling for geographic distance. Geographic distance is the primary determinant of how similar languages are, but migrations are also an indispensable part of any explanation for why languages over large distances can end up being similar, and hence why large linguistic areas are formed.

3. Areas for further research

In many ways, the studies in this thesis are preliminary and offer clear paths for improvement and further research.

Chapter 2 raises issues that have to do with what non-linguistic factors can affect phonological properties such as tone. If it is true that tonal languages are more likely to be found in humid environments, or in populations that have a higher frequency of particular genetic variants, then this is a potential confound to using these properties to study language contact. In order to properly test these causal claims, it is best to study variation among speakers of a language, for instance, if it could be shown that speakers in dry environments really do find it harder to speak a tonal language than in humid environments.

As it stands, this type of evidence for their causal claims is lacking (the latter experiment has not been done for instance), and language contact is a confound to their claims. These methods for demonstrating this statistically are preliminary, and do not exploit the richness of the results of Chapter 3, which identifies structured linguistic areas rather than proxies for contact such as geographical distance. A particularly interesting possibility is using the phylogenetic trees discovered in Chapter 3 to conduct a phylogenetically controlled test of whether tone correlates with humidity or genetic variants (using the same technique as described in Dunn et al. 2011 for word order correlations).

The study in Chapter 3 on detecting linguistic areas could also be improved in some ways, such as taking into account functional dependencies between linguistic features. A more radical improvement would be to build a phylogenetic model which uses more than one tree. The mathematical framework for this method exists, having been described in Pagel and Meade (2004). Gray, Greenhill and Ross (2010) allude to this framework, adding that this method has not been applied to cultural data: 'We suggest that Bayesian phylogenetic mixture models (Pagel and Meade 2004) could be used to investigate complex histories at the level of specific characters. Instead of forcing all the characters onto a single tree, these mixture models allow different models of evolution to be applied to each character in the data. Essentially this allows the characters to "choose" between alternative trees. A multiple topology mixture model is currently implemented in Bayes Phylogenies (Pagel and Meade 2004) but, to the best of our knowledge, has not yet been used in studies of cultural evolution.' This remains true at the time of writing, and the application of a multiple topology model is particularly appropriate for structural data. A mixture model which also constrains trees to be correlated with each other would be especially effective, since features rarely have entirely independent transmission histories.

The case study in Chapter 4 presented fieldwork on syntactic and semantic structures. Although there is some quantitative support for the relationship between Tai structural features and the presence of Tai speakers, more features and languages would be needed for a proper statistical demonstration. It may also be good in future to conduct fieldwork on phonology, given the demonstration in Chapter 3 that phonological features are informative about history. More investigation would be needed on the demographics of Tai speakers, and genetics of these populations would help clarify how much Tai influence on Palaungic languages correlates Tai admixture with Palaungic speakers.

Chapter 5 attempted a quantitative study of the relationship between migration and linguistic similarity. While the analysis focused on mtDNA as an exercise in phylogeography, as a way to reconstruct the migrations of individuals, using whole genome data or SNPs is a better way of modelling the history of populations. More generally, using models of migrations with whole genome data and time calibrations from ancient DNA (such as from Haak et al. 2015 on the spread of Indo-European speakers) would be a better basis for comparing genetics to linguistics in the future.

A more ambitious project would therefore be the use of mixture models with correlated phylogenies to model language data, perhaps drawn from fieldwork, with a greater range of structural features (semantic and phonological in particular) and at a more fine-grained dialectal level; and alongside greater demographic detail, perhaps with genetic data and more sophisticated modelling of migration.

This thesis hopes to have demonstrated the rationale for pursuing these more sophisticated projects. Linguistic areas are statistically supported, geographically coherent clusters that emerge from phylogenetics of linguistic structures. Particular linguistic areas receive strong support in the analysis presented here, such as Southeast Asia, northern Eurasia, and the Caucasus/Middle-East, which like language families can provide frameworks for understanding the history of Eurasia. In particular, it shows the history of migration, as demonstrated on a local scale in the fieldwork study on how different features have been spreading with Tai speakers among the Palaungic languages, and on a continental scale in the correlation between migrations and linguistic similarity. Finally, the fieldwork study also showed the surprising diversity of linguistic structures among closely related dialects, especially in less explored domains such as semantics. More local, fine-grained linguistic studies, along with more detailed demographic and genetic data, will provide richer accounts of language contact and human prehistory.

Bibliography

Aikhenvald, A.Y. 2001. Areal diffusion, genetic inheritance, and problems of subgrouping: a North Arawak case study. In Aikhenvald, A.Y. ed. *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford: Oxford University Press. 167-194.

Ammerman, A.J., L.L.Cavalli-Sforza. 1984. *The Neolithic transition and the genetics of populations in Europe.* Princeton: Princeton University Press.

Anderson, G.D.S. 2006. Towards a typology of the Siberian linguistic area. In Matras, Y., A.McMahon, N.Vincent eds. *Linguistic Areas. Convergence in Historical and Typological Perspective*. London: Palgrave MacMillan.

Andrews, R.M., I.Kubacka, P.F.Chinnery, R.N.Lightowlers, D.M.Turnbull, N. Howell. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* 23, no. 2. 147.

Atkinson, Q. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346-349.

Bates, D., M.Maechler, B.Bolker, S.Walker, R.H.Bojesen Christensen, H.Singmann, B.Dai, F.Scheipl, G.Grothendieck, P.Green, J.Fox. 2015. lme4. R package in CRAN Repository.

Bates, D., M.Maechler, B.Bolker, S.Walker, R.H.Bojesen Christensen, H.Singmann, B.Dai, F.Scheipl, G.Grothendieck, P.Green, J.Fox. 2015. Geiger. R package in CRAN Repository.

Bellwood, P. 2013. *First migrants: ancient migration in global perspective.* Hoboken, New Jersey: John Wiley and Sons.

Benson, D.A, M.Cavanaugh, K.Clark, I.Karsch-Mizrachi, D.J.Lipman, J.Ostell, E.W.Sayers. 2013. GenBank. *Nucleic Acids Resolution*, 41. D36-42.

Bickel, B., J.Nichols. 2013. Autotyp. Project description online at http://www.autotyp.uzh.ch.

Bickel, B., J.Nichols. 2006. Oceania, the Pacific Rim and the theory of linguistic areas. *Annual Meeting of the Berkeley Linguistics Society*, vol. 32, no. 2. 3-15.

Bickel, B., J.Nichols. 2009. The geography of case. In Malchukov, A., A.Spencer. 2009. *The Oxford handbook of case*. Oxford: Oxford University Press. 479-493

Bickel, B. 2013. Distributional biases in language families. In Bickel, B., L-a.Grenoble, D.A.Peterson, A.Timberlake. 2013. *Language typology and historical contingency*. Amsterdam: John Benjamins Publishing Co. 415-444.

Bickel, B., J.Nichols, T.Zakharko, A.Witzlack-Makarevich, K.Hildebrandt, M.Rießler, L.Bierkandt, F.Zúñiga, J.B.Lowe. 2017. The AUTOTYP typological databases. Version 0.1.0. https://github.com/autotyp/autotyp-data/tree/0.1.0.

Birchall, J. 2014. *Argument realization in the languages of South America.* Nijmegen: Radboud University Nijmegen dissertation.

Blust, R. 2013. *The Austronesian Languages.* Asia-Pacific Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University. Available online at: https://openresearch-repository.anu.edu.au/handle/1885/10191

Bouckaert, R, P.Lemey, M.Dunn, S.J.Greenhill, A.V.Alekseyenko, A.J.Drummond, R.D.Gray, M.A.Suchard, and Q.D.Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337, no. 6097. 957-960.

Bouckaert, R., J.Heled, D.Kühnert, T.Vaughan, C-H.Wu, D.Xie, M.A.Suchard, A.Rambaut, A.J.Drummond. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4), e1003537. doi:10.1371/journal.pcbi.1003537.

Bouckaert, R. 2016. Phylogeography by diffusion on a sphere. *PeerJ* 4:e2406.

Brown, C.H., C.R.Clement, P.Epps, E.Luedeling, S.Wichmann. 2014. The Paleobiolinguistics of Maize. *Ethnobiology Letters* 5. 52-64.

Brown, S., P.E.Savage, A.Ko, M.Stoneking, Y-C.Ko, J-H.Loo, J.A.Trejaut. 2014. Correlations in the population structure of music, genes and language. *Proc. R. Soc. B* 281, no. 1774: 20132072.

Brownrigg, R. 2018. Maps. R package in CRAN Repository.

Burenhult, N. 2005. *A Grammar of Jahai.* Canberra: Pacific linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Campbell, L., T.Kaufman, T.C.Smith-Stark. 1986. Meso-America as a linguistic area. *Language* 62(3). 530-570.

Campbell, L. 2004. *Historical Linguistics (2nd Edition)*. Cambridge, MA: MIT Press.

Campbell, L. 2006. Areal linguistics: a closer scrutiny. In Matras, Y., A.McMahon, N.Vincent eds. *Linguistic Areas. Convergenge in Historical and Typological Perspective,* London: Palgrave MacMillan. 1–31.

Campbell, L. 2013. *Historical Linguistics (3rd edition)*. Edinburgh: Edinburgh University Press.

Casella, G., R.L.Berger. 2001 Hypothesis testing. In Smesler, N.J., P.B.Baltes eds. *The International Encyclopedia of the Social and Behavioral Sciences.* New York: Elsevier. 7115-7121.

Cavalli-Sforza, L.L, A.Piazza, P.Menozzi, J.Mountain. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* 85, no. 16. 6002-6006.

Cavalli-Sforza, L.L., E.Minch, J.L.Mountain. 1992. Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences* 89, no. 12. 5620-5624.

Chang, W., L.Michael. 2014. A relaxed admixture model of contact. *Language Dynamics and Change* 4(1). 1-26.

Chang, W., C.Cathcart, D.Hall, A.Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91, no. 1. 194-244.

Chikhi, L., R.A.Nichols, G.Barbujani, M.A.Beaumont. 2002. Y genetic data support the Neolithic demic diffusion model. *Proceedings of the National Academy of Sciences* 99, no. 17. 11008-11013.

Chirikba, V.A. 2008. The problem of the Caucasian Sprachbund. In Muysken, P. (ed.). *From linguistic areas to areal linguistics*. Vol. 90. Amsterdam: John Benjamins Publishing. 25-93.

Collins, J. 2017. Real and spurious correlations involving tonal languages. In: Enfield, N. J. (ed.) *Dependencies in language: On the causal ontology of linguistic systems.* (Studies in Diversity Linguistics 14). Berlin: Language Science Press. 129-139.

Creanza, N., M.Ruhlen, T.J.Pemberton, N.A.Rosenberg, M.W.Feldman, S.Ramachandran. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* 112, no. 5. 1265-1272.

Crevels, M., H.Van der Voort. 2008. The Guaporé-Mamoré region as a linguistic area. In: Muysken, P. (ed.). *From linguistic areas to areal linguistics.* Vol. 90. Amsterdam: John Benjamins Publishing. 151-179.

Daumé III, H.C. 2009. Non-parametric Bayesian areal linguistics. In Ostendorf, M., M.Collins, S.Narayanan, D.W.Oard, L.Vanderwende (eds.) *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* 593-601

Dahl, O. 2008. An exercise in 'a posteriori' language sampling. *Sprachtypologie und Universalienforschung* 61(3). 208–220.

Dawkins, R. 2004. *The ancestor's tale: a pilgrimage to the dawn of life.* London: Weidenfeld & Nicolson.

de Filippo, C., K.Bostoen, M.Stoneking, B.Pakendorf. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B* 279. 3256-3263.

Dediu, D. 2007. *Non-spurious correlations between genetic and linguistic diversities in the context of human evolution.* Edinburgh: PhD Thesis, University of Edinburgh.

Dediu, D. 2015. Language family classifications as Newick trees with branch length [database and software tool]. GitHub. https://github.com/ddediu/lgfam-newick.

Dediu, D., D.R.Ladd. 2007. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes. *Proceedings of the National Academy of Sciences*, 104. 10944-10949.

Dediu, D. (2011). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. Proceedings of the Royal Society of London/B, 278(1704). 474-479.

Dediu, D., S.C.Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLOS One*, 7(9), e45198. doi:10.1371/journal.pone.0045198.

Dediu, D., M.A.Cysouw. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLOS One*, 8: e55009. doi:10.1371/journal.pone.0055009.

Dediu, D, R.Janssen, S.R.Moisik. 2017. Language is not isolated from its wider environment: vocal tract influences on the evolution of speech and language. *Language & Communication* 54. 9-20.

Denham, T.P., S.G. Haberle, C.Lentfer, R.Fullagar, J.Field, M.Therin, N.Porch, B.Winsborough. 2003. Origins of agriculture at Kuk Swamp in the highlands of New Guinea. *Science* 11. 189-193.

de Roulet, E.D. 2018. The Sino-Korean influence on Middle Korean vowel harmony: a usage-based perspective. Buckeye East Asian Linguistics 4 (BEAL 4).

Diamond, J. 1997. *Guns, Germs, and Steel: The Fates of Human Societies.* New York: W.W. Norton & Company.

Diamond, J., P. Bellwood. 2003. Farmers and their languages: the first expansions. *Science* 300. 597-603.

Dixon, R.M.W. 2001. The Australian linguistic area. In: Aikhenvald, A.Y. ed. 2001. *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford: Oxford University Press. 64-104.

Donohue, M., R.Hetherington, J.McElvenny & V.Dawson. 2013. World phonotactics database. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au. Accessed 15/6/2015.

Drummond, A., S.Y.W.Ho, M.J.Phillips, A.Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4(5): e88.

Dryer, M.S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257-292.

Dryer, M.S. 1992. The Greenbergian word order correlations. *Language* 68. 81-138.

Dryer, M.S., M.Haspelmath (eds.). 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/116. Accessed on 2016-01-26.)

Dunn, M., A.Terrill, G.Reesink, R.A.Foley, S.C.Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743). 2072-2075.

Dunn, M. 2009. Contact and phylogeny in Island Melanesia. *Lingua* 119. 1664–1678.

Dunn, M., S. Greenhill, S.C.Levinson & R.D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word order universals. *Nature* 473. 79-82.

Endicott, P., S.Y.W Ho. 2008. A Bayesian evaluation of human mitochondrial substitution rates. *The American Journal of Human Genetics* 82, no. 4. 895-902.

Emeneau, M.B. 1956. India as a lingustic area. *Language* 32, no. 1. 3-16.

Enfield, N.J. 2005. Areal linguistics and Mainland Southeast Asia. *Annual Review of Anthropology* 34. 181-206.

Enfield, N.J. 2015. *Natural causes of language: Frames, biases, and cultural transmission.* Berlin: Language Science Press.

Everett, E., D.Blasi, S.G.Roberts. 2015. Climate, vocal folds and tonal languages: connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112. 1322-1327.

Felsenstein, J. 1978. The number of evolutionary trees. *Systematic Zoology* 27. 27-33.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17, no. 6. 368-376.

Forkel, R., M.List, S.J.Greenhill, C.Rzymski, S.Bank, M.Cysouw, H.Hammarström, M.Haspelmath, G.Kaiping, R.D.Gray. 2018. Cross-Linguistic Data Formats, advancing data-sharing and re-use in comparative linguistics. *Scientific Data* 5, no.180205.

Forster, P., C.Renfrew. Mother tongue and Y chromosomes. *Science* 333, no. 6048. 1390-1391.

Fuller, D.Q., L.Qin, Y.Zheng, Z.Zhao, X.Chen, L.A.Hosoya, G.-P.Sun. 2009. The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze. *Science* 323. 1607–1610 .

Gil, D. 2015. The Mekong-Mamberamo linguistic area. In Enfield, N.J, B. Comrie. *Languages of Mainland Southeast Asia: The state of the art.* Berlin: Mouton de Gruyter. 266-355.

Gilks, W.R., P.Wild. 1992. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* 41 (2). 337–348.

Goh, C.C.M. 1998. The level tone in Singapore English. *English Today* 14, no. 1. 50-53.

Gottfried, T.L., T.L.Suiter. 1997. Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics* 25. 207-231.

Gravel, A., D.Ablashi, L.Flamand. 2013. Complete genome sequence of early passaged human herpesvirus 6A (GS strain) isolated from North America. *Genome Announcements* 1, no. 3. e00012-13.

Gray, R.D., F.M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790). 1052-5.

Gray, R.D., Q.D.Atkinson. 2001. Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435-439

Gray, R.D., A.J.Drummond, S.J.Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323, no. 5913. 479-483.

Gray, R.D., S.J.Greenhill, R.M.Ross. 2010. The pleasures and perils of Darwinizing culture (with phylogenies). In Linquist, S. (ed.) *The Evolution of Culture*, vol.4. London: Routledge. 360-375.

Green, R.E., A.S.Malaspinas, J.Krause, A.W.Briggs, P.L.F.Johnson, C.Uhler, M.Meyer, J.M.Good, T.Maricic, U.Stenzel, K.Pruefer, M.Siebauer, H.A.Burbano, M.Ronan, J.M.Rothberg, M.Egholm, P.Rudan, D.Brajkovic, Z.Kucan, I.Gusic, M.Wikstrom, L.Laakkonen, J.Kelso, M.Slatkin, S.Paabo. 2008. A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134, issue 3. 416-426.

Greenberg, J. 1966. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Greenberg, J. (ed.) *Universals of Language.* London: MIT Press. 110-113.

Greenhill, S.J., C-H.Wu, X.Hua, M.Dunn, S.C.Levinson, R.D.Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114, no. 42. E8822-E8829.

Greenhill, S.J. 2021. Do languages and genes share cultural evolutionary history? *Science Advances* 7: eabm2472.

Grollemund, R., S.Branford, K.Bostoen, A.Meade, C.Venditti, M.Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Science*, 112. 13296–13301.

Gruen, B., F.Leisch. 2008. FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* 28(4). 1-35.

Güldemann, T., M.Stoneking. 2008. A historical appraisal of clicks: a linguistic and genetic population perspective. *Annual Review of Anthropology* 37. 93-109.

Güldemann, T. 2010. Sprachraum and geography: linguistic macro-areas in Africa. In Lameli, A., R.Kehrein, S.Rabanus (eds.). *Language and space: an international handbook of linguistic variation, volume 2: language mapping.* Berlin: Mouton de Gruyter. 561-585.

Gumperz, J.J., R.Wilson. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border. In Hymes, D. (ed.) *Pidginization and creolization of languages*. Cambridge: Cambridge University Press. 151-167.

Haak, W, I.Lazaridis, N.Patterson, N.Rohland, S.Mallick, B.Llamas, G.Brandt, S.Nordenfelt, E.Harney, K.Stewardson, Q.Fu. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, no. 7555. 207.

Hammarström, H., R.Forkel, M.Haspelmath, S.Bank. 2014. Glottolog 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hammarström, H., R.Forkel, M.Haspelmath, S.Bank. 2018. Glottolog 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hansen Edwards, J.G., M.L.Zampini, eds. 2008. *Phonology and second language acquisition.* Vol. 36. Amsterdam: John Benjamins Publishing.

Harries, L. 1964. The Arabs and Swahili Culture. *Africa* 34, no. 3. 224-229.

Hasegawa, M., H.Kishino, T-A.Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22, no. 2. 160-174.

Hashimoto, M. 1986. The Altaicization of Northern Chinese. In: McCoy, J., T.Light (eds.) *Contributions to Sino-Tibetan Studies.* Leiden: E.J.Brill. 76-97

Haspelmath, M. 2001. The European linguistic area: Standard Average European. In: Haspelmath, M., E.König, W.Oesterreicher, W.Raible (eds.). *Language Typology and Language Universals: An International Handbook*. Vol.2. Berlin: Mouton de Gruyter. 1492-1519

Haudricourt, A-G. 2018 (1954). The origin of tones in Vietnamese. HAL ID: halshs-01678018f.

Heine, B., D.Nurse, eds. 2008. *A Linguistic Geography of Africa*. Cambridge: Cambridge University Press.

Heyer, E., E.Zietkiewicz, A.Rochowski, V.Yotova, J.Puymirat, D. Labuda. 2001. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *The American Journal of Human Genetics* 69, no. 5. 1113-1126.

Hijmans, R.J, E.Williams, C.Vennes. 2015. Geosphere. R package in CRAN Repository.

Holman, E.W., C.H.Brown, S.Wichmann, A.Müller, V.Velupillai, H.Hammarström, S.Sauppe, H.Jung, D.Bakker, P.Brown, O.Belyaev. 2011. Automated dating of the world's language families based on lexical similarities. *Current Anthropology* 52 (6). 841-875.

Howell, N., C.Bogolin Smejkal, D.A.Mackey, P.F.Chinnery, D.M.Turnbull, C.Herrnstadt. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *The American Journal of Human Genetics* 72, no. 3. 659-670.

Huang, X., N.Kurata, Z.X.Wang, A.Wang, Q.Zhao, Y.Zhao, K.Liu, H.Lu, W.Li, Y.Guo, Y.Lu. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490. 497-501.

Huelsenbeck J.P., F.Ronquist, R.Nielsen, J.P.Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294. 2310-2314.

Hung, T.T.N. 2000. Towards a phonology of Hong Kong English. *World Englishes* 19(3). 337-356.

Huson, D.H., D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23. 254–267.

Ioannidis, A.G., J.Blanco-Portillo, K.Sandoval, E.Hagelberg, C.Barberena-Jonas, A.V.S.Hill, J.E.Rodríguez-Rodríguez, K.Fox, K.Robson, S.Haoa-Cardinali, C.D.Quinto-Cordés, J.F.Miquel-Poblete, K.Auckland, T.Parks, A.S.M.Sofro, M.C.Ávila-Arcos, A.Sockell, J.R.Homburger, C.Eng, C.Huntsman, E.G.Burchard, C.R.Gignoux, R.A.Verdugo, M.Moraga, C.D.Bustamante, A.J.Mentzer, A.Moreno-Estrada. 2021. Paths and timings of the peopling of Polynesia inferred from genomic networks. *Nature* 597: 522-526.

Ingman, M., H.Kaessmann, S.Pääbo, U.Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408, no. 6813. 708.

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

Jäger, G., S.Wichmann. 2016. Inferring the world tree of languages from word lists. In Roberts, S., C.Cluskey, L.McCrohon, L.Barceló-Coblijn, O.Feher, T.Verhoef. 2016. *Proceedings of the 11th International Conference on the Evolution of Language*.

Jäger, G. 2018. Global scale linguistic phylogenetic inference from lexical resources. *Scientific Data* 5: 180189.

Jaeger, T.F., P.Graff, W.Croft, D.Pontillo. 2011. Mixed effects models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15. 281-320/

Jenks, P., P.Pittayaporn. 2017. Kra-Dai Languages. Oxford Bibliographies. Available online at https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0178.xml. Accessed on 23/01/2022.

Jenny, M. 2015. The far west of Southeast Asia: 'Give' and 'get' in the languages of Myanmar. In Enfield, N.J., B.Comrie eds. *Languages of Mainland Southeast Asia: The state of the art*. Berlin: Mouton de Gruyter. 155-208.

Johanson, L., C.Bulut. Eds. 2006. *Turkic-Iranian Contact Areas. Historical and Linguistic Aspects.* Wiesbaden: Harrassowitz.

Jones, W. 1786. The Third Anniversary Discourse. In: Shore, J. (ed.). 1807. *The Works of Sir William Jones. With a Life of the Author.* Vol. III. John Stockdale and John Walker. 24–46.

Joseph, B.D. 1983. Language use in the Balkans: The contributions of historical linguistics. *Anthropological Linguistics* 25, no. 3. 275-287.

Joseph, B.D. 1983. The synchrony and diachrony of the Balkan infinitive. A study in areal, general and historical linguistics. *Cambridge Studies in Linguistics: Supplementary Volume*. Cambridge: Cambridge University Press.

Kalnay, E. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*. Available at http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/.Diagnostic/.above_ground/.

Kass, R.E., A.E.Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association,* 90 (430). 773-795.

Katz, H. 1975. *Generative Phonologie und phonologische Sprachbünde des Ostjakischen un Samojedischen.* Munich: Wilhelm Fink.

Kayser, M., Y.Choi, M.van Oven, S.Mona, S.Brauer, R.J.Trent, D.Suarkia, W. Schiefenhövel, M.Stoneking. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Molecular Biology and Evolution* 25, no. 7. 1362-1374.

Kelleher, J., Y.Wong, P.Albers, A.W.Wohns, G.McVean. 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics* 51: 1330-1338.

Kivisild, T. 2015. Maternal ancestry and population history from whole mitochondrial genomes. *Investigative Genetics* 6, no. 1: 3.

Klimov, G.A. 1978. Strukturnye obščnosti kavkazkix jazykov. [The structural affinities between Caucasian languages]. Moscow: Nauka.

Krause, J., Q.Fu, J.M.Good, B.Viola, M.V.Shunkov, A.P.Derevianko, S.Pääbo. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 464, no. 7290: 894.

Kruspe, N. 2004. *A grammar of Semelai.* Cambridge: Cambridge University Press.

Kumar, S., C.Bellis, M.Zlojutro, P.E.Melton, J.Blangero, J.E.Curran. *BMC Evolutionary Biology*, no. 11: 293.

Lanfear, R., H.Xia, D.L.Warren. 2016. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biology and Evolution* 8(8): 2319-2332.

LaPolla, R.J. 2010. Language contact and language change in the history of Sinitic languages. *Procedia Social and Behavioura Sciences* 2: 6858-6868.

Lemey, P., A.Rambaut, A.J.Drummond, M.A.Suchard. 2009. Bayesian phylogeography finds its roots. *PLoS Computational Biology* 5, e1000520.

Lick, J. 2018. Mthap. MTDNA haplogroup classification tool available at https://dna.jameslick.com/mthap/

List, J-M., R.Forkel. 2016. LingPy. A Python library for historical linguistics. Version 2.4. URL: http://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. With collaborations by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Simon Greenhill. Jena: Max Planck Institute for the Science of Human History.

Llamas, B., L.Fehren-Schmitz, G.Valverde, J.Soubrier, S.Mallick, N.Rohland, S.Nordenfelt, C.Valdiosera, S.M.Richards, A.Rohrlach, M.I.B.Romero. 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances* 2, no. 4: e1501385.

Longobardi, G., S.Ghirotto, C.Guardiano, F.Tassi, A.Benazzo, A.Ceolin, G.Barbujani. 2015. Across language families: Genome diversity mirrors linguistic variation within Europe. *American Journal of Physical Anthropology* 157, no. 4. 630-640.

Lu, H., J.Zhang, K.B.Liu, N.Wu, Y.Li, K.Zhou, M.Ye, T.Zhang, H.Zhang, X.Yang, L.Shen. 2009. Earliest domestication of common millet (Panicum miliaceum) in East Asia extended to 10,000 years ago. *Proceedings of the National Academy of Sciences* 106 (18). 7367–72.

Lupyan, G., R.Dale. 2010. Language structure is partly determined by social structure. *PLOS One* 5, no. 1: e8559.

Maddieson, I. 2013. Tone. In M. Dryer, M. Haspelmath, eds. World Atlas of Language Structures. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/13. Accessed on 2016-01-26.)

Madrigal, L., M.Melendez-Obando, R.Villegas-Palma, R.Barrantes, H.Raventos, R.Pereira, D.Luiselli, D.Pettener, G.Barbujani. 2012. High mitochondrial mutation rates estimated from deep-rooting costa rican pedigrees. *American Journal of Physical Anthropology* 148, no. 3. 327-333.

Magee, D., M.A.Suchard, M.Scotch. 2017. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. *PLoS Computational Biology* 13, no. 2: e1005389.

Majid, A., J.S. Boster, M.Bowerman. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition* 109, no. 2 (2008). 235-250.

Mallory, J.P., D.Q.Adams. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World.* Oxford: Oxford University Press.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27, no. 2 part 1. 209-220.

Masica, C.P. 1976. *Defining a Linguistic Area: South Asia.* Chicago: University of Chicago Press.

Masica, C.P. 2001. The definition and significance of linguistic areas: methods, pitfalls, and possibilities (with special reference to the validity of South Asia as a linguistic area). In Singh, R., P.Bhaskararao, K.V.Subbarao eds. *The Yearbook of South Asian Languages and Linguistics.* Tokyo symposium on South Asian languages: contact, convergence and typology. New Delhi: Sage.

Matisoff, J.A. 1999. Tibeto-Burman tonology in an areal context. In Kaji, S. ed. *Proceedings of the Symposium Cross-Linguistic Studies of Tonal Phenomena, Tonogenesis, Typology, and Related Topics.* Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies. 3-31.

Matisoff, J.A. 2015. Tibeto-Burman languages. *Encyclopedia Britannica*, 13th March 2015. Available online at https://www.britannica.com/topic/Tibeto-Burman-languages. Accessed 15 June 2022.

Matsumae, H., P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel. 2021. Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Science Advances* 7: abd9223.

Maurits, L.. R.Forkel, G.A.Kaiping, Q.D.Atkinson. 2017. BEASTLING: A software tool for linguistic phylogenetics using BEAST 2. PLoS ONE 12(8): e0180908

Maurits, L., R.Bouckaert, J.Heled, R.Gray, Q.Atkinson. Forthcoming. Structural phylogeography and the evolution of Eurasian languages. (Unpublished).

Mekel-Bobrov, N., D. Posthuma, S.L.Gilbert, P.Lind, M.Florencia Gosso, M.Luciano, S.E.Harris, T.C.Bates, T.J.C.Polderman, L.J.Whalley, H.Fox, J.M.Starr, P.D.Evans, G.W.Montgomery, C.Fernandes, P.Heutink, N.G.Martin, D.I.Boomsma, I.J.Deary, M.J. Wright, E.J.C. de Geus, Bruce T. Lahn. 2007. The ongoing adaptive evolution of ASPM and Microcephalin is not explained by increased intelligence. *Human Molecular Genetics* 16. 600–608.

Metropolis, N., A.W.Rosenbluth, M.N.Rosenbluth, A.H. Teller, E.Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21, no. 6. 1087-1092.

Maurer, P., S.Michaelis, M.Haspelmath, M. Huber. 2013. Atlas of Pidgin and Creole Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Miller, L. 1997. *Wasei eigo*: English 'loanwords' coined in Japan. In J.H.Hill, P.J.Mistry, L.Campbell (eds.) *The Life of Language: Papers in Linguistics in Honor of William Bright*. Mouton/De Gruyter. 123–139.

Molina, J., M.Sikora, N.Garud, J.M.Flowers, S.Rubinstein, A.Reynolds, P.Huang, S.Jackson, B.A.Schaal, C.D.Bustamante, A.R.Boyko. 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences* 108. 8351–8356.

Moran, S., D. McCloy, R. Wright. 2014. PHOIBLE Online. https://phoible.org.

Morey, S.. 2006. Constituent order change in the Tai languages of Assam. *Linguistic Typology* 10, no. 3. 327-367.

Muysken, P., N.Smith. 1995. The Study of Pidgin and Creole Languages. In Arends, J., P.Muysken, N.Smith eds. *Pidgins and Creoles: An Introduction*. Amsterdam: John Benjamins Publishing. 3-14.

Muysken, P. 2008. Introduction: Conceptual and methodological issues in areal linguistics. In Muysken, P. ed. *From Linguistic Areas to Areal Linguistics.* Amsterdam: John Benjamins Publishing. 1-23.

Muysken, P., H.Hammarström, J.Birchall, S.Danielsen, L. Eriksen, A.Vilacy Galucio, R.van Gijn, S.van de Kerke, V.Kolipakam, O.Krasnoukhova, N.Müller, L.O'Connor. 2014. The Languages of South America: Deep families, areal relationships, and language contact. In Muysken, P., L.O'Connor eds. *The Native Languages of South America: Origins, Development, Typology.* Cambridge: Cambridge University Press. 552-599.

Nebel, A., E.Landau-Tasseron, D.Filon, A.Oppenheim, M.Faerman. 2002. Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. *American Journal of Human Genetics* 70(6). 1594-1596.

Newman, J. 2009. A cross-linguistic overview of 'eat'and 'drink'. In: Newman, J. 2009. *The Linguistics of Eating and Drinking.* Amsterdam: John Benjamins. 1-26.

Ngai, S.S. 2015. On the origin of special numerals. In Chappell, H.M. ed. *Diversity in Sinitic languages*. Oxford: Oxford University Press. 190-225.

Nichols, J. 1992. *Linguistic Diversity in Space and Time.* Chicago: University of Chicago Press.

Oliver, J. 2008. Archeology of agriculture in ancient Amazonia. In Silverman, H., W.Isbell. ed. *Handbook of South American archaeology.* Berlin: Springer Science & Business Media. 185-216.

O'Reilly, J.E, P.C.J.Donogue. 2017. The Efficacy of Consensus Tree Methods for Summarizing Phylogenetic Relationships from a Posterior Sample of Trees Estimated from Morphological Data. *Systematic Biology* 67(2): 354–362.

Page, R.D.M, M.A.Charleston. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution* 13, no. 9. 356-359.

Pagel, M., A.Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 53, no. 4. 571-581.

Pagel, M., Q.D.Atkinson, A.Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449. 717-720.

Pagel, M., Q.D.Atkinson, A.S.Calude, A.Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences* 2013 110 (21). 8471-8476.

Pakendorf, B. 2007. Contact in the prehistory of the Sakha (Yakuts): linguistic and genetic perspectives. *LOT Dissertation Series.* LOT, Utrecht. Retrieved from https://hdl.handle.net/1887/12492

Pakendorf, B. 2014. Coevolution of languages and genes. *Current Opinion in Genetics & Development* 29. 39-44.

Paradis, E., S.Blomberg, B.Bolker, J.Brown, J.Claude, H.S.Cuong, R.Desper, G.Didier, B.Durand, J.Dutheil, R.J.Ewing, O.Gascuel, T.Guillerme, C.Heibl, A.Ives, B.Jones, F.Krah, D.Lawson, V.Lefort, P.Legendre, J.Lemon, E.Marcon, R.McCloskey, J.Nylander, R.Opgen-Rhein, A-A.Popescu, M.Royer-Carenzi, K.Schliep, K.Strimmer, D.de Vienne. Ape. R package in CRAN Repository.

Patterson Giersch, C. 2006. Asian Borderlands: The Transformation of Qing China's Yunnan Frontier. Harvard University Press.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82. 669–709.

Peretz, I., D.T.Vuvan. 2017. Prevalence of congenital amusia. *European Journal of Human Genetics* 25. 625-630.

Pinker, S. 1994. *The Language Instinct.* London: Penguin Publishing.

Pritchard, J.K., M.Stephens, P.Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155. 945-959.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rabel, L. 1958. *Khasi, a language of Assam.* Berkeley: Doctoral dissertation, University of California at Berkeley.

Reesink, G., R.Singer, M.Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7, no. 11: e1000241.

Revell, L.J. 2015. Phylogenetic tools for comparative biology (and other things). R package in CRAN Repository.

Rieux, A., A.Eriksson, M. Li, B.Sobkowiak, L.A.Weinert, V.Warmuth, A.Ruiz-Linares, A.Manica, F.Balloux. 2014. Improved calibration of the human mitochondrial clock using ancient genomes. *Molecular Biology and Evolution* 31, no. 10. 2780-2792.

Rischel, J. 1995. *Minor Mlabri: A Hunter-Gatherer Language of Northern Indochina.* Copenhagen: Museum Tusculanum Press.

Robbeets, M. 2005. *Is Japanese related to Korean, Tungusic, Mongolic and Turkic?* Wiesbaden: Harrassowitz.

Robinson, A. 2007. The Last Man Who Knew Everything: Thomas Young, the Anonymous Genius who Proved Newton Wrong and Deciphered the Rosetta Stone, among Other Surprising Feats. Penguin Publishing.

Ross, M. 1996. Contact-induced change and the comparative method: cases from Papua New Guinea. In Durie, M., M.Ross (eds.) *The Comparative Method Reviewed. Regularity and Irregularity in Language Change.* New York, Oxford: Oxford University Press. 180-217.

Rosser, Z.H., T.Zerjal, M.E.Hurles, M.Adojaan, D.Alavantic, A.Amorim, W.Amos, M.Armenteros, E.Arroyo, G.Barbujani, G.Beckman. 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *The American Journal of Human Genetics* 67, no. 6. 1526-1543.

Russell, T., F.Silva, J.Steele. 2014. Modelling the spread of farming in the Bantu-speaking regions of Africa: An archaeology-based phylogeography. *PLoS One* 9, e87854.

Rzymski, C., T.Tresoldi. 2019. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. DOI: 10.1038/s41597-019-0341-x. Available online at https://clics.clld.org/. Accessed on 28/01/2022.

Sagart, L., G.Jacques, Y.Lai, R.J.Ryder, V.Thouzeau, S.J.Greenhill, J-M.List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *PNAS* 116 (21). 10317-10322.

Sawyer, S. 2013. The high-coverage genome sequence of a Neandertal individual from Denisova cave in the Altai. Unpublished (submitted for publication in 2013). Sequence accession number KC879692 in GenBank, https://www.ncbi.nlm.nih.gov/nuccore/KC879692.

Schmidt, W. 1899. *Über das Verhälltniss der melanesischen Sprachen zu den polynesischen und untereinander. Sitzungsberichte der kaiserlichen Akademie der Wissenschaften, philosophish-historisch Classe*, vol. CXL. Vienna. [Citation from Blust 2013: 818.]

Schmueli, G. 2010. To explain or to predict? *Statistical Science* 25, no.3. 289-310.

Schönig, C. 2003. Turko-Mongolic relations. In: Janhunen, J. (ed.) *The Mongolic Languages.* London: Routledge. 403-419.

Schwarz, G.E. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6** (2). 461-464.

Séguy, I. 2019. Current trends in Roman demography, and empirical approaches to the dynamics of the *Limes* populations. In Verhagen, P., J.Joyce, M.Groenhuijzen (eds.). *Finding the limits of the Limes.* Springer, Cham.

Sherzer, J. 1973. Areal linguistics in North America. In: Sebeok, T.A. 1973. *Linguistics in North America*. Mouton, The Hague. 749-795.

Sidwell, P. 2011. Proto-Khasian and Khasi-Palaungic. *Journal of the Southeast Asian Linguistics Society* 4. 144-168.

Sidwell, P. 2015. Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies* (Notes, Reviews, Data-Papers) 44. lxviii-ccclvii.

Sidwell, P. 2015b. Phylogeny, innovations, and correlations in the prehistory of Austroasiatic. Paper presented at the workshop Integrating inferences about our past: new findings and current issues in the peopling of the Pacific and South East Asia, 22 June – 23 June 2015, Max Planck Institute for the Science of Human History, Jena, Germany.

Sigurðardottir, S., A.Helgason, J.R.Gulcher, K.Stefansson, P.Donnelly. 2000. The mutation rate in the human mtDNA control region. *The American Journal of Human Genetics* 66, no. 5. 1599-1609.

Silva, F., C.J.Stevens, A.Weisskopf, C.Castillo, L.Qin, A.Bevan, D.Q.Fuller. 2015. Modelling the geographical origin of rice cultivation in Asia using the rice archaeological database. *PLOS One*, https://doi.org/10.1371/journal.pone.0137024.

Skirgård, H., S.G.Roberts, L.Yencken. 2017. Why are some languages confused for others? Investigating data from the Great Language Game. *PLOS One* 12, no. 4: e0165934.

Simms, S.S. 2017. *Great lords of the sky: Burma's Shan aristocracy. Asian Highlands Perspectives*, 48.

Simonsen, M., T.Mailund, C.N.S.Pedersen. 2010. Inference of large phylogenies using neighbour-joining. In A.Fred, J.Filipe, H.Gamboa (eds) *Biomedical Engineering*

Systems and Technologies. BIOSTEC 2010. Communications in Computer and Information Science, 127.

Soares, P., D.Abrantes, T.Rito, N.Thomson, P.Radivojac, B.Li, V.Macaulay, D.C.Samuels, L.Pereira. 2013. Evaluating purifying selection in the mitochondrial DNA of various mammalian species. *PLOS One* 8, no. 3: e58993.

Sokal, R.R., C.D.Michener. 1958. *The University of Kansas Science Bulletin* 38. 1409-1438.

Sokal, R.R., N.L.Oden, B.A.Thomson. 1992. Origins of the Indo-Europeans: genetic evidence. *Proceedings of the National Academy of Sciences* 89, no. 16. 7669-7673.

Starostin, S. 1998. The Tower of Babel. Available online at http://starling.rinet.ru.

Stolz, T. 2006. All or nothing. In: In: Matras, Y., A.McMahon, N.Vincent (eds.), *Linguistic Areas. Convergenge in Historical and Typological Perspective*. London: Palgrave MacMillan.

Stuessy, T.F., C.König. 2008. Patrocladistic classification. *Taxon*. 57 (2). 594–601.

Swadesh, M. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452-463.

Tehrani, J.J. 2013. The phylogeny of little red riding hood. *PLOS One* 8, no. 11: e78871.

Thomason, S.G, T.Kaufman. 1992. *Language contact, creolization, and genetic linguistics.* Berkeley: University of California Press.

Thomason, S.G. 2000. Linguistic areas and language history. In Gilbers, D.G., J.Nerbonne, J.Schaeken. eds. *Languages in Contact.* Amsterdam: Brill. 311-317.

Thomason, S.G. 2001. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

Thompson, J.D., Higgins D.G., Gibson T.J. 1994. (November 1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 22. 4673–80.

Thurgood, G. 2003. A sub-grouping of Sino-Tibetan languages: the interaction between language contact, change and inheritance. In: Thurgood, G., R. LaPolla. 2003. *The Sino-Tibetan languages*. London: Routledge. 3-21

Trejaut, J.A., T.Kivisild, Loo J.H., Lee C.L., He C.L., Hsu C.Y., Lee Z.Y, M.Lin. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biology* 3(8): e247.

Trubetzkoy, N.S. 1928. Proposition 16. In: *Acts of the First International Congress of Linguists*. 17-18.

Urban, M. 2010. ‘Sun’='Eye of the Day': A linguistic pattern of Southeast Asia and Oceania. Oceanic Linguistics 49, no. 2 (2010): 568-579.Urban, Matthias. 'Sun'='Eye of the Day': A linguistic pattern of Southeast Asia and Oceania. *Oceanic Linguistics* 49, no. 2. 568-579.

van Gijn, R., P.Muysken. 2016. Linguistic Areas. Oxford Bibliographies.

Van Oven, M., M.Kayser. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* 30(2):E386-E394. http://www.phylotree.org.

Vianello D., S.F., G.Castellani, L.Lomartire, M.Capri, C.Franceschi. 2013. HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. *Human Mutation* 34(9). 1189-94.

Watkins, C. 2001. An Indo-European area and its characteristics: a challenge to the Comparative Method? In Aikhenvald, A.Y., R.M.W.Dixon eds. *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics.* Oxford: Oxford University Press.

Wichmann, S., E.W.Holmann, D.Bakker, C.Brown. 2010. Evaluating linguistic distance measures. *Physica A.* 389. 3632-3629.

Wichmann, S., E.W. Holman, C.H. Brown. 2016. The ASJP Database (version 17). Available online at http://asjp.clld.org.

Wohns, A.W., Y.Wong, B.Jeffery, A.Akbari, S.Mallick, R.Pinhasi, N.Patterson, D.Reich, J.Kelleher, G.McVean. 2022. A unified genealogy of modern and ancient genomes. *Science* 375, 6583: eabi8264.

Wong, P.C.M, B.Chandrasekaran, J.Zheng. 2012. The derived allele of ASPM is associated with lexical tone perception. *PLOS One* 7(4): e34243

Yunusbayev, B., M.Metspalu, E.Metspalu, A.Valeev, S.Litvinov, R.Valiev, V.Akhmetova, E.Balanovska, O.Balanovsky, S.Turdikulova, D.Dalimova, P.Nymadawa, A.Bahmanimehr, H.Sahakyan, K.Tambets, S.Fedorova, N.Barashkov, I.Khidiyatova, E.Mihailov, R.Khusainova, L.Damba, M.Derenko, B.Malyarchuk, L.Osipova, M.Voevoda, L.Yepiskoposyan, T.Kivisild, E.Khusnutdinova, R.Villems. 2015. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genetics* 2015 Apr 21;11(4):e1005068. doi: 10.1371/journal.pgen.1005068. PMID: 25898006; PMCID: PMC4405460.

Yun N., Y.Sun, I.Peretz. 2020. Congenital amusia in speakers of a tone language. *Brain*. 133: 9. 2635-2642.

Zavjalov, O. 1978. Some phonological aspects of the Dungan dialects. *Computational Analyses of Asian and African Languages* 9. 1-24.

Zhang, X., S.Liao, X.Qi, J.Liu, J.Kampuansai, H.Zhang, Z.Yang, B.Serey, T.Sovannary, L.Bunnath, H.S.Aun. 2015. Y-chromosome diversity suggests southern origin and Paleolithic backwave migration of Austro-Asiatic speakers from eastern Asia to the Indian subcontinent. *Scientific Reports* 5: 15486.

Zhang, M., S.Yan, W.Pan, L.Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* 569. 112–115.

Zhao, J.X. 1995. 东巴象形文 [*Dongba Xiangxingwen;* Dongba Ideograms]. Yunnan Renmin Chubanshe.

Zvelebil, M. 2000. The social context of the agricultural transition in Europe. Archaeogenetics: DNA and the population prehistory of Europe. 57-79.

Appendix 1: Supplementary Information

1.1 Material on GitHub

Data and code is available on GitHub, at https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials, aside from data taken from the World Phonotactics Database, which was publicly available at the time of writing but has since been taken offline. This data can be obtained by contacting the author requesting data.csv for Chapter 3 or text files for Chapter 2; see the Github repository for the latest status of the data availability.

The following section provides directories for individual chapters, and a summary of the files.

1.2 Chapter 2

The directory is at https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials/tree/master/Chapter%205/Agricultural%20Spreads%20and%20Humidity. The following files are included:

1.R: R script for analysing the relationship between tone and humidity, and phylogeographic analyses.

africatree.txt, asiatree.txt, glottologtree.txt: tree files for use in the phylogeographic analysis.

epicentersafrica.txt, epicentersasia.txt, epicentersmexico.txt, epicenterspapua.txt, epicentersworld.txt: data on number of tones in different regions (Africa, Asia, Mexico, Papua New Guinea, and globally) from the World Phonotactics Database. Please contact the author for access to these files, or see the Github repository for the latest status.

humidity.csv, humidityiso.csv: data on humidity.

1.3 Chapter 3

The directory is at https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials/tree/master/Chapter%202. The following files are included:

1.log, 1.tree, 1.trees, 1.xml, 1.xml.state: files relating to the run of the phylogenetic analysis of languages without clade constraints. These are the log file of states, the consensus tree, the trees, the xml file, and the xml.state files respectively.

2.log, 2.tree, 2.trees, 2.xml, 2.xml.state: files relating to the clade-constrained phylogenetic analysis.

data.csv: data from the World Phontactics Database (Donohue et al. 2013) in csv format: please contact the author for access to this file, or see the Github repository for the latest status.

nameeditor.py: a python script for editing names in tree files.

point_generator.py: a python script for generating maps from a set of coordinates.

template_beginning.xml: a xml script for use in Google Earth for generating maps.

additional_code/main.py: script for producing a dendrogram using the UPGMA algorithm

1.4 Chapter 4

The directory is at https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials/tree/master/Chapter%204. The following files are included:

analyses.txt: a transcription of data used in the chapter.

classifier_list.txt: a list of classifiers elicited in each language.

classifiers.kml: a kml map of number of classifiers in each language.

coordinates.txt: the coordinates of each language location.

demography.kml: a kml file of cities from which demographic data came.

demography_coordinates.txt: a table of languages and the nearest city for which there is demographic data.

demography_ratio.kml: a kml file of the ratio of Tai speakers to speakers of Palaungic languages.

demonstratives.kml: a kml file of demonstrative-noun order in each language.

eat_rice.kml: a kml file of 'eat rice' in each language.

eat_verbs.txt: a table summarising verbs relating to 'eat/drink' in each language.

eatverbs.png: a figure showing the geographical distribution of 'eat/drink' verb systems.

eatverbscoloured.numbers: a table of probabilities relating to usage of 'eat/drink' verbs.

eye_of_the_day.kml: a kml file of the word for 'sun' in each language.

glottolog.kml: a kml file of Glottolog names for each language.

template_beginning.xml: a template for maps in Google Earth.

additional_code/main.py: script for finding distance of languages to nearest towns

An additional repository https://github.com/JeremyCollinsMPI/palaungic-data-cldf contains files in cross-linguistic data format (CLDF 1.2, Forkel et al. 2018): see the README in this repository for an updated description of the files.

1.5 Chapter 5

The directory is at https://github.com/JeremyCollinsMPI/Thesis-Supplementary-Materials/tree/master/Chapter%203. The following files are included:

/Data and Preprocessing:

31485 documents.gb: mtDNA sequences for 31,845 individuals downloaded from GenBank (Benson et al. 2013).

all-aligned.tar.gz: a file showing aligned mtDNA sequences used in this chapter.

all.txt: a file of unaligned mtDNA sequences used in this chapter.

ancient.py, ancient.txt, ancient_check.py, ancient_check.txt, ancient_check_results.txt, ancient_taxa.txt: files relating to giving time calibrations to ancient DNA samples.

cambridge.py, cambridge.txt: files relating to partitioning sequences.

countries_new.txt: a file containing names of locations of sequences.

longconverter.py, longitude_hack.py: a file for preprocessing longitudes before the phylogeographic analysis.

mtdnaanalysis.py, pruner.py: a script for sampling sequences for the phylogenetic analysis and preparing the nexus file.

native_names.txt: names of Native American mtDNA sequences included.

/Run 1:

1.xml, 1.xml.state, output.tree, output.trees, output.kml: files relating to the time-calibrated analysis of mtDNA sequences.

ages.txt: time calibrations for sequences.

eurasianlanguages.kml: a kml file for languages in Eurasia.

kmleditor.py: a script for producing kml files used in the chapter.

southeastasianlanguages.kml: a kml file for languages in Southeast Asia.

/Run 2:

1.xml, 1.xml.state, beast.log, output.kml, output.tree, output.trees: files relating to the phylogeographic analysis of 2000 sequences.

migrations.r: R script for analysing migration routes.

migrations7.r: R script for predicting language similarity from number of migrations between languages.

movement_model.py: python script for processing migration route data.

movements2000.csv: csv file of reconstructed migration routes.

squares.kml, squares10.kml, squares11.kml, squares2.kml, squares3.kml, squares4.kml, squares5.kml , squares6.kml, squares7.kml, squares8.kml, squares9.kml: kml files showing tendencies in direction of migration in Eurasia and Africa.

additional_code/test.py: script for submitting fasta sequences to MtHap to find the haplogroups

1.6 Demographic Data for Chapter 4

This section gives data from a set of online sources, and the calculation to produce the ratio of Tai to Palaungic speakers.

Menglian

Source: http://webcache.googleusercontent.com/search?q=cache:kKhEKV36K4YJ:xxgk.yn.gov.cn/Z_M_011/Info_Detail.aspx%3FDocumentKeyID%3D2E22EBC4A6D84AB2B1461B5B34122A61+&cd=1&hl=en&ct=clnk&gl=hk

Original text: ‘民族与人口 孟连傣族拉祜族佤族自治县成立于1954年6月16日。县内居住着傣族、拉祜族、佤族、汉族、哈尼族、彝族、景颇族等民族。2010年全县总人口13.55万人，少数民族人口为10.71万人，占总人口的79.01%。其中傣族人口为2.56万人，占总人口的18.85%；拉祜族人口为3.81万人，占总人口的28.13%；佤族人口为2.75万人，占总人口的20.31%’

Summarised translation: In 2010 the total population was 135,500 people, of which 107,100 are ethnic minorities…of which Tai people number 25,600, and Wa number 27,500.

Ratio of Tai to Wa speakers: 25600/27500 = **0.93**

Ximeng

Source: https://zh.wikipedia.org/wiki/西盟佤族自治县

Original text: ‘县内居住着以佤族为主的24个少数民族，佤族、拉祜族和傣族为世居少数民族，少数民族占总人口的94%，其中：佤族人口占总人口的72%，分布在全县各乡（镇）；拉祜族占全县总人口的18%’

Summarised translation: Ethnic minorities make up 94% of the population, including Wa who make up 72% of the population, and Lahu (non-Tai) who make up 18% of the population.

Ratio of Tai to Wa speakers: the maximum number of Tai speakers would be 4% of the population (given that 94% of the population is made up of ethnic minorities and 90% is Wa and Lahu). Assuming the maximum number, then the ratio is 4%/72% = **0.06**.

Menghai

Source: https://baike.baidu.com/item/勐海县#5

Original text: ‘2007年，全县总人口73846户302943人，少数民族中，傣族117168人，占总人数的38.67%，哈尼族63231人，占20.87%，拉祜族39801人，占13.13%，布朗族32624人，占10.76%’

Summarised translation: There are 117,168 Tai people, and 32,624 Bulang people.

Ratio of Tai to Bulang speakers: 117168/32624 = **3.59**

Jinghong

Source: https://baike.baidu.com/item/%E6%99%AF%E6%B4%AA/400271?fromtitle=%E6%99%AF%E6%B4%AA%E5%B8%82&fromid=2278149#4

Original text: ‘2010年末，景洪市有12.23万户，总人口（户籍）39.92万人。其中，非农业人口16.2万人，占总人口的40.57%；少数民族人口28.01万人，占总人口的70.17%。居住有傣族、哈尼族、拉祜族、布朗族、彝族、基诺族、瑶族、壮族、回族、苗族等13个世居民族。主要少数民族人口：傣族13.85万人，占总人口的34.68%；哈尼族7.03万人，占总人口的17.62%；布朗族8230人，占总人口的2.06%；基诺族2.91万人，占总人口的0.55%。人口自然比2009年增长率为4.12%’

Summarised translation: There are 138,500 Tai people, and 8,230 Bulang people.

Ratio of Tai to Bulang speakers: 138500/8230 = **16.83**

Dehong

Source: https://zh.wikipedia.org/zh-hk/德宏傣族景颇族自治州#民族

Original text: 傣族 349840, 德昂族 14436, 佤族 1203

Summarised translation: Tai 349840, De’ang (Palaungic) 14436, Wa 1203

Ratio of Tai to De’ang and Wa speakers: 349840/(14436 + 1203) = **22.37**

Baoshan

Source: https://zh.wikipedia.org/zh-hk/保山市#民族

Original text: 傣族 43049, 布朗族 9834, 佤族 4833

Summarised translation: Tai 43049, Bulang 9834, Wa 4833

Ratio of Tai to Bulang and Wa speakers: 43049/(9834 + 4833) = **2.94**

Lincang

Source: https://zh.wikipedia.org/zh-hant/临沧市#民族

Original text: 傣族 114312, 佤族 235165, 布朗族 40434

Summarised translation: Tai 114312, Wa 235165, Bulang 40434

Ratio of Tai to Wa and Bulang speakers: 114312/(235165 + 40434) = **0.41**

Cangyuan

Source: https://baike.baidu.com/item/沧源佤族自治县

Original text: ‘2010全县辖6乡4镇，93个村民委员会，一个国营勐省农场，总人口18万人，少数民族人口占93.4%，佤族人口占总人口的85.1%’

Summarised translation: Ethnic minorities make up 93.4% of the population, and Wa people make up 85.1% of the population.

Ratio of Tai to Wa speakers: a maximum of 8.3% are Tai (93.4% - 95.1 %). Assuming this maximum number, then the ratio is 8.3%/85.1% = **0.1**

Pu’er

Source: https://zh.wikipedia.org/wiki/%E6%99%AE%E6%B4%B1%E5%B8%82#%E6%B0%91%E6%97%8F

Original text: 傣族 144117, 佤族 150164, 布朗族 15543

Summarised translation: Tai 144117, Wa 150164, Bulang 15543

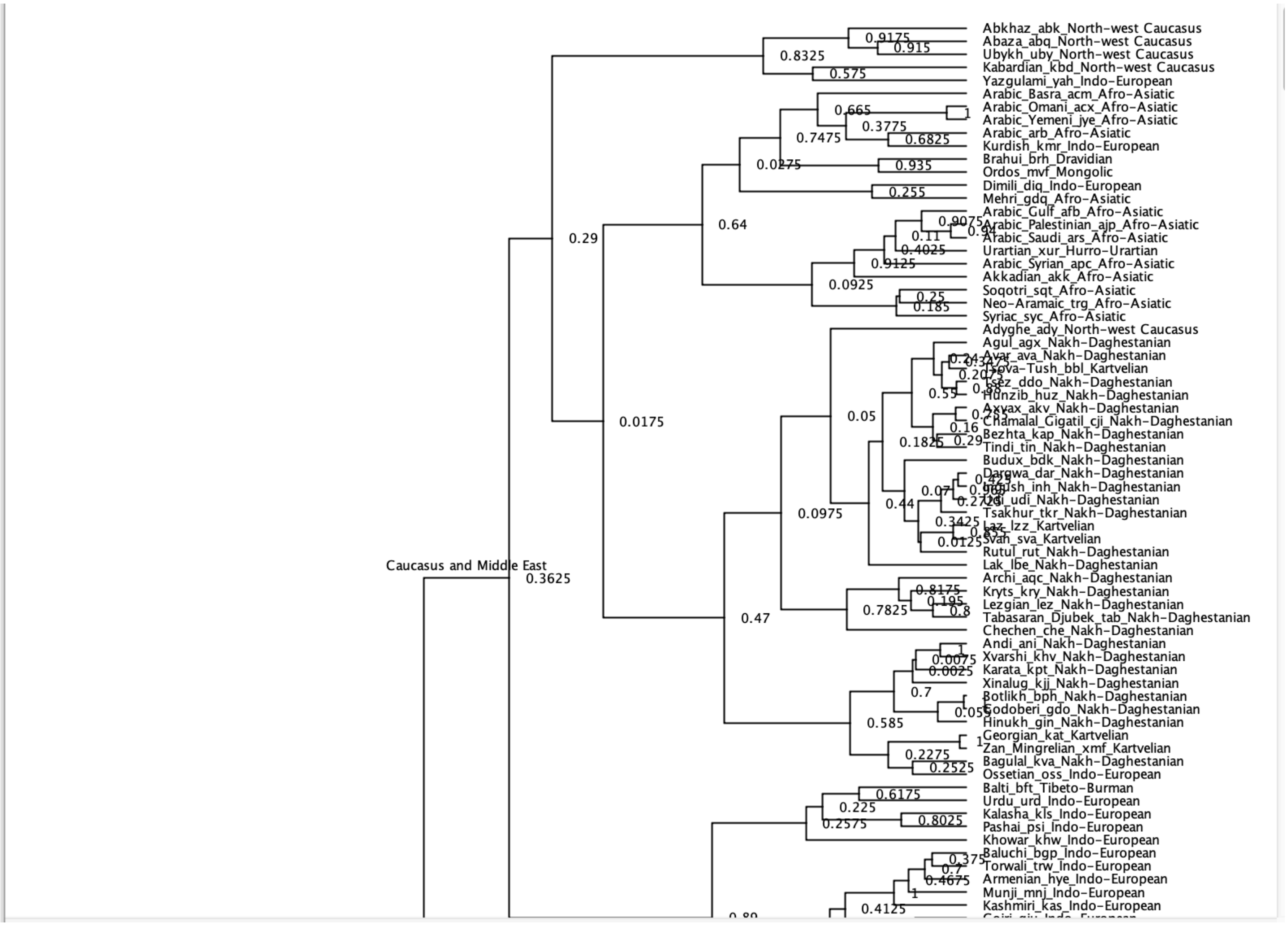
Ratio of Tai to Wa and Bulang speakers: 144117/(150164 + 15543) = **0.87**

Location	Nearest Town with Demographic Data	Distance (km)
Kunge	Jinghong	23
Bulangshan	Menghai	58
Zhanglang	Menghai	44
Bada	Menghai	53
Wengwa	Menglian	46
Manghong	Menghai	30
Mangjing	Ximeng	36
Menglian	Menglian	0
Gongxin	Menglian	27
Ximeng	Ximeng	0
Wengding	Cangyuan	15
Bangxie	Lincang	46
Xiaomenge	Lincang	44

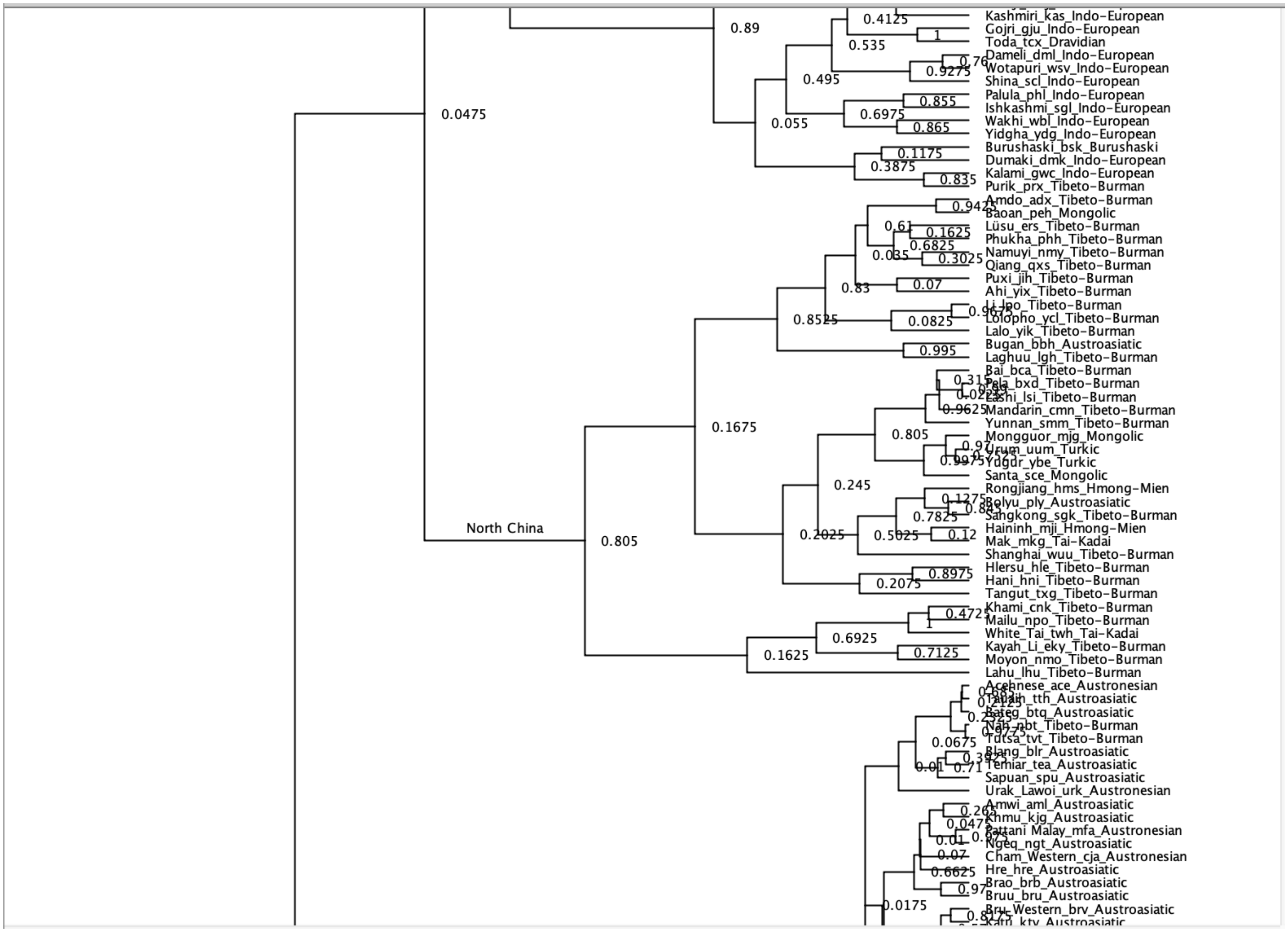
Table S1: Distances of locations to nearest towns with demographic data.

Appendix 2: Supplementary Figures

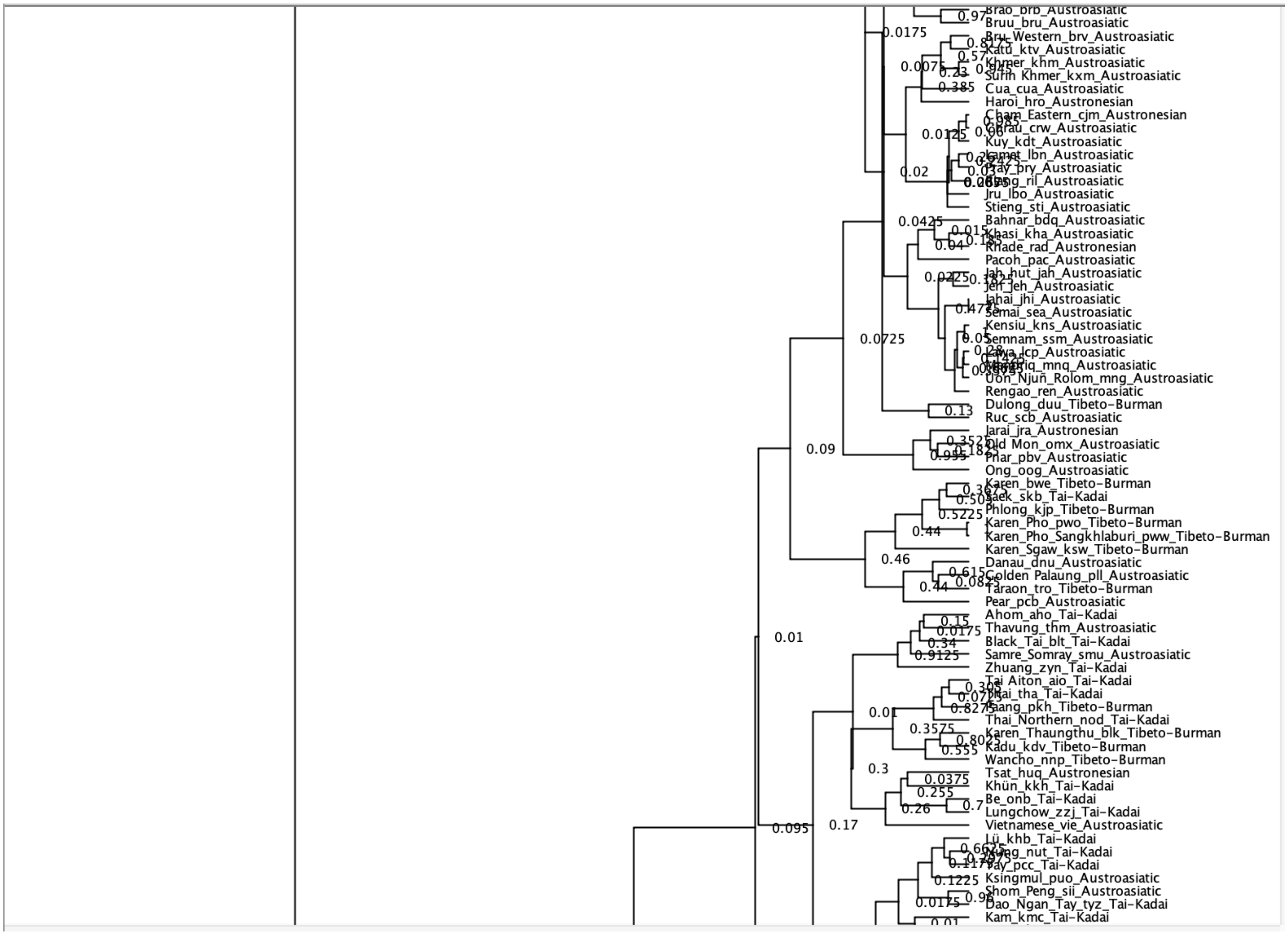
Supplementary Figure S1:1-11 Phonotactics tree using Bayesian inference


Abkhaz_abk_North-west Caucasus
Abaza_abq_North-west Caucasus
Ubykh_uby_North-west Caucasus
Kabardian_kbd_North-west Caucasus
Yazgulami_yah_Indo-European
Arabic_Basra_acm_Afro-Asiatic
Arabic_Omani_acx_Afro-Asiatic
Arabic_Yemeni_jye_Afro-Asiatic
Arabic_arb_Afro-Asiatic
Kurdish_kmr_Indo-European
Brahui_brh_Dravidian
Ordos_mvf_Mongolic
Dimili_diq_Indo-European
Mehri_gdq_Afro-Asiatic
Arabic_Gulf_afb_Afro-Asiatic
Arabic_Palestinian_ajp_Afro-Asiatic
Arabic_Saudi_ars_Afro-Asiatic
Urartian_xur_Hurro-Urartian
Arabic_Syrian_apc_Afro-Asiatic
Akkadian_akk_Afro-Asiatic
Soqotri_sqt_Afro-Asiatic
Neo-Aramaic_trg_Afro-Asiatic
Syriac_syc_Afro-Asiatic
Adyghe_ady_North-west Caucasus
Agul_agx_Nakh-Daghestanian
Avar_ava_Nakh-Daghestanian
Tsova-Tush_bbl_Kartvelian
Tsez_ddo_Nakh-Daghestanian
Hunzib_huz_Nakh-Daghestanian
Axvax_akv_Nakh-Daghestanian
Chamalal_Gigatil_cji_Nakh-Daghestanian
Bezhta_kap_Nakh-Daghestanian
Tindi_tin_Nakh-Daghestanian
Budux_bdk_Nakh-Daghestanian
Dargwa_dar_Nakh-Daghestanian
Ingush_inh_Nakh-Daghestanian
Udi_udi_Nakh-Daghestanian
Tsakhur_tkr_Nakh-Daghestanian
Laz_lzz_Kartvelian
Svan_sva_Kartvelian
Rutul_rut_Nakh-Daghestanian
Lak_lbe_Nakh-Daghestanian
Archi_aqc_Nakh-Daghestanian
Kryts_kry_Nakh-Daghestanian
Lezgian_lez_Nakh-Daghestanian
Tabasaran_Djubek_tab_Nakh-Daghestanian
Chechen_che_Nakh-Daghestanian
Andi_ani_Nakh-Daghestanian
Xvarshi_khv_Nakh-Daghestanian
Karata_kpt_Nakh-Daghestanian
Xinalug_kjj_Nakh-Daghestanian
Botlikh_bph_Nakh-Daghestanian
Godoberi_gdo_Nakh-Daghestanian
Hinukh_gin_Nakh-Daghestanian
Georgian_kat_Kartvelian
Zan_Mingrelian_xmf_Kartvelian
Bagulal_kva_Nakh-Daghestanian
Ossetian_oss_Indo-European
Balti_bft_Tibeto-Burman
Urdu_urd_Indo-European
Kalasha_kls_Indo-European
Pashai_psi_Indo-European
Khowar_khw_Indo-European
Baluchi_bgp_Indo-European
Torwali_trw_Indo-European
Armenian_hye_Indo-European
Munji_mnj_Indo-European
Kashmiri_kas_Indo-European
Gojri_gju_Indo-European
Caucasus and Middle East
0.9175
0.915
0.8325
0.575
0.665
1
0.3775
0.7475
0.6825
0.0275
0.935
0.255
0.29
0.64
0.9075
0.94
0.11
0.4025
0.9125
0.0925
0.25
0.185
0.3625
0.0175
0.24
0.3475
0.2075
0.55
0.88
0.755
0.05
0.16
0.1825
0.29
0.425
0.07
0.965
0.2725
0.44
0.0975
0.3425
0.855
0.0125
0.8175
0.195
0.7825
0.8
0.47
1
0.0075
0.0025
0.7
1
0.055
0.585
1
0.2275
0.2525
0.6175
0.225
0.2575
0.8025
0.375
0.7
0.4675
1
0.4125
0.89

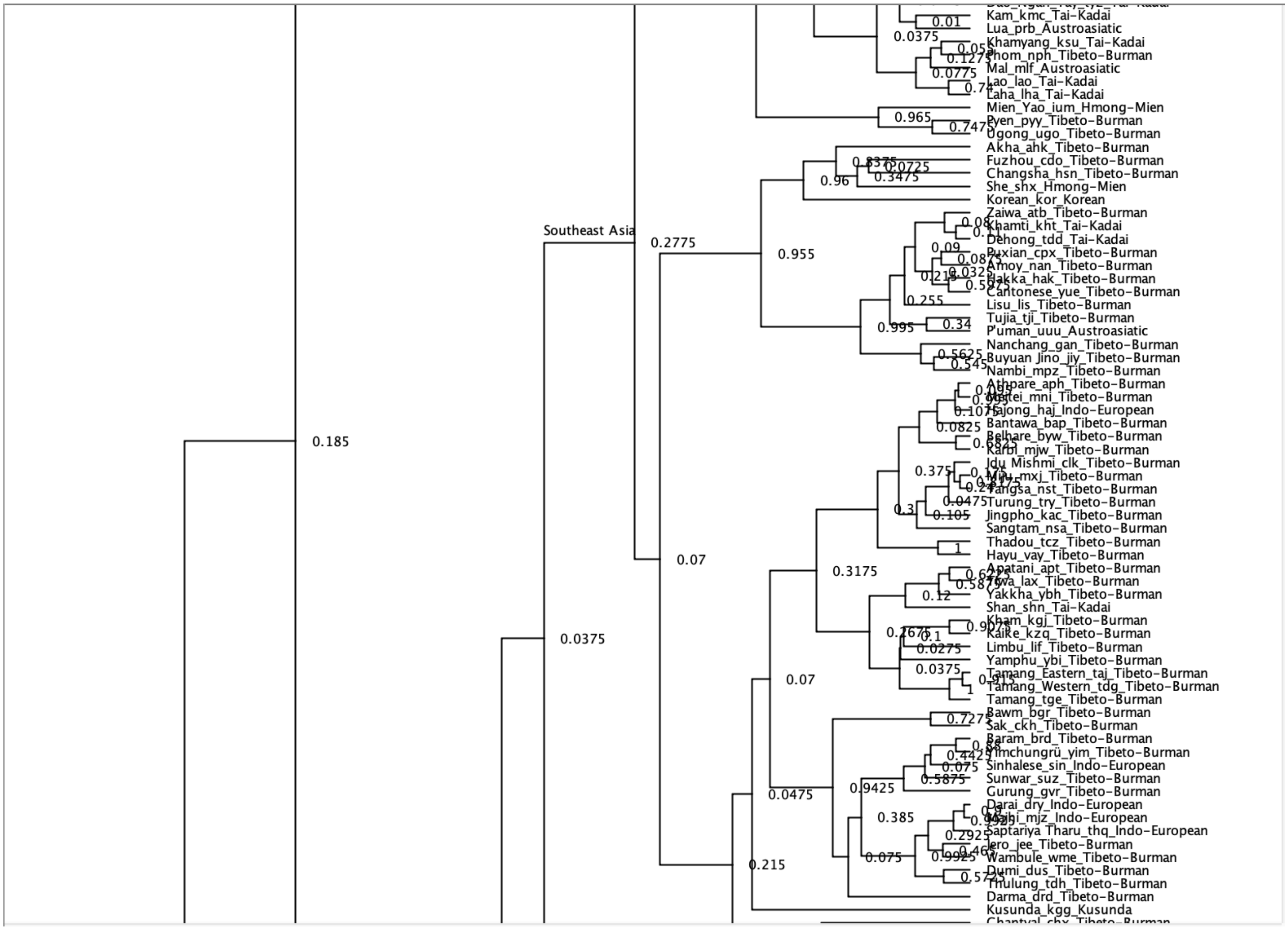
S1.1


0.89
0.0475
0.4125
Kashmiri_kas_Indo-European
Gojri_gju_Indo-European
1
Toda_tcx_Dravidian
0.535
Dameli_dml_Indo-European
0.76
Wotapuri_wsv_Indo-European
0.9275
0.495
Shina_scl_Indo-European
Palula_phl_Indo-European
0.855
Ishkashmi_sgl_Indo-European
0.6975
Wakhi_wbl_Indo-European
0.055
0.865
Yidgha_ydg_Indo-European
Burushaski_bsk_Burushaski
0.1175
Dumaki_dmk_Indo-European
0.3875
Kalami_gwc_Indo-European
0.835
Purik_prx_Tibeto-Burman
Amdo_adx_Tibeto-Burman
0.9425
Baoan_peh_Mongolic
0.61
Lüsu_ers_Tibeto-Burman
0.1625
Phukha_phh_Tibeto-Burman
0.6825
Namuyi_nmy_Tibeto-Burman
0.035
0.3025
Qiang_qxs_Tibeto-Burman
Puxi_jih_Tibeto-Burman
0.83
0.07
Ahi_yix_Tibeto-Burman
Li_lpo_Tibeto-Burman
0.9675
0.8525
0.0825
Lolopho_ycl_Tibeto-Burman
Lalo_yik_Tibeto-Burman
Bugan_bbh_Austroasiatic
0.995
Laghuu_lgh_Tibeto-Burman
Bai_bca_Tibeto-Burman
0.315
Pela_bxd_Tibeto-Burman
0.99
1
0.0225
Lashi_lsi_Tibeto-Burman
0.9625
Mandarin_cmn_Tibeto-Burman
0.1675
Yunnan_smm_Tibeto-Burman
0.805
Mongguor_mjg_Mongolic
0.97
Urum_uum_Turkic
0.7525
0.9975
Yugur_ybe_Turkic
Santa_sce_Mongolic
0.245
Rongjiang_hms_Hmong-Mien
0.1275
Bolyu_ply_Austroasiatic
0.845
0.7825
Sangkong_sgk_Tibeto-Burman
Haininh_mji_Hmong-Mien
North China
0.2025
0.5025
0.12
Mak_mkg_Tai-Kadai
0.805
Shanghai_wuu_Tibeto-Burman
Hlersu_hle_Tibeto-Burman
0.8975
0.2075
Hani_hni_Tibeto-Burman
Tangut_txg_Tibeto-Burman
Khami_cnk_Tibeto-Burman
0.4725
1
Mailu_npo_Tibeto-Burman
White_Tai_twh_Tai-Kadai
0.6925
Kayah_Li_eky_Tibeto-Burman
0.1625
0.7125
Moyon_nmo_Tibeto-Burman
Lahu_lhu_Tibeto-Burman
Acehnese_ace_Austronesian
0.685
Tauaih_tth_Austroasiatic
0.2125
Bateg_btq_Austroasiatic
0.2325
Nah_nbt_Tibeto-Burman
0.9775
0.0675
Tutsa_tvt_Tibeto-Burman
Blang_blr_Austroasiatic
0.3925
0.01
0.71
Temiar_tea_Austroasiatic
Sapuan_spu_Austroasiatic
Urak_Lawoi_urk_Austronesian
Amwi_aml_Austroasiatic
0.265
Khmu_kjg_Austroasiatic
0.0475
Pattani Malay_mfa_Austronesian
0.01
0.975
Ngeq_ngt_Austroasiatic
0.07
Cham_Western_cja_Austronesian
0.6625
Hre_hre_Austroasiatic
Brao_brb_Austroasiatic
0.97
Bruu_bru_Austroasiatic
0.0175
Bru_Western_brv_Austroasiatic
0.8175
Katu_ktv_Austroasiatic

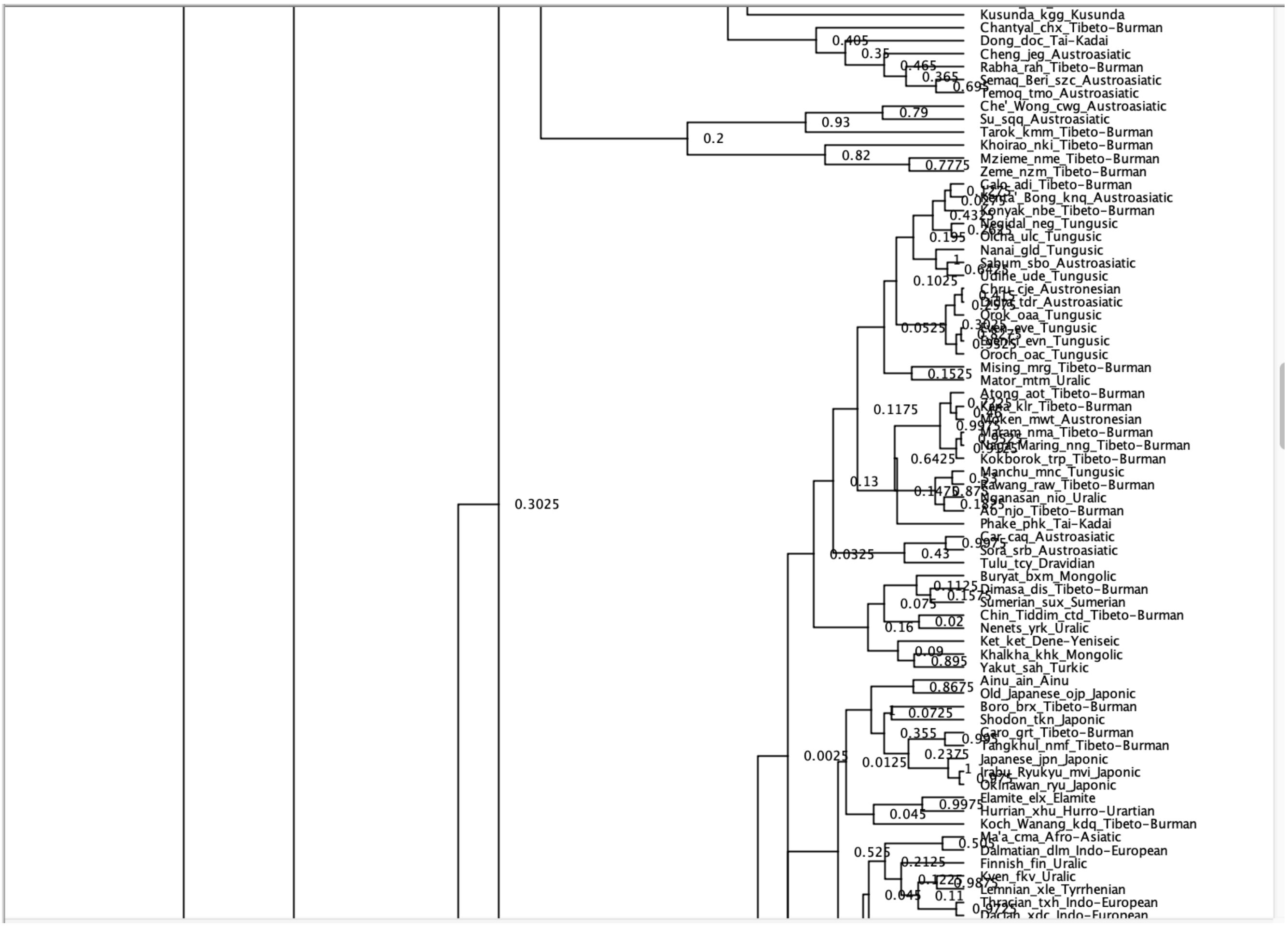
S1.2


Brao_brb_Austroasiatic
0.97
Bruu_bru_Austroasiatic
0.0175
Bru_Western_brv_Austroasiatic
0.8175
Katu_ktv_Austroasiatic
0.57
Khmer_khm_Austroasiatic
0.0075
0.23
0.945
Surin Khmer_kxm_Austroasiatic
0.385
Cua_cua_Austroasiatic
Haroi_hro_Austronesian
Cham_Eastern_cjm_Austronesian
0.985
Chrau_crw_Austroasiatic
0.0125
0.06
Kuy_kdt_Austroasiatic
Lamet_lbn_Austroasiatic
0.2425
Bray_bry_Austroasiatic
0.03
0.02
Riang_ril_Austroasiatic
Jru_lbo_Austroasiatic
Stieng_sti_Austroasiatic
0.0425
Bahnar_bdq_Austroasiatic
0.015
Khasi_kha_Austroasiatic
0.04
0.185
Rhade_rad_Austronesian
Pacoh_pac_Austroasiatic
0.0225
0.1825
Jah_hut_jah_Austroasiatic
Jeh_jeh_Austroasiatic
Jahai_jhi_Austroasiatic
0.4775
Semai_sea_Austroasiatic
Kensiu_kns_Austroasiatic
0.0725
0.05
Semnam_ssm_Austroasiatic
0.28
Lawa_lcp_Austroasiatic
0.1425
Mendriq_mnq_Austroasiatic
Uon_Njuñ_Rolom_mng_Austroasiatic
Rengao_ren_Austroasiatic
Dulong_duu_Tibeto-Burman
0.13
Ruc_scb_Austroasiatic
Jarai_jra_Austronesian
0.3525
0.09
Old Mon_omx_Austroasiatic
0.1825
0.955
Phar_pbv_Austroasiatic
Ong_oog_Austroasiatic
Karen_bwe_Tibeto-Burman
0.3675
0.505
Saek_skb_Tai-Kadai
Phlong_kjp_Tibeto-Burman
0.5225
Karen_Pho_pwo_Tibeto-Burman
0.44
Karen_Pho_Sangkhlaburi_pww_Tibeto-Burman
Karen_Sgaw_ksw_Tibeto-Burman
0.46
Danau_dnu_Austroasiatic
0.615
Golden Palaung_pll_Austroasiatic
0.0825
0.44
Taraon_tro_Tibeto-Burman
Pear_pcb_Austroasiatic
Ahom_aho_Tai-Kadai
0.15
0.0175
Thavung_thm_Austroasiatic
0.01
0.34
Black_Tai_blt_Tai-Kadai
0.9125
Samre_Somray_smu_Austroasiatic
Zhuang_zyn_Tai-Kadai
Tai Aiton_aio_Tai-Kadai
0.305
0.0725
Phai_tha_Tai-Kadai
0.8275
Saang_pkh_Tibeto-Burman
0.01
Thai_Northern_nod_Tai-Kadai
0.3575
Karen_Thaungthu_blk_Tibeto-Burman
0.8025
0.555
Kadu_kdv_Tibeto-Burman
Wancho_nnp_Tibeto-Burman
0.3
Tsat_huq_Austronesian
0.0375
Khün_kkh_Tai-Kadai
0.255
Be_onb_Tai-Kadai
0.26
0.7
Lungchow_zzj_Tai-Kadai
0.095
0.17
Vietnamese_vie_Austroasiatic
Lü_khb_Tai-Kadai
0.6625
Nung_nut_Tai-Kadai
0.7975
0.1175
Tay_pcc_Tai-Kadai
0.1225
Ksingmul_puo_Austroasiatic
Shom_Peng_sii_Austroasiatic
0.0175
0.98
Dao_Ngan_Tay_tyz_Tai-Kadai
Kam_kmc_Tai-Kadai
0.01

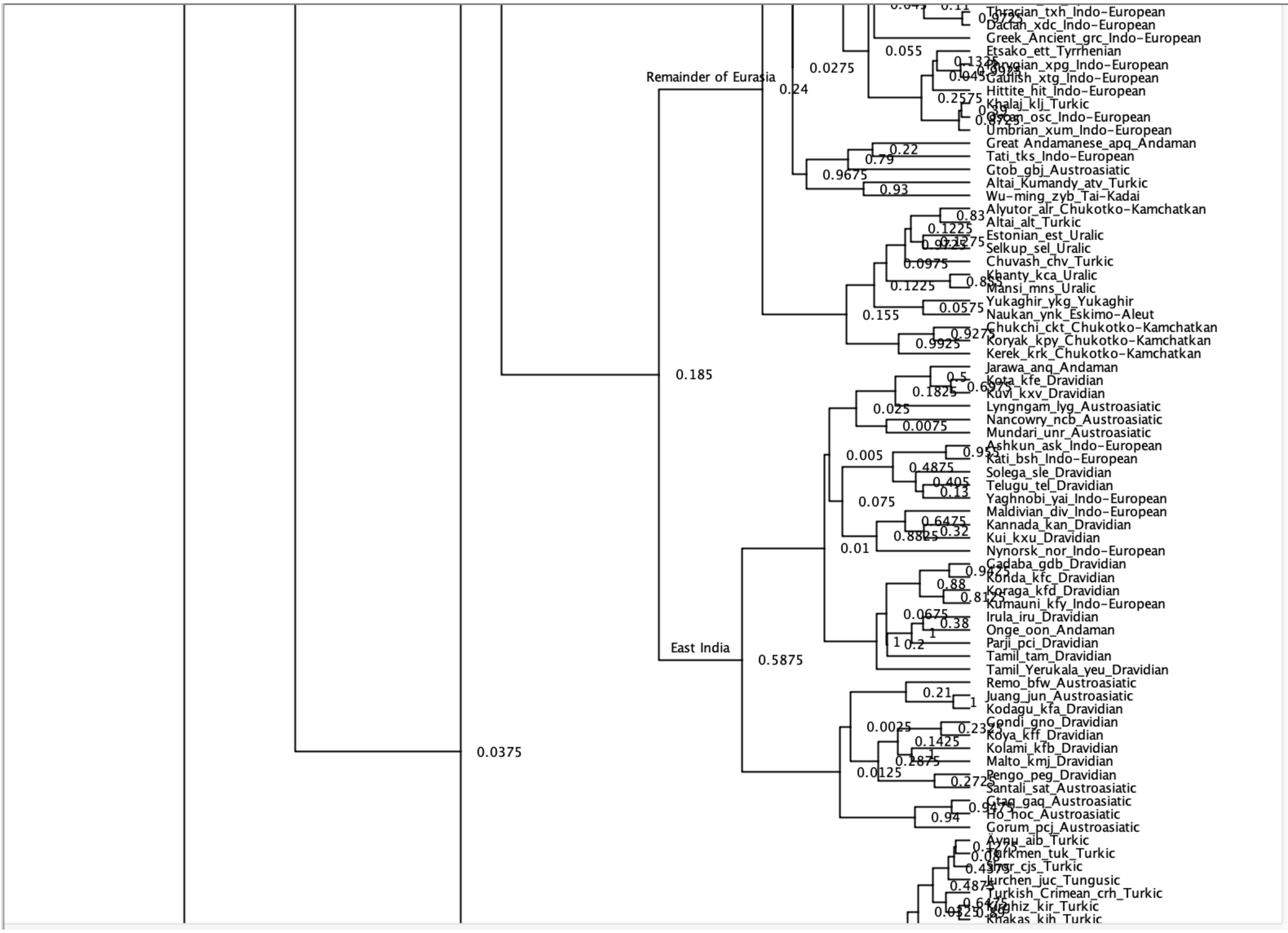
S1.3


Kam_kmc_Tai-Kadai
0.01
Lua_prb_Austroasiatic
0.0375
Khamyang_ksu_Tai-Kadai
0.055
Phom_nph_Tibeto-Burman
0.1275
Mal_mlf_Austroasiatic
0.0775
Lao_lao_Tai-Kadai
0.74
Laha_lha_Tai-Kadai
Mien_Yao_ium_Hmong-Mien
0.965
Pyen_pyy_Tibeto-Burman
0.7475
Ugong_ugo_Tibeto-Burman
Akha_ahk_Tibeto-Burman
0.8375
Fuzhou_cdo_Tibeto-Burman
0.0725
Changsha_hsn_Tibeto-Burman
0.3475
0.96
She_shx_Hmong-Mien
Korean_kor_Korean
Zaiwa_atb_Tibeto-Burman
0.08
Khamti_kht_Tai-Kadai
0.11
Dehong_tdd_Tai-Kadai
0.09
Puxian_cpx_Tibeto-Burman
0.0875
Amoy_nan_Tibeto-Burman
0.0325
0.215
Hakka_hak_Tibeto-Burman
0.5975
Cantonese_yue_Tibeto-Burman
0.255
Lisu_lis_Tibeto-Burman
Tujia_tji_Tibeto-Burman
0.34
0.995
P'uman_uuu_Austroasiatic
Nanchang_gan_Tibeto-Burman
0.5625
Buyuan Jino_jiy_Tibeto-Burman
0.545
Nambi_mpz_Tibeto-Burman
Southeast Asia
0.2775
0.955
0.185
Athpare_aph_Tibeto-Burman
0.095
Meitei_mni_Tibeto-Burman
0.995
0.1075
Hajong_haj_Indo-European
0.0825
Bantawa_bap_Tibeto-Burman
Belhare_byw_Tibeto-Burman
0.6825
Karbi_mjw_Tibeto-Burman
Idu Mishmi_clk_Tibeto-Burman
0.375
0.175
Miju_mxj_Tibeto-Burman
0.8175
0.24
Tangsa_nst_Tibeto-Burman
0.0475
Turung_try_Tibeto-Burman
0.3
0.105
Jingpho_kac_Tibeto-Burman
Sangtam_nsa_Tibeto-Burman
1
Thadou_tcz_Tibeto-Burman
Hayu_vay_Tibeto-Burman
0.07
0.3175
Apatani_apt_Tibeto-Burman
0.6225
Tiwa_lax_Tibeto-Burman
0.5875
0.12
Yakkha_ybh_Tibeto-Burman
Shan_shn_Tai-Kadai
0.0375
Kham_kgj_Tibeto-Burman
0.9075
0.2675
0.1
Kaike_kzq_Tibeto-Burman
0.0275
Limbu_lif_Tibeto-Burman
Yamphu_ybi_Tibeto-Burman
0.0375
0.07
Tamang_Eastern_taj_Tibeto-Burman
0.915
1
Tamang_Western_tdg_Tibeto-Burman
Tamang_tge_Tibeto-Burman
Bawm_bgr_Tibeto-Burman
0.7275
Sak_ckh_Tibeto-Burman
Baram_brd_Tibeto-Burman
0.88
0.4425
Yimchungrü_yim_Tibeto-Burman
0.075
Sinhalese_sin_Indo-European
0.0475
0.9425
0.5875
Sunwar_suz_Tibeto-Burman
Gurung_gvr_Tibeto-Burman
Darai_dry_Indo-European
0.9
0.385
0.9925
Majhi_mjz_Indo-European
0.2925
Saptariya Tharu_thq_Indo-European
Jero_jee_Tibeto-Burman
0.465
0.075
0.9925
Wambule_wme_Tibeto-Burman
Dumi_dus_Tibeto-Burman
0.5725
Thulung_tdh_Tibeto-Burman
0.215
Darma_drd_Tibeto-Burman
Kusunda_kgg_Kusunda
Chantyal_chx_Tibeto-Burman

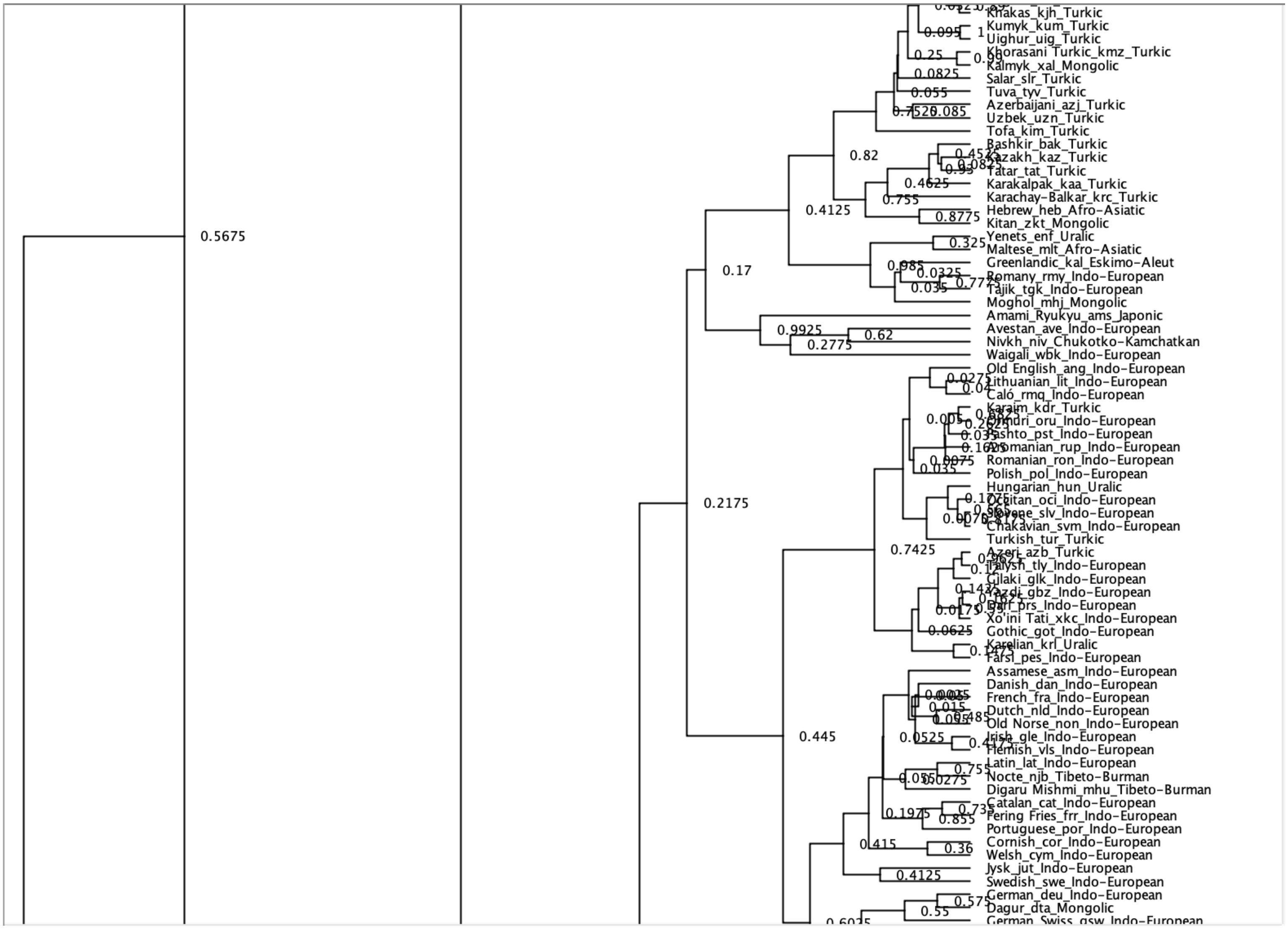
S1.4


0.3025
0.2
0.405
0.35
0.465
0.365
0.695
0.93
0.79
0.82
0.7775
Kusunda_kgg_Kusunda
Chantyal_chx_Tibeto-Burman
Dong_doc_Tai-Kadai
Cheng_jeg_Austroasiatic
Rabha_rah_Tibeto-Burman
Semaq_Beri_szc_Austroasiatic
Temoq_tmo_Austroasiatic
Che'_Wong_cwg_Austroasiatic
Su_sqq_Austroasiatic
Tarok_kmm_Tibeto-Burman
Khoirao_nki_Tibeto-Burman
Mzieme_nme_Tibeto-Burman
Zeme_nzm_Tibeto-Burman
0.1275
0.0275
0.4325
0.2625
0.195
Galo_adi_Tibeto-Burman
Kenta'_Bong_knq_Austroasiatic
Konyak_nbe_Tibeto-Burman
Negidal_neg_Tungusic
Olcha_ulc_Tungusic
1
0.6425
0.1025
Nanai_gld_Tungusic
Sabum_sbo_Austroasiatic
Udihe_ude_Tungusic
0.415
0.2975
0.0525
0.3025
0.8275
0.9525
Chru_cje_Austronesian
Didra_tdr_Austroasiatic
Orok_oaa_Tungusic
Even_eve_Tungusic
Evenki_evn_Tungusic
Oroch_oac_Tungusic
0.1525
Mising_mrg_Tibeto-Burman
Mator_mtm_Uralic
0.1175
0.7225
0.46
0.9975
0.9525
0.9925
0.6425
Atong_aot_Tibeto-Burman
Kena_klr_Tibeto-Burman
Moken_mwt_Austronesian
Maram_nma_Tibeto-Burman
Naga_Maring_nng_Tibeto-Burman
Kokborok_trp_Tibeto-Burman
0.13
0.53
0.1475
0.875
0.1825
Manchu_mnc_Tungusic
Rawang_raw_Tibeto-Burman
Nganasan_nio_Uralic
Ao_njo_Tibeto-Burman
Phake_phk_Tai-Kadai
0.9975
0.0325
0.43
Car_caq_Austroasiatic
Sora_srb_Austroasiatic
Tulu_tcy_Dravidian
0.1125
0.1575
0.075
0.02
0.16
Buryat_bxm_Mongolic
Dimasa_dis_Tibeto-Burman
Sumerian_sux_Sumerian
Chin_Tiddim_ctd_Tibeto-Burman
Nenets_yrk_Uralic
0.09
0.895
Ket_ket_Dene-Yeniseic
Khalkha_khk_Mongolic
Yakut_sah_Turkic
0.8675
Ainu_ain_Ainu
Old_Japanese_ojp_Japonic
1
0.0725
0.355
0.995
0.0025
0.2375
0.0125
1
0.975
Boro_brx_Tibeto-Burman
Shodon_tkn_Japonic
Garo_grt_Tibeto-Burman
Tangkhul_nmf_Tibeto-Burman
Japanese_jpn_Japonic
Irabu_Ryukyu_mvi_Japonic
Okinawan_ryu_Japonic
0.9975
0.045
Elamite_elx_Elamite
Hurrian_xhu_Hurro-Urartian
Koch_Wanang_kdq_Tibeto-Burman
0.505
0.525
0.2125
0.1225
0.9875
0.045
0.11
0.9725
Ma'a_cma_Afro-Asiatic
Dalmatian_dlm_Indo-European
Finnish_fin_Uralic
Kven_fkv_Uralic
Lemnian_xle_Tyrrhenian
Thracian_txh_Indo-European
Dacian_xdc_Indo-European

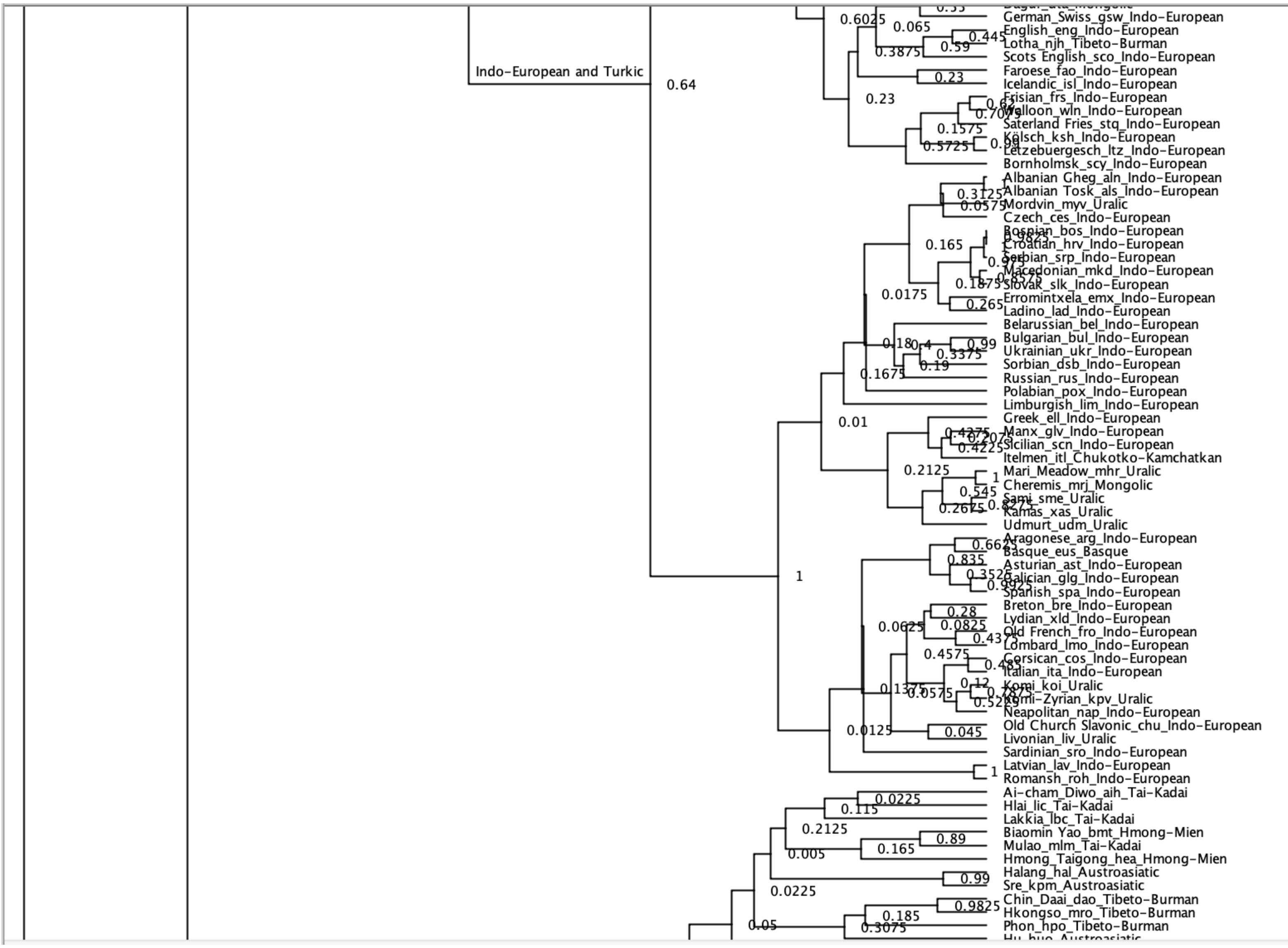
S1.5


0.045
0.11
Thracian_txh_Indo-European
0.9725
Dacian_xdc_Indo-European
Greek_Ancient_grc_Indo-European
0.055
Etsako_ett_Tyrrhenian
0.0275
0.1325
Phrygian_xpg_Indo-European
0.9925
0.045
Gaulish_xtg_Indo-European
Remainder of Eurasia
0.24
Hittite_hit_Indo-European
0.2575
Khalaj_klj_Turkic
0.39
Oscan_osc_Indo-European
0.8725
Umbrian_xum_Indo-European
Great Andamanese_apq_Andaman
0.22
Tati_tks_Indo-European
0.79
Gtob_gbj_Austroasiatic
0.9675
Altai_Kumandy_atv_Turkic
0.93
Wu-ming_zyb_Tai-Kadai
Alyutor_alr_Chukotko-Kamchatkan
0.83
Altai_alt_Turkic
0.1225
Estonian_est_Uralic
0.1275
0.9725
Selkup_sel_Uralic
Chuvash_chv_Turkic
0.0975
Khanty_kca_Uralic
0.855
0.1225
Mansi_mns_Uralic
Yukaghir_ykg_Yukaghir
0.0575
0.155
Naukan_ynk_Eskimo-Aleut
Chukchi_ckt_Chukotko-Kamchatkan
0.9275
Koryak_kpy_Chukotko-Kamchatkan
0.9925
Kerek_krk_Chukotko-Kamchatkan
0.185
Jarawa_anq_Andaman
0.5
Kota_kfe_Dravidian
0.6975
0.1825
Kuvi_kxv_Dravidian
Lyngngam_lyg_Austroasiatic
0.025
Nancowry_ncb_Austroasiatic
0.0075
Mundari_unr_Austroasiatic
Ashkun_ask_Indo-European
0.005
0.955
Kati_bsh_Indo-European
0.4875
Solega_sle_Dravidian
0.405
Telugu_tel_Dravidian
0.13
0.075
Yaghnobi_yai_Indo-European
Maldivian_div_Indo-European
0.6475
Kannada_kan_Dravidian
0.32
0.8825
Kui_kxu_Dravidian
0.01
Nynorsk_nor_Indo-European
Gadaba_gdb_Dravidian
0.9425
Konda_kfc_Dravidian
0.88
Koraga_kfd_Dravidian
0.8125
Kumauni_kfy_Indo-European
0.0675
Irula_iru_Dravidian
0.38
1
Onge_oon_Andaman
1
0.2
Parji_pci_Dravidian
East India
Tamil_tam_Dravidian
0.5875
Tamil_Yerukala_yeu_Dravidian
Remo_bfw_Austroasiatic
0.21
Juang_jun_Austroasiatic
1
Kodagu_kfa_Dravidian
0.0025
Gondi_gno_Dravidian
0.2325
Koya_kff_Dravidian
0.1425
Kolami_kfb_Dravidian
1
0.0375
0.2875
Malto_kmj_Dravidian
0.0125
Pengo_peg_Dravidian
0.2725
Santali_sat_Austroasiatic
Gtaq_gaq_Austroasiatic
0.9475
Ho_hoc_Austroasiatic
0.94
Gorum_pcj_Austroasiatic
Äynu_aib_Turkic
0.1275
Turkmen_tuk_Turkic
0.08
Shor_cjs_Turkic
0.4375
Jurchen_juc_Tungusic
0.4875
Turkish_Crimean_crh_Turkic
0.6475
Kirghiz_kir_Turkic
0.0325
0.89
Khakas_kjh_Turkic

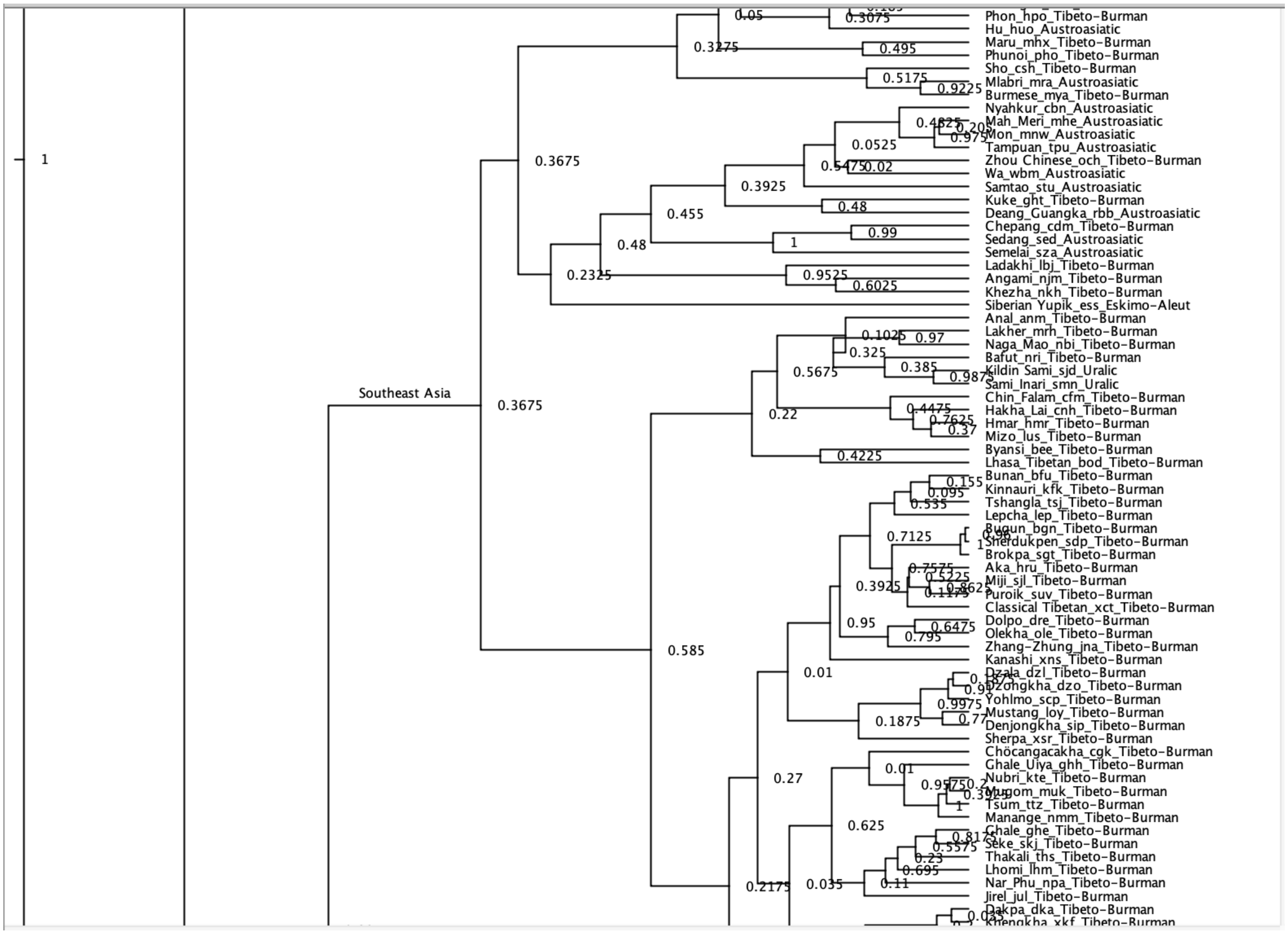
S1.6


0.5675
0.17
0.2175
0.445
0.82
0.4125
0.7425
0.6025
Khakas_kjh_Turkic
0.095
1
Kumyk_kum_Turkic
Uighur_uig_Turkic
0.25
0.99
Khorasani Turkic_kmz_Turkic
Kalmyk_xal_Mongolic
0.0825
Salar_slr_Turkic
0.055
Tuva_tyv_Turkic
0.7525
0.085
Azerbaijani_azj_Turkic
Uzbek_uzn_Turkic
Tofa_kim_Turkic
0.4525
0.0825
0.95
Bashkir_bak_Turkic
Kazakh_kaz_Turkic
Tatar_tat_Turkic
0.4625
Karakalpak_kaa_Turkic
0.755
Karachay-Balkar_krc_Turkic
0.8775
Hebrew_heb_Afro-Asiatic
Kitan_zkt_Mongolic
0.325
Yenets_enf_Uralic
Maltese_mlt_Afro-Asiatic
0.985
Greenlandic_kal_Eskimo-Aleut
0.0325
0.7775
Romany_rmy_Indo-European
0.035
Tajik_tgk_Indo-European
Moghol_mhj_Mongolic
Amami_Ryukyu_ams_Japonic
0.9925
0.62
Avestan_ave_Indo-European
0.2775
Nivkh_niv_Chukotko-Kamchatkan
Waigali_wbk_Indo-European
Old English_ang_Indo-European
0.0275
0.04
Lithuanian_lit_Indo-European
Caló_rmq_Indo-European
0.6825
Karaim_kdr_Turkic
0.005
0.2625
Ormuri_oru_Indo-European
0.035
Pashto_pst_Indo-European
0.1625
Aromanian_rup_Indo-European
0.0075
Romanian_ron_Indo-European
0.035
Polish_pol_Indo-European
Hungarian_hun_Uralic
0.1775
Occitan_oci_Indo-European
0.565
0.0075
0.8175
Slovene_slv_Indo-European
Chakavian_svm_Indo-European
Turkish_tur_Turkic
0.9625
Azeri_azb_Turkic
0.12
Talysh_tly_Indo-European
Gilaki_glk_Indo-European
0.1425
Yazdi_gbz_Indo-European
0.1625
0.0175
0.55
Dari_prs_Indo-European
Xo'ini Tati_xkc_Indo-European
0.0625
Gothic_got_Indo-European
0.1475
Karelian_krl_Uralic
Farsi_pes_Indo-European
Assamese_asm_Indo-European
Danish_dan_Indo-European
0.0025
0.005
French_fra_Indo-European
0.015
Dutch_nld_Indo-European
0.095
0.485
Old Norse_non_Indo-European
0.0525
0.4175
Irish_gle_Indo-European
Flemish_vls_Indo-European
0.755
Latin_lat_Indo-European
0.055
0.0275
Nocte_njb_Tibeto-Burman
Digaru Mishmi_mhu_Tibeto-Burman
0.735
Catalan_cat_Indo-European
0.1975
0.855
Fering Fries_frr_Indo-European
Portuguese_por_Indo-European
0.415
0.36
Cornish_cor_Indo-European
Welsh_cym_Indo-European
Jysk_jut_Indo-European
0.4125
Swedish_swe_Indo-European
0.575
German_deu_Indo-European
0.55
Dagur_dta_Mongolic
German Swiss_gsw_Indo-European

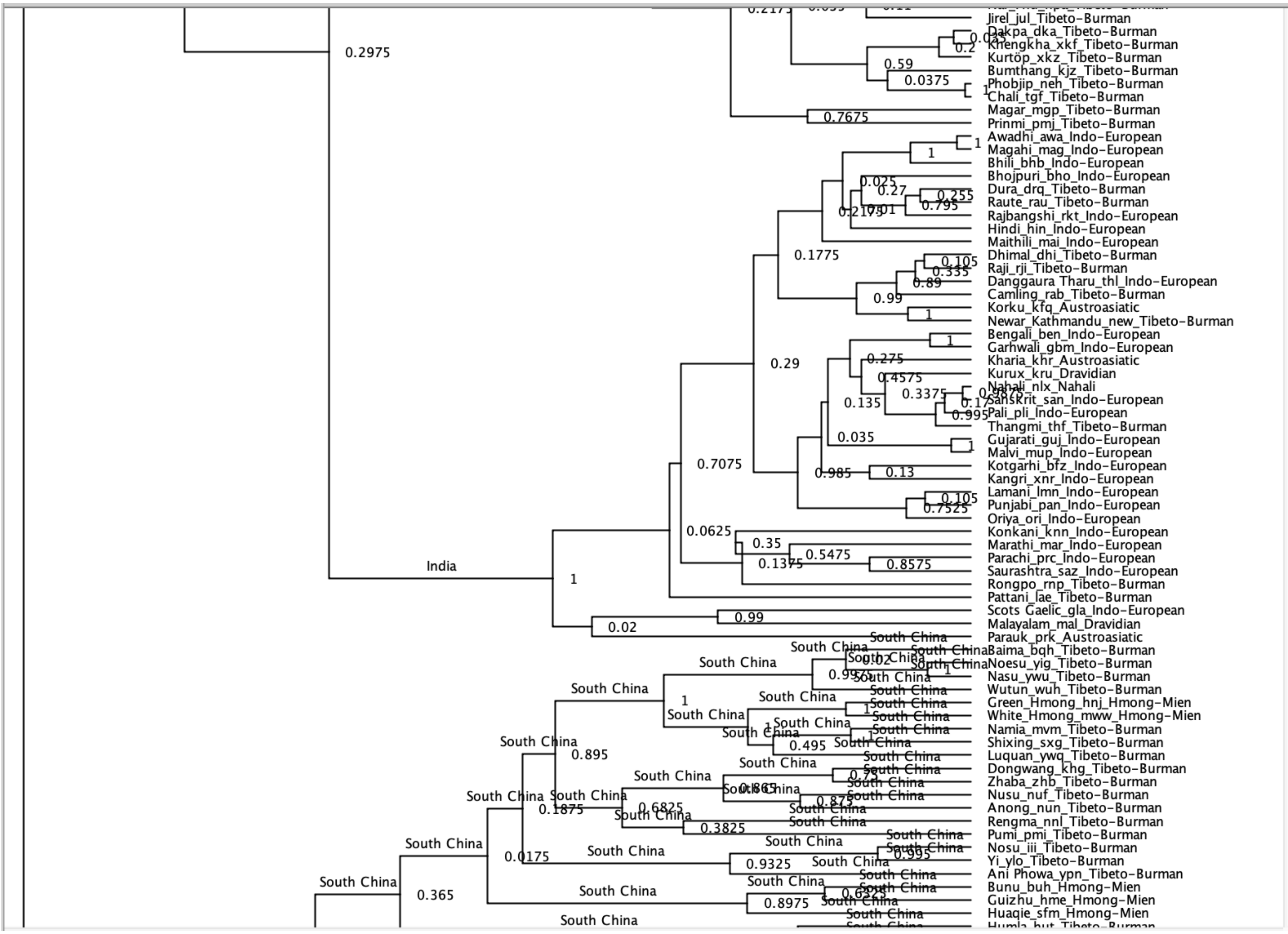
S1.7


German_Swiss_gsw_Indo-European
0.6025
0.065
English_eng_Indo-European
0.445
Lotha_njh_Tibeto-Burman
0.59
0.3875
Scots English_sco_Indo-European
Faroese_fao_Indo-European
0.23
Icelandic_isl_Indo-European
Indo-European and Turkic
0.64
0.23
Frisian_frs_Indo-European
0.62
0.7075
Walloon_wln_Indo-European
Saterland Fries_stq_Indo-European
0.1575
Kölsch_ksh_Indo-European
0.99
0.5725
Letzebuergesch_ltz_Indo-European
Bornholmsk_scy_Indo-European
Albanian Gheg_aln_Indo-European
1
0.3125
Albanian Tosk_als_Indo-European
0.0575
Mordvin_myv_Uralic
Czech_ces_Indo-European
Bosnian_bos_Indo-European
0.9825
0.165
1
Croatian_hrv_Indo-European
Serbian_srp_Indo-European
0.975
Macedonian_mkd_Indo-European
0.8575
0.1875
Slovak_slk_Indo-European
0.0175
Errommintxela_emx_Indo-European
0.265
Ladino_lad_Indo-European
Belarussian_bel_Indo-European
Bulgarian_bul_Indo-European
0.18
0.4
0.99
Ukrainian_ukr_Indo-European
0.3375
0.19
Sorbian_dsb_Indo-European
0.1675
Russian_rus_Indo-European
Polabian_pox_Indo-European
Limburgish_lim_Indo-European
0.01
Greek_ell_Indo-European
0.4275
Manx_glv_Indo-European
0.2075
Sicilian_scn_Indo-European
0.4225
Itelmen_itl_Chukotko-Kamchatkan
0.2125
Mari_Meadow_mhr_Uralic
1
Cheremis_mrj_Mongolic
0.545
Sami_sme_Uralic
0.8275
0.2675
Kamas_xas_Uralic
Udmurt_udm_Uralic
Aragonese_arg_Indo-European
0.6625
Basque_eus_Basque
0.835
Asturian_ast_Indo-European
1
0.3525
Galician_glg_Indo-European
0.9925
Spanish_spa_Indo-European
Breton_bre_Indo-European
0.28
Lydian_xld_Indo-European
0.0625
0.0825
Old French_fro_Indo-European
0.4375
Lombard_lmo_Indo-European
0.4575
Corsican_cos_Indo-European
0.485
Italian_ita_Indo-European
0.12
Komi_koi_Uralic
0.1375
0.0575
0.7875
Komi-Zyrian_kpv_Uralic
0.5225
Neapolitan_nap_Indo-European
Old Church Slavonic_chu_Indo-European
0.0125
0.045
Livonian_liv_Uralic
Sardinian_sro_Indo-European
Latvian_lav_Indo-European
1
Romansh_roh_Indo-European
Ai-cham_Diwo_aih_Tai-Kadai
0.0225
0.115
Hlai_lic_Tai-Kadai
Lakkia_lbc_Tai-Kadai
0.2125
Biaomin Yao_bmt_Hmong-Mien
0.89
0.005
0.165
Mulao_mlm_Tai-Kadai
Hmong_Taigong_hea_Hmong-Mien
Halang_hal_Austroasiatic
0.99
Sre_kpm_Austroasiatic
0.0225
Chin_Daai_dao_Tibeto-Burman
0.9825
0.185
Hkongso_mro_Tibeto-Burman
0.05
Phon_hpo_Tibeto-Burman
0.3075
Hu_huo_Austroasiatic

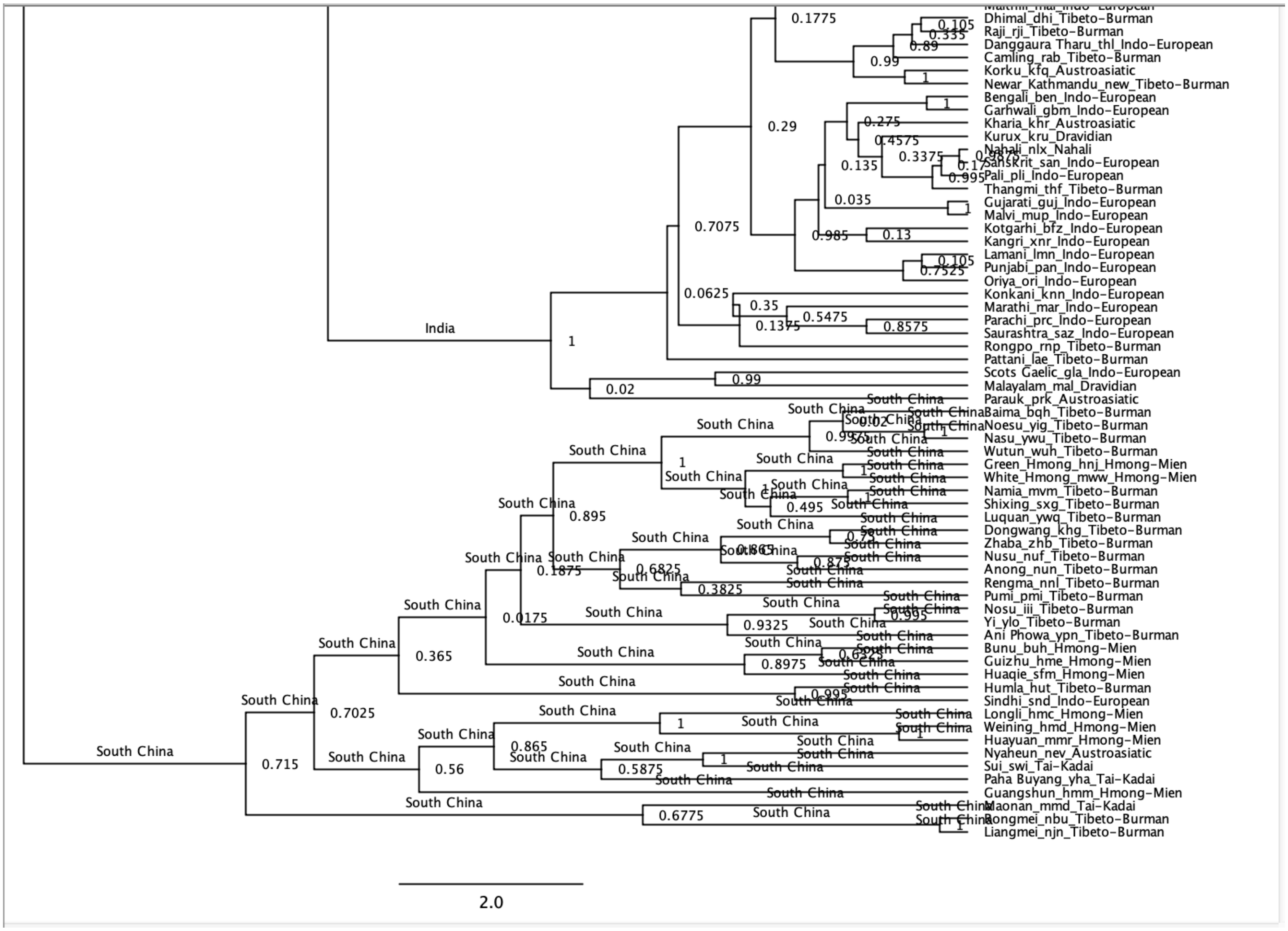
S1.8


1
Southeast Asia
0.3675
0.3675
0.585
0.05
0.3075
0.3275
0.495
0.5175
0.9225
0.4825
0.205
0.975
0.0525
0.5475
0.02
0.3925
0.48
0.455
1
0.99
0.48
0.9525
0.6025
0.2325
0.1025
0.97
0.325
0.385
0.9875
0.5675
0.4475
0.7625
0.37
0.22
0.4225
0.155
0.095
0.535
0.7125
0.98
1
0.7575
0.5225
0.8625
0.1175
0.3925
0.95
0.6475
0.795
0.01
0.91
0.9975
0.77
0.1875
0.27
0.01
0.9575
0.2
0.3925
1
0.625
0.8175
0.5575
0.23
0.695
0.035
0.11
0.2175
0.035
Phon_hpo_Tibeto-Burman
Hu_huo_Austroasiatic
Maru_mhx_Tibeto-Burman
Phunoi_pho_Tibeto-Burman
Sho_csh_Tibeto-Burman
Mlabri_mra_Austroasiatic
Burmese_mya_Tibeto-Burman
Nyahkur_cbn_Austroasiatic
Mah_Meri_mhe_Austroasiatic
Mon_mnw_Austroasiatic
Tampuan_tpu_Austroasiatic
Zhou Chinese_och_Tibeto-Burman
Wa_wbm_Austroasiatic
Samtao_stu_Austroasiatic
Kuke_ght_Tibeto-Burman
Deang_Guangka_rbb_Austroasiatic
Chepang_cdm_Tibeto-Burman
Sedang_sed_Austroasiatic
Semelai_sza_Austroasiatic
Ladakhi_lbj_Tibeto-Burman
Angami_njm_Tibeto-Burman
Khezha_nkh_Tibeto-Burman
Siberian Yupik_ess_Eskimo-Aleut
Anal_anm_Tibeto-Burman
Lakher_mrh_Tibeto-Burman
Naga_Mao_nbi_Tibeto-Burman
Bafut_nri_Tibeto-Burman
Kildin Sami_sjd_Uralic
Sami_Inari_smn_Uralic
Chin_Falam_cfm_Tibeto-Burman
Hakha_Lai_cnh_Tibeto-Burman
Hmar_hmr_Tibeto-Burman
Mizo_lus_Tibeto-Burman
Byansi_bee_Tibeto-Burman
Lhasa_Tibetan_bod_Tibeto-Burman
Bunan_bfu_Tibeto-Burman
Kinnauri_kfk_Tibeto-Burman
Tshangla_tsj_Tibeto-Burman
Lepcha_lep_Tibeto-Burman
Bugun_bgn_Tibeto-Burman
Sherdukpen_sdp_Tibeto-Burman
Brokpa_sgt_Tibeto-Burman
Aka_hru_Tibeto-Burman
Miji_sjl_Tibeto-Burman
Puroik_suv_Tibeto-Burman
Classical Tibetan_xct_Tibeto-Burman
Dolpo_dre_Tibeto-Burman
Olekha_ole_Tibeto-Burman
Zhang-Zhung_jna_Tibeto-Burman
Kanashi_xns_Tibeto-Burman
Dzala_dzl_Tibeto-Burman
Dzongkha_dzo_Tibeto-Burman
Yohlmo_scp_Tibeto-Burman
Mustang_loy_Tibeto-Burman
Denjongkha_sip_Tibeto-Burman
Sherpa_xsr_Tibeto-Burman
Chöcangacakha_cgk_Tibeto-Burman
Ghale_Uiya_ghh_Tibeto-Burman
Nubri_kte_Tibeto-Burman
Mugom_muk_Tibeto-Burman
Tsum_ttz_Tibeto-Burman
Manange_nmm_Tibeto-Burman
Ghale_ghe_Tibeto-Burman
Seke_skj_Tibeto-Burman
Thakali_ths_Tibeto-Burman
Lhomi_lhm_Tibeto-Burman
Nar_Phu_npa_Tibeto-Burman
Jirel_jul_Tibeto-Burman
Dakpa_dka_Tibeto-Burman
Khengkha_xkf_Tibeto-Burman

S1.9

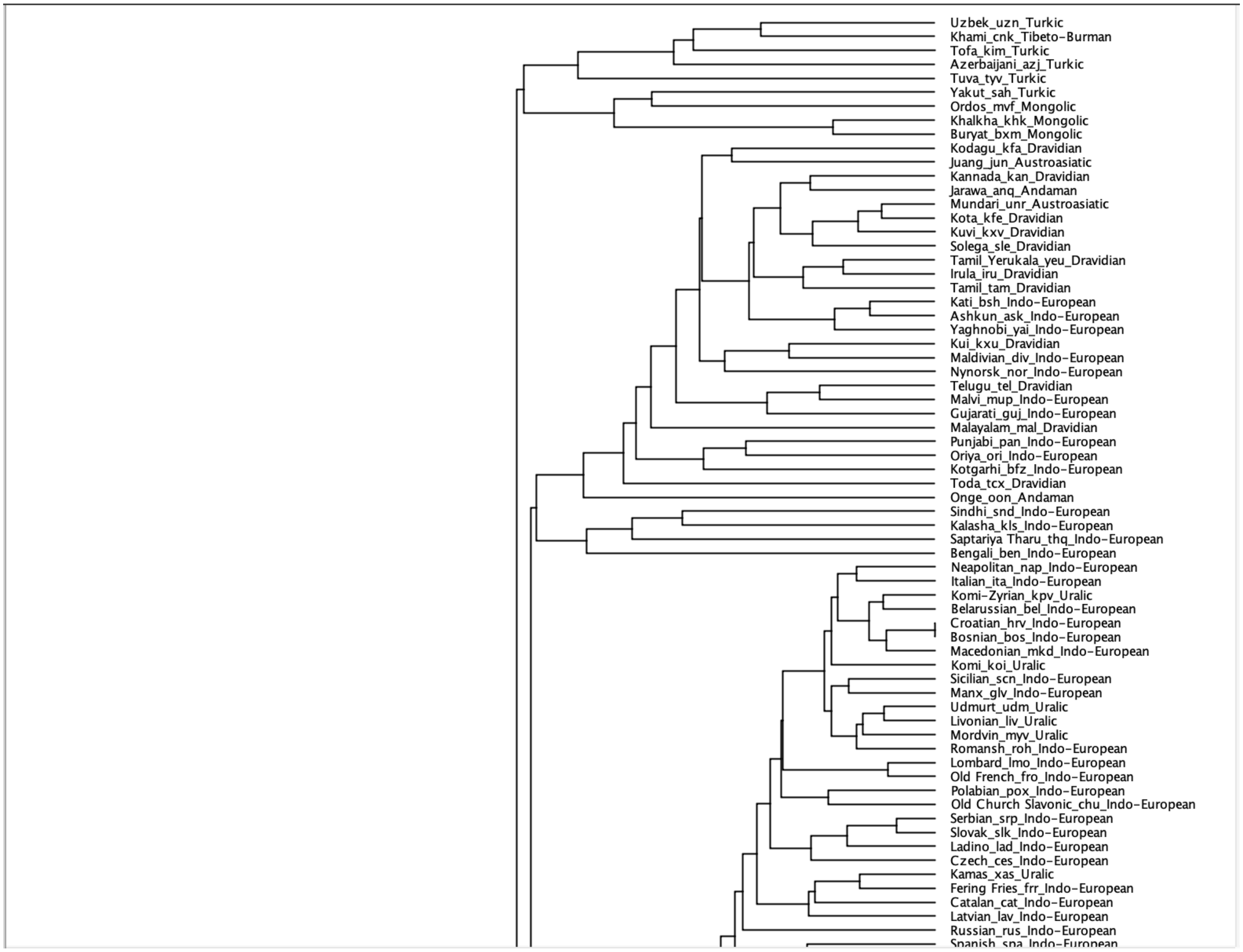

0.2175
0.035
0.11
Jirel_jul_Tibeto-Burman
Dakpa_dka_Tibeto-Burman
0.035
Khengkha_xkf_Tibeto-Burman
0.2
Kurtöp_xkz_Tibeto-Burman
0.2975
0.59
Bumthang_kjz_Tibeto-Burman
0.0375
Phobjip_neh_Tibeto-Burman
1
Chali_tgf_Tibeto-Burman
Magar_mgp_Tibeto-Burman
0.7675
Prinmi_pmj_Tibeto-Burman
Awadhi_awa_Indo-European
1
1
Magahi_mag_Indo-European
Bhili_bhb_Indo-European
0.025
Bhojpuri_bho_Indo-European
0.27
Dura_drq_Tibeto-Burman
0.255
Raute_rau_Tibeto-Burman
0.795
0.01
0.2175
Rajbangshi_rkt_Indo-European
Hindi_hin_Indo-European
Maithili_mai_Indo-European
0.1775
Dhimal_dhi_Tibeto-Burman
0.105
Raji_rji_Tibeto-Burman
0.335
0.89
Danggaura Tharu_thl_Indo-European
Camling_rab_Tibeto-Burman
0.99
Korku_kfq_Austroasiatic
1
Newar_Kathmandu_new_Tibeto-Burman
Bengali_ben_Indo-European
1
Garhwali_gbm_Indo-European
0.29
0.275
Kharia_khr_Austroasiatic
0.4575
Kurux_kru_Dravidian
Nahali_nlx_Nahali
0.3375
0.9875
0.17
Sanskrit_san_Indo-European
0.135
Pali_pli_Indo-European
0.995
Thangmi_thf_Tibeto-Burman
0.035
Gujarati_guj_Indo-European
1
Malvi_mup_Indo-European
0.7075
Kotgarhi_bfz_Indo-European
0.985
0.13
Kangri_xnr_Indo-European
Lamani_lmn_Indo-European
0.105
Punjabi_pan_Indo-European
0.7525
Oriya_ori_Indo-European
0.0625
Konkani_knn_Indo-European
0.35
Marathi_mar_Indo-European
0.5475
Parachi_prc_Indo-European
0.1375
0.8575
Saurashtra_saz_Indo-European
India
1
Rongpo_rnp_Tibeto-Burman
Pattani_lae_Tibeto-Burman
Scots Gaelic_gla_Indo-European
0.99
Malayalam_mal_Dravidian
0.02
South China
Parauk_prk_Austroasiatic
South China
South China
Baima_bqh_Tibeto-Burman
South China
0.02
South China
South China
Noesu_yig_Tibeto-Burman
0.9975
South China
1
Nasu_ywu_Tibeto-Burman
South China
South China
Wutun_wuh_Tibeto-Burman
1
South China
South China
Green_Hmong_hnj_Hmong-Mien
1
South China
South China
White_Hmong_mww_Hmong-Mien
South China
1
South China
South China
Namia_mvm_Tibeto-Burman
South China
1
South China
Shixing_sxg_Tibeto-Burman
0.895
0.495
South China
South China
Luquan_ywq_Tibeto-Burman
South China
South China
Dongwang_khg_Tibeto-Burman
South China
0.73
South China
Zhaba_zhb_Tibeto-Burman
South China
South China
0.865
South China
Nusu_nuf_Tibeto-Burman
0.1875
0.6825
0.875
South China
Anong_nun_Tibeto-Burman
South China
South China
Rengma_nnl_Tibeto-Burman
0.3825
South China
Pumi_pmi_Tibeto-Burman
South China
South China
South China
Nosu_iii_Tibeto-Burman
0.0175
South China
0.995
Yi_ylo_Tibeto-Burman
0.9325
South China
Ani Phowa_ypn_Tibeto-Burman
South China
South China
South China
Bunu_buh_Hmong-Mien
0.365
South China
0.6525
South China
Guizhu_hme_Hmong-Mien
0.8975
South China
South China
Huaqie_sfm_Hmong-Mien
South China
South China
Humla_hut_Tibeto-Burman

S1.10

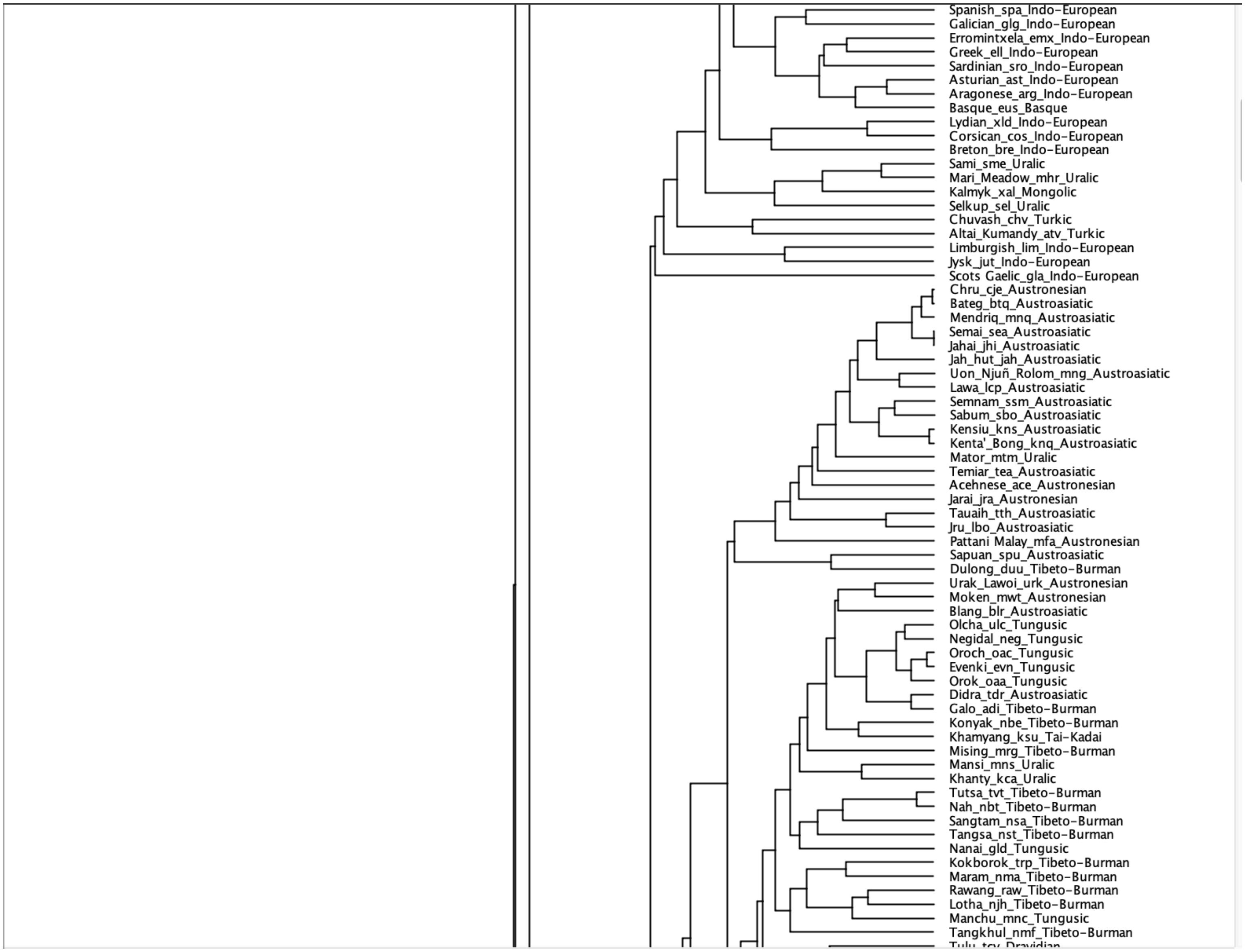

Maithili_mai_Indo-European
Dhimal_dhi_Tibeto-Burman
Raji_rji_Tibeto-Burman
Danggaura Tharu_thl_Indo-European
Camling_rab_Tibeto-Burman
Korku_kfq_Austroasiatic
Newar_Kathmandu_new_Tibeto-Burman
Bengali_ben_Indo-European
Garhwali_gbm_Indo-European
Kharia_khr_Austroasiatic
Kurux_kru_Dravidian
Nahali_nlx_Nahali
Sanskrit_san_Indo-European
Pali_pli_Indo-European
Thangmi_thf_Tibeto-Burman
Gujarati_guj_Indo-European
Malvi_mup_Indo-European
Kotgarhi_bfz_Indo-European
Kangri_xnr_Indo-European
Lamani_lmn_Indo-European
Punjabi_pan_Indo-European
Oriya_ori_Indo-European
Konkani_knn_Indo-European
Marathi_mar_Indo-European
Parachi_prc_Indo-European
Saurashtra_saz_Indo-European
Rongpo_rnp_Tibeto-Burman
Pattani_lae_Tibeto-Burman
Scots Gaelic_gla_Indo-European
Malayalam_mal_Dravidian
Parauk_prk_Austroasiatic
Baima_bqh_Tibeto-Burman
Noesu_yig_Tibeto-Burman
Nasu_ywu_Tibeto-Burman
Wutun_wuh_Tibeto-Burman
Green_Hmong_hnj_Hmong-Mien
White_Hmong_mww_Hmong-Mien
Namia_mvm_Tibeto-Burman
Shixing_sxg_Tibeto-Burman
Luquan_ywq_Tibeto-Burman
Dongwang_khg_Tibeto-Burman
Zhaba_zhb_Tibeto-Burman
Nusu_nuf_Tibeto-Burman
Anong_nun_Tibeto-Burman
Rengma_nnl_Tibeto-Burman
Pumi_pmi_Tibeto-Burman
Nosu_iii_Tibeto-Burman
Yi_ylo_Tibeto-Burman
Ani Phowa_ypn_Tibeto-Burman
Bunu_buh_Hmong-Mien
Guizhu_hme_Hmong-Mien
Huaqie_sfm_Hmong-Mien
Humla_hut_Tibeto-Burman
Sindhi_snd_Indo-European
Longli_hmc_Hmong-Mien
Weining_hmd_Hmong-Mien
Huayuan_mmr_Hmong-Mien
Nyaheun_nev_Austroasiatic
Sui_swi_Tai-Kadai
Paha Buyang_yha_Tai-Kadai
Guangshun_hmm_Hmong-Mien
Maonan_mmd_Tai-Kadai
Rongmei_nbu_Tibeto-Burman
Liangmei_njn_Tibeto-Burman
0.1775
0.105
0.335
0.89
0.99
1
0.29
0.275
0.4575
0.3375
0.135
0.995
0.035
0.7075
0.985
0.13
0.7525
0.0625
0.35
0.5475
0.1375
0.8575
India
1
0.99
0.02
South China
0.9975
0.895
0.495
0.865
0.875
0.1875
0.6825
0.3825
0.0175
0.9325
0.995
0.365
0.8975
0.7025
0.715
0.865
0.56
0.5875
0.6775
2.0

S1.11

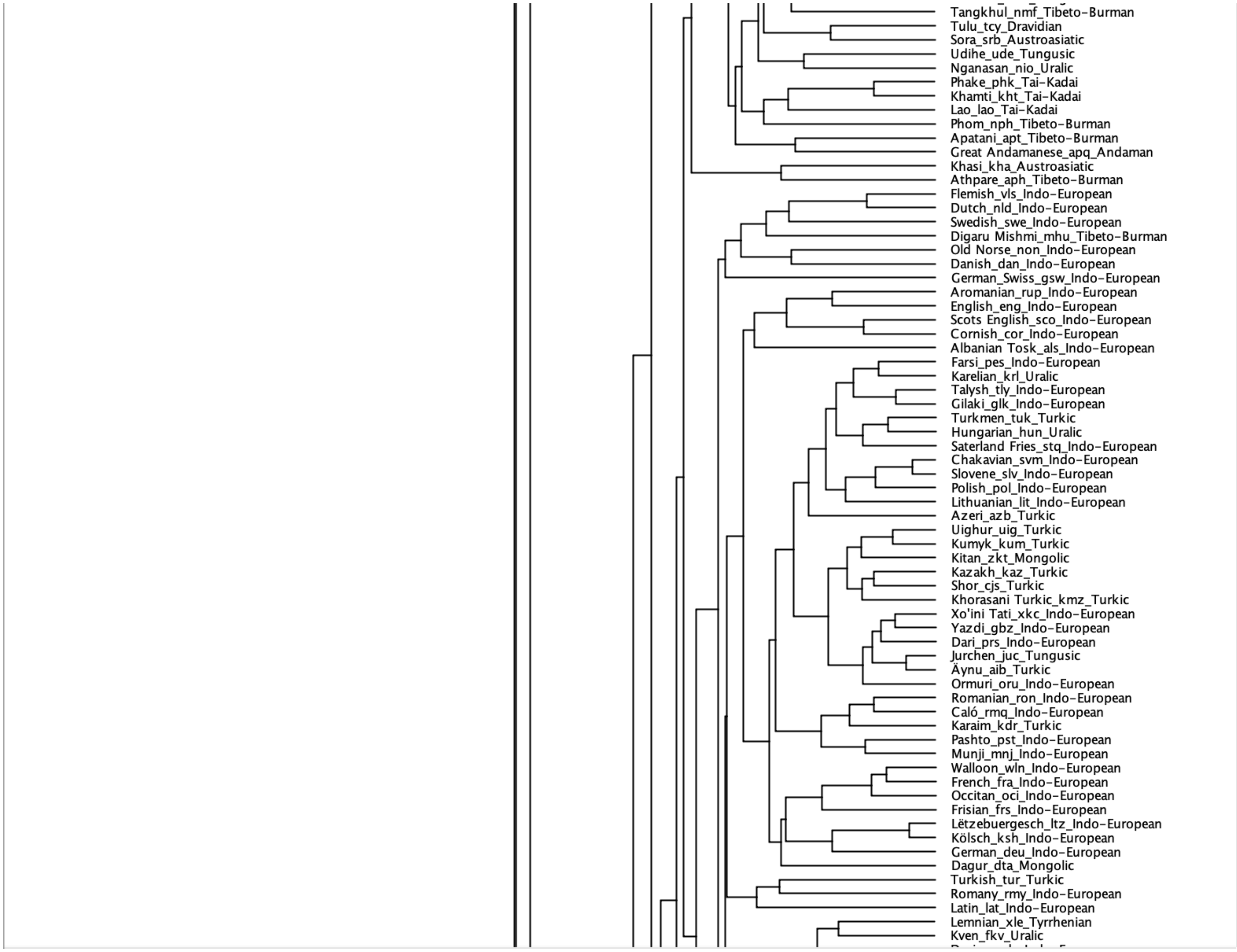
Supplementary Figure S2:1-11 Phonotactics tree using UPGMA


Uzbek_uzn_Turkic
Khami_cnk_Tibeto-Burman
Tofa_kim_Turkic
Azerbaijani_azj_Turkic
Tuva_tyv_Turkic
Yakut_sah_Turkic
Ordos_mvf_Mongolic
Khalkha_khk_Mongolic
Buryat_bxm_Mongolic
Kodagu_kfa_Dravidian
Juang_jun_Austroasiatic
Kannada_kan_Dravidian
Jarawa_anq_Andaman
Mundari_unr_Austroasiatic
Kota_kfe_Dravidian
Kuvi_kxv_Dravidian
Solega_sle_Dravidian
Tamil_Yerukala_yeu_Dravidian
Irula_iru_Dravidian
Tamil_tam_Dravidian
Kati_bsh_Indo-European
Ashkun_ask_Indo-European
Yaghnobi_yai_Indo-European
Kui_kxu_Dravidian
Maldivian_div_Indo-European
Nynorsk_nor_Indo-European
Telugu_tel_Dravidian
Malvi_mup_Indo-European
Gujarati_guj_Indo-European
Malayalam_mal_Dravidian
Punjabi_pan_Indo-European
Oriya_ori_Indo-European
Kotgarhi_bfz_Indo-European
Toda_tcx_Dravidian
Onge_oon_Andaman
Sindhi_snd_Indo-European
Kalasha_kls_Indo-European
Saptariya Tharu_thq_Indo-European
Bengali_ben_Indo-European
Neapolitan_nap_Indo-European
Italian_ita_Indo-European
Komi-Zyrian_kpv_Uralic
Belarussian_bel_Indo-European
Croatian_hrv_Indo-European
Bosnian_bos_Indo-European
Macedonian_mkd_Indo-European
Komi_koi_Uralic
Sicilian_scn_Indo-European
Manx_glv_Indo-European
Udmurt_udm_Uralic
Livonian_liv_Uralic
Mordvin_myv_Uralic
Romansh_roh_Indo-European
Lombard_lmo_Indo-European
Old French_fro_Indo-European
Polabian_pox_Indo-European
Old Church Slavonic_chu_Indo-European
Serbian_srp_Indo-European
Slovak_slk_Indo-European
Ladino_lad_Indo-European
Czech_ces_Indo-European
Kamas_xas_Uralic
Fering Fries_frr_Indo-European
Catalan_cat_Indo-European
Latvian_lav_Indo-European
Russian_rus_Indo-European
Spanish_spa_Indo-European

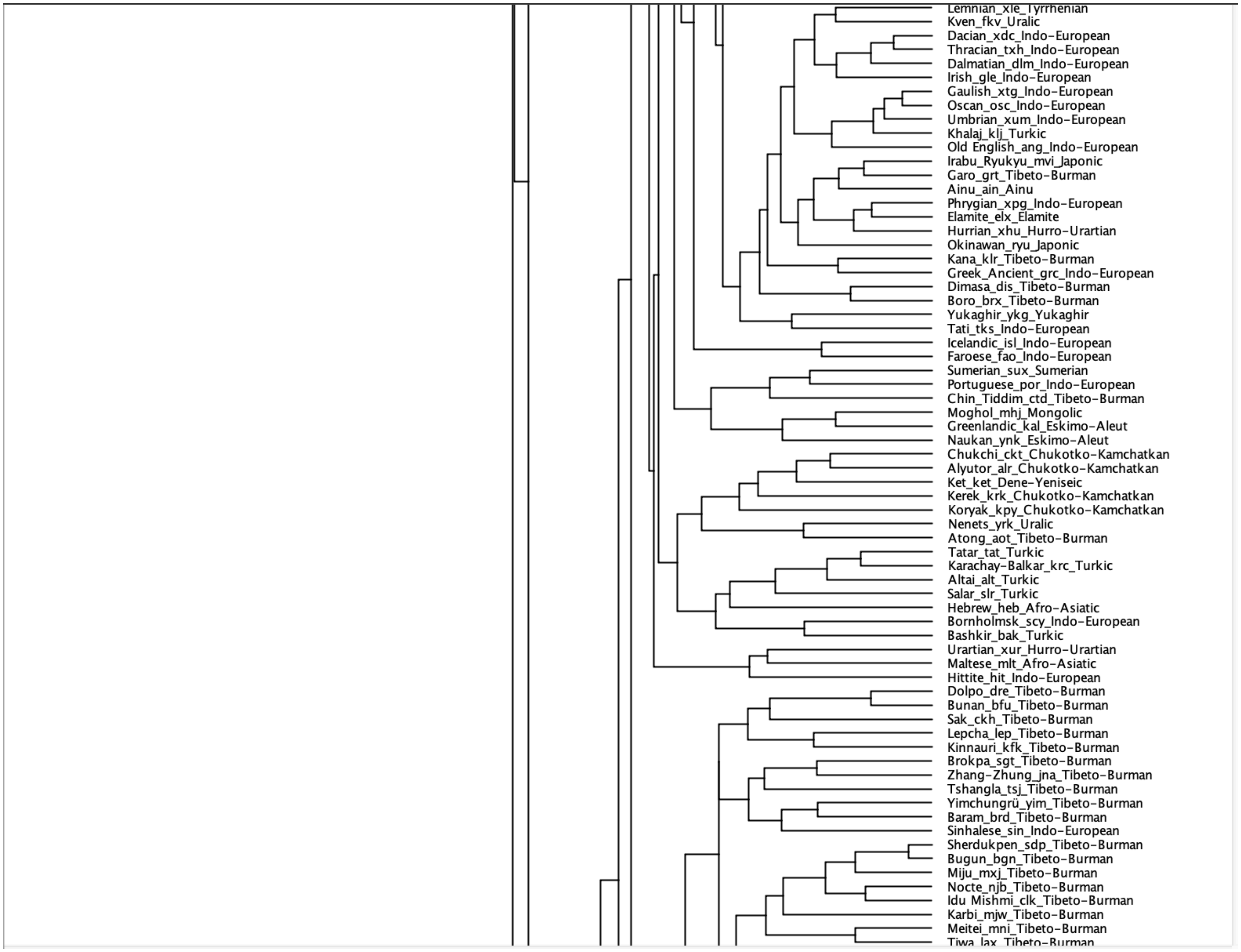
S2.1


Spanish_spa_Indo-European
Galician_glg_Indo-European
Erromintxela_emx_Indo-European
Greek_ell_Indo-European
Sardinian_sro_Indo-European
Asturian_ast_Indo-European
Aragonese_arg_Indo-European
Basque_eus_Basque
Lydian_xld_Indo-European
Corsican_cos_Indo-European
Breton_bre_Indo-European
Sami_sme_Uralic
Mari_Meadow_mhr_Uralic
Kalmyk_xal_Mongolic
Selkup_sel_Uralic
Chuvash_chv_Turkic
Altai_Kumandy_atv_Turkic
Limburgish_lim_Indo-European
Jysk_jut_Indo-European
Scots Gaelic_gla_Indo-European
Chru_cje_Austronesian
Bateg_btq_Austroasiatic
Mendriq_mnq_Austroasiatic
Semai_sea_Austroasiatic
Jahai_jhi_Austroasiatic
Jah_hut_jah_Austroasiatic
Uon_Njuñ_Rolom_mng_Austroasiatic
Lawa_lcp_Austroasiatic
Semnam_ssm_Austroasiatic
Sabum_sbo_Austroasiatic
Kensiu_kns_Austroasiatic
Kenta'_Bong_knq_Austroasiatic
Mator_mtm_Uralic
Temiar_tea_Austroasiatic
Acehnese_ace_Austronesian
Jarai_jra_Austronesian
Tauaih_tth_Austroasiatic
Jru_lbo_Austroasiatic
Pattani Malay_mfa_Austronesian
Sapuan_spu_Austroasiatic
Dulong_duu_Tibeto-Burman
Urak_Lawoi_urk_Austronesian
Moken_mwt_Austronesian
Blang_blr_Austroasiatic
Olcha_ulc_Tungusic
Negidal_neg_Tungusic
Oroch_oac_Tungusic
Evenki_evn_Tungusic
Orok_oaa_Tungusic
Didra_tdr_Austroasiatic
Galo_adi_Tibeto-Burman
Konyak_nbe_Tibeto-Burman
Khamyang_ksu_Tai-Kadai
Mising_mrg_Tibeto-Burman
Mansi_mns_Uralic
Khanty_kca_Uralic
Tutsa_tvt_Tibeto-Burman
Nah_nbt_Tibeto-Burman
Sangtam_nsa_Tibeto-Burman
Tangsa_nst_Tibeto-Burman
Nanai_gld_Tungusic
Kokborok_trp_Tibeto-Burman
Maram_nma_Tibeto-Burman
Rawang_raw_Tibeto-Burman
Lotha_njh_Tibeto-Burman
Manchu_mnc_Tungusic
Tangkhul_nmf_Tibeto-Burman
Tulu_tcy_Dravidian

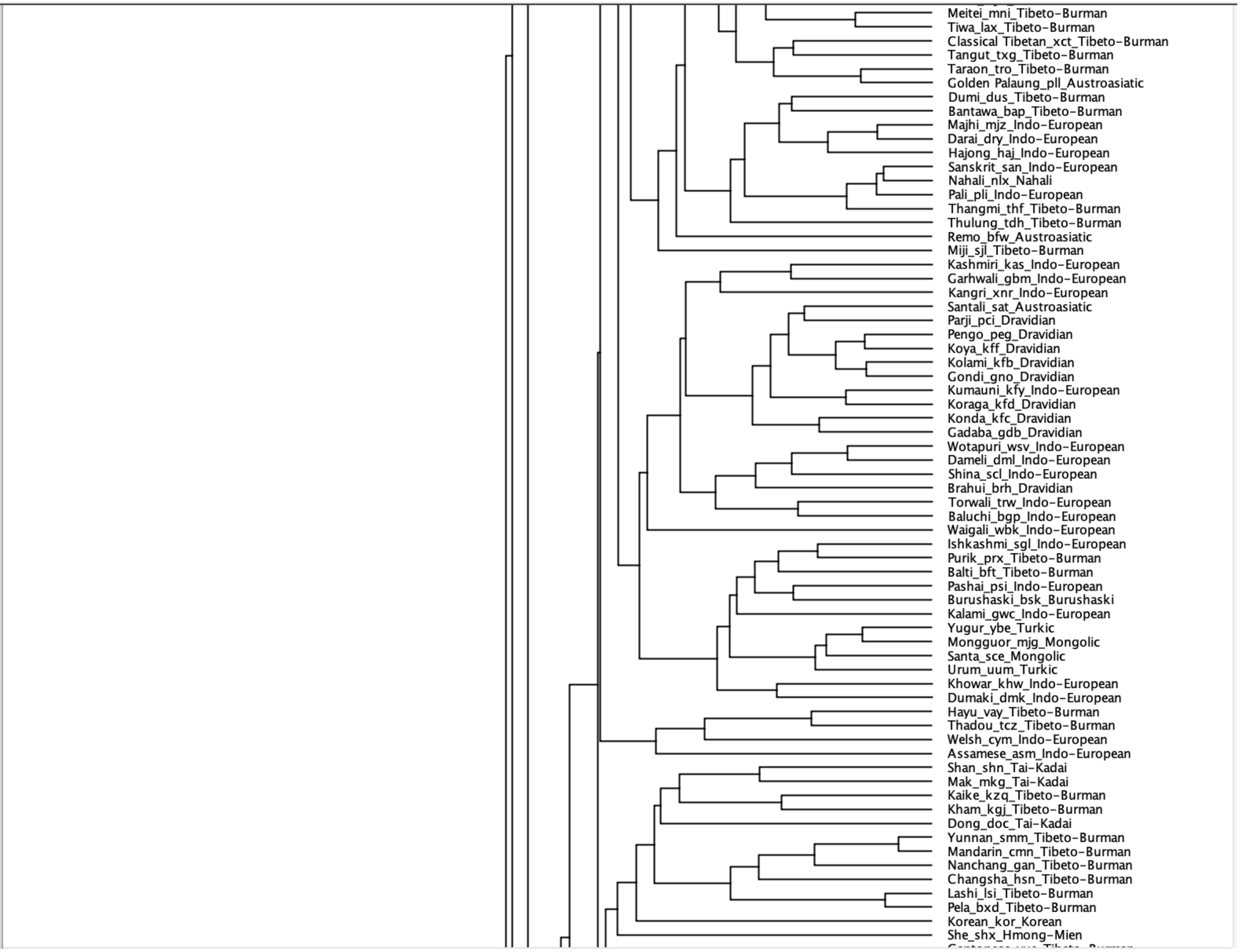
S2.2


Tangkhul_nmf_Tibeto-Burman
Tulu_tcy_Dravidian
Sora_srb_Austroasiatic
Udihe_ude_Tungusic
Nganasan_nio_Uralic
Phake_phk_Tai-Kadai
Khamti_kht_Tai-Kadai
Lao_lao_Tai-Kadai
Phom_nph_Tibeto-Burman
Apatani_apt_Tibeto-Burman
Great Andamanese_apq_Andaman
Khasi_kha_Austroasiatic
Athpare_aph_Tibeto-Burman
Flemish_vls_Indo-European
Dutch_nld_Indo-European
Swedish_swe_Indo-European
Digaru Mishmi_mhu_Tibeto-Burman
Old Norse_non_Indo-European
Danish_dan_Indo-European
German_Swiss_gsw_Indo-European
Aromanian_rup_Indo-European
English_eng_Indo-European
Scots English_sco_Indo-European
Cornish_cor_Indo-European
Albanian Tosk_als_Indo-European
Farsi_pes_Indo-European
Karelian_krl_Uralic
Talysh_tly_Indo-European
Gilaki_glk_Indo-European
Turkmen_tuk_Turkic
Hungarian_hun_Uralic
Saterland Fries_stq_Indo-European
Chakavian_svm_Indo-European
Slovene_slv_Indo-European
Polish_pol_Indo-European
Lithuanian_lit_Indo-European
Azeri_azb_Turkic
Uighur_uig_Turkic
Kumyk_kum_Turkic
Kitan_zkt_Mongolic
Kazakh_kaz_Turkic
Shor_cjs_Turkic
Khorasani Turkic_kmz_Turkic
Xo'ini Tati_xkc_Indo-European
Yazdi_gbz_Indo-European
Dari_prs_Indo-European
Jurchen_juc_Tungusic
Äynu_aib_Turkic
Ormuri_oru_Indo-European
Romanian_ron_Indo-European
Caló_rmq_Indo-European
Karaim_kdr_Turkic
Pashto_pst_Indo-European
Munji_mnj_Indo-European
Walloon_wln_Indo-European
French_fra_Indo-European
Occitan_oci_Indo-European
Frisian_frs_Indo-European
Lëtzebuergesch_ltz_Indo-European
Kölsch_ksh_Indo-European
German_deu_Indo-European
Dagur_dta_Mongolic
Turkish_tur_Turkic
Romany_rmy_Indo-European
Latin_lat_Indo-European
Lemnian_xle_Tyrrhenian
Kven_fkv_Uralic

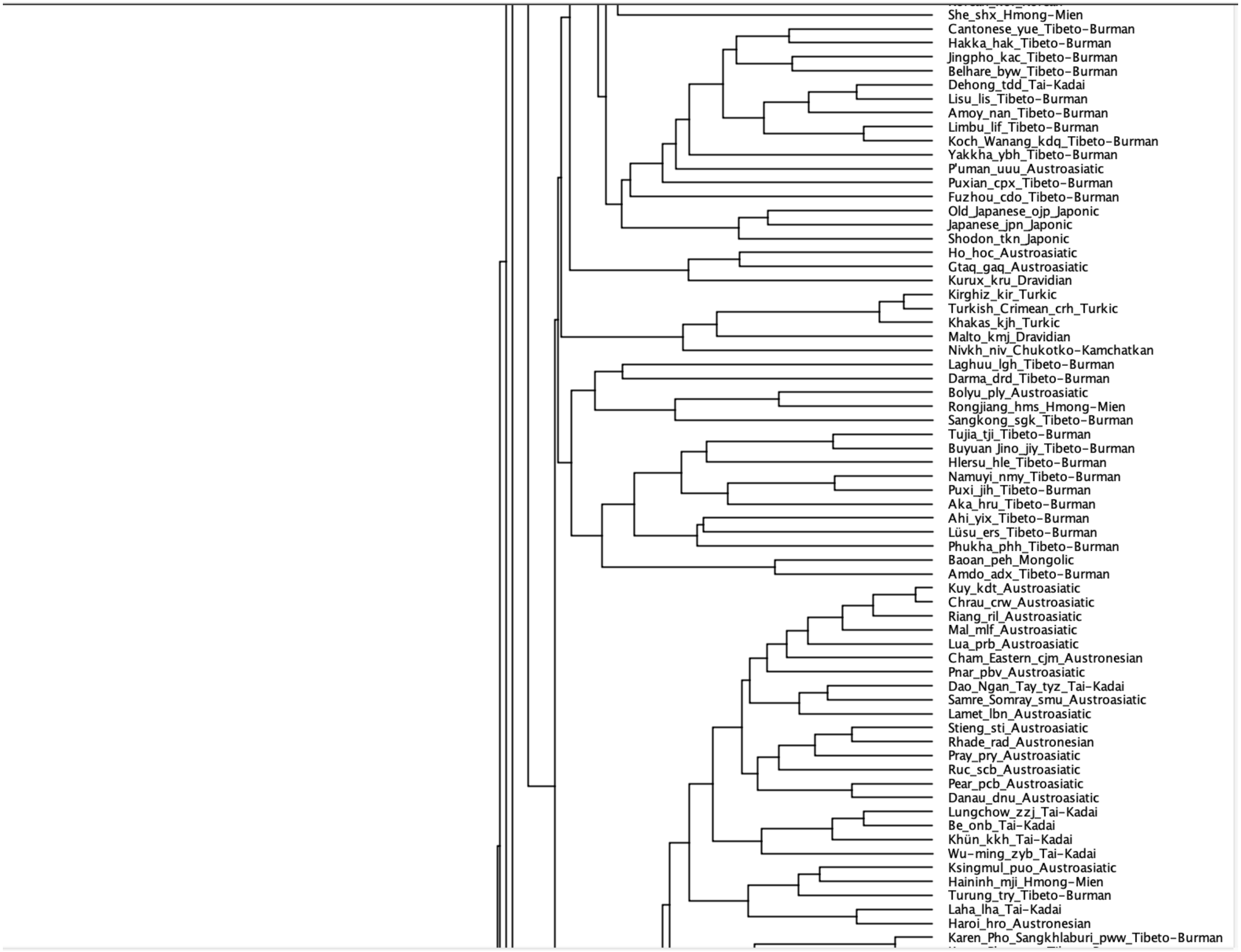
S2.3


Lemnian_xle_Tyrrhenian
Kven_fkv_Uralic
Dacian_xdc_Indo-European
Thracian_txh_Indo-European
Dalmatian_dlm_Indo-European
Irish_gle_Indo-European
Gaulish_xtg_Indo-European
Oscan_osc_Indo-European
Umbrian_xum_Indo-European
Khalaj_klj_Turkic
Old English_ang_Indo-European
Irabu_Ryukyu_mvi_Japonic
Garo_grt_Tibeto-Burman
Ainu_ain_Ainu
Phrygian_xpg_Indo-European
Elamite_elx_Elamite
Hurrian_xhu_Hurro-Urartian
Okinawan_ryu_Japonic
Kana_klr_Tibeto-Burman
Greek_Ancient_grc_Indo-European
Dimasa_dis_Tibeto-Burman
Boro_brx_Tibeto-Burman
Yukaghir_ykg_Yukaghir
Tati_tks_Indo-European
Icelandic_isl_Indo-European
Faroese_fao_Indo-European
Sumerian_sux_Sumerian
Portuguese_por_Indo-European
Chin_Tiddim_ctd_Tibeto-Burman
Moghol_mhj_Mongolic
Greenlandic_kal_Eskimo-Aleut
Naukan_ynk_Eskimo-Aleut
Chukchi_ckt_Chukotko-Kamchatkan
Alyutor_alr_Chukotko-Kamchatkan
Ket_ket_Dene-Yeniseic
Kerek_krk_Chukotko-Kamchatkan
Koryak_kpy_Chukotko-Kamchatkan
Nenets_yrk_Uralic
Atong_aot_Tibeto-Burman
Tatar_tat_Turkic
Karachay-Balkar_krc_Turkic
Altai_alt_Turkic
Salar_slr_Turkic
Hebrew_heb_Afro-Asiatic
Bornholmsk_scy_Indo-European
Bashkir_bak_Turkic
Urartian_xur_Hurro-Urartian
Maltese_mlt_Afro-Asiatic
Hittite_hit_Indo-European
Dolpo_dre_Tibeto-Burman
Bunan_bfu_Tibeto-Burman
Sak_ckh_Tibeto-Burman
Lepcha_lep_Tibeto-Burman
Kinnauri_kfk_Tibeto-Burman
Brokpa_sgt_Tibeto-Burman
Zhang-Zhung_jna_Tibeto-Burman
Tshangla_tsj_Tibeto-Burman
Yimchungrü_yim_Tibeto-Burman
Baram_brd_Tibeto-Burman
Sinhalese_sin_Indo-European
Sherdukpen_sdp_Tibeto-Burman
Bugun_bgn_Tibeto-Burman
Miju_mxj_Tibeto-Burman
Nocte_njb_Tibeto-Burman
Idu Mishmi_clk_Tibeto-Burman
Karbi_mjw_Tibeto-Burman
Meitei_mni_Tibeto-Burman
Tiwa_lax_Tibeto-Burman

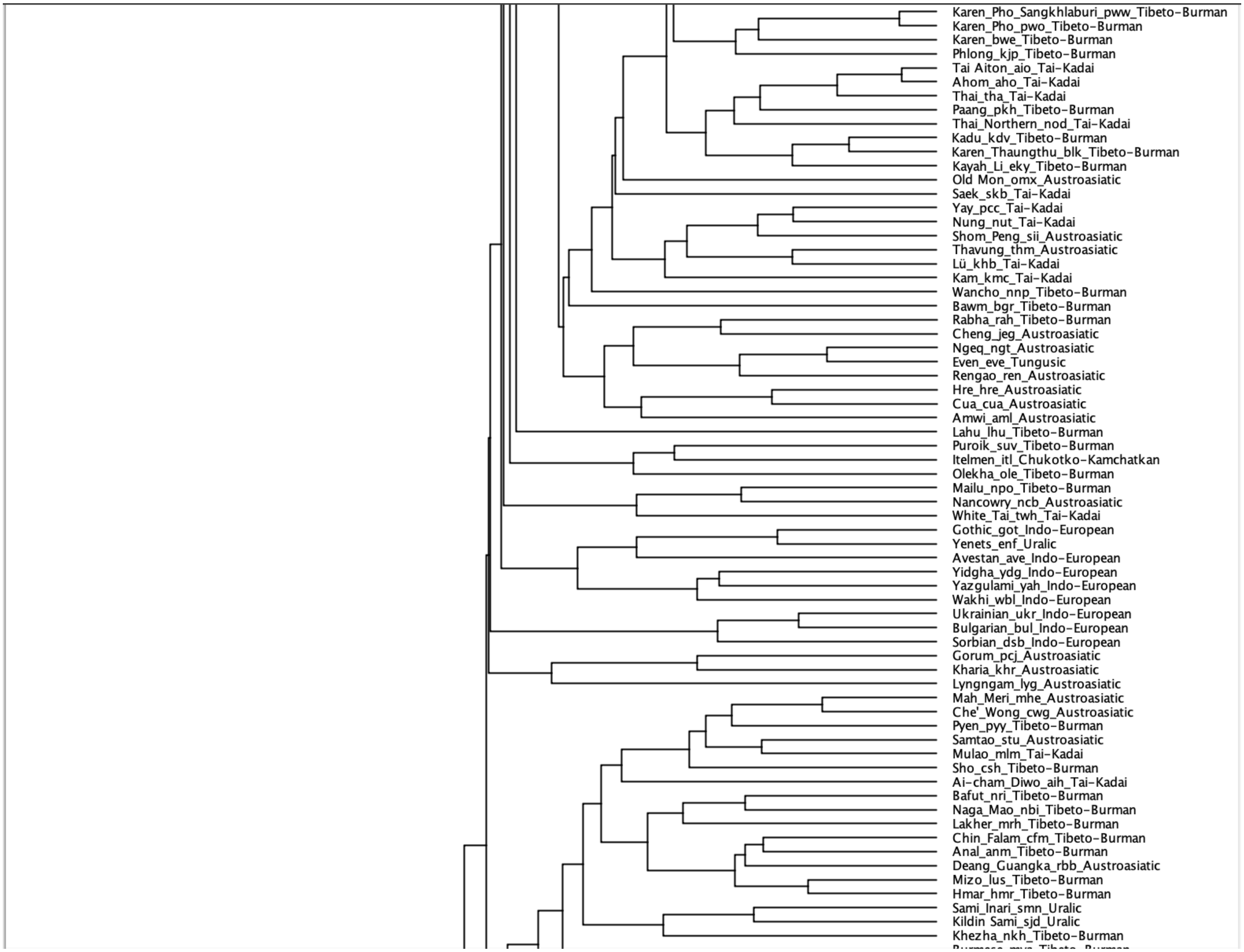
S2.4


Meitei_mni_Tibeto-Burman
Tiwa_lax_Tibeto-Burman
Classical Tibetan_xct_Tibeto-Burman
Tangut_txg_Tibeto-Burman
Taraon_tro_Tibeto-Burman
Golden Palaung_pll_Austroasiatic
Dumi_dus_Tibeto-Burman
Bantawa_bap_Tibeto-Burman
Majhi_mjz_Indo-European
Darai_dry_Indo-European
Hajong_haj_Indo-European
Sanskrit_san_Indo-European
Nahali_nlx_Nahali
Pali_pli_Indo-European
Thangmi_thf_Tibeto-Burman
Thulung_tdh_Tibeto-Burman
Remo_bfw_Austroasiatic
Miji_sjl_Tibeto-Burman
Kashmiri_kas_Indo-European
Garhwali_gbm_Indo-European
Kangri_xnr_Indo-European
Santali_sat_Austroasiatic
Parji_pci_Dravidian
Pengo_peg_Dravidian
Koya_kff_Dravidian
Kolami_kfb_Dravidian
Gondi_gno_Dravidian
Kumauni_kfy_Indo-European
Koraga_kfd_Dravidian
Konda_kfc_Dravidian
Gadaba_gdb_Dravidian
Wotapuri_wsv_Indo-European
Dameli_dml_Indo-European
Shina_scl_Indo-European
Brahui_brh_Dravidian
Torwali_trw_Indo-European
Baluchi_bgp_Indo-European
Waigali_wbk_Indo-European
Ishkashmi_sgl_Indo-European
Purik_prx_Tibeto-Burman
Balti_bft_Tibeto-Burman
Pashai_psi_Indo-European
Burushaski_bsk_Burushaski
Kalami_gwc_Indo-European
Yugur_ybe_Turkic
Mongguor_mjg_Mongolic
Santa_sce_Mongolic
Urum_uum_Turkic
Khowar_khw_Indo-European
Dumaki_dmk_Indo-European
Hayu_vay_Tibeto-Burman
Thadou_tcz_Tibeto-Burman
Welsh_cym_Indo-European
Assamese_asm_Indo-European
Shan_shn_Tai-Kadai
Mak_mkg_Tai-Kadai
Kaike_kzq_Tibeto-Burman
Kham_kgj_Tibeto-Burman
Dong_doc_Tai-Kadai
Yunnan_smm_Tibeto-Burman
Mandarin_cmn_Tibeto-Burman
Nanchang_gan_Tibeto-Burman
Changsha_hsn_Tibeto-Burman
Lashi_lsi_Tibeto-Burman
Pela_bxd_Tibeto-Burman
Korean_kor_Korean
She_shx_Hmong-Mien

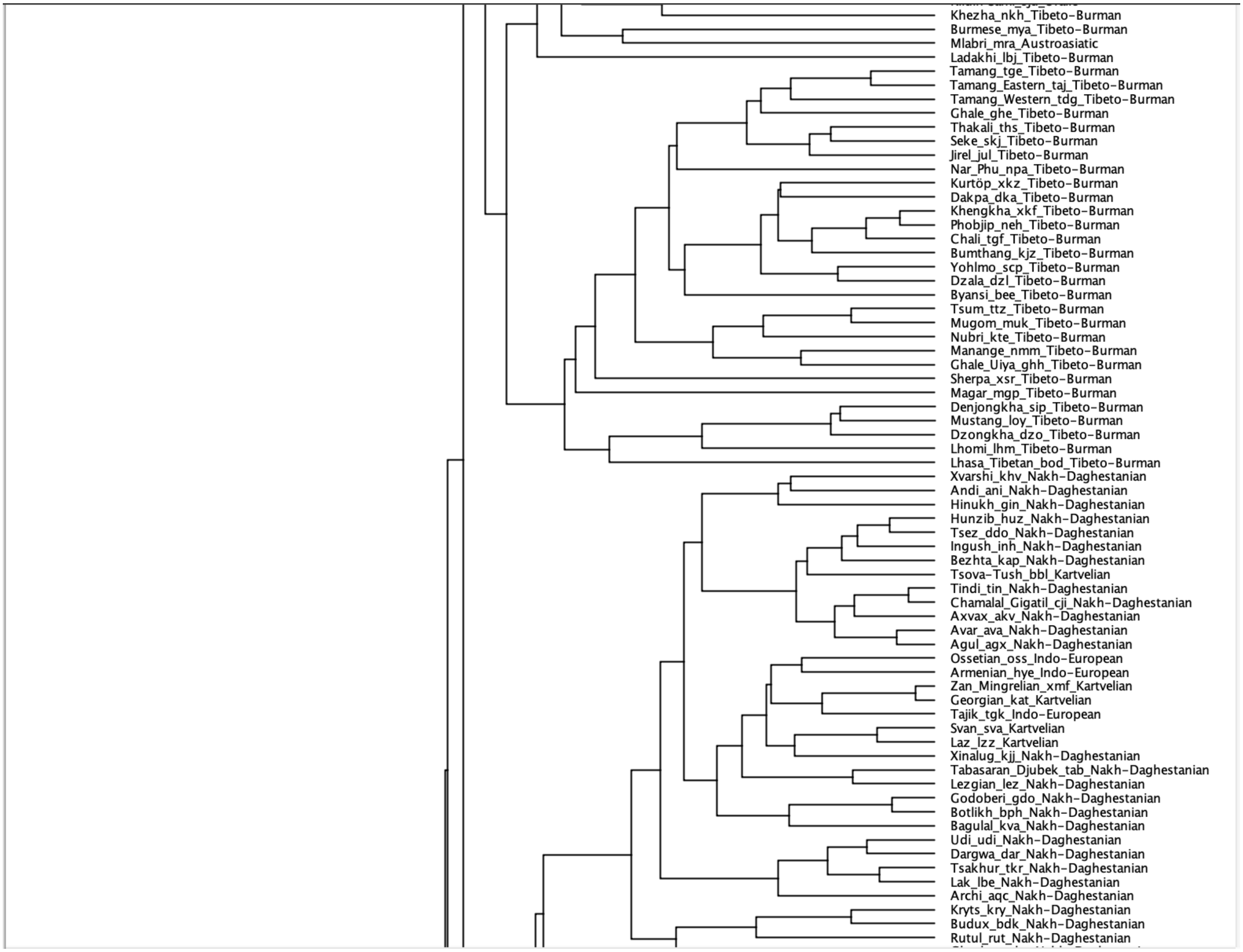
S2.5


She_shx_Hmong-Mien
Cantonese_yue_Tibeto-Burman
Hakka_hak_Tibeto-Burman
Jingpho_kac_Tibeto-Burman
Belhare_byw_Tibeto-Burman
Dehong_tdd_Tai-Kadai
Lisu_lis_Tibeto-Burman
Amoy_nan_Tibeto-Burman
Limbu_lif_Tibeto-Burman
Koch_Wanang_kdq_Tibeto-Burman
Yakkha_ybh_Tibeto-Burman
P'uman_uuu_Austroasiatic
Puxian_cpx_Tibeto-Burman
Fuzhou_cdo_Tibeto-Burman
Old_Japanese_ojp_Japonic
Japanese_jpn_Japonic
Shodon_tkn_Japonic
Ho_hoc_Austroasiatic
Gtaq_gaq_Austroasiatic
Kurux_kru_Dravidian
Kirghiz_kir_Turkic
Turkish_Crimean_crh_Turkic
Khakas_kjh_Turkic
Malto_kmj_Dravidian
Nivkh_niv_Chukotko-Kamchatkan
Laghuu_lgh_Tibeto-Burman
Darma_drd_Tibeto-Burman
Bolyu_ply_Austroasiatic
Rongjiang_hms_Hmong-Mien
Sangkong_sgk_Tibeto-Burman
Tujia_tji_Tibeto-Burman
Buyuan Jino_jiy_Tibeto-Burman
Hlersu_hle_Tibeto-Burman
Namuyi_nmy_Tibeto-Burman
Puxi_jih_Tibeto-Burman
Aka_hru_Tibeto-Burman
Ahi_yix_Tibeto-Burman
Lüsu_ers_Tibeto-Burman
Phukha_phh_Tibeto-Burman
Baoan_peh_Mongolic
Amdo_adx_Tibeto-Burman
Kuy_kdt_Austroasiatic
Chrau_crw_Austroasiatic
Riang_ril_Austroasiatic
Mal_mlf_Austroasiatic
Lua_prb_Austroasiatic
Cham_Eastern_cjm_Austronesian
Pnar_pbv_Austroasiatic
Dao_Ngan_Tay_tyz_Tai-Kadai
Samre_Somray_smu_Austroasiatic
Lamet_lbn_Austroasiatic
Stieng_sti_Austroasiatic
Rhade_rad_Austronesian
Pray_pry_Austroasiatic
Ruc_scb_Austroasiatic
Pear_pcb_Austroasiatic
Danau_dnu_Austroasiatic
Lungchow_zzj_Tai-Kadai
Be_onb_Tai-Kadai
Khün_kkh_Tai-Kadai
Wu-ming_zyb_Tai-Kadai
Ksingmul_puo_Austroasiatic
Haininh_mji_Hmong-Mien
Turung_try_Tibeto-Burman
Laha_lha_Tai-Kadai
Haroi_hro_Austronesian
Karen_Pho_Sangkhlaburi_pww_Tibeto-Burman

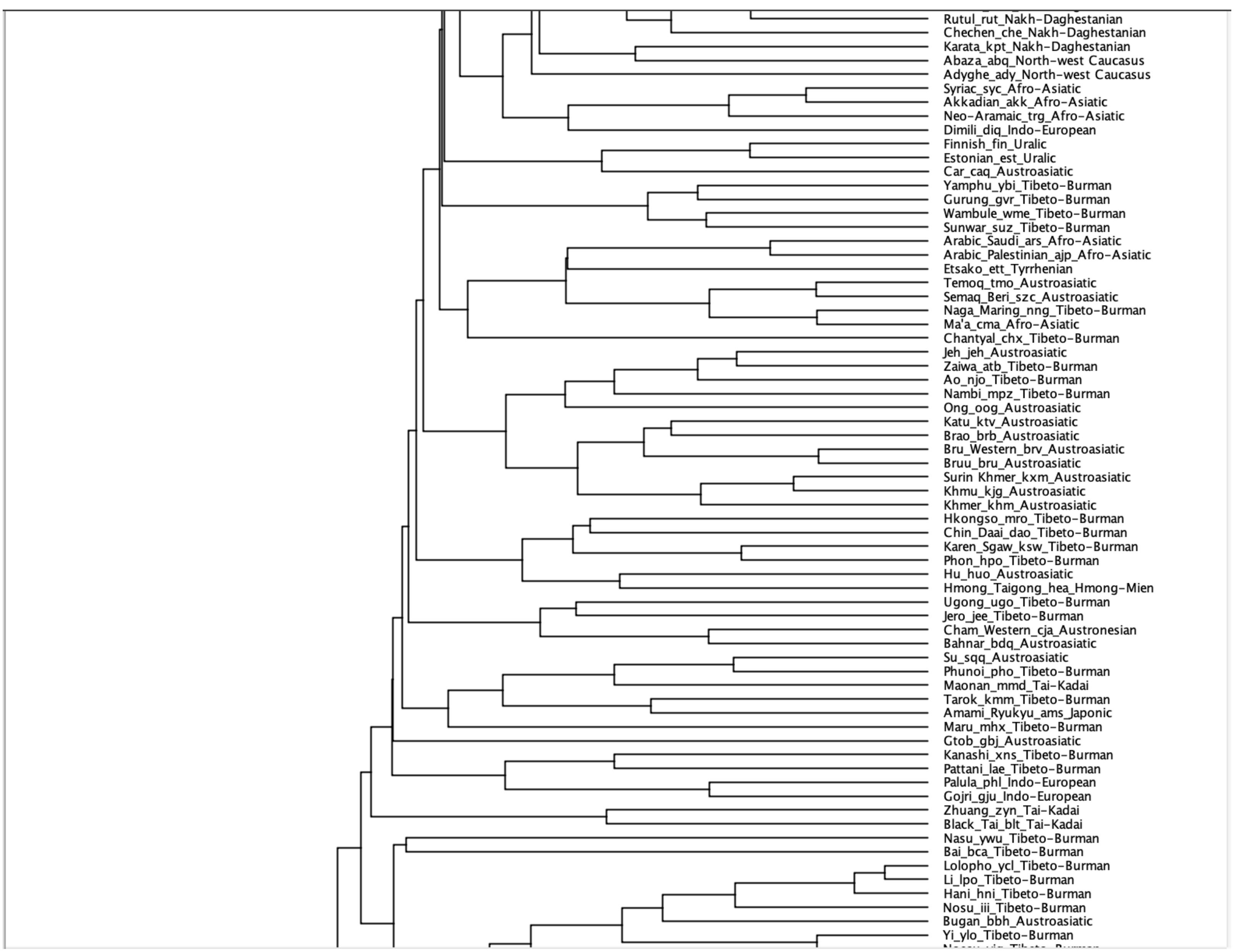
S2.6


Karen_Pho_Sangkhlaburi_pww_Tibeto-Burman
Karen_Pho_pwo_Tibeto-Burman
Karen_bwe_Tibeto-Burman
Phlong_kjp_Tibeto-Burman
Tai Aiton_aio_Tai-Kadai
Ahom_aho_Tai-Kadai
Thai_tha_Tai-Kadai
Paang_pkh_Tibeto-Burman
Thai_Northern_nod_Tai-Kadai
Kadu_kdv_Tibeto-Burman
Karen_Thaungthu_blk_Tibeto-Burman
Kayah_Li_eky_Tibeto-Burman
Old Mon_omx_Austroasiatic
Saek_skb_Tai-Kadai
Yay_pcc_Tai-Kadai
Nung_nut_Tai-Kadai
Shom_Peng_sii_Austroasiatic
Thavung_thm_Austroasiatic
Lü_khb_Tai-Kadai
Kam_kmc_Tai-Kadai
Wancho_nnp_Tibeto-Burman
Bawm_bgr_Tibeto-Burman
Rabha_rah_Tibeto-Burman
Cheng_jeg_Austroasiatic
Ngeq_ngt_Austroasiatic
Even_eve_Tungusic
Rengao_ren_Austroasiatic
Hre_hre_Austroasiatic
Cua_cua_Austroasiatic
Amwi_aml_Austroasiatic
Lahu_lhu_Tibeto-Burman
Puroik_suv_Tibeto-Burman
Itelmen_itl_Chukotko-Kamchatkan
Olekha_ole_Tibeto-Burman
Mailu_npo_Tibeto-Burman
Nancowry_ncb_Austroasiatic
White_Tai_twh_Tai-Kadai
Gothic_got_Indo-European
Yenets_enf_Uralic
Avestan_ave_Indo-European
Yidgha_ydg_Indo-European
Yazgulami_yah_Indo-European
Wakhi_wbl_Indo-European
Ukrainian_ukr_Indo-European
Bulgarian_bul_Indo-European
Sorbian_dsb_Indo-European
Gorum_pcj_Austroasiatic
Kharia_khr_Austroasiatic
Lyngngam_lyg_Austroasiatic
Mah_Meri_mhe_Austroasiatic
Che'_Wong_cwg_Austroasiatic
Pyen_pyy_Tibeto-Burman
Samtao_stu_Austroasiatic
Mulao_mlm_Tai-Kadai
Sho_csh_Tibeto-Burman
Ai-cham_Diwo_aih_Tai-Kadai
Bafut_nri_Tibeto-Burman
Naga_Mao_nbi_Tibeto-Burman
Lakher_mrh_Tibeto-Burman
Chin_Falam_cfm_Tibeto-Burman
Anal_anm_Tibeto-Burman
Deang_Guangka_rbb_Austroasiatic
Mizo_lus_Tibeto-Burman
Hmar_hmr_Tibeto-Burman
Sami_Inari_smn_Uralic
Kildin Sami_sjd_Uralic
Khezha_nkh_Tibeto-Burman

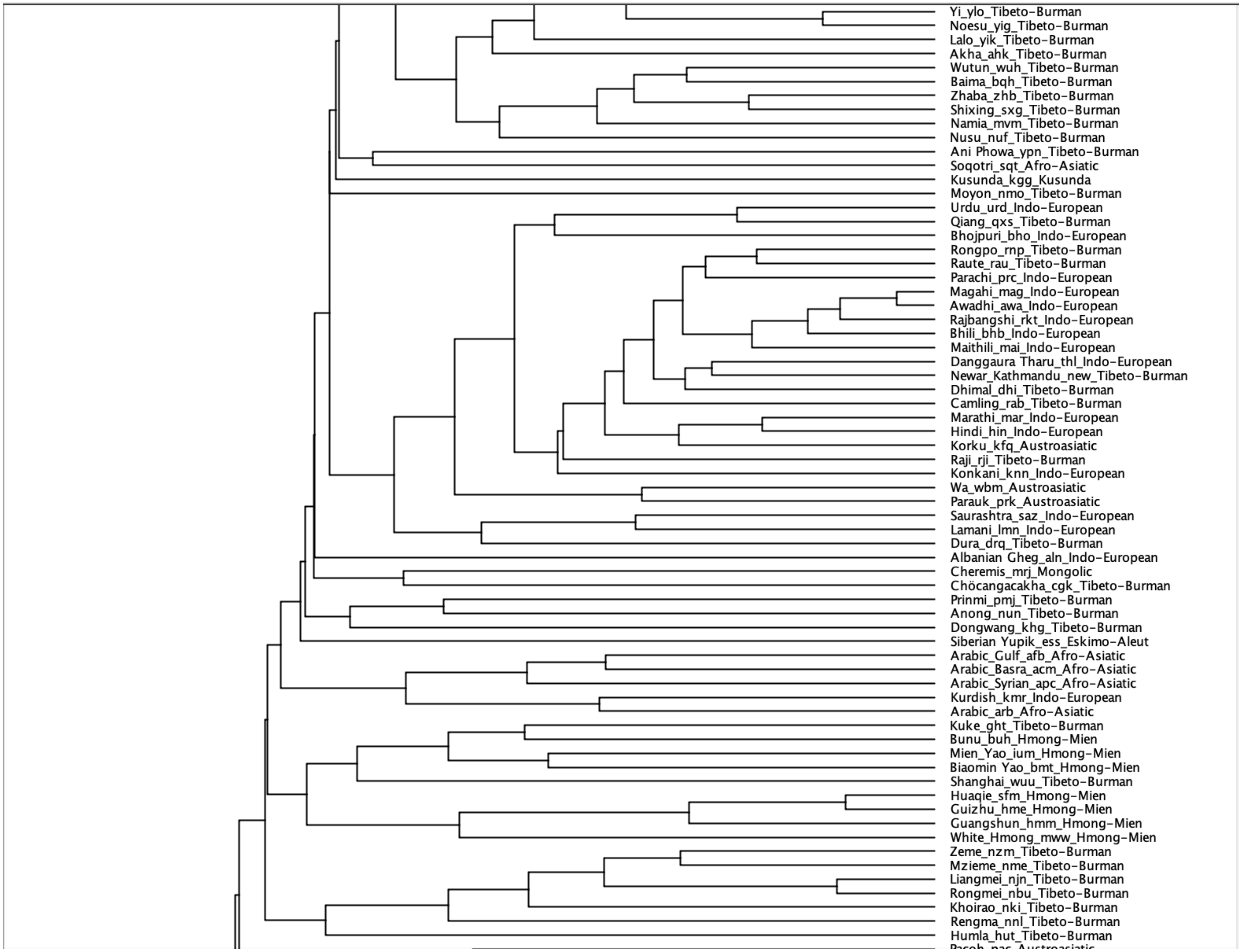
S2.7


Khezha_nkh_Tibeto-Burman
Burmese_mya_Tibeto-Burman
Mlabri_mra_Austroasiatic
Ladakhi_lbj_Tibeto-Burman
Tamang_tge_Tibeto-Burman
Tamang_Eastern_taj_Tibeto-Burman
Tamang_Western_tdg_Tibeto-Burman
Ghale_ghe_Tibeto-Burman
Thakali_ths_Tibeto-Burman
Seke_skj_Tibeto-Burman
Jirel_jul_Tibeto-Burman
Nar_Phu_npa_Tibeto-Burman
Kurtöp_xkz_Tibeto-Burman
Dakpa_dka_Tibeto-Burman
Khengkha_xkf_Tibeto-Burman
Phobjip_neh_Tibeto-Burman
Chali_tgf_Tibeto-Burman
Bumthang_kjz_Tibeto-Burman
Yohlmo_scp_Tibeto-Burman
Dzala_dzl_Tibeto-Burman
Byansi_bee_Tibeto-Burman
Tsum_ttz_Tibeto-Burman
Mugom_muk_Tibeto-Burman
Nubri_kte_Tibeto-Burman
Manange_nmm_Tibeto-Burman
Ghale_Uiya_ghh_Tibeto-Burman
Sherpa_xsr_Tibeto-Burman
Magar_mgp_Tibeto-Burman
Denjongkha_sip_Tibeto-Burman
Mustang_loy_Tibeto-Burman
Dzongkha_dzo_Tibeto-Burman
Lhomi_lhm_Tibeto-Burman
Lhasa_Tibetan_bod_Tibeto-Burman
Xvarshi_khv_Nakh-Daghestanian
Andi_ani_Nakh-Daghestanian
Hinukh_gin_Nakh-Daghestanian
Hunzib_huz_Nakh-Daghestanian
Tsez_ddo_Nakh-Daghestanian
Ingush_inh_Nakh-Daghestanian
Bezhta_kap_Nakh-Daghestanian
Tsova-Tush_bbl_Kartvelian
Tindi_tin_Nakh-Daghestanian
Chamalal_Gigatil_cji_Nakh-Daghestanian
Axvax_akv_Nakh-Daghestanian
Avar_ava_Nakh-Daghestanian
Agul_agx_Nakh-Daghestanian
Ossetian_oss_Indo-European
Armenian_hye_Indo-European
Zan_Mingrelian_xmf_Kartvelian
Georgian_kat_Kartvelian
Tajik_tgk_Indo-European
Svan_sva_Kartvelian
Laz_lzz_Kartvelian
Xinalug_kjj_Nakh-Daghestanian
Tabasaran_Djubek_tab_Nakh-Daghestanian
Lezgian_lez_Nakh-Daghestanian
Godoberi_gdo_Nakh-Daghestanian
Botlikh_bph_Nakh-Daghestanian
Bagulal_kva_Nakh-Daghestanian
Udi_udi_Nakh-Daghestanian
Dargwa_dar_Nakh-Daghestanian
Tsakhur_tkr_Nakh-Daghestanian
Lak_lbe_Nakh-Daghestanian
Archi_aqc_Nakh-Daghestanian
Kryts_kry_Nakh-Daghestanian
Budux_bdk_Nakh-Daghestanian
Rutul_rut_Nakh-Daghestanian

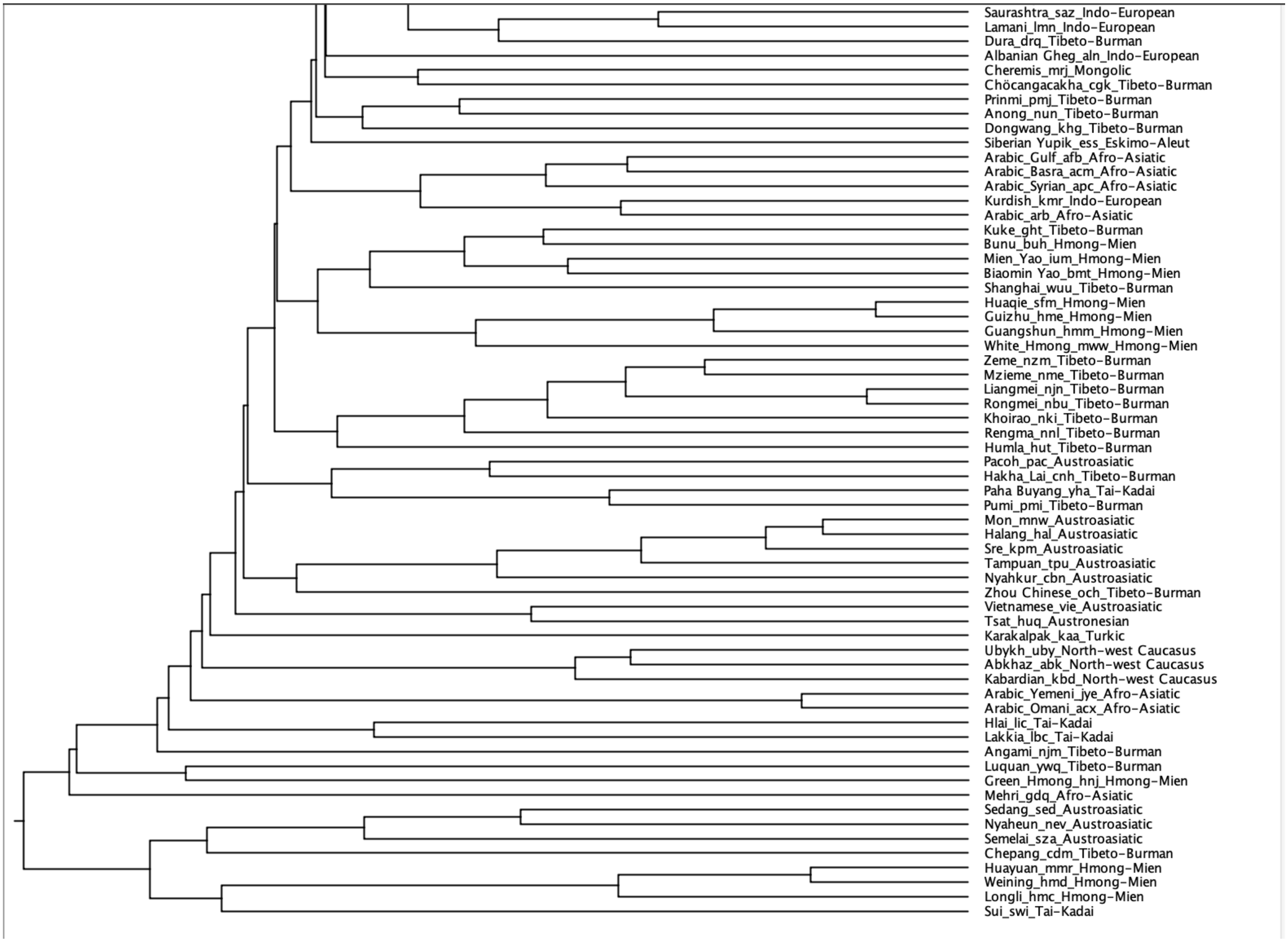
S2.8


Rutul_rut_Nakh-Daghestanian
Chechen_che_Nakh-Daghestanian
Karata_kpt_Nakh-Daghestanian
Abaza_abq_North-west Caucasus
Adyghe_ady_North-west Caucasus
Syriac_syc_Afro-Asiatic
Akkadian_akk_Afro-Asiatic
Neo-Aramaic_trg_Afro-Asiatic
Dimili_diq_Indo-European
Finnish_fin_Uralic
Estonian_est_Uralic
Car_caq_Austroasiatic
Yamphu_ybi_Tibeto-Burman
Gurung_gvr_Tibeto-Burman
Wambule_wme_Tibeto-Burman
Sunwar_suz_Tibeto-Burman
Arabic_Saudi_ars_Afro-Asiatic
Arabic_Palestinian_ajp_Afro-Asiatic
Etsako_ett_Tyrrhenian
Temoq_tmo_Austroasiatic
Semaq_Beri_szc_Austroasiatic
Naga_Maring_nng_Tibeto-Burman
Ma'a_cma_Afro-Asiatic
Chantyal_chx_Tibeto-Burman
Jeh_jeh_Austroasiatic
Zaiwa_atb_Tibeto-Burman
Ao_njo_Tibeto-Burman
Nambi_mpz_Tibeto-Burman
Ong_oog_Austroasiatic
Katu_ktv_Austroasiatic
Brao_brb_Austroasiatic
Bru_Western_brv_Austroasiatic
Bruu_bru_Austroasiatic
Surin Khmer_kxm_Austroasiatic
Khmu_kjg_Austroasiatic
Khmer_khm_Austroasiatic
Hkongso_mro_Tibeto-Burman
Chin_Daai_dao_Tibeto-Burman
Karen_Sgaw_ksw_Tibeto-Burman
Phon_hpo_Tibeto-Burman
Hu_huo_Austroasiatic
Hmong_Taigong_hea_Hmong-Mien
Ugong_ugo_Tibeto-Burman
Jero_jee_Tibeto-Burman
Cham_Western_cja_Austronesian
Bahnar_bdq_Austroasiatic
Su_sqq_Austroasiatic
Phunoi_pho_Tibeto-Burman
Maonan_mmd_Tai-Kadai
Tarok_kmm_Tibeto-Burman
Amami_Ryukyu_ams_Japonic
Maru_mhx_Tibeto-Burman
Gtob_gbj_Austroasiatic
Kanashi_xns_Tibeto-Burman
Pattani_lae_Tibeto-Burman
Palula_phl_Indo-European
Gojri_gju_Indo-European
Zhuang_zyn_Tai-Kadai
Black_Tai_blt_Tai-Kadai
Nasu_ywu_Tibeto-Burman
Bai_bca_Tibeto-Burman
Lolopho_ycl_Tibeto-Burman
Li_lpo_Tibeto-Burman
Hani_hni_Tibeto-Burman
Nosu_iii_Tibeto-Burman
Bugan_bbh_Austroasiatic
Yi_ylo_Tibeto-Burman

S2.9


Yi_ylo_Tibeto-Burman
Noesu_yig_Tibeto-Burman
Lalo_yik_Tibeto-Burman
Akha_ahk_Tibeto-Burman
Wutun_wuh_Tibeto-Burman
Baima_bqh_Tibeto-Burman
Zhaba_zhb_Tibeto-Burman
Shixing_sxg_Tibeto-Burman
Namia_mvm_Tibeto-Burman
Nusu_nuf_Tibeto-Burman
Ani Phowa_ypn_Tibeto-Burman
Soqotri_sqt_Afro-Asiatic
Kusunda_kgg_Kusunda
Moyon_nmo_Tibeto-Burman
Urdu_urd_Indo-European
Qiang_qxs_Tibeto-Burman
Bhojpuri_bho_Indo-European
Rongpo_rnp_Tibeto-Burman
Raute_rau_Tibeto-Burman
Parachi_prc_Indo-European
Magahi_mag_Indo-European
Awadhi_awa_Indo-European
Rajbangshi_rkt_Indo-European
Bhili_bhb_Indo-European
Maithili_mai_Indo-European
Danggaura Tharu_thl_Indo-European
Newar_Kathmandu_new_Tibeto-Burman
Dhimal_dhi_Tibeto-Burman
Camling_rab_Tibeto-Burman
Marathi_mar_Indo-European
Hindi_hin_Indo-European
Korku_kfq_Austroasiatic
Raji_rji_Tibeto-Burman
Konkani_knn_Indo-European
Wa_wbm_Austroasiatic
Parauk_prk_Austroasiatic
Saurashtra_saz_Indo-European
Lamani_lmn_Indo-European
Dura_drq_Tibeto-Burman
Albanian Gheg_aln_Indo-European
Cheremis_mrj_Mongolic
Chöcangacakha_cgk_Tibeto-Burman
Prinmi_pmj_Tibeto-Burman
Anong_nun_Tibeto-Burman
Dongwang_khg_Tibeto-Burman
Siberian Yupik_ess_Eskimo-Aleut
Arabic_Gulf_afb_Afro-Asiatic
Arabic_Basra_acm_Afro-Asiatic
Arabic_Syrian_apc_Afro-Asiatic
Kurdish_kmr_Indo-European
Arabic_arb_Afro-Asiatic
Kuke_ght_Tibeto-Burman
Bunu_buh_Hmong-Mien
Mien_Yao_ium_Hmong-Mien
Biaomin Yao_bmt_Hmong-Mien
Shanghai_wuu_Tibeto-Burman
Huaqie_sfm_Hmong-Mien
Guizhu_hme_Hmong-Mien
Guangshun_hmm_Hmong-Mien
White_Hmong_mww_Hmong-Mien
Zeme_nzm_Tibeto-Burman
Mzieme_nme_Tibeto-Burman
Liangmei_njn_Tibeto-Burman
Rongmei_nbu_Tibeto-Burman
Khoirao_nki_Tibeto-Burman
Rengma_nnl_Tibeto-Burman
Humla_hut_Tibeto-Burman
Pacoh_pac_Austroasiatic

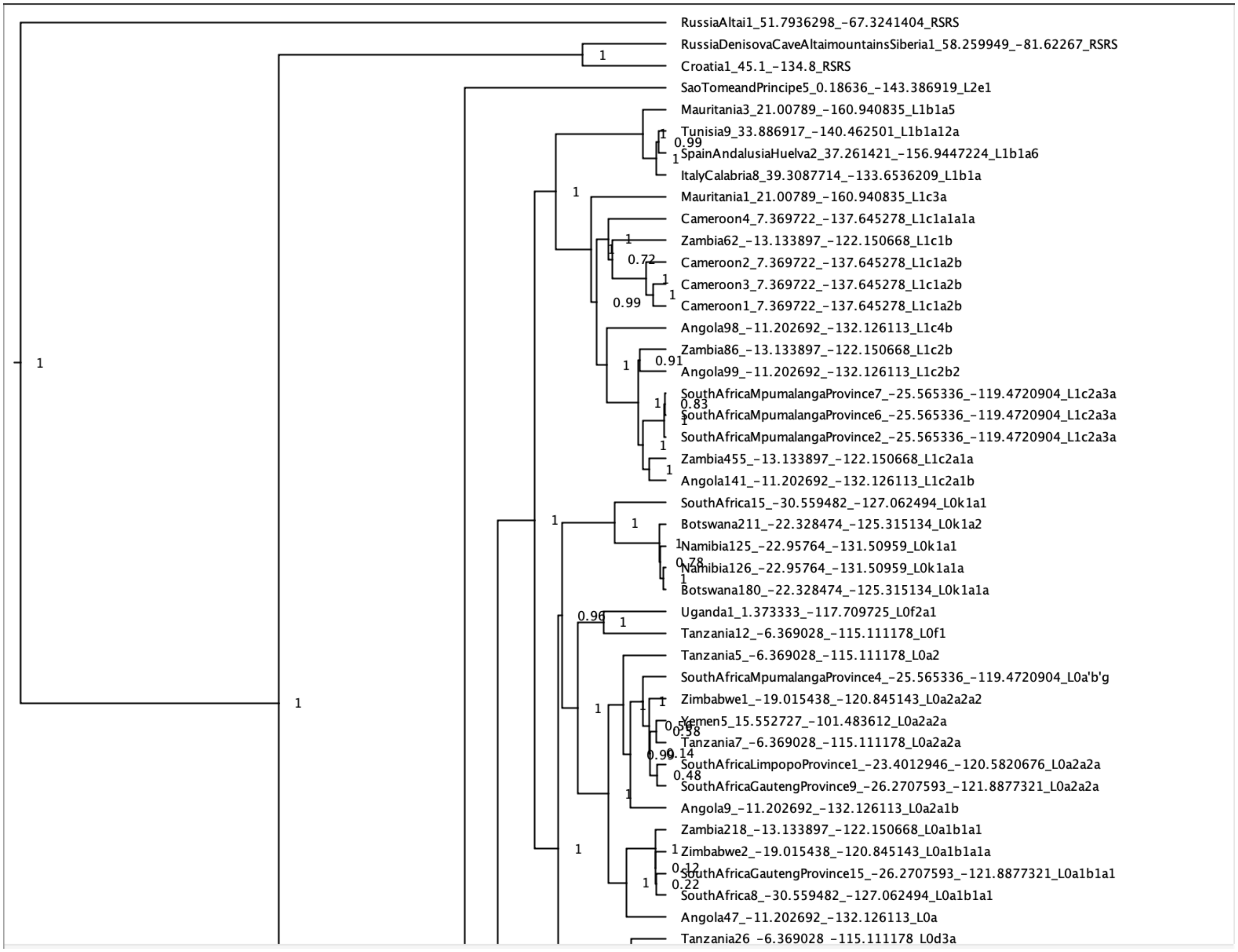
S2.10


Saurashtra_saz_Indo-European
Lamani_lmn_Indo-European
Dura_drq_Tibeto-Burman
Albanian Gheg_aln_Indo-European
Cheremis_mrj_Mongolic
Chöcangacakha_cgk_Tibeto-Burman
Prinmi_pmj_Tibeto-Burman
Anong_nun_Tibeto-Burman
Dongwang_khg_Tibeto-Burman
Siberian Yupik_ess_Eskimo-Aleut
Arabic_Gulf_afb_Afro-Asiatic
Arabic_Basra_acm_Afro-Asiatic
Arabic_Syrian_apc_Afro-Asiatic
Kurdish_kmr_Indo-European
Arabic_arb_Afro-Asiatic
Kuke_ght_Tibeto-Burman
Bunu_buh_Hmong-Mien
Mien_Yao_ium_Hmong-Mien
Biaomin Yao_bmt_Hmong-Mien
Shanghai_wuu_Tibeto-Burman
Huaqie_sfm_Hmong-Mien
Guizhu_hme_Hmong-Mien
Guangshun_hmm_Hmong-Mien
White_Hmong_mww_Hmong-Mien
Zeme_nzm_Tibeto-Burman
Mzieme_nme_Tibeto-Burman
Liangmei_njn_Tibeto-Burman
Rongmei_nbu_Tibeto-Burman
Khoirao_nki_Tibeto-Burman
Rengma_nnl_Tibeto-Burman
Humla_hut_Tibeto-Burman
Pacoh_pac_Austroasiatic
Hakha_Lai_cnh_Tibeto-Burman
Paha Buyang_yha_Tai-Kadai
Pumi_pmi_Tibeto-Burman
Mon_mnw_Austroasiatic
Halang_hal_Austroasiatic
Sre_kpm_Austroasiatic
Tampuan_tpu_Austroasiatic
Nyahkur_cbn_Austroasiatic
Zhou Chinese_och_Tibeto-Burman
Vietnamese_vie_Austroasiatic
Tsat_huq_Austronesian
Karakalpak_kaa_Turkic
Ubykh_uby_North-west Caucasus
Abkhaz_abk_North-west Caucasus
Kabardian_kbd_North-west Caucasus
Arabic_Yemeni_jye_Afro-Asiatic
Arabic_Omani_acx_Afro-Asiatic
Hlai_lic_Tai-Kadai
Lakkia_lbc_Tai-Kadai
Angami_njm_Tibeto-Burman
Luquan_ywq_Tibeto-Burman
Green_Hmong_hnj_Hmong-Mien
Mehri_gdq_Afro-Asiatic
Sedang_sed_Austroasiatic
Nyaheun_nev_Austroasiatic
Semelai_sza_Austroasiatic
Chepang_cdm_Tibeto-Burman
Huayuan_mmr_Hmong-Mien
Weining_hmd_Hmong-Mien
Longli_hmc_Hmong-Mien
Sui_swi_Tai-Kadai

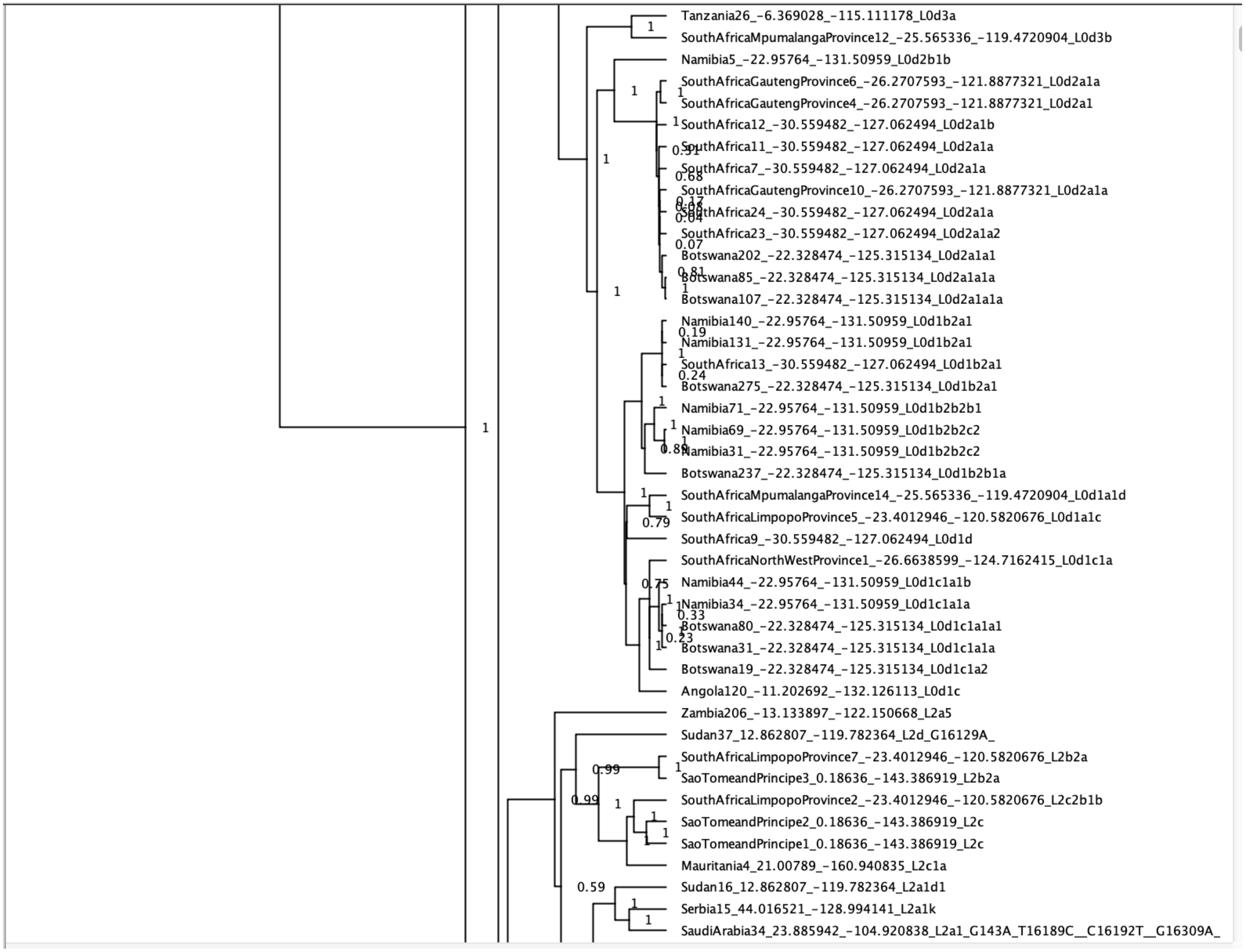
S2.11

Supplementary Figure S3:1-52 MtDNA Maximum Clade Credibility Tree

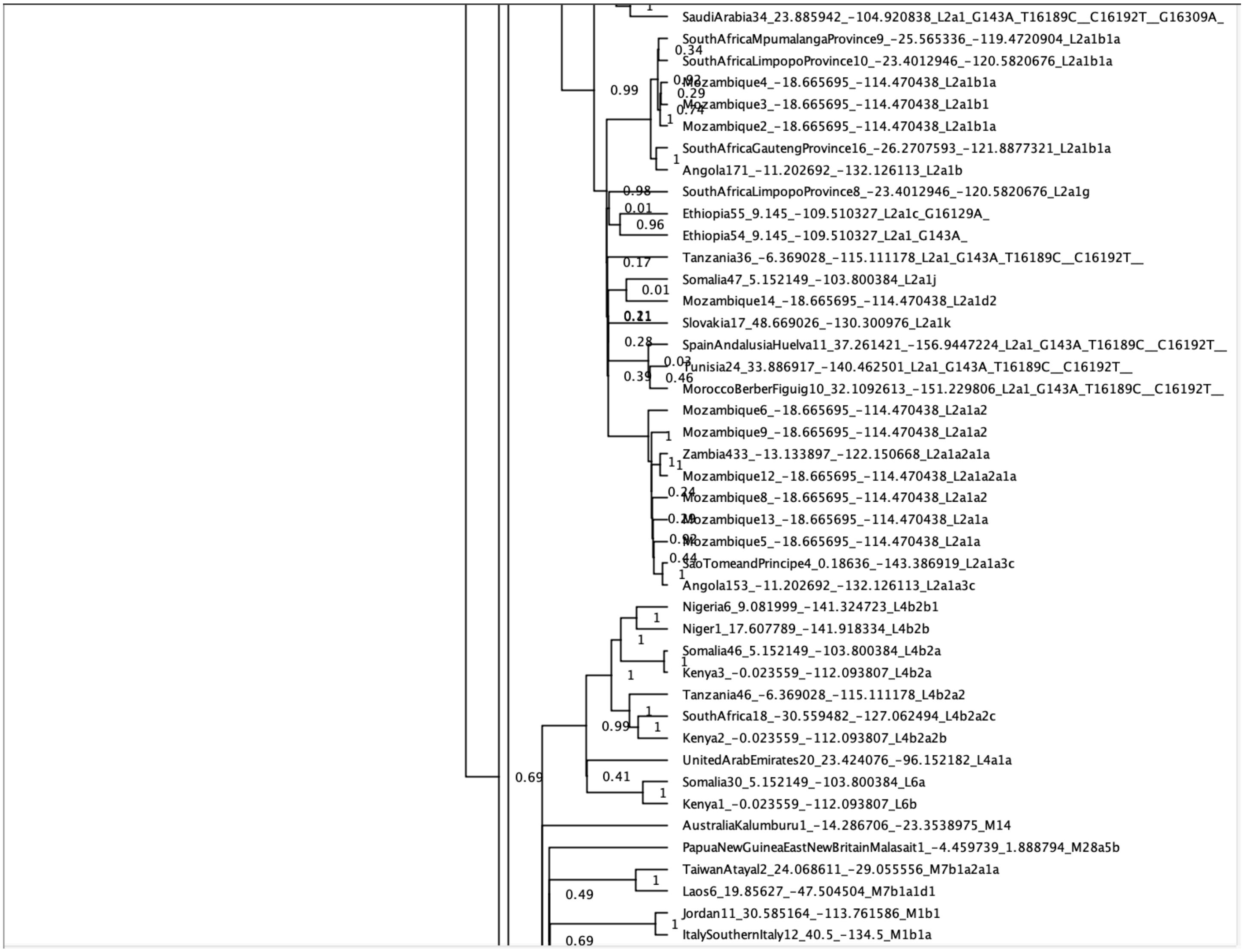
Numbers on nodes are posterior probabilities; each taxon has the format {location name}_{latitude}_{longitude, transformed in order to center the Pacific}_{haplogroup according to Mthap}.


RussiaAltai1_51.7936298_-67.3241404_RSRS
RussiaDenisovaCaveAltaimountainsSiberia1_58.259949_-81.62267_RSRS
1
Croatia1_45.1_-134.8_RSRS
SaoTomeandPrincipe5_0.18636_-143.386919_L2e1
Mauritania3_21.00789_-160.940835_L1b1a5
1
Tunisia9_33.886917_-140.462501_L1b1a12a
0.99
SpainAndalusiaHuelva2_37.261421_-156.9447224_L1b1a6
1
ItalyCalabria8_39.3087714_-133.6536209_L1b1a
1
Mauritania1_21.00789_-160.940835_L1c3a
Cameroon4_7.369722_-137.645278_L1c1a1a1a
1
Zambia62_-13.133897_-122.150668_L1c1b
1
0.72
Cameroon2_7.369722_-137.645278_L1c1a2b
1
Cameroon3_7.369722_-137.645278_L1c1a2b
1
0.99
Cameroon1_7.369722_-137.645278_L1c1a2b
Angola98_-11.202692_-132.126113_L1c4b
1
Zambia86_-13.133897_-122.150668_L1c2b
0.91
Angola99_-11.202692_-132.126113_L1c2b2
SouthAfricaMpumalangaProvince7_-25.565336_-119.4720904_L1c2a3a
1
0.83
SouthAfricaMpumalangaProvince6_-25.565336_-119.4720904_L1c2a3a
1
SouthAfricaMpumalangaProvince2_-25.565336_-119.4720904_L1c2a3a
1
Zambia455_-13.133897_-122.150668_L1c2a1a
1
Angola141_-11.202692_-132.126113_L1c2a1b
1
SouthAfrica15_-30.559482_-127.062494_L0k1a1
1
1
Botswana211_-22.328474_-125.315134_L0k1a2
1
Namibia125_-22.95764_-131.50959_L0k1a1
0.78
Namibia126_-22.95764_-131.50959_L0k1a1a
1
Botswana180_-22.328474_-125.315134_L0k1a1a
0.96
Uganda1_1.373333_-117.709725_L0f2a1
1
Tanzania12_-6.369028_-115.111178_L0f1
1
Tanzania5_-6.369028_-115.111178_L0a2
SouthAfricaMpumalangaProvince4_-25.565336_-119.4720904_L0a'b'g
1
1
1
Zimbabwe1_-19.015438_-120.845143_L0a2a2a2
0.56
Yemen5_15.552727_-101.483612_L0a2a2a
0.58
Tanzania7_-6.369028_-115.111178_L0a2a2a
0.99
0.14
SouthAfricaLimpopoProvince1_-23.4012946_-120.5820676_L0a2a2a
0.48
SouthAfricaGautengProvince9_-26.2707593_-121.8877321_L0a2a2a
1
Angola9_-11.202692_-132.126113_L0a2a1b
Zambia218_-13.133897_-122.150668_L0a1b1a1
1
1
Zimbabwe2_-19.015438_-120.845143_L0a1b1a1a
0.12
SouthAfricaGautengProvince15_-26.2707593_-121.8877321_L0a1b1a1
1
0.22
SouthAfrica8_-30.559482_-127.062494_L0a1b1a1
Angola47_-11.202692_-132.126113_L0a
Tanzania26 -6.369028 -115.111178 L0d3a

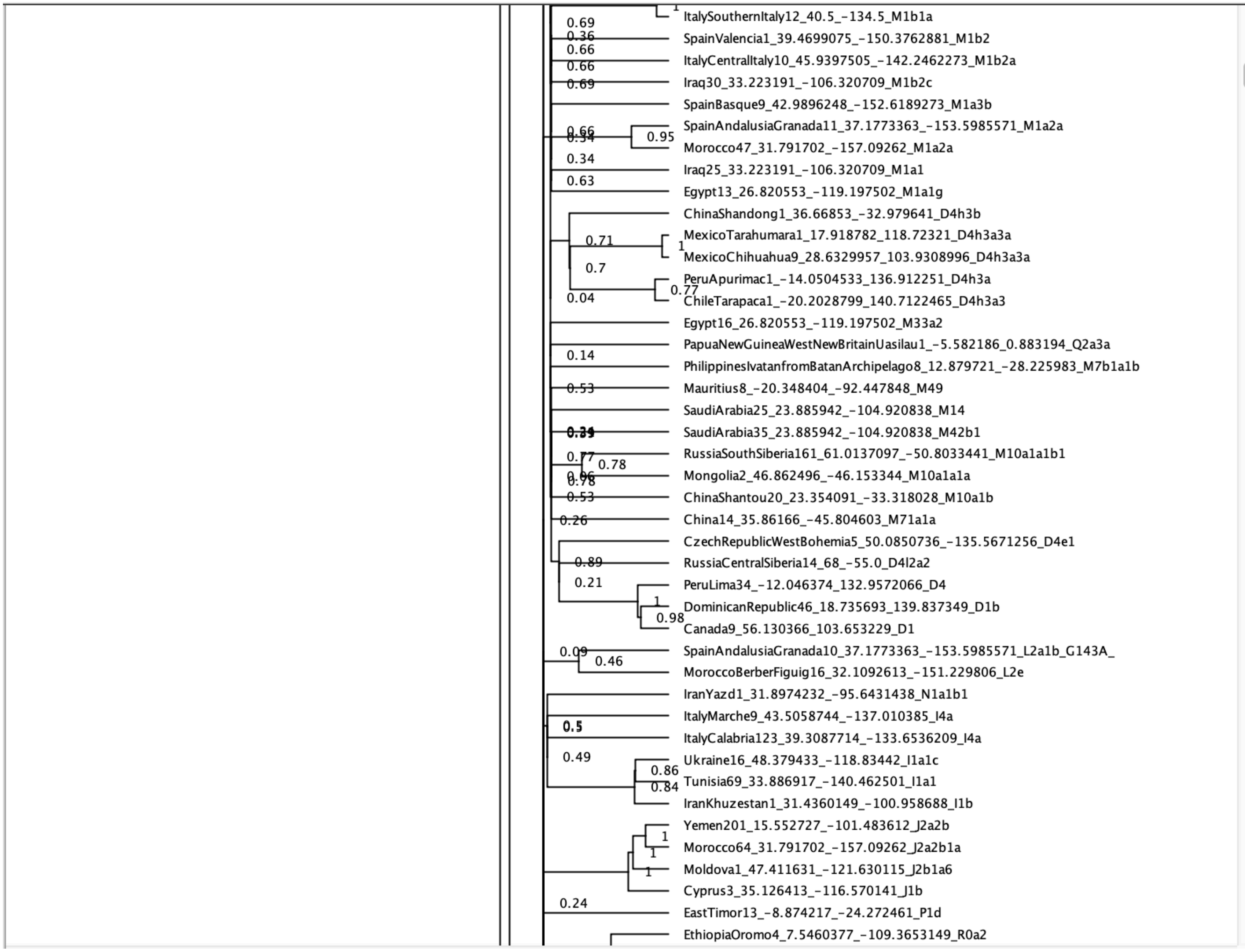
S3.1


Tanzania26_-6.369028_-115.111178_L0d3a
1
SouthAfricaMpumalangaProvince12_-25.565336_-119.4720904_L0d3b
Namibia5_-22.95764_-131.50959_L0d2b1b
1
SouthAfricaGautengProvince6_-26.2707593_-121.8877321_L0d2a1a
1
SouthAfricaGautengProvince4_-26.2707593_-121.8877321_L0d2a1
1
SouthAfrica12_-30.559482_-127.062494_L0d2a1b
SouthAfrica11_-30.559482_-127.062494_L0d2a1a
0.31
1
SouthAfrica7_-30.559482_-127.062494_L0d2a1a
0.68
SouthAfricaGautengProvince10_-26.2707593_-121.8877321_L0d2a1a
0.17
0.08
SouthAfrica24_-30.559482_-127.062494_L0d2a1a
0.04
SouthAfrica23_-30.559482_-127.062494_L0d2a1a2
0.07
Botswana202_-22.328474_-125.315134_L0d2a1a1
0.81
Botswana85_-22.328474_-125.315134_L0d2a1a1a
1
1
Botswana107_-22.328474_-125.315134_L0d2a1a1a
Namibia140_-22.95764_-131.50959_L0d1b2a1
0.19
Namibia131_-22.95764_-131.50959_L0d1b2a1
1
SouthAfrica13_-30.559482_-127.062494_L0d1b2a1
0.24
Botswana275_-22.328474_-125.315134_L0d1b2a1
1
Namibia71_-22.95764_-131.50959_L0d1b2b2b1
1
1
Namibia69_-22.95764_-131.50959_L0d1b2b2c2
1
0.89
Namibia31_-22.95764_-131.50959_L0d1b2b2c2
Botswana237_-22.328474_-125.315134_L0d1b2b1a
1
SouthAfricaMpumalangaProvince14_-25.565336_-119.4720904_L0d1a1d
1
0.79
SouthAfricaLimpopoProvince5_-23.4012946_-120.5820676_L0d1a1c
SouthAfrica9_-30.559482_-127.062494_L0d1d
SouthAfricaNorthWestProvince1_-26.6638599_-124.7162415_L0d1c1a
0.75
Namibia44_-22.95764_-131.50959_L0d1c1a1b
1
1
Namibia34_-22.95764_-131.50959_L0d1c1a1a
0.33
Botswana80_-22.328474_-125.315134_L0d1c1a1a1
0.23
1
Botswana31_-22.328474_-125.315134_L0d1c1a1a
Botswana19_-22.328474_-125.315134_L0d1c1a2
Angola120_-11.202692_-132.126113_L0d1c
Zambia206_-13.133897_-122.150668_L2a5
Sudan37_12.862807_-119.782364_L2d_G16129A_
1
SouthAfricaLimpopoProvince7_-23.4012946_-120.5820676_L2b2a
0.99
1
SaoTomeandPrincipe3_0.18636_-143.386919_L2b2a
SouthAfricaLimpopoProvince2_-23.4012946_-120.5820676_L2c2b1b
0.99
1
1
SaoTomeandPrincipe2_0.18636_-143.386919_L2c
1
1
SaoTomeandPrincipe1_0.18636_-143.386919_L2c
Mauritania4_21.00789_-160.940835_L2c1a
0.59
Sudan16_12.862807_-119.782364_L2a1d1
1
Serbia15_44.016521_-128.994141_L2a1k
1
SaudiArabia34_23.885942_-104.920838_L2a1_G143A_T16189C__C16192T__G16309A_

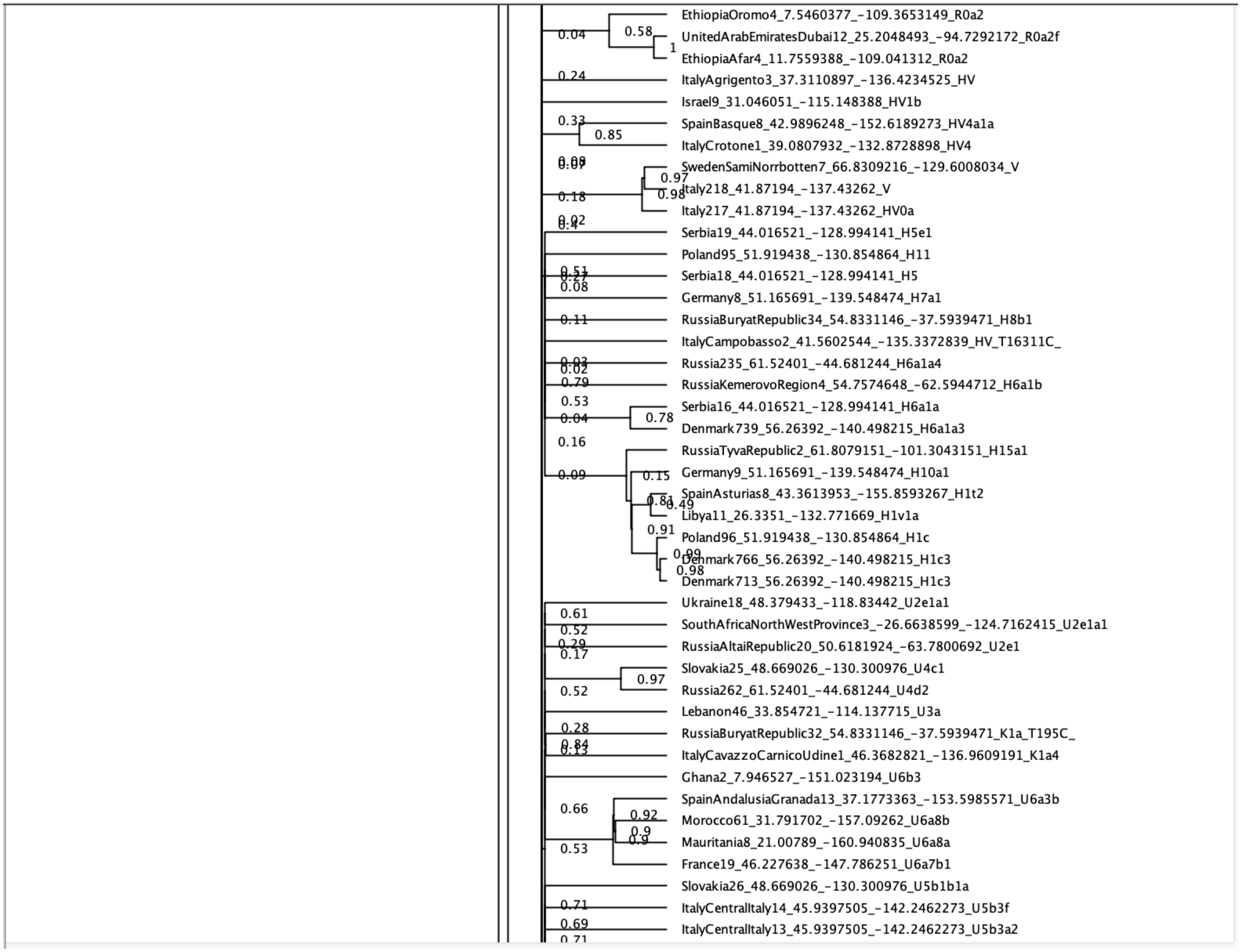
S3.2


1
SaudiArabia34_23.885942_-104.920838_L2a1_G143A_T16189C__C16192T__G16309A_
SouthAfricaMpumalangaProvince9_-25.565336_-119.4720904_L2a1b1a
0.34
SouthAfricaLimpopoProvince10_-23.4012946_-120.5820676_L2a1b1a
0.99
0.92
Mozambique4_-18.665695_-114.470438_L2a1b1a
0.29
Mozambique3_-18.665695_-114.470438_L2a1b1
0.74
1
Mozambique2_-18.665695_-114.470438_L2a1b1a
SouthAfricaGautengProvince16_-26.2707593_-121.8877321_L2a1b1a
1
Angola171_-11.202692_-132.126113_L2a1b
0.98
SouthAfricaLimpopoProvince8_-23.4012946_-120.5820676_L2a1g
0.01
Ethiopia55_9.145_-109.510327_L2a1c_G16129A_
0.96
Ethiopia54_9.145_-109.510327_L2a1_G143A_
Tanzania36_-6.369028_-115.111178_L2a1_G143A_T16189C__C16192T__
0.17
Somalia47_5.152149_-103.800384_L2a1j
0.01
Mozambique14_-18.665695_-114.470438_L2a1d2
0.21
Slovakia17_48.669026_-130.300976_L2a1k
0.28
SpainAndalusiaHuelva11_37.261421_-156.9447224_L2a1_G143A_T16189C__C16192T__
0.03
Tunisia24_33.886917_-140.462501_L2a1_G143A_T16189C__C16192T__
0.39
0.46
MoroccoBerberFiguig10_32.1092613_-151.229806_L2a1_G143A_T16189C__C16192T__
Mozambique6_-18.665695_-114.470438_L2a1a2
1
Mozambique9_-18.665695_-114.470438_L2a1a2
Zambia433_-13.133897_-122.150668_L2a1a2a1a
1 1
Mozambique12_-18.665695_-114.470438_L2a1a2a1a
0.24
Mozambique8_-18.665695_-114.470438_L2a1a2
0.29
Mozambique13_-18.665695_-114.470438_L2a1a
0.92
Mozambique5_-18.665695_-114.470438_L2a1a
0.44
SaoTomeandPrincipe4_0.18636_-143.386919_L2a1a3c
1
Angola153_-11.202692_-132.126113_L2a1a3c
Nigeria6_9.081999_-141.324723_L4b2b1
1
Niger1_17.607789_-141.918334_L4b2b
1
Somalia46_5.152149_-103.800384_L4b2a
1
1
Kenya3_-0.023559_-112.093807_L4b2a
Tanzania46_-6.369028_-115.111178_L4b2a2
1
SouthAfrica18_-30.559482_-127.062494_L4b2a2c
0.99
1
Kenya2_-0.023559_-112.093807_L4b2a2b
UnitedArabEmirates20_23.424076_-96.152182_L4a1a
0.69
0.41
Somalia30_5.152149_-103.800384_L6a
1
Kenya1_-0.023559_-112.093807_L6b
AustraliaKalumburu1_-14.286706_-23.3538975_M14
PapuaNewGuineaEastNewBritainMalasait1_-4.459739_1.888794_M28a5b
TaiwanAtayal2_24.068611_-29.055556_M7b1a2a1a
1
0.49
Laos6_19.85627_-47.504504_M7b1a1d1
Jordan11_30.585164_-113.761586_M1b1
1
ItalySouthernItaly12_40.5_-134.5_M1b1a
0.69

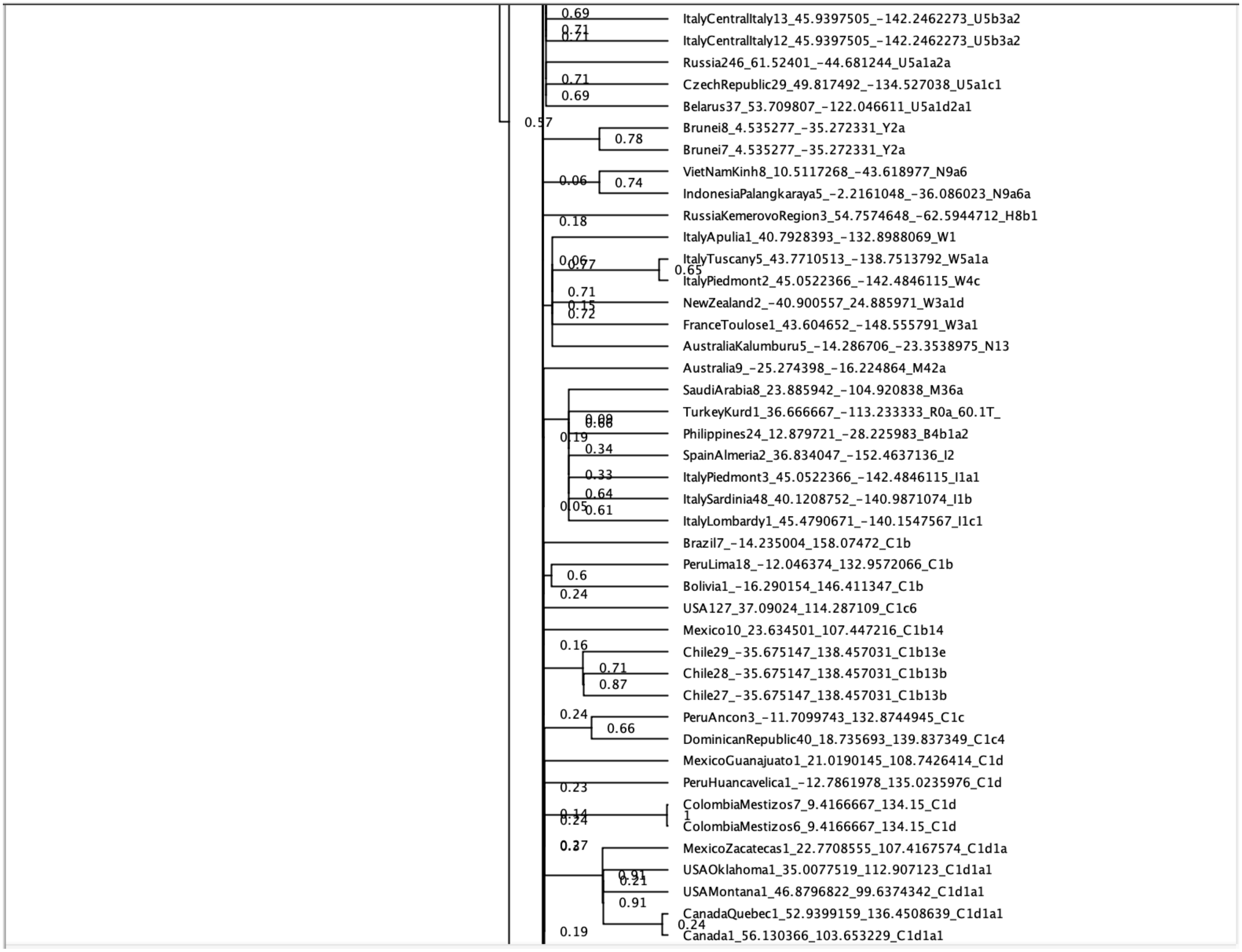
S3.3


1
ItalySouthernItaly12_40.5_-134.5_M1b1a
0.69
0.36
SpainValencia1_39.4699075_-150.3762881_M1b2
0.66
ItalyCentralItaly10_45.9397505_-142.2462273_M1b2a
0.66
0.69
Iraq30_33.223191_-106.320709_M1b2c
SpainBasque9_42.9896248_-152.6189273_M1a3b
0.66
0.34
0.95
SpainAndalusiaGranada11_37.1773363_-153.5985571_M1a2a
Morocco47_31.791702_-157.09262_M1a2a
0.34
Iraq25_33.223191_-106.320709_M1a1
0.63
Egypt13_26.820553_-119.197502_M1a1g
ChinaShandong1_36.66853_-32.979641_D4h3b
0.71
1
MexicoTarahumara1_17.918782_118.72321_D4h3a3a
MexicoChihuahua9_28.6329957_103.9308996_D4h3a3a
0.7
0.77
PeruApurimac1_-14.0504533_136.912251_D4h3a
0.04
ChileTarapaca1_-20.2028799_140.7122465_D4h3a3
Egypt16_26.820553_-119.197502_M33a2
PapuaNewGuineaWestNewBritainUasilau1_-5.582186_0.883194_Q2a3a
0.14
PhilippinesIvatanfromBatanArchipelago8_12.879721_-28.225983_M7b1a1b
0.53
Mauritius8_-20.348404_-92.447848_M49
SaudiArabia25_23.885942_-104.920838_M14
0.31
SaudiArabia35_23.885942_-104.920838_M42b1
0.77
0.78
RussiaSouthSiberia161_61.0137097_-50.8033441_M10a1a1b1
0.06
0.78
Mongolia2_46.862496_-46.153344_M10a1a1a
0.53
ChinaShantou20_23.354091_-33.318028_M10a1b
0.26
China14_35.86166_-45.804603_M71a1a
CzechRepublicWestBohemia5_50.0850736_-135.5671256_D4e1
0.89
RussiaCentralSiberia14_68_-55.0_D4l2a2
0.21
PeruLima34_-12.046374_132.9572066_D4
1
DominicanRepublic46_18.735693_139.837349_D1b
0.98
Canada9_56.130366_103.653229_D1
0.09
0.46
SpainAndalusiaGranada10_37.1773363_-153.5985571_L2a1b_G143A_
MoroccoBerberFiguig16_32.1092613_-151.229806_L2e
IranYazd1_31.8974232_-95.6431438_N1a1b1
ItalyMarche9_43.5058744_-137.010385_I4a
0.5
ItalyCalabria123_39.3087714_-133.6536209_I4a
0.49
Ukraine16_48.379433_-118.83442_I1a1c
0.86
0.84
Tunisia69_33.886917_-140.462501_I1a1
IranKhuzestan1_31.4360149_-100.958688_I1b
Yemen201_15.552727_-101.483612_J2a2b
1
1
Morocco64_31.791702_-157.09262_J2a2b1a
1
Moldova1_47.411631_-121.630115_J2b1a6
Cyprus3_35.126413_-116.570141_J1b
0.24
EastTimor13_-8.874217_-24.272461_P1d
EthiopiaOromo4_7.5460377_-109.3653149_R0a2

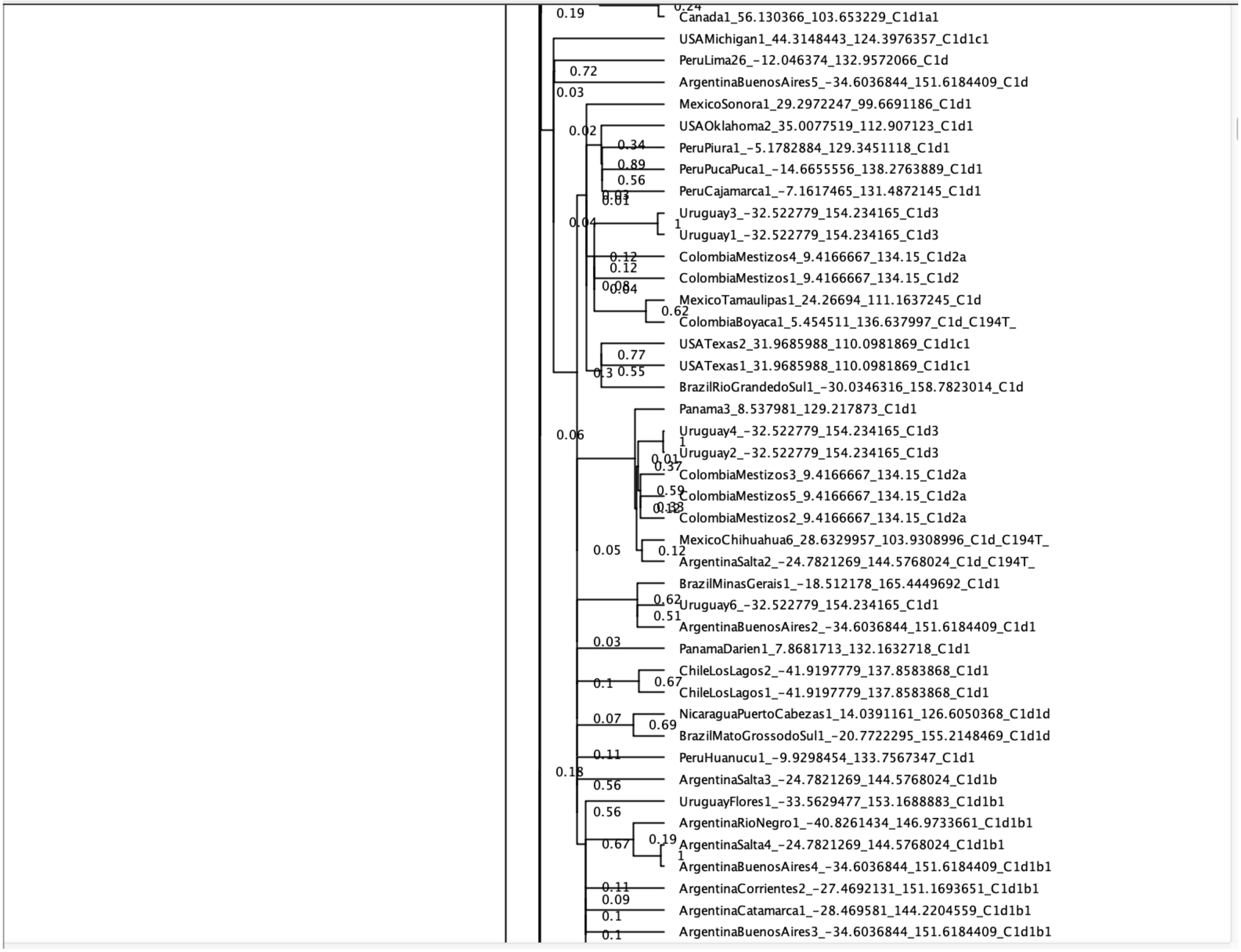
S3.4


EthiopiaOromo4_7.5460377_-109.3653149_R0a2
UnitedArabEmiratesDubai12_25.2048493_-94.7292172_R0a2f
EthiopiaAfar4_11.7559388_-109.041312_R0a2
ItalyAgrigento3_37.3110897_-136.4234525_HV
Israel9_31.046051_-115.148388_HV1b
SpainBasque8_42.9896248_-152.6189273_HV4a1a
ItalyCrotone1_39.0807932_-132.8728898_HV4
SwedenSamiNorrbotten7_66.8309216_-129.6008034_V
Italy218_41.87194_-137.43262_V
Italy217_41.87194_-137.43262_HV0a
Serbia19_44.016521_-128.994141_H5e1
Poland95_51.919438_-130.854864_H11
Serbia18_44.016521_-128.994141_H5
Germany8_51.165691_-139.548474_H7a1
RussiaBuryatRepublic34_54.8331146_-37.5939471_H8b1
ItalyCampobasso2_41.5602544_-135.3372839_HV_T16311C_
Russia235_61.52401_-44.681244_H6a1a4
RussiaKemerovoRegion4_54.7574648_-62.5944712_H6a1b
Serbia16_44.016521_-128.994141_H6a1a
Denmark739_56.26392_-140.498215_H6a1a3
RussiaTyvaRepublic2_61.8079151_-101.3043151_H15a1
Germany9_51.165691_-139.548474_H10a1
SpainAsturias8_43.3613953_-155.8593267_H1t2
Libya11_26.3351_-132.771669_H1v1a
Poland96_51.919438_-130.854864_H1c
Denmark766_56.26392_-140.498215_H1c3
Denmark713_56.26392_-140.498215_H1c3
Ukraine18_48.379433_-118.83442_U2e1a1
SouthAfricaNorthWestProvince3_-26.6638599_-124.7162415_U2e1a1
RussiaAltaiRepublic20_50.6181924_-63.7800692_U2e1
Slovakia25_48.669026_-130.300976_U4c1
Russia262_61.52401_-44.681244_U4d2
Lebanon46_33.854721_-114.137715_U3a
RussiaBuryatRepublic32_54.8331146_-37.5939471_K1a_T195C_
ItalyCavazzoCarnicoUdine1_46.3682821_-136.9609191_K1a4
Ghana2_7.946527_-151.023194_U6b3
SpainAndalusiaGranada13_37.1773363_-153.5985571_U6a3b
Morocco61_31.791702_-157.09262_U6a8b
Mauritania8_21.00789_-160.940835_U6a8a
France19_46.227638_-147.786251_U6a7b1
Slovakia26_48.669026_-130.300976_U5b1b1a
ItalyCentralItaly14_45.9397505_-142.2462273_U5b3f
ItalyCentralItaly13_45.9397505_-142.2462273_U5b3a2
0.04
0.58
1
0.24
0.33
0.85
0.09
0.07
0.97
0.98
0.18
0.02
0.4
0.51
0.27
0.08
0.11
0.03
0.02
0.79
0.53
0.04
0.78
0.16
0.09
0.15
0.81
0.49
0.91
0.99
0.98
0.61
0.52
0.29
0.17
0.97
0.52
0.28
0.84
0.13
0.66
0.92
0.9
0.9
0.53
0.71
0.69
0.71

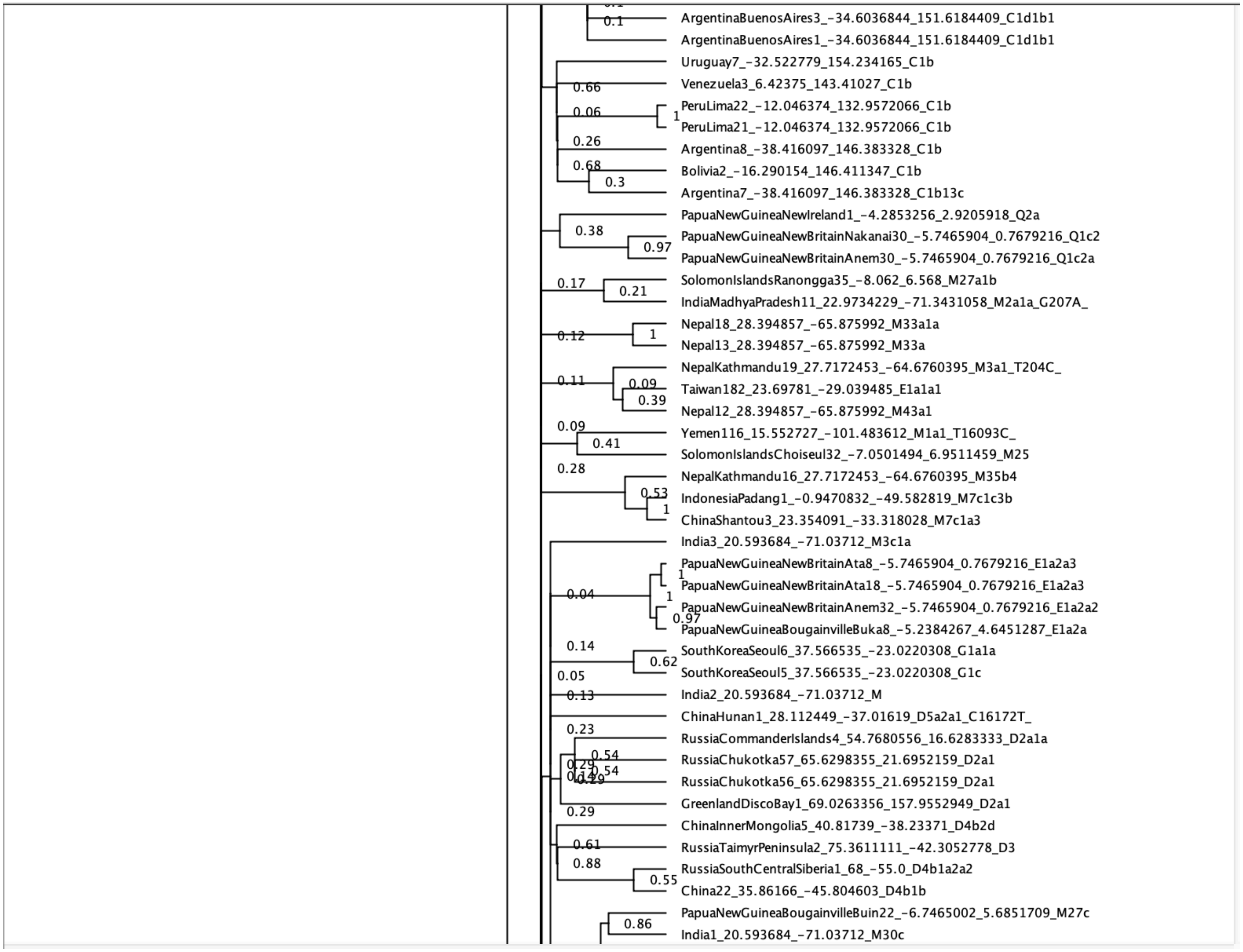
S3.5


0.69
ItalyCentralItaly13_45.9397505_-142.2462273_U5b3a2
0.71
0.71
ItalyCentralItaly12_45.9397505_-142.2462273_U5b3a2
Russia246_61.52401_-44.681244_U5a1a2a
0.71
CzechRepublic29_49.817492_-134.527038_U5a1c1
0.69
Belarus37_53.709807_-122.046611_U5a1d2a1
0.57
Brunei8_4.535277_-35.272331_Y2a
0.78
Brunei7_4.535277_-35.272331_Y2a
VietNamKinh8_10.5117268_-43.618977_N9a6
0.06
0.74
IndonesiaPalangkaraya5_-2.2161048_-36.086023_N9a6a
RussiaKemerovoRegion3_54.7574648_-62.5944712_H8b1
0.18
ItalyApulia1_40.7928393_-132.8988069_W1
0.06
0.77
ItalyTuscany5_43.7710513_-138.7513792_W5a1a
0.65
ItalyPiedmont2_45.0522366_-142.4846115_W4c
0.71
0.15
0.72
NewZealand2_-40.900557_24.885971_W3a1d
FranceToulose1_43.604652_-148.555791_W3a1
AustraliaKalumburu5_-14.286706_-23.3538975_N13
Australia9_-25.274398_-16.224864_M42a
SaudiArabia8_23.885942_-104.920838_M36a
TurkeyKurd1_36.666667_-113.233333_R0a_60.1T_
0.08
0.06
Philippines24_12.879721_-28.225983_B4b1a2
0.19
0.34
SpainAlmeria2_36.834047_-152.4637136_I2
0.33
ItalyPiedmont3_45.0522366_-142.4846115_I1a1
0.64
ItalySardinia48_40.1208752_-140.9871074_I1b
0.05
0.61
ItalyLombardy1_45.4790671_-140.1547567_I1c1
Brazil7_-14.235004_158.07472_C1b
PeruLima18_-12.046374_132.9572066_C1b
0.6
Bolivia1_-16.290154_146.411347_C1b
0.24
USA127_37.09024_114.287109_C1c6
Mexico10_23.634501_107.447216_C1b14
0.16
Chile29_-35.675147_138.457031_C1b13e
0.71
Chile28_-35.675147_138.457031_C1b13b
0.87
Chile27_-35.675147_138.457031_C1b13b
0.24
PeruAncon3_-11.7099743_132.8744945_C1c
0.66
DominicanRepublic40_18.735693_139.837349_C1c4
MexicoGuanajuato1_21.0190145_108.7426414_C1d
PeruHuancavelica1_-12.7861978_135.0235976_C1d
0.23
ColombiaMestizos7_9.4166667_134.15_C1d
0.14
0.24
1
ColombiaMestizos6_9.4166667_134.15_C1d
0.3
0.27
MexicoZacatecas1_22.7708555_107.4167574_C1d1a
USAOklahoma1_35.0077519_112.907123_C1d1a1
0.91
0.21
USAMontana1_46.8796822_99.6374342_C1d1a1
0.91
CanadaQuebec1_52.9399159_136.4508639_C1d1a1
0.24
0.19
Canada1_56.130366_103.653229_C1d1a1

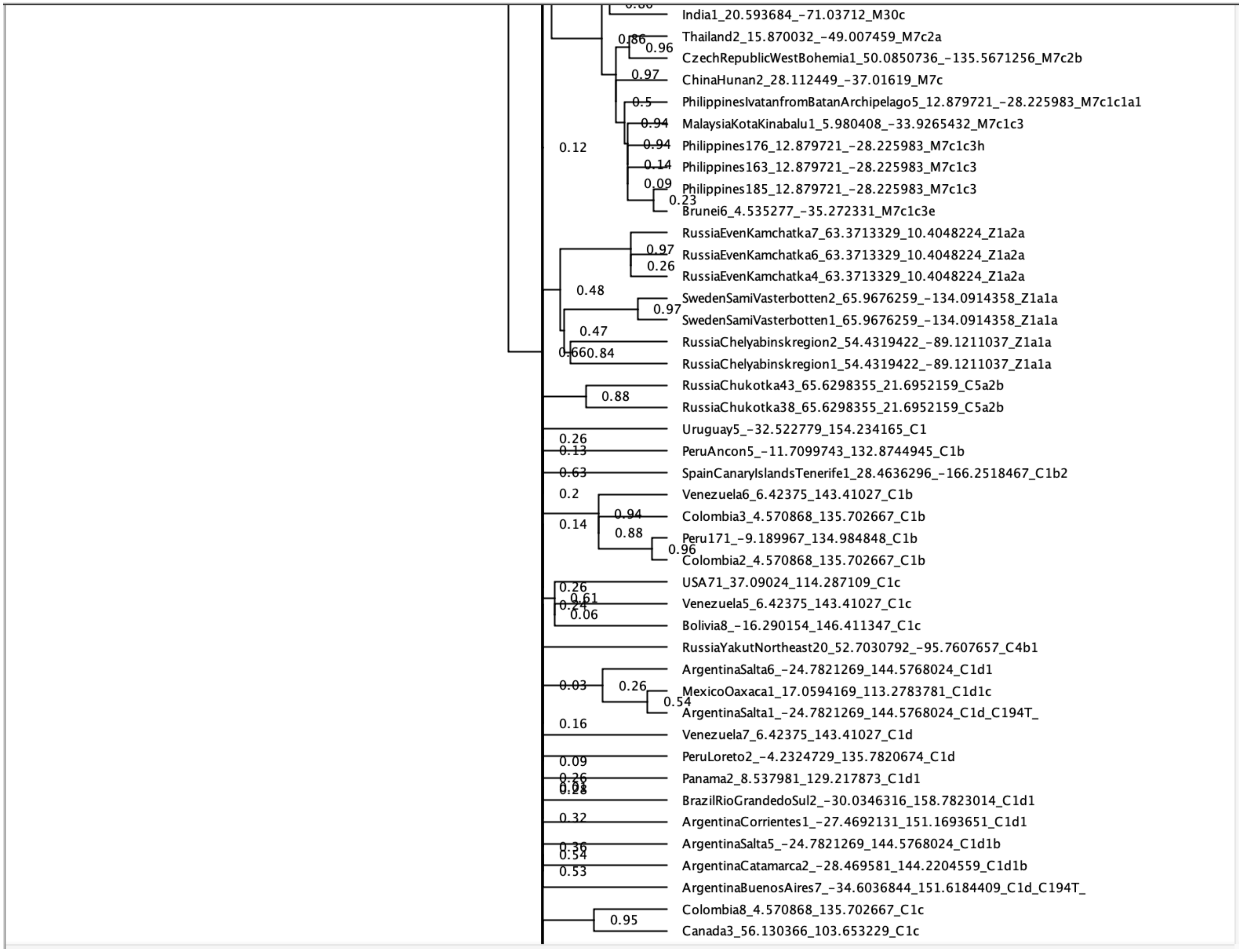
S3.6


0.24
0.19
Canada1_56.130366_103.653229_C1d1a1
USAMichigan1_44.3148443_124.3976357_C1d1c1
PeruLima26_-12.046374_132.9572066_C1d
0.72
ArgentinaBuenosAires5_-34.6036844_151.6184409_C1d
0.03
MexicoSonora1_29.2972247_99.6691186_C1d1
0.02
USAOklahoma2_35.0077519_112.907123_C1d1
0.34
PeruPiura1_-5.1782884_129.3451118_C1d1
0.89
PeruPucaPuca1_-14.6655556_138.2763889_C1d1
0.56
PeruCajamarca1_-7.1617465_131.4872145_C1d1
0.03
0.01
Uruguay3_-32.522779_154.234165_C1d3
0.04
1
Uruguay1_-32.522779_154.234165_C1d3
0.12
ColombiaMestizos4_9.4166667_134.15_C1d2a
0.12
ColombiaMestizos1_9.4166667_134.15_C1d2
0.08
0.04
MexicoTamaulipas1_24.26694_111.1637245_C1d
0.62
ColombiaBoyaca1_5.454511_136.637997_C1d_C194T_
USATexas2_31.9685988_110.0981869_C1d1c1
0.77
USATexas1_31.9685988_110.0981869_C1d1c1
0.3
0.55
BrazilRioGrandedoSul1_-30.0346316_158.7823014_C1d
Panama3_8.537981_129.217873_C1d1
0.06
Uruguay4_-32.522779_154.234165_C1d3
1
Uruguay2_-32.522779_154.234165_C1d3
0.01
0.37
ColombiaMestizos3_9.4166667_134.15_C1d2a
0.59
ColombiaMestizos5_9.4166667_134.15_C1d2a
0.13
0.12
ColombiaMestizos2_9.4166667_134.15_C1d2a
MexicoChihuahua6_28.6329957_103.9308996_C1d_C194T_
0.05
0.12
ArgentinaSalta2_-24.7821269_144.5768024_C1d_C194T_
BrazilMinasGerais1_-18.512178_165.4449692_C1d1
0.62
Uruguay6_-32.522779_154.234165_C1d1
0.51
ArgentinaBuenosAires2_-34.6036844_151.6184409_C1d1
0.03
PanamaDarien1_7.8681713_132.1632718_C1d1
ChileLosLagos2_-41.9197779_137.8583868_C1d1
0.1
0.67
ChileLosLagos1_-41.9197779_137.8583868_C1d1
0.07
NicaraguaPuertoCabezas1_14.0391161_126.6050368_C1d1d
0.69
BrazilMatoGrossodoSul1_-20.7722295_155.2148469_C1d1d
0.11
PeruHuanucu1_-9.9298454_133.7567347_C1d1
0.18
ArgentinaSalta3_-24.7821269_144.5768024_C1d1b
0.56
UruguayFlores1_-33.5629477_153.1688883_C1d1b1
0.56
ArgentinaRioNegro1_-40.8261434_146.9733661_C1d1b1
0.19
0.67
ArgentinaSalta4_-24.7821269_144.5768024_C1d1b1
1
ArgentinaBuenosAires4_-34.6036844_151.6184409_C1d1b1
0.11
ArgentinaCorrientes2_-27.4692131_151.1693651_C1d1b1
0.09
ArgentinaCatamarca1_-28.469581_144.2204559_C1d1b1
0.1
ArgentinaBuenosAires3_-34.6036844_151.6184409_C1d1b1
0.1

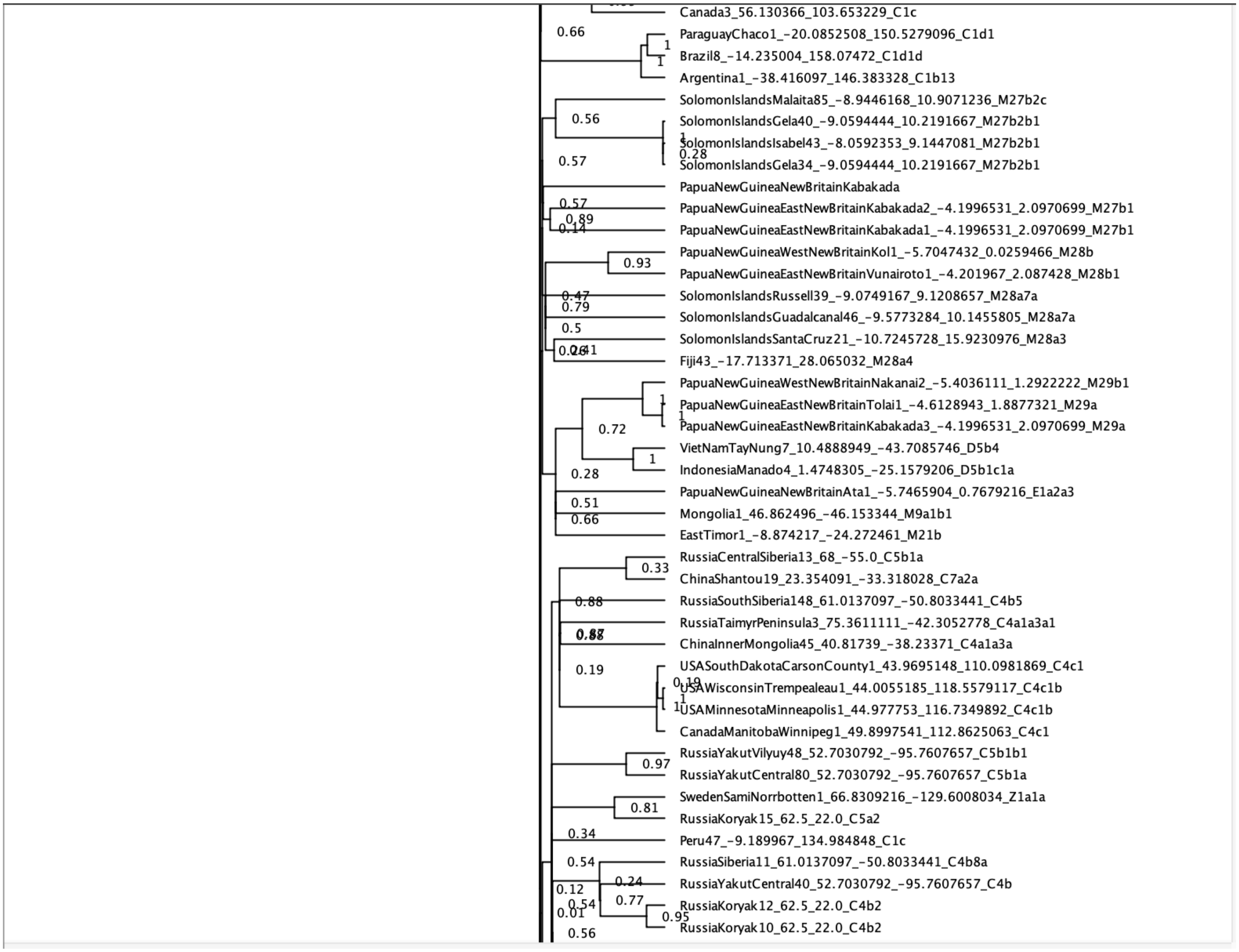
S3.7


0.1
ArgentinaBuenosAires3_-34.6036844_151.6184409_C1d1b1
ArgentinaBuenosAires1_-34.6036844_151.6184409_C1d1b1
Uruguay7_-32.522779_154.234165_C1b
0.66
Venezuela3_6.42375_143.41027_C1b
0.06
PeruLima22_-12.046374_132.9572066_C1b
1
PeruLima21_-12.046374_132.9572066_C1b
0.26
Argentina8_-38.416097_146.383328_C1b
0.68
Bolivia2_-16.290154_146.411347_C1b
0.3
Argentina7_-38.416097_146.383328_C1b13c
PapuaNewGuineaNewIreland1_-4.2853256_2.9205918_Q2a
0.38
PapuaNewGuineaNewBritainNakanai30_-5.7465904_0.7679216_Q1c2
0.97
PapuaNewGuineaNewBritainAnem30_-5.7465904_0.7679216_Q1c2a
0.17
SolomonIslandsRanongga35_-8.062_6.568_M27a1b
0.21
IndiaMadhyaPradesh11_22.9734229_-71.3431058_M2a1a_G207A_
Nepal18_28.394857_-65.875992_M33a1a
0.12
1
Nepal13_28.394857_-65.875992_M33a
NepalKathmandu19_27.7172453_-64.6760395_M3a1_T204C_
0.11
0.09
Taiwan182_23.69781_-29.039485_E1a1a1
0.39
Nepal12_28.394857_-65.875992_M43a1
0.09
Yemen116_15.552727_-101.483612_M1a1_T16093C_
0.41
SolomonIslandsChoiseul32_-7.0501494_6.9511459_M25
0.28
NepalKathmandu16_27.7172453_-64.6760395_M35b4
0.53
IndonesiaPadang1_-0.9470832_-49.582819_M7c1c3b
1
ChinaShantou3_23.354091_-33.318028_M7c1a3
India3_20.593684_-71.03712_M3c1a
PapuaNewGuineaNewBritainAta8_-5.7465904_0.7679216_E1a2a3
1
0.04
1
PapuaNewGuineaNewBritainAta18_-5.7465904_0.7679216_E1a2a3
PapuaNewGuineaNewBritainAnem32_-5.7465904_0.7679216_E1a2a2
0.97
PapuaNewGuineaBougainvilleBuka8_-5.2384267_4.6451287_E1a2a
0.14
SouthKoreaSeoul6_37.566535_-23.0220308_G1a1a
0.62
SouthKoreaSeoul5_37.566535_-23.0220308_G1c
0.05
0.13
India2_20.593684_-71.03712_M
ChinaHunan1_28.112449_-37.01619_D5a2a1_C16172T_
0.23
RussiaCommanderIslands4_54.7680556_16.6283333_D2a1a
0.54
RussiaChukotka57_65.6298355_21.6952159_D2a1
0.29
0.54
0.14
0.29
RussiaChukotka56_65.6298355_21.6952159_D2a1
GreenlandDiscoBay1_69.0263356_157.9552949_D2a1
0.29
ChinaInnerMongolia5_40.81739_-38.23371_D4b2d
0.61
RussiaTaimyrPeninsula2_75.3611111_-42.3052778_D3
0.88
RussiaSouthCentralSiberia1_68_-55.0_D4b1a2a2
0.55
China22_35.86166_-45.804603_D4b1b
PapuaNewGuineaBougainvilleBuin22_-6.7465002_5.6851709_M27c
0.86
India1_20.593684_-71.03712_M30c

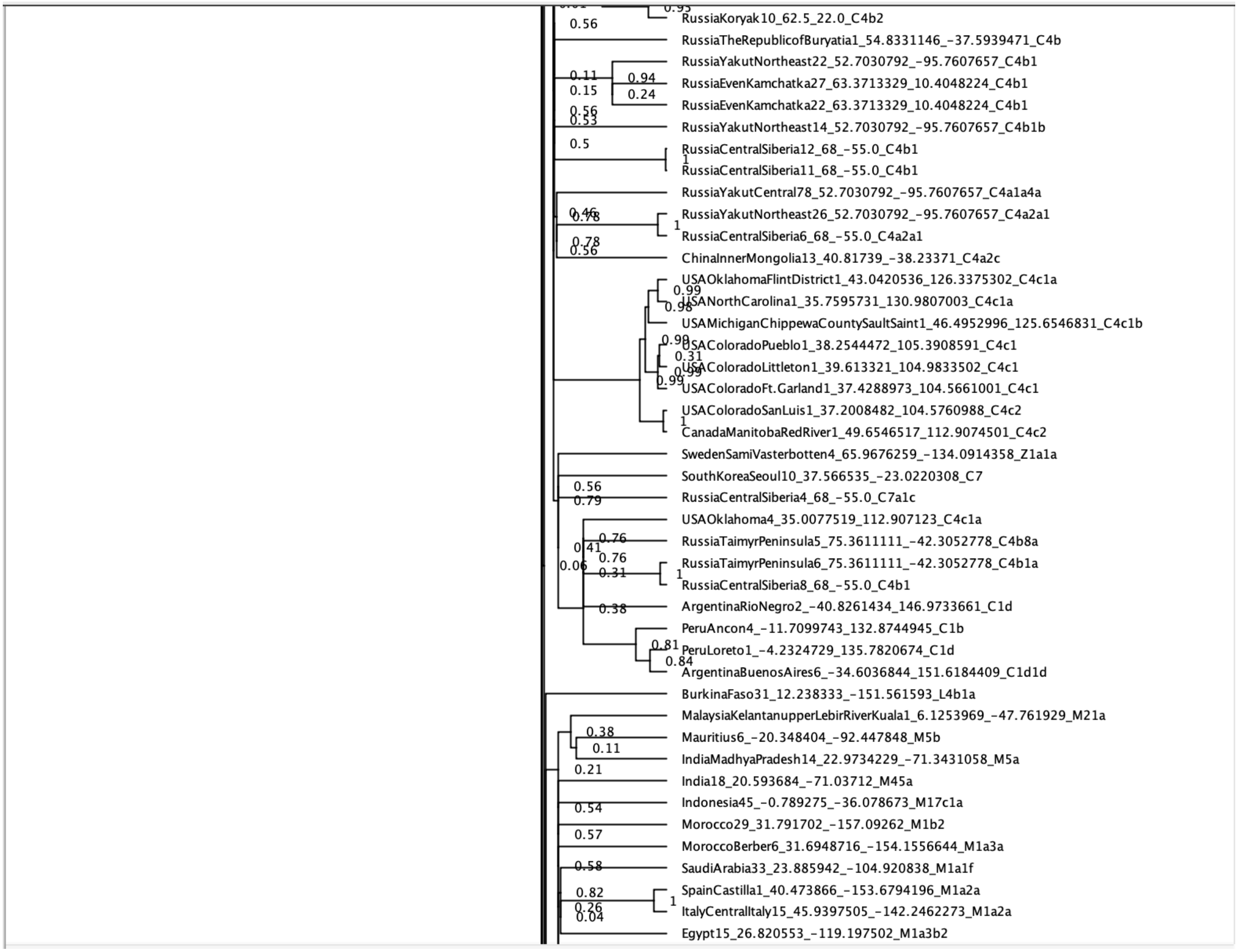
S3.8


0.86
India1_20.593684_-71.03712_M30c
0.86
Thailand2_15.870032_-49.007459_M7c2a
0.96
CzechRepublicWestBohemia1_50.0850736_-135.5671256_M7c2b
0.97
ChinaHunan2_28.112449_-37.01619_M7c
0.5
PhilippinesIvatanfromBatanArchipelago5_12.879721_-28.225983_M7c1c1a1
0.94
MalaysiaKotaKinabalu1_5.980408_-33.9265432_M7c1c3
0.12
0.94
Philippines176_12.879721_-28.225983_M7c1c3h
0.14
Philippines163_12.879721_-28.225983_M7c1c3
0.09
Philippines185_12.879721_-28.225983_M7c1c3
0.23
Brunei6_4.535277_-35.272331_M7c1c3e
RussiaEvenKamchatka7_63.3713329_10.4048224_Z1a2a
0.97
RussiaEvenKamchatka6_63.3713329_10.4048224_Z1a2a
0.26
RussiaEvenKamchatka4_63.3713329_10.4048224_Z1a2a
0.48
SwedenSamiVasterbotten2_65.9676259_-134.0914358_Z1a1a
0.97
SwedenSamiVasterbotten1_65.9676259_-134.0914358_Z1a1a
0.47
RussiaChelyabinskregion2_54.4319422_-89.1211037_Z1a1a
0.66
0.84
RussiaChelyabinskregion1_54.4319422_-89.1211037_Z1a1a
RussiaChukotka43_65.6298355_21.6952159_C5a2b
0.88
RussiaChukotka38_65.6298355_21.6952159_C5a2b
0.26
Uruguay5_-32.522779_154.234165_C1
0.13
PeruAncon5_-11.7099743_132.8744945_C1b
0.63
SpainCanaryIslandsTenerife1_28.4636296_-166.2518467_C1b2
0.2
Venezuela6_6.42375_143.41027_C1b
0.94
0.14
Colombia3_4.570868_135.702667_C1b
0.88
Peru171_-9.189967_134.984848_C1b
0.96
Colombia2_4.570868_135.702667_C1b
0.26
USA71_37.09024_114.287109_C1c
0.61
0.24
Venezuela5_6.42375_143.41027_C1c
0.06
Bolivia8_-16.290154_146.411347_C1c
RussiaYakutNortheast20_52.7030792_-95.7607657_C4b1
ArgentinaSalta6_-24.7821269_144.5768024_C1d1
0.03
0.26
MexicoOaxaca1_17.0594169_113.2783781_C1d1c
0.54
ArgentinaSalta1_-24.7821269_144.5768024_C1d_C194T_
0.16
Venezuela7_6.42375_143.41027_C1d
0.09
PeruLoreto2_-4.2324729_135.7820674_C1d
0.26
Panama2_8.537981_129.217873_C1d1
0.28
BrazilRioGrandedoSul2_-30.0346316_158.7823014_C1d1
0.32
ArgentinaCorrientes1_-27.4692131_151.1693651_C1d1
0.36
ArgentinaSalta5_-24.7821269_144.5768024_C1d1b
0.54
ArgentinaCatamarca2_-28.469581_144.2204559_C1d1b
0.53
ArgentinaBuenosAires7_-34.6036844_151.6184409_C1d_C194T_
Colombia8_4.570868_135.702667_C1c
0.95
Canada3_56.130366_103.653229_C1c

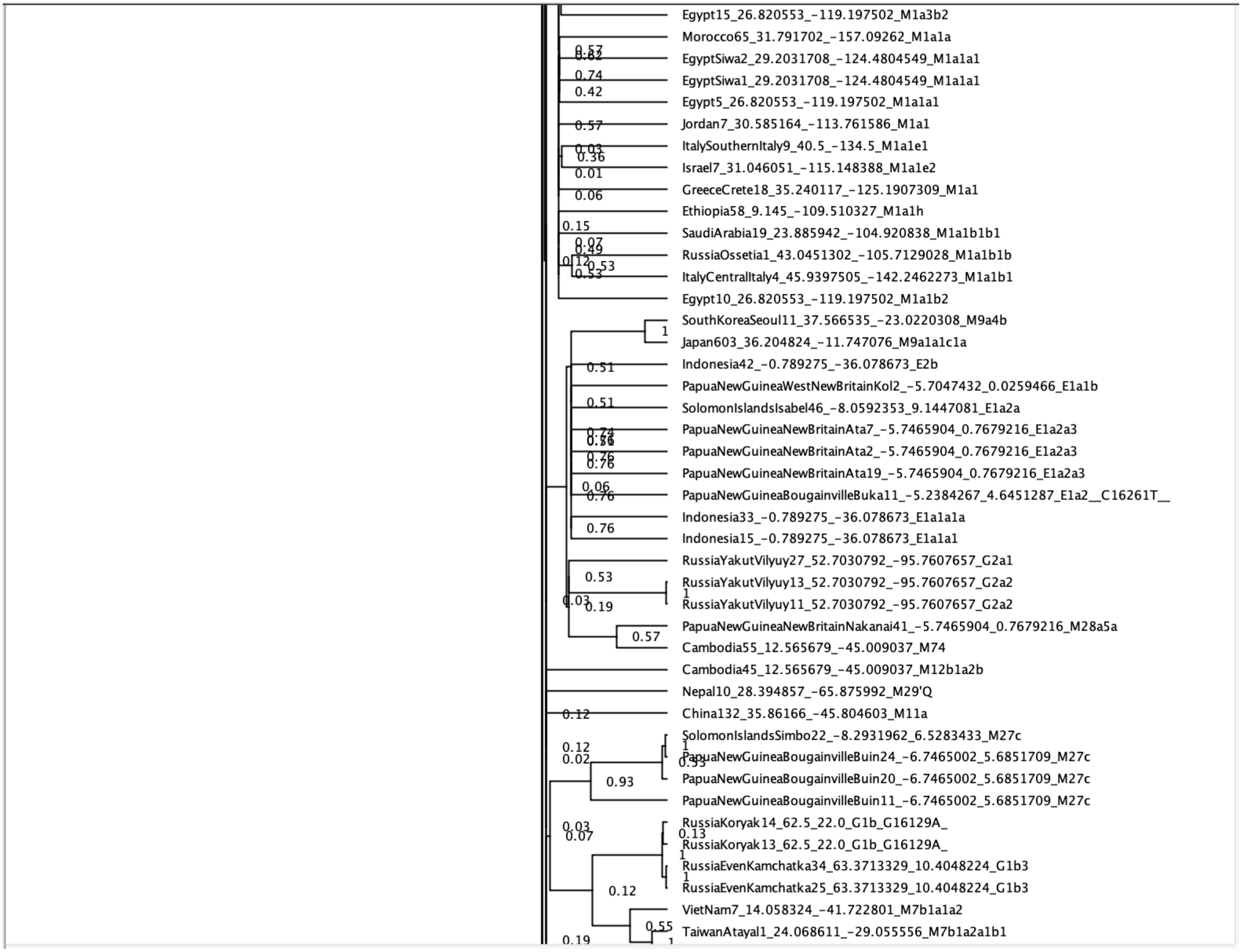
S3.9


Canada3_56.130366_103.653229_C1c
0.66
ParaguayChaco1_-20.0852508_150.5279096_C1d1
1
Brazil8_-14.235004_158.07472_C1d1d
1
Argentina1_-38.416097_146.383328_C1b13
SolomonIslandsMalaita85_-8.9446168_10.9071236_M27b2c
0.56
SolomonIslandsGela40_-9.0594444_10.2191667_M27b2b1
1
SolomonIslandsIsabel43_-8.0592353_9.1447081_M27b2b1
0.28
0.57
SolomonIslandsGela34_-9.0594444_10.2191667_M27b2b1
PapuaNewGuineaNewBritainKabakada
0.57
PapuaNewGuineaEastNewBritainKabakada2_-4.1996531_2.0970699_M27b1
0.89
0.14
PapuaNewGuineaEastNewBritainKabakada1_-4.1996531_2.0970699_M27b1
PapuaNewGuineaWestNewBritainKol1_-5.7047432_0.0259466_M28b
0.93
PapuaNewGuineaEastNewBritainVunairoto1_-4.201967_2.087428_M28b1
0.47
SolomonIslandsRussell39_-9.0749167_9.1208657_M28a7a
0.79
SolomonIslandsGuadalcanal46_-9.5773284_10.1455805_M28a7a
0.5
SolomonIslandsSantaCruz21_-10.7245728_15.9230976_M28a3
0.26
0.41
Fiji43_-17.713371_28.065032_M28a4
PapuaNewGuineaWestNewBritainNakanai2_-5.4036111_1.2922222_M29b1
1
PapuaNewGuineaEastNewBritainTolai1_-4.6128943_1.8877321_M29a
1
PapuaNewGuineaEastNewBritainKabakada3_-4.1996531_2.0970699_M29a
0.72
VietNamTayNung7_10.4888949_-43.7085746_D5b4
1
IndonesiaManado4_1.4748305_-25.1579206_D5b1c1a
0.28
PapuaNewGuineaNewBritainAta1_-5.7465904_0.7679216_E1a2a3
0.51
Mongolia1_46.862496_-46.153344_M9a1b1
0.66
EastTimor1_-8.874217_-24.272461_M21b
RussiaCentralSiberia13_68_-55.0_C5b1a
0.33
ChinaShantou19_23.354091_-33.318028_C7a2a
0.88
RussiaSouthSiberia148_61.0137097_-50.8033441_C4b5
RussiaTaimyrPeninsula3_75.3611111_-42.3052778_C4a1a3a1
0.87
0.88
ChinaInnerMongolia45_40.81739_-38.23371_C4a1a3a
0.19
USASouthDakotaCarsonCounty1_43.9695148_110.0981869_C4c1
0.19
USAWisconsinTrempealeau1_44.0055185_118.5579117_C4c1b
1
1
USAMinnesotaMinneapolis1_44.977753_116.7349892_C4c1b
CanadaManitobaWinnipeg1_49.8997541_112.8625063_C4c1
RussiaYakutVilyuy48_52.7030792_-95.7607657_C5b1b1
0.97
RussiaYakutCentral80_52.7030792_-95.7607657_C5b1a
SwedenSamiNorrbotten1_66.8309216_-129.6008034_Z1a1a
0.81
RussiaKoryak15_62.5_22.0_C5a2
0.34
Peru47_-9.189967_134.984848_C1c
0.54
RussiaSiberia11_61.0137097_-50.8033441_C4b8a
0.24
RussiaYakutCentral40_52.7030792_-95.7607657_C4b
0.12
0.54
0.77
RussiaKoryak12_62.5_22.0_C4b2
0.01
0.95
RussiaKoryak10_62.5_22.0_C4b2
0.56

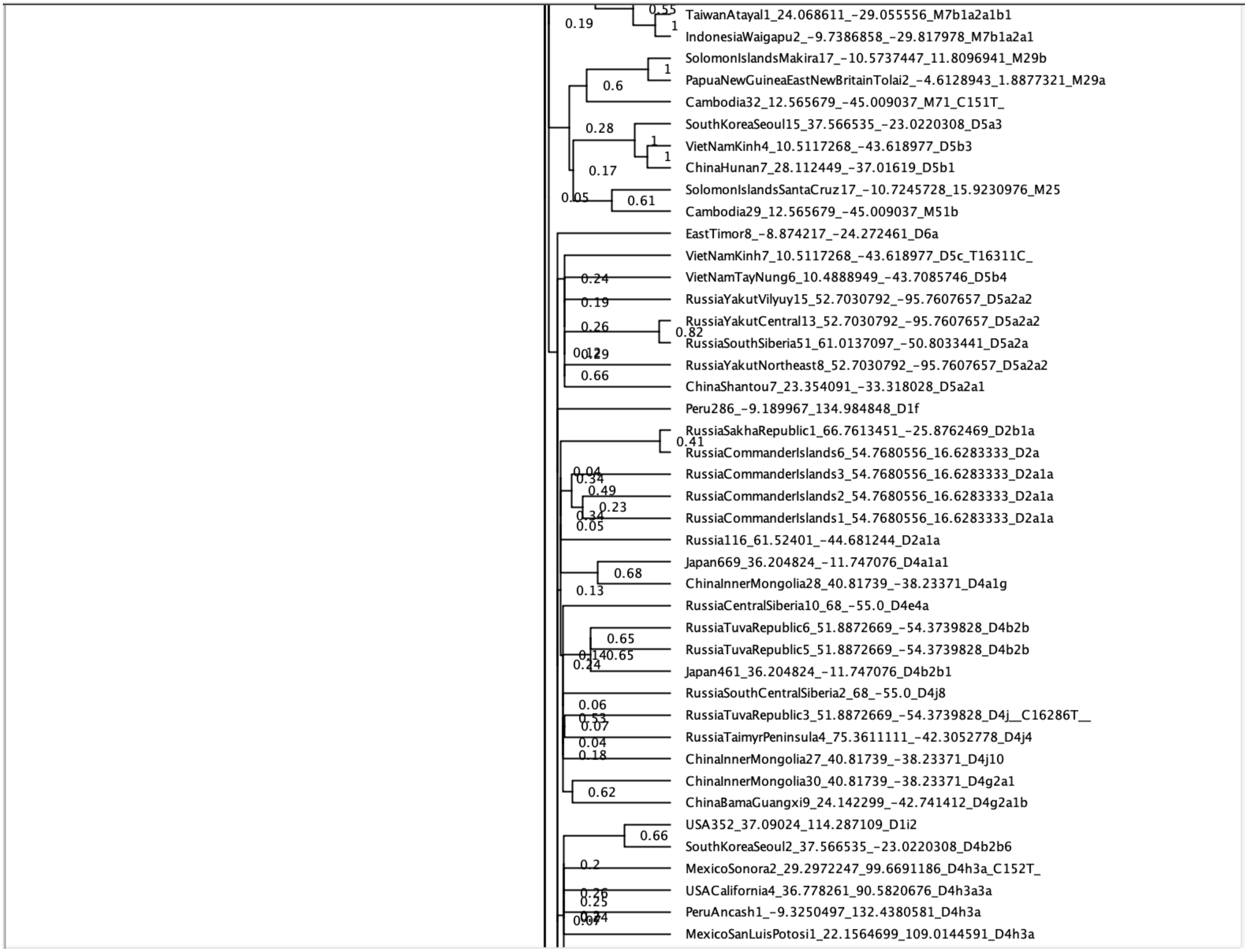
S3.10


0.56
RussiaKoryak10_62.5_22.0_C4b2
RussiaTheRepublicofBuryatia1_54.8331146_-37.5939471_C4b
RussiaYakutNortheast22_52.7030792_-95.7607657_C4b1
0.11
0.15
0.94
0.24
RussiaEvenKamchatka27_63.3713329_10.4048224_C4b1
RussiaEvenKamchatka22_63.3713329_10.4048224_C4b1
0.56
0.53
RussiaYakutNortheast14_52.7030792_-95.7607657_C4b1b
0.5
RussiaCentralSiberia12_68_-55.0_C4b1
1
RussiaCentralSiberia11_68_-55.0_C4b1
RussiaYakutCentral78_52.7030792_-95.7607657_C4a1a4a
0.46
0.78
RussiaYakutNortheast26_52.7030792_-95.7607657_C4a2a1
1
RussiaCentralSiberia6_68_-55.0_C4a2a1
0.78
0.56
ChinaInnerMongolia13_40.81739_-38.23371_C4a2c
USAOklahomaFlintDistrict1_43.0420536_126.3375302_C4c1a
0.99
0.98
USANorthCarolina1_35.7595731_130.9807003_C4c1a
USAMichiganChippewaCountySaultSaint1_46.4952996_125.6546831_C4c1b
0.99
USAColoradoPueblo1_38.2544472_105.3908591_C4c1
0.31
0.99
USAColoradoLittleton1_39.613321_104.9833502_C4c1
0.99
USAColoradoFt.Garland1_37.4288973_104.5661001_C4c1
USAColoradoSanLuis1_37.2008482_104.5760988_C4c2
1
CanadaManitobaRedRiver1_49.6546517_112.9074501_C4c2
SwedenSamiVasterbotten4_65.9676259_-134.0914358_Z1a1a
SouthKoreaSeoul10_37.566535_-23.0220308_C7
0.56
0.79
RussiaCentralSiberia4_68_-55.0_C7a1c
USAOklahoma4_35.0077519_112.907123_C4c1a
0.76
0.41
RussiaTaimyrPeninsula5_75.3611111_-42.3052778_C4b8a
0.76
0.06
0.31
RussiaTaimyrPeninsula6_75.3611111_-42.3052778_C4b1a
1
RussiaCentralSiberia8_68_-55.0_C4b1
0.38
ArgentinaRioNegro2_-40.8261434_146.9733661_C1d
PeruAncon4_-11.7099743_132.8744945_C1b
0.81
PeruLoreto1_-4.2324729_135.7820674_C1d
0.84
ArgentinaBuenosAires6_-34.6036844_151.6184409_C1d1d
BurkinaFaso31_12.238333_-151.561593_L4b1a
MalaysiaKelantanupperLebirRiverKuala1_6.1253969_-47.761929_M21a
0.38
Mauritius6_-20.348404_-92.447848_M5b
0.11
IndiaMadhyaPradesh14_22.9734229_-71.3431058_M5a
0.21
India18_20.593684_-71.03712_M45a
Indonesia45_-0.789275_-36.078673_M17c1a
0.54
Morocco29_31.791702_-157.09262_M1b2
0.57
MoroccoBerber6_31.6948716_-154.1556644_M1a3a
0.58
SaudiArabia33_23.885942_-104.920838_M1a1f
0.82
SpainCastilla1_40.473866_-153.6794196_M1a2a
1
0.26
0.04
ItalyCentralItaly15_45.9397505_-142.2462273_M1a2a
Egypt15_26.820553_-119.197502_M1a3b2

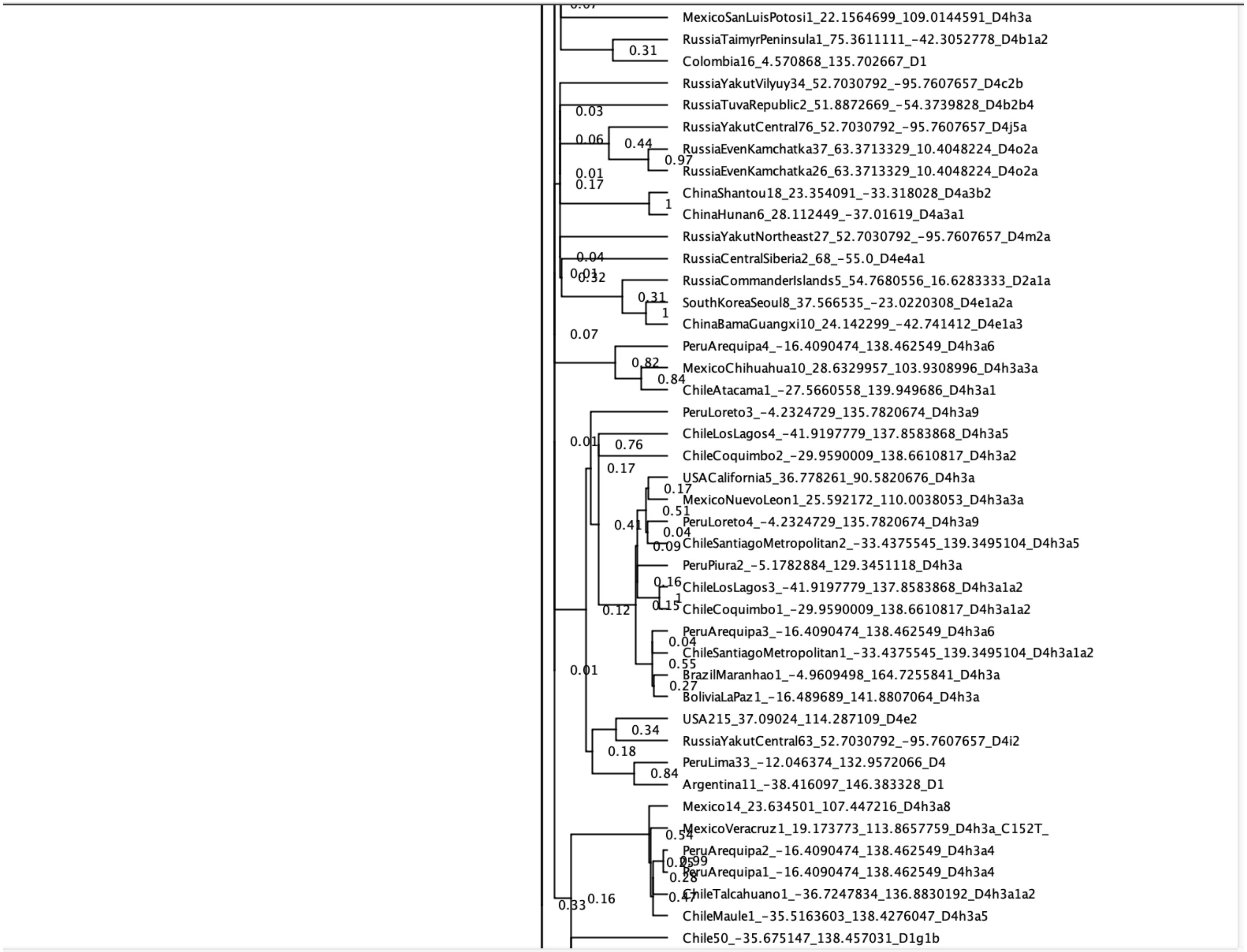
S3.11


Egypt15_26.820553_-119.197502_M1a3b2
Morocco65_31.791702_-157.09262_M1a1a
EgyptSiwa2_29.2031708_-124.4804549_M1a1a1
EgyptSiwa1_29.2031708_-124.4804549_M1a1a1
Egypt5_26.820553_-119.197502_M1a1a1
Jordan7_30.585164_-113.761586_M1a1
ItalySouthernItaly9_40.5_-134.5_M1a1e1
Israel7_31.046051_-115.148388_M1a1e2
GreeceCrete18_35.240117_-125.1907309_M1a1
Ethiopia58_9.145_-109.510327_M1a1h
SaudiArabia19_23.885942_-104.920838_M1a1b1b1
RussiaOssetia1_43.0451302_-105.7129028_M1a1b1b
ItalyCentralItaly4_45.9397505_-142.2462273_M1a1b1
Egypt10_26.820553_-119.197502_M1a1b2
SouthKoreaSeoul11_37.566535_-23.0220308_M9a4b
Japan603_36.204824_-11.747076_M9a1a1c1a
Indonesia42_-0.789275_-36.078673_E2b
PapuaNewGuineaWestNewBritainKol2_-5.7047432_0.0259466_E1a1b
SolomonIslandsIsabel46_-8.0592353_9.1447081_E1a2a
PapuaNewGuineaNewBritainAta7_-5.7465904_0.7679216_E1a2a3
PapuaNewGuineaNewBritainAta2_-5.7465904_0.7679216_E1a2a3
PapuaNewGuineaNewBritainAta19_-5.7465904_0.7679216_E1a2a3
PapuaNewGuineaBougainvilleBuka11_-5.2384267_4.6451287_E1a2__C16261T__
Indonesia33_-0.789275_-36.078673_E1a1a1a
Indonesia15_-0.789275_-36.078673_E1a1a1
RussiaYakutVilyuy27_52.7030792_-95.7607657_G2a1
RussiaYakutVilyuy13_52.7030792_-95.7607657_G2a2
RussiaYakutVilyuy11_52.7030792_-95.7607657_G2a2
PapuaNewGuineaNewBritainNakanai41_-5.7465904_0.7679216_M28a5a
Cambodia55_12.565679_-45.009037_M74
Cambodia45_12.565679_-45.009037_M12b1a2b
Nepal10_28.394857_-65.875992_M29'Q
China132_35.86166_-45.804603_M11a
SolomonIslandsSimbo22_-8.2931962_6.5283433_M27c
PapuaNewGuineaBougainvilleBuin24_-6.7465002_5.6851709_M27c
PapuaNewGuineaBougainvilleBuin20_-6.7465002_5.6851709_M27c
PapuaNewGuineaBougainvilleBuin11_-6.7465002_5.6851709_M27c
RussiaKoryak14_62.5_22.0_G1b_G16129A_
RussiaKoryak13_62.5_22.0_G1b_G16129A_
RussiaEvenKamchatka34_63.3713329_10.4048224_G1b3
RussiaEvenKamchatka25_63.3713329_10.4048224_G1b3
VietNam7_14.058324_-41.722801_M7b1a1a2
TaiwanAtayal1_24.068611_-29.055556_M7b1a2a1b1
0.57
0.62
0.74
0.42
0.57
0.03
0.36
0.01
0.06
0.15
0.07
0.49
0.12
0.53
0.53
1
0.51
0.51
0.74
0.56
0.76
0.76
0.06
0.76
0.76
0.53
0.03
0.19
1
0.57
0.12
0.12
0.02
1
0.33
0.93
0.03
0.07
0.13
1
1
0.12
0.55
1
0.19

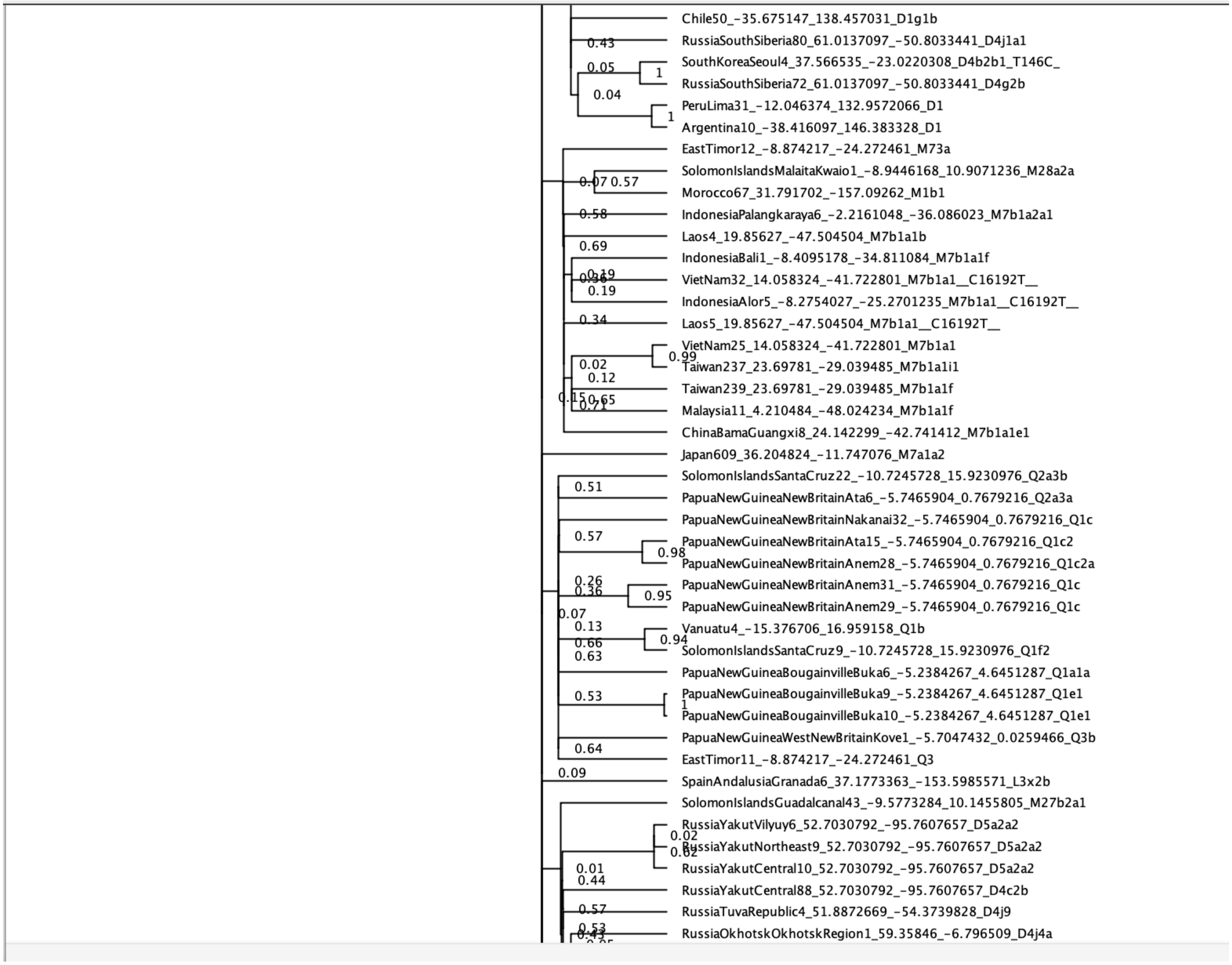
S3.12


0.55
TaiwanAtayal1_24.068611_-29.055556_M7b1a2a1b1
0.19
1
IndonesiaWaigapu2_-9.7386858_-29.817978_M7b1a2a1
SolomonIslandsMakira17_-10.5737447_11.8096941_M29b
1
0.6
PapuaNewGuineaEastNewBritainTolai2_-4.6128943_1.8877321_M29a
Cambodia32_12.565679_-45.009037_M71_C151T_
0.28
SouthKoreaSeoul15_37.566535_-23.0220308_D5a3
1
VietNamKinh4_10.5117268_-43.618977_D5b3
1
0.17
ChinaHunan7_28.112449_-37.01619_D5b1
0.05
SolomonIslandsSantaCruz17_-10.7245728_15.9230976_M25
0.61
Cambodia29_12.565679_-45.009037_M51b
EastTimor8_-8.874217_-24.272461_D6a
VietNamKinh7_10.5117268_-43.618977_D5c_T16311C_
0.24
VietNamTayNung6_10.4888949_-43.7085746_D5b4
0.19
RussiaYakutVilyuy15_52.7030792_-95.7607657_D5a2a2
0.26
RussiaYakutCentral13_52.7030792_-95.7607657_D5a2a2
0.82
RussiaSouthSiberia51_61.0137097_-50.8033441_D5a2a
0.12
0.29
RussiaYakutNortheast8_52.7030792_-95.7607657_D5a2a2
0.66
ChinaShantou7_23.354091_-33.318028_D5a2a1
Peru286_-9.189967_134.984848_D1f
RussiaSakhaRepublic1_66.7613451_-25.8762469_D2b1a
0.41
RussiaCommanderIslands6_54.7680556_16.6283333_D2a
0.04
0.34
RussiaCommanderIslands3_54.7680556_16.6283333_D2a1a
0.49
RussiaCommanderIslands2_54.7680556_16.6283333_D2a1a
0.23
0.34
0.05
RussiaCommanderIslands1_54.7680556_16.6283333_D2a1a
Russia116_61.52401_-44.681244_D2a1a
Japan669_36.204824_-11.747076_D4a1a1
0.68
ChinaInnerMongolia28_40.81739_-38.23371_D4a1g
0.13
RussiaCentralSiberia10_68_-55.0_D4e4a
RussiaTuvaRepublic6_51.8872669_-54.3739828_D4b2b
0.65
RussiaTuvaRepublic5_51.8872669_-54.3739828_D4b2b
0.14
0.65
0.24
Japan461_36.204824_-11.747076_D4b2b1
RussiaSouthCentralSiberia2_68_-55.0_D4j8
0.06
RussiaTuvaRepublic3_51.8872669_-54.3739828_D4j__C16286T__
0.53
0.07
RussiaTaimyrPeninsula4_75.3611111_-42.3052778_D4j4
0.04
0.18
ChinaInnerMongolia27_40.81739_-38.23371_D4j10
ChinaInnerMongolia30_40.81739_-38.23371_D4g2a1
0.62
ChinaBamaGuangxi9_24.142299_-42.741412_D4g2a1b
USA352_37.09024_114.287109_D1i2
0.66
SouthKoreaSeoul2_37.566535_-23.0220308_D4b2b6
0.2
MexicoSonora2_29.2972247_99.6691186_D4h3a_C152T_
0.26
0.25
USACalifornia4_36.778261_90.5820676_D4h3a3a
0.07
0.24
PeruAncash1_-9.3250497_132.4380581_D4h3a
MexicoSanLuisPotosi1_22.1564699_109.0144591_D4h3a

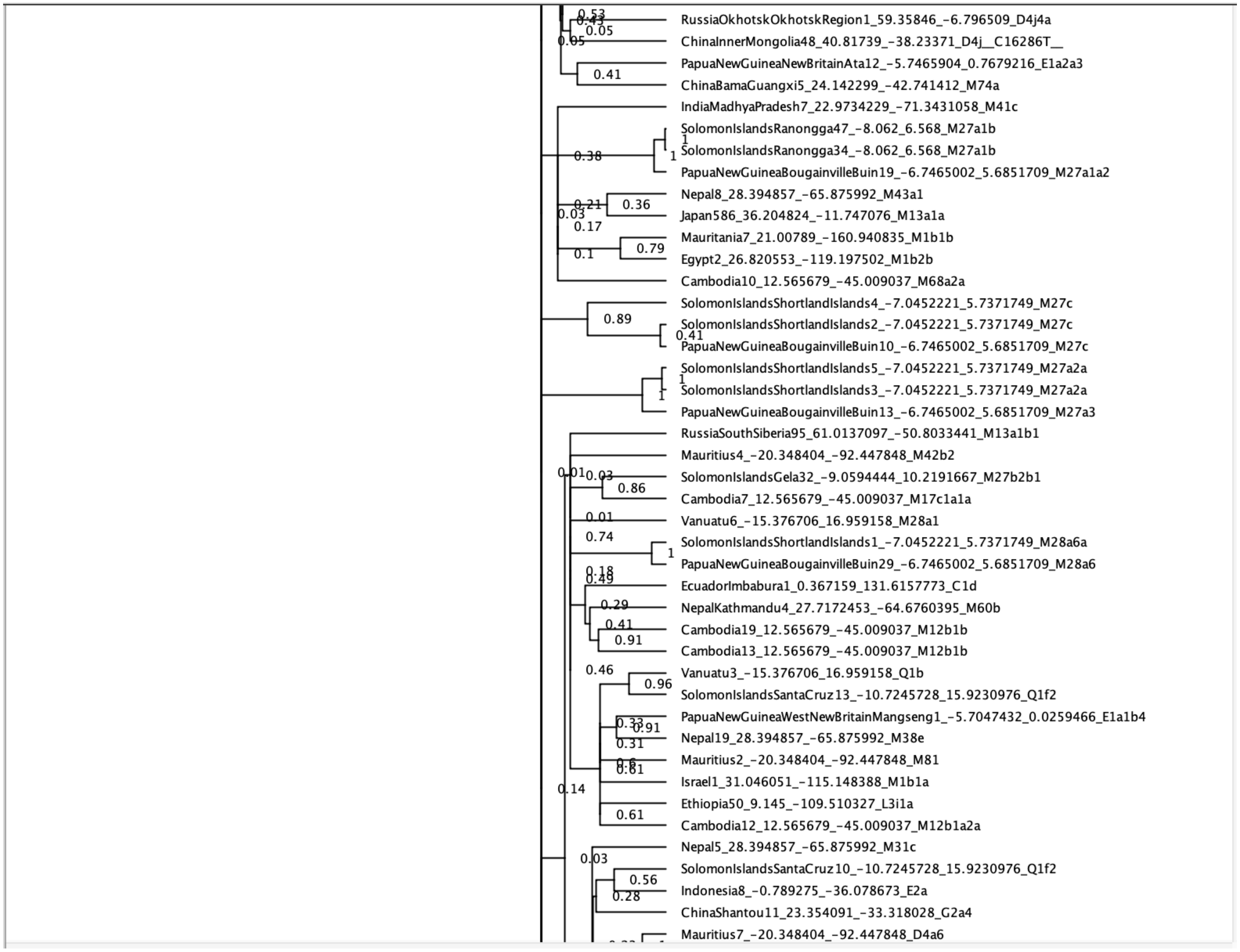
S3.13


MexicoSanLuisPotosi1_22.1564699_109.0144591_D4h3a
RussiaTaimyrPeninsula1_75.3611111_-42.3052778_D4b1a2
0.31
Colombia16_4.570868_135.702667_D1
RussiaYakutVilyuy34_52.7030792_-95.7607657_D4c2b
0.03
RussiaTuvaRepublic2_51.8872669_-54.3739828_D4b2b4
RussiaYakutCentral76_52.7030792_-95.7607657_D4j5a
0.06
0.44
RussiaEvenKamchatka37_63.3713329_10.4048224_D4o2a
0.97
0.01
RussiaEvenKamchatka26_63.3713329_10.4048224_D4o2a
0.17
ChinaShantou18_23.354091_-33.318028_D4a3b2
1
ChinaHunan6_28.112449_-37.01619_D4a3a1
RussiaYakutNortheast27_52.7030792_-95.7607657_D4m2a
0.04
RussiaCentralSiberia2_68_-55.0_D4e4a1
0.03
0.32
RussiaCommanderIslands5_54.7680556_16.6283333_D2a1a
0.31
SouthKoreaSeoul8_37.566535_-23.0220308_D4e1a2a
1
ChinaBamaGuangxi10_24.142299_-42.741412_D4e1a3
0.07
PeruArequipa4_-16.4090474_138.462549_D4h3a6
0.82
MexicoChihuahua10_28.6329957_103.9308996_D4h3a3a
0.84
ChileAtacama1_-27.5660558_139.949686_D4h3a1
PeruLoreto3_-4.2324729_135.7820674_D4h3a9
0.01
ChileLosLagos4_-41.9197779_137.8583868_D4h3a5
0.76
ChileCoquimbo2_-29.9590009_138.6610817_D4h3a2
0.17
USACalifornia5_36.778261_90.5820676_D4h3a
0.17
MexicoNuevoLeon1_25.592172_110.0038053_D4h3a3a
0.51
0.41
PeruLoreto4_-4.2324729_135.7820674_D4h3a9
0.04
0.09
ChileSantiagoMetropolitan2_-33.4375545_139.3495104_D4h3a5
PeruPiura2_-5.1782884_129.3451118_D4h3a
0.16
ChileLosLagos3_-41.9197779_137.8583868_D4h3a1a2
1
0.15
0.12
ChileCoquimbo1_-29.9590009_138.6610817_D4h3a1a2
PeruArequipa3_-16.4090474_138.462549_D4h3a6
0.04
ChileSantiagoMetropolitan1_-33.4375545_139.3495104_D4h3a1a2
0.55
0.01
BrazilMaranhao1_-4.9609498_164.7255841_D4h3a
0.27
BoliviaLaPaz1_-16.489689_141.8807064_D4h3a
USA215_37.09024_114.287109_D4e2
0.34
RussiaYakutCentral63_52.7030792_-95.7607657_D4i2
0.18
PeruLima33_-12.046374_132.9572066_D4
0.84
Argentina11_-38.416097_146.383328_D1
Mexico14_23.634501_107.447216_D4h3a8
0.54
MexicoVeracruz1_19.173773_113.8657759_D4h3a_C152T_
PeruArequipa2_-16.4090474_138.462549_D4h3a4
0.25
0.99
PeruArequipa1_-16.4090474_138.462549_D4h3a4
0.28
0.47
ChileTalcahuano1_-36.7247834_136.8830192_D4h3a1a2
0.33
0.16
ChileMaule1_-35.5163603_138.4276047_D4h3a5
Chile50_-35.675147_138.457031_D1g1b

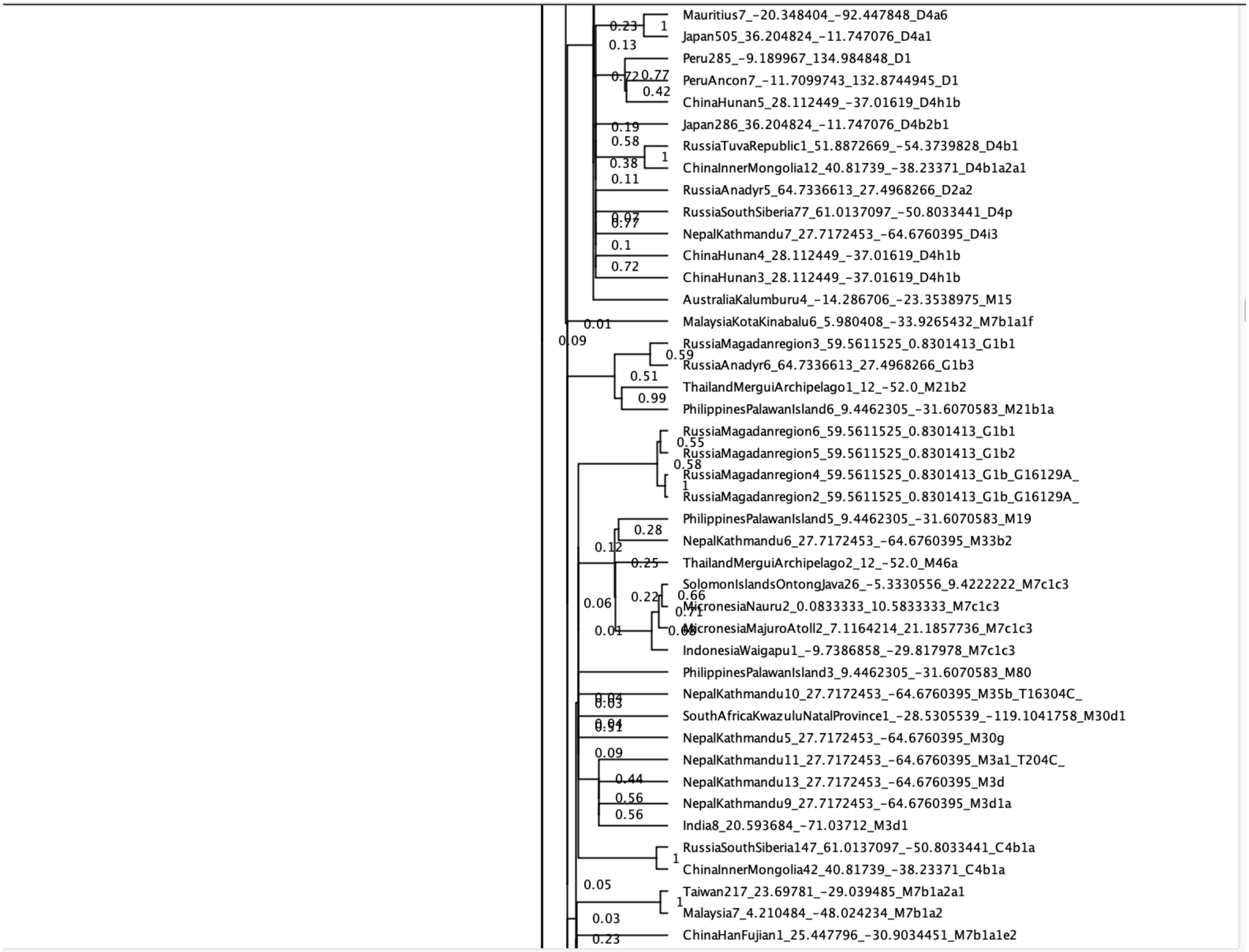
S3.14


Chile50_-35.675147_138.457031_D1g1b
0.43
RussiaSouthSiberia80_61.0137097_-50.8033441_D4j1a1
0.05
SouthKoreaSeoul4_37.566535_-23.0220308_D4b2b1_T146C_
1
RussiaSouthSiberia72_61.0137097_-50.8033441_D4g2b
0.04
PeruLima31_-12.046374_132.9572066_D1
1
Argentina10_-38.416097_146.383328_D1
EastTimor12_-8.874217_-24.272461_M73a
SolomonIslandsMalaitaKwaio1_-8.9446168_10.9071236_M28a2a
0.07 0.57
Morocco67_31.791702_-157.09262_M1b1
0.58
IndonesiaPalangkaraya6_-2.2161048_-36.086023_M7b1a2a1
Laos4_19.85627_-47.504504_M7b1a1b
0.69
IndonesiaBali1_-8.4095178_-34.811084_M7b1a1f
0.19
0.36
VietNam32_14.058324_-41.722801_M7b1a1__C16192T__
0.19
IndonesiaAlor5_-8.2754027_-25.2701235_M7b1a1__C16192T__
0.34
Laos5_19.85627_-47.504504_M7b1a1__C16192T__
VietNam25_14.058324_-41.722801_M7b1a1
0.99
0.02
Taiwan237_23.69781_-29.039485_M7b1a1i1
0.12
Taiwan239_23.69781_-29.039485_M7b1a1f
0.15
0.65
0.71
Malaysia11_4.210484_-48.024234_M7b1a1f
ChinaBamaGuangxi8_24.142299_-42.741412_M7b1a1e1
Japan609_36.204824_-11.747076_M7a1a2
SolomonIslandsSantaCruz22_-10.7245728_15.9230976_Q2a3b
0.51
PapuaNewGuineaNewBritainAta6_-5.7465904_0.7679216_Q2a3a
PapuaNewGuineaNewBritainNakanai32_-5.7465904_0.7679216_Q1c
0.57
PapuaNewGuineaNewBritainAta15_-5.7465904_0.7679216_Q1c2
0.98
PapuaNewGuineaNewBritainAnem28_-5.7465904_0.7679216_Q1c2a
0.26
0.36
PapuaNewGuineaNewBritainAnem31_-5.7465904_0.7679216_Q1c
0.95
PapuaNewGuineaNewBritainAnem29_-5.7465904_0.7679216_Q1c
0.07
0.13
Vanuatu4_-15.376706_16.959158_Q1b
0.66
0.94
0.63
SolomonIslandsSantaCruz9_-10.7245728_15.9230976_Q1f2
PapuaNewGuineaBougainvilleBuka6_-5.2384267_4.6451287_Q1a1a
0.53
PapuaNewGuineaBougainvilleBuka9_-5.2384267_4.6451287_Q1e1
1
PapuaNewGuineaBougainvilleBuka10_-5.2384267_4.6451287_Q1e1
PapuaNewGuineaWestNewBritainKove1_-5.7047432_0.0259466_Q3b
0.64
EastTimor11_-8.874217_-24.272461_Q3
0.09
SpainAndalusiaGranada6_37.1773363_-153.5985571_L3x2b
SolomonIslandsGuadalcanal43_-9.5773284_10.1455805_M27b2a1
RussiaYakutVilyuy6_52.7030792_-95.7607657_D5a2a2
0.02
RussiaYakutNortheast9_52.7030792_-95.7607657_D5a2a2
0.62
0.01
RussiaYakutCentral10_52.7030792_-95.7607657_D5a2a2
0.44
RussiaYakutCentral88_52.7030792_-95.7607657_D4c2b
0.57
RussiaTuvaRepublic4_51.8872669_-54.3739828_D4j9
0.53
0.43
RussiaOkhotskOkhotskRegion1_59.35846_-6.796509_D4j4a

S3.15

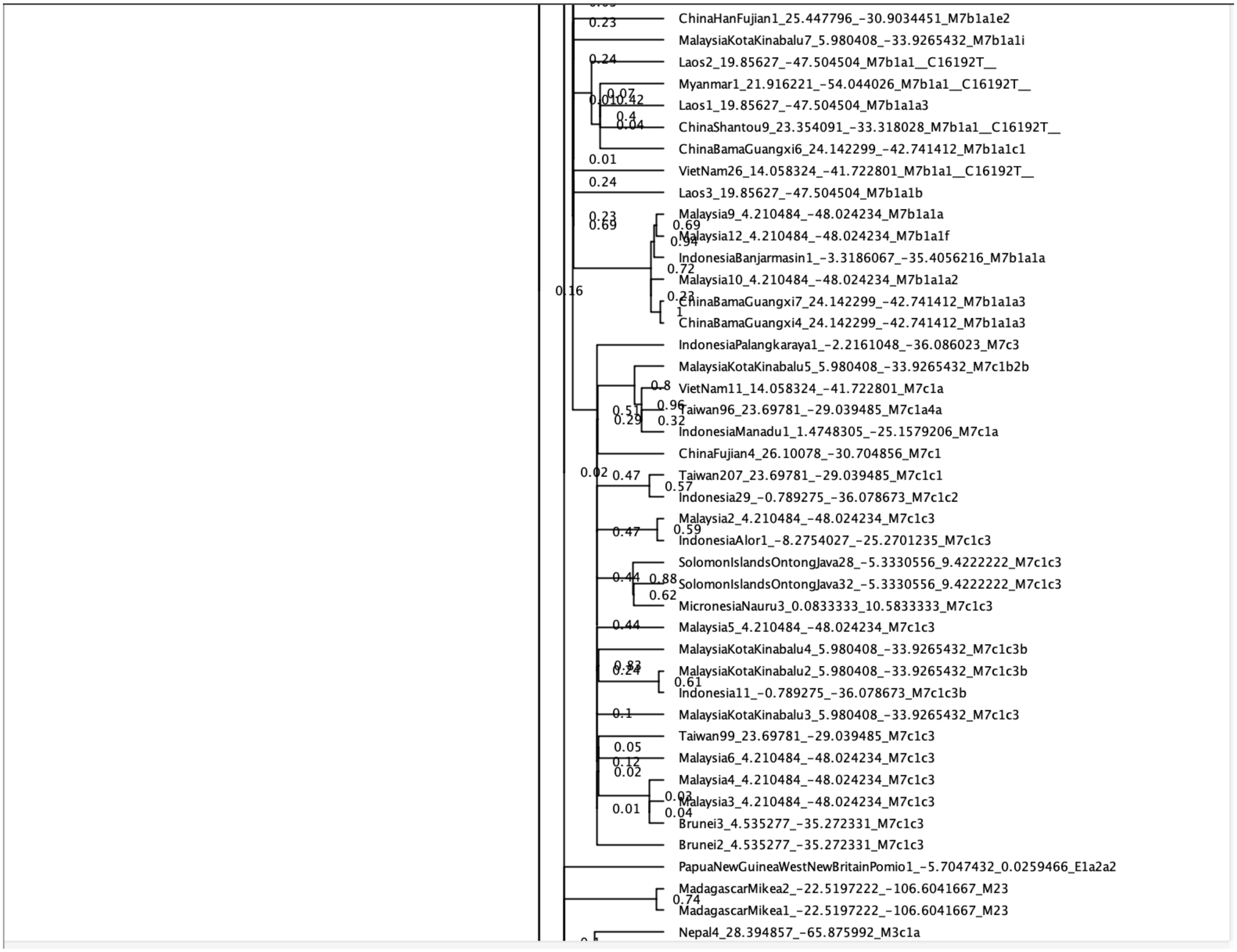

RussiaOkhotskOkhotskRegion1_59.35846_-6.796509_D4j4a
ChinaInnerMongolia48_40.81739_-38.23371_D4j__C16286T__
PapuaNewGuineaNewBritainAta12_-5.7465904_0.7679216_E1a2a3
ChinaBamaGuangxi5_24.142299_-42.741412_M74a
IndiaMadhyaPradesh7_22.9734229_-71.3431058_M41c
SolomonIslandsRanongga47_-8.062_6.568_M27a1b
SolomonIslandsRanongga34_-8.062_6.568_M27a1b
PapuaNewGuineaBougainvilleBuin19_-6.7465002_5.6851709_M27a1a2
Nepal8_28.394857_-65.875992_M43a1
Japan586_36.204824_-11.747076_M13a1a
Mauritania7_21.00789_-160.940835_M1b1b
Egypt2_26.820553_-119.197502_M1b2b
Cambodia10_12.565679_-45.009037_M68a2a
SolomonIslandsShortlandIslands4_-7.0452221_5.7371749_M27c
SolomonIslandsShortlandIslands2_-7.0452221_5.7371749_M27c
PapuaNewGuineaBougainvilleBuin10_-6.7465002_5.6851709_M27c
SolomonIslandsShortlandIslands5_-7.0452221_5.7371749_M27a2a
SolomonIslandsShortlandIslands3_-7.0452221_5.7371749_M27a2a
PapuaNewGuineaBougainvilleBuin13_-6.7465002_5.6851709_M27a3
RussiaSouthSiberia95_61.0137097_-50.8033441_M13a1b1
Mauritius4_-20.348404_-92.447848_M42b2
SolomonIslandsGela32_-9.0594444_10.2191667_M27b2b1
Cambodia7_12.565679_-45.009037_M17c1a1a
Vanuatu6_-15.376706_16.959158_M28a1
SolomonIslandsShortlandIslands1_-7.0452221_5.7371749_M28a6a
PapuaNewGuineaBougainvilleBuin29_-6.7465002_5.6851709_M28a6
EcuadorImbabura1_0.367159_131.6157773_C1d
NepalKathmandu4_27.7172453_-64.6760395_M60b
Cambodia19_12.565679_-45.009037_M12b1b
Cambodia13_12.565679_-45.009037_M12b1b
Vanuatu3_-15.376706_16.959158_Q1b
SolomonIslandsSantaCruz13_-10.7245728_15.9230976_Q1f2
PapuaNewGuineaWestNewBritainMangseng1_-5.7047432_0.0259466_E1a1b4
Nepal19_28.394857_-65.875992_M38e
Mauritius2_-20.348404_-92.447848_M81
Israel1_31.046051_-115.148388_M1b1a
Ethiopia50_9.145_-109.510327_L3i1a
Cambodia12_12.565679_-45.009037_M12b1a2a
Nepal5_28.394857_-65.875992_M31c
SolomonIslandsSantaCruz10_-10.7245728_15.9230976_Q1f2
Indonesia8_-0.789275_-36.078673_E2a
ChinaShantou11_23.354091_-33.318028_G2a4
Mauritius7_-20.348404_-92.447848_D4a6

S3.16

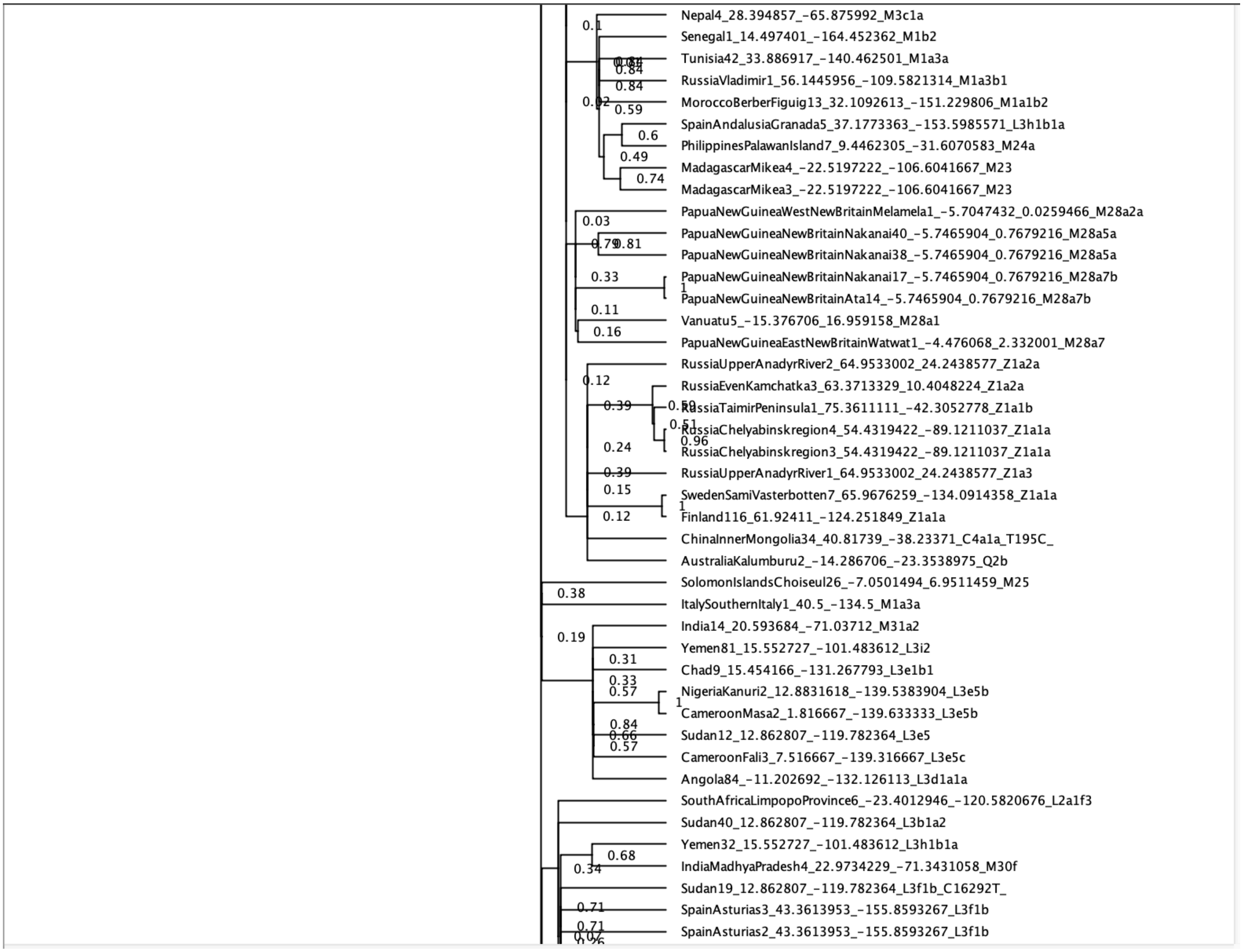


Mauritius7_-20.348404_-92.447848_D4a6
Japan505_36.204824_-11.747076_D4a1
Peru285_-9.189967_134.984848_D1
PeruAncon7_-11.7099743_132.8744945_D1
ChinaHunan5_28.112449_-37.01619_D4h1b
Japan286_36.204824_-11.747076_D4b2b1
RussiaTuvaRepublic1_51.8872669_-54.3739828_D4b1
ChinaInnerMongolia12_40.81739_-38.23371_D4b1a2a1
RussiaAnadyr5_64.7336613_27.4968266_D2a2
RussiaSouthSiberia77_61.0137097_-50.8033441_D4p
NepalKathmandu7_27.7172453_-64.6760395_D4i3
ChinaHunan4_28.112449_-37.01619_D4h1b
ChinaHunan3_28.112449_-37.01619_D4h1b
AustraliaKalumburu4_-14.286706_-23.3538975_M15
MalaysiaKotaKinabalu6_5.980408_-33.9265432_M7b1a1f
RussiaMagadanregion3_59.5611525_0.8301413_G1b1
RussiaAnadyr6_64.7336613_27.4968266_G1b3
ThailandMerguiArchipelago1_12_-52.0_M21b2
PhilippinesPalawanIsland6_9.4462305_-31.6070583_M21b1a
RussiaMagadanregion6_59.5611525_0.8301413_G1b1
RussiaMagadanregion5_59.5611525_0.8301413_G1b2
RussiaMagadanregion4_59.5611525_0.8301413_G1b_G16129A_
RussiaMagadanregion2_59.5611525_0.8301413_G1b_G16129A_
PhilippinesPalawanIsland5_9.4462305_-31.6070583_M19
NepalKathmandu6_27.7172453_-64.6760395_M33b2
ThailandMerguiArchipelago2_12_-52.0_M46a
SolomonIslandsOntongJava26_-5.3330556_9.4222222_M7c1c3
MicronesiaNauru2_0.0833333_10.5833333_M7c1c3
MicronesiaMajuroAtoll2_7.1164214_21.1857736_M7c1c3
IndonesiaWaigapu1_-9.7386858_-29.817978_M7c1c3
PhilippinesPalawanIsland3_9.4462305_-31.6070583_M80
NepalKathmandu10_27.7172453_-64.6760395_M35b_T16304C_
SouthAfricaKwazuluNatalProvince1_-28.5305539_-119.1041758_M30d1
NepalKathmandu5_27.7172453_-64.6760395_M30g
NepalKathmandu11_27.7172453_-64.6760395_M3a1_T204C_
NepalKathmandu13_27.7172453_-64.6760395_M3d
NepalKathmandu9_27.7172453_-64.6760395_M3d1a
India8_20.593684_-71.03712_M3d1
RussiaSouthSiberia147_61.0137097_-50.8033441_C4b1a
ChinaInnerMongolia42_40.81739_-38.23371_C4b1a
Taiwan217_23.69781_-29.039485_M7b1a2a1
Malaysia7_4.210484_-48.024234_M7b1a2
ChinaHanFujian1_25.447796_-30.9034451_M7b1a1e2
0.23
1
0.13
0.72
0.77
0.42
0.19
0.58
1
0.38
0.11
0.07
0.77
0.1
0.72
0.01
0.09
0.59
0.51
0.99
0.55
0.58
1
0.28
0.12
0.25
0.22
0.66
0.71
0.06
0.01
0.08
0.04
0.03
0.04
0.91
0.09
0.44
0.56
0.56
1
0.05
1
0.03
0.23

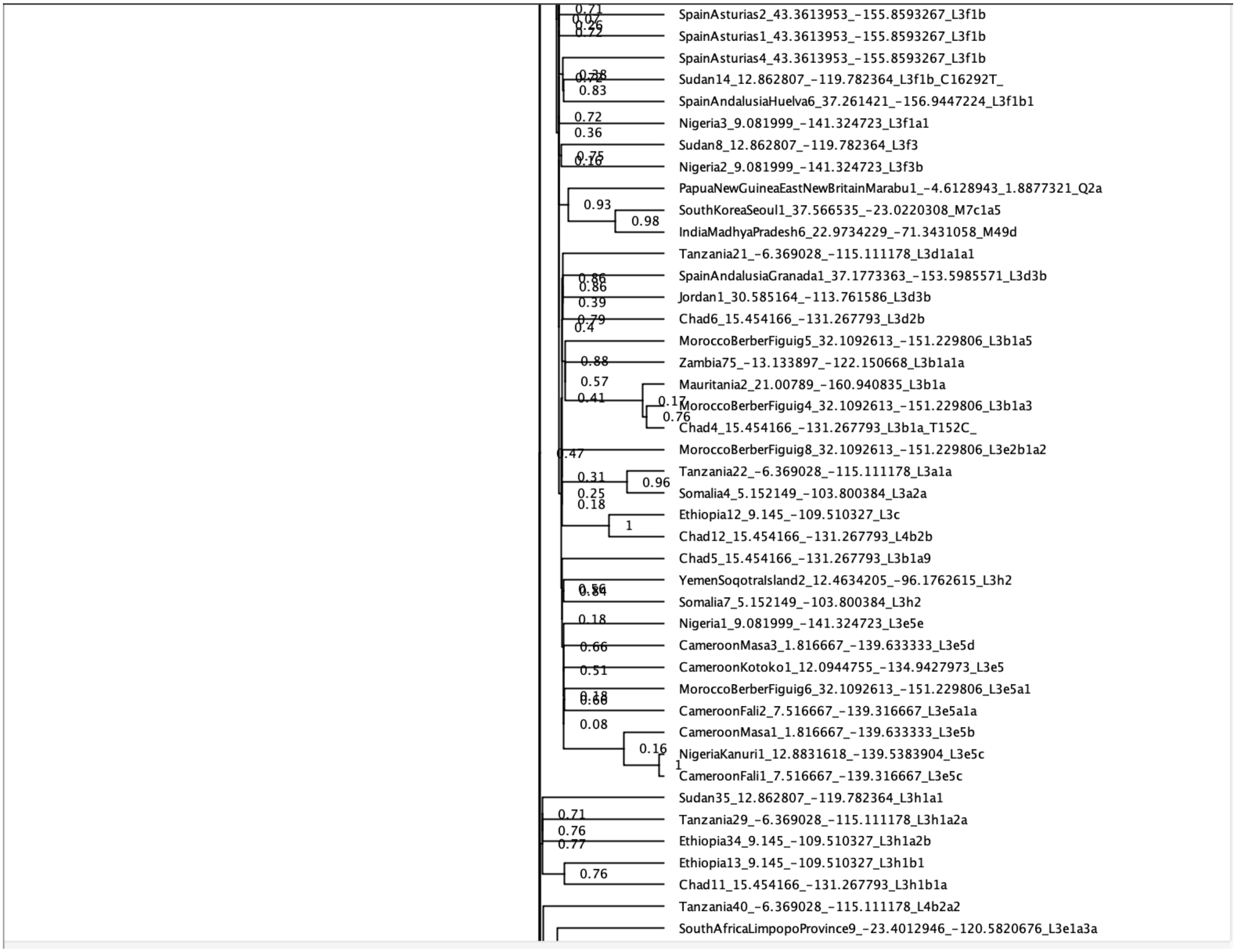
S3.17


0.23
ChinaHanFujian1_25.447796_-30.9034451_M7b1a1e2
MalaysiaKotaKinabalu7_5.980408_-33.9265432_M7b1a1i
0.24
Laos2_19.85627_-47.504504_M7b1a1__C16192T__
0.07
Myanmar1_21.916221_-54.044026_M7b1a1__C16192T__
0.010.42
Laos1_19.85627_-47.504504_M7b1a1a3
0.4
0.04
ChinaShantou9_23.354091_-33.318028_M7b1a1__C16192T__
ChinaBamaGuangxi6_24.142299_-42.741412_M7b1a1c1
0.01
VietNam26_14.058324_-41.722801_M7b1a1__C16192T__
0.24
Laos3_19.85627_-47.504504_M7b1a1b
0.23
0.69
Malaysia9_4.210484_-48.024234_M7b1a1a
0.69
0.94
Malaysia12_4.210484_-48.024234_M7b1a1f
IndonesiaBanjarmasin1_-3.3186067_-35.4056216_M7b1a1a
0.72
Malaysia10_4.210484_-48.024234_M7b1a1a2
0.16
0.23
ChinaBamaGuangxi7_24.142299_-42.741412_M7b1a1a3
1
ChinaBamaGuangxi4_24.142299_-42.741412_M7b1a1a3
IndonesiaPalangkaraya1_-2.2161048_-36.086023_M7c3
MalaysiaKotaKinabalu5_5.980408_-33.9265432_M7c1b2b
0.8
VietNam11_14.058324_-41.722801_M7c1a
0.51
0.96
Taiwan96_23.69781_-29.039485_M7c1a4a
0.29
0.32
IndonesiaManadu1_1.4748305_-25.1579206_M7c1a
ChinaFujian4_26.10078_-30.704856_M7c1
0.02
0.47
Taiwan207_23.69781_-29.039485_M7c1c1
0.57
Indonesia29_-0.789275_-36.078673_M7c1c2
Malaysia2_4.210484_-48.024234_M7c1c3
0.47
0.59
IndonesiaAlor1_-8.2754027_-25.2701235_M7c1c3
SolomonIslandsOntongJava28_-5.3330556_9.4222222_M7c1c3
0.44
0.88
SolomonIslandsOntongJava32_-5.3330556_9.4222222_M7c1c3
0.62
MicronesiaNauru3_0.0833333_10.5833333_M7c1c3
0.44
Malaysia5_4.210484_-48.024234_M7c1c3
MalaysiaKotaKinabalu4_5.980408_-33.9265432_M7c1c3b
0.83
0.24
MalaysiaKotaKinabalu2_5.980408_-33.9265432_M7c1c3b
0.61
Indonesia11_-0.789275_-36.078673_M7c1c3b
0.1
MalaysiaKotaKinabalu3_5.980408_-33.9265432_M7c1c3
Taiwan99_23.69781_-29.039485_M7c1c3
0.05
Malaysia6_4.210484_-48.024234_M7c1c3
0.12
0.02
Malaysia4_4.210484_-48.024234_M7c1c3
0.03
Malaysia3_4.210484_-48.024234_M7c1c3
0.01
0.04
Brunei3_4.535277_-35.272331_M7c1c3
Brunei2_4.535277_-35.272331_M7c1c3
PapuaNewGuineaWestNewBritainPomio1_-5.7047432_0.0259466_E1a2a2
MadagascarMikea2_-22.5197222_-106.6041667_M23
0.74
MadagascarMikea1_-22.5197222_-106.6041667_M23
Nepal4_28.394857_-65.875992_M3c1a

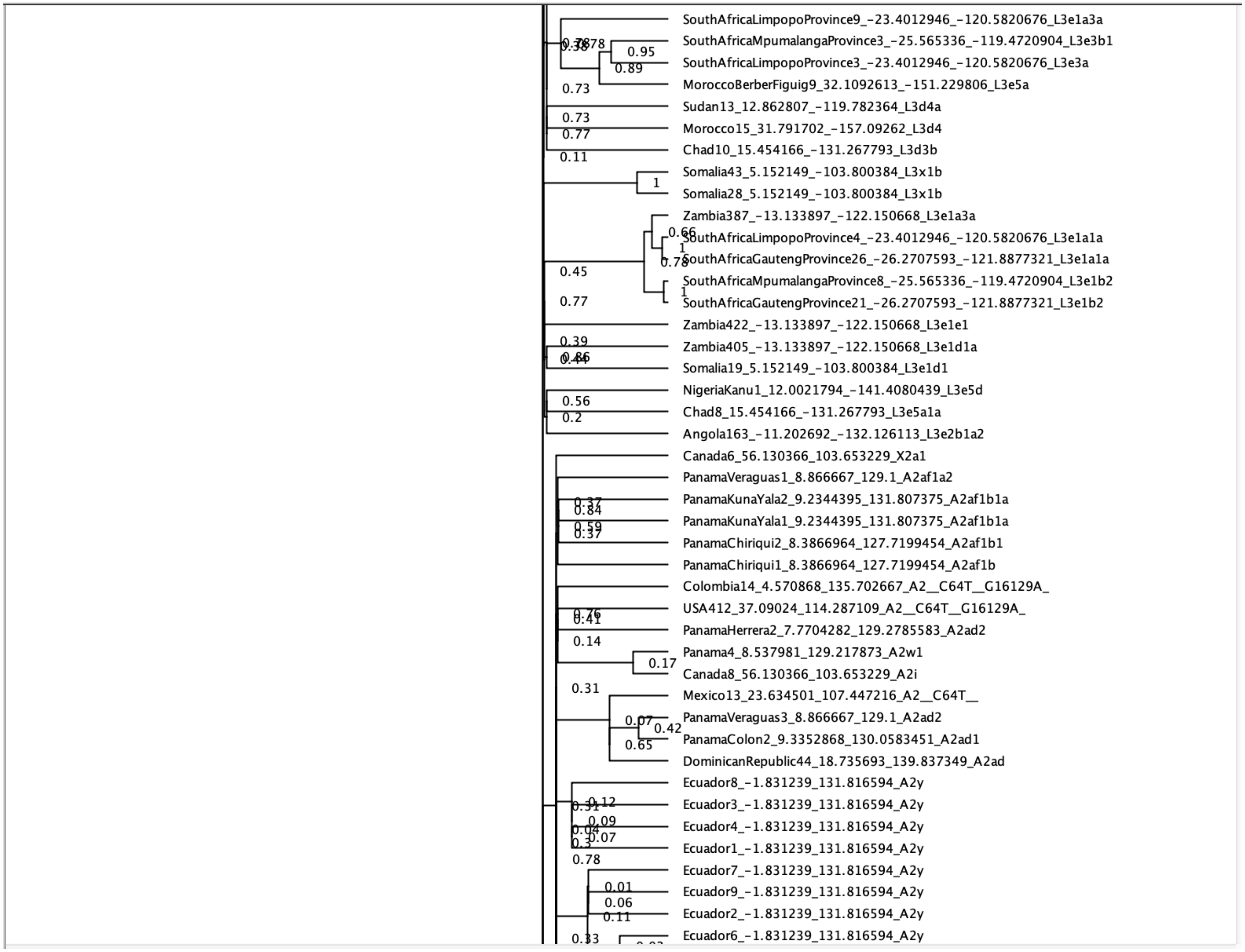
S3.18


Nepal4_28.394857_-65.875992_M3c1a
0.1
Senegal1_14.497401_-164.452362_M1b2
Tunisia42_33.886917_-140.462501_M1a3a
0.84
RussiaVladimir1_56.1445956_-109.5821314_M1a3b1
0.84
0.02
MoroccoBerberFiguig13_32.1092613_-151.229806_M1a1b2
0.59
SpainAndalusiaGranada5_37.1773363_-153.5985571_L3h1b1a
0.6
PhilippinesPalawanIsland7_9.4462305_-31.6070583_M24a
0.49
MadagascarMikea4_-22.5197222_-106.6041667_M23
0.74
MadagascarMikea3_-22.5197222_-106.6041667_M23
PapuaNewGuineaWestNewBritainMelamela1_-5.7047432_0.0259466_M28a2a
0.03
PapuaNewGuineaNewBritainNakanai40_-5.7465904_0.7679216_M28a5a
0.81
PapuaNewGuineaNewBritainNakanai38_-5.7465904_0.7679216_M28a5a
0.33
PapuaNewGuineaNewBritainNakanai17_-5.7465904_0.7679216_M28a7b
1
PapuaNewGuineaNewBritainAta14_-5.7465904_0.7679216_M28a7b
0.11
Vanuatu5_-15.376706_16.959158_M28a1
0.16
PapuaNewGuineaEastNewBritainWatwat1_-4.476068_2.332001_M28a7
RussiaUpperAnadyrRiver2_64.9533002_24.2438577_Z1a2a
0.12
RussiaEvenKamchatka3_63.3713329_10.4048224_Z1a2a
0.39
RussiaTaimirPeninsula1_75.3611111_-42.3052778_Z1a1b
0.51
RussiaChelyabinskregion4_54.4319422_-89.1211037_Z1a1a
0.96
0.24
RussiaChelyabinskregion3_54.4319422_-89.1211037_Z1a1a
0.39
RussiaUpperAnadyrRiver1_64.9533002_24.2438577_Z1a3
0.15
SwedenSamiVasterbotten7_65.9676259_-134.0914358_Z1a1a
1
0.12
Finland116_61.92411_-124.251849_Z1a1a
ChinaInnerMongolia34_40.81739_-38.23371_C4a1a_T195C_
AustraliaKalumburu2_-14.286706_-23.3538975_Q2b
SolomonIslandsChoiseul26_-7.0501494_6.9511459_M25
0.38
ItalySouthernItaly1_40.5_-134.5_M1a3a
India14_20.593684_-71.03712_M31a2
0.19
Yemen81_15.552727_-101.483612_L3i2
0.31
Chad9_15.454166_-131.267793_L3e1b1
0.33
0.57
NigeriaKanuri2_12.8831618_-139.5383904_L3e5b
1
CameroonMasa2_1.816667_-139.633333_L3e5b
0.84
0.66
Sudan12_12.862807_-119.782364_L3e5
0.57
CameroonFali3_7.516667_-139.316667_L3e5c
Angola84_-11.202692_-132.126113_L3d1a1a
SouthAfricaLimpopoProvince6_-23.4012946_-120.5820676_L2a1f3
Sudan40_12.862807_-119.782364_L3b1a2
Yemen32_15.552727_-101.483612_L3h1b1a
0.68
0.34
IndiaMadhyaPradesh4_22.9734229_-71.3431058_M30f
Sudan19_12.862807_-119.782364_L3f1b_C16292T_
0.71
SpainAsturias3_43.3613953_-155.8593267_L3f1b
0.71
SpainAsturias2_43.3613953_-155.8593267_L3f1b

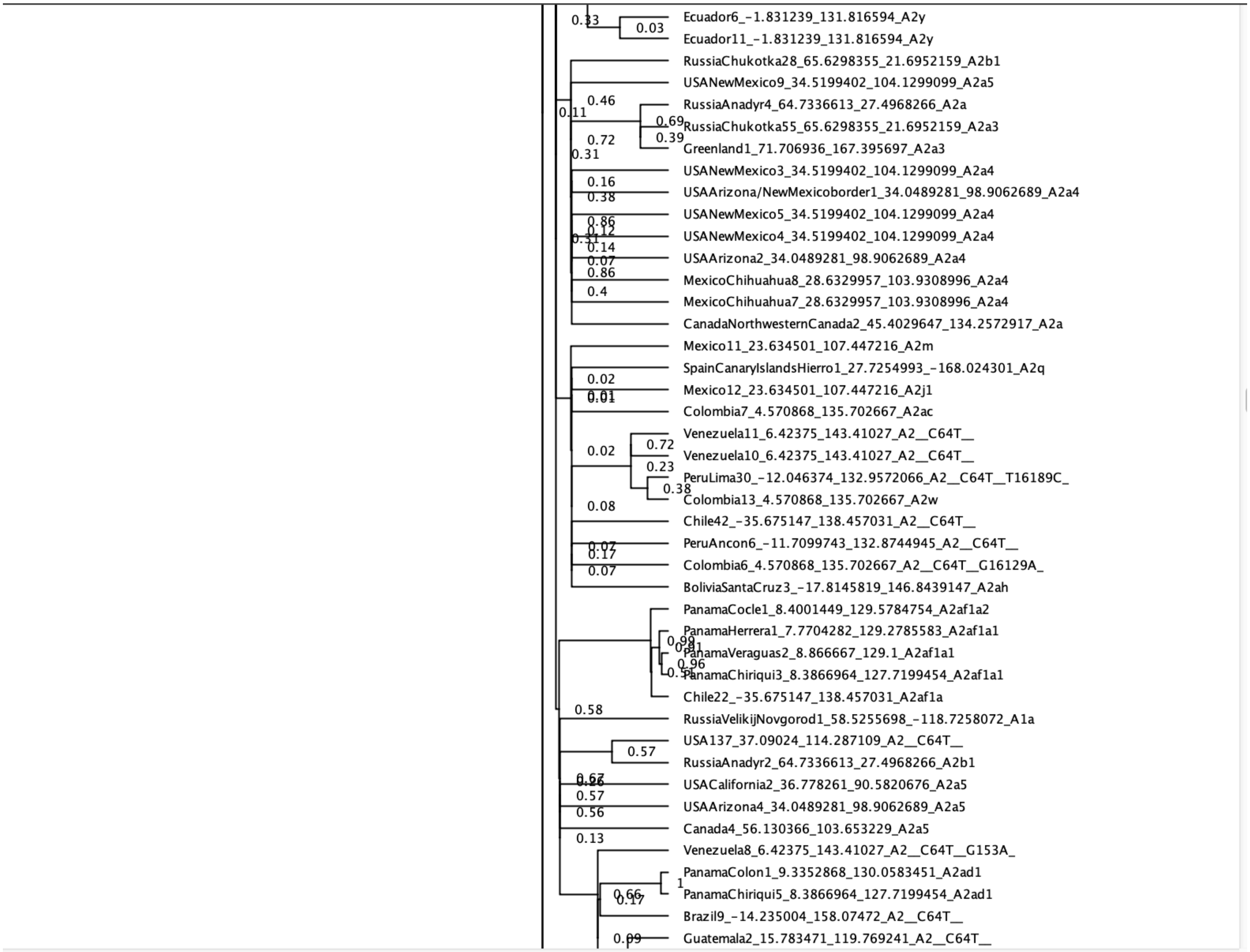
S3.19


0.71
0.07
0.26
0.72
SpainAsturias2_43.3613953_-155.8593267_L3f1b
SpainAsturias1_43.3613953_-155.8593267_L3f1b
SpainAsturias4_43.3613953_-155.8593267_L3f1b
0.38
0.72
0.83
Sudan14_12.862807_-119.782364_L3f1b_C16292T_
SpainAndalusiaHuelva6_37.261421_-156.9447224_L3f1b1
0.72
0.36
Nigeria3_9.081999_-141.324723_L3f1a1
Sudan8_12.862807_-119.782364_L3f3
0.75
0.16
Nigeria2_9.081999_-141.324723_L3f3b
PapuaNewGuineaEastNewBritainMarabu1_-4.6128943_1.8877321_Q2a
0.93
SouthKoreaSeoul1_37.566535_-23.0220308_M7c1a5
0.98
IndiaMadhyaPradesh6_22.9734229_-71.3431058_M49d
Tanzania21_-6.369028_-115.111178_L3d1a1a1
0.86
0.86
SpainAndalusiaGranada1_37.1773363_-153.5985571_L3d3b
0.39
Jordan1_30.585164_-113.761586_L3d3b
0.79
0.4
Chad6_15.454166_-131.267793_L3d2b
MoroccoBerberFiguig5_32.1092613_-151.229806_L3b1a5
0.88
Zambia75_-13.133897_-122.150668_L3b1a1a
0.57
Mauritania2_21.00789_-160.940835_L3b1a
0.41
0.17
MoroccoBerberFiguig4_32.1092613_-151.229806_L3b1a3
0.76
Chad4_15.454166_-131.267793_L3b1a_T152C_
0.47
MoroccoBerberFiguig8_32.1092613_-151.229806_L3e2b1a2
0.31
Tanzania22_-6.369028_-115.111178_L3a1a
0.96
0.25
Somalia4_5.152149_-103.800384_L3a2a
0.18
Ethiopia12_9.145_-109.510327_L3c
1
Chad12_15.454166_-131.267793_L4b2b
Chad5_15.454166_-131.267793_L3b1a9
YemenSoqotraIsland2_12.4634205_-96.1762615_L3h2
0.56
0.84
Somalia7_5.152149_-103.800384_L3h2
0.18
Nigeria1_9.081999_-141.324723_L3e5e
0.66
CameroonMasa3_1.816667_-139.633333_L3e5d
0.51
CameroonKotoko1_12.0944755_-134.9427973_L3e5
MoroccoBerberFiguig6_32.1092613_-151.229806_L3e5a1
0.18
0.66
CameroonFali2_7.516667_-139.316667_L3e5a1a
0.08
CameroonMasa1_1.816667_-139.633333_L3e5b
0.16
NigeriaKanuri1_12.8831618_-139.5383904_L3e5c
1
CameroonFali1_7.516667_-139.316667_L3e5c
Sudan35_12.862807_-119.782364_L3h1a1
0.71
Tanzania29_-6.369028_-115.111178_L3h1a2a
0.76
0.77
Ethiopia34_9.145_-109.510327_L3h1a2b
Ethiopia13_9.145_-109.510327_L3h1b1
0.76
Chad11_15.454166_-131.267793_L3h1b1a
Tanzania40_-6.369028_-115.111178_L4b2a2
SouthAfricaLimpopoProvince9_-23.4012946_-120.5820676_L3e1a3a

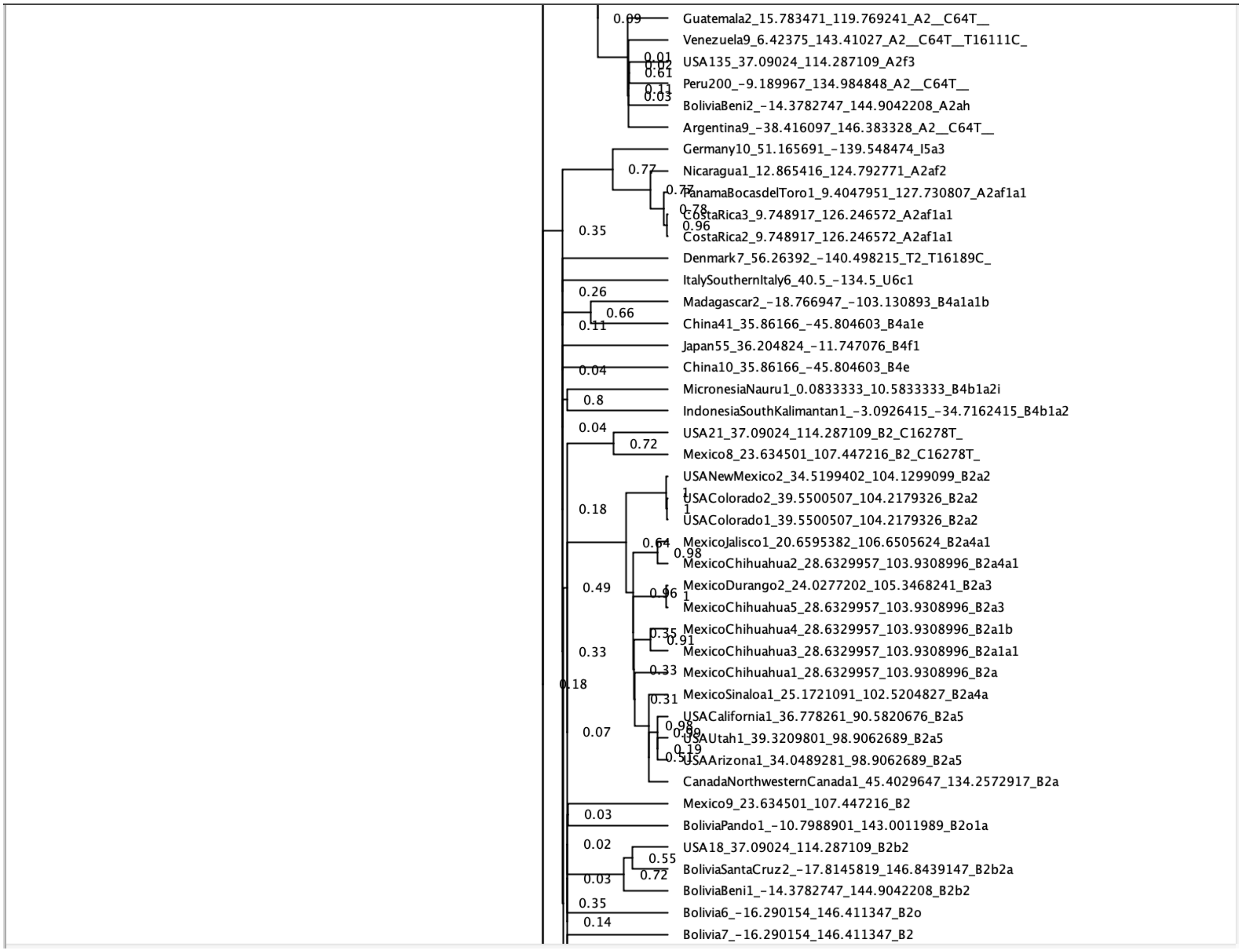
S3.20


SouthAfricaLimpopoProvince9_-23.4012946_-120.5820676_L3e1a3a
0.38
0.78
0.78
SouthAfricaMpumalangaProvince3_-25.565336_-119.4720904_L3e3b1
0.95
SouthAfricaLimpopoProvince3_-23.4012946_-120.5820676_L3e3a
0.89
MoroccoBerberFiguig9_32.1092613_-151.229806_L3e5a
0.73
Sudan13_12.862807_-119.782364_L3d4a
0.73
Morocco15_31.791702_-157.09262_L3d4
0.77
Chad10_15.454166_-131.267793_L3d3b
0.11
Somalia43_5.152149_-103.800384_L3x1b
1
Somalia28_5.152149_-103.800384_L3x1b
Zambia387_-13.133897_-122.150668_L3e1a3a
0.66
SouthAfricaLimpopoProvince4_-23.4012946_-120.5820676_L3e1a1a
1
0.78
SouthAfricaGautengProvince26_-26.2707593_-121.8877321_L3e1a1a
0.45
SouthAfricaMpumalangaProvince8_-25.565336_-119.4720904_L3e1b2
1
0.77
SouthAfricaGautengProvince21_-26.2707593_-121.8877321_L3e1b2
Zambia422_-13.133897_-122.150668_L3e1e1
0.39
Zambia405_-13.133897_-122.150668_L3e1d1a
0.86
0.44
Somalia19_5.152149_-103.800384_L3e1d1
NigeriaKanu1_12.0021794_-141.4080439_L3e5d
0.56
Chad8_15.454166_-131.267793_L3e5a1a
0.2
Angola163_-11.202692_-132.126113_L3e2b1a2
Canada6_56.130366_103.653229_X2a1
PanamaVeraguas1_8.866667_129.1_A2af1a2
0.37
0.84
PanamaKunaYala2_9.2344395_131.807375_A2af1b1a
0.59
0.37
PanamaKunaYala1_9.2344395_131.807375_A2af1b1a
PanamaChiriqui2_8.3866964_127.7199454_A2af1b1
PanamaChiriqui1_8.3866964_127.7199454_A2af1b
Colombia14_4.570868_135.702667_A2__C64T__G16129A_
0.76
0.41
USA412_37.09024_114.287109_A2__C64T__G16129A_
PanamaHerrera2_7.7704282_129.2785583_A2ad2
0.14
Panama4_8.537981_129.217873_A2w1
0.17
Canada8_56.130366_103.653229_A2i
0.31
Mexico13_23.634501_107.447216_A2__C64T__
0.07
0.42
PanamaVeraguas3_8.866667_129.1_A2ad2
0.65
PanamaColon2_9.3352868_130.0583451_A2ad1
DominicanRepublic44_18.735693_139.837349_A2ad
Ecuador8_-1.831239_131.816594_A2y
0.31
0.12
Ecuador3_-1.831239_131.816594_A2y
0.09
0.04
Ecuador4_-1.831239_131.816594_A2y
0.3
0.07
Ecuador1_-1.831239_131.816594_A2y
0.78
Ecuador7_-1.831239_131.816594_A2y
0.01
Ecuador9_-1.831239_131.816594_A2y
0.06
Ecuador2_-1.831239_131.816594_A2y
0.11
0.33
Ecuador6_-1.831239_131.816594_A2y

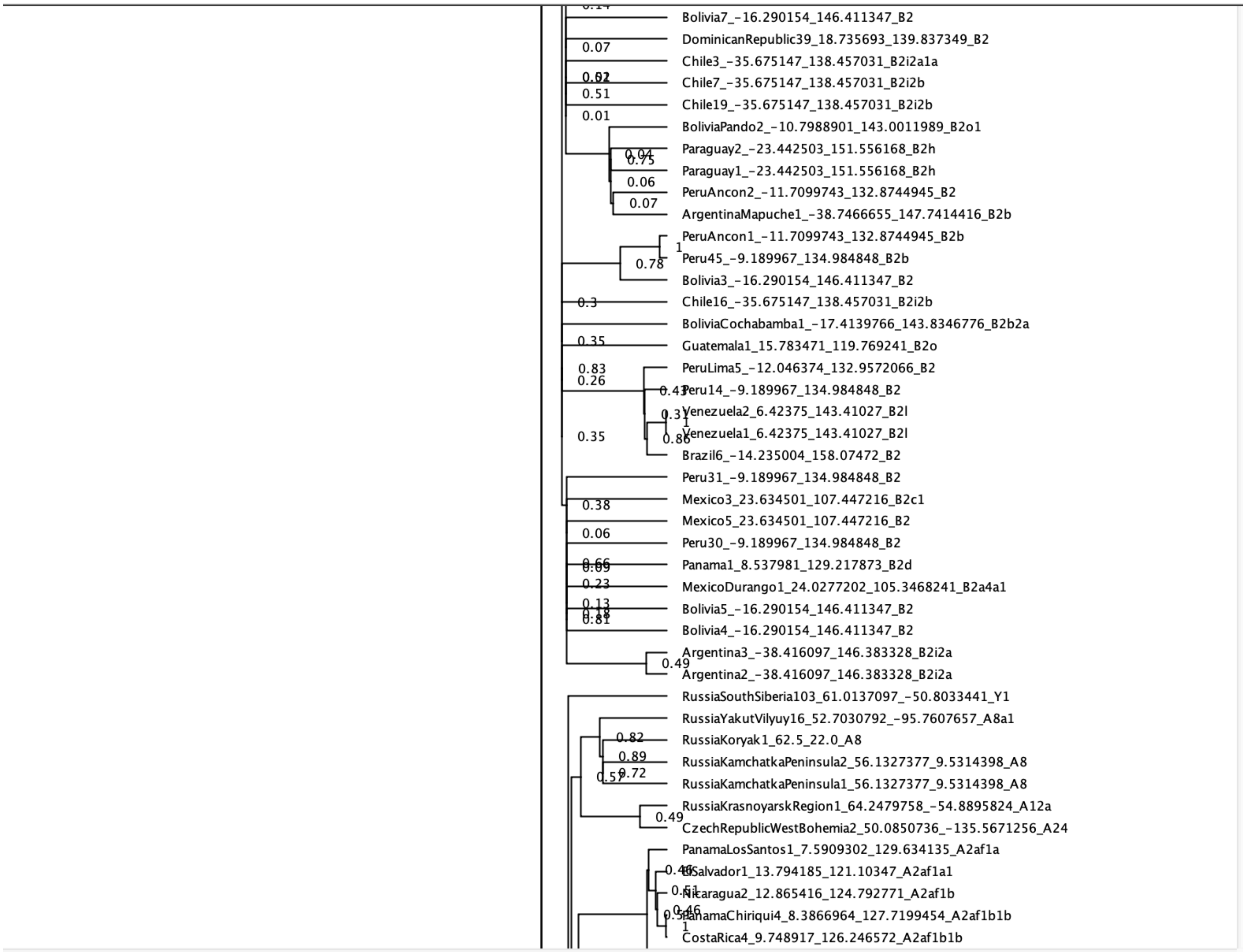
S3.21


0.33
0.03
Ecuador6_-1.831239_131.816594_A2y
Ecuador11_-1.831239_131.816594_A2y
RussiaChukotka28_65.6298355_21.6952159_A2b1
USANewMexico9_34.5199402_104.1299099_A2a5
0.46
RussiaAnadyr4_64.7336613_27.4968266_A2a
0.11
0.69
RussiaChukotka55_65.6298355_21.6952159_A2a3
0.72
0.39
Greenland1_71.706936_167.395697_A2a3
0.31
USANewMexico3_34.5199402_104.1299099_A2a4
0.16
USAArizona/NewMexicoborder1_34.0489281_98.9062689_A2a4
0.38
USANewMexico5_34.5199402_104.1299099_A2a4
0.86
0.12
USANewMexico4_34.5199402_104.1299099_A2a4
0.31
0.14
USAArizona2_34.0489281_98.9062689_A2a4
0.07
0.86
MexicoChihuahua8_28.6329957_103.9308996_A2a4
0.4
MexicoChihuahua7_28.6329957_103.9308996_A2a4
CanadaNorthwesternCanada2_45.4029647_134.2572917_A2a
Mexico11_23.634501_107.447216_A2m
SpainCanaryIslandsHierro1_27.7254993_-168.024301_A2q
0.02
Mexico12_23.634501_107.447216_A2j1
0.01
0.01
Colombia7_4.570868_135.702667_A2ac
Venezuela11_6.42375_143.41027_A2__C64T__
0.72
0.02
Venezuela10_6.42375_143.41027_A2__C64T__
0.23
PeruLima30_-12.046374_132.9572066_A2__C64T__T16189C_
0.38
Colombia13_4.570868_135.702667_A2w
0.08
Chile42_-35.675147_138.457031_A2__C64T__
0.07
PeruAncon6_-11.7099743_132.8744945_A2__C64T__
0.17
Colombia6_4.570868_135.702667_A2__C64T__G16129A_
0.07
BoliviaSantaCruz3_-17.8145819_146.8439147_A2ah
PanamaCocle1_8.4001449_129.5784754_A2af1a2
PanamaHerrera1_7.7704282_129.2785583_A2af1a1
0.99
0.81
PanamaVeraguas2_8.866667_129.1_A2af1a1
0.96
0.51
PanamaChiriqui3_8.3866964_127.7199454_A2af1a1
Chile22_-35.675147_138.457031_A2af1a
0.58
RussiaVelikijNovgorod1_58.5255698_-118.7258072_A1a
USA137_37.09024_114.287109_A2__C64T__
0.57
RussiaAnadyr2_64.7336613_27.4968266_A2b1
0.67
0.26
USACalifornia2_36.778261_90.5820676_A2a5
0.57
USAArizona4_34.0489281_98.9062689_A2a5
0.56
Canada4_56.130366_103.653229_A2a5
0.13
Venezuela8_6.42375_143.41027_A2__C64T__G153A_
PanamaColon1_9.3352868_130.0583451_A2ad1
1
0.66
0.17
PanamaChiriqui5_8.3866964_127.7199454_A2ad1
Brazil9_-14.235004_158.07472_A2__C64T__
0.09
Guatemala2_15.783471_119.769241_A2__C64T__

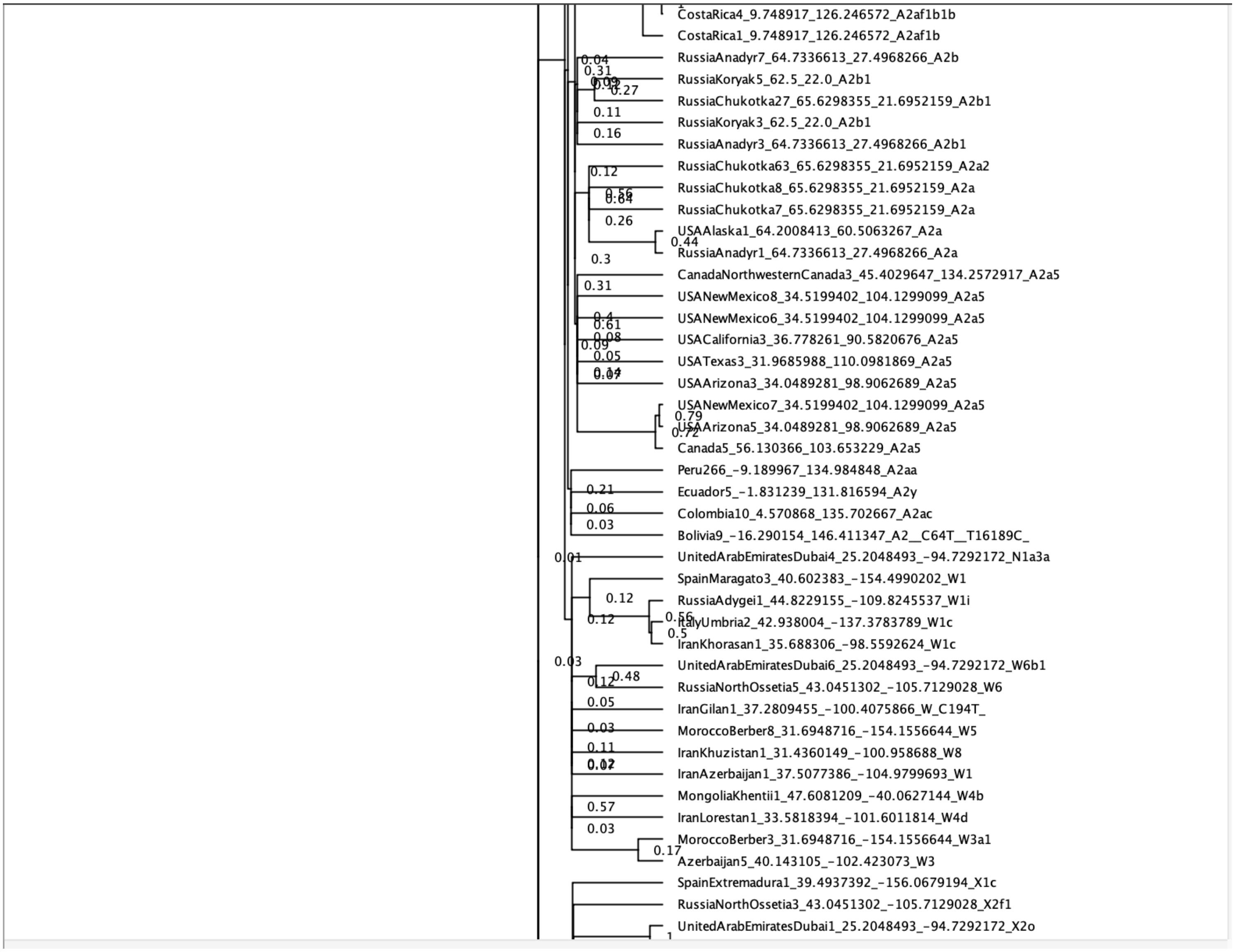
S3.22


0.09
Guatemala2_15.783471_119.769241_A2__C64T__
Venezuela9_6.42375_143.41027_A2__C64T__T16111C_
0.01
0.02
USA135_37.09024_114.287109_A2f3
0.61
Peru200_-9.189967_134.984848_A2__C64T__
0.11
0.03
BoliviaBeni2_-14.3782747_144.9042208_A2ah
Argentina9_-38.416097_146.383328_A2__C64T__
Germany10_51.165691_-139.548474_I5a3
0.77
Nicaragua1_12.865416_124.792771_A2af2
0.77
PanamaBocasdelToro1_9.4047951_127.730807_A2af1a1
0.78
CostaRica3_9.748917_126.246572_A2af1a1
0.35
0.96
CostaRica2_9.748917_126.246572_A2af1a1
Denmark7_56.26392_-140.498215_T2_T16189C_
ItalySouthernItaly6_40.5_-134.5_U6c1
0.26
Madagascar2_-18.766947_-103.130893_B4a1a1b
0.66
0.11
China41_35.86166_-45.804603_B4a1e
Japan55_36.204824_-11.747076_B4f1
0.04
China10_35.86166_-45.804603_B4e
MicronesiaNauru1_0.0833333_10.5833333_B4b1a2i
0.8
IndonesiaSouthKalimantan1_-3.0926415_-34.7162415_B4b1a2
0.04
USA21_37.09024_114.287109_B2_C16278T_
0.72
Mexico8_23.634501_107.447216_B2_C16278T_
USANewMexico2_34.5199402_104.1299099_B2a2
1
USAColorado2_39.5500507_104.2179326_B2a2
0.18
1
USAColorado1_39.5500507_104.2179326_B2a2
0.64
MexicoJalisco1_20.6595382_106.6505624_B2a4a1
0.98
MexicoChihuahua2_28.6329957_103.9308996_B2a4a1
0.49
MexicoDurango2_24.0277202_105.3468241_B2a3
0.96
1
MexicoChihuahua5_28.6329957_103.9308996_B2a3
0.35
MexicoChihuahua4_28.6329957_103.9308996_B2a1b
0.91
0.33
MexicoChihuahua3_28.6329957_103.9308996_B2a1a1
0.33
MexicoChihuahua1_28.6329957_103.9308996_B2a
0.18
0.31
MexicoSinaloa1_25.1721091_102.5204827_B2a4a
USACalifornia1_36.778261_90.5820676_B2a5
0.98
0.99
0.07
USAUtah1_39.3209801_98.9062689_B2a5
0.19
0.51
USAArizona1_34.0489281_98.9062689_B2a5
CanadaNorthwesternCanada1_45.4029647_134.2572917_B2a
Mexico9_23.634501_107.447216_B2
0.03
BoliviaPando1_-10.7988901_143.0011989_B2o1a
0.02
USA18_37.09024_114.287109_B2b2
0.55
BoliviaSantaCruz2_-17.8145819_146.8439147_B2b2a
0.03
0.72
BoliviaBeni1_-14.3782747_144.9042208_B2b2
0.35
Bolivia6_-16.290154_146.411347_B2o
0.14
Bolivia7_-16.290154_146.411347_B2

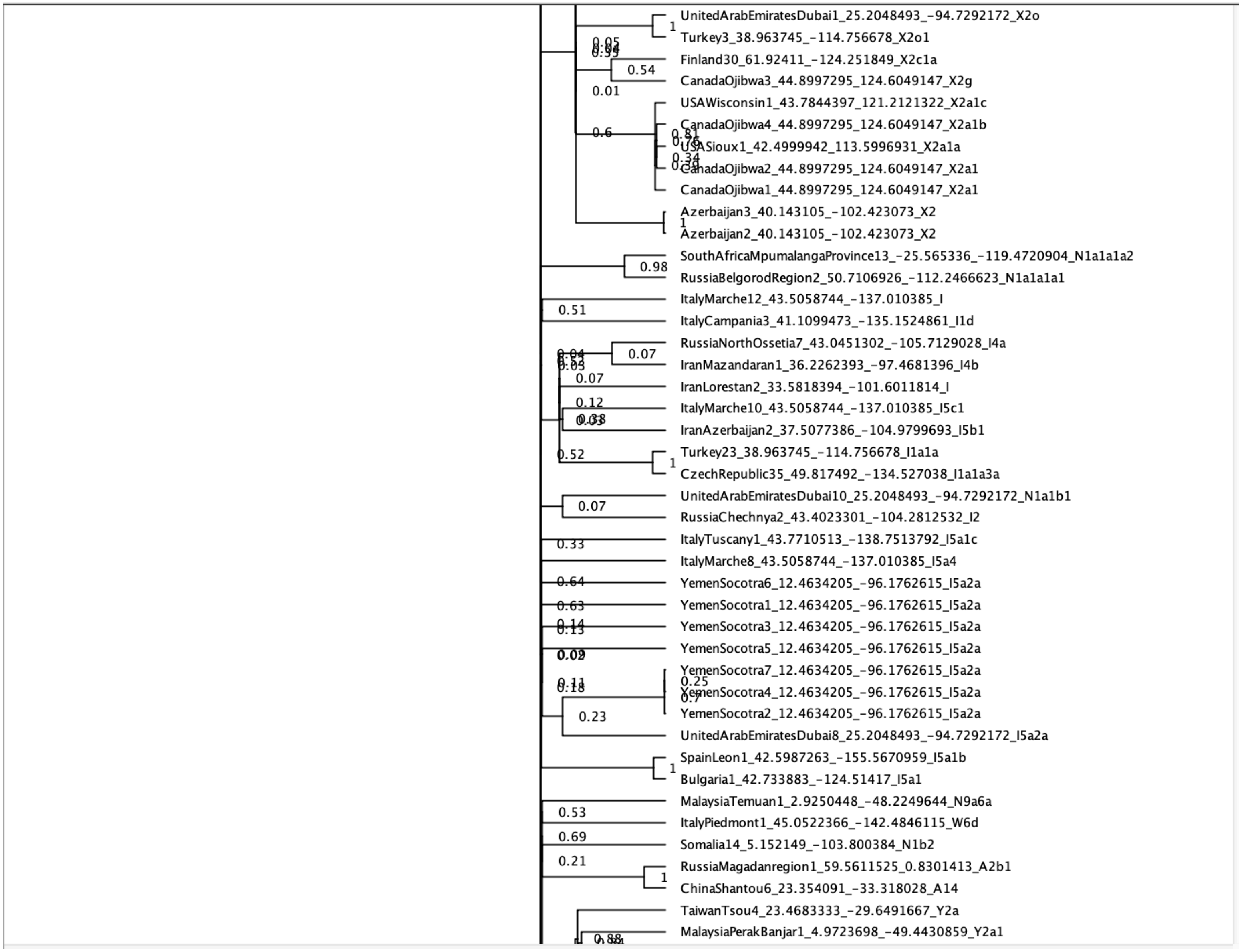
S3.23


Bolivia7_-16.290154_146.411347_B2
DominicanRepublic39_18.735693_139.837349_B2
0.07
Chile3_-35.675147_138.457031_B2i2a1a
Chile7_-35.675147_138.457031_B2i2b
0.51
Chile19_-35.675147_138.457031_B2i2b
0.01
BoliviaPando2_-10.7988901_143.0011989_B2o1
Paraguay2_-23.442503_151.556168_B2h
0.75
Paraguay1_-23.442503_151.556168_B2h
0.06
PeruAncon2_-11.7099743_132.8744945_B2
0.07
ArgentinaMapuche1_-38.7466655_147.7414416_B2b
PeruAncon1_-11.7099743_132.8744945_B2b
1
Peru45_-9.189967_134.984848_B2b
0.78
Bolivia3_-16.290154_146.411347_B2
0.3
Chile16_-35.675147_138.457031_B2i2b
BoliviaCochabamba1_-17.4139766_143.8346776_B2b2a
0.35
Guatemala1_15.783471_119.769241_B2o
0.83
PeruLima5_-12.046374_132.9572066_B2
0.26
0.43
Peru14_-9.189967_134.984848_B2
0.31
Venezuela2_6.42375_143.41027_B2l
1
0.86
Venezuela1_6.42375_143.41027_B2l
0.35
Brazil6_-14.235004_158.07472_B2
Peru31_-9.189967_134.984848_B2
0.38
Mexico3_23.634501_107.447216_B2c1
Mexico5_23.634501_107.447216_B2
0.06
Peru30_-9.189967_134.984848_B2
0.66
Panama1_8.537981_129.217873_B2d
0.23
MexicoDurango1_24.0277202_105.3468241_B2a4a1
0.13
Bolivia5_-16.290154_146.411347_B2
0.81
Bolivia4_-16.290154_146.411347_B2
Argentina3_-38.416097_146.383328_B2i2a
0.49
Argentina2_-38.416097_146.383328_B2i2a
RussiaSouthSiberia103_61.0137097_-50.8033441_Y1
RussiaYakutVilyuy16_52.7030792_-95.7607657_A8a1
0.82
RussiaKoryak1_62.5_22.0_A8
0.89
RussiaKamchatkaPeninsula2_56.1327377_9.5314398_A8
0.57
0.72
RussiaKamchatkaPeninsula1_56.1327377_9.5314398_A8
RussiaKrasnoyarskRegion1_64.2479758_-54.8895824_A12a
0.49
CzechRepublicWestBohemia2_50.0850736_-135.5671256_A24
PanamaLosSantos1_7.5909302_129.634135_A2af1a
0.46
ElSalvador1_13.794185_121.10347_A2af1a1
0.51
Nicaragua2_12.865416_124.792771_A2af1b
0.46
PanamaChiriqui4_8.3866964_127.7199454_A2af1b1b
1
CostaRica4_9.748917_126.246572_A2af1b1b

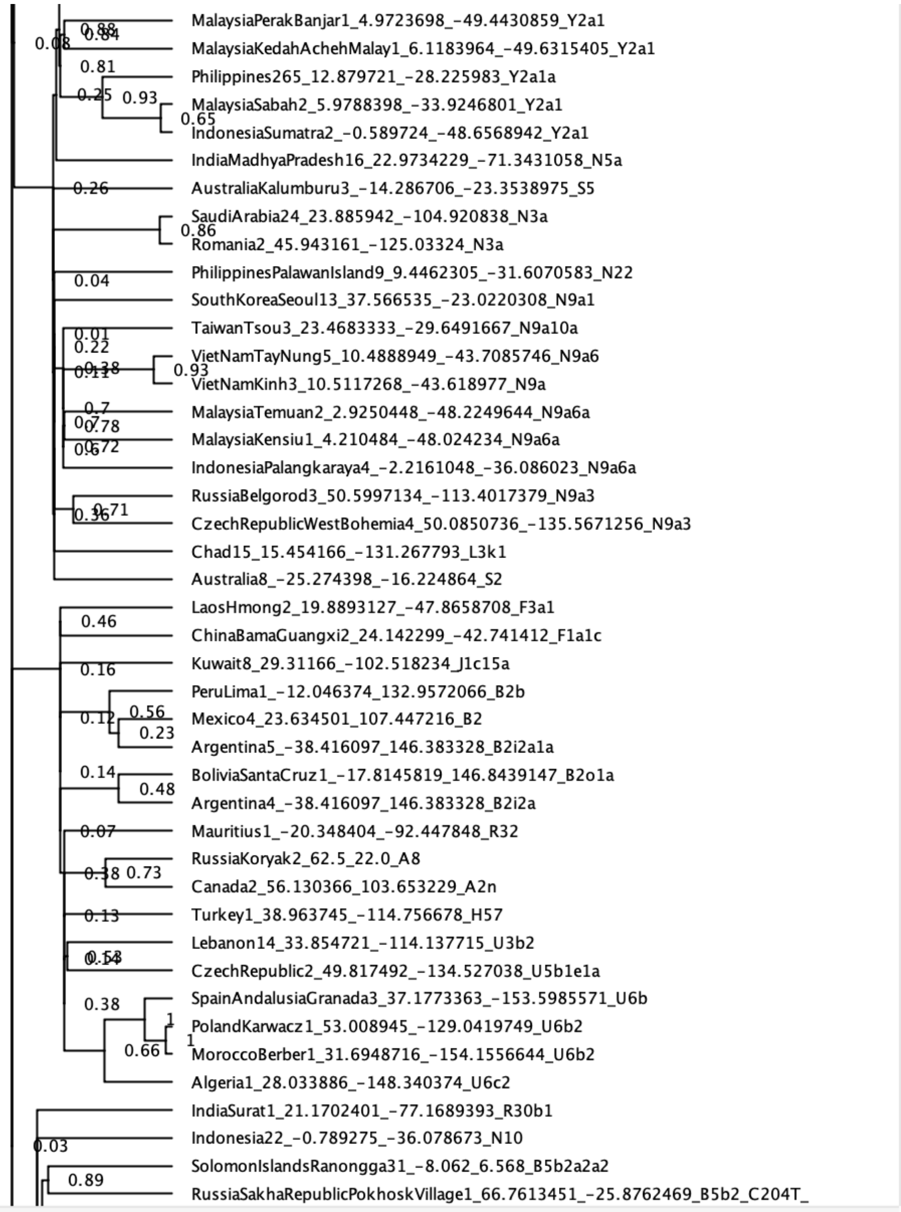
S3.24


CostaRica4_9.748917_126.246572_A2af1b1b
CostaRica1_9.748917_126.246572_A2af1b
RussiaAnadyr7_64.7336613_27.4968266_A2b
RussiaKoryak5_62.5_22.0_A2b1
RussiaChukotka27_65.6298355_21.6952159_A2b1
RussiaKoryak3_62.5_22.0_A2b1
RussiaAnadyr3_64.7336613_27.4968266_A2b1
RussiaChukotka63_65.6298355_21.6952159_A2a2
RussiaChukotka8_65.6298355_21.6952159_A2a
RussiaChukotka7_65.6298355_21.6952159_A2a
USAAlaska1_64.2008413_60.5063267_A2a
RussiaAnadyr1_64.7336613_27.4968266_A2a
CanadaNorthwesternCanada3_45.4029647_134.2572917_A2a5
USANewMexico8_34.5199402_104.1299099_A2a5
USANewMexico6_34.5199402_104.1299099_A2a5
USACalifornia3_36.778261_90.5820676_A2a5
USATexas3_31.9685988_110.0981869_A2a5
USAArizona3_34.0489281_98.9062689_A2a5
USANewMexico7_34.5199402_104.1299099_A2a5
USAArizona5_34.0489281_98.9062689_A2a5
Canada5_56.130366_103.653229_A2a5
Peru266_-9.189967_134.984848_A2aa
Ecuador5_-1.831239_131.816594_A2y
Colombia10_4.570868_135.702667_A2ac
Bolivia9_-16.290154_146.411347_A2__C64T__T16189C_
UnitedArabEmiratesDubai4_25.2048493_-94.7292172_N1a3a
SpainMaragato3_40.602383_-154.4990202_W1
RussiaAdygei1_44.8229155_-109.8245537_W1i
ItalyUmbria2_42.938004_-137.3783789_W1c
IranKhorasan1_35.688306_-98.5592624_W1c
UnitedArabEmiratesDubai6_25.2048493_-94.7292172_W6b1
RussiaNorthOssetia5_43.0451302_-105.7129028_W6
IranGilan1_37.2809455_-100.4075866_W_C194T_
MoroccoBerber8_31.6948716_-154.1556644_W5
IranKhuzistan1_31.4360149_-100.958688_W8
IranAzerbaijan1_37.5077386_-104.9799693_W1
MongoliaKhentii1_47.6081209_-40.0627144_W4b
IranLorestan1_33.5818394_-101.6011814_W4d
MoroccoBerber3_31.6948716_-154.1556644_W3a1
Azerbaijan5_40.143105_-102.423073_W3
SpainExtremadura1_39.4937392_-156.0679194_X1c
RussiaNorthOssetia3_43.0451302_-105.7129028_X2f1
UnitedArabEmiratesDubai1_25.2048493_-94.7292172_X2o
1
0.04
0.31
0.09
0.27
0.11
0.16
0.12
0.56
0.64
0.26
0.44
0.3
0.31
0.4
0.61
0.08
0.09
0.05
0.14
0.07
0.79
0.72
0.21
0.06
0.03
0.01
0.12
0.12
0.56
0.5
0.03
0.48
0.12
0.05
0.03
0.11
0.02
0.07
0.57
0.03
0.17
1

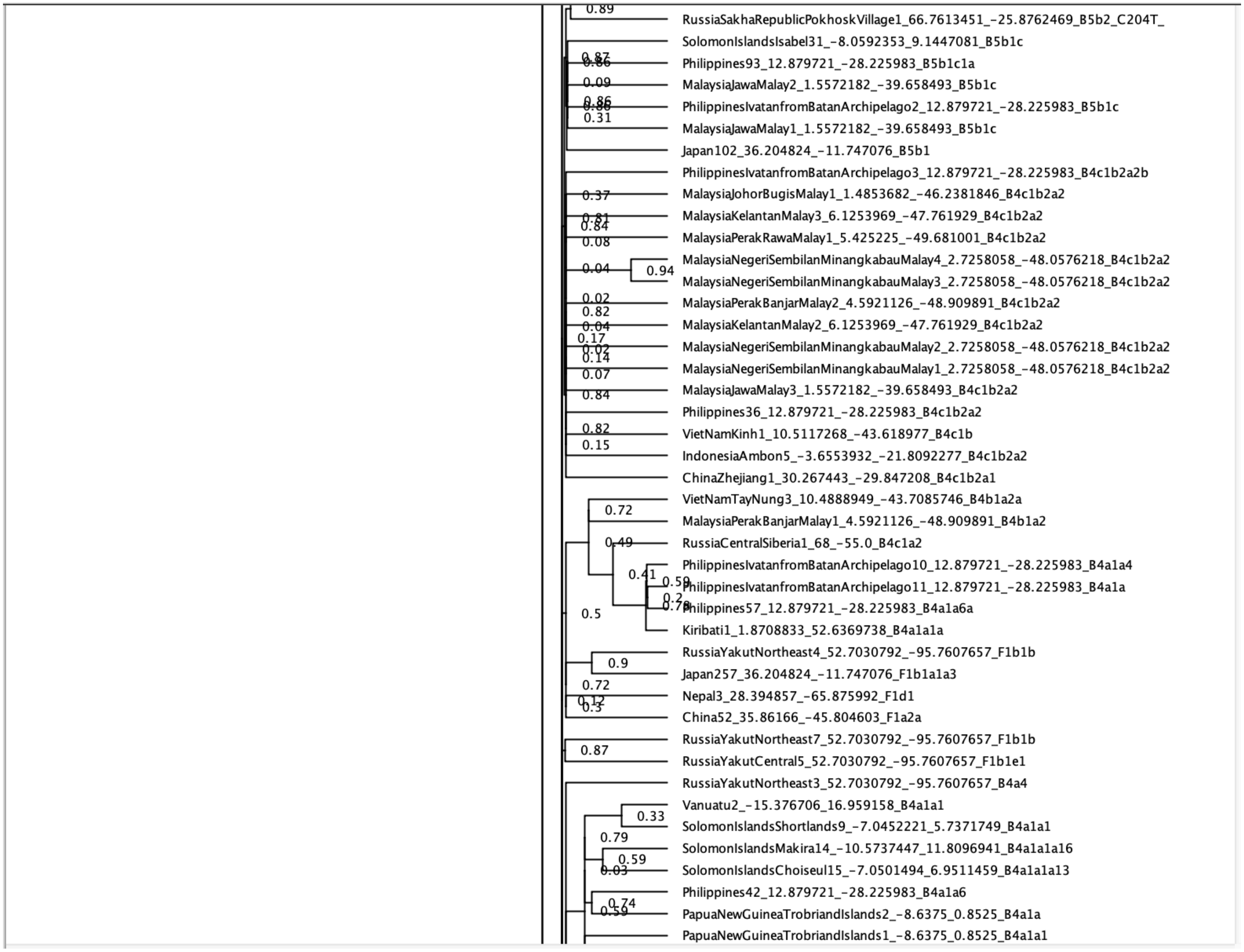
S3.25


UnitedArabEmiratesDubai1_25.2048493_-94.7292172_X2o
1
Turkey3_38.963745_-114.756678_X2o1
0.05
0.54
0.55
Finland30_61.92411_-124.251849_X2c1a
0.54
CanadaOjibwa3_44.8997295_124.6049147_X2g
0.01
USAWisconsin1_43.7844397_121.2121322_X2a1c
CanadaOjibwa4_44.8997295_124.6049147_X2a1b
0.6
0.81
0.76
USASioux1_42.4999942_113.5996931_X2a1a
0.34
0.34
CanadaOjibwa2_44.8997295_124.6049147_X2a1
CanadaOjibwa1_44.8997295_124.6049147_X2a1
Azerbaijan3_40.143105_-102.423073_X2
1
Azerbaijan2_40.143105_-102.423073_X2
SouthAfricaMpumalangaProvince13_-25.565336_-119.4720904_N1a1a1a2
0.98
RussiaBelgorodRegion2_50.7106926_-112.2466623_N1a1a1a1
ItalyMarche12_43.5058744_-137.010385_I
0.51
ItalyCampania3_41.1099473_-135.1524861_I1d
RussiaNorthOssetia7_43.0451302_-105.7129028_I4a
0.04
0.52
0.03
0.07
IranMazandaran1_36.2262393_-97.4681396_I4b
0.07
IranLorestan2_33.5818394_-101.6011814_I
0.12
ItalyMarche10_43.5058744_-137.010385_I5c1
0.38
0.03
IranAzerbaijan2_37.5077386_-104.9799693_I5b1
0.52
Turkey23_38.963745_-114.756678_I1a1a
1
CzechRepublic35_49.817492_-134.527038_I1a1a3a
UnitedArabEmiratesDubai10_25.2048493_-94.7292172_N1a1b1
0.07
RussiaChechnya2_43.4023301_-104.2812532_I2
ItalyTuscany1_43.7710513_-138.7513792_I5a1c
0.33
ItalyMarche8_43.5058744_-137.010385_I5a4
0.64
YemenSocotra6_12.4634205_-96.1762615_I5a2a
0.63
YemenSocotra1_12.4634205_-96.1762615_I5a2a
0.14
0.13
YemenSocotra3_12.4634205_-96.1762615_I5a2a
YemenSocotra5_12.4634205_-96.1762615_I5a2a
0.09
0.02
YemenSocotra7_12.4634205_-96.1762615_I5a2a
0.11
0.18
0.25
YemenSocotra4_12.4634205_-96.1762615_I5a2a
0.7
YemenSocotra2_12.4634205_-96.1762615_I5a2a
0.23
UnitedArabEmiratesDubai8_25.2048493_-94.7292172_I5a2a
SpainLeon1_42.5987263_-155.5670959_I5a1b
1
Bulgaria1_42.733883_-124.51417_I5a1
MalaysiaTemuan1_2.9250448_-48.2249644_N9a6a
0.53
ItalyPiedmont1_45.0522366_-142.4846115_W6d
0.69
Somalia14_5.152149_-103.800384_N1b2
0.21
RussiaMagadanregion1_59.5611525_0.8301413_A2b1
1
ChinaShantou6_23.354091_-33.318028_A14
TaiwanTsou4_23.4683333_-29.6491667_Y2a
MalaysiaPerakBanjar1_4.9723698_-49.4430859_Y2a1
0.88
0.88

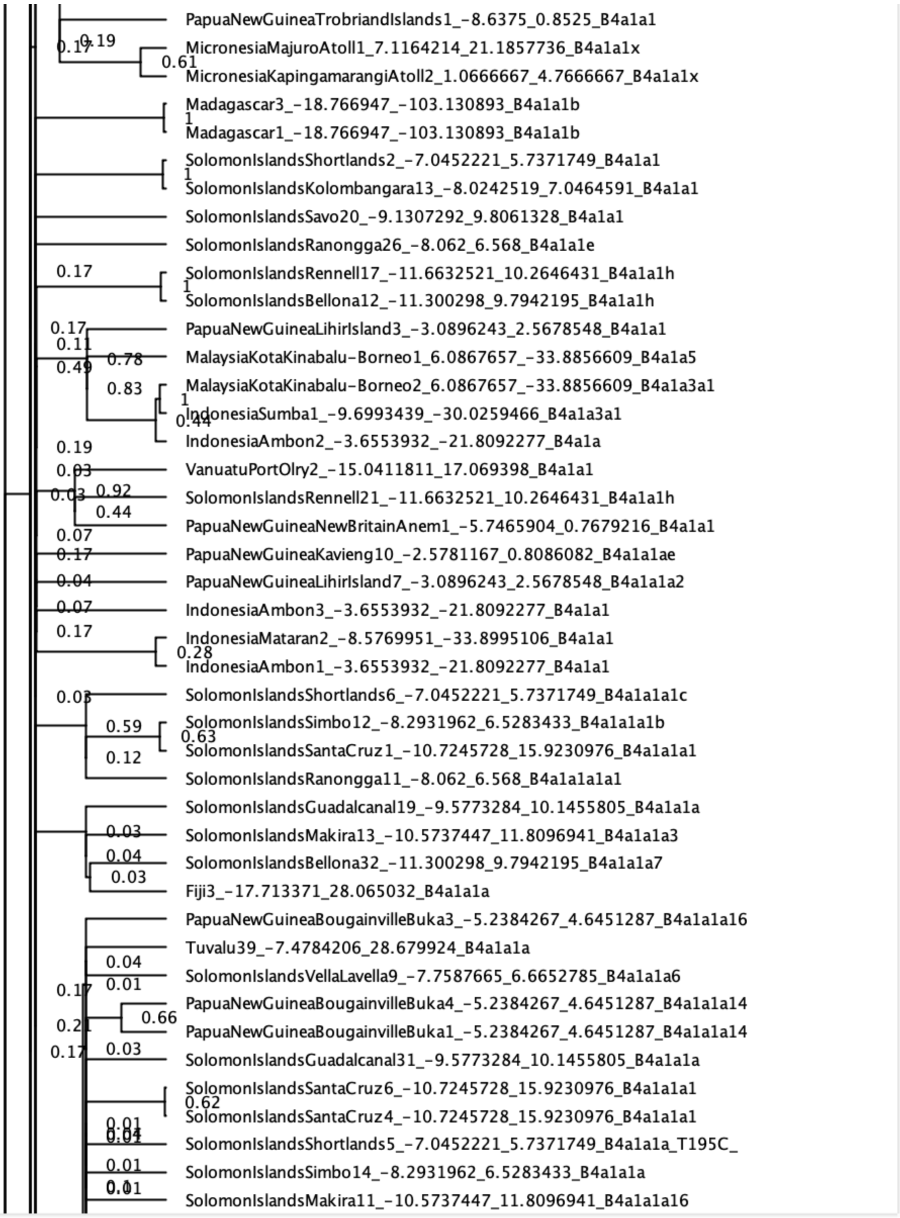
S3.26


MalaysiaPerakBanjar1_4.9723698_-49.4430859_Y2a1
MalaysiaKedahAchehMalay1_6.1183964_-49.6315405_Y2a1
Philippines265_12.879721_-28.225983_Y2a1a
MalaysiaSabah2_5.9788398_-33.9246801_Y2a1
IndonesiaSumatra2_-0.589724_-48.6568942_Y2a1
IndiaMadhyaPradesh16_22.9734229_-71.3431058_N5a
AustraliaKalumburu3_-14.286706_-23.3538975_S5
SaudiArabia24_23.885942_-104.920838_N3a
Romania2_45.943161_-125.03324_N3a
PhilippinesPalawanIsland9_9.4462305_-31.6070583_N22
SouthKoreaSeoul13_37.566535_-23.0220308_N9a1
TaiwanTsou3_23.4683333_-29.6491667_N9a10a
VietNamTayNung5_10.4888949_-43.7085746_N9a6
VietNamKinh3_10.5117268_-43.618977_N9a
MalaysiaTemuan2_2.9250448_-48.2249644_N9a6a
MalaysiaKensiu1_4.210484_-48.024234_N9a6a
IndonesiaPalangkaraya4_-2.2161048_-36.086023_N9a6a
RussiaBelgorod3_50.5997134_-113.4017379_N9a3
CzechRepublicWestBohemia4_50.0850736_-135.5671256_N9a3
Chad15_15.454166_-131.267793_L3k1
Australia8_-25.274398_-16.224864_S2
LaosHmong2_19.8893127_-47.8658708_F3a1
ChinaBamaGuangxi2_24.142299_-42.741412_F1a1c
Kuwait8_29.31166_-102.518234_J1c15a
PeruLima1_-12.046374_132.9572066_B2b
Mexico4_23.634501_107.447216_B2
Argentina5_-38.416097_146.383328_B2i2a1a
BoliviaSantaCruz1_-17.8145819_146.8439147_B2o1a
Argentina4_-38.416097_146.383328_B2i2a
Mauritius1_-20.348404_-92.447848_R32
RussiaKoryak2_62.5_22.0_A8
Canada2_56.130366_103.653229_A2n
Turkey1_38.963745_-114.756678_H57
Lebanon14_33.854721_-114.137715_U3b2
CzechRepublic2_49.817492_-134.527038_U5b1e1a
SpainAndalusiaGranada3_37.1773363_-153.5985571_U6b
PolandKarwacz1_53.008945_-129.0419749_U6b2
MoroccoBerber1_31.6948716_-154.1556644_U6b2
Algeria1_28.033886_-148.340374_U6c2
IndiaSurat1_21.1702401_-77.1689393_R30b1
Indonesia22_-0.789275_-36.078673_N10
SolomonIslandsRanongga31_-8.062_6.568_B5b2a2a2
RussiaSakhaRepublicPokhoskVillage1_66.7613451_-25.8762469_B5b2_C204T_
0.88
0.84
0.08
0.81
0.25
0.93
0.65
0.26
0.86
0.04
0.01
0.22
0.38
0.11
0.93
0.7
0.7
0.78
0.6
0.72
0.36
0.71
0.46
0.16
0.12
0.56
0.23
0.14
0.48
0.07
0.38
0.73
0.13
0.58
0.14
0.38
1
1
0.66
0.03
0.89

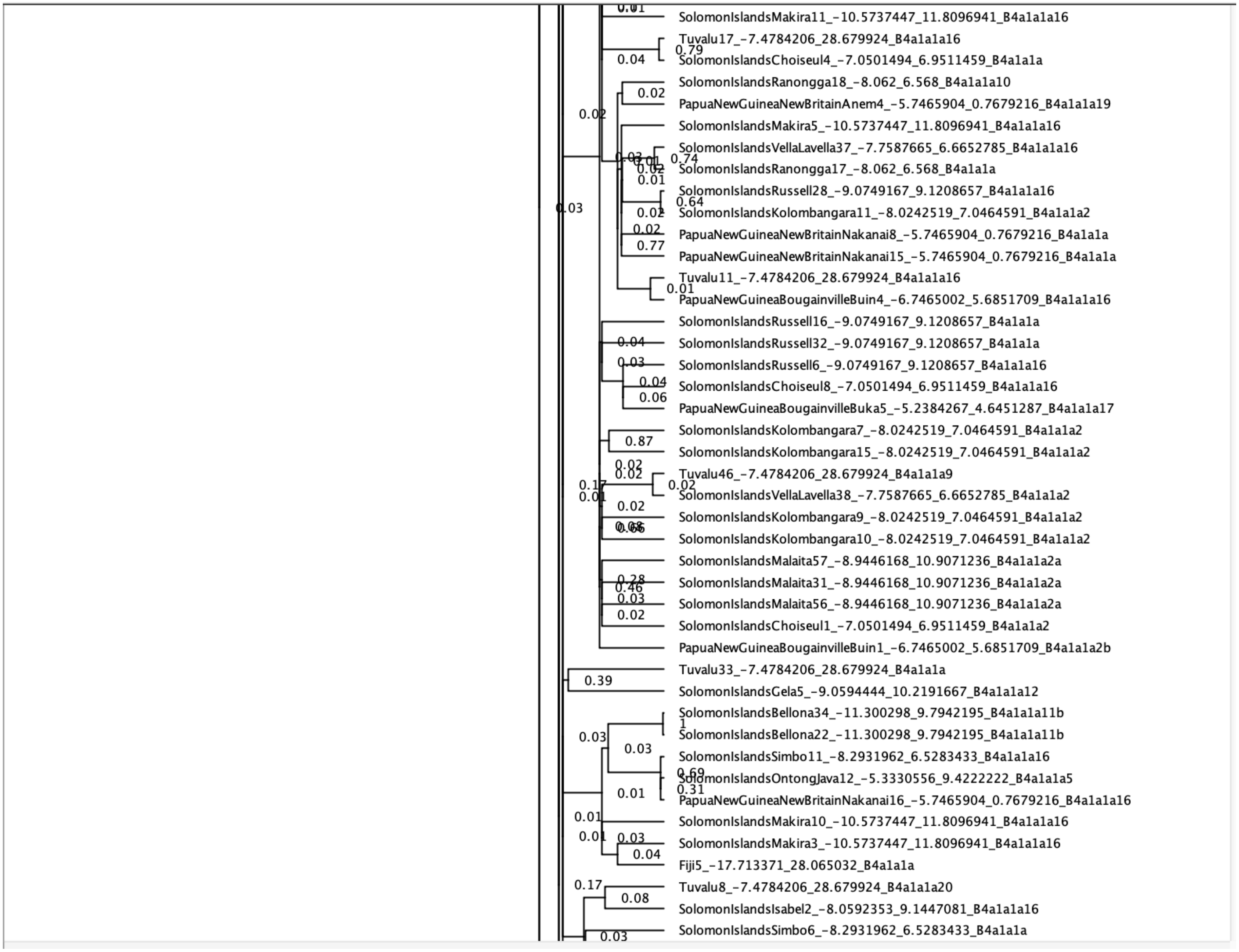
S3.27


0.89
RussiaSakhaRepublicPokhoskVillage1_66.7613451_-25.8762469_B5b2_C204T_
SolomonIslandsIsabel31_-8.0592353_9.1447081_B5b1c
0.87
0.86
Philippines93_12.879721_-28.225983_B5b1c1a
0.09
MalaysiaJawaMalay2_1.5572182_-39.658493_B5b1c
0.86
0.88
PhilippinesIvatanfromBatanArchipelago2_12.879721_-28.225983_B5b1c
0.31
MalaysiaJawaMalay1_1.5572182_-39.658493_B5b1c
Japan102_36.204824_-11.747076_B5b1
PhilippinesIvatanfromBatanArchipelago3_12.879721_-28.225983_B4c1b2a2b
0.37
MalaysiaJohorBugisMalay1_1.4853682_-46.2381846_B4c1b2a2
0.81
0.84
MalaysiaKelantanMalay3_6.1253969_-47.761929_B4c1b2a2
0.08
MalaysiaPerakRawaMalay1_5.425225_-49.681001_B4c1b2a2
0.04
0.94
MalaysiaNegeriSembilanMinangkabauMalay4_2.7258058_-48.0576218_B4c1b2a2
MalaysiaNegeriSembilanMinangkabauMalay3_2.7258058_-48.0576218_B4c1b2a2
0.02
0.82
MalaysiaPerakBanjarMalay2_4.5921126_-48.909891_B4c1b2a2
0.04
0.17
MalaysiaKelantanMalay2_6.1253969_-47.761929_B4c1b2a2
0.02
0.14
MalaysiaNegeriSembilanMinangkabauMalay2_2.7258058_-48.0576218_B4c1b2a2
0.07
MalaysiaNegeriSembilanMinangkabauMalay1_2.7258058_-48.0576218_B4c1b2a2
0.84
MalaysiaJawaMalay3_1.5572182_-39.658493_B4c1b2a2
Philippines36_12.879721_-28.225983_B4c1b2a2
0.82
VietNamKinh1_10.5117268_-43.618977_B4c1b
0.15
IndonesiaAmbon5_-3.6553932_-21.8092277_B4c1b2a2
ChinaZhejiang1_30.267443_-29.847208_B4c1b2a1
0.72
VietNamTayNung3_10.4888949_-43.7085746_B4b1a2a
MalaysiaPerakBanjarMalay1_4.5921126_-48.909891_B4b1a2
0.49
RussiaCentralSiberia1_68_-55.0_B4c1a2
PhilippinesIvatanfromBatanArchipelago10_12.879721_-28.225983_B4a1a4
0.41
0.59
PhilippinesIvatanfromBatanArchipelago11_12.879721_-28.225983_B4a1a
0.2
0.78
Philippines57_12.879721_-28.225983_B4a1a6a
0.5
Kiribati1_1.8708833_52.6369738_B4a1a1a
RussiaYakutNortheast4_52.7030792_-95.7607657_F1b1b
0.9
Japan257_36.204824_-11.747076_F1b1a1a3
0.72
Nepal3_28.394857_-65.875992_F1d1
0.12
0.3
China52_35.86166_-45.804603_F1a2a
RussiaYakutNortheast7_52.7030792_-95.7607657_F1b1b
0.87
RussiaYakutCentral5_52.7030792_-95.7607657_F1b1e1
RussiaYakutNortheast3_52.7030792_-95.7607657_B4a4
Vanuatu2_-15.376706_16.959158_B4a1a1
0.33
SolomonIslandsShortlands9_-7.0452221_5.7371749_B4a1a1
0.79
SolomonIslandsMakira14_-10.5737447_11.8096941_B4a1a1a16
0.59
0.03
SolomonIslandsChoiseul15_-7.0501494_6.9511459_B4a1a1a13
Philippines42_12.879721_-28.225983_B4a1a6
0.74
0.59
PapuaNewGuineaTrobriandIslands2_-8.6375_0.8525_B4a1a
PapuaNewGuineaTrobriandIslands1_-8.6375_0.8525_B4a1a1

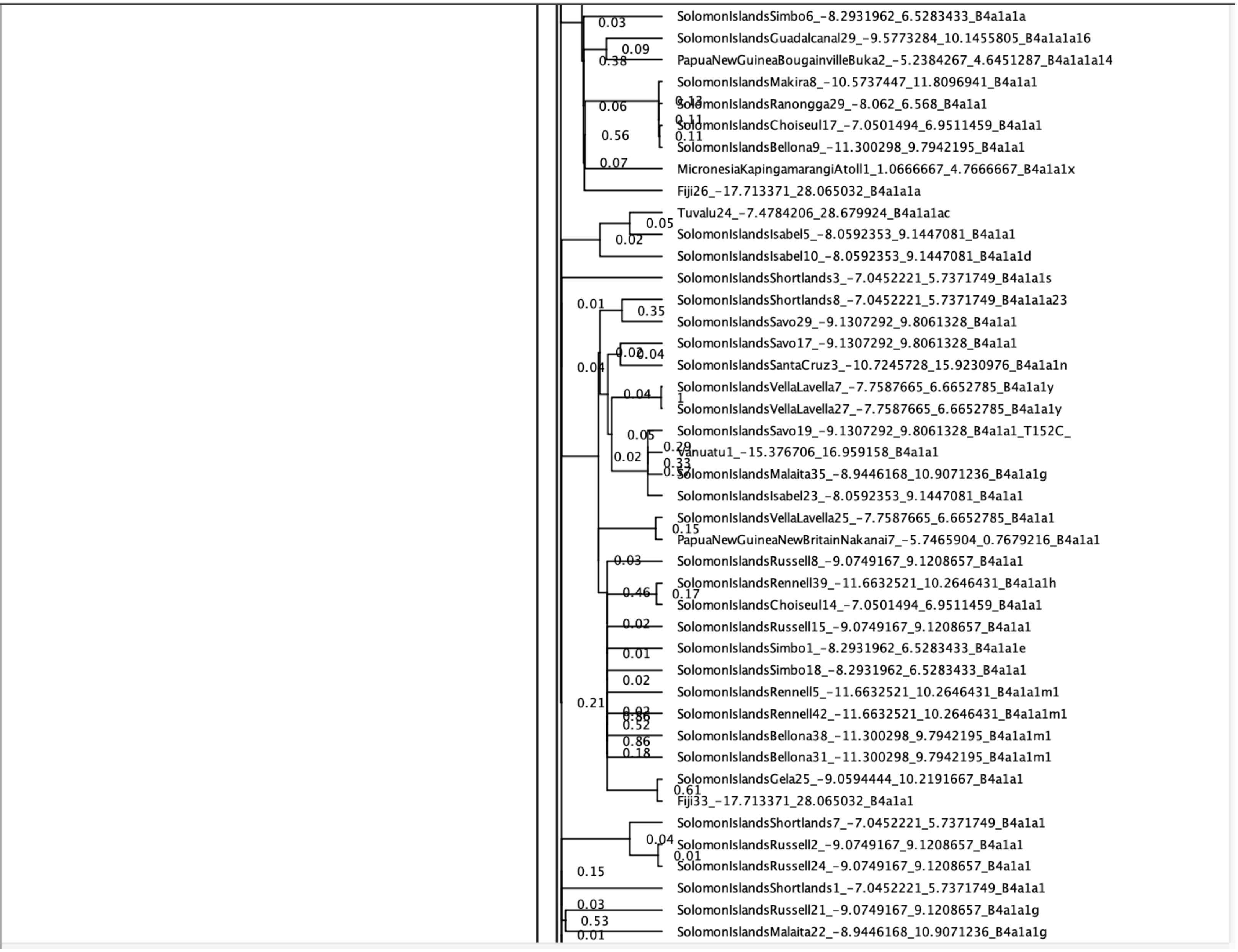
S3.28


PapuaNewGuineaTrobriandIslands1_-8.6375_0.8525_B4a1a1
0.17
0.19
MicronesiaMajuroAtoll1_7.1164214_21.1857736_B4a1a1x
0.61
MicronesiaKapingamarangiAtoll2_1.0666667_4.7666667_B4a1a1x
Madagascar3_-18.766947_-103.130893_B4a1a1b
1
Madagascar1_-18.766947_-103.130893_B4a1a1b
SolomonIslandsShortlands2_-7.0452221_5.7371749_B4a1a1
1
SolomonIslandsKolombangara13_-8.0242519_7.0464591_B4a1a1
SolomonIslandsSavo20_-9.1307292_9.8061328_B4a1a1
SolomonIslandsRanongga26_-8.062_6.568_B4a1a1e
0.17
SolomonIslandsRennell17_-11.6632521_10.2646431_B4a1a1h
1
SolomonIslandsBellona12_-11.300298_9.7942195_B4a1a1h
0.17
PapuaNewGuineaLihirIsland3_-3.0896243_2.5678548_B4a1a1
0.11
0.78
MalaysiaKotaKinabalu-Borneo1_6.0867657_-33.8856609_B4a1a5
0.49
0.83
MalaysiaKotaKinabalu-Borneo2_6.0867657_-33.8856609_B4a1a3a1
1
IndonesiaSumba1_-9.6993439_-30.0259466_B4a1a3a1
0.44
IndonesiaAmbon2_-3.6553932_-21.8092277_B4a1a
0.19
0.03
VanuatuPortOlry2_-15.0411811_17.069398_B4a1a1
0.03
0.92
SolomonIslandsRennell21_-11.6632521_10.2646431_B4a1a1h
0.44
PapuaNewGuineaNewBritainAnem1_-5.7465904_0.7679216_B4a1a1
0.07
0.17
PapuaNewGuineaKavieng10_-2.5781167_0.8086082_B4a1a1ae
0.04
PapuaNewGuineaLihirIsland7_-3.0896243_2.5678548_B4a1a1a2
0.07
IndonesiaAmbon3_-3.6553932_-21.8092277_B4a1a1
0.17
IndonesiaMataran2_-8.5769951_-33.8995106_B4a1a1
0.28
IndonesiaAmbon1_-3.6553932_-21.8092277_B4a1a1
SolomonIslandsShortlands6_-7.0452221_5.7371749_B4a1a1a1c
0.03
0.59
SolomonIslandsSimbo12_-8.2931962_6.5283433_B4a1a1a1b
0.63
SolomonIslandsSantaCruz1_-10.7245728_15.9230976_B4a1a1a1
0.12
SolomonIslandsRanongga11_-8.062_6.568_B4a1a1a1a1
SolomonIslandsGuadalcanal19_-9.5773284_10.1455805_B4a1a1a
0.03
SolomonIslandsMakira13_-10.5737447_11.8096941_B4a1a1a3
0.04
SolomonIslandsBellona32_-11.300298_9.7942195_B4a1a1a7
0.03
Fiji3_-17.713371_28.065032_B4a1a1a
PapuaNewGuineaBougainvilleBuka3_-5.2384267_4.6451287_B4a1a1a16
Tuvalu39_-7.4784206_28.679924_B4a1a1a
0.04
0.01
SolomonIslandsVellaLavella9_-7.7587665_6.6652785_B4a1a1a6
0.17
PapuaNewGuineaBougainvilleBuka4_-5.2384267_4.6451287_B4a1a1a14
0.66
0.21
PapuaNewGuineaBougainvilleBuka1_-5.2384267_4.6451287_B4a1a1a14
0.17
0.03
SolomonIslandsGuadalcanal31_-9.5773284_10.1455805_B4a1a1a
SolomonIslandsSantaCruz6_-10.7245728_15.9230976_B4a1a1a1
0.62
SolomonIslandsSantaCruz4_-10.7245728_15.9230976_B4a1a1a1
0.01
0.04
SolomonIslandsShortlands5_-7.0452221_5.7371749_B4a1a1a_T195C_
0.01
SolomonIslandsSimbo14_-8.2931962_6.5283433_B4a1a1a
0.01
SolomonIslandsMakira11_-10.5737447_11.8096941_B4a1a1a16

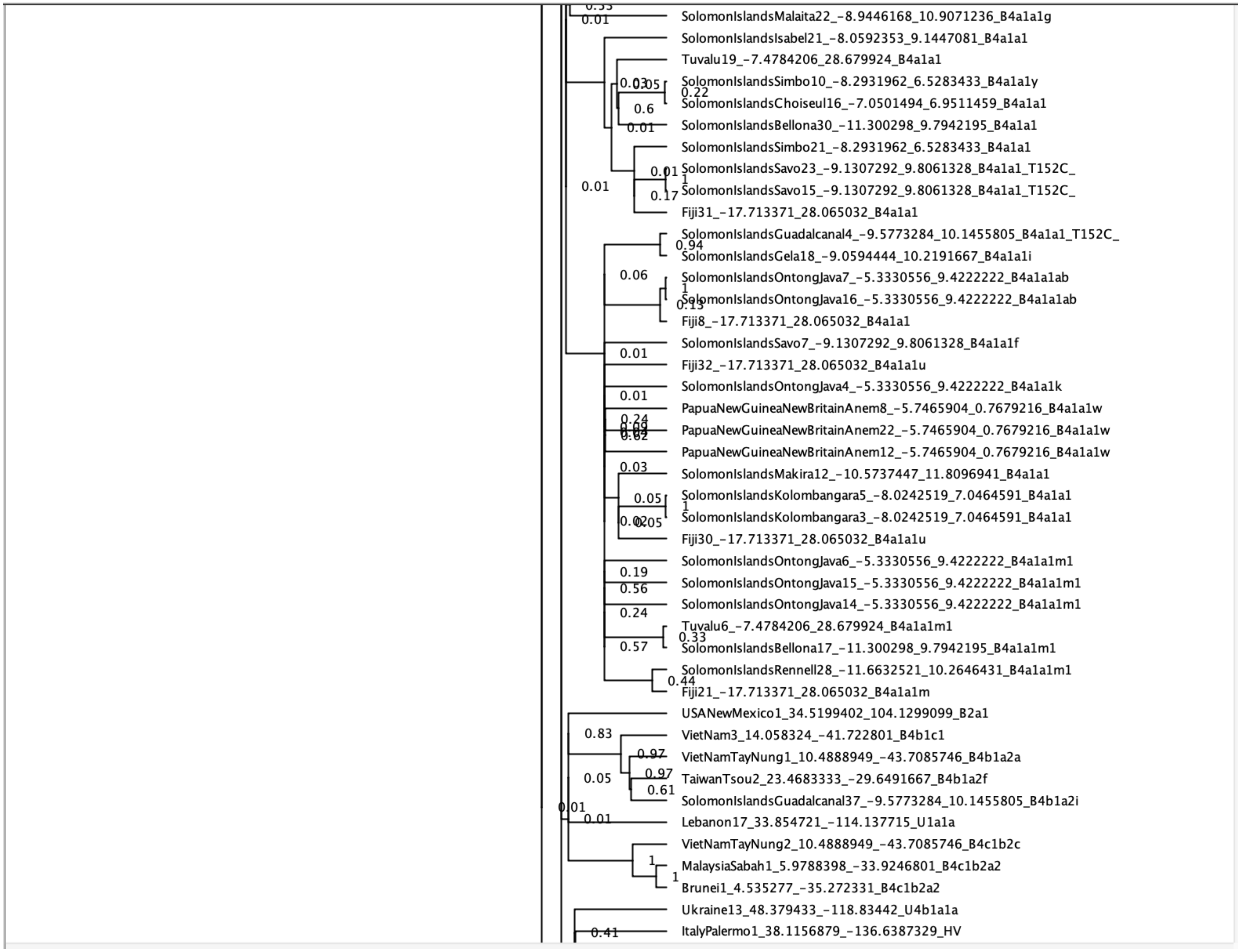
S3.29


0.01
SolomonIslandsMakira11_-10.5737447_11.8096941_B4a1a1a16
Tuvalu17_-7.4784206_28.679924_B4a1a1a16
0.79
0.04
SolomonIslandsChoiseul4_-7.0501494_6.9511459_B4a1a1a
SolomonIslandsRanongga18_-8.062_6.568_B4a1a1a10
0.02
PapuaNewGuineaNewBritainAnem4_-5.7465904_0.7679216_B4a1a1a19
0.02
SolomonIslandsMakira5_-10.5737447_11.8096941_B4a1a1a16
SolomonIslandsVellaLavella37_-7.7587665_6.6652785_B4a1a1a16
0.03
0.01
0.74
0.02
SolomonIslandsRanongga17_-8.062_6.568_B4a1a1a
0.01
SolomonIslandsRussell28_-9.0749167_9.1208657_B4a1a1a16
0.64
0.03
0.02
SolomonIslandsKolombangara11_-8.0242519_7.0464591_B4a1a1a2
0.02
PapuaNewGuineaNewBritainNakanai8_-5.7465904_0.7679216_B4a1a1a
0.77
PapuaNewGuineaNewBritainNakanai15_-5.7465904_0.7679216_B4a1a1a
Tuvalu11_-7.4784206_28.679924_B4a1a1a16
0.01
PapuaNewGuineaBougainvilleBuin4_-6.7465002_5.6851709_B4a1a1a16
SolomonIslandsRussell16_-9.0749167_9.1208657_B4a1a1a
0.04
SolomonIslandsRussell32_-9.0749167_9.1208657_B4a1a1a
0.03
SolomonIslandsRussell6_-9.0749167_9.1208657_B4a1a1a16
0.04
SolomonIslandsChoiseul8_-7.0501494_6.9511459_B4a1a1a16
0.06
PapuaNewGuineaBougainvilleBuka5_-5.2384267_4.6451287_B4a1a1a17
SolomonIslandsKolombangara7_-8.0242519_7.0464591_B4a1a1a2
0.87
SolomonIslandsKolombangara15_-8.0242519_7.0464591_B4a1a1a2
0.02
0.02
Tuvalu46_-7.4784206_28.679924_B4a1a1a9
0.17
0.02
0.01
SolomonIslandsVellaLavella38_-7.7587665_6.6652785_B4a1a1a2
0.02
SolomonIslandsKolombangara9_-8.0242519_7.0464591_B4a1a1a2
0.08
0.06
SolomonIslandsKolombangara10_-8.0242519_7.0464591_B4a1a1a2
SolomonIslandsMalaita57_-8.9446168_10.9071236_B4a1a1a2a
0.28
0.46
SolomonIslandsMalaita31_-8.9446168_10.9071236_B4a1a1a2a
0.03
SolomonIslandsMalaita56_-8.9446168_10.9071236_B4a1a1a2a
0.02
SolomonIslandsChoiseul1_-7.0501494_6.9511459_B4a1a1a2
PapuaNewGuineaBougainvilleBuin1_-6.7465002_5.6851709_B4a1a1a2b
Tuvalu33_-7.4784206_28.679924_B4a1a1a
0.39
SolomonIslandsGela5_-9.0594444_10.2191667_B4a1a1a12
SolomonIslandsBellona34_-11.300298_9.7942195_B4a1a1a11b
1
0.03
SolomonIslandsBellona22_-11.300298_9.7942195_B4a1a1a11b
0.03
SolomonIslandsSimbo11_-8.2931962_6.5283433_B4a1a1a16
0.69
SolomonIslandsOntongJava12_-5.3330556_9.4222222_B4a1a1a5
0.31
0.01
PapuaNewGuineaNewBritainNakanai16_-5.7465904_0.7679216_B4a1a1a16
0.01
SolomonIslandsMakira10_-10.5737447_11.8096941_B4a1a1a16
0.01
0.03
SolomonIslandsMakira3_-10.5737447_11.8096941_B4a1a1a16
0.04
Fiji5_-17.713371_28.065032_B4a1a1a
0.17
Tuvalu8_-7.4784206_28.679924_B4a1a1a20
0.08
SolomonIslandsIsabel2_-8.0592353_9.1447081_B4a1a1a16
SolomonIslandsSimbo6_-8.2931962_6.5283433_B4a1a1a
0.03

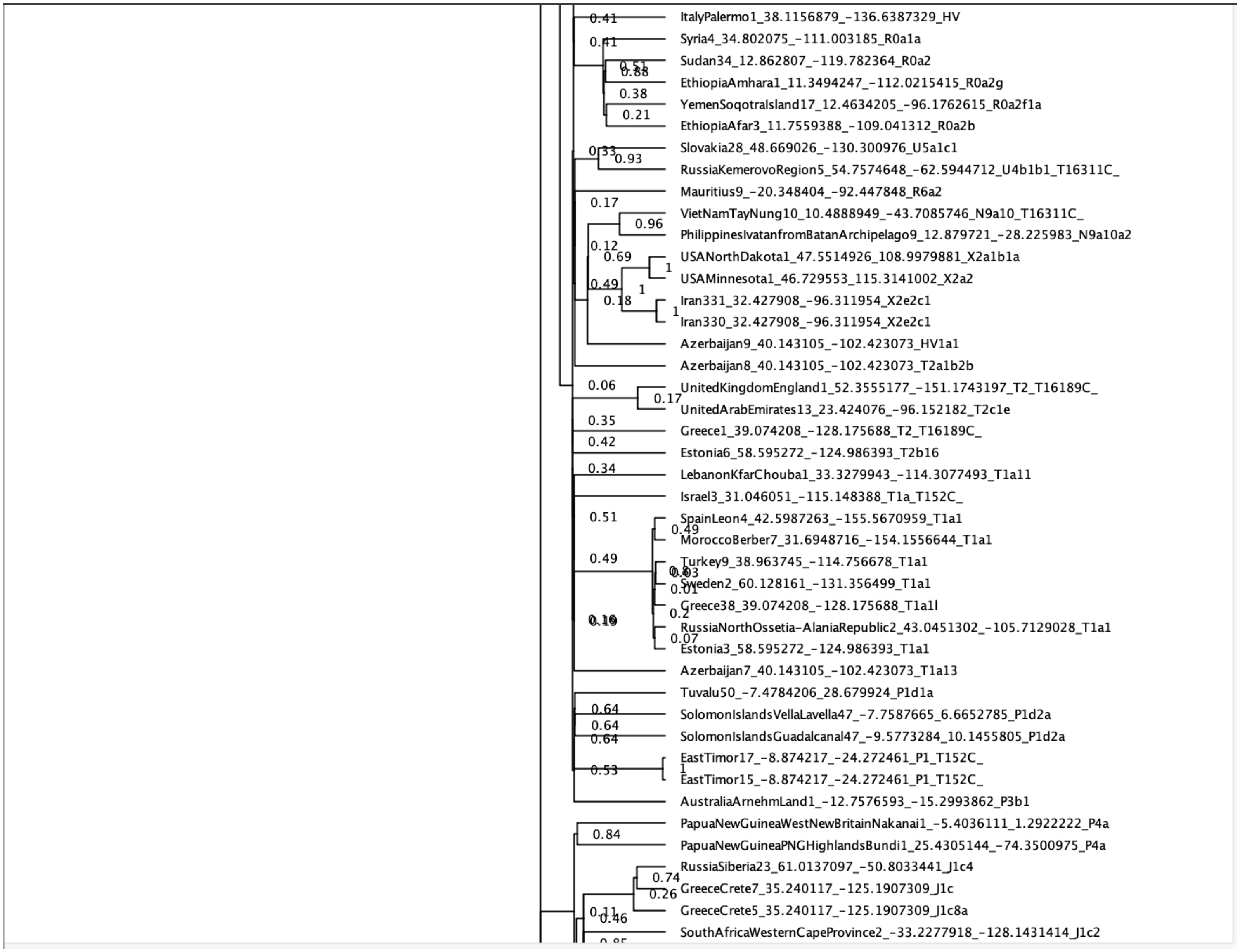
S3.30


0.03
SolomonIslandsSimbo6_-8.2931962_6.5283433_B4a1a1a
0.09
SolomonIslandsGuadalcanal29_-9.5773284_10.1455805_B4a1a1a16
0.38
PapuaNewGuineaBougainvilleBuka2_-5.2384267_4.6451287_B4a1a1a14
SolomonIslandsMakira8_-10.5737447_11.8096941_B4a1a1
0.06
0.13
SolomonIslandsRanongga29_-8.062_6.568_B4a1a1
0.11
SolomonIslandsChoiseul17_-7.0501494_6.9511459_B4a1a1
0.56
0.11
SolomonIslandsBellona9_-11.300298_9.7942195_B4a1a1
0.07
MicronesiaKapingamarangiAtoll1_1.0666667_4.7666667_B4a1a1x
Fiji26_-17.713371_28.065032_B4a1a1a
Tuvalu24_-7.4784206_28.679924_B4a1a1ac
0.05
SolomonIslandsIsabel5_-8.0592353_9.1447081_B4a1a1
0.02
SolomonIslandsIsabel10_-8.0592353_9.1447081_B4a1a1d
SolomonIslandsShortlands3_-7.0452221_5.7371749_B4a1a1s
0.01
SolomonIslandsShortlands8_-7.0452221_5.7371749_B4a1a1a23
0.35
SolomonIslandsSavo29_-9.1307292_9.8061328_B4a1a1
SolomonIslandsSavo17_-9.1307292_9.8061328_B4a1a1
0.02
0.04
0.04
SolomonIslandsSantaCruz3_-10.7245728_15.9230976_B4a1a1n
0.04
SolomonIslandsVellaLavella7_-7.7587665_6.6652785_B4a1a1y
1
SolomonIslandsVellaLavella27_-7.7587665_6.6652785_B4a1a1y
0.05
SolomonIslandsSavo19_-9.1307292_9.8061328_B4a1a1_T152C_
0.29
Vanuatu1_-15.376706_16.959158_B4a1a1
0.02
0.33
0.57
SolomonIslandsMalaita35_-8.9446168_10.9071236_B4a1a1g
SolomonIslandsIsabel23_-8.0592353_9.1447081_B4a1a1
SolomonIslandsVellaLavella25_-7.7587665_6.6652785_B4a1a1
0.15
PapuaNewGuineaNewBritainNakanai7_-5.7465904_0.7679216_B4a1a1
0.03
SolomonIslandsRussell8_-9.0749167_9.1208657_B4a1a1
SolomonIslandsRennell39_-11.6632521_10.2646431_B4a1a1h
0.46
0.17
SolomonIslandsChoiseul14_-7.0501494_6.9511459_B4a1a1
0.02
SolomonIslandsRussell15_-9.0749167_9.1208657_B4a1a1
SolomonIslandsSimbo1_-8.2931962_6.5283433_B4a1a1e
0.01
SolomonIslandsSimbo18_-8.2931962_6.5283433_B4a1a1
0.02
SolomonIslandsRennell5_-11.6632521_10.2646431_B4a1a1m1
0.21
0.02
0.86
SolomonIslandsRennell42_-11.6632521_10.2646431_B4a1a1m1
0.52
SolomonIslandsBellona38_-11.300298_9.7942195_B4a1a1m1
0.86
0.18
SolomonIslandsBellona31_-11.300298_9.7942195_B4a1a1m1
SolomonIslandsGela25_-9.0594444_10.2191667_B4a1a1
0.61
Fiji33_-17.713371_28.065032_B4a1a1
SolomonIslandsShortlands7_-7.0452221_5.7371749_B4a1a1
0.04
SolomonIslandsRussell2_-9.0749167_9.1208657_B4a1a1
0.01
SolomonIslandsRussell24_-9.0749167_9.1208657_B4a1a1
0.15
SolomonIslandsShortlands1_-7.0452221_5.7371749_B4a1a1
0.03
SolomonIslandsRussell21_-9.0749167_9.1208657_B4a1a1g
0.53
SolomonIslandsMalaita22_-8.9446168_10.9071236_B4a1a1g
0.01

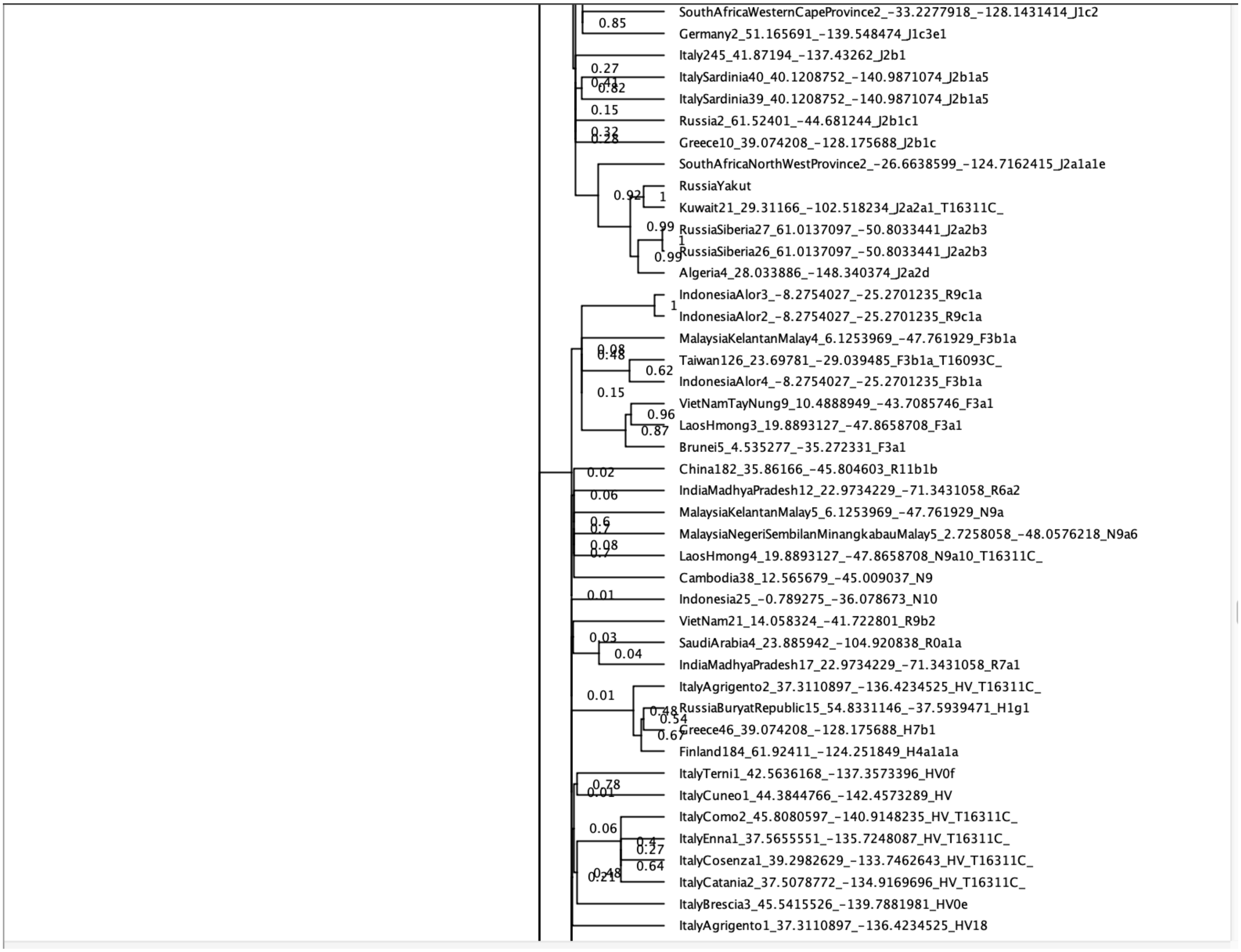
S3.31


0.33
0.01
SolomonIslandsMalaita22_-8.9446168_10.9071236_B4a1a1g
SolomonIslandsIsabel21_-8.0592353_9.1447081_B4a1a1
Tuvalu19_-7.4784206_28.679924_B4a1a1
0.03
0.05
SolomonIslandsSimbo10_-8.2931962_6.5283433_B4a1a1y
0.22
SolomonIslandsChoiseul16_-7.0501494_6.9511459_B4a1a1
0.6
0.01
SolomonIslandsBellona30_-11.300298_9.7942195_B4a1a1
SolomonIslandsSimbo21_-8.2931962_6.5283433_B4a1a1
0.01
1
SolomonIslandsSavo23_-9.1307292_9.8061328_B4a1a1_T152C_
0.01
0.17
SolomonIslandsSavo15_-9.1307292_9.8061328_B4a1a1_T152C_
Fiji31_-17.713371_28.065032_B4a1a1
SolomonIslandsGuadalcanal4_-9.5773284_10.1455805_B4a1a1_T152C_
0.94
SolomonIslandsGela18_-9.0594444_10.2191667_B4a1a1i
0.06
SolomonIslandsOntongJava7_-5.3330556_9.4222222_B4a1a1ab
1
SolomonIslandsOntongJava16_-5.3330556_9.4222222_B4a1a1ab
0.13
Fiji8_-17.713371_28.065032_B4a1a1
SolomonIslandsSavo7_-9.1307292_9.8061328_B4a1a1f
0.01
Fiji32_-17.713371_28.065032_B4a1a1u
SolomonIslandsOntongJava4_-5.3330556_9.4222222_B4a1a1k
0.01
PapuaNewGuineaNewBritainAnem8_-5.7465904_0.7679216_B4a1a1w
0.24
0.09
0.02
PapuaNewGuineaNewBritainAnem22_-5.7465904_0.7679216_B4a1a1w
PapuaNewGuineaNewBritainAnem12_-5.7465904_0.7679216_B4a1a1w
0.03
SolomonIslandsMakira12_-10.5737447_11.8096941_B4a1a1
0.05
SolomonIslandsKolombangara5_-8.0242519_7.0464591_B4a1a1
1
SolomonIslandsKolombangara3_-8.0242519_7.0464591_B4a1a1
0.02
0.05
Fiji30_-17.713371_28.065032_B4a1a1u
SolomonIslandsOntongJava6_-5.3330556_9.4222222_B4a1a1m1
0.19
SolomonIslandsOntongJava15_-5.3330556_9.4222222_B4a1a1m1
0.56
SolomonIslandsOntongJava14_-5.3330556_9.4222222_B4a1a1m1
0.24
Tuvalu6_-7.4784206_28.679924_B4a1a1m1
0.33
0.57
SolomonIslandsBellona17_-11.300298_9.7942195_B4a1a1m1
SolomonIslandsRennell28_-11.6632521_10.2646431_B4a1a1m1
0.44
Fiji21_-17.713371_28.065032_B4a1a1m
USANewMexico1_34.5199402_104.1299099_B2a1
0.83
VietNam3_14.058324_-41.722801_B4b1c1
0.97
VietNamTayNung1_10.4888949_-43.7085746_B4b1a2a
0.05
0.97
TaiwanTsou2_23.4683333_-29.6491667_B4b1a2f
0.61
SolomonIslandsGuadalcanal37_-9.5773284_10.1455805_B4b1a2i
0.01
0.01
Lebanon17_33.854721_-114.137715_U1a1a
VietNamTayNung2_10.4888949_-43.7085746_B4c1b2c
1
MalaysiaSabah1_5.9788398_-33.9246801_B4c1b2a2
1
Brunei1_4.535277_-35.272331_B4c1b2a2
Ukraine13_48.379433_-118.83442_U4b1a1a
0.41
ItalyPalermo1_38.1156879_-136.6387329_HV

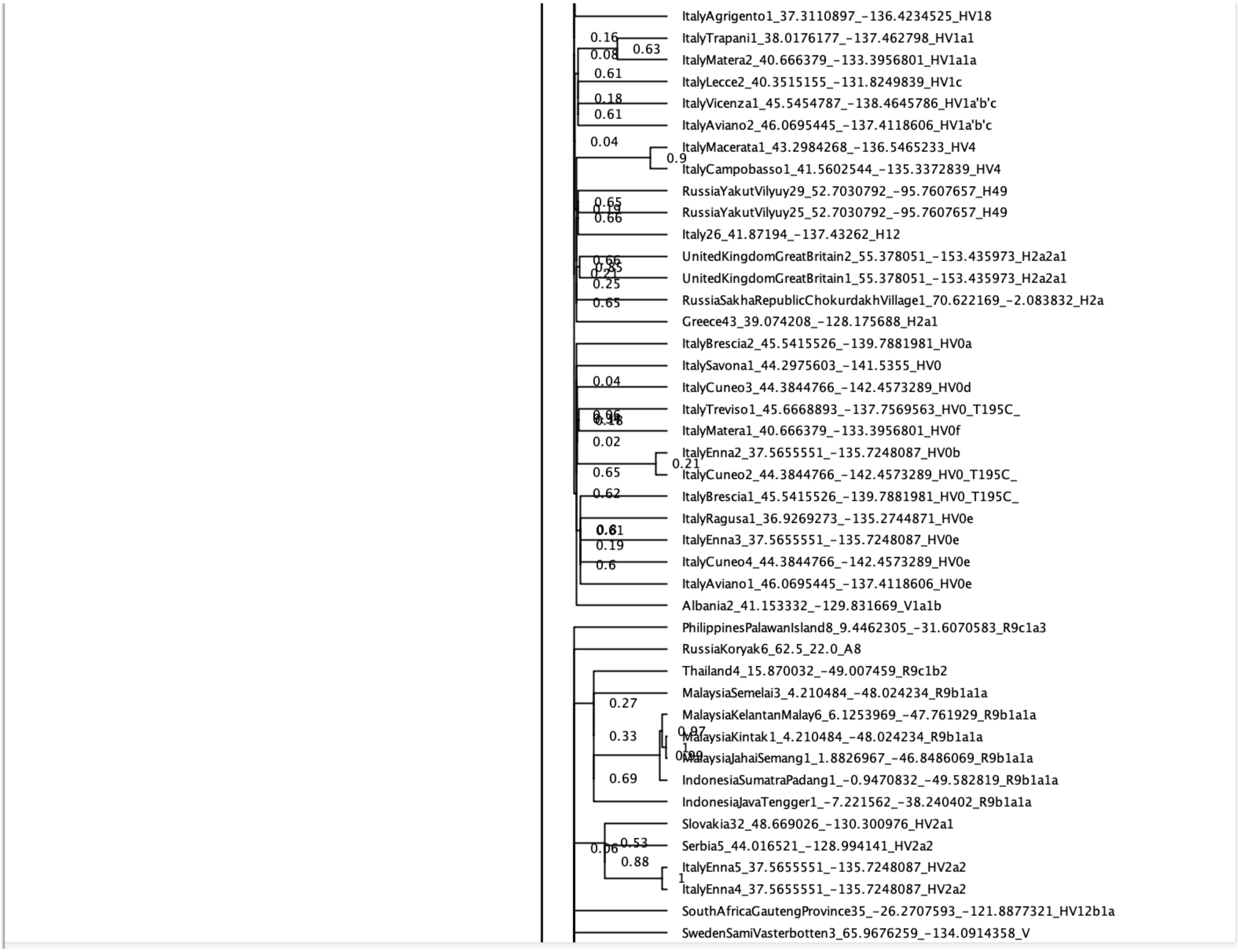
S3.32


ItalyPalermo1_38.1156879_-136.6387329_HV
Syria4_34.802075_-111.003185_R0a1a
Sudan34_12.862807_-119.782364_R0a2
EthiopiaAmhara1_11.3494247_-112.0215415_R0a2g
YemenSoqotraIsland17_12.4634205_-96.1762615_R0a2f1a
EthiopiaAfar3_11.7559388_-109.041312_R0a2b
Slovakia28_48.669026_-130.300976_U5a1c1
RussiaKemerovoRegion5_54.7574648_-62.5944712_U4b1b1_T16311C_
Mauritius9_-20.348404_-92.447848_R6a2
VietNamTayNung10_10.4888949_-43.7085746_N9a10_T16311C_
PhilippinesIvatanfromBatanArchipelago9_12.879721_-28.225983_N9a10a2
USANorthDakota1_47.5514926_108.9979881_X2a1b1a
USAMinnesota1_46.729553_115.3141002_X2a2
Iran331_32.427908_-96.311954_X2e2c1
Iran330_32.427908_-96.311954_X2e2c1
Azerbaijan9_40.143105_-102.423073_HV1a1
Azerbaijan8_40.143105_-102.423073_T2a1b2b
UnitedKingdomEngland1_52.3555177_-151.1743197_T2_T16189C_
UnitedArabEmirates13_23.424076_-96.152182_T2c1e
Greece1_39.074208_-128.175688_T2_T16189C_
Estonia6_58.595272_-124.986393_T2b16
LebanonKfarChouba1_33.3279943_-114.3077493_T1a11
Israel3_31.046051_-115.148388_T1a_T152C_
SpainLeon4_42.5987263_-155.5670959_T1a1
MoroccoBerber7_31.6948716_-154.1556644_T1a1
Turkey9_38.963745_-114.756678_T1a1
Sweden2_60.128161_-131.356499_T1a1
Greece38_39.074208_-128.175688_T1a1l
RussiaNorthOssetia-AlaniaRepublic2_43.0451302_-105.7129028_T1a1
Estonia3_58.595272_-124.986393_T1a1
Azerbaijan7_40.143105_-102.423073_T1a13
Tuvalu50_-7.4784206_28.679924_P1d1a
SolomonIslandsVellaLavella47_-7.7587665_6.6652785_P1d2a
SolomonIslandsGuadalcanal47_-9.5773284_10.1455805_P1d2a
EastTimor17_-8.874217_-24.272461_P1_T152C_
EastTimor15_-8.874217_-24.272461_P1_T152C_
AustraliaArnehmLand1_-12.7576593_-15.2993862_P3b1
PapuaNewGuineaWestNewBritainNakanai1_-5.4036111_1.2922222_P4a
PapuaNewGuineaPNGHighlandsBundi1_25.4305144_-74.3500975_P4a
RussiaSiberia23_61.0137097_-50.8033441_J1c4
GreeceCrete7_35.240117_-125.1907309_J1c
GreeceCrete5_35.240117_-125.1907309_J1c8a
SouthAfricaWesternCapeProvince2_-33.2277918_-128.1431414_J1c2
0.41
0.41
0.51
0.88
0.38
0.21
0.33
0.93
0.17
0.96
0.12
0.69
1
0.49
1
0.18
1
0.06
0.17
0.35
0.42
0.34
0.51
0.49
0.49
0.03
0.01
0.10
0.2
0.07
0.64
0.64
0.64
1
0.53
0.84
0.74
0.26
0.11
0.46
0.85

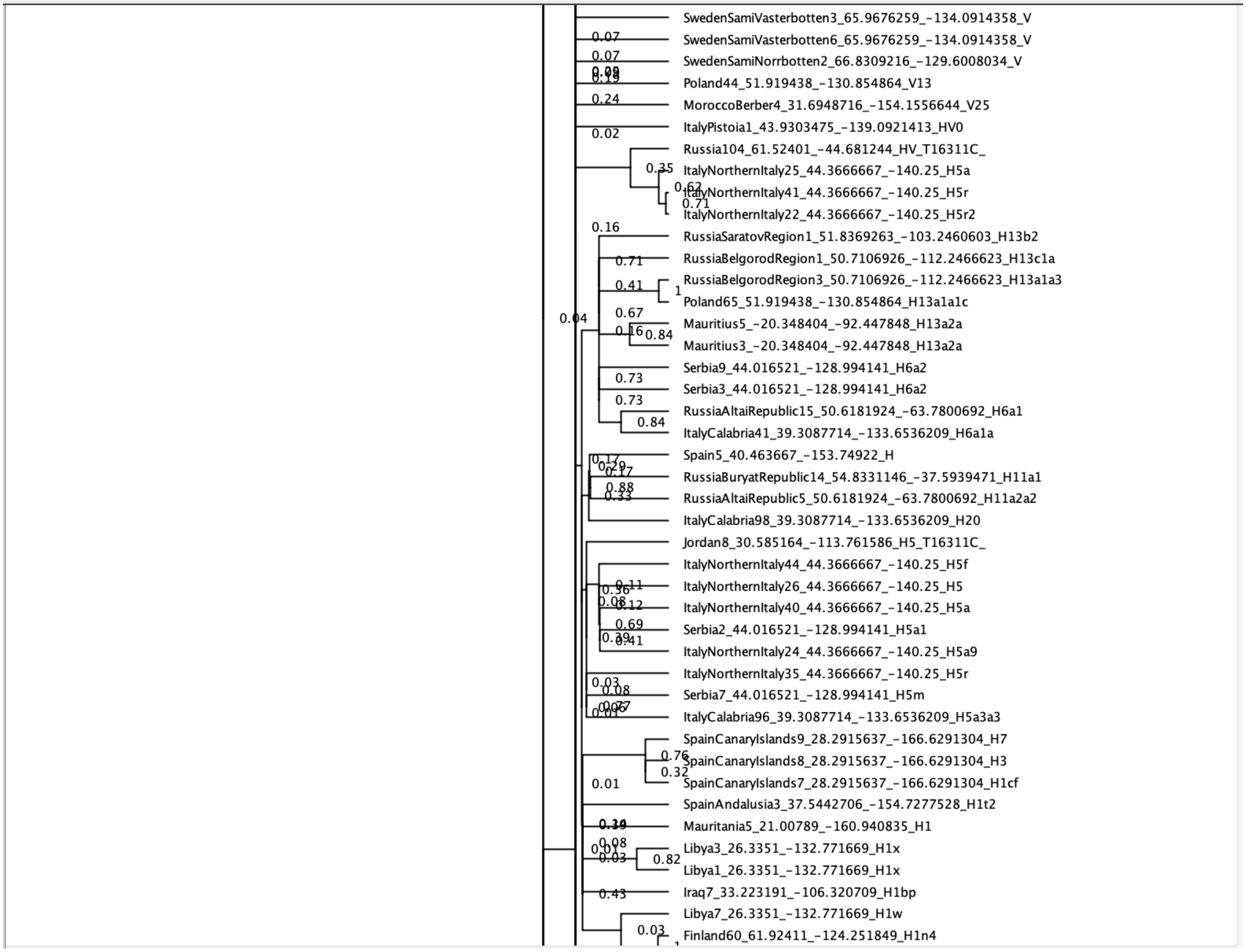
S3.33


SouthAfricaWesternCapeProvince2_-33.2277918_-128.1431414_J1c2
0.85
Germany2_51.165691_-139.548474_J1c3e1
Italy245_41.87194_-137.43262_J2b1
0.27
0.41
0.82
ItalySardinia40_40.1208752_-140.9871074_J2b1a5
ItalySardinia39_40.1208752_-140.9871074_J2b1a5
0.15
Russia2_61.52401_-44.681244_J2b1c1
0.32
0.28
Greece10_39.074208_-128.175688_J2b1c
SouthAfricaNorthWestProvince2_-26.6638599_-124.7162415_J2a1a1e
RussiaYakut
0.92
1
Kuwait21_29.31166_-102.518234_J2a2a1_T16311C_
0.99
RussiaSiberia27_61.0137097_-50.8033441_J2a2b3
1
0.99
RussiaSiberia26_61.0137097_-50.8033441_J2a2b3
Algeria4_28.033886_-148.340374_J2a2d
IndonesiaAlor3_-8.2754027_-25.2701235_R9c1a
1
IndonesiaAlor2_-8.2754027_-25.2701235_R9c1a
MalaysiaKelantanMalay4_6.1253969_-47.761929_F3b1a
0.08
0.48
Taiwan126_23.69781_-29.039485_F3b1a_T16093C_
0.62
IndonesiaAlor4_-8.2754027_-25.2701235_F3b1a
0.15
VietNamTayNung9_10.4888949_-43.7085746_F3a1
0.96
LaosHmong3_19.8893127_-47.8658708_F3a1
0.87
Brunei5_4.535277_-35.272331_F3a1
0.02
China182_35.86166_-45.804603_R11b1b
0.06
IndiaMadhyaPradesh12_22.9734229_-71.3431058_R6a2
MalaysiaKelantanMalay5_6.1253969_-47.761929_N9a
0.6
0.7
MalaysiaNegeriSembilanMinangkabauMalay5_2.7258058_-48.0576218_N9a6
0.08
0.7
LaosHmong4_19.8893127_-47.8658708_N9a10_T16311C_
Cambodia38_12.565679_-45.009037_N9
0.01
Indonesia25_-0.789275_-36.078673_N10
VietNam21_14.058324_-41.722801_R9b2
0.03
SaudiArabia4_23.885942_-104.920838_R0a1a
0.04
IndiaMadhyaPradesh17_22.9734229_-71.3431058_R7a1
ItalyAgrigento2_37.3110897_-136.4234525_HV_T16311C_
0.01
0.48
RussiaBuryatRepublic15_54.8331146_-37.5939471_H1g1
0.54
Greece46_39.074208_-128.175688_H7b1
0.67
Finland184_61.92411_-124.251849_H4a1a1a
ItalyTerni1_42.5636168_-137.3573396_HV0f
0.78
0.01
ItalyCuneo1_44.3844766_-142.4573289_HV
ItalyComo2_45.8080597_-140.9148235_HV_T16311C_
0.06
ItalyEnna1_37.5655551_-135.7248087_HV_T16311C_
0.4
0.27
ItalyCosenza1_39.2982629_-133.7462643_HV_T16311C_
0.64
0.48
0.21
ItalyCatania2_37.5078772_-134.9169696_HV_T16311C_
ItalyBrescia3_45.5415526_-139.7881981_HV0e
ItalyAgrigento1_37.3110897_-136.4234525_HV18

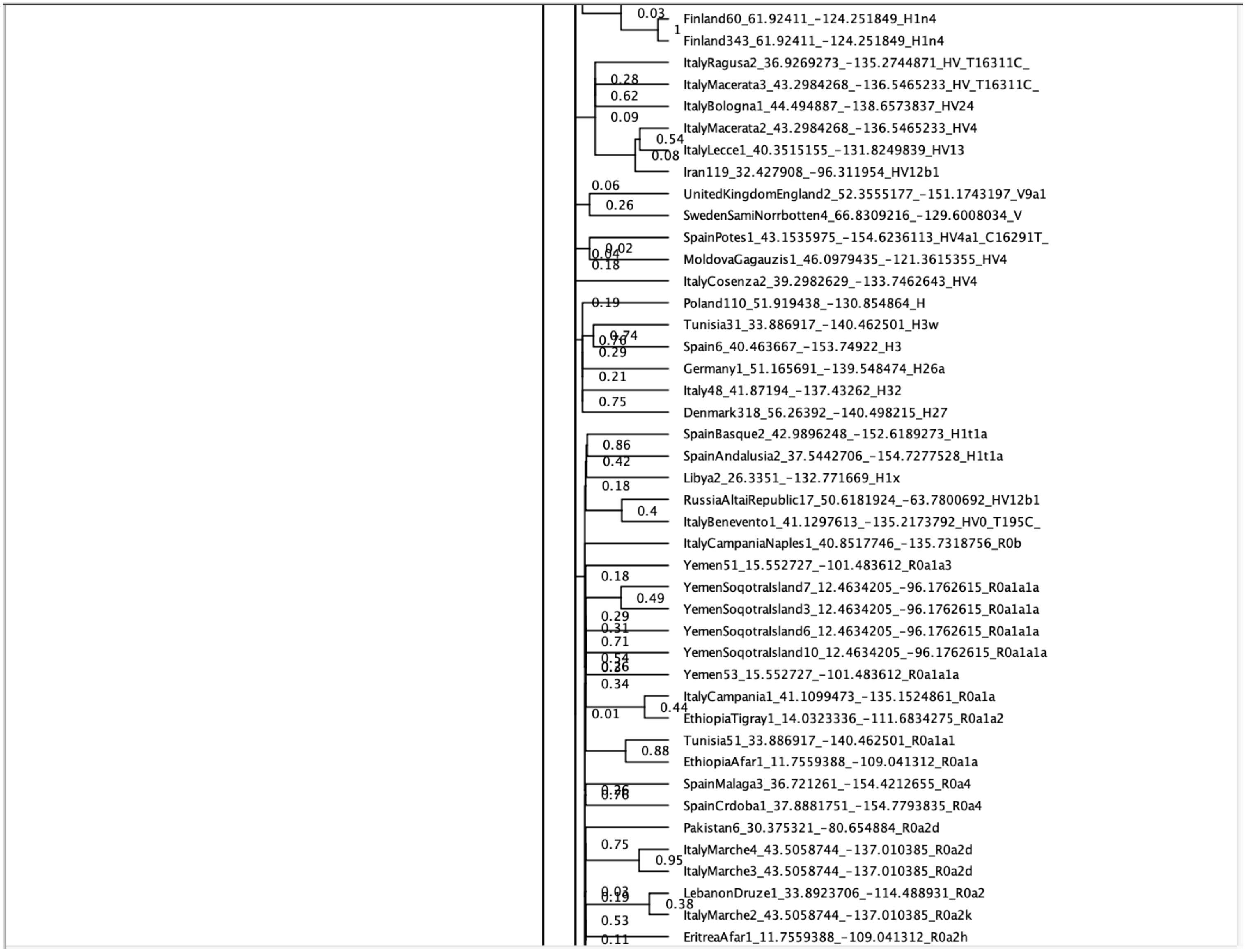
S3.34


ItalyAgrigento1_37.3110897_-136.4234525_HV18
0.16
ItalyTrapani1_38.0176177_-137.462798_HV1a1
0.63
0.08
ItalyMatera2_40.666379_-133.3956801_HV1a1a
0.61
ItalyLecce2_40.3515155_-131.8249839_HV1c
0.18
ItalyVicenza1_45.5454787_-138.4645786_HV1a'b'c
0.61
ItalyAviano2_46.0695445_-137.4118606_HV1a'b'c
0.04
ItalyMacerata1_43.2984268_-136.5465233_HV4
0.9
ItalyCampobasso1_41.5602544_-135.3372839_HV4
RussiaYakutVilyuy29_52.7030792_-95.7607657_H49
0.65
0.19
RussiaYakutVilyuy25_52.7030792_-95.7607657_H49
0.66
Italy26_41.87194_-137.43262_H12
0.66
0.35
UnitedKingdomGreatBritain2_55.378051_-153.435973_H2a2a1
0.21
UnitedKingdomGreatBritain1_55.378051_-153.435973_H2a2a1
0.25
RussiaSakhaRepublicChokurdakhVillage1_70.622169_-2.083832_H2a
0.65
Greece43_39.074208_-128.175688_H2a1
ItalyBrescia2_45.5415526_-139.7881981_HV0a
ItalySavona1_44.2975603_-141.5355_HV0
0.04
ItalyCuneo3_44.3844766_-142.4573289_HV0d
ItalyTreviso1_45.6668893_-137.7569563_HV0_T195C_
0.06
0.18
ItalyMatera1_40.666379_-133.3956801_HV0f
0.02
ItalyEnna2_37.5655551_-135.7248087_HV0b
0.21
0.65
ItalyCuneo2_44.3844766_-142.4573289_HV0_T195C_
0.62
ItalyBrescia1_45.5415526_-139.7881981_HV0_T195C_
ItalyRagusa1_36.9269273_-135.2744871_HV0e
0.61
0.6
0.19
ItalyEnna3_37.5655551_-135.7248087_HV0e
0.6
ItalyCuneo4_44.3844766_-142.4573289_HV0e
ItalyAviano1_46.0695445_-137.4118606_HV0e
Albania2_41.153332_-129.831669_V1a1b
PhilippinesPalawanIsland8_9.4462305_-31.6070583_R9c1a3
RussiaKoryak6_62.5_22.0_A8
Thailand4_15.870032_-49.007459_R9c1b2
MalaysiaSemelai3_4.210484_-48.024234_R9b1a1a
0.27
MalaysiaKelantanMalay6_6.1253969_-47.761929_R9b1a1a
0.97
0.33
MalaysiaKintak1_4.210484_-48.024234_R9b1a1a
1
0.99
MalaysiaJahaiSemang1_1.8826967_-46.8486069_R9b1a1a
0.69
IndonesiaSumatraPadang1_-0.9470832_-49.582819_R9b1a1a
IndonesiaJavaTengger1_-7.221562_-38.240402_R9b1a1a
Slovakia32_48.669026_-130.300976_HV2a1
0.53
0.06
Serbia5_44.016521_-128.994141_HV2a2
0.88
ItalyEnna5_37.5655551_-135.7248087_HV2a2
1
ItalyEnna4_37.5655551_-135.7248087_HV2a2
SouthAfricaGautengProvince35_-26.2707593_-121.8877321_HV12b1a
SwedenSamiVasterbotten3_65.9676259_-134.0914358_V

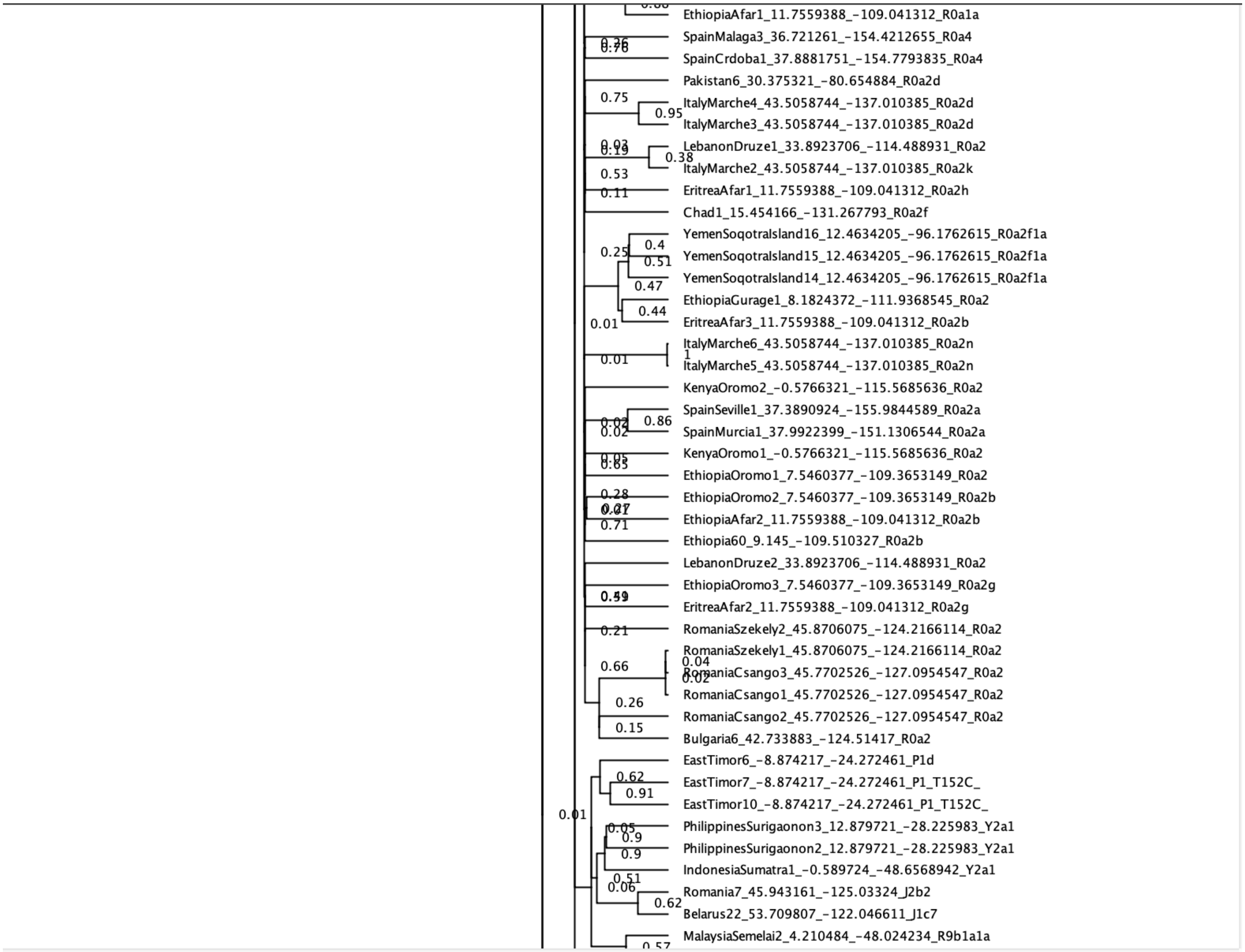
S3.35


SwedenSamiVasterbotten3_65.9676259_-134.0914358_V
0.07
SwedenSamiVasterbotten6_65.9676259_-134.0914358_V
0.07
SwedenSamiNorrbotten2_66.8309216_-129.6008034_V
Poland44_51.919438_-130.854864_V13
0.24
MoroccoBerber4_31.6948716_-154.1556644_V25
ItalyPistoia1_43.9303475_-139.0921413_HV0
0.02
Russia104_61.52401_-44.681244_HV_T16311C_
0.35
ItalyNorthernItaly25_44.3666667_-140.25_H5a
0.62
ItalyNorthernItaly41_44.3666667_-140.25_H5r
0.71
ItalyNorthernItaly22_44.3666667_-140.25_H5r2
0.16
RussiaSaratovRegion1_51.8369263_-103.2460603_H13b2
0.71
RussiaBelgorodRegion1_50.7106926_-112.2466623_H13c1a
0.41
1
RussiaBelgorodRegion3_50.7106926_-112.2466623_H13a1a3
Poland65_51.919438_-130.854864_H13a1a1c
0.04
0.67
Mauritius5_-20.348404_-92.447848_H13a2a
0.16
0.84
Mauritius3_-20.348404_-92.447848_H13a2a
Serbia9_44.016521_-128.994141_H6a2
0.73
Serbia3_44.016521_-128.994141_H6a2
0.73
RussiaAltaiRepublic15_50.6181924_-63.7800692_H6a1
0.84
ItalyCalabria41_39.3087714_-133.6536209_H6a1a
Spain5_40.463667_-153.74922_H
RussiaBuryatRepublic14_54.8331146_-37.5939471_H11a1
0.88
RussiaAltaiRepublic5_50.6181924_-63.7800692_H11a2a2
ItalyCalabria98_39.3087714_-133.6536209_H20
Jordan8_30.585164_-113.761586_H5_T16311C_
ItalyNorthernItaly44_44.3666667_-140.25_H5f
ItalyNorthernItaly26_44.3666667_-140.25_H5
ItalyNorthernItaly40_44.3666667_-140.25_H5a
0.69
Serbia2_44.016521_-128.994141_H5a1
ItalyNorthernItaly24_44.3666667_-140.25_H5a9
ItalyNorthernItaly35_44.3666667_-140.25_H5r
Serbia7_44.016521_-128.994141_H5m
ItalyCalabria96_39.3087714_-133.6536209_H5a3a3
SpainCanaryIslands9_28.2915637_-166.6291304_H7
0.76
SpainCanaryIslands8_28.2915637_-166.6291304_H3
0.32
0.01
SpainCanaryIslands7_28.2915637_-166.6291304_H1cf
SpainAndalusia3_37.5442706_-154.7277528_H1t2
Mauritania5_21.00789_-160.940835_H1
Libya3_26.3351_-132.771669_H1x
0.82
Libya1_26.3351_-132.771669_H1x
0.43
Iraq7_33.223191_-106.320709_H1bp
Libya7_26.3351_-132.771669_H1w
0.03
Finland60_61.92411_-124.251849_H1n4

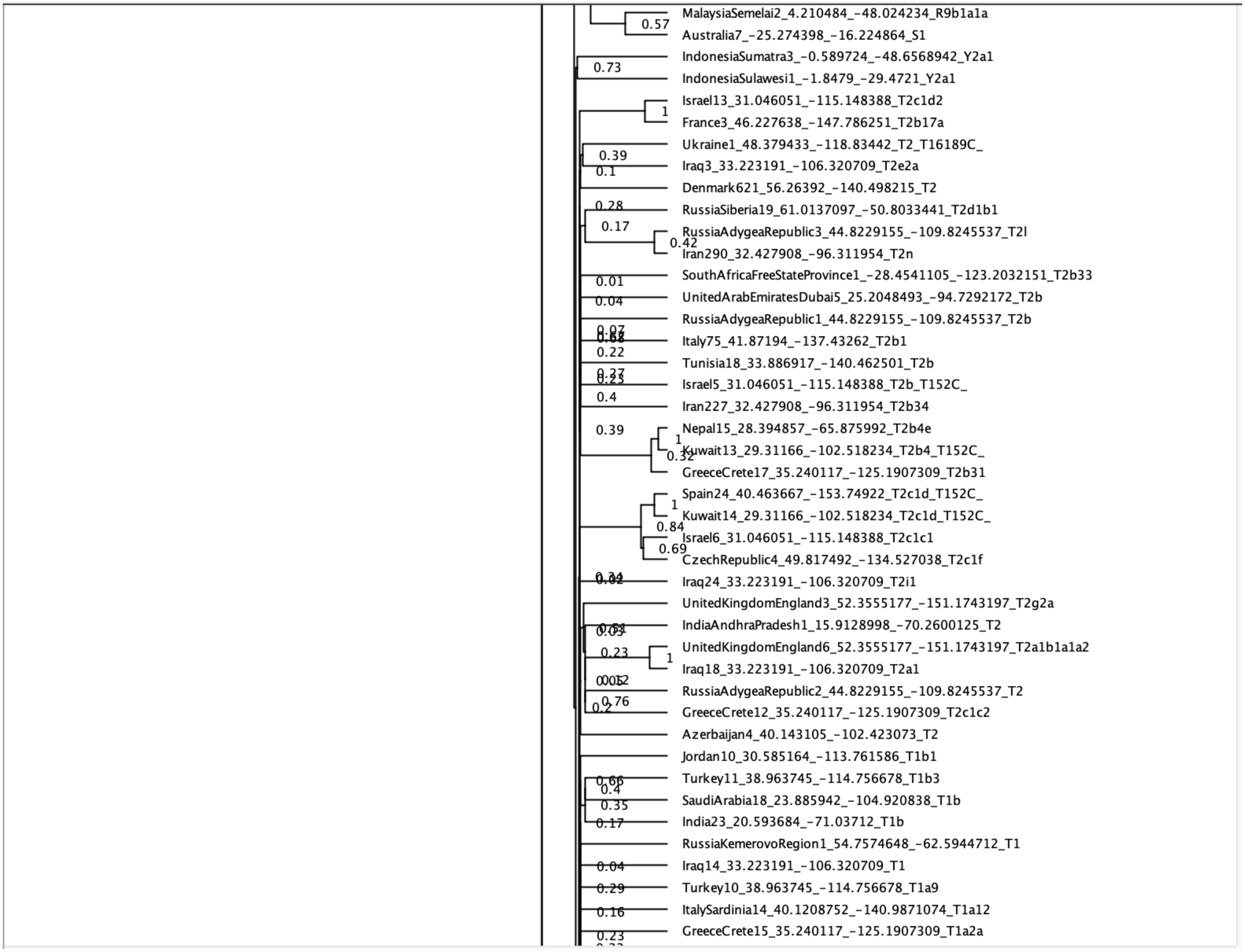
S3.36


0.03
1
Finland60_61.92411_-124.251849_H1n4
Finland343_61.92411_-124.251849_H1n4
ItalyRagusa2_36.9269273_-135.2744871_HV_T16311C_
0.28
ItalyMacerata3_43.2984268_-136.5465233_HV_T16311C_
0.62
ItalyBologna1_44.494887_-138.6573837_HV24
0.09
ItalyMacerata2_43.2984268_-136.5465233_HV4
0.54
ItalyLecce1_40.3515155_-131.8249839_HV13
0.08
Iran119_32.427908_-96.311954_HV12b1
0.06
UnitedKingdomEngland2_52.3555177_-151.1743197_V9a1
0.26
SwedenSamiNorrbotten4_66.8309216_-129.6008034_V
SpainPotes1_43.1535975_-154.6236113_HV4a1_C16291T_
0.02
0.04
MoldovaGagauzis1_46.0979435_-121.3615355_HV4
0.18
ItalyCosenza2_39.2982629_-133.7462643_HV4
0.19
Poland110_51.919438_-130.854864_H
Tunisia31_33.886917_-140.462501_H3w
0.74
0.76
Spain6_40.463667_-153.74922_H3
0.29
Germany1_51.165691_-139.548474_H26a
0.21
Italy48_41.87194_-137.43262_H32
0.75
Denmark318_56.26392_-140.498215_H27
SpainBasque2_42.9896248_-152.6189273_H1t1a
0.86
SpainAndalusia2_37.5442706_-154.7277528_H1t1a
0.42
Libya2_26.3351_-132.771669_H1x
0.18
RussiaAltaiRepublic17_50.6181924_-63.7800692_HV12b1
0.4
ItalyBenevento1_41.1297613_-135.2173792_HV0_T195C_
ItalyCampaniaNaples1_40.8517746_-135.7318756_R0b
Yemen51_15.552727_-101.483612_R0a1a3
0.18
YemenSoqotraIsland7_12.4634205_-96.1762615_R0a1a1a
0.49
YemenSoqotraIsland3_12.4634205_-96.1762615_R0a1a1a
0.29
0.31
YemenSoqotraIsland6_12.4634205_-96.1762615_R0a1a1a
0.71
YemenSoqotraIsland10_12.4634205_-96.1762615_R0a1a1a
0.54
0.26
Yemen53_15.552727_-101.483612_R0a1a1a
0.34
ItalyCampania1_41.1099473_-135.1524861_R0a1a
0.44
0.01
EthiopiaTigray1_14.0323336_-111.6834275_R0a1a2
Tunisia51_33.886917_-140.462501_R0a1a1
0.88
EthiopiaAfar1_11.7559388_-109.041312_R0a1a
SpainMalaga3_36.721261_-154.4212655_R0a4
0.26
0.76
SpainCrdoba1_37.8881751_-154.7793835_R0a4
Pakistan6_30.375321_-80.654884_R0a2d
0.75
ItalyMarche4_43.5058744_-137.010385_R0a2d
0.95
ItalyMarche3_43.5058744_-137.010385_R0a2d
0.03
0.19
LebanonDruze1_33.8923706_-114.488931_R0a2
0.38
ItalyMarche2_43.5058744_-137.010385_R0a2k
0.53
EritreaAfar1_11.7559388_-109.041312_R0a2h
0.11

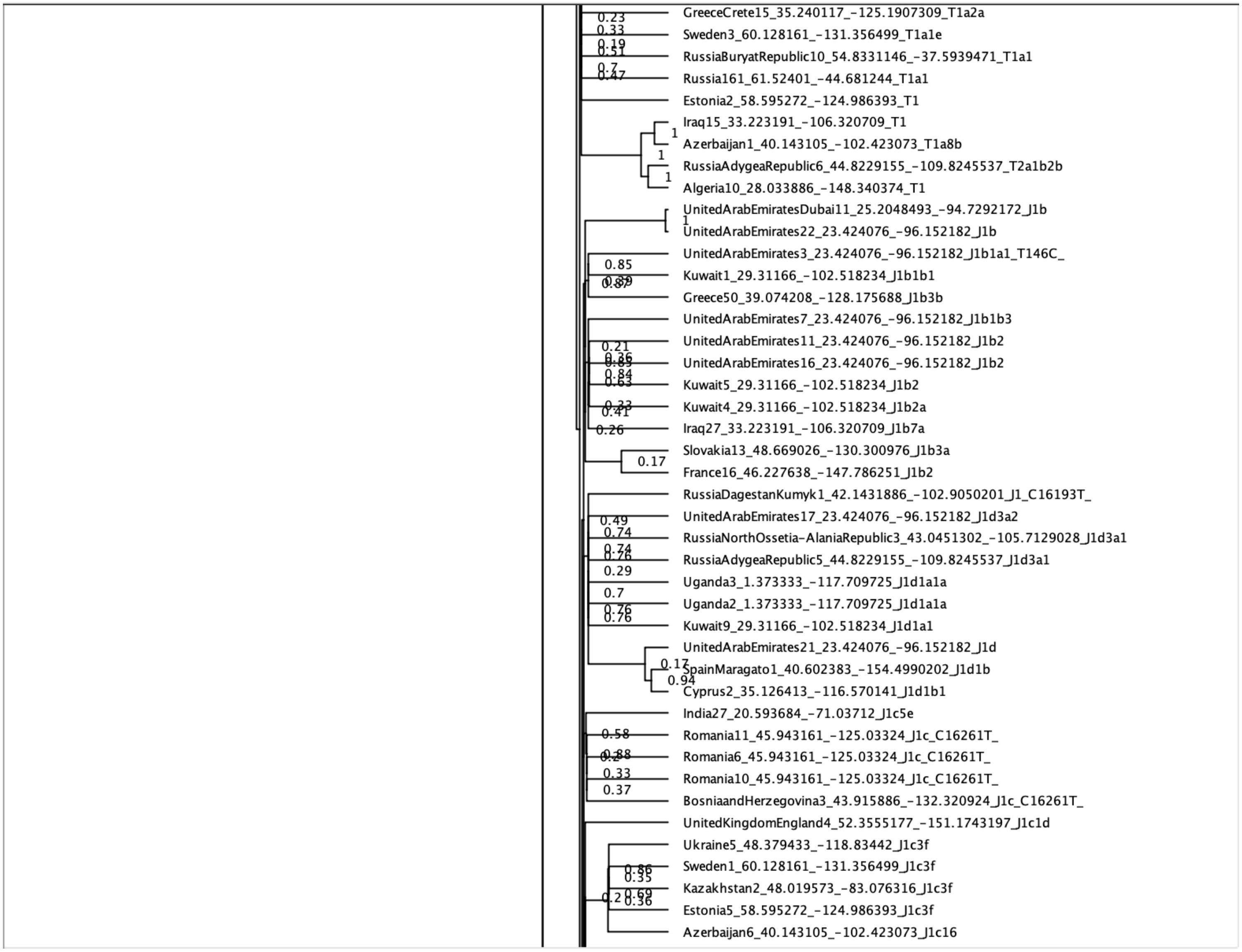
S3.37


EthiopiaAfar1_11.7559388_-109.041312_R0a1a
0.36
0.78
SpainMalaga3_36.721261_-154.4212655_R0a4
SpainCrdoba1_37.8881751_-154.7793835_R0a4
Pakistan6_30.375321_-80.654884_R0a2d
0.75
ItalyMarche4_43.5058744_-137.010385_R0a2d
0.95
ItalyMarche3_43.5058744_-137.010385_R0a2d
0.03
0.19
LebanonDruze1_33.8923706_-114.488931_R0a2
0.38
ItalyMarche2_43.5058744_-137.010385_R0a2k
0.53
0.11
EritreaAfar1_11.7559388_-109.041312_R0a2h
Chad1_15.454166_-131.267793_R0a2f
YemenSoqotraIsland16_12.4634205_-96.1762615_R0a2f1a
0.4
0.25
0.51
YemenSoqotraIsland15_12.4634205_-96.1762615_R0a2f1a
YemenSoqotraIsland14_12.4634205_-96.1762615_R0a2f1a
0.47
EthiopiaGurage1_8.1824372_-111.9368545_R0a2
0.44
0.01
EritreaAfar3_11.7559388_-109.041312_R0a2b
ItalyMarche6_43.5058744_-137.010385_R0a2n
1
0.01
ItalyMarche5_43.5058744_-137.010385_R0a2n
KenyaOromo2_-0.5766321_-115.5685636_R0a2
SpainSeville1_37.3890924_-155.9844589_R0a2a
0.86
0.02
0.02
SpainMurcia1_37.9922399_-151.1306544_R0a2a
0.05
0.65
KenyaOromo1_-0.5766321_-115.5685636_R0a2
EthiopiaOromo1_7.5460377_-109.3653149_R0a2
0.28
EthiopiaOromo2_7.5460377_-109.3653149_R0a2b
0.27
0.01
0.71
EthiopiaAfar2_11.7559388_-109.041312_R0a2b
Ethiopia60_9.145_-109.510327_R0a2b
LebanonDruze2_33.8923706_-114.488931_R0a2
EthiopiaOromo3_7.5460377_-109.3653149_R0a2g
0.59
0.41
EritreaAfar2_11.7559388_-109.041312_R0a2g
0.21
RomaniaSzekely2_45.8706075_-124.2166114_R0a2
RomaniaSzekely1_45.8706075_-124.2166114_R0a2
0.04
0.66
0.02
RomaniaCsango3_45.7702526_-127.0954547_R0a2
RomaniaCsango1_45.7702526_-127.0954547_R0a2
0.26
RomaniaCsango2_45.7702526_-127.0954547_R0a2
0.15
Bulgaria6_42.733883_-124.51417_R0a2
EastTimor6_-8.874217_-24.272461_P1d
0.62
EastTimor7_-8.874217_-24.272461_P1_T152C_
0.91
EastTimor10_-8.874217_-24.272461_P1_T152C_
0.01
0.05
PhilippinesSurigaonon3_12.879721_-28.225983_Y2a1
0.9
PhilippinesSurigaonon2_12.879721_-28.225983_Y2a1
0.9
IndonesiaSumatra1_-0.589724_-48.6568942_Y2a1
0.51
0.06
Romania7_45.943161_-125.03324_J2b2
0.62
Belarus22_53.709807_-122.046611_J1c7
MalaysiaSemelai2_4.210484_-48.024234_R9b1a1a
0.57

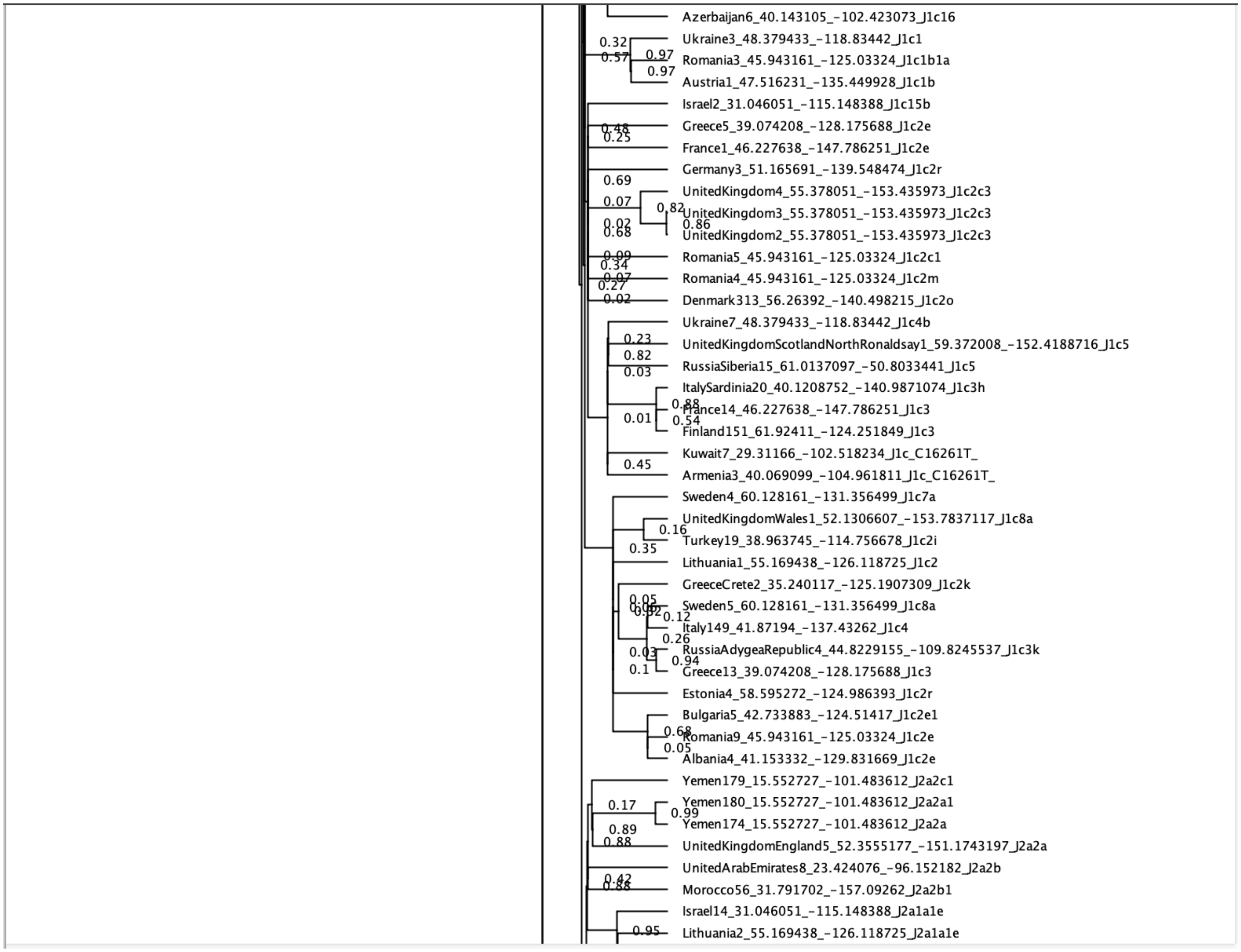
S3.38


MalaysiaSemelai2_4.210484_-48.024234_R9b1a1a
0.57
Australia7_-25.274398_-16.224864_S1
IndonesiaSumatra3_-0.589724_-48.6568942_Y2a1
0.73
IndonesiaSulawesi1_-1.8479_-29.4721_Y2a1
Israel13_31.046051_-115.148388_T2c1d2
1
France3_46.227638_-147.786251_T2b17a
Ukraine1_48.379433_-118.83442_T2_T16189C_
0.39
Iraq3_33.223191_-106.320709_T2e2a
0.1
Denmark621_56.26392_-140.498215_T2
0.28
RussiaSiberia19_61.0137097_-50.8033441_T2d1b1
0.17
RussiaAdygeaRepublic3_44.8229155_-109.8245537_T2l
0.42
Iran290_32.427908_-96.311954_T2n
SouthAfricaFreeStateProvince1_-28.4541105_-123.2032151_T2b33
0.01
UnitedArabEmiratesDubai5_25.2048493_-94.7292172_T2b
0.04
RussiaAdygeaRepublic1_44.8229155_-109.8245537_T2b
0.07
Italy75_41.87194_-137.43262_T2b1
0.22
Tunisia18_33.886917_-140.462501_T2b
0.27
0.23
Israel5_31.046051_-115.148388_T2b_T152C_
0.4
Iran227_32.427908_-96.311954_T2b34
0.39
Nepal15_28.394857_-65.875992_T2b4e
1
Kuwait13_29.31166_-102.518234_T2b4_T152C_
0.32
GreeceCrete17_35.240117_-125.1907309_T2b31
Spain24_40.463667_-153.74922_T2c1d_T152C_
1
Kuwait14_29.31166_-102.518234_T2c1d_T152C_
0.84
Israel6_31.046051_-115.148388_T2c1c1
0.69
CzechRepublic4_49.817492_-134.527038_T2c1f
Iraq24_33.223191_-106.320709_T2i1
UnitedKingdomEngland3_52.3555177_-151.1743197_T2g2a
0.51
0.03
IndiaAndhraPradesh1_15.9128998_-70.2600125_T2
0.23
UnitedKingdomEngland6_52.3555177_-151.1743197_T2a1b1a1a2
1
Iraq18_33.223191_-106.320709_T2a1
RussiaAdygeaRepublic2_44.8229155_-109.8245537_T2
0.76
0.2
GreeceCrete12_35.240117_-125.1907309_T2c1c2
Azerbaijan4_40.143105_-102.423073_T2
Jordan10_30.585164_-113.761586_T1b1
0.66
0.4
Turkey11_38.963745_-114.756678_T1b3
0.35
SaudiArabia18_23.885942_-104.920838_T1b
0.17
India23_20.593684_-71.03712_T1b
RussiaKemerovoRegion1_54.7574648_-62.5944712_T1
0.04
Iraq14_33.223191_-106.320709_T1
0.29
Turkey10_38.963745_-114.756678_T1a9
0.16
ItalySardinia14_40.1208752_-140.9871074_T1a12
0.23
GreeceCrete15_35.240117_-125.1907309_T1a2a

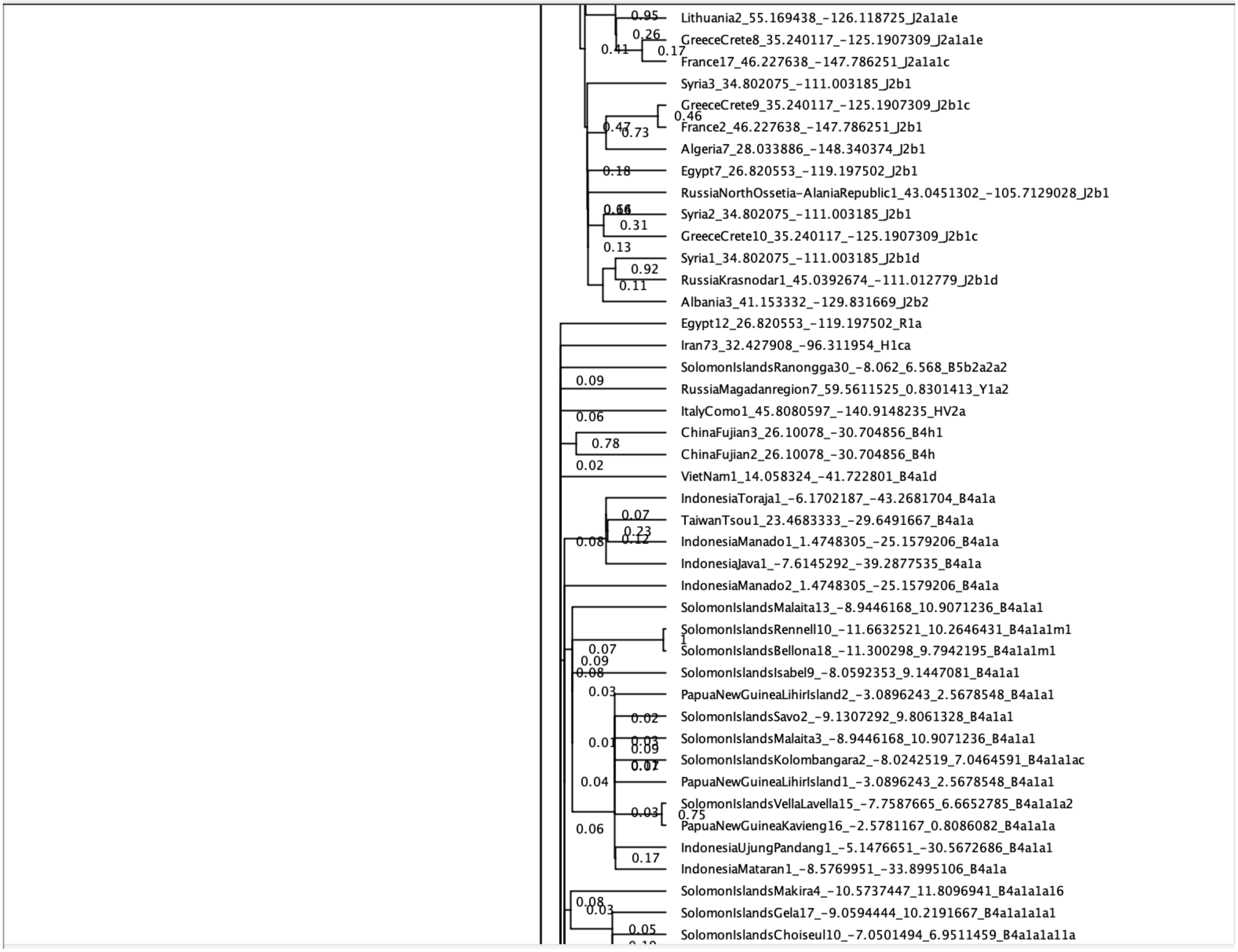
S3.39


GreeceCrete15_35.240117_-125.1907309_T1a2a
Sweden3_60.128161_-131.356499_T1a1e
RussiaBuryatRepublic10_54.8331146_-37.5939471_T1a1
Russia161_61.52401_-44.681244_T1a1
Estonia2_58.595272_-124.986393_T1
Iraq15_33.223191_-106.320709_T1
Azerbaijan1_40.143105_-102.423073_T1a8b
RussiaAdygeaRepublic6_44.8229155_-109.8245537_T2a1b2b
Algeria10_28.033886_-148.340374_T1
UnitedArabEmiratesDubai11_25.2048493_-94.7292172_J1b
UnitedArabEmirates22_23.424076_-96.152182_J1b
UnitedArabEmirates3_23.424076_-96.152182_J1b1a1_T146C_
Kuwait1_29.31166_-102.518234_J1b1b1
Greece50_39.074208_-128.175688_J1b3b
UnitedArabEmirates7_23.424076_-96.152182_J1b1b3
UnitedArabEmirates11_23.424076_-96.152182_J1b2
UnitedArabEmirates16_23.424076_-96.152182_J1b2
Kuwait5_29.31166_-102.518234_J1b2
Kuwait4_29.31166_-102.518234_J1b2a
Iraq27_33.223191_-106.320709_J1b7a
Slovakia13_48.669026_-130.300976_J1b3a
France16_46.227638_-147.786251_J1b2
RussiaDagestanKumyk1_42.1431886_-102.9050201_J1_C16193T_
UnitedArabEmirates17_23.424076_-96.152182_J1d3a2
RussiaNorthOssetia-AlaniaRepublic3_43.0451302_-105.7129028_J1d3a1
RussiaAdygeaRepublic5_44.8229155_-109.8245537_J1d3a1
Uganda3_1.373333_-117.709725_J1d1a1a
Uganda2_1.373333_-117.709725_J1d1a1a
Kuwait9_29.31166_-102.518234_J1d1a1
UnitedArabEmirates21_23.424076_-96.152182_J1d
SpainMaragato1_40.602383_-154.4990202_J1d1b
Cyprus2_35.126413_-116.570141_J1d1b1
India27_20.593684_-71.03712_J1c5e
Romania11_45.943161_-125.03324_J1c_C16261T_
Romania6_45.943161_-125.03324_J1c_C16261T_
Romania10_45.943161_-125.03324_J1c_C16261T_
BosniaandHerzegovina3_43.915886_-132.320924_J1c_C16261T_
UnitedKingdomEngland4_52.3555177_-151.1743197_J1c1d
Ukraine5_48.379433_-118.83442_J1c3f
Sweden1_60.128161_-131.356499_J1c3f
Kazakhstan2_48.019573_-83.076316_J1c3f
Estonia5_58.595272_-124.986393_J1c3f
Azerbaijan6_40.143105_-102.423073_J1c16
0.23
0.33
0.19
0.51
0.7
0.47
1
1
1
1
0.85
0.87
0.21
0.36
0.85
0.84
0.63
0.33
0.41
0.26
0.17
0.49
0.74
0.74
0.76
0.29
0.7
0.76
0.76
0.17
0.94
0.58
0.88
0.2
0.33
0.37
0.86
0.35
0.69
0.36
0.2

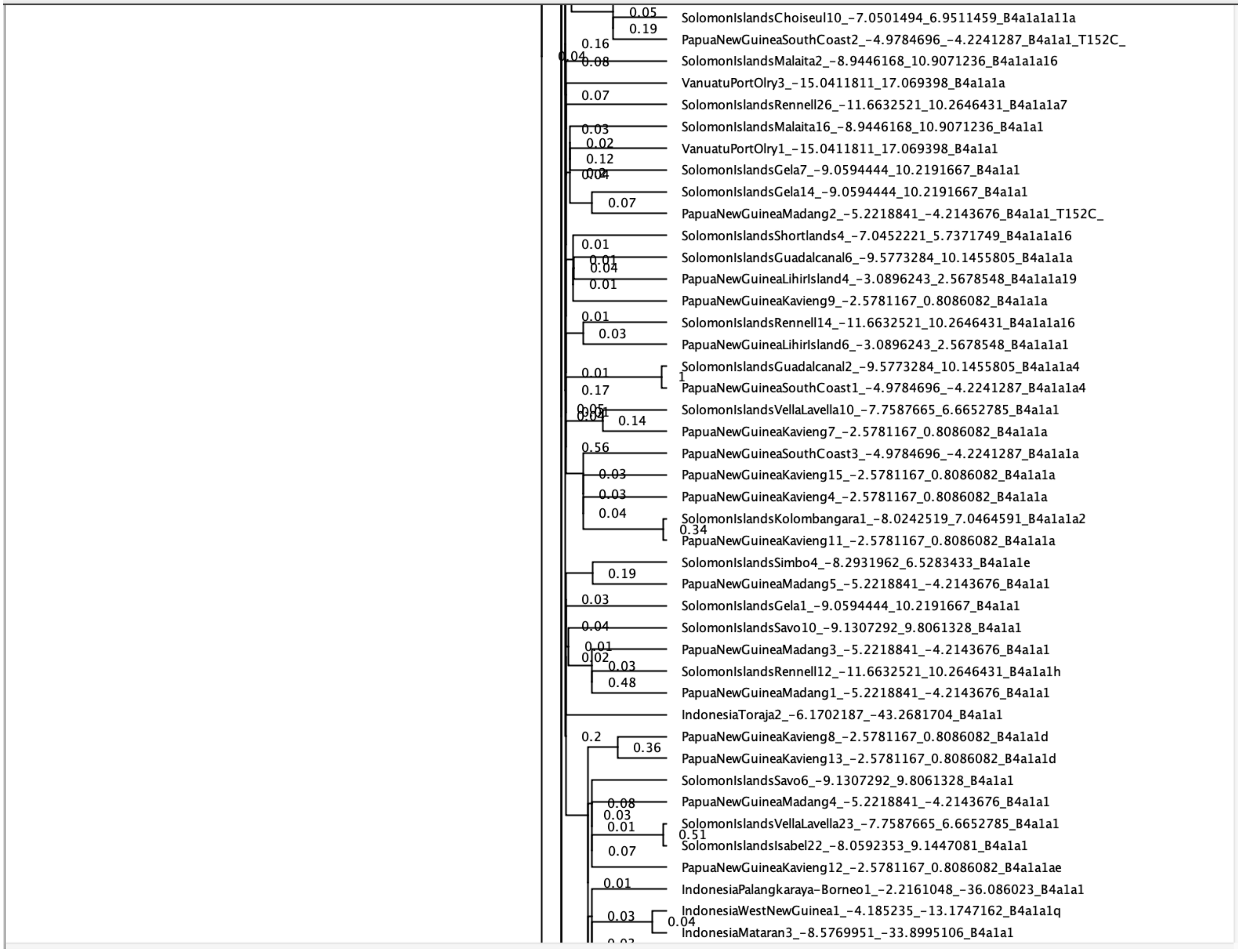
S3.40


Azerbaijan6_40.143105_-102.423073_J1c16
0.32
Ukraine3_48.379433_-118.83442_J1c1
0.57
0.97
Romania3_45.943161_-125.03324_J1c1b1a
0.97
Austria1_47.516231_-135.449928_J1c1b
Israel2_31.046051_-115.148388_J1c15b
0.48
Greece5_39.074208_-128.175688_J1c2e
0.25
France1_46.227638_-147.786251_J1c2e
Germany3_51.165691_-139.548474_J1c2r
0.69
UnitedKingdom4_55.378051_-153.435973_J1c2c3
0.07
0.82
UnitedKingdom3_55.378051_-153.435973_J1c2c3
0.02
0.86
0.68
UnitedKingdom2_55.378051_-153.435973_J1c2c3
0.09
Romania5_45.943161_-125.03324_J1c2c1
0.34
0.07
Romania4_45.943161_-125.03324_J1c2m
0.27
Denmark313_56.26392_-140.498215_J1c2o
0.02
Ukraine7_48.379433_-118.83442_J1c4b
0.23
UnitedKingdomScotlandNorthRonaldsay1_59.372008_-152.4188716_J1c5
0.82
RussiaSiberia15_61.0137097_-50.8033441_J1c5
0.03
ItalySardinia20_40.1208752_-140.9871074_J1c3h
0.88
France14_46.227638_-147.786251_J1c3
0.01
0.54
Finland151_61.92411_-124.251849_J1c3
Kuwait7_29.31166_-102.518234_J1c_C16261T_
0.45
Armenia3_40.069099_-104.961811_J1c_C16261T_
Sweden4_60.128161_-131.356499_J1c7a
UnitedKingdomWales1_52.1306607_-153.7837117_J1c8a
0.16
Turkey19_38.963745_-114.756678_J1c2i
0.35
Lithuania1_55.169438_-126.118725_J1c2
GreeceCrete2_35.240117_-125.1907309_J1c2k
0.05
0.06
0.92
Sweden5_60.128161_-131.356499_J1c8a
0.12
Italy149_41.87194_-137.43262_J1c4
0.26
0.03
RussiaAdygeaRepublic4_44.8229155_-109.8245537_J1c3k
0.94
0.1
Greece13_39.074208_-128.175688_J1c3
Estonia4_58.595272_-124.986393_J1c2r
Bulgaria5_42.733883_-124.51417_J1c2e1
0.68
Romania9_45.943161_-125.03324_J1c2e
0.05
Albania4_41.153332_-129.831669_J1c2e
Yemen179_15.552727_-101.483612_J2a2c1
0.17
Yemen180_15.552727_-101.483612_J2a2a1
0.99
0.89
Yemen174_15.552727_-101.483612_J2a2a
0.88
UnitedKingdomEngland5_52.3555177_-151.1743197_J2a2a
UnitedArabEmirates8_23.424076_-96.152182_J2a2b
0.42
0.88
Morocco56_31.791702_-157.09262_J2a2b1
Israel14_31.046051_-115.148388_J2a1a1e
0.95
Lithuania2_55.169438_-126.118725_J2a1a1e

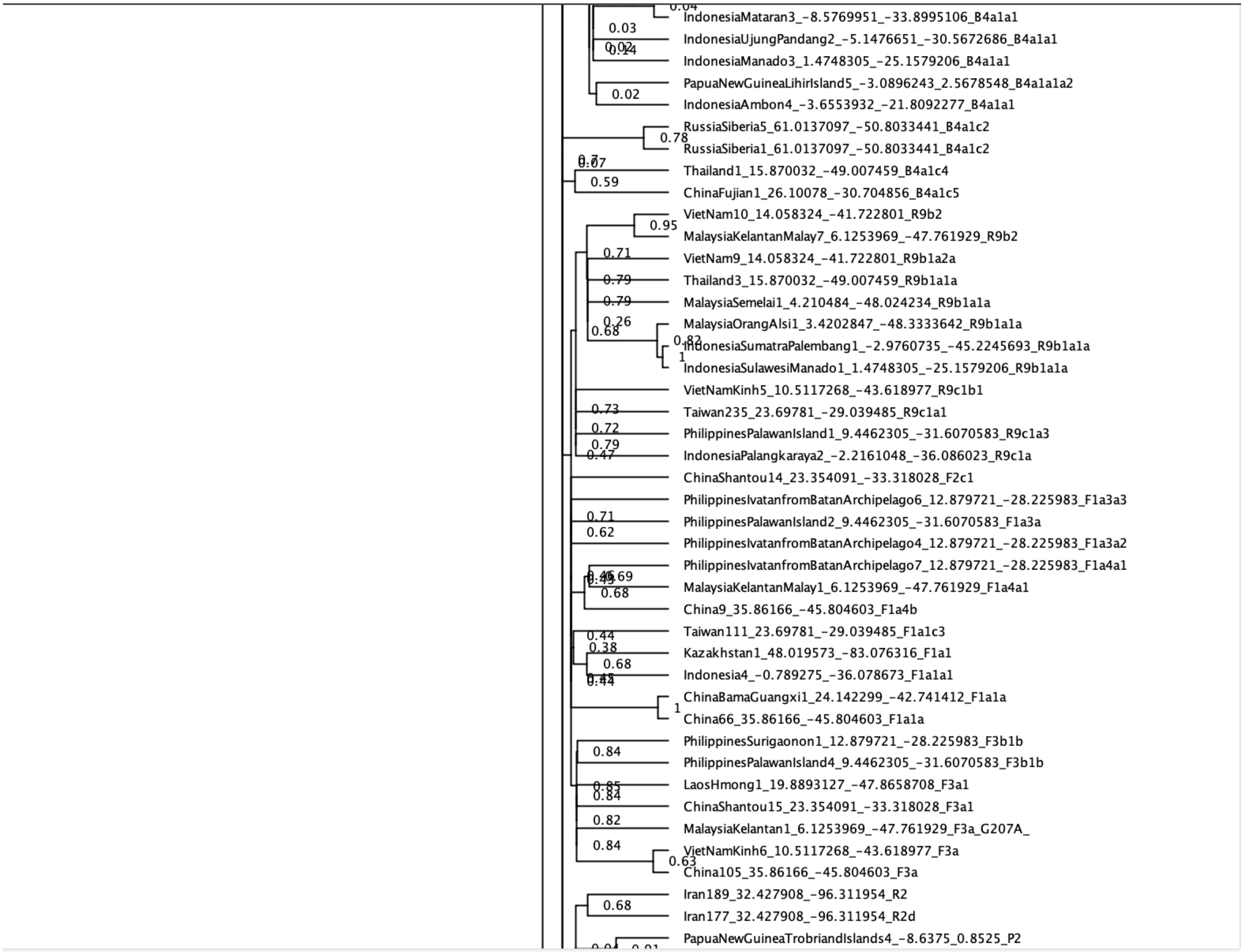
S3.41


0.95
Lithuania2_55.169438_-126.118725_J2a1a1e
0.26
GreeceCrete8_35.240117_-125.1907309_J2a1a1e
0.41
0.17
France17_46.227638_-147.786251_J2a1a1c
Syria3_34.802075_-111.003185_J2b1
GreeceCrete9_35.240117_-125.1907309_J2b1c
0.46
0.47
0.73
France2_46.227638_-147.786251_J2b1
Algeria7_28.033886_-148.340374_J2b1
0.18
Egypt7_26.820553_-119.197502_J2b1
RussiaNorthOssetia-AlaniaRepublic1_43.0451302_-105.7129028_J2b1
Syria2_34.802075_-111.003185_J2b1
0.31
GreeceCrete10_35.240117_-125.1907309_J2b1c
0.13
Syria1_34.802075_-111.003185_J2b1d
0.92
RussiaKrasnodar1_45.0392674_-111.012779_J2b1d
0.11
Albania3_41.153332_-129.831669_J2b2
Egypt12_26.820553_-119.197502_R1a
Iran73_32.427908_-96.311954_H1ca
SolomonIslandsRanongga30_-8.062_6.568_B5b2a2a2
0.09
RussiaMagadanregion7_59.5611525_0.8301413_Y1a2
ItalyComo1_45.8080597_-140.9148235_HV2a
0.06
ChinaFujian3_26.10078_-30.704856_B4h1
0.78
ChinaFujian2_26.10078_-30.704856_B4h
0.02
VietNam1_14.058324_-41.722801_B4a1d
IndonesiaToraja1_-6.1702187_-43.2681704_B4a1a
0.07
TaiwanTsou1_23.4683333_-29.6491667_B4a1a
0.23
0.08
0.12
IndonesiaManado1_1.4748305_-25.1579206_B4a1a
IndonesiaJava1_-7.6145292_-39.2877535_B4a1a
IndonesiaManado2_1.4748305_-25.1579206_B4a1a
SolomonIslandsMalaita13_-8.9446168_10.9071236_B4a1a1
SolomonIslandsRennell10_-11.6632521_10.2646431_B4a1a1m1
1
0.07
SolomonIslandsBellona18_-11.300298_9.7942195_B4a1a1m1
0.09
0.08
SolomonIslandsIsabel9_-8.0592353_9.1447081_B4a1a1
0.03
PapuaNewGuineaLihirIsland2_-3.0896243_2.5678548_B4a1a1
0.02
SolomonIslandsSavo2_-9.1307292_9.8061328_B4a1a1
0.01
0.03
SolomonIslandsMalaita3_-8.9446168_10.9071236_B4a1a1
0.09
SolomonIslandsKolombangara2_-8.0242519_7.0464591_B4a1a1ac
0.04
PapuaNewGuineaLihirIsland1_-3.0896243_2.5678548_B4a1a1
SolomonIslandsVellaLavella15_-7.7587665_6.6652785_B4a1a1a2
0.03
0.75
PapuaNewGuineaKavieng16_-2.5781167_0.8086082_B4a1a1a
0.06
IndonesiaUjungPandang1_-5.1476651_-30.5672686_B4a1a1
0.17
IndonesiaMataran1_-8.5769951_-33.8995106_B4a1a
SolomonIslandsMakira4_-10.5737447_11.8096941_B4a1a1a16
0.08
0.03
SolomonIslandsGela17_-9.0594444_10.2191667_B4a1a1a1a1
0.05
SolomonIslandsChoiseul10_-7.0501494_6.9511459_B4a1a1a11a

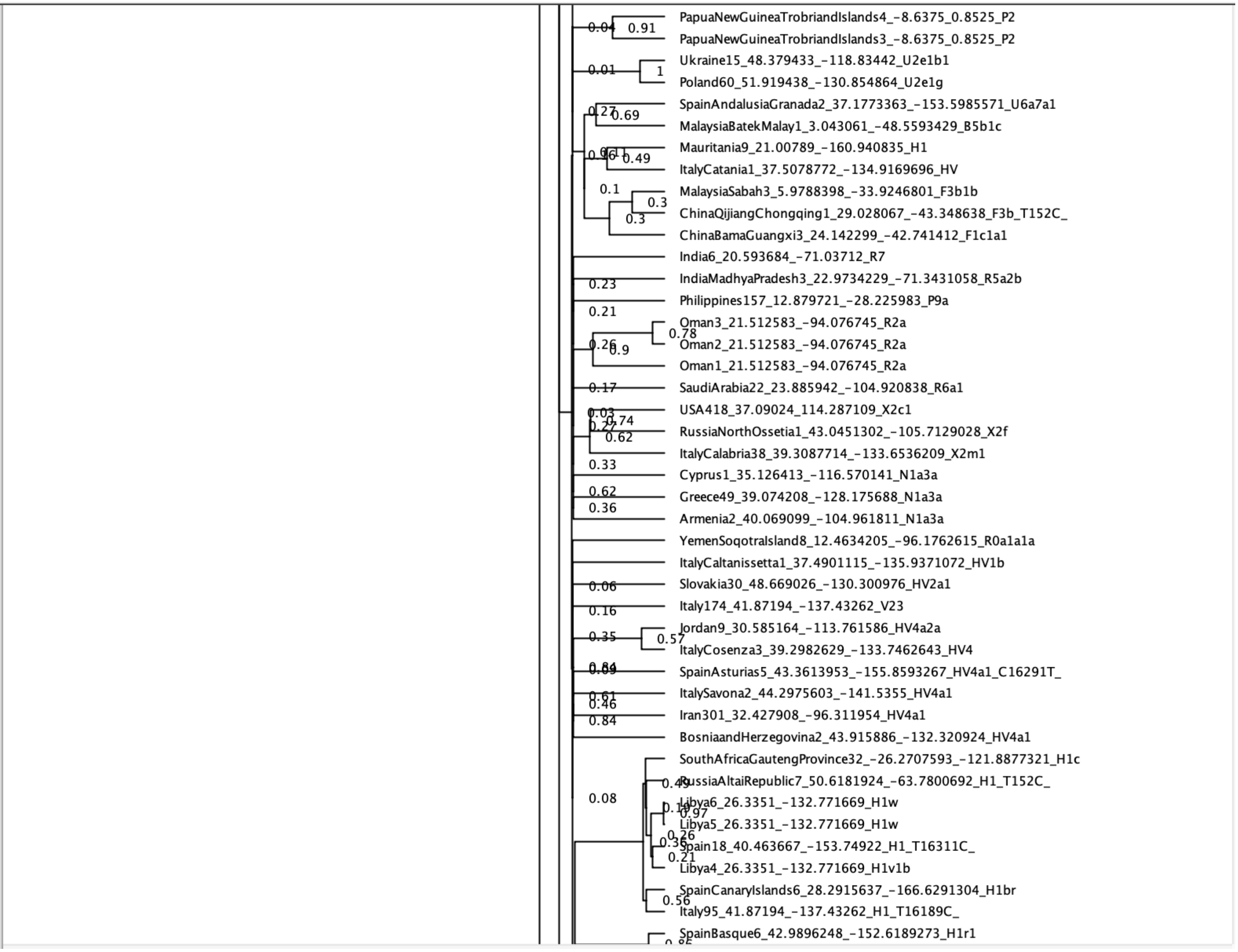
S3.42


0.05
SolomonIslandsChoiseul10_-7.0501494_6.9511459_B4a1a1a11a
0.19
PapuaNewGuineaSouthCoast2_-4.9784696_-4.2241287_B4a1a1_T152C_
0.16
0.04
0.08
SolomonIslandsMalaita2_-8.9446168_10.9071236_B4a1a1a16
VanuatuPortOlry3_-15.0411811_17.069398_B4a1a1a
0.07
SolomonIslandsRennell26_-11.6632521_10.2646431_B4a1a1a7
0.03
SolomonIslandsMalaita16_-8.9446168_10.9071236_B4a1a1
0.02
VanuatuPortOlry1_-15.0411811_17.069398_B4a1a1
0.12
SolomonIslandsGela7_-9.0594444_10.2191667_B4a1a1
0.04
0.2
SolomonIslandsGela14_-9.0594444_10.2191667_B4a1a1
0.07
PapuaNewGuineaMadang2_-5.2218841_-4.2143676_B4a1a1_T152C_
SolomonIslandsShortlands4_-7.0452221_5.7371749_B4a1a1a16
0.01
0.01
SolomonIslandsGuadalcanal6_-9.5773284_10.1455805_B4a1a1a
0.04
PapuaNewGuineaLihirIsland4_-3.0896243_2.5678548_B4a1a1a19
0.01
PapuaNewGuineaKavieng9_-2.5781167_0.8086082_B4a1a1a
0.01
SolomonIslandsRennell14_-11.6632521_10.2646431_B4a1a1a16
0.03
PapuaNewGuineaLihirIsland6_-3.0896243_2.5678548_B4a1a1a1
0.01
SolomonIslandsGuadalcanal2_-9.5773284_10.1455805_B4a1a1a4
1
0.17
PapuaNewGuineaSouthCoast1_-4.9784696_-4.2241287_B4a1a1a4
0.05
0.04
0.01
SolomonIslandsVellaLavella10_-7.7587665_6.6652785_B4a1a1
0.14
PapuaNewGuineaKavieng7_-2.5781167_0.8086082_B4a1a1a
0.56
PapuaNewGuineaSouthCoast3_-4.9784696_-4.2241287_B4a1a1a
0.03
PapuaNewGuineaKavieng15_-2.5781167_0.8086082_B4a1a1a
0.03
PapuaNewGuineaKavieng4_-2.5781167_0.8086082_B4a1a1a
0.04
SolomonIslandsKolombangara1_-8.0242519_7.0464591_B4a1a1a2
0.34
PapuaNewGuineaKavieng11_-2.5781167_0.8086082_B4a1a1a
SolomonIslandsSimbo4_-8.2931962_6.5283433_B4a1a1e
0.19
PapuaNewGuineaMadang5_-5.2218841_-4.2143676_B4a1a1
0.03
SolomonIslandsGela1_-9.0594444_10.2191667_B4a1a1
0.04
SolomonIslandsSavo10_-9.1307292_9.8061328_B4a1a1
0.01
0.02
PapuaNewGuineaMadang3_-5.2218841_-4.2143676_B4a1a1
0.03
SolomonIslandsRennell12_-11.6632521_10.2646431_B4a1a1h
0.48
PapuaNewGuineaMadang1_-5.2218841_-4.2143676_B4a1a1
IndonesiaToraja2_-6.1702187_-43.2681704_B4a1a1
0.2
PapuaNewGuineaKavieng8_-2.5781167_0.8086082_B4a1a1d
0.36
PapuaNewGuineaKavieng13_-2.5781167_0.8086082_B4a1a1d
SolomonIslandsSavo6_-9.1307292_9.8061328_B4a1a1
0.08
PapuaNewGuineaMadang4_-5.2218841_-4.2143676_B4a1a1
0.03
0.01
SolomonIslandsVellaLavella23_-7.7587665_6.6652785_B4a1a1
0.51
SolomonIslandsIsabel22_-8.0592353_9.1447081_B4a1a1
0.07
PapuaNewGuineaKavieng12_-2.5781167_0.8086082_B4a1a1ae
0.01
IndonesiaPalangkaraya-Borneo1_-2.2161048_-36.086023_B4a1a1
0.03
IndonesiaWestNewGuinea1_-4.185235_-13.1747162_B4a1a1q
0.04
IndonesiaMataran3_-8.5769951_-33.8995106_B4a1a1

S3.43

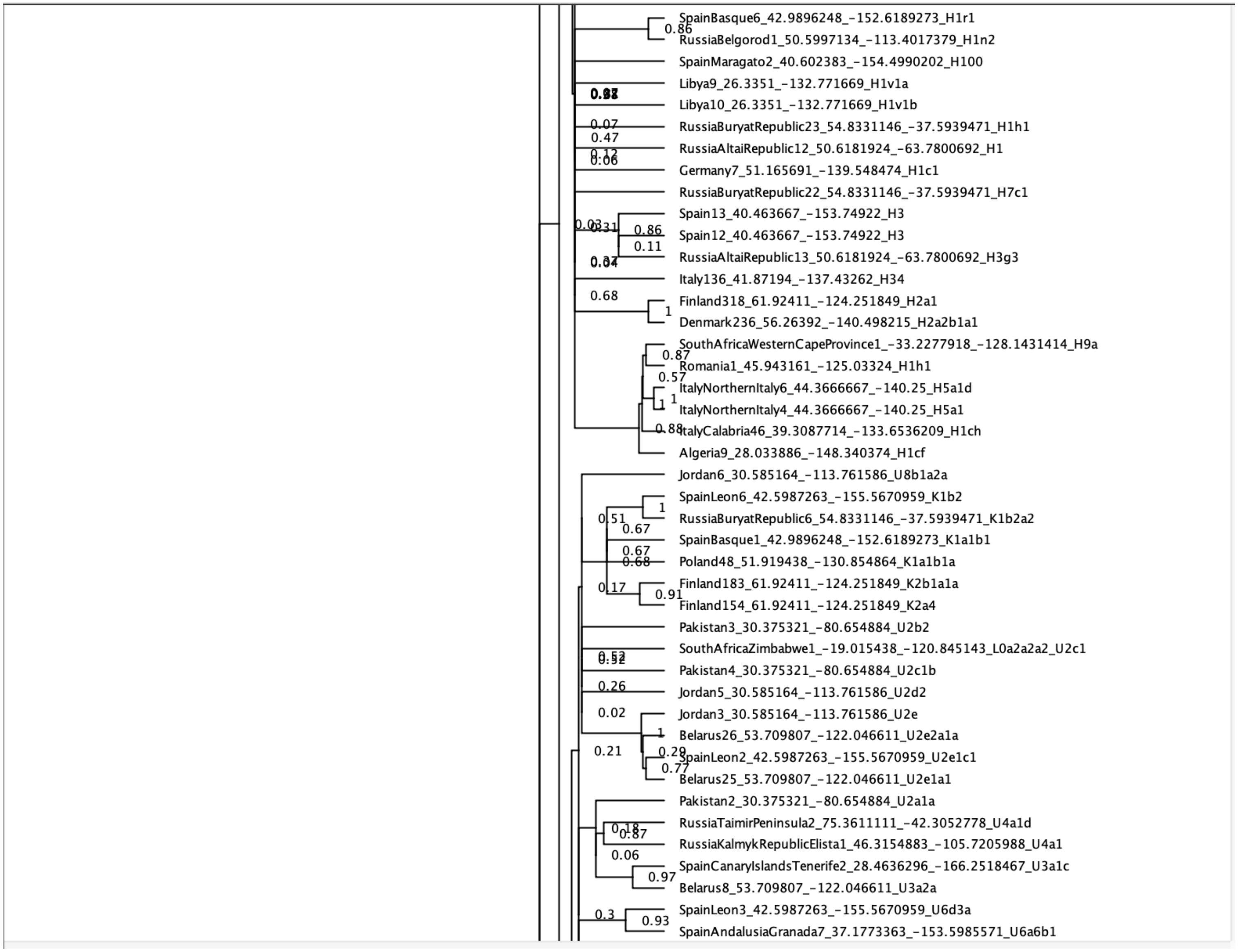

0.04
IndonesiaMataran3_-8.5769951_-33.8995106_B4a1a1
0.03
IndonesiaUjungPandang2_-5.1476651_-30.5672686_B4a1a1
0.02
0.14
IndonesiaManado3_1.4748305_-25.1579206_B4a1a1
PapuaNewGuineaLihirIsland5_-3.0896243_2.5678548_B4a1a1a2
0.02
IndonesiaAmbon4_-3.6553932_-21.8092277_B4a1a1
RussiaSiberia5_61.0137097_-50.8033441_B4a1c2
0.78
RussiaSiberia1_61.0137097_-50.8033441_B4a1c2
0.7
0.07
Thailand1_15.870032_-49.007459_B4a1c4
0.59
ChinaFujian1_26.10078_-30.704856_B4a1c5
VietNam10_14.058324_-41.722801_R9b2
0.95
MalaysiaKelantanMalay7_6.1253969_-47.761929_R9b2
0.71
VietNam9_14.058324_-41.722801_R9b1a2a
0.79
Thailand3_15.870032_-49.007459_R9b1a1a
0.79
MalaysiaSemelai1_4.210484_-48.024234_R9b1a1a
0.26
0.68
MalaysiaOrangAlsi1_3.4202847_-48.3333642_R9b1a1a
0.82
IndonesiaSumatraPalembang1_-2.9760735_-45.2245693_R9b1a1a
1
IndonesiaSulawesiManado1_1.4748305_-25.1579206_R9b1a1a
VietNamKinh5_10.5117268_-43.618977_R9c1b1
0.73
Taiwan235_23.69781_-29.039485_R9c1a1
0.72
PhilippinesPalawanIsland1_9.4462305_-31.6070583_R9c1a3
0.79
0.47
IndonesiaPalangkaraya2_-2.2161048_-36.086023_R9c1a
ChinaShantou14_23.354091_-33.318028_F2c1
PhilippinesIvatanfromBatanArchipelago6_12.879721_-28.225983_F1a3a3
0.71
PhilippinesPalawanIsland2_9.4462305_-31.6070583_F1a3a
0.62
PhilippinesIvatanfromBatanArchipelago4_12.879721_-28.225983_F1a3a2
PhilippinesIvatanfromBatanArchipelago7_12.879721_-28.225983_F1a4a1
0.46
0.45
0.69
MalaysiaKelantanMalay1_6.1253969_-47.761929_F1a4a1
0.68
China9_35.86166_-45.804603_F1a4b
Taiwan111_23.69781_-29.039485_F1a1c3
0.44
0.38
Kazakhstan1_48.019573_-83.076316_F1a1
0.68
Indonesia4_-0.789275_-36.078673_F1a1a1
0.45
0.44
ChinaBamaGuangxi1_24.142299_-42.741412_F1a1a
1
China66_35.86166_-45.804603_F1a1a
PhilippinesSurigaonon1_12.879721_-28.225983_F3b1b
0.84
PhilippinesPalawanIsland4_9.4462305_-31.6070583_F3b1b
LaosHmong1_19.8893127_-47.8658708_F3a1
0.85
0.84
ChinaShantou15_23.354091_-33.318028_F3a1
0.82
MalaysiaKelantan1_6.1253969_-47.761929_F3a_G207A_
0.84
VietNamKinh6_10.5117268_-43.618977_F3a
0.63
China105_35.86166_-45.804603_F3a
Iran189_32.427908_-96.311954_R2
0.68
Iran177_32.427908_-96.311954_R2d
PapuaNewGuineaTrobriandIslands4_-8.6375_0.8525_P2
0.04
0.01

S3.44

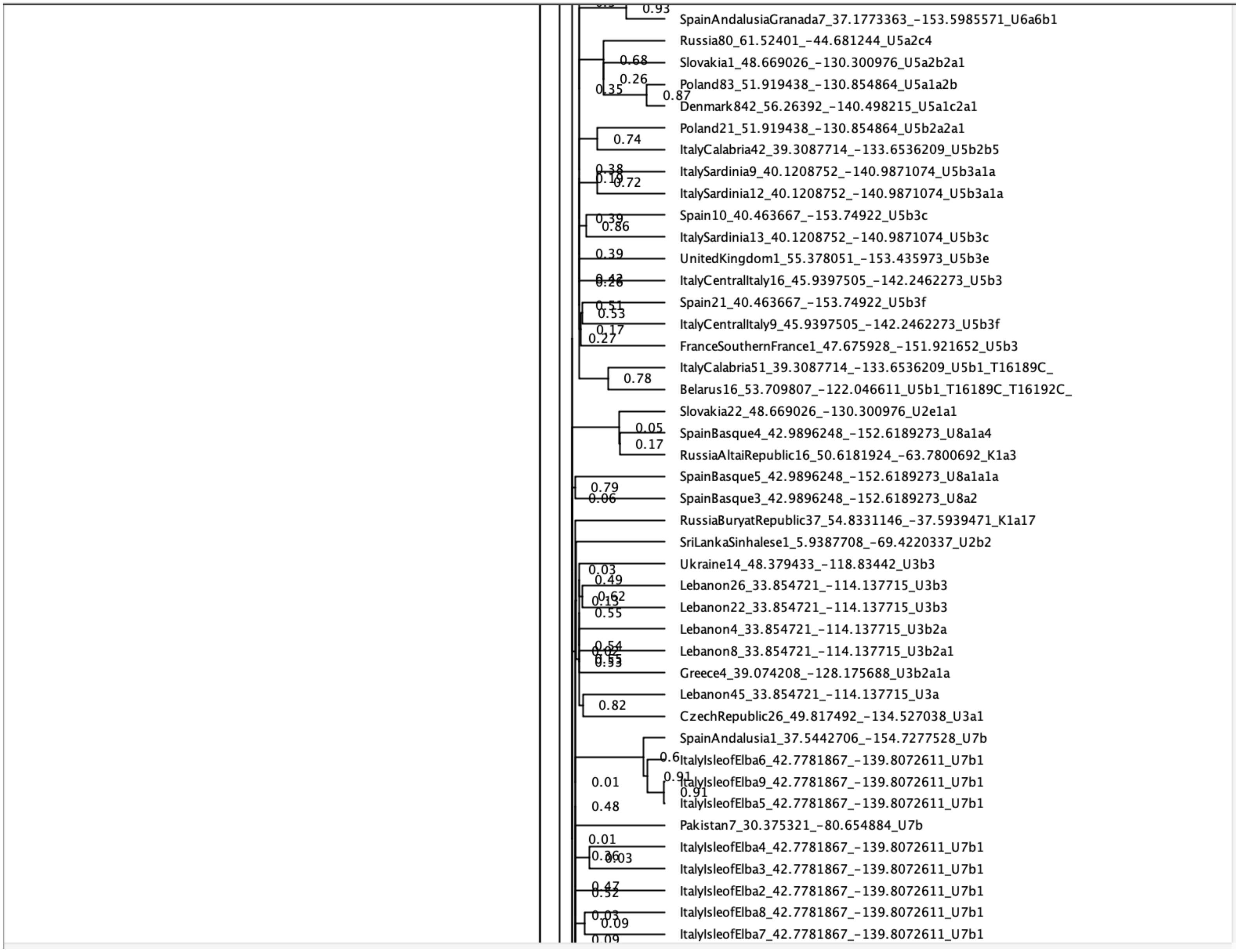


PapuaNewGuineaTrobriandIslands4_-8.6375_0.8525_P2
PapuaNewGuineaTrobriandIslands3_-8.6375_0.8525_P2
Ukraine15_48.379433_-118.83442_U2e1b1
Poland60_51.919438_-130.854864_U2e1g
SpainAndalusiaGranada2_37.1773363_-153.5985571_U6a7a1
MalaysiaBatekMalay1_3.043061_-48.5593429_B5b1c
Mauritania9_21.00789_-160.940835_H1
ItalyCatania1_37.5078772_-134.9169696_HV
MalaysiaSabah3_5.9788398_-33.9246801_F3b1b
ChinaQijiangChongqing1_29.028067_-43.348638_F3b_T152C_
ChinaBamaGuangxi3_24.142299_-42.741412_F1c1a1
India6_20.593684_-71.03712_R7
IndiaMadhyaPradesh3_22.9734229_-71.3431058_R5a2b
Philippines157_12.879721_-28.225983_P9a
Oman3_21.512583_-94.076745_R2a
Oman2_21.512583_-94.076745_R2a
Oman1_21.512583_-94.076745_R2a
SaudiArabia22_23.885942_-104.920838_R6a1
USA418_37.09024_114.287109_X2c1
RussiaNorthOssetia1_43.0451302_-105.7129028_X2f
ItalyCalabria38_39.3087714_-133.6536209_X2m1
Cyprus1_35.126413_-116.570141_N1a3a
Greece49_39.074208_-128.175688_N1a3a
Armenia2_40.069099_-104.961811_N1a3a
YemenSoqotraIsland8_12.4634205_-96.1762615_R0a1a1a
ItalyCaltanissetta1_37.4901115_-135.9371072_HV1b
Slovakia30_48.669026_-130.300976_HV2a1
Italy174_41.87194_-137.43262_V23
Jordan9_30.585164_-113.761586_HV4a2a
ItalyCosenza3_39.2982629_-133.7462643_HV4
SpainAsturias5_43.3613953_-155.8593267_HV4a1_C16291T_
ItalySavona2_44.2975603_-141.5355_HV4a1
Iran301_32.427908_-96.311954_HV4a1
BosniaandHerzegovina2_43.915886_-132.320924_HV4a1
SouthAfricaGautengProvince32_-26.2707593_-121.8877321_H1c
RussiaAltaiRepublic7_50.6181924_-63.7800692_H1_T152C_
Libya6_26.3351_-132.771669_H1w
Libya5_26.3351_-132.771669_H1w
Spain18_40.463667_-153.74922_H1_T16311C_
Libya4_26.3351_-132.771669_H1v1b
SpainCanaryIslands6_28.2915637_-166.6291304_H1br
Italy95_41.87194_-137.43262_H1_T16189C_
SpainBasque6_42.9896248_-152.6189273_H1r1
0.04
0.91
0.01
1
0.27
0.69
0.49
0.1
0.3
0.3
0.23
0.21
0.78
0.26
0.9
0.17
0.03
0.74
0.27
0.62
0.33
0.62
0.36
0.06
0.16
0.35
0.57
0.84
0.69
0.61
0.46
0.84
0.08
0.49
0.19
0.97
0.26
0.35
0.21
0.56

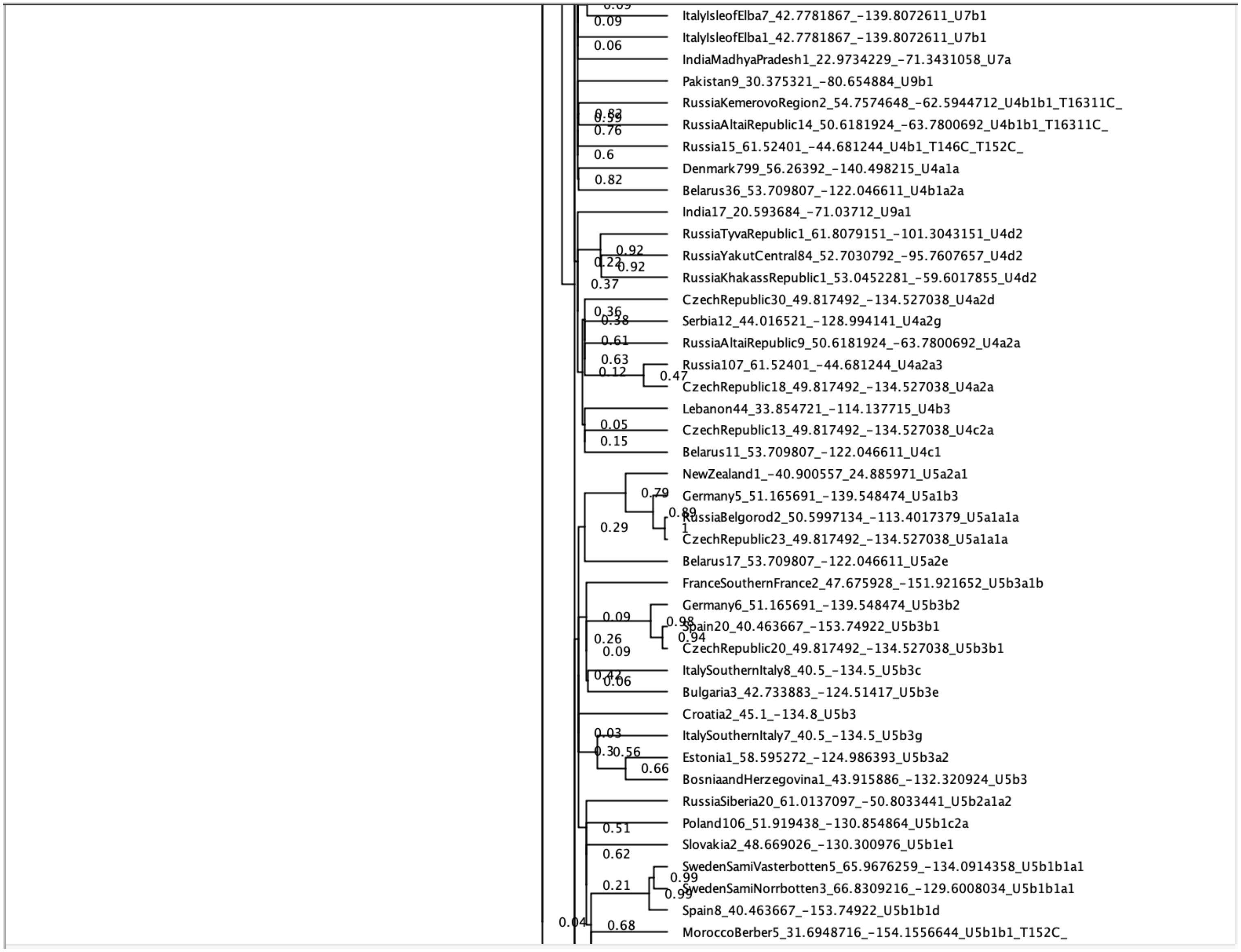
S3.45


SpainBasque6_42.9896248_-152.6189273_H1r1
0.86
RussiaBelgorod1_50.5997134_-113.4017379_H1n2
SpainMaragato2_40.602383_-154.4990202_H100
Libya9_26.3351_-132.771669_H1v1a
Libya10_26.3351_-132.771669_H1v1b
0.07
RussiaBuryatRepublic23_54.8331146_-37.5939471_H1h1
0.47
RussiaAltaiRepublic12_50.6181924_-63.7800692_H1
0.12
0.06
Germany7_51.165691_-139.548474_H1c1
RussiaBuryatRepublic22_54.8331146_-37.5939471_H7c1
Spain13_40.463667_-153.74922_H3
0.86
Spain12_40.463667_-153.74922_H3
0.11
RussiaAltaiRepublic13_50.6181924_-63.7800692_H3g3
Italy136_41.87194_-137.43262_H34
0.68
Finland318_61.92411_-124.251849_H2a1
1
Denmark236_56.26392_-140.498215_H2a2b1a1
SouthAfricaWesternCapeProvince1_-33.2277918_-128.1431414_H9a
0.87
Romania1_45.943161_-125.03324_H1h1
0.57
ItalyNorthernItaly6_44.3666667_-140.25_H5a1d
1
1
ItalyNorthernItaly4_44.3666667_-140.25_H5a1
0.88
ItalyCalabria46_39.3087714_-133.6536209_H1ch
Algeria9_28.033886_-148.340374_H1cf
Jordan6_30.585164_-113.761586_U8b1a2a
SpainLeon6_42.5987263_-155.5670959_K1b2
1
0.51
RussiaBuryatRepublic6_54.8331146_-37.5939471_K1b2a2
0.67
SpainBasque1_42.9896248_-152.6189273_K1a1b1
0.67
0.68
Poland48_51.919438_-130.854864_K1a1b1a
0.17
Finland183_61.92411_-124.251849_K2b1a1a
0.91
Finland154_61.92411_-124.251849_K2a4
Pakistan3_30.375321_-80.654884_U2b2
SouthAfricaZimbabwe1_-19.015438_-120.845143_L0a2a2a2_U2c1
Pakistan4_30.375321_-80.654884_U2c1b
0.26
Jordan5_30.585164_-113.761586_U2d2
0.02
Jordan3_30.585164_-113.761586_U2e
1
Belarus26_53.709807_-122.046611_U2e2a1a
0.21
0.29
SpainLeon2_42.5987263_-155.5670959_U2e1c1
0.77
Belarus25_53.709807_-122.046611_U2e1a1
Pakistan2_30.375321_-80.654884_U2a1a
0.18
0.87
RussiaTaimirPeninsula2_75.3611111_-42.3052778_U4a1d
RussiaKalmykRepublicElista1_46.3154883_-105.7205988_U4a1
0.06
SpainCanaryIslandsTenerife2_28.4636296_-166.2518467_U3a1c
0.97
Belarus8_53.709807_-122.046611_U3a2a
0.3
SpainLeon3_42.5987263_-155.5670959_U6d3a
0.93
SpainAndalusiaGranada7_37.1773363_-153.5985571_U6a6b1

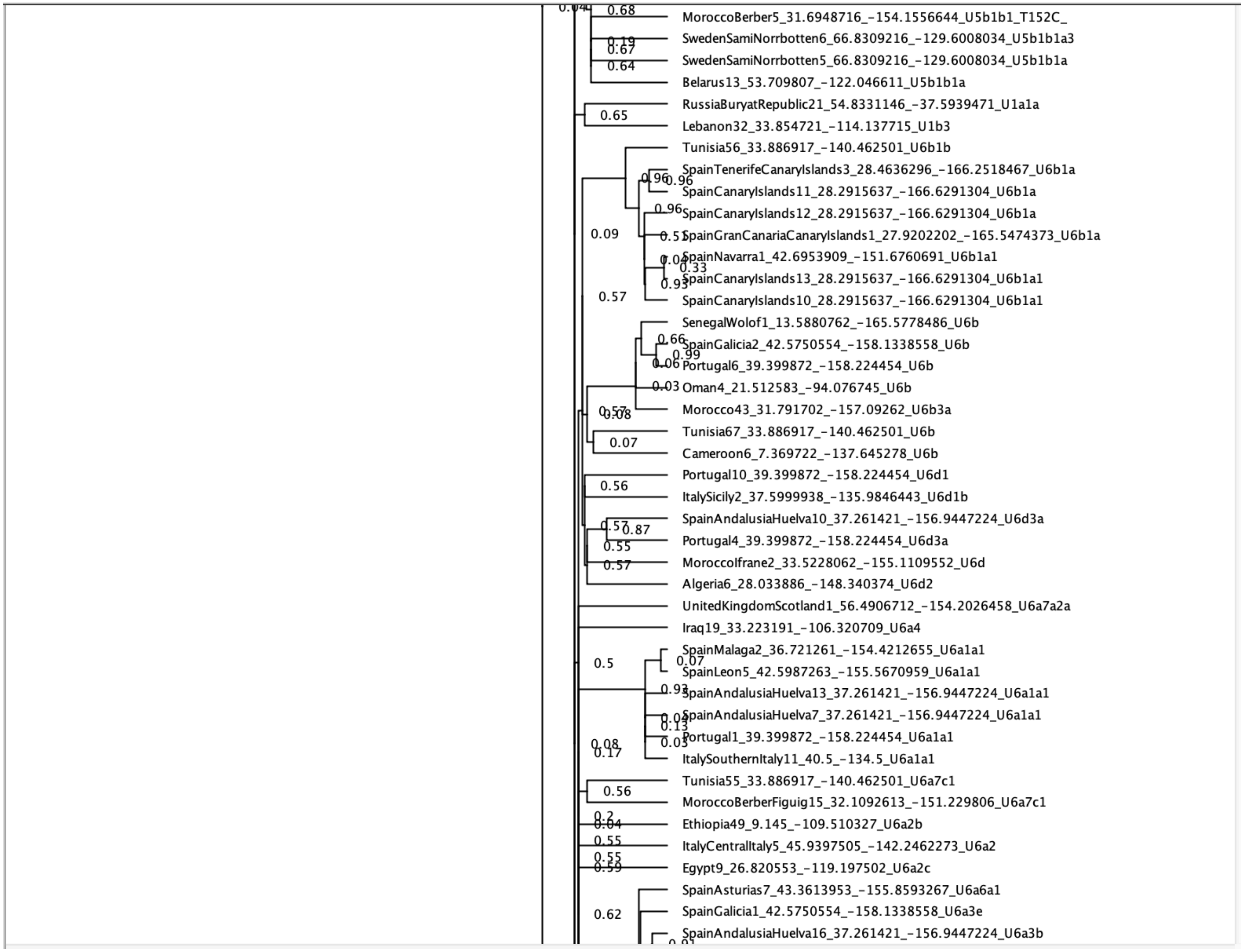
S3.46


0.93
SpainAndalusiaGranada7_37.1773363_-153.5985571_U6a6b1
Russia80_61.52401_-44.681244_U5a2c4
0.68
Slovakia1_48.669026_-130.300976_U5a2b2a1
0.26
0.35
Poland83_51.919438_-130.854864_U5a1a2b
0.87
Denmark842_56.26392_-140.498215_U5a1c2a1
Poland21_51.919438_-130.854864_U5b2a2a1
0.74
ItalyCalabria42_39.3087714_-133.6536209_U5b2b5
0.38
0.10
0.72
ItalySardinia9_40.1208752_-140.9871074_U5b3a1a
ItalySardinia12_40.1208752_-140.9871074_U5b3a1a
0.39
0.86
Spain10_40.463667_-153.74922_U5b3c
ItalySardinia13_40.1208752_-140.9871074_U5b3c
0.39
UnitedKingdom1_55.378051_-153.435973_U5b3e
0.42
0.26
ItalyCentralItaly16_45.9397505_-142.2462273_U5b3
0.51
0.53
Spain21_40.463667_-153.74922_U5b3f
0.17
ItalyCentralItaly9_45.9397505_-142.2462273_U5b3f
0.27
FranceSouthernFrance1_47.675928_-151.921652_U5b3
ItalyCalabria51_39.3087714_-133.6536209_U5b1_T16189C_
0.78
Belarus16_53.709807_-122.046611_U5b1_T16189C_T16192C_
Slovakia22_48.669026_-130.300976_U2e1a1
0.05
SpainBasque4_42.9896248_-152.6189273_U8a1a4
0.17
RussiaAltaiRepublic16_50.6181924_-63.7800692_K1a3
SpainBasque5_42.9896248_-152.6189273_U8a1a1a
0.79
0.06
SpainBasque3_42.9896248_-152.6189273_U8a2
RussiaBuryatRepublic37_54.8331146_-37.5939471_K1a17
SriLankaSinhalese1_5.9387708_-69.4220337_U2b2
0.03
Ukraine14_48.379433_-118.83442_U3b3
0.49
Lebanon26_33.854721_-114.137715_U3b3
0.62
0.13
Lebanon22_33.854721_-114.137715_U3b3
0.55
Lebanon4_33.854721_-114.137715_U3b2a
0.54
0.02
0.55
0.33
Lebanon8_33.854721_-114.137715_U3b2a1
Greece4_39.074208_-128.175688_U3b2a1a
Lebanon45_33.854721_-114.137715_U3a
0.82
CzechRepublic26_49.817492_-134.527038_U3a1
SpainAndalusia1_37.5442706_-154.7277528_U7b
0.6
ItalyIsleofElba6_42.7781867_-139.8072611_U7b1
0.01
0.91
ItalyIsleofElba9_42.7781867_-139.8072611_U7b1
0.91
0.48
ItalyIsleofElba5_42.7781867_-139.8072611_U7b1
Pakistan7_30.375321_-80.654884_U7b
0.01
ItalyIsleofElba4_42.7781867_-139.8072611_U7b1
0.36
0.03
ItalyIsleofElba3_42.7781867_-139.8072611_U7b1
0.47
0.32
ItalyIsleofElba2_42.7781867_-139.8072611_U7b1
0.03
0.09
ItalyIsleofElba8_42.7781867_-139.8072611_U7b1
ItalyIsleofElba7_42.7781867_-139.8072611_U7b1
0.09

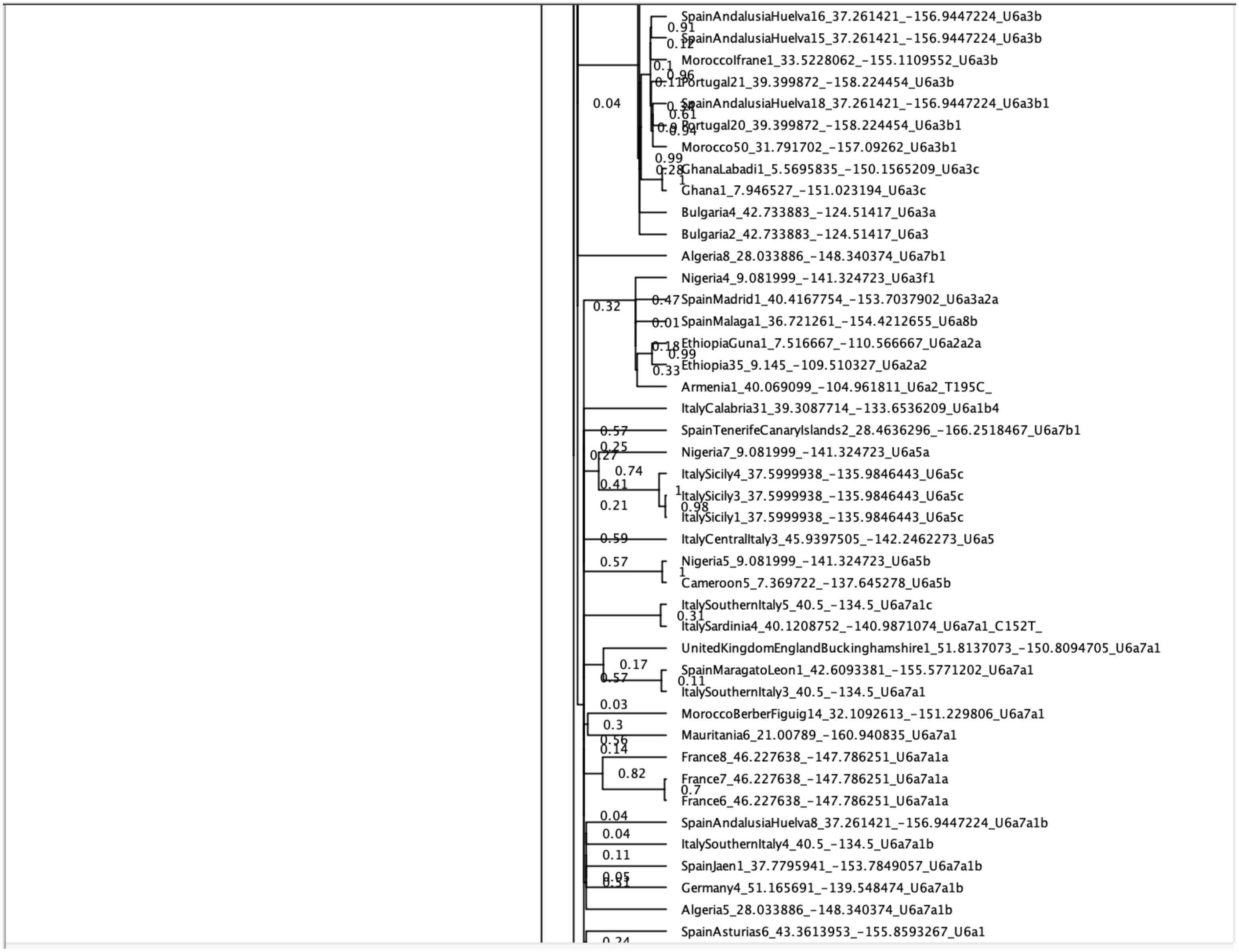
S3.47


0.09
ItalyIsleofElba7_42.7781867_-139.8072611_U7b1
0.06
ItalyIsleofElba1_42.7781867_-139.8072611_U7b1
IndiaMadhyaPradesh1_22.9734229_-71.3431058_U7a
Pakistan9_30.375321_-80.654884_U9b1
RussiaKemerovoRegion2_54.7574648_-62.5944712_U4b1b1_T16311C_
0.83
0.59
0.76
RussiaAltaiRepublic14_50.6181924_-63.7800692_U4b1b1_T16311C_
0.6
Russia15_61.52401_-44.681244_U4b1_T146C_T152C_
0.82
Denmark799_56.26392_-140.498215_U4a1a
Belarus36_53.709807_-122.046611_U4b1a2a
India17_20.593684_-71.03712_U9a1
RussiaTyvaRepublic1_61.8079151_-101.3043151_U4d2
0.92
0.22
0.92
RussiaYakutCentral84_52.7030792_-95.7607657_U4d2
0.37
RussiaKhakassRepublic1_53.0452281_-59.6017855_U4d2
CzechRepublic30_49.817492_-134.527038_U4a2d
0.36
0.38
Serbia12_44.016521_-128.994141_U4a2g
0.61
RussiaAltaiRepublic9_50.6181924_-63.7800692_U4a2a
0.63
Russia107_61.52401_-44.681244_U4a2a3
0.12
0.47
CzechRepublic18_49.817492_-134.527038_U4a2a
Lebanon44_33.854721_-114.137715_U4b3
0.05
CzechRepublic13_49.817492_-134.527038_U4c2a
0.15
Belarus11_53.709807_-122.046611_U4c1
NewZealand1_-40.900557_24.885971_U5a2a1
0.79
Germany5_51.165691_-139.548474_U5a1b3
0.89
RussiaBelgorod2_50.5997134_-113.4017379_U5a1a1a
0.29
1
CzechRepublic23_49.817492_-134.527038_U5a1a1a
Belarus17_53.709807_-122.046611_U5a2e
FranceSouthernFrance2_47.675928_-151.921652_U5b3a1b
Germany6_51.165691_-139.548474_U5b3b2
0.09
0.98
Spain20_40.463667_-153.74922_U5b3b1
0.26
0.94
0.09
CzechRepublic20_49.817492_-134.527038_U5b3b1
ItalySouthernItaly8_40.5_-134.5_U5b3c
0.42
0.06
Bulgaria3_42.733883_-124.51417_U5b3e
Croatia2_45.1_-134.8_U5b3
0.03
ItalySouthernItaly7_40.5_-134.5_U5b3g
0.3
0.56
Estonia1_58.595272_-124.986393_U5b3a2
0.66
BosniaandHerzegovina1_43.915886_-132.320924_U5b3
RussiaSiberia20_61.0137097_-50.8033441_U5b2a1a2
0.51
Poland106_51.919438_-130.854864_U5b1c2a
0.62
Slovakia2_48.669026_-130.300976_U5b1e1
SwedenSamiVasterbotten5_65.9676259_-134.0914358_U5b1b1a1
0.99
0.21
0.99
SwedenSamiNorrbotten3_66.8309216_-129.6008034_U5b1b1a1
Spain8_40.463667_-153.74922_U5b1b1d
0.04
0.68
MoroccoBerber5_31.6948716_-154.1556644_U5b1b1_T152C_

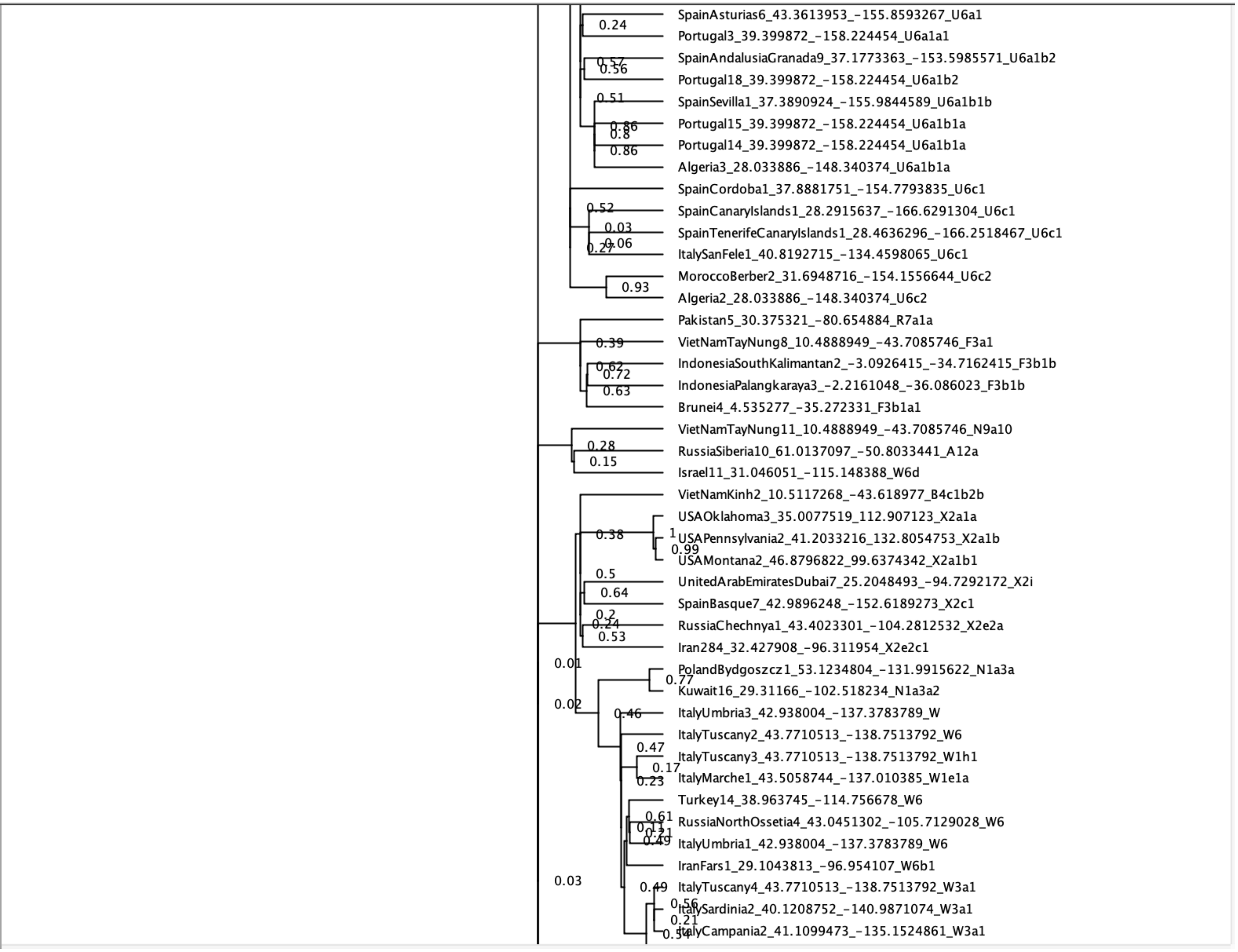
S3.48


0.04
0.68
MoroccoBerber5_31.6948716_-154.1556644_U5b1b1_T152C_
0.19
0.67
SwedenSamiNorrbotten6_66.8309216_-129.6008034_U5b1b1a3
SwedenSamiNorrbotten5_66.8309216_-129.6008034_U5b1b1a
0.64
Belarus13_53.709807_-122.046611_U5b1b1a
RussiaBuryatRepublic21_54.8331146_-37.5939471_U1a1a
0.65
Lebanon32_33.854721_-114.137715_U1b3
Tunisia56_33.886917_-140.462501_U6b1b
SpainTenerifeCanaryIslands3_28.4636296_-166.2518467_U6b1a
0.96
0.96
SpainCanaryIslands11_28.2915637_-166.6291304_U6b1a
0.96
SpainCanaryIslands12_28.2915637_-166.6291304_U6b1a
0.09
0.51
SpainGranCanariaCanaryIslands1_27.9202202_-165.5474373_U6b1a
0.04
SpainNavarra1_42.6953909_-151.6760691_U6b1a1
0.33
0.93
SpainCanaryIslands13_28.2915637_-166.6291304_U6b1a1
0.57
SpainCanaryIslands10_28.2915637_-166.6291304_U6b1a1
SenegalWolof1_13.5880762_-165.5778486_U6b
0.66
SpainGalicia2_42.5750554_-158.1338558_U6b
0.99
0.06
Portugal6_39.399872_-158.224454_U6b
0.03
Oman4_21.512583_-94.076745_U6b
0.57
0.08
Morocco43_31.791702_-157.09262_U6b3a
Tunisia67_33.886917_-140.462501_U6b
0.07
Cameroon6_7.369722_-137.645278_U6b
Portugal10_39.399872_-158.224454_U6d1
0.56
ItalySicily2_37.5999938_-135.9846443_U6d1b
SpainAndalusiaHuelva10_37.261421_-156.9447224_U6d3a
0.57
0.87
Portugal4_39.399872_-158.224454_U6d3a
0.55
0.57
MoroccoIfrane2_33.5228062_-155.1109552_U6d
Algeria6_28.033886_-148.340374_U6d2
UnitedKingdomScotland1_56.4906712_-154.2026458_U6a7a2a
Iraq19_33.223191_-106.320709_U6a4
SpainMalaga2_36.721261_-154.4212655_U6a1a1
0.5
0.07
SpainLeon5_42.5987263_-155.5670959_U6a1a1
0.93
SpainAndalusiaHuelva13_37.261421_-156.9447224_U6a1a1
0.09
SpainAndalusiaHuelva7_37.261421_-156.9447224_U6a1a1
0.13
0.08
0.03
Portugal1_39.399872_-158.224454_U6a1a1
0.17
ItalySouthernItaly11_40.5_-134.5_U6a1a1
Tunisia55_33.886917_-140.462501_U6a7c1
0.56
MoroccoBerberFiguig15_32.1092613_-151.229806_U6a7c1
0.2
0.04
Ethiopia49_9.145_-109.510327_U6a2b
0.55
ItalyCentralItaly5_45.9397505_-142.2462273_U6a2
0.55
0.59
Egypt9_26.820553_-119.197502_U6a2c
SpainAsturias7_43.3613953_-155.8593267_U6a6a1
0.62
SpainGalicia1_42.5750554_-158.1338558_U6a3e
SpainAndalusiaHuelva16_37.261421_-156.9447224_U6a3b
0.91

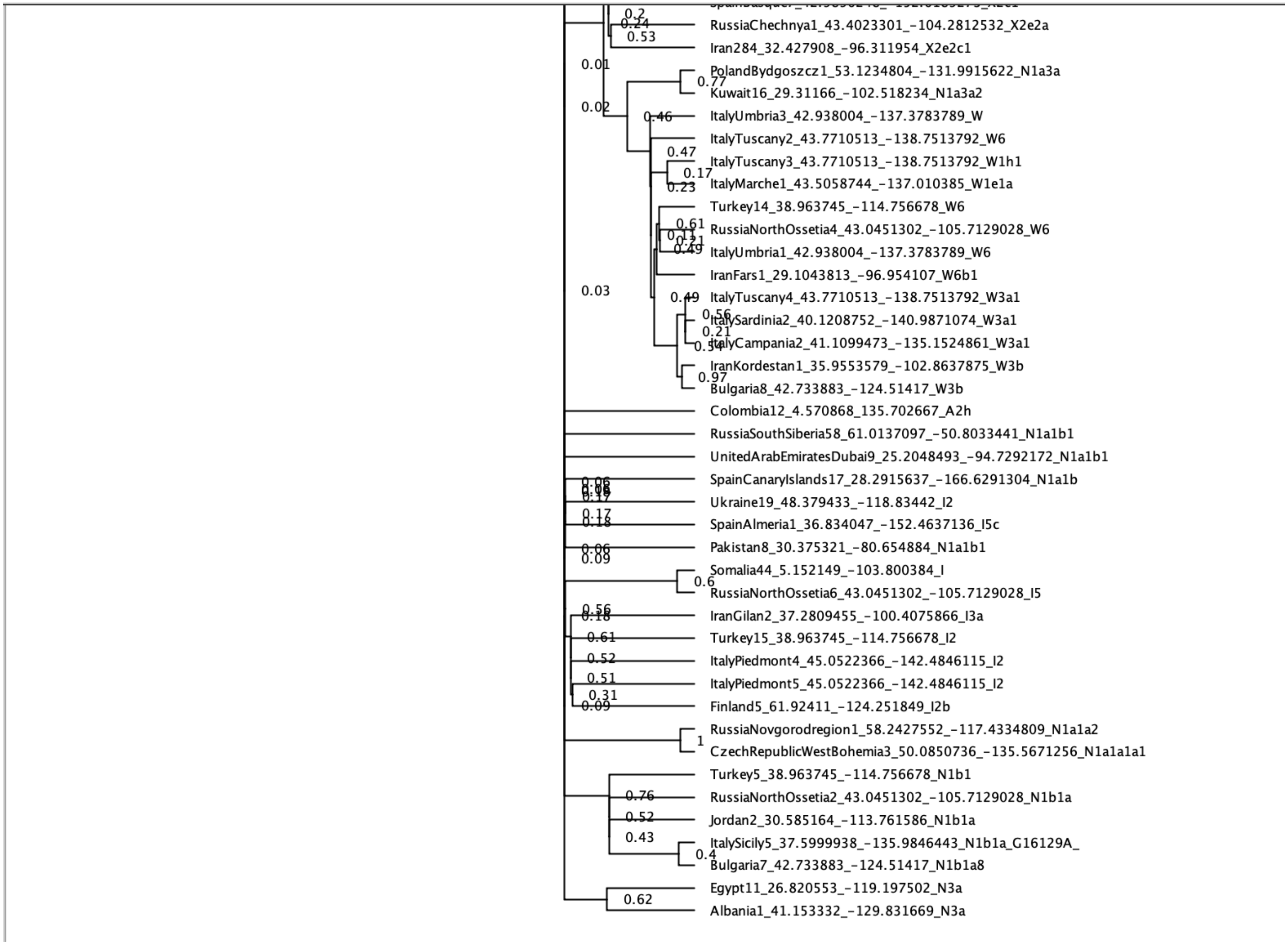
S3.49


SpainAndalusiaHuelva16_37.261421_-156.9447224_U6a3b
0.91
SpainAndalusiaHuelva15_37.261421_-156.9447224_U6a3b
0.12
MoroccoIfrane1_33.5228062_-155.1109552_U6a3b
0.1
0.96
0.11
Portugal21_39.399872_-158.224454_U6a3b
0.04
SpainAndalusiaHuelva18_37.261421_-156.9447224_U6a3b1
0.34
0.61
0.9
Portugal20_39.399872_-158.224454_U6a3b1
0.94
Morocco50_31.791702_-157.09262_U6a3b1
0.99
0.28
GhanaLabadi1_5.5695835_-150.1565209_U6a3c
1
Ghana1_7.946527_-151.023194_U6a3c
Bulgaria4_42.733883_-124.51417_U6a3a
Bulgaria2_42.733883_-124.51417_U6a3
Algeria8_28.033886_-148.340374_U6a7b1
Nigeria4_9.081999_-141.324723_U6a3f1
0.47
SpainMadrid1_40.4167754_-153.7037902_U6a3a2a
0.32
0.01
SpainMalaga1_36.721261_-154.4212655_U6a8b
0.18
EthiopiaGuna1_7.516667_-110.566667_U6a2a2a
0.99
Ethiopia35_9.145_-109.510327_U6a2a2
0.33
Armenia1_40.069099_-104.961811_U6a2_T195C_
ItalyCalabria31_39.3087714_-133.6536209_U6a1b4
0.57
SpainTenerifeCanaryIslands2_28.4636296_-166.2518467_U6a7b1
0.25
0.27
Nigeria7_9.081999_-141.324723_U6a5a
0.74
ItalySicily4_37.5999938_-135.9846443_U6a5c
0.41
1
ItalySicily3_37.5999938_-135.9846443_U6a5c
0.21
0.98
ItalySicily1_37.5999938_-135.9846443_U6a5c
0.59
ItalyCentralItaly3_45.9397505_-142.2462273_U6a5
0.57
Nigeria5_9.081999_-141.324723_U6a5b
1
Cameroon5_7.369722_-137.645278_U6a5b
ItalySouthernItaly5_40.5_-134.5_U6a7a1c
0.31
ItalySardinia4_40.1208752_-140.9871074_U6a7a1_C152T_
UnitedKingdomEnglandBuckinghamshire1_51.8137073_-150.8094705_U6a7a1
0.17
SpainMaragatoLeon1_42.6093381_-155.5771202_U6a7a1
0.57
0.11
ItalySouthernItaly3_40.5_-134.5_U6a7a1
0.03
MoroccoBerberFiguig14_32.1092613_-151.229806_U6a7a1
0.3
Mauritania6_21.00789_-160.940835_U6a7a1
0.56
0.14
France8_46.227638_-147.786251_U6a7a1a
0.82
France7_46.227638_-147.786251_U6a7a1a
0.7
France6_46.227638_-147.786251_U6a7a1a
0.04
SpainAndalusiaHuelva8_37.261421_-156.9447224_U6a7a1b
0.04
ItalySouthernItaly4_40.5_-134.5_U6a7a1b
0.11
SpainJaen1_37.7795941_-153.7849057_U6a7a1b
0.05
0.31
Germany4_51.165691_-139.548474_U6a7a1b
Algeria5_28.033886_-148.340374_U6a7a1b
SpainAsturias6_43.3613953_-155.8593267_U6a1
0.24

S3.50

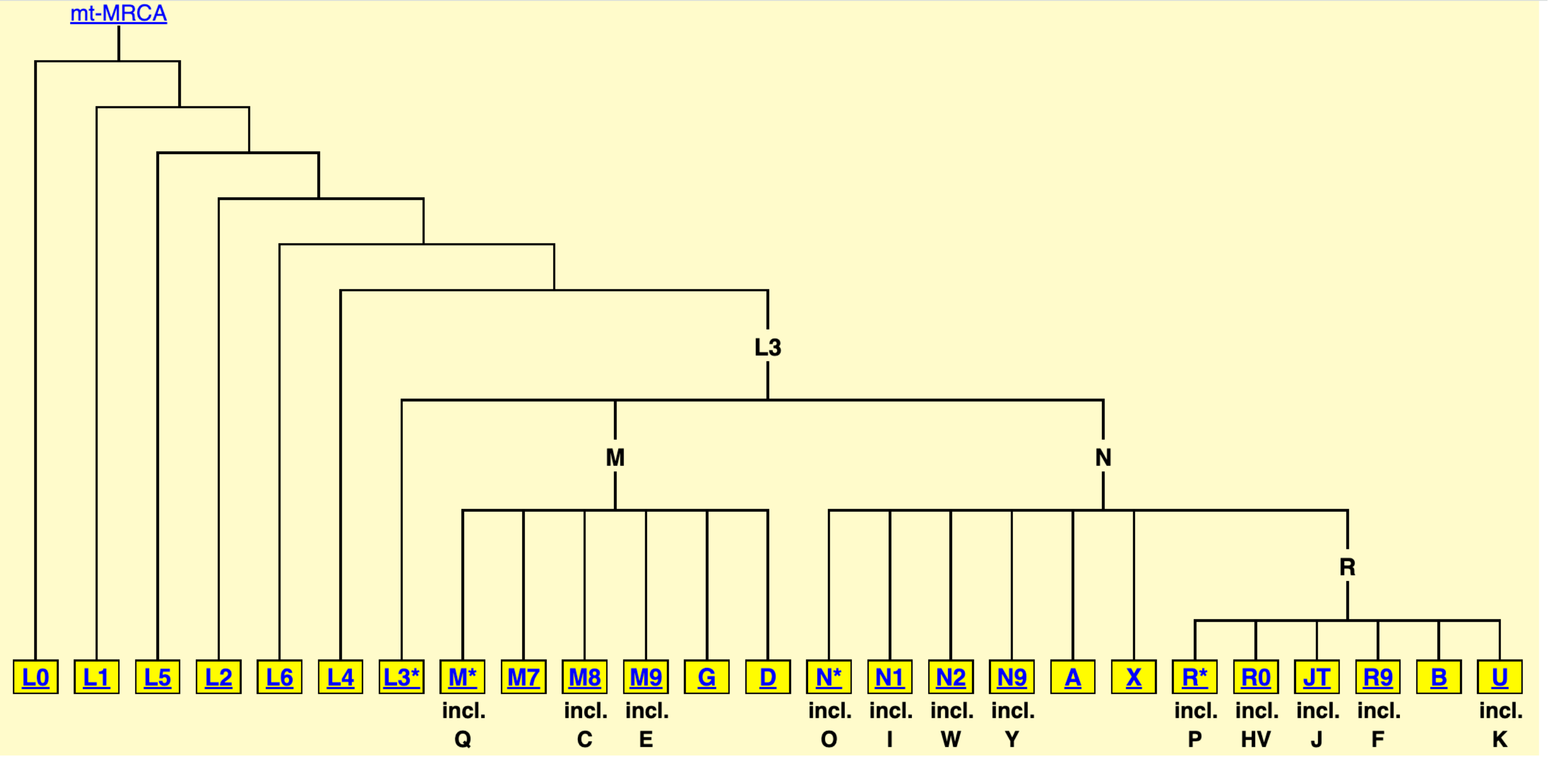

SpainAsturias6_43.3613953_-155.8593267_U6a1
0.24
Portugal3_39.399872_-158.224454_U6a1a1
0.57
0.56
SpainAndalusiaGranada9_37.1773363_-153.5985571_U6a1b2
Portugal18_39.399872_-158.224454_U6a1b2
0.51
SpainSevilla1_37.3890924_-155.9844589_U6a1b1b
0.86
0.8
Portugal15_39.399872_-158.224454_U6a1b1a
0.86
Portugal14_39.399872_-158.224454_U6a1b1a
Algeria3_28.033886_-148.340374_U6a1b1a
SpainCordoba1_37.8881751_-154.7793835_U6c1
0.52
SpainCanaryIslands1_28.2915637_-166.6291304_U6c1
0.03
SpainTenerifeCanaryIslands1_28.4636296_-166.2518467_U6c1
0.06
0.27
ItalySanFele1_40.8192715_-134.4598065_U6c1
MoroccoBerber2_31.6948716_-154.1556644_U6c2
0.93
Algeria2_28.033886_-148.340374_U6c2
Pakistan5_30.375321_-80.654884_R7a1a
0.39
VietNamTayNung8_10.4888949_-43.7085746_F3a1
0.62
0.72
IndonesiaSouthKalimantan2_-3.0926415_-34.7162415_F3b1b
0.63
IndonesiaPalangkaraya3_-2.2161048_-36.086023_F3b1b
Brunei4_4.535277_-35.272331_F3b1a1
VietNamTayNung11_10.4888949_-43.7085746_N9a10
0.28
RussiaSiberia10_61.0137097_-50.8033441_A12a
0.15
Israel11_31.046051_-115.148388_W6d
VietNamKinh2_10.5117268_-43.618977_B4c1b2b
USAOklahoma3_35.0077519_112.907123_X2a1a
0.38
1
USAPennsylvania2_41.2033216_132.8054753_X2a1b
0.99
USAMontana2_46.8796822_99.6374342_X2a1b1
0.5
UnitedArabEmiratesDubai7_25.2048493_-94.7292172_X2i
0.64
SpainBasque7_42.9896248_-152.6189273_X2c1
0.2
0.24
RussiaChechnya1_43.4023301_-104.2812532_X2e2a
0.53
Iran284_32.427908_-96.311954_X2e2c1
0.01
PolandBydgoszcz1_53.1234804_-131.9915622_N1a3a
0.77
Kuwait16_29.31166_-102.518234_N1a3a2
0.02
0.46
ItalyUmbria3_42.938004_-137.3783789_W
ItalyTuscany2_43.7710513_-138.7513792_W6
0.47
ItalyTuscany3_43.7710513_-138.7513792_W1h1
0.17
0.23
ItalyMarche1_43.5058744_-137.010385_W1e1a
Turkey14_38.963745_-114.756678_W6
0.61
0.11
0.21
RussiaNorthOssetia4_43.0451302_-105.7129028_W6
0.49
ItalyUmbria1_42.938004_-137.3783789_W6
IranFars1_29.1043813_-96.954107_W6b1
0.03
0.49
ItalyTuscany4_43.7710513_-138.7513792_W3a1
0.56
ItalySardinia2_40.1208752_-140.9871074_W3a1
0.21
0.54
ItalyCampania2_41.1099473_-135.1524861_W3a1

S3.51


0.2
0.24
RussiaChechnya1_43.4023301_-104.2812532_X2e2a
0.53
Iran284_32.427908_-96.311954_X2e2c1
0.01
PolandBydgoszcz1_53.1234804_-131.9915622_N1a3a
0.77
Kuwait16_29.31166_-102.518234_N1a3a2
0.02
0.46
ItalyUmbria3_42.938004_-137.3783789_W
ItalyTuscany2_43.7710513_-138.7513792_W6
0.47
ItalyTuscany3_43.7710513_-138.7513792_W1h1
0.17
0.23
ItalyMarche1_43.5058744_-137.010385_W1e1a
Turkey14_38.963745_-114.756678_W6
0.61
RussiaNorthOssetia4_43.0451302_-105.7129028_W6
0.11
0.21
0.49
ItalyUmbria1_42.938004_-137.3783789_W6
IranFars1_29.1043813_-96.954107_W6b1
0.03
0.49
ItalyTuscany4_43.7710513_-138.7513792_W3a1
0.56
ItalySardinia2_40.1208752_-140.9871074_W3a1
0.21
0.54
ItalyCampania2_41.1099473_-135.1524861_W3a1
IranKordestan1_35.9553579_-102.8637875_W3b
0.97
Bulgaria8_42.733883_-124.51417_W3b
Colombia12_4.570868_135.702667_A2h
RussiaSouthSiberia58_61.0137097_-50.8033441_N1a1b1
UnitedArabEmiratesDubai9_25.2048493_-94.7292172_N1a1b1
0.06
SpainCanaryIslands17_28.2915637_-166.6291304_N1a1b
0.17
Ukraine19_48.379433_-118.83442_I2
0.17
0.18
SpainAlmeria1_36.834047_-152.4637136_I5c
0.06
0.09
Pakistan8_30.375321_-80.654884_N1a1b1
Somalia44_5.152149_-103.800384_I
0.6
RussiaNorthOssetia6_43.0451302_-105.7129028_I5
0.56
0.18
IranGilan2_37.2809455_-100.4075866_I3a
0.61
Turkey15_38.963745_-114.756678_I2
0.52
ItalyPiedmont4_45.0522366_-142.4846115_I2
0.51
ItalyPiedmont5_45.0522366_-142.4846115_I2
0.31
0.09
Finland5_61.92411_-124.251849_I2b
RussiaNovgorodregion1_58.2427552_-117.4334809_N1a1a2
1
CzechRepublicWestBohemia3_50.0850736_-135.5671256_N1a1a1a1
Turkey5_38.963745_-114.756678_N1b1
0.76
RussiaNorthOssetia2_43.0451302_-105.7129028_N1b1a
0.52
Jordan2_30.585164_-113.761586_N1b1a
0.43
ItalySicily5_37.5999938_-135.9846443_N1b1a_G16129A_
0.4
Bulgaria7_42.733883_-124.51417_N1b1a8
Egypt11_26.820553_-119.197502_N3a
0.62
Albania1_41.153332_-129.831669_N3a

S3.52

Supplementary Figure S4: Primary branches of the MtDNA Phylogeny from Phylotree (image from https://www.phylotree.org/tree/index.htm, van Oven and Kayer 2009).



Summary

This thesis explores using language structures, such as grammatical and phonological features, to understand language history and human migration patterns. In particular it investigates language families in Mainland Southeast Asia, such as Austroasiatic, Tai-Kadai, and Sino-Tibetan, focusing on structural data from linguistic databases and fieldwork.

Chapter 1 introduces concepts such as linguistic areas (groupings of languages that are often unrelated but show similarities in their grammatical or phonological structures) and Bayesian phylogenetics, an approach from evolutionary biology for reconstructing the way that species are related to each other which has been adapted to studying the history of language families.

Chapter 2 examines the history of one particular linguistic feature that has a striking geographical distribution, namely lexical tone, the property of using pitch to differentiate words. Tonal languages are predominantly found in equatorial, humid climates such as in Africa and Southeast Asia. It is proposed that these clusters reflect both the history of large language families and the influence of language contact, where features like tone spread between different language families. A hypothesis explored in this section is that the complexity of tonal languages should decrease the further away they are from where tone was originally innovated in that region. This is because an expanding society speaking tonal languages would encounter populations speaking both tonal and non-tonal languages; when this happens, the number of contrasting tones typically decreases, because tones are difficult for second language learners to acquire if their native language is non- tonal (Gottfried and Suiter 1997), and they tend to simplify the tonal system. This hypothesis is plausibly borne out in a phylogenetic analysis of Sino-Tibetan and Niger-Congo as they expanded in their respective regions. This chapter also explores the hypothesis that climate influences the development of tonal languages (Everett, Blasi and Roberts 2015), arguing through the use of simulations that this additional explanation may not be needed once language contact is accounted for.

Chapter 3 employs a phylogenetic method to analyse data from the World Phonotactics Database to find how languages cluster based on 184 features, and using these to understand language history and the impact of language contact. This analysis reveals significant clusters across Eurasia, indicating interactions between unrelated languages in areas such as the Caucasus, India, and Southeast Asia. These clusters reveal language contact that is described in a non-quantitative way in previous studies such as Chirikba (2008) on the Caucasus, or Emeneau (1956) and Masica (1976) on India. Some larger linguistic areas also receive strong support in the analysis presented here, such as

Southeast Asia, Northern Eurasia, and the Caucasus/Middle-East. Like language families, these clusters can provide frameworks for understanding the history of Eurasia.

Chapter 4 examines recent language contact through fieldwork on East Palaungic languages in Southwest China in Chapter 4, revealing diversity in linguistic features influenced by nearby Tai-Kadai languages. There are differing numbers of Tai speakers in each Palaungic-speaking region: this is reflected in cultural differences between these regions, such as with Buddhist temples being mainly found further east in Tai-majority areas such as in Bada, Zhanglang and Mangjing; while traditional Wa symbols such as the bull skull is used on clothes and buildings further west, such as Ximeng, Wengding and Gongxin. The proportion of Tai speakers in each region is shown to predict some linguistic properties such as the use of Subject-Verb rather than Verb-Subject order, the use of the construction ‘eye of the day’ for ‘sun’, and the use of Tai numerals. In addition, data is presented on the surprising way that the verbs ‘eat’ and ‘drink’ vary across East Palaungic languages, in which languages vary from one extreme of having different verbs for different types of object of the verb ‘eat’ (e.g. ‘eat meat’, ‘eat rice’ and ‘eat vegetables’) to a more Tai-like system of using a single verb covering all types of eating and drinking.

Chapter 5 links linguistic structures with human migration patterns, using genetic data to reconstruct migrations and comparing this with linguistic similarity. The study uses mitochondrial DNA (mtDNA) data from GenBank (Benson et al. 2013) and Bayesian phylogeography to reconstruct common migration routes. A set of 2000 mtDNA sequences from 423 locations around the world was analysed using a spatial model implemented in BEAST by Bouckaert (2016). This chapter may be the first model-based reconstruction of mtDNA migration routes, despite the large literature on analysing the geographical distribution of mtDNA haplogroups (e.g. Harcourt 2016), and despite the use of Bayesian phylogeography for other domains (such as the transmission of viruses, Lemey et al. 2009, Magee et al. 2017; and the spread of languages, Bouckaert et al. 2012). The relationship between migration and languages is then examined, by modelling how they are correlated with each other and with geographical distance. The relationship between geographical distance and linguistic distance is first modelled, by using the results of the phylogenetic analysis of Eurasian languages in Chapter 3. The number of mtDNA migrations between linguistic communities is then included in the model, to find out whether they increase its predictive accuracy. It is found that there is a large increase in accuracy once migrations are included, in the case of predicting linguistic similarity from the clade-constrained phylogenetic analysis; but that there is no increase in accuracy in predicting linguistic similarity from the non-constrained phylogenetic analysis, which turns out to be very well correlated with geographical distance. The main conclusion of this is that, from one measure of linguistic similarity, linguistic areas have been shaped in part by migrations and not just by sharing of features by linguistic communities in close proximity.

In summary, this thesis provides insights into language history, with the main findings being that linguistic areas are statistically supported, geographically coherent clusters that emerge from phylogenetics of linguistic structures; and that language contact in some cases reflects the history of migration, as demonstrated on a local scale in the fieldwork study on how different features have been spreading with Tai speakers among the Palaungic languages, and on a continental scale in the correlation between migrations and linguistic similarity.

Samenvatting

Deze dissertatie onderzoekt het gebruik van taalstructuren, zoals grammaticale en fonologische kenmerken, om taalgeschiedenis en menselijke migratiepatronen te begrijpen. Het richt zich in het bijzonder op taalfamilies in Zuidoost-Azië, zoals Austroaziatisch, Tai-Kadai en Sino-Tibetaans, met de focus op structurele gegevens uit taalkundige databases en veldwerk.

Hoofdstuk 1 introduceert concepten zoals taalgebieden (groeperingen van talen die vaak ongerelateerd zijn maar overeenkomsten tonen in hun grammaticale of fonologische structuren) en Bayesiaanse fylogenetica, een benadering uit de evolutionaire biologie voor het reconstrueren van de manier waarop soorten aan elkaar gerelateerd zijn, wat is aangepast voor het bestuderen van de geschiedenis van taalfamilies.

Hoofdstuk 2 onderzoekt de geschiedenis van een specifieke taalkundige eigenschap met een opvallende geografische distributie, namelijk lexicale toon, de eigenschap van het gebruik van toonhoogte om woorden te onderscheiden. Tonale talen worden voornamelijk gevonden in equatoriale, vochtige klimaten zoals in Afrika en Zuidoost-Azië. Er wordt voorgesteld dat deze clusters zowel de geschiedenis van grote taalfamilies als de invloed van taalcontact weerspiegelen, waarbij kenmerken zoals toon zich verspreiden tussen verschillende taalfamilies. Een hypothese die in deze sectie wordt onderzocht, is dat de complexiteit van tonale talen zou moeten afnemen naarmate ze verder verwijderd zijn van waar toon oorspronkelijk werd geïnnoveerd in die regio. Dit komt omdat een uitbreidende samenleving die tonale talen spreekt, populaties zou tegenkomen die zowel tonale als niet-tonale talen spreken; wanneer dit gebeurt, neemt het aantal contrasterende tonen doorgaans af, omdat tonen moeilijk te verwerven zijn voor tweedetaalleerders als hun moedertaal niet-tonaal is (Gottfried en Suiter 1997), en ze hebben de neiging om het tonale systeem te vereenvoudigen. Deze hypothese wordt aannemelijk ondersteund in een fylogenetische analyse van Sino-Tibetaans en Niger-Congo terwijl ze zich uitbreidden in hun respectievelijke regio's. Dit hoofdstuk onderzoekt ook de hypothese dat klimaat de ontwikkeling van tonale talen beïnvloedt (Everett, Blasi en Roberts 2015), waarbij wordt betoogd via het gebruik van simulaties dat deze extra verklaring mogelijk niet nodig is zodra taalcontact wordt meegenomen.

Hoofdstuk 3 maakt gebruik van een fylogenetische methode om gegevens van de World Phonotactics Database te analyseren om te ontdekken hoe talen clusteren op basis van 184 kenmerken, en deze te gebruiken om taalgeschiedenis en de impact van taalcontact te begrijpen. Deze analyse onthult significante clusters door heel Eurazië, wat wijst op interacties tussen niet-verwante talen in gebieden zoals de Kaukasus, India en Zuidoost-Azië. Deze clusters onthullen taalcontact dat op een niet-kwantitatieve manier wordt beschreven in eerdere studies, zoals Chirikba (2008) over de Kaukasus, of Emeneau (1956) en Masica (1976) over India. Sommige grotere taalgebieden krijgen ook sterke ondersteuning in de analyse die hier wordt gepresenteerd, zoals Zuidoost-Azië, Noord-Eurazië en de Kaukasus/Midden-Oosten. Net als taalfamilies kunnen deze clusters kaders bieden voor het begrijpen van de geschiedenis van Eurazië.

Hoofdstuk 4 onderzoekt recent taalcontact door veldwerk over Oost-Palaungische talen in Zuidwest-China, waarbij diversiteit in taalkundige kenmerken wordt onthuld die zijn beïnvloed door nabijgelegen Tai-Kadai-talen. Er zijn verschillende aantallen Tai-sprekers in elke Palaungisch-sprekende regio: dit weerspiegelt zich in culturele verschillen tussen deze regio's, zoals met boeddhistische tempels die voornamelijk verder naar het oosten worden gevonden in Tai-meerderheidsgebieden zoals in Bada, Zhanglang en Mangjing; terwijl traditionele Wa-symbolen zoals de stierenkop gebruikt worden op kleding en gebouwen verder naar het westen, zoals in Ximeng, Wengding en Gongxin. Het aandeel Tai-sprekers in elke regio blijkt enkele taalkundige eigenschappen te voorspellen, zoals het gebruik van de Subject-Werkwoord in plaats van Werkwoord-Subject volgorde, het gebruik van de constructie ‘oog van de dag’ voor ‘zon’ en het gebruik van Tai-nummers. Daarnaast worden gegevens gepresenteerd over de verrassende manier waarop de werkwoorden ‘eten’ en ‘drinken’ variëren in Oost-Palaungische talen, waarin talen variëren van het ene uiterste van het hebben van verschillende werkwoorden voor verschillende soorten object van het werkwoord ‘eten’ (bijvoorbeeld ‘vlees eten’, ‘rijst eten’ en ‘groenten eten’) tot een meer Tai-achtig systeem van het gebruik van één werkwoord dat alle soorten eten en drinken omvat.

Hoofdstuk 5 koppelt taalstructuren aan menselijke migratiepatronen, waarbij genetische gegevens worden gebruikt om migraties te reconstrueren en dit te vergelijken met taalkundige gelijkenis. De studie maakt gebruik van mitochondriaal DNA (mtDNA) gegevens van GenBank (Benson et al. 2013) en Bayesiaanse fylogeografie om gemeenschappelijke migratieroutes te reconstrueren. Een set van 2000 mtDNA-sequenties van 423 locaties over de hele wereld werd geanalyseerd met behulp van een ruimtelijk model geïmplementeerd in BEAST door Bouckaert (2016). Dit hoofdstuk kan de eerste op modellen gebaseerde reconstructie zijn van mtDNA-migratieroutes, ondanks de grote literatuur over het analyseren van de geografische distributie van mtDNA-haplogroepen (bijvoorbeeld Harcourt 2016), en ondanks het gebruik van Bayesiaanse fylogeografie voor andere domeinen (zoals de overdracht van virussen, Lemey et al. 2009, Magee et al. 2017; en de verspreiding van talen, Bouckaert et al. 2012). De relatie tussen migratie en talen wordt vervolgens onderzocht, door te modelleren hoe ze met elkaar en met geografische afstand gecorreleerd zijn. De relatie tussen geografische afstand en taalkundige afstand wordt eerst gemodelleerd, door gebruik te maken van de resultaten van de fylogenetische analyse van Euraziatische talen in Hoofdstuk 3. Het aantal mtDNA-migraties tussen taalgemeenschappen wordt vervolgens opgenomen in het model, om te ontdekken of ze de voorspellende nauwkeurigheid ervan verhogen. Er wordt geconstateerd dat er een grote toename is in nauwkeurigheid zodra migraties worden opgenomen, in het geval van het voorspellen van taalkundige gelijkenis uit de clade-beperkte fylogenetische analyse; maar dat er geen toename in nauwkeurigheid is bij het voorspellen van taalkundige gelijkenis uit de niet-beperkte fylogenetische analyse, die blijkt zeer goed gecorreleerd te zijn met geografische afstand. De belangrijkste conclusie hiervan is dat, vanuit één maatstaf van taalkundige gelijkenis, taalgebieden gedeeltelijk zijn gevormd door migraties en niet alleen door het delen van kenmerken door taalgemeenschappen in nauwe nabijheid.

Samengevat biedt deze dissertatie inzichten in de taalgeschiedenis, waarbij de belangrijkste bevindingen zijn dat linguïstische gebieden statistisch ondersteunde, geografisch samenhangende clusters zijn die voortkomen uit de fylogenetica van taalstructuren; en dat taalcontact in sommige gevallen de geschiedenis van migratie weerspiegelt, zoals aangetoond op lokale schaal in de veldstudie over hoe verschillende kenmerken zich hebben verspreid onder de Tai-sprekers binnen de Palaungische talen, en op continentaal niveau in de correlatie tussen migraties en taalkundige gelijkenis.

Curriculum Vitae

Jeremy Collins (born 1987) obtained a BA(Hons) in Chinese at Wadham College, Oxford and MPhil in Linguistics from the University of Hong Kong in 2012, for which he received the Li Ka Shing Outstanding Research Student Award for best MPhil dissertation in the humanities. He worked in the Centre for Language Studies at Radboud University and the Max Planck Institute for Psycholinguistics in Nijmegen during his PhD from 2012 to 2016, and in the Glottobank Consortium as a designer of the Grambank database. He has worked in Hong Kong since 2018 in HSBC and other companies as a machine learning engineer, and is currently working as lead data engineer in Rida, a logistics company in Singapore.