



**HAL**  
open science

# Phylogenetic Perspectives on the Relative Importance of Identity by Descent versus Borrowing in the Lexica of Altaic Languages

Kenichi W Okamoto

► **To cite this version:**

Kenichi W Okamoto. Phylogenetic Perspectives on the Relative Importance of Identity by Descent versus Borrowing in the Lexica of Altaic Languages. 2021. hal-03241668

**HAL Id: hal-03241668**

**<https://hal.science/hal-03241668>**

Preprint submitted on 28 May 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Phylogenetic Perspectives on the Relative Importance of Identity  
by Descent versus Borrowing in the Lexica of Altaic Languages

Kenichi W. Okamoto

†*Department of Biology*

*University of St. Thomas, St. Paul MN 55105 USA*

May 17, 2021

## Abstract

The relationship among five North Eurasian language families - Turkic, Mongolic, Japonic, Korean and Tungusic - (so-called “Altaic” languages) has long troubled historical linguists. Although the hypothesis that these language families share a unique common ancestor languished, computational phylogenetics involving lexical datasets suggest some may share a common ancestor. Still, lexical similarity can readily arise from word borrowing; distinguishing potential loan words from genealogical relation is critical to assessing the Altaic hypothesis. Here we aim to fill this major gap by explicitly accounting for lexical borrowing to phylogenetically assess relationships among the Altaic language families. Using methods developed to compare phylogenetic networks with horizontal gene transfer, we statistically evaluate genealogical hypotheses in the presence of potentially extensive borrowing. Examining over 800 concepts in a well-attested database as unique etymological topologies, we find evidence of considerable lexical borrowing among the Altaic language families. Nevertheless, our results also suggest borrowing alone does not account for lexical similarity - in particular, Japonic and Korean likely share a common ancestor distinct from other families, and Mongolic is likely related to Tungusic or Turkic, but not both. We highlight implications of these results for historical linguists more broadly.

# 1 Introduction

2 Few ideas in the historical linguistics of Eurasia have proven as persistently controversial as the  
3 Altaic hypothesis (e.g., [1–14]). This hypothesis holds that the Turkic, Mongolic and Tungusic  
4 languages (a “micro-Altaic” grouping), and, occasionally, the Korean and Japonic languages  
5 (“macro-Altaic”), constitute a larger language family - that is, a group of languages which derive  
6 from a common ancestor language and which are more closely related to each other than any of  
7 them are to other languages ([15]). In its original formulation in the late 19<sup>th</sup> century, the Altaic  
8 hypothesis built upon two main arguments ([4]). First, that the five language families share a  
9 similar grammatical structure, especially as it regards their rules for modifying and situating  
10 verbs internally (“verb morphology”; - e.g., [12]). Second, early proponents of Altaic highlighted  
11 the lexical similarity across language families (e.g., [16]). When such correspondences reflect a  
12 shared, common ancestor for a word across multiple languages, the words are said to exhibit  
13 “genetic similarity”, as distinct from genetic similarity in the biological sense of sharing nucleic  
14 acid alleles ([17]).

15 Despite general early acceptance, the Altaic hypothesis came under widespread and sus-  
16 tained criticism in the latter half of the 20<sup>th</sup> century and the early 21<sup>st</sup> century (e.g., [1, 7,  
17 18, 19]). A major critique focuses on the purported lexical similarity ([1, 20]). In particu-  
18 lar, skeptics argue how words which appear akin to each other among the Altaic language  
19 families probably result from word borrowing, rather than a shared etymology, and that the  
20 languages seem to exhibit considerable word borrowings with each other (the Altaic *sprachbund*,  
21 or linguistic area, hypothesis - e.g., [18]). Several lines of evidence support this criticism. For  
22 instance, a close comparative reading of some of the earliest written records of Japonic (e.g.,  
23 the Man’yōshū poetry collection), Turkic (the Orkhon scripts) and Mongolic (the Mongγol-un  
24 niγuča tobčiyān, or the Secret History of the Mongols) appeared to suggest less, rather than  
25 greater, lexical similarity as would be expected if lexical similarity reflected common ancestry  
26 ([1, 4, 20]). Additionally, alleged word similarities have been shown to be inconsistent across  
27 and within some language families ([20, 21]). Were the words to share a genealogical origin,  
28 such discordant patterns of cross-familial similarities would not be expected.

29 Nevertheless, there is growing reexamination of genetic similarities in the lexica of the

30 Altaic languages ([9, 13, 22–24]), spurred by increased acceptance of automated, computational  
31 methods among historical linguists as a tool for examining language change ([25–29]; see also [30]  
32 for a comprehensive review, including a discussion of some of the limitations of this approach).  
33 This shift follows the considerable success of computational phylogenetics in the life sciences.  
34 The ability to reconstruct evolutionary trajectories of biological populations from nucleic and  
35 amino acid sequence data has been leveraged to tackle a wide range of long standing questions,  
36 such as reconstructing the proteins used by the very earliest life forms (e.g., [31]), quantifying  
37 the extent of Neanderthal DNA in modern human lineages (e.g., [32]) and mapping disease  
38 spread ([33, 34]).

39 To the extent that words in a language can be analogized to nucleic acid sequences in  
40 a genome, one can apply these increasingly sophisticated molecular phylogenetic methods to  
41 historical linguistics (e.g., [29, 35–37]). Manually comparing several languages from different  
42 families is a daunting task that requires considerable specialized expertise ([37]). By contrast,  
43 incorporating the tools of computational phylogenetics allows replicable, consistent and rela-  
44 tively fast results even among researchers not intimately familiar with the history of each of the  
45 languages they examine ([28]). Yet perhaps one of the strongest advantages of this approach  
46 to historical linguistics is the ability to use common metrics to evaluate competing hypotheses.  
47 Proposed language relations, like biological relations among taxa, are, at their core, hypotheses  
48 about what happened in the past. Modern phylogenetics provides a systematic framework for  
49 testing and comparing such hypotheses ([38–42]).

50 Briefly, these analyses begin with a database of words and meanings from different languages.  
51 They then evaluate multiple hypothesized language phylogenies by comparing how consistently  
52 language families are classified together in each phylogeny based on their lexical similarity. The  
53 application of computational approaches to reconstruct language families suggests qualified  
54 support for genetic relatedness among at least some of the Altaic language families (especially  
55 among the micro-Altaic families - [9, 13, 14, 24, 26, 29]).

56 Nevertheless, one limitation of this approach is that if there is extensive borrowing between  
57 language families, then it is not always clear that the fact that languages were grouped together  
58 on the basis of lexical similarity implies that they are genetically related to each other (e.g.,  
59 [43, 44]). If the computed phylogenies show that the Altaic languages exhibit a high degree

60 of lexical similarity to each other relative to other languages, this pattern could still result  
61 from extensive word borrowing. To be sure, a hypothesis of genetic relatedness among the  
62 Altaic languages is also potentially compatible with rampant borrowing among the language  
63 families (as occurs among several Indo-European languages - [45, 46]). This question can only  
64 be resolved by characterizing the relative probabilities that (i) lexical similarity in Altaic results  
65 from borrowing within the group, (ii) lexical similarity results from genetic relatedness, or (iii)  
66 both. To our knowledge, a systematic, quantitative comparison of these possibilities for Altaic  
67 has never been conducted.

68 Here we begin to address this major gap. As has been observed by others (e.g., [46]),  
69 bacterial phylogenetics presents an analogous problem of partitioning the effects of borrowing  
70 from identity by descent. Bacteria populations can exchange genetic material with distantly  
71 related populations through a process known as horizontal, or lateral gene transfer ([47–49]).  
72 Thus, a high degree of nucleic acid sequence similarity among bacteria taxa still does not  
73 conclusively establish phylogenetic relatedness. To address this issue, there is a rapidly growing  
74 literature in molecular phylogenetics on incorporating lateral gene transfer into inferences about  
75 phylogenies ([50–56]). Similar inferential difficulties arise when there is hybridization between  
76 Eukaryotic lineages (reviewed in, e.g., [57]), and this has spurred an interest in techniques  
77 that can test phylogenetic hypotheses by comparing reticulating phylogenetic networks from  
78 molecular sequencing data ([58–60]). Our aim is to apply some of these methods to evaluate  
79 the relative role of word borrowing, on the one hand, and genetic similarity, on the other, in  
80 structuring the relationships of the Altaic vocabularies to each other and to other vocabularies.

## 81 **Materials and Methods**

82 We collated word lists for all languages in each of the five purported Altaic language families  
83 that were available in NorthEuraLex v. 0.9 ([61]). NorthEuraLex is a lexicostatistical database  
84 of over 100 Eurasian languages, with lists of over 1000 concepts. These word lists form the  
85 basis of our analysis described below.

86 Testing the genetic relatedness of Altaic languages involves assessing whether the language  
87 families are more related to each other than any of them are to other languages. To quantify

88 the relatedness of the Altaic language families relative to other languages, we also included an  
 89 outgroup into our phylogenetic analyses, using Dravidian, a well-established language family. In  
 90 addition to being a well-documented language family in NorthEuraLex v0.9, one advantage of  
 91 using Dravidian as an outgroup is that although it is widely regarded to be genetically unrelated  
 92 to Japonic, Korean, Mongolic, Turkic or Tungusic, (but see [62, 63] viz. Japonic and Korean)  
 93 we think it is plausible that of the language families in NorthEuraLex v. 0.9, Dravidian is likely  
 94 to be one of the larger language families to have had among the least amount of direct loan  
 95 words into or from the Altaic languages (in comparison to, say, Uralic or Sino-Tibetan). Table  
 96 1 lists all the languages used in the analyses.

Table 1	
Language Family	Languages
Turkic	Turkish, North Azeri, Kazakh, Bashkir, Tatar, Sakha, Chuvash, Southern Uzbek
97 Mongolic	Khalkha Mongolian, Russian Buriat, Kalmyk
Tungusic	Evenki, Manchu, Nanai
Japonic	Japanese
Korean	Korean
Dravidian (outgroup)	Kannada, Malayalam, Tamil, Telugu

98 Table 1. Languages used in the analyses. For Turkic, the languages were further grouped by subfamily  
 99 [61] in all analyses.

100 For all languages, the collated word lists were converted into a QLC-formatted table using  
 101 a custom R ([64]) script, and the resulting QLC tables were pre-processed using the `Lingpy`  
 102 `Python` library ([37, 65]) to create distance matrices for all the words in the QLC table. Briefly,  
 103 the QLC table was imported into `Python` as a `LexStat` object, and we used the `get_scorer`  
 104 method to characterize the transition probabilities among sound classes for the languages ([65]).  
 105 For each concept in the database, we then called the `align_pairs` function to perform a pairwise  
 106 Sound-Class Based Phonetic Alignment (SCA; [66]) for each pair of words across two languages  
 107 using the `lexstat` method ([65]). The resulting alignment score was used as our measure of  
 108 pairwise distance among words in different languages, enabling us to construct a distance matrix  
 109 for all concepts. For each word, the languages were then clustered according to this similarity  
 110 measure separately for each concept using the unweighted pair group method with arithmetic  
 111 mean (UPGMA; [67]) implemented in the `upgma` function from the `phangorn` package in R ([68]).  
 112 By treating the historical trajectory of each word as it’s own phylogenetic tree, we could group  
 113 the source and destination language of a borrowed word together even if those languages are

114 genetically distinct. Put differently, this process enables us to characterize how each language’s  
 115 word has a lineage distinct from the historical development of the language as a whole.

116 We then generated a series of progressively more complex topologies involving a range of  
 117 genealogical scenarios across the language families. Table 2 illustrates the criteria used for  
 118 generating the hypothesized phylogenetic relationships; to facilitate intuition, Supplementary  
 119 Figure S1 sketches the genealogical hypotheses examined.

Table 2

Scenario	Characteristic	Possible values
	genealogical relationship among micro-Altaic language families	Present/Absent
	genealogical groupings among micro-Altaic language families	Mongolic and Turkic/Mongolic and Tungusic/Tungusic and Turkic/Absent
120	genealogical relationship among macro-Altaic language families	Present/Absent/Microaltaic nested within macro-Altaic/micro-Altaic as own language family with Japonic, Korean as language isolates/micro-Altaic as own language family with Japonic, Korean as own, separate language family
	genealogical groupings and subsets of macro-Altaic language families	Japonic and micro-Altaic subnetwork/Japonic and Korean/ Korean and micro-Altaic subnetwork/Absent

121 Table 2. Criteria used to guide construction of alternative phylogenetic networks among the language  
 122 families. The phylogenetic genealogies we evaluated are permutations of the above scenarios. For in-  
 123 stance, a hypothesized evolutionary trajectory where all macro-Altaic languages are genetically related,  
 124 but are equally divergent from each other, will assume the micro-Altaic groupings and genealogical re-  
 125 lationships to be absent, and while a genealogical relationship among macro-Altaic languages will be  
 126 present, genealogical subgroups will also be absent.

127 For analyses that excluded and included Japonic and Korean, the `lebor` plugin ([69]) to `lingpy`  
 128 was used to identify cognates within each micro-Altaic language family and within the outgroup  
 129 using the `internal.cognates` function with a threshold value of 0.5, as in the case study  
 130 presented on `lebor`’s `github` page ([69]). External cognate detection was then conducted  
 131 using `lebor`’s `external.cognates` function, this time with a threshold value of 0.3 and the  
 132 method set to “parsody”, again following the case study example. Once potential cognates were  
 133 identified, for each scenario described in Table 2, a minimal lateral network (MLN; [46, 70])  
 134 was constructed using `lingpy`’s `PhyBo` class’s `analyze` and `get_MLN` methods using majority  
 135 rule. To keep subsequent analyses computationally tractable, an edge-exclusion threshold of 10  
 136 suspected borrowing events was applied. Each MLN was generated assuming a phylogenetic  
 137 tree whose topological structure follows the variations described in Table 2; effectively, this  
 138 allowed differing borrowing scenarios to be grafted onto each hypothesized language genealogy  
 139 to construct the final network.



140 To evaluate the relative likelihoods of differing scenarios involving borrowing and identity  
141 by descent among Altaic languages, we used **Phylonet** ([52, 71]), a tool for inferring and quan-  
142 tifying the likelihoods of phylogenetic networks. **Phylonet** enables calculating the likelihood of  
143 network topologies which involve reticulated evolutionary events (such as horizontal gene trans-  
144 fer) between branches of a phylogenetic tree. The likelihood for each MLN was evaluated using  
145 the **CalGTProb** function from **Phylonet** ([72–74]). Briefly, **CalGTProb** calculates the probability  
146 that the observed array of gene trees (i.e., a representation of the evolutionary history of a  
147 particular gene) and their topologies result from a hypothesized phylogenetic network. In our  
148 context, the UPGMA-classified word-specific tree constitutes a gene tree. Thus, we evaluated  
149 the likelihood that the array of our observed word trees resulted from the MLN describing both  
150 genetic relatedness and borrowing among the Altaic language families. We further specified  
151 **CalGTProb** to optimize the branch lengths and inheritance probabilities of each network, and  
152 set the maximum number of rounds for branch-length optimization to 1000.

153 Our assessment of the phylogenetic network likelihoods involved two further steps. We eval-  
154 uated the likelihoods under two variants of the Altaic hypothesis (Table 2). The first variant  
155 describes genetic relatedness among Mongolic, Tungusic and Turkic, but is non-committal about  
156 whether any of them are related to Japonic and Korean (the “micro-Altaic” hypothesis). The  
157 second posits all five language families belong to a single lineage (the “macro-Altaic” hypothe-  
158 sis). Thus, we structured likelihood evaluations under each variant as follows. We first assessed  
159 the phylogenetic networks among the three micro-Altaic language families to identify potential  
160 sister families (i.e., two language families more closely related to each other than either are to  
161 the other language families analyzed) and whether all three micro-Altaic languages derive from  
162 a more recent common ancestor language (i.e., form a clade) than the other language families  
163 (Table 2; Suppl. Fig S1a-S1h). We then nested the resulting micro-Altaic network with the  
164 highest likelihood into our analyses of macro-Altaic (Table 2; Suppl. Fig. S1i-S1l).

165 All scripts and code (including the Rich Newick-formatted phylogenetic network input files;  
166 [75]), used in the analysis are publicly available at [github.com/kewok/altaic](https://github.com/kewok/altaic) and released  
167 under the GNU Public License ([76]).

168 **Results**

169 Table 3 summarizes the total log probability for each of the genealogical hypotheses among the  
 170 micro-Altaic languages.

Table 3

Hypothesis about micro-Altaic Language Families	Hypothesized Sister Families	Total Log Probability
Genetic relationship absent		
All language families unre-	lated	<b>8182.82</b>
Only Mongolic and Turkic	related	<b>10335.73</b>
Only Mongolic and Tungu-	sic related	<b>10335.74</b>
Only Turkic and Tungusic	related	<b>8249.65</b>
Genetic relationship present		
No structure within micro-	altaic	<b>9977.38</b>
Mongolic/Tungusic as sister	taxa	<b>9890.62</b>
Mongolic/Turkic as sister	taxa	<b>9890.73</b>
Turkic/Tungusic as sister	taxa	<b>7077.37</b>

171  
 172 Table 3 The log-likelihoods for each language network for the micro-Altaic languages at select quintiles (color)  
 173 with Dravidian language families as the outgroup. The color scheme represents the hypothesized network’s log-  
 174 likelihood percentile across all hypothesized language networks (ex-Japonic and Korean) from lowest (coolest  
 175 color) to highest (warmest color).

176 The analyses suggest strongest support for two of the three micro-Altaic families being genet-  
 177 ically related despite extensive borrowing. Yet no undisputed pair of sister taxa among the  
 178 micro-Altaic language families emerges. In particular, the total log probability of Mongolic  
 179 and Tungusic constituting a single sister group, with Turkic being unrelated, is almost identi-  
 180 cal to the total log probability of Mongolic and Turkic constituting a single sister group, with  
 181 Tungusic being unrelated (Table 3).

182 Thus, we conducted our analyses including all macro-Altaic families (Japonic, Korean, Mon-  
 183 golic, Tungusic and Turkic) under both hypotheses with the highest total log probabilities in  
 184 Table 3. Table 4 summarizes the total log probabilities for each of the hypotheses.

Table 3

Hypothesized Relationship among micro-Altaic Language Families	Hypothesized Relationship	Log-Likelihood
--	---------------------------	----------------

Mongolic/Tungusic Sister Taxa and Turkic Unrelated	Japonic/Korean Sister Taxa Grouped with Mongolic/Tungusic	<b>11803.47</b>
	Japonic, Korean Unrelated but Grouped with Mongolic/Tungusic	<b>4879.55</b>
	Japonic/Korean Sister Taxa Unrelated to Mongolic/Tungusic	<b>12782.26</b>
	Japonic, Korean Unrelated to each other and Mongolic/Tungusic	<b>10012.05</b>
Mongolic/Turkic Sister Taxa and Tungusic Unrelated	Japonic/Korean Sister Taxa Grouped with Mongolic/Turkic	<b>10954.64</b>
	Japonic, Korean Unrelated but Grouped with Mongolic/Turkic	<b>5062.79</b>
	Japonic/Korean Sister Taxa Unrelated to Mongolic/Turkic	<b>11255.98</b>
	Japonic, Korean Unrelated to each other and Mongolic/Turkic	<b>1583.47</b>

185

186 Table 4 The log-likelihoods for each language network for the Macro-Altaic languages at select quintiles (color).  
 187 As in Table 3, Dravidian language families are the outgroup and the color scheme represents the hypothesized  
 188 network’s loglikelihood percentile across all hypothesized language networks from lowest (coolest color) to highest  
 189 (warmest color).  
 190 The key result is that regardless of whether Mongolic forms a sister group with Tungusic or  
 191 Turkic, Japonic and Korean are less likely to form a genetic grouping with the micro-Altaic  
 192 sister group. Moreover, there is a higher total log probability that Japonic and Korean share  
 193 a more recent common ancestor than any do with the other languages. Figure 1 illustrates the  
 194 hypothesized networks with the highest total log probabilities.

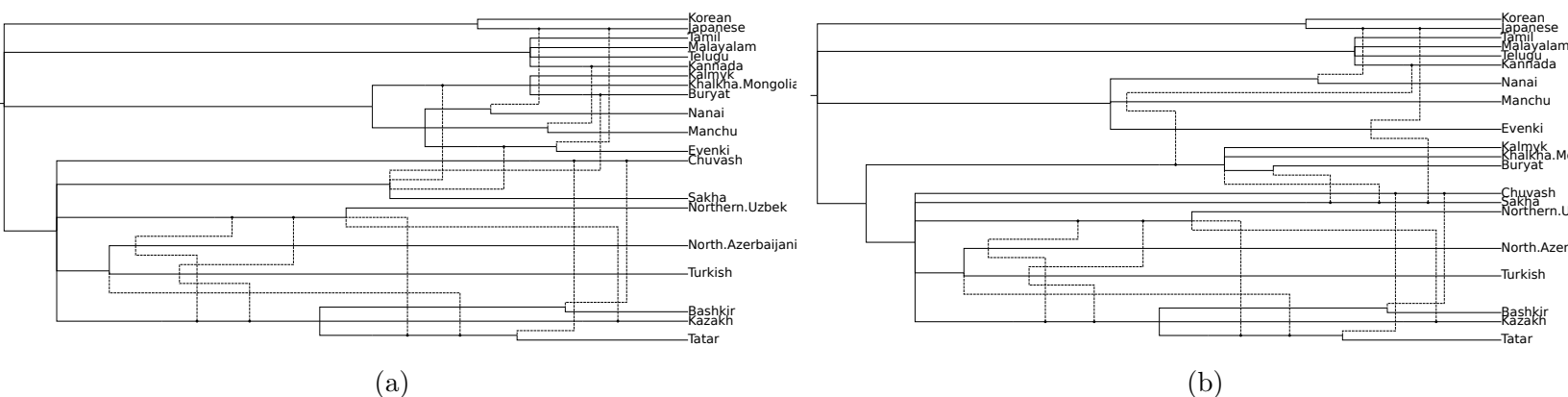


Figure 1: Phylogenetic network among “Altaic” and Dravidian (outgroup) languages with the highest total log-likelihoods when (a) Mongolic groups with Tungusic and (b) Mongolic groups with Turkic; generated via IcyTree ([77]). Vertical lines represent more than 10 word borrowings identified by the phonetic alignment.

## 195 Discussion

196 As all human languages likely had a single common ancestor ([78]), the Altaic languages are  
 197 of course genetically related in a certain sense. And given the demographic history of Altaic  
 198 speakers (e.g., [79–81]), it also strikes us as plausible that these languages are more closely  
 199 related to each other than any of them are to, say, the Khoe-Kwadi languages of southern  
 200 Africa. Thus, we propose that the key question may not be whether the lexical similarities in  
 201 Altaic are a result of genetic relatedness or borrowing; rather, the issue is how much of this  
 202 lexical similarity reflects borrowing, how much reflects descent from a common ancestor, and

203 how much is mere chance correspondence.

204 Among lexical similarities reflecting a common ancestor, the question then becomes whether  
205 there are likelier to be more such similarities among the Altaic language families than among any  
206 of those languages families and other language families in Eurasia. This question is inherently  
207 quantitative, and phylogenetic methods that can systematically integrate the probabilities of  
208 genetic relatedness and borrowing provide one way forward.

209 Our main results (summarized in Fig. 1) suggest the following conclusions. First, based on  
210 word-tree topology, borrowing alone seems less likely to account for lexical similarity among the  
211 Altaic language families. Nevertheless, Tungusic and Turkic are likely genetically unrelated.  
212 An important implication of this result is that borrowing appears likelier to account for lexical  
213 similarity between these two language families. Second, Mongolic is likely to be a sister language  
214 family to either Tungusic or Turkic (a micro-Altaic sister group), but not both. This implies  
215 that while there may be considerable borrowing between Mongolic, Turkic and Tungusic, there  
216 is evidence for genetic relatedness between Mongolic and one of the other micro-Altaic language  
217 families even after accounting for potential borrowing. Finally, our results suggest that Japonic  
218 and Korean, while related to each other, are less likely to exhibit a genetic relationship with  
219 a micro-Altaic sister group. Thus, our results support the likelihood that lexical similarity  
220 between Japonic and Korean, on the one hand, and the micro-Altaic language families, on the  
221 other, is primarily due to borrowing, rather than identity by descent.

222 Our analyses leave open the important issue of whether Tungusic or Turkic is much likelier  
223 to share a common ancestor with Mongolic. We suggest some potential approaches to resolving  
224 this question. Adopting a more extensive word list might be one approach, although the  
225 almost identical total log probabilities mean that insight from including more words, particularly  
226 beyond commonly used words, may prove marginal. Including more Mongolic and Tungusic  
227 languages may therefore prove more fruitful. A major advantage of our computational approach  
228 is that it readily allows iteratively refining hypotheses as more data (words and languages) are  
229 incorporated. Advances in phylogenetic analyses extending beyond lexical datasets (e.g., [82])  
230 can also potentially provide a complementary datapoint that sheds light on this question.

231 Another particularly promising approach would be to compare robustly reconstructed proto-  
232 Mongolic, proto-Tungusic and proto-Turkic languages (e.g., [11]). In principle, the same method-

233 ological approach we present could be used on the proto languages - MLNs could be identified for  
234 hypotheses grouping proto-Mongolic with proto-Tungusic or proto-Mongolic with proto-Turkic,  
235 respectively, and the resulting MLNs could then be scored using the topological structure that  
236 exists among the proto-words for the three languages. Our results further suggest that proto-  
237 Japonic or proto-Korean, or even a proto language for the two families (e.g., [83]), may serve  
238 as a candidate outgroup.

239 A highly speculative hypothesis that might, in theory, account for the very similar log like-  
240 lihood probabilities linking Mongolic to Turkic and to Tungusic is that Mongolic might be a  
241 creole derived from the other two language families. If two unrelated language families led to the  
242 formation of a creole, then this would, in principle, be consistent with the result that Mongolic  
243 appears equally likely to be related to the two unrelated language families. To be sure, assessing  
244 such a hypothesis in earnest is beyond the scope of the current study, and much more work  
245 examining Mongolic from the lens of creole historical linguistics (e.g., [84–86]) would be needed  
246 before the hypothesis can be taken seriously. From a computational perspective, identifying  
247 hybrid origins, rather than lateral gene transfer following earlier divergence, is a notoriously  
248 challenging problem in phylogenetic systematics (e.g., [87]), although there is increasing appre-  
249 ciation for the potential of phylogenetic tools to study historical creole linguistics (e.g., [88, 89]).  
250 For now, we merely wish to raise it as one possible hypothesis for why we found Mongolic to ap-  
251 pear equally likely to share a genealogical relationship with two apparently unrelated language  
252 families.

253 Like other computational methods in historical linguistics, our approach enables repro-  
254 ducible comparisons. Nevertheless, a caveat is that the results rely upon an automated char-  
255 acterization of phylogenetic trees for individual words. Using alignment strategies that aim  
256 to reliably identify cognates across languages, as we have done here, can constrain the ad-  
257 verse effects of chance phonetic alignments among words for subsequent analyses ([66]). Yet no  
258 tractable, automated method for grouping words based on their phonetic profiles can defini-  
259 tively differentiate chance correspondence from cognates (whether borrowed or from a common  
260 ancestor) for every word. Ultimately, only a close, humanistic examination of the historical, an-  
261 thropological and/or literary records can reconstruct robust etymological profiles ([7]). To the  
262 extent that our subsequent characterizations of the likelihoods for each network depend on the

263 phylogenetic topology of the underlying words, the grouping of words from distant languages  
264 may, at least in some instances, reflect mere coincidence rather than borrowing or common  
265 ancestry (as may have occurred, despite our high borrowing threshold, between Manchu and  
266 Kannada).

267 Despite this limitation, on balance we argue that the benefits of a systematic, tractable  
268 approach to testing hypotheses about the interplay between borrowing and genealogy should  
269 not be discounted. An analogous problem arises in the comparison of biological sequence data,  
270 where chance mutations can lead to similar nucleic acid sequences in particular gene families  
271 across unrelated taxa. Yet biologists routinely use sequence alignments across loci to calculate  
272 likelihoods for evaluating phylogenetic hypotheses, including hypotheses about lateral gene  
273 transfer. Systematists have long recognized how including more gene trees in their analyses  
274 can mitigate the effects of spurious concordances ([52, 54, 74]). Our use of a very large, well-  
275 curated dataset of over 800 words ([61]) helps reduce the effect that any one chance phonetic  
276 correspondence leads to a grouping of words across languages that distorts our overall likelihood  
277 calculations.

278 Finally, we highlight that although the present study focuses on the alleged Altaic language  
279 families, the underlying approach enables testing different phylogenetic hypotheses while also  
280 accounting for potential borrowing. We therefore think our method can prove useful in other  
281 Sprachbunds where distinguishing genealogy from borrowing has proven challenging such as  
282 among pre-Colombian North American languages of the Pacific Northwest.

## 283 **Acknowledgments**

284 I would like to thank C. Heisler and T. Stickler for assistance with earlier rounds of data  
285 curation, D. F. Morales-Briones for valuable discussion on phylonet, and the Minnesota Super-  
286 computing Institute at the University of Minnesota for allowing me to run the analyses on their  
287 clusters.

## References

- [1] Gerard Clauson. The case against the Altaic theory. *Central Asiatic Journal*, pages 181–187, 1956.
- [2] Roy Andrew Miller. Anti-Altaicists contra Altaicists. *Ural-altaische Jahrbucher I Ural-Altaic Yearbook*, pages 5–62, 1991.
- [3] T. I. Mills. Var mi, yok mu? (“Does it or doesn’t it exist?”): the Altaic dilemma (or: Aru, nai?). *Calgary Working Papers in Linguistics*, 20(Winter):55–72, 1998.
- [4] Stefan Georg, Peter A. Michalove, Alexis Manaster Ramer, and Paul J. Sidwell. Telling general linguists about Altaic. *Journal of Linguistics*, 35(1):65–98, 1999.
- [5] Schönig C. Turko-Mongolic relations. In *The Mongolic Languages*, page 18. Routledge, London, 2003.
- [6] Sergei Starostin. Review of Stefan Georg’s: Etymological Dictionary of the Altaic Languages. *Diachronica*, 22(2):451–454, dec 2005.
- [7] Alexander Vovin. The end of the Altaic controversy: In memory of Gerhard Doerfer. *Central Asiatic Journal*, 49(1):71–132, 2005.
- [8] Yuri Tambovtsev. The Altaic Language Taxon: Language Family or Language Union? *California Linguistic Notes*, 34(1):1–25, 2009.
- [9] Peter Turchin, Ilia Peiros, and Murray Gell-Mann. Analyzing Genetic Connections between Languages by Matching Consonant Classes. *Journal of Language Relationship*, pages 117–126, 2010.
- [10] Martine Robbeets. Transeurasian: Can verbal morphology end the controversy. In Lars Johanson and Martine Irma Robbeets, editors, *Transeurasian verbal morphology in a comparative perspective: Genealogy, contact, chance*, pages 81–114. Harrassowitz Verlag Wiesbaden, Wiesbaden, Germany, 2010.



- [11] Aleksey A. Burykin. Methods of comparative linguistics, the Altaic theory and the Turkic-Mongolic language relations (Remark on Valentin Rassadin’s article). *Ural-Altai Studies*, 19(4):93–105, 2015.
- [12] George Starostin. Altaic Languages. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics.*, page 9780199384655.013.35. Oxford University Press, New York, NY, 2016.
- [13] Martine Robbeets and Remco Bouckaert. Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. *Journal of Language Evolution*, 3(2):145–162, 2018.
- [14] Anna Dybo. New trends in European studies on the Altaic problem. *Journal of Language Relationship*, 14(1-2):71–106, 2020.
- [15] Robert McColl Millar and Larry Trask. *Trask’s Historical Linguistics*. 2015.
- [16] Roy Andrew Miller. Genetic Connections Among the Altaic Languages. In Sydney M. Lamb Mitchell. and E. Douglass Mitchell, editors, *Sprung from Some Common Source: Investigations Into the Prehistory of Languages*. Stanford University Press, Palo Alto CA, 1991.
- [17] Lyle Campbell. Do Languages and Genes Correlate?: Some Methodological Issues. *Language Dynamics and Change*, 5(2):202–226, 2015.
- [18] Alexander Vovin. 24 Northeastern and Central Asia: “Altaic” linguistic history. In *The Encyclopedia of Global Human Migration*. 2013.
- [19] Alexander Vovin. Why Koreanic is not Demonstrably Related to Tungusic. In Soo Hee Too and Kwang Chung, editors, *Korean within Altaic Languages and Altaic Languages within Korean. Altaic Series 2, ACSI.*, pages 97–117. Aleum Publishers, Seoul, ROK, 2014.
- [20] Alexander Vovin. Personal pronouns: a pillar or a pillory of the ‘Altaic’ hypothesis? *Türk Dili Araştırmaları*, 21(2):251–278, 2013.
- [21] Béla Kempf. Mongolic čilagun: Turkic tāš. *Turkic Languages*, 14:103–112, 2010.

- [22] Martine Irma Robbeets. *Is Japanese related to Korean, Tungusic, Mongolic and Turkic?*, volume 64. Harrassowitz Verlag Wiesbaden, Wiesbaden, Germany, 2005.
- [23] A. Pereltsvaig. *Languages of the world: An introduction*. Cambridge University Press, New York, NY, 2017.
- [24] A. Ceolin. Significance testing of the Altaic family. *Diachronica*, 36(3):299–336, 2019.
- [25] Norman I. Platnick and H. Don Cameron. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Biology*, 26(4):380–385, 1977.
- [26] Gerhard Jäger. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12752–12757, 2015.
- [27] Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244, 2015.
- [28] Johann Mattis List, Simon J. Greenhill, and Russell D. Gray. The Potential of Automatic Word Comparison for Historical Linguistics. *PLoS ONE*, 12(1):e0170046, 2017.
- [29] Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1):1–16, 2018.
- [30] Claire Bowern. Computational Phylogenetics. *Annual Review of Linguistics*, 4:281–296, 2018.
- [31] Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi, and William F. Martin. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1(9):1–8, 2016.
- [32] Benjamin Vernot and Joshua M. Akey. Complex history of admixture between modern humans and neandertals. *American Journal of Human Genetics*, 96(3):448–453, 2015.
- [33] Tetyana I. Vasylyeva, Samuel R. Friedman, Dimitrios Paraskevis, and Gkikas Magiorkinis. Integrating molecular epidemiology and social network analysis to study infectious diseases:

- Towards a socio-molecular era for public health. *Infection, Genetics and Evolution*, 46:248–255, 2016.
- [34] Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9241–9243, 2020.
- [35] Quentin D. Atkinson and Russell D. Gray. Curious parallels and curious connections - Phylogenetic thinking in biology and historical linguistics, 2005.
- [36] Mark Pagel, Quentin D. Atkinson, and Andrew Meade. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163):717–720, 2007.
- [37] Johann Mattis List, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144, 2018.
- [38] J. Felsenstein. Statistical inference of phylogenies. *Journal of the Royal Statistical Society: Series A (General)*, 146(3):246–262, 1983.
- [39] J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125(1):1–15, 1985.
- [40] John P. Huelsenbeck and Keith A. Crandall. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28:437–466, 1997.
- [41] Mark Pagel. Inferring the historical patterns of biological evolution, 1999.
- [42] Stacey D. Smith, Matthew W. Pennell, Casey W. Dunn, and Scott V. Edwards. Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends in Ecology and Evolution*, 35(5):415–425, 2020.
- [43] Luay Nakhleh, Don Ringe, and Tandy Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420, 2005.

- [44] Alexandre Francois. Trees, waves and linkages: Models of language diversification. In Claire Bowerman and Bethwyn Evans, editors, *The Routledge Handbook of Historical Linguistics*, pages 161–189. Routledge, London, 2014.
- [45] Nelson Sathi Shijulal, Johann Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 278(1713):1794–1803, 2011.
- [46] Johann Mattis List, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays*, 36(2):141–150, 2014.
- [47] K Ochiai, To Yamanaka, K Kimura, and O Sawada. Inheritance of drug resistance (and its transfer) between Shigella strains and Between Shigella and E. coli strains. *Hihon Iji Shimpou*, 1861:34, 1959.
- [48] Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7):3801–3806, 1999.
- [49] Shannon M. Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: Building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.
- [50] Daniel H. Huson and David Bryant. Application of phylogenetic networks in evolutionary studies, 2006.
- [51] Simone Linz, Achim Radtke, and Arndt Von Haeseler. A likelihood framework to measure horizontal gene transfer. *Molecular Biology and Evolution*, 24(6):1312–1319, 2007.
- [52] Cuong Than, Derek Ruths, and Luay Nakhleh. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9:322, 2008.
- [53] Alix Boc, Hervé Philippe, and Vladimir Makarenkov. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology*, 59(2):195–211, 2010.

- [54] Claudia Solís-Lemus and Cécile Ané. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLoS Genetics*, 12(3):e1005896, 2016.
- [55] Gavin M. Douglas and Morgan G.I. Langille. Current and promising approaches to identify horizontal gene transfer events in metagenomes. *Genome Biology and Evolution*, 11(10):2750–2766, 2019.
- [56] Gur Sevillya, Orit Adato, and Sagi Snir. Detecting horizontal gene transfer: A probabilistic approach. *BMC Genomics*, 21:1–11, 2020.
- [57] Patrick J. Keeling and Jeffrey D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.
- [58] Olga K. Kamneva, John Syring, Aaron Liston, and Noah A. Rosenberg. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, 17(1):1–19, 2017.
- [59] Diego F. Morales-Briones, Aaron Liston, and David C. Tank. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytologist*, 218(4):1668–1684, 2018.
- [60] Miguel Caparros and Sandrine Prat. A Phylogenetic Networks perspective on reticulate human evolution. *iScience*, 24(4):102359, 2021.
- [61] Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54:273–301, 2020.
- [62] Morgan E. Clippinger. Korean and Dravidian: Lexical Evidence for an Old Theory. *Korean Studies*, 8(1):1–57, 1984.
- [63] Susumu Ohno. The Genealogy of the Japanese Language - Tamil and Japanese. *Gengo Kenkyuu*, 1989(95):32–63, 1989.

- [64] R Development Core Team. R: A Language and Environment for Statistical Computing, 2008.
- [65] Johann-Mattis List and Robert Forkel. LingPy. A Python library for historical linguistics., 2021.
- [66] Johann Mattis List. SCA: Phonetic alignment based on sound classes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.
- [67] Robert R Sokal. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438, 1958.
- [68] Klaus Peter Schliep. phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.
- [69] Johann-Mattis List. Lebor: Lexical Borrowing Detection with LingPy, 2019.
- [70] Tal Dagan, Yael Artzy-Randrup, and William Martin. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):10039–10044, 2008.
- [71] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):735–740, 2018.
- [72] Yun Yu, James H. Degnan, and Luay Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, 8(4):e1002660, 2012.
- [73] Yun Yu, Nikola Ristic, and Luay Nakhleh. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC bioinformatics*, 14 Suppl 1:1–10, 2013.
- [74] Yun Yu, Jianrong Dong, Kevin J. Liu, and Luay Nakhleh. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the United States of America*, 111(46):16448–16453, 2014.
- [75] Robert Matthew Barnett. Overview of Rich Newick Strings, 2012.

- [76] Richard Stallman. GNU General Public License v3, 2007.
- [77] Timothy G Vaughan. Icytree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15):2392–2394, 2017.
- [78] J. D Bengtson and M. Ruhlen. Global Etymologies. In *On the Origin of Languages: Studies in Linguistic Taxonomy*, pages 277–336. Stanford University Press, Stanford, CA, 1994.
- [79] Hao Sun, Chi Zhou, Xiaoqin Huang, Shuyuan Liu, Keqin Lin, Liang Yu, Kai Huang, Jiayou Chu, and Zhaoqing Yang. Correlation between the linguistic affinity and genetic diversity of Chinese ethnic groups. *Journal of Human Genetics*, 58(10):686–693, 2013.
- [80] Ana T. Duggan, Mark Whitten, Victor Wiebe, Michael Crawford, Anne Butthof, Victor Spitsyn, Sergey Makarov, Innokentiy Novgorodov, Vladimir Osakovsky, and Brigitte Pakendorf. Investigating the prehistory of Tungusic peoples of Siberia and the Amur-Ussuri region with complete mtDNA genome sequences and Y-chromosomal markers. *PLoS ONE*, 8(12):e83570, 2013.
- [81] Bayazit Yunusbayev, Mait Metspalu, Ene Metspalu, Albert Valeev, Sergei Litvinov, Ruslan Valiev, Vita Akhmetova, Elena Balanovska, Oleg Balanovsky, Shahlo Turdikulova, Dilbar Dalimova, Pagbajabyn Nymadawa, Ardeshir Bahmanimehr, Hovhannes Sahakyan, Kristiina Tambets, Sardana Fedorova, Nikolay Barashkov, Irina Khidiyatova, Evelin Mikhailov, Rita Khusainova, Larisa Damba, Miroslava Derenko, Boris Malyarchuk, Ludmila Osipova, Mikhail Voevoda, Levon Yepiskoposyan, Toomas Kivisild, Elza Khusnutdinova, and Richard Villems. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genetics*, 11(4):e1005068, 2015.
- [82] Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011.
- [83] Alexander Takenobu Francis-Ratte and James M Unger. *Proto-Korean-Japanese: A New Reconstruction of the Common Origin of the Japanese and Korean Languages*. PhD thesis, 2016.

- [84] Sarah G. Thomason. Pidgins/Creoles and Historical Linguistics. In Silvia Kouwenberg and John Victor Singler, editors, *The Handbook of Pidgin and Creole Studies*, page 242. Blackwell Publishing Ltd, Hoboken, NJ, 2009.
- [85] John H. McWhorter. *The Creole Debate*. Cambridge University Press, New York, NY, 2018.
- [86] Silvia Kouwenberg and John Victor Singler. Creolization in Context: Historical and Typological Perspectives. *Annual Review of Linguistics*, 4:213–232, 2018.
- [87] Joseph B. Slowinski and Roderic D.M. Page. How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48(4):814–825, 1999.
- [88] Yugo Murawaki. Statistical modeling of creole genesis. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 1329–1339, 2016.
- [89] Peter Bakker, Finn Borchsenius, Carsten Levisen, and Eeva Sippola. *Creole Studies – Phylogenetic Approaches*. John Benjamins Publishing Company, Philadelphia, PA, 2017.

Supplementary Figure S1