Environmental factors affect the evolution of linguistic subgroups in Borneo

Alexander D. Smith and Taraka Rama
The Chinese University of Hong Kong | University of North Texas

This study investigates the relatedness and history of the Austronesian languages of Borneo, which is the third largest island in the world and home to significant linguistic diversity. We apply Bayesian phylogenetic dating methods to lexical cognate data based on four historical calibration points to infer a dated phylogeny of 87 languages. The inferred tree topology agrees with the mid and lower-level subgrouping proposals based on the classical comparative method, but suggests a different higher-level organization. The root age of the dated tree is shallower than the archaeological estimates but agrees with a hypothesis of a past linguistic leveling event. The inferred homelands of the major linguistic subgroups from a Bayesian phylogeographic analysis agree with the homeland proposals from archaeology and linguistics. The inferred homelands for four of the eight subgroups support the riverine homeland hypothesis whereby the major linguistic subgroups developed initially in communities situated along Borneo's major rivers.

Keywords: Bayesian phylogenetics, homeland, Austronesian, Borneo

1. Introduction

Borneo lies at a cross-roads in Island Southeast Asia (ISEA). The island sits on the easternmost extension of the Sunda Shelf, a part of the Southeast Asian continental shelf that connects the large Greater Sunda Islands to Mainland Southeast Asia during periods of glacial expansion and low sea-levels. Although humans have occupied this area for tens of thousands of years, the current dominant ethnolinguistic group in ISEA only reached Borneo in the last 4,000 years as a result of the Austronesian expansion out of Taiwan, into the Philippines, and beyond (Bellwood 2007; Blust 1985–1986). The initial population expansion took place over the ocean, with Austronesian settlers in ISEA utilizing advanced sea-faring technology to expand quickly over a large area (Blust 2019a). Over the course of only a few hundred years, nearly the entirety of ISEA had been settled by speakers

https://doi.org/10.1075/dia.20024.smi | Published online: 19 January 2022
Diachronica ISSN 0176-4225 | E-ISSN 1569-9714

of Austronesian languages. This resulted in the displacement, expulsion, or incorporation of existing populations into the larger Austronesian ethno-linguistic and cultural society, which resulted in the high linguistic diversity found in Borneo. Although nearly all languages of Borneo appear to descend directly from the first Austronesian inhabitants, thousands of years of linguistic evolution and population movement have created a great deal of linguistic diversity.

Geographically, Borneo is roughly the size of the US state of Texas (about 743,330 km^2). It is highly mountainous and was historically covered by an ancient old-growth tropical rainforest, although the vitality of Borneo's rainforests has been greatly reduced due to logging activities. The mountainous terrain and tropical rainforest climate have given rise to an island-wide system of large rivers; the largest rivers in ISEA are all located on Borneo. These environmental factors appear to have had a substantial impact on the linguistic and cultural history of the island.

Bornean languages belong to the larger Austronesian language family which has been analyzed extensively using both the traditional comparative method (Blust 2014) and modern phylogenetic methods (Gray et al. 2009, 2010). There is a near consensus on the placement of the Austronesian homeland in Taiwan (Blust 2019a), although the specifics of higher-order subgrouping remain a topic of debate. Regarding Borneo and its languages, the comparative analysis of the Bornean languages in Smith (2017a) has recently proposed subgroupings and their homelands for all languages of Borneo. The major subgroups consist of the Southwest Sabah, Northeast Sabah, North Sarawak, Central Sarawak, Kayanic, Land Dayak, Malayic, and Barito groups, but this recent hypothesis has not yet been verified using Bayesian phylogenetic techniques. For instance, in the Bayesian phylogenetic dating study (Gray et al. 2009) involving the Austronesian language family, only 20 languages belonging to North Sarawak, Barito, Malayic Dayak, and Kayanic subgroups were included, whereas our study includes all known linguistic subgroups of Borneo. We therefore test both the validity of the higher-order subgroups, as well as the composition of mid and lower-level subgroups as they are presented in Smith (2017a).

Our Bayesian phylogenetic analysis is performed with lexical cognate data based on the word lists collected in Smith (2017a). Typically, Bayesian phylogenetic studies – for various families such as Austronesian (Gray et al. 2009), Pama-Nyungan (Bowern & Atkinson 2012; Bouckaert et al. 2018), Indo-European (Bouckaert et al. 2012; Chang et al. 2015), Sino-Tibetan (Sagart et al. 2019; Zhang et al. 2019), and Bantu (Grollemund et al. 2015) – employ expert cognate judgments for inferring dated phylogenetic trees. In this paper, we test an automated cognate detection method (List et al. 2017; Rama et al. 2018). The automatic detection method may be useful for specialist linguists as a means to reduce workload,

by first detecting cognates automatically, then allowing for an expert to come in later and correct any inaccurate cognate judgments. In this study, automatic cognate detection corrections were provided by the first author, allowing our cognate assignment process to be transparent and time-efficient.

Next, our study addresses the role of the Bornean environment on the development of subgroups. There have been multiple studies (Nettle 1999; Greenhill 2014; Gavin et al. 2013) examining the role of ecological factors such as rainfall (Nettle 1998), latitude (Mace & Pagel 1995), river density (Axelsen & Manrubia 2014), and climate (Hua et al. 2019) in shaping the linguistic diversity in different areas of the world. Recently, a large scale study (Bentz et al. 2018) – involving phylogenetic trees inferred from multiple phylogenetic inference techniques for 46 language families of the world – applied two different phylogenetic signal techniques to study the effect to which environmental factors such as latitudes, longitudes, elevation, and distances to water bodies drive the evolution of language families. Based on the range and significance of the signals, the authors propose that environmental factors shaped the evolution of language families. We utilize these methods to test if similar environmental factors are at play in Borneo.

Another class of techniques (phylogeographical methods) explicitly reconstruct the internal nodes' homelands and can be utilized to give further insights into the migration patterns of Austronesian settlers in Borneo. Bayesian phylogeographical techniques have been applied to reconstruct the Indo-European homeland (Bouckaert et al. 2012), identify the migration routes of Bantu subgroups (Currie et al. 2013; Grollemund et al. 2015), and the expansion of the Pama-Nyungan family (Bouckaert et al. 2018). We apply Bayesian phylogeographic techniques to reconstruct the geographical locations of the internal nodes of the dated phylogenetic tree and compare the reconstructed homelands with the proposed homelands of the Bornean languages' subgroups. Further, we test the proximity of the major subgroups' reconstructed homelands to different water bodies and show that rivers are significantly closer to the reconstructed homelands than other geographical markers. As such, the migration of the proto-language speakers of a majority of the subgroups took place close to the rivers and not to other waterbodies such as lakes or coastline (the RIVERINE HOMELAND HYPOTHESIS).

2. Materials and methods

2.1 Data

The data for our analysis come from three sources. The majority are from Smith's (2017a) dissertation, which provides word lists from 78 linguistic communities

on Borneo. These data were gathered between 2014 and 2016 and include languages spoken throughout central and southern Borneo. Additional data sets are included from Lobel's *North Borneo Sourcebook* (Lobel 2016), which includes languages from the northern state of Sabah, Malaysia. Lobel's data are included to fill a gap from Smith (2017a), which did not include data from languages of this area. Additionally, a small data set for languages in the Berawan-Lower Baram subgroup was provided to Smith by Robert Blust from his unpublished field notes, which were used in the original analysis in Smith (2017a).[1]

2.2 Bornean Subgrouping from Smith (2017a)

Our study reexamines the conclusions drawn in Smith's dissertation with computational methods, and as such an overview of his study is warranted. The primary linguistic division in Borneo is between Greater North Borneo (GNB), which includes Northeast Sabah, Southwest Sabah, North Sarawak, Central Sarawak, Kayanic, Land Dayak, and Malayic, vs. the Basap-Barito subgroup which includes Basap and Barito (Blust 2010; Smith 2017a; but see Adelaar 2005 for an alternative proposal). Blust did not recognize the Central Sarawak subgroup nor the inclusion of Basap with Barito, but other facets of Smith's proposal align with those from Blust. The principal piece of evidence for the GNB hypothesis is an innovation in the numerals, where Proto-Malayo-Polynesian (PMP) *pitu 'seven' was replaced with a reflex of PMP *tuzuq 'to point', which became PGNB *tuju? 'seven' whereas *pitu remains unchanged in Barito. Additional lexical evidence for GNB is presented by Smith (2017a). An overview of Smith's internal divisions at middle and lower levels is given here:

- **Southwest Sabah** is divided into Greater Dusunic (Bisaya-Lotud-Dusunic and Paitanic) and Greater Murutic (Tatana, Papar, and Murutic). The center of diversity for Southwest Sabah is around the area of the city of Kota Kinabalu.
- **Northeast Sabah** is divided into Bonggi, a single language spoken on an island just off the northernmost tip of Borneo, and Idaanic, a group of languages spoken in far eastern Sabah.
- **North Sarawak** is divided into four groups, Dayic (Kelabit and Lun Dayeh), Kenyah (separated into Highland and Lowland varieties, and including the language of the nomadic Penan), Berawan-Lower Baram (including Miri,

1. Electronic supplementary material containing the data and the programs along with notes is available here: https://figshare.com/articles/dataset/Environmental_factors_affect_the_evolution_of_linguistic_subgroups_in_Borneo/13309121/3

Narum, and Kiput in the Lower Baram group and various Berawan languages), and Bintulu, a single language.

- **Central Sarawak** consists of the Melanau languages, Kajang, Punan, and Muller-Schwaner. Punan and Muller-Schwaner are further grouped together as Punan-Müller-Schwaner in Smith's analysis.
- **Land Dayak** languages are spoken in southern Sarawak and throughout the northwestern and north-central areas of West Kalimantan. They are divided into the Benyadu-Bekati' group and the Bidayuh-Southern Land Dayak group.
- **Malayic** is spoken throughout western Borneo, in most major cities on the island, as well as in areas outside of Borneo. Standard Malay and Indonesian are national languages, but there is a diversity of Malayic languages in Borneo that implies that Borneo is the ultimate homeland of the Malays. The internal subgrouping of Malayic is complex, but Ibanic and Kendayan are two major Malayic branches in Borneo, along with other smaller Malayic varieties. Because of the presence of several different types of Malay and Malayic on the island, when "Malay" or "Indonesian" is cited as a source of borrowing or interference, it refers not only to standard varieties but to the many other dialects/languages spoken throughout the island. Local varieties of Malay have been influencing languages in Borneo for quite some time and that influence is now being compounded by the spread of national standard varieties. The minutiae of Malay dialect diversity are beyond the scope of this paper, but see Adelaar (1992) for more on Malay and Malayic varieties.
- The last Greater North Bornean subgroup is **Kayanic,** which consists of the Kayan-Murik group (divisible into Kayan and Murik-Merap), and the Segai-Modang group (divisible into Segai and Modang). Kayanic languages are found throughout central Borneo, but Segai Modang and Murik-Merap are both found primarily in eastern Kalimantan on the Indonesian side of the island.
- **Basap-Barito** is the final subgroup, and the only non-Greater North Bornean subgroup located on the island, according to Smith. Basap languages are only minimally documented and are spoken in small pockets throughout far eastern areas of North and East Kalimantan. Barito languages, in contrast to Basap, form a large, well-studied group. They consist of the languages of southern Borneo as well as Malagasy, which is spoken throughout the island of Madagascar. Malagasy was shown to be most closely related to Ma'anyan and other Barito languages in the Southeast Barito group (Dahl 1951). Note that we do not include Malagasy in our study, but its close relationship to Ma'anyan means that its exclusion will not impact the tree. Recently, Barito

was reclassified as an innovation-defined linkage, rather than a traditional subgroup (Smith 2018).

2.3 Cognate detection

There has been quite some literature on the development of cognate detection methods (List et al. 2017) and testing their utility for inferring phylogenetic trees (Rama et al. 2018).

We infer the cognate judgments using a family-agnostic cognate detection method whose output was then corrected by the first author. Our cognate detection system is based on a linear SVM (Support Vector Machine) classifier that combines data-driven segment similarity scores (Jäger 2013) with a segment similarity score computed using sound changes that are weighted for prominence (List 2012). The SVM classifier computes a similarity score for all the word pairs belonging to a meaning which is then clustered using the similarity dependent Chinese Restaurant Process clustering algorithm (Rama 2018) that automatically tunes the clustering threshold for each meaning. After the cognate detection step followed by correction by the specialist linguist, we added data for the following languages and assigned cognate judgments: Rungus, Kadazan Kimanis, Dumpas, Lingkabau, Lobu Lanas, Tatana, Papar, Timugon Murut, and Tidung Sumbol. In summary, our data consist of 87 languages with 2966 unique cognate sets. The evaluation procedure of the cognate detection experiment is given in §3.1.

2.4 Bayesian phylogenetic dating

Lexical evolution model

All our cognate data is binary coded and has two states: *1* if a language is present in a cognate set and *0* otherwise. We model the loss or gain of a cognate using the binary Continuous Time Markov Chain (BinCTMC) model that allows an arbitrary number of transitions between the binary states. The rate variation across sites is modeled using a discrete Gamma distribution with four rate categories (Yang 1994) correcting for all-absence sites (Felsenstein 1992) in all the phylogenetic analyses. This model is known as BinCTMC + Γ model.

We utilize the BinCTMC + Γ over other character substitution models such as Stochastic Dollo and Covarion models. A recent study (Ritchie & Ho 2019) comparing different substitution models applied to four different language families showed that there is no clear winner among the four models in terms of Bayes Factor. This study is at odds with some individual works that find support for some models over others (Gray et al. 2009; Bowern & Atkinson 2012); there is

no consensus in the field. We therefore utilize the BinCTMC model in this study, since it has only a single parameter, whereas the covarion model has three parameters. This makes the BinCTMC model less complex overall.

Tree Prior

All our Bayesian phylogenetic analyses were performed with MrBayes 3.2.7 (Ronquist et al. 2012). We used a Fossilized Birth-Death tree prior (Zhang et al. 2015) with fossilization rate set to 0, as our dataset does not have any extinct languages. The birth-death model handles incomplete language sampling through a parameter $p = n/N$, where n is the number of languages in the sample and N is the total number of extant languages in the family. Here, we fix p to 1 since our dataset covers all the languages belonging to Western Indonesian branch of languages spoken in Borneo. In this paper, we use an Independent Gamma Rate (Lepage et al. 2007) relaxed clock model (IGR) where the rate of each branch comes from a Gamma distribution whose mean is 1.0 and variance is proportional to the inverse of the branch length. The base clock rate is drawn from a lognormal distribution with $\mu = -7$ and $\sigma = 0.6$ which would have a mean and standard deviation of 0.001 substitutions per thousand years.

Calibration points

Historical documentation with hard dates regarding the known histories of languages may serve as calibration points which may then inform phylogenetic dating. In our study we were able to determine calibration points for several nodes, utilizing written historical documents and linguistic insights outlined in Smith (2017a) and other sources (see Table 1). Our calibration points indicate the most recent common ancestor. Ages indicate the latest age at which the given node split up. The Ukit-Buket and Punan calibration points were inferred through analysis of historical documents dating from the time of early British presence in Sarawak (Kaboy 1974; Sandin 1994; Sellato 1994, 2001). These documents outline the recent history of these groups from histories collected from native speakers. The Tunjung-Benuaq calibration point was determined through analysis of texts from the Nāgara-Kĕrtāgama document which originates from the ancient Majapahit Kingdom (Coedès 1968; Pigeaud 1962). In this document, a group of people referred to as *Tuñjung-Kute* are mentioned around the Mahakam river. The document was written in the mid-1300s. The Ma'anyan-Dusun calibration point is informed by the linguistic and historical analysis in Adelaar (1989). In that paper, Adelaar estimates the date of Malagasy migration from Borneo at around 1,300–1,400 years ago, which allows us to establish a calibration point for Malagasy's closest relatives, Ma'anyan and Dusun. All calibration points' priors are drawn from a uniform distribution with the lower and upper bounds given in

Table 1. The root age is drawn from a uniform distribution ranging between 0 and 10,000 years before present (BP). We test the effect of the calibration points on the root age through a leave-one-out phylogenetic analysis where each calibration point is excluded from the phylogenetic analysis.

Table 1. Calibration points used in the analysis

Constraint Name	Languages	Age range (years BP)
Ukit-Buket	Ukit, Buket	250–350
Punan	Punan Aput, Punan Bah, Punan Lisum, Punan Tuvu, Ukit, Buket	500–600
Tunjung-Benuaq	Tunjung, Benuaq	600–5000
Ma'anyan-Dusun	Ma'anyan, Dusun	1300–1400
Root	All languages	0–10000

Monte-Carlo Markov chain settings

We performed two independent runs and sampled parameters by running one cold and three hot chains in parallel. The hot chains allow the MCMC to explore the parameter landscape efficiently by moving across the peaks and not getting stuck in a local optimum. We ran the chains for 50×10^6 generations and sampled the chain at every 2000th generation in order to reduce auto-correlation. We assessed the convergence of branch lengths using the Potential Scale Reduction Factor (Gelman et al. 2013), whose value should approach 1 across independent runs should the runs converge. The independence between the samples for parameters such as tree height, birth-death rates, and the IGR variance parameter is assessed using Estimated Sample Size, which is expected to be at least 100 for all the parameters (Gelman et al. 2013).

Topological constraints

We apply topological constraints on the shape of our tree to control for borrowing, which may give false cognate scores between subgroups with a known history of contact. Smith (2017a) has performed detailed analysis of borrowing relationships between subgroups in Borneo, which was utilized in our study for the creation of topological constraints. The main source of noise from borrowing in our study is Malay/Indonesian, the dominant lingua franca of the area. We tested the effect of topological constraints on the results of the dating analysis. We test if the results from the Bayesian dating are affected if topological constraints are not included in our analysis.

We test both the effect of calibration points and topological constraints through a sampling through priors analysis.

2.5 Geographical reconstruction

According to Blust (2019a) and Smith (2017a), the speakers of the ancestral language were supposed to have entered Borneo through Palawan island and then spread into the rest of Borneo by moving through the east and west coast of Borneo followed by subsequent inland migration.

We reconstructed the homelands for the major subgroups using both fixed rates and variable rates geographical models implemented in BayesTraits v.2.6.3.[2] The fixed rates model as implemented in BayesTraits (BTF)[3] is a Brownian motion model where the latitudes and longitudes are mapped to the three dimensional Cartesian coordinate system. The three dimensional coordinates are treated independently in this model. In the BTF model, there is a single parameter, the variance of the normal distribution, which is sampled using a Monte Carlo Markov Chain (MCMC) procedure along with the three dimensional coordinates. The BayesTraits models take a single tree with branch lengths and the geographical coordinates as input and then reconstructs the internal nodes' geographical locations using the MCMC procedure. We ran the BTF model for 1 million iterations sampling at every 1000th iteration. This run was preceded by a burnin of 10,000 iterations.

The variable rates model (BTV) is a relaxation of the single parameter Brownian motion which assumes that the rate of change is fixed across all the branches. In this model, the branch lengths are allowed to shrink or expand reflecting large movements in space. In contrast to the MCMC models, where the number of parameters is fixed through the sampling process, the BTV model has two parameter changes: whether to scale a branch and to sample the scaling parameter of a particular branch. The decision to scale a branch increases the number of parameters by 1 and requires the use of a Reverse Jump MCMC (RJMCMC) procedure to sample parameters. The RJMCMC procedure does not easily converge, requiring long running times depending on the number of languages. In this paper, we run the model for 11 million iterations (sampled at every 2000th iteration) preceded by a burnin of 1 million iterations.

We also test a third model, dubbed the Northern homeland model (NorBTF), which is an extension to the BTF model where the majority consensus tree root's geographical coordinates are fixed to test for linguistic hypothesis that the first

2. http://www.evolution.rdg.ac.uk/BayesTraits.html

3. We follow the abbreviations of the model names in Wichmann and Rama (2020).

Table 2. Topological constraints used in the analysis

Constraint	Languages
Barito	Kadorih, Ngaju, Bakumpai, Dusun, Ma'anyan, Tawoyan, Benuaq, Tunjung, Paser, Basap Lebo
Malayic	Kendayan 2, Ketapang, Keninjal, Seberuang, Upper Kapuas Iban, Mualang
Land Dayak	Benyadu, Bekati, Hliboi, Bidayuh, Sungkung, Ribun, Jangkang, Pangkodan Sanggau, Golik
Sabahan	Burusu, SBisaya, BrDusun, LBisaya, Lotud, Bonggi, Idaan, Begak, Seguliud, Bulungan, Rungus, Kadazan, Kimanis, Dumpas, Lingkabau, Lobu, Lanas, Tatana, Papar, Timugon, Murut, Tidung, Sumbol
Malayic–Land Dayak (negative)	Malayic and Land Dayak should not occur together
Malayic–Barito–Sabahan (negative)	Malayic, Barito and Sabahan should not occur together

speakers of Austronesian languages entered into Borneo via Palawan. Similar to the BTF model, we ran the NorBTF model for 1 million iterations sampling at every 1000th iteration, preceded by a burnin of 10,000 iterations.

We assess the support for the three different models using Bayes Factor, which involves the computation of the marginal likelihood. The log marginal likelihood is computed using the stepping stone sampler (Xie et al. 2011) implemented in the BayesTraits software with 20 stones and 1000 iterations for each stone. Then, the logarithm of the Bayes Factor is computed as twice the difference between the logarithm of the marginal likelihoods. The results of this analysis are given in Table 6. A difference greater than 10 is considered to be *very strong* for choosing a complex model (variable rates model) over the simpler fixed rates model.

2.6 Test of riverine hypothesis

Based on the placement of the subgroups, it was proposed in Smith (2017a) that rivers played a dominant role in the development of major subgroups given in Table 3. We test the effect of ecological factors such as lakes, rivers, and coastlines on the development of major subgroups through a statistical significance test to determine if the geographical distances of the reconstructed homelands of the major subgroups to water bodies are lesser than the distances from randomly selected land points within Borneo. For each major subgroup, we extract the 1000 best homelands and determine if the reconstructed geographical points are sig-

Table 3. Major subgroups and the list of languages. Information regarding Land Dayak, Barito, and Malayic subgroups is given in Table 2

Subgroup	Languages
Central Sarawak	Aoheng, Seputan, Hovongan, Kereho, Buket, Ukit, Punan Aput, Punan Lisum, Punan Bah, Punan Tuvu, Dalat Melanau, Kanowit, Kejaman, Sekapan, Lahanan
Kayanic	Apo Kayan Kayan, Baram Kayan, Busang, Bahau Saq, Mpraa, Ngorek, Gaai, Kelai, Long Gelat, Modang Woeg Helag
North Sarawak	Badeng, Tau, Pawe, Sawa, Laang, Gah, Penan Beku, Penan Jeki-tan, Penan Mubui, Sebop (Old), Vo, Berawan Terawan, Berawan jeegan, Kiput, Miri, Narum, Kelabit, Lun-Dayeh
Southwest Sabah	BrDusun, LBisaya, SBisaya, Burusu, Tidung Sumbol, Timugon Murut, Papar, Tatana, Dumpas, Lingkabau, Lobu Lanas, Rungus, Kadazan Kimanis, Lotud
Northeast Sabah	Begak, Idaan, Seguliud, Bonggi

nificantly affected by the presence of water bodies. For the statistically significant subgroups, we outline the nearest major rivers along with their frequencies.

3. Results

3.1 Cognate detection

We performed automatic cognate detection on 49,560 lexical items, out of which 17,092 lexical items for 221 meanings were double checked by Smith, who corrected incorrect automatic cognate judgments based on his expert knowledge of the languages of Borneo and their historical phonological development. The 17,092 items were chosen because they had no missing entries for each concept. The remaining 32,468 are excluded from the phylogenetic analysis. Roughly 40 hours were spent correcting errors in the original cognate judgments, which, for a data set of this size, is significantly more time-efficient than complete manual cognate detection.

We evaluate the quality of the inferred cognate clusters against the expert cognate judgments using B-cubed F-scores (Amigó et al. 2009) which is a standard measure to evaluate the performance of cognate detection systems (Rama 2018). The F-score is computed as the harmonic mean of precision and recall. Precision measures if all the words within a cluster are cognate with each other. Recall measures if the detection algorithm is good at putting all the cognate words within the same cluster. We obtain a precision of 0.77, recall of 0.94, and a F_1-score of 0.84

suggesting that the cognate detection system has about 80% agreement with the expert cognate judgments. In total, 8,510 judgments were adjusted to reflect corrected cognate judgments and an additional 1350 items were added to the database along with cognate judgments. The corrected judgments were then used to perform the phylogenetic analysis.

We compare the performance of our cognate detection system against two other systems: lumper and splitter (Rama et al. 2018). The lumper system puts all synonyms into a single cognate set (every word is cognate with each other) whereas the splitter system puts each word into its own cognate set. For each word, the lumper system has a perfect recall of 1 whereas the precision is not 1 since there are words in the cognate set that are not cognate with the word. The splitter system has a precision of 1, since every word is in its own set, whereas the recall is not 1, since words that are cognate are placed in their own cognate sets, leading to a very low recall. The results of the lumper and splitter systems are given in Table 4. The SVM-CRP system produces better results than lumper and splitter systems in terms of precision and recall.

Table 4. Performance of different automated cognate detection systems

System	Precision	Recall	F_1-score
Lumper	0.42	1	0.59
Splitter	1	0.19	0.31
SVM-CRP	0.77	0.94	0.84

3.2 Network analysis

The non-treelike signal in the data is assessed by inferring a NeighborNet network using the SplitsTree software (Huson & Bryant 2006). We assess the tree-signal in the data using δ score (Holland et al. 2002) and Q-residual score. The mean δ score of 0.28 is between the reported scores of Indo-European (0.21) and Austronesian (0.33) in Gray et al. (2010). We obtain a Q-residual score of 0.0046 which is within the scores reported for other language families. The δ score for each of the Bornean languages ranges from 0.234–0.365.

Only 17 languages have a δ-score greater than 0.3, which is outside the range of mean summed with standard deviation. Of these, six are Barito languages which form a linkage relationship (Smith 2018). As defined by Ross (1988), linkages lack many traditional treelike characteristics, including a lack of internal structure and clearly-defined subgroup-wide innovations. The high δ-scores with these languages therefore stem from the nature of their relatedness. Additionally, five languages from within North Sarawak with δ-scores greater than 0.3 were evalu-

ated with data from secondary resources. These secondary resources have lower-than-average lexical attestation (fewer recorded lexemes for comparison), which contributes to worse reticulation scores. Bulungan had the highest score, although Bulungan is neither in a linkage relationship with its sister languages nor were Bulungan data from a secondary source. Bulungan's high score can rather be explained as a product of its intense borrowing relationship with Malay. The name Bulungan follows from the sultanate of Bulungan, a pre-colonial regional power which used Malay as a regional lingua franca (although Bulungan itself is not a Malayic language. It rather subgroups with the Sabahan languages.) Borrowing can explain many of the other languages with high scores, including, for example, Punan Tuvu, which is geographically far-separated from other Punan languages and has a high borrowing rate from Kayan as detailed in Smith (2017a, 2018). Even though the overall number of languages with a δ-score of greater than 0.3 is already small, examination of the languages in this set has shown that the scores are easily explained. The resulting network (Supplementary Material 3; Figure 1) correctly groups languages into the major subgroups hypothesized in Smith (2017a), with Sabahan, North Sarawak, Central Sarawak, Kayanic, Malayic, Land Dayak, and Barito all forming recognizable clusters.

3.3 Dating

3.3.1 *Tree topology accuracy*

The majority consensus tree provided in Figure 1 was inferred using the four calibration points given in Table 1. It has been noted that the accuracy of a phylogenetic model relative to a model from the comparative method is best analyzed by comparing the similarity of mid and lower level subgroups between the two models, as well as analyzing the presence of widely-accepted subgroups within the phylogenetic model (Nichols & Warnow 2008). Widely-accepted mid-level subgroups in Borneo are Northeast Sabah, Southwest Sabah, North Sarawak, Kayanic, Central Sarawak, Malayic, Land Dayak, and Barito (Tables 2 and 3), as listed in Smith (2017a). Our tree agrees with Smith's subgrouping in several important respects. At the mid and lower levels, the majority consensus tree without topological constraints supports all subgroups with the exception of Barito. Disagreement over the placement of Barito languages is a result of the nature of their relatedness as a linkage rather than a traditional subgroup, which results in low cognate scores (Smith 2018). Topological constraints on the majority consensus tree in Figure 2 take Barito's status into account, resulting in a discrete Barito group. For a full discussion on how the different trees agree, see §4.1 where we discuss how the trees agree, how they differ, and the implications of the new model.

3.3.2 *Effect of topological constraints on root age*

The result of the phylogenetic analysis with calibration points and topological constraints is given in Figure 3b. The root of the tree without the constraints (Table 2) has a median age of 2692 with a 95% highest posterior density (HPD) interval of [2085–3421]. The median root age of the tree with constraints (Figure 2) is slightly older at 2992 years with a HPD interval of [2318–3900]. It has to be noted that we provided a very wide possible age range as the prior for the root age. As a matter of fact, the upper bound of the root age prior is larger than the Austronesian age inferred in Gray et al. (2009). We note that both the root ages are close to the Barito subgroup age of 3200 years inferred in the larger Austronesian Bayesian phylogenetic study (Gray et al. 2009).

3.3.3 *Leave-one out prediction of calibration points*

As shown in the case of Indo-European Bayesian phylogenetic analyses (Chang et al. 2015), the selection of calibration points can be quite influential in any phylogenetic analysis. Therefore, we evaluate the effect of calibration points on the root age by excluding each calibration point iteratively and then performing a complete phylogenetic analysis. We also compare the predicted age of the excluded calibration point against its gold standard age in Table 5. We observe that the Punan and Ma'anyan-Dusun calibration points influence the root age predictions. The predicted age of Punan is much older than our calibrated age range. On the other hand, Ma'anyan-Dusun's predicted age is quite young compared to our calibrated age range. The MrBayes commands and the majority consensus tree files are provided in Supplementary Material 2.2.

3.3.4 *Priors only analysis*

We perform priors only analyses for both the phylogenetic analyses to explore the effect of priors on the distribution of root ages. The results of this experiment are given in Figure 3. The histograms of the root ages are based on the post burnin of 25% of the samples. In both the analyses, the root age in the prior samples is distributed across a wide range from about 1,500 years to 10,000 years. The lower bound of 1,500 years is due to the Ma'anyan-Dusun calibration point which is upper bounded to 1,400 years. In both the analyses, the posterior samples (those which utilize lexical data in our analysis) are distributed between 2,000 and 4,000 years, suggesting that the root age is dependent on the lexical data.

Table 5. Leave-one-out calibration point experiment showing the median age and the 95% HPD Intervals for both the calibration point and the root age

Calibration point	Age range	Predicted Subgroup Age	Predicted Root age
Ukit-Buket	250–350	248 [79–419]	3027 [2335–3846]
Punan	500–600	2130 [1383–3060]	4518 [3125–6131]
Tunjung-Benuaq	600–5000	518 [187–855]	2903 [2238–3715]
Ma'anyan-Dusun	1300–1400	209 [67–375]	1610 [1250–2044]

3.4 Test of geographical reconstruction methods

The log marginal likelihoods of the different geographical models computed from the stepping stone analysis are given in Table 6. The best model turns out to be the BTV model where there is very strong evidence (difference greater than 10; Kass and Raftery 1995) for preference over the BTF and NorBTF models. We also compare the distances between the MAP *(maximum a posteriori)* homelands inferred by different methods for each major subgroup in Table 7. As Table 7 indicates, the BTF model with homeland fixed in the north affects the homelands of Land Dayak and Malayic subgroups the most. Moreover, the Northern homeland is more than 1,000 kms apart from the best fitting variable rates model. The MAP homeland (Figure 4) inferred using the variable rates model is placed in the South-west region of the island. The difference between the homelands inferred by the BTF and BTV models ranges from 24 kms to 194 kms for the major subgroups and 283 kms for the root. Therefore, we use the inference from the BTV model in the following geographical homeland analysis.

Table 6. Log marginal likelihoods for the different geographical reconstruction models ordered by the best model

Model	Log marginal likelihood
Variable Rates (BTV)	−2537
Fixed Rates (BTF)	−2731
Fixed Rates model with Northern homeland (NorBTF)	−2893

3.5 Riverine hypothesis test

The riverine hypothesis (described in §2.6) proposes that rivers influenced the development of the homelands of the major subgroups. Apart from rivers, there

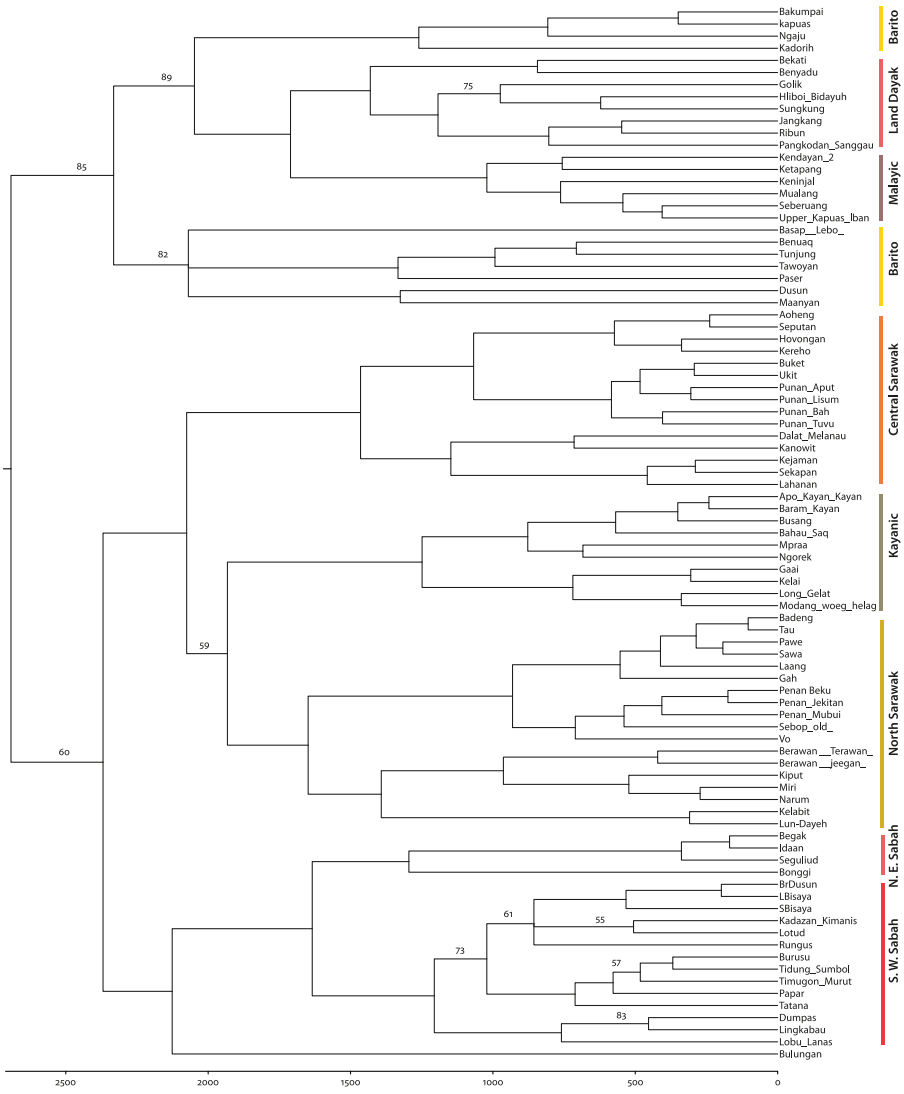


Figure 1. Majority consensus tree with calibrations without topological constraints

are other water bodies such as coastlines and lakes that could also be influential in the dispersal of the major subgroups.

We reconstructed the homelands for the major subgroups using the variable rates geographical model implemented in BayesTraits. We extracted the best 1000 coordinates ranked by their likelihood. We tested the significance of the distances of these reconstructed coordinates to waterways by choosing 1000 random land points and then computing the distances to waterways. A Wilcoxon rank sum

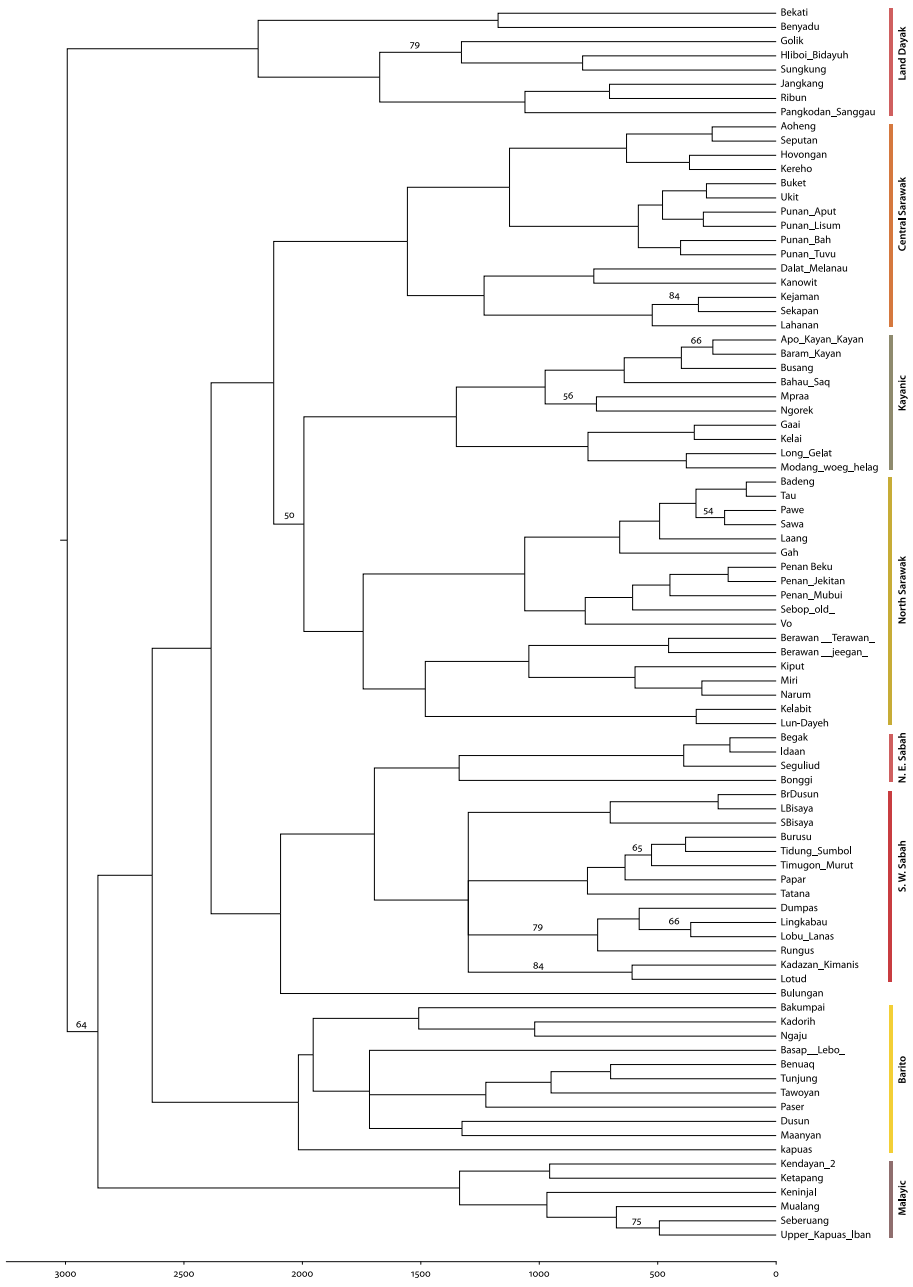
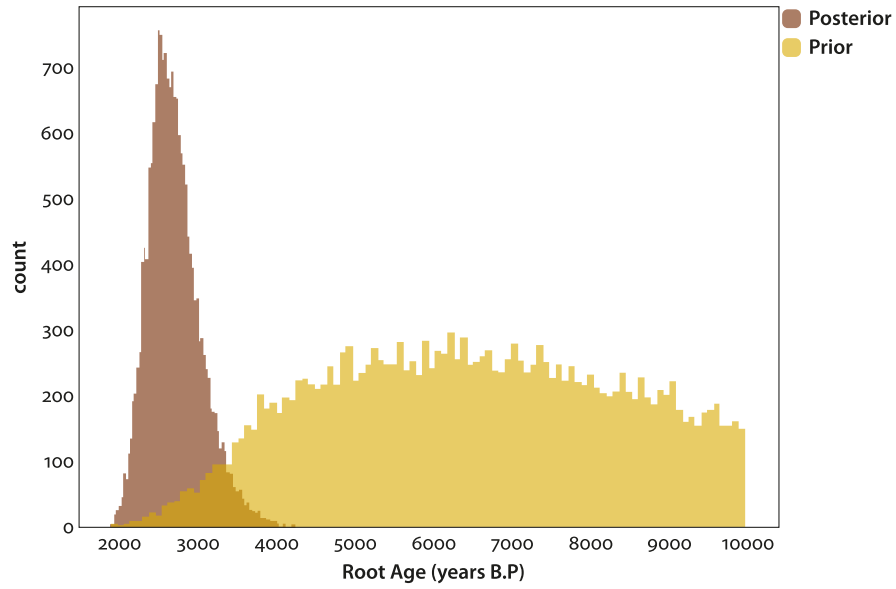
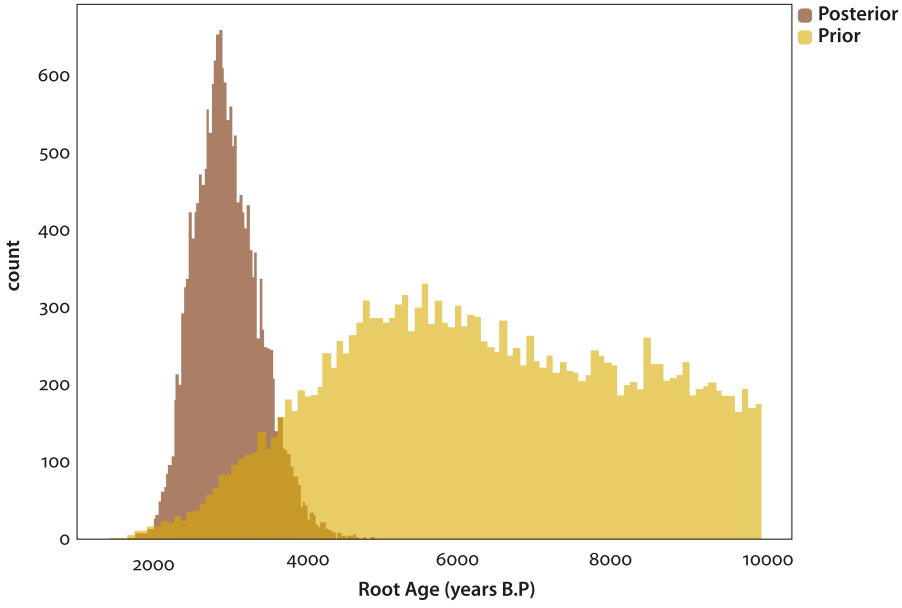


Figure 2. Majority consensus tree with calibrations and topological constraints. Only those branches which are found less than 90% are annotated. All the trees are drawn using `Figtree` software (Rambaut 2012)



a. Calibration points only



b. Including topological constraints

Figure 3. Posterior distribution of the root age against the prior distribution of the root ages for dating analyses without and with topological constraints

Table 7. Distances (in kilometers) between the outputs of different models. BTF: BayesTraits Fixed Rates model; BTV: BayesTraits Variable Rates model; NorBTF: BayesTraits Fixed Rates with Northern homeland

Subgroup	NorBTF vs. BTF	BTV vs. BTF	BTV vs. NorBTF
Barito	171	164	318
Central Sarawak	64	24	78
Kayanic	38	82	69
Land Dayak	400	194	535
Malayic	197	82	271
North Sarawak	86	45	53
Northeast Sabah	84	122	117
Southwest Sabah	34	81	49
Root	783	283	1010

test at $p < 0.001$ suggests that the reconstructed coordinates of Kayanic, Central Sarawak, North Sarawak, and Barito are significantly closer to rivers than the random points. In the case of Southwestern Sabah, both lakes and coastline are significant at $p < 0.001$ whereas in the case of Northeastern Sabah, only the minimum distance to coastline is significantly less than the random points. For each major subgroup, we also list the closest river along with its frequency to each of the 1000 inferred homelands in the Table 8.[4] The code and the files required to perform the distance computations and significance testing are given in Supplementary Materials 2.5 and 2.6.

4. Discussion

4.1 Subgrouping

With respect to topological accuracy, we found that the majority consensus tree agrees in important respects with already long-established subgroups. Beyond its agreement with long-established subgroups, the majority consensus tree also agrees with more recent subgrouping proposals (Smith 2017a; Blust 2010). We also find disagreements between our tree and previous hypotheses from the comparative method, particularly in the higher-level nodes.

4. Note that there are two "Kapuas Rivers" in Borneo. The Kapuas listed in the Barito row is a different river from that listed in the Land Dayak and Malayic rows.

Table 8. Top 3 closest rivers and their frequencies. The subgroups in bold show statistically significant support for the riverine hypothesis

Subgroup	River 1	River 2	River 3
Barito	Barito (572)	Sungai Teweh (218)	Sungai Kapuas (118)
Central Sarawak	Sungai Rajang (802)	Sungai Belaga (190)	
Land Dayak	Sungai Kapuas (663)	Batang Sadong (273)	
Malayic	Sungai Kapuas (458)	Sungai Melawi (236)	
Kayanic	Sungai Kayan (447)	Baram (259)	Sungai Bahau (235)
North Sarawak	Baram (609)	Sungai Akah (343)	
Southwest Sabah	Sungai Padas (620)	Sungai Sugut (248)	Sungai Kinabatangan (132)
Northeast Sabah	Sungai Sugut (557)	Sungai Kinabatangan (329)	

To review, Blust has proposed that the languages of Borneo descend from a common ancestor, Proto-Western Indonesian, which splits into at least two large subgroups on Borneo; Greater North Borneo (GNB) and Barito (Blust 2010). Blust's proposal includes additional languages which he considers part of Western Indonesian (WIn), but which form an as-yet undetermined number of subgroups, all of which are spoken outside of Borneo and therefore not included in the present analysis. At this first-order level, the majority consensus tree and Blust's proposals disagree. Our tree recognizes an early split between Land Dayak and other Bornean languages, with a second split separating Malayic from the remaining languages. We recall from §2.2 that both Land Dayak and Malayic are included in GNB according to previous studies. Our tree therefore has the effect of expelling Malayic and Land Dayak from GNB and also does not recognize the primary division between GNB and Barito. The third split in our tree separates Barito from the remaining "Greater North Bornean" languages which then divide along conventional lines.

From an Austronesian specialist's perspective, the early split of Land Dayak follows from the subgroup's overall divergence; it is one of the most divergent subgroups in Borneo (Smith 2019b). Our analysis finds the subgroup's divergence significant enough to warrant an early split. It was mentioned in §2.2 that the principal piece of evidence for the GNB hypothesis is a semantic shift where PMP *pitu 'seven' was replaced with a reflex of PMP *tuzuq 'to point', which then became PGNB *tuju? 'seven'. As Smith (2017a) points out, however, Land Dayak does not contain reflexes of *tuzuq as 'seven'. Rather, Proto-Land Dayak (PLD) 'seven' is reconstructed as *iju?, a superficially similar but likely unrelated innovation. The only supporting evidence given for including Land Dayak in GNB is a set of ten lexical innovations and a single phonological innovation: the prothesis

of a support vowel on words that became monosyllabic through other historical changes (for example, PMP *buhək 'head hair' > PGNB *əbbuk, and ultimately PLD *abuk). However, this single piece of phonological evidence is weakened by the presence of support vowels in reflexes of this word in other, non-Bornean languages such as Giangan *obbuk* and Toba Batak *obuk*. The reliance on this single piece of phonological evidence for including Land Dayak in GNB in earlier studies is therefore problematic. Our main takeaway from these higher-level subgrouping disagreements is that the GNB hypothesis itself, while plausible, lacks quantitatively robust evidence and that Land Dayak in particular is not well supported as a member of GNB.

Significant agreements between our tree and recent hypotheses arise in the middle level of the tree. This includes the Central Sarawak subgroup and its internal divisions Melanau, Kajang, and Punan-Müller-Schwaner, as well as the inclusion of Basap into the Greater Barito subgroup, both proposed quite recently (Smith 2017a, 2018). This is in addition to the current tree's agreement with longer-standing subgroups discussed earlier in §3.3.1. Even at lower-level nodes, the phylogenetic analysis gave results that complement recent hypotheses from the comparative method, earlier discussed in §2.2. The North Sarawak subgroup is divided into Kenyah, Berawan-Lower Baram, and Dayic subgroups, which follows from Blust (2010) and Smith (2017a). Kenyah is itself split into two groups which align with those proposed in Smith (2015b) and Smith (2015a). Berawan-Lower Baram subgroup is also internally identical to previous hypotheses (Blust 2010). The phylogenetic analysis recognizes a Kayanic subgroup with a Kayan-Murik and Segai-Modang division, agreeing with recent works (Blust 1974; Smith 2019a). The analysis of Land Dayak, proposed in Smith (2017a), as being comprised of a Benyadu-Bekati' group and a Bidayuh-Southern Land Dayak group is also supported in the new tree. These agreements hold true with both the topologically constrained and unconstrained trees since our constraints are general and do not interfere with the grouping of nodes at these lower levels. Thus, we observe more agreement in mid and lower level nodes, despite some disagreements in the higher-level nodes.

4.2 Dating

Archaeological evidence suggests that the Austronesian expansion out of Taiwan and into the northern Philippines began 4,000–4,200 BP with the settlement of the Batanes islands (Blust 2019a; Bellwood & Dizon 2005). This movement of people appears to have been rapid, and most of ISEA was settled by the descendants of these initial settlers within several hundred years (Bellwood 2007). The rapid expansion of Austronesian speaking people can be seen in the tree structure

of Malayo-Polynesian, which lacks a "nested" internal structure and instead has a "rake-like" structure indicative of rapid population movements (Gray et al. 2009; Smith 2017b). Settlement of Borneo would therefore have taken place not long after the settlement of the Philippines and we may expect evidence for Austronesian settlement to begin appearing in the archaeological record between 3,500 and 4,000 BP. This is indeed the case, and archaeological evidence routinely places the settlement of Borneo within this range (Bellwood 1995, 2007).

Our tree with constraints (Figure 2) places median root age at 2,992 with HPD of [2,318–3,900] and is thus shallower than archaeological estimates, but consistent with estimates from earlier phylogenetic studies (Gray et al. 2009). The shallow phylogenetic dating in Borneo parallels those found in the Philippines, where it was noted that a likely Greater Central Philippine expansion results in shallower-than-expected dates (Gray et al. 2009; Blust 2005). A similar history may explain shallow dates in Borneo. This possibility is discussed in greater detail in §4.3.

Finally, we note that the dating method in Gray et al. (2009) is different from the dating methods used in this paper. The methodology in Gray et al. (2009) consists of inferring an initial unrooted tree that is rooted with Old Chinese as outgroup followed by application of rate smoothing method for dating the tree. In this paper, we perform both dating and tree inference jointly.

4.3 Homeland inference

It is understood that homeland inferences may disagree with historical homelands due to the effects of subsequent population movements and linguistic leveling events. For example, Blust (2019b) has shown that a Philippine subgroup likely formed when a group of people speaking a putative Proto-Philippine language expanded from the central Philippines, reducing the expected level of linguistic diversity in the Philippines and possibly covering up linguistic evidence which would otherwise place the Malayo-Polynesian homeland in the Northern Philippines. Inferences on Malayo-Polynesian homelands are thus mainly informed by archaeology, since the linguistic evidence has been lost.

We find a similar situation in Borneo. Although the first Austronesian speaking people to enter the island almost certainly traveled through the Philippines and entered from the north, our variable rates model places the root in southwestern Borneo. This follows from the observation that the first several divisions in the Majority consensus tree with calibrations and topological constraints are between Land Dayak (first division), Malayic (second division) and Barito (third division). All of these subgroups are in southern Borneo, with Land Dayak and

Malayic in southwestern Borneo. This suggests that the area of highest linguistic diversity is in the southwest, rather than the north.

With both a shallower-than-expected root age and a homeland placement in southwestern rather than northern Borneo, the evidence from our study strongly suggests a linguistic leveling event. The possibility of such a leveling event in Borneo has already been suggested in Smith (2017a), where it is said that "Greater North Borneo could not have been spoken along the entire area where its daughter languages were located in the late 1700s. Its current distribution must have been the result of an expansion from a more compact area, although the location of that area is not clear" (Smith 2017a: 417). Smith goes on to state that "It is possible that the west coast of Borneo was home to more than one primary branch of [Western Indonesian], but that past diversity was leveled after the expansion of [Proto-Greater North Borneo]."

Smith's observation was limited to Greater North Borneo, but the present study calls into question the validity of this subgroup, and instead suggests that the leveling event occurs with the root node. Under such a hypothesis, the initial settlement of Borneo would have involved the gradual settling of the coast and development of multiple direct descendant languages dispersed throughout the island roughly 4,000 BP, consistent with archaeological evidence. Just over 1,000 years later, people from southwestern Borneo came to dominate the island, and their language replaced any past diversity. This scenario finds supporting evidence in both the shallow age and southwestern homeland indicated in our model.

4.4 Riverine hypothesis and the dispersal of major subgroups

We find that the placement of the Barito, Central Sarawak, Kayanic, and North Sarawak homelands are significantly affected by large rivers (see Table 8). Importantly, the rivers that affect these subgroups are precisely those which are suggested in Smith (2017a). The Barito homeland is significantly affected by the Barito river, the Central Sarawak homeland by the Rajang river (alternatively spelled "Rejang"), the Kayanic homeland by the Kayan river, and the North Sarawak homeland by the Baram river. Other homelands, such as the Malayic and Land Dayak homelands, are placed near the Kapuas river in agreement with Smith (2017a), however these homelands fail to reach significance of $p < 0.001$. The two Sabahan subgroups, on the other hand, are significantly affected by other factors such as coastlines and lakes. The significance of coastlines in Sabah appears to follow from the geography of Sabah: there are fewer large rivers in Sabah and rivers seem to have played a smaller role in Sabah than elsewhere in Borneo.

Based on the MAP estimates of the homelands, we trace the migration routes of the major subgroups from their homelands to the immediate descendants in Figure 4. We find that the immediate descendants' homelands are close to the major river systems and furthermore, that the migration routes suggested by the model can provide insights into the histories of the different subgroups. Since we traced the homelands of each node, we are able to discuss the inferred migration routes for each subgroup and its descendants. In the following discussion we make reference to the maps and homeland points in Figure 4.

4.4.1 *Malayic and Land Dayak*

These two subgroups have had some of the most intense contact of any two subgroups in Borneo. Specifically, Land Dayak has borrowed a large amount of its vocabulary from Malayic languages (Smith 2019b). Both subgroups' homelands are placed near the Kapuas river (① for Malayic, ② for Land Dayak). Proto-Malayic and Proto-Land Dayak diversified in such a way that their daughter nodes remained close to the Kapuas river (ⓐ, ⓒ, ⓓ), although the Benyadu-Bekati' division within Land Dayak involved migration away from the Kapuas (ⓑ).

4.4.2 *North Sarawak*

We place the Proto-North Sarawak homeland near the Baram river (③), noting that its immediate daughter nodes are all concentrated around the same area (ⓔ, ⓕ, ⓖ). We also note that Dayic appears to have followed the course of the Baram river all the way to its source in the Barito highlands (ⓗ). The Baram river was therefore a major factor in the movements of speakers of North Sarawak languages and their speakers.

4.4.3 *Central Sarawak*

The Central Sarawak homeland is placed along the Rejang river (④). From there, two migration paths separate Melanau-Kajang (⓵) and Punan-Müller Schwaner (⓶). The placement of the latter is especially interesting, since we were able to accurately place the Punan-Müller Schwaner homeland in the upper Balui/Baleh area, a location that matches oral histories gathered by anthropologists (Sellato 2001, 1994). We note that Punan languages are currently quite widely dispersed throughout Borneo, with no clear center of dispersal. We were able to find a homeland that agrees with oral histories despite this dispersion.

4.4.4 *Kayanic*

We place the Proto-Kayanic homeland in the Upper Pujungan and Apo Kayan highlands of the Kayan river system (⑤). From here, there is a split between

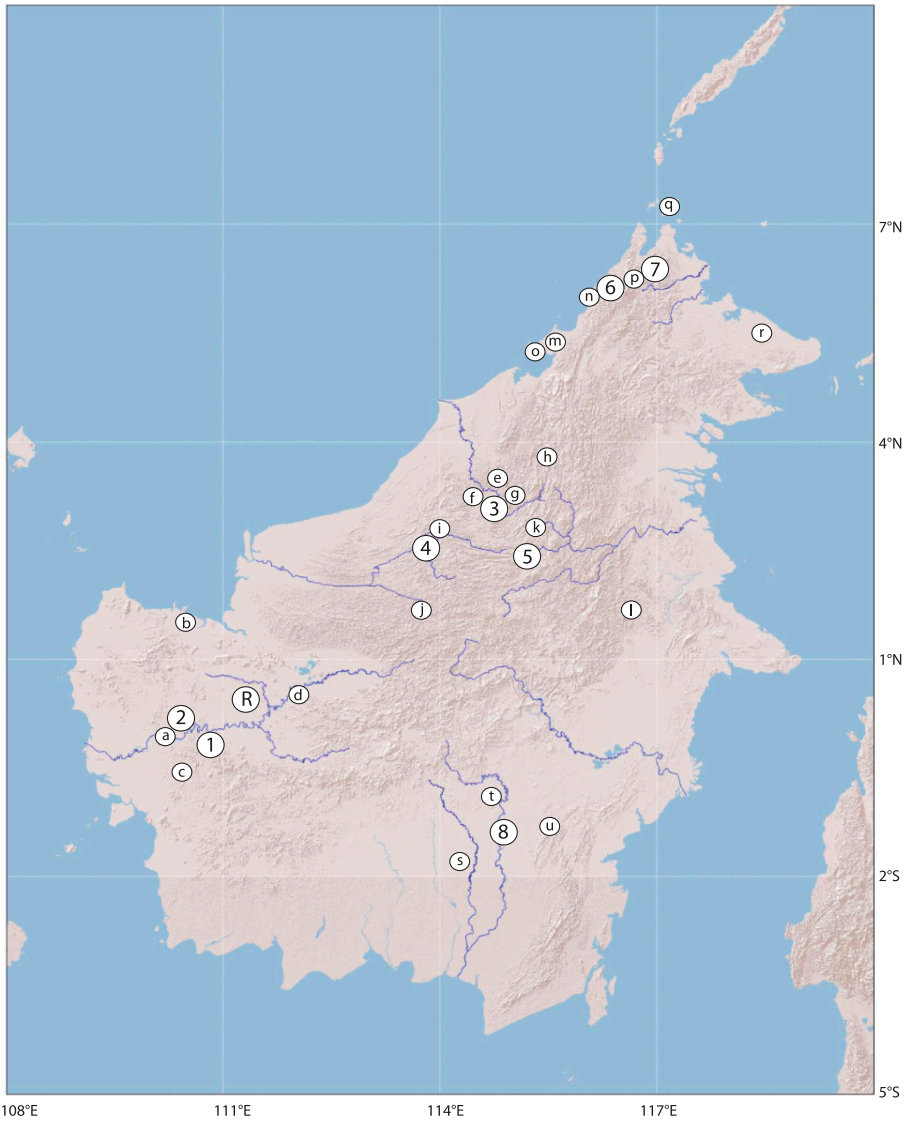


Figure 4. a = Bidayuh-Southern Land Dayak, b = Benyadu-Bekati', c = Kendayan-Ketapang, d = Keninjal-Iban, e = Berawan-Lower Baram, f = Lowland Kenyah, g = Highland Kenyah, h = Dayic, i = Melanau-Kajang, j = Punan-Müller-Schwaner, k = Kayan-Murik, l = Segai-Modang, m = BrDusun-SBisaya, n = Kadazan-Lotud, o = Burusu-Timugon Murut, p = Dumpas-Rungus, q = Bonggi, r = Idaan, s = Kapuas, t = Bakumpai-Ngaju, u = Basap-Tunjung

a group that heads towards the Kayan river headwaters (⑭), and another that heads downriver (⑪) We find that the upriver group is the Kayan-Murik subgroup and the downriver group is the Segai-Modang subgroup. These migration routs are intuitive, since Kayan-Murik languages remain mostly highland, and Segai-Modang have moved into lowland areas.

Furthermore, the downriver movement of Segai-Modang explains the separation of Basap from other Barito languages. Smith (2017a) first hypothesized that the movement of Segai-Modang speakers into their current position displaced existing Basap language speakers. Apparent borrowings between the two groups provided initial evidence, and our model suggests that the migration routes also support this history.

4.4.5 *Southwest Sabah and Northeast Sabah*

The homeland of Southwest Sabah was *coastal*, not riverine (⑥). This results in a mostly coastal spread as the proto-language began to diversify (⑬, ⑫, ⑩, ⑮).

Northeast Sabah is also placed near the coast, close to the Labuk river. The migration of speakers from this homeland has two patterns. First is Bonggi (⑨), which lies on an island separated from Borneo. The Idaan branch (⑦) is also quite far separated from the Proto-Northeast Sabah homeland, but as mentioned in Blust (2010), this may have been because of the movement of speakers of Southwest Sabah languages into historically Idaan territory.

4.4.6 *Barito*

The Barito homeland is placed along the middle course of the Barito river (⑧). Barito diversified along the Barito and other major rivers in the area. We note that the Tunjung-Basap node (⑪) is farther to the east due to the influence of Basap, which is currently located in far eastern Borneo.

4.4.7 *Overall migration patterns*

We find a strong correlation between the movement of speakers, and the paths of river systems. In some cases, the inferred homelands and migration patterns match what has been recorded in anthropological oral histories as well as what has been hypothesized through the linguistic comparative method. This gives us strong confidence in the accuracy of homeland inferences even for lower-level subgroups. The routes posited for Kayanic, Central Sarawak, Malayic, and Southwest Sabah, for example, complement the already well-understood histories of these subgroups, and give insight into how these subgroups interacted and the geographical impact that Borneo has had on the development of subgroups from early in the history of Bornean languages.

One issue with the phylogeographic model's inference, however, is the somewhat greater-than-expected inland orientation of the subgroups' root nodes. For example, North and Central Sarawak are placed in inland locations despite a likely history of more coastal alignment (these rivers were almost certainly settled from the coast, with migration from the coast traveling upriver). Determining the homelands for these groups is made difficult by the absence of a settle-from-the-coast constraint. A model that is better able to incorporate different geographical environments and constrain movement accordingly may result in more accurate homeland placement.[5]

5. Conclusion

We performed a first-of-its-kind Bayesian phylogenetic and phylogeographic analysis on a comprehensive dataset focused on the languages of Borneo. Our study is unique in both its focus and scope; it is focused on one area of Austronesian studies but within its focus has greater scope than any previous study. The output of the automatic cognate detection method was double checked by an expert linguist for accuracy and found to have a high level of accuracy at detecting cognates and also improved overall efficiency in the manual cognate detection task. Our analyses show support for several hypotheses regarding the composition of subgroups and the effect of geographical factors on homeland and migration.

The phylogenetic analysis showed support for several hypotheses on linguistic subgroups in Borneo, including both long-established subgroups as well as more recent proposals regarding mid and lower-level subgroups. Higher-order subgroups, on the other hand, differed from past hypotheses, particularly with regard to the division between Greater North Borneo and Barito, which did not appear in our tree. These results demonstrate the difficulties associated with making inferences on higher-order subgroups in Borneo. The GNB hypothesis, while plausible and supported by interesting qualitative evidence, does not appear in our tree.

Furthermore, we utilized phylogeographic methods and found that rivers were a significant factor in homeland placement and subsequent population movement for major Bornean subgroups. Conversely, the ultimate homeland (root node) was placed in southwest Borneo, which disagrees with what we know about the history of Austronesian settlement of the island. We posit a linguistic leveling event, which spread from southwest Borneo. Also, the strictly Bornean

5. An example of such a model has been attempted in Bouckaert et al. (2018).

focus of the study may influence root-node placement, and we hypothesize that a follow-up study inclusive of non-Bornean data will result in more historically accurate homeland placement without interference from the proposed leveling event.

Our phylogenetic dating resulted in a shallower-than-expected tree age. At nearly 3,000 years, our root age undershoots archaeological evidence by several centuries. Again, we interpret this as support for a leveling event in Borneo's linguistic past, consistent with statements from Smith (2017a). This implies that the subgroups found on Borneo today represent a fraction of the diversity that initially formed after initial Austronesian settlement.

We can identify areas for future research that will expand upon the analysis in this study. First, to infer a root node for Bornean languages more accurately, we will need to include non-Bornean languages in the analysis, and infer a tree with Bornean languages nested within the larger Malayo-Polynesian subgroup. We expect that including non-Bornean out-groups will both yield a northern entry point for the Bornean root node and also aid in assessing the linguistic position of Malayic and Land Dayak which were both expelled from the GNB subgroup in this study. Second, we recognize the need to develop a method to force a settle-from-the-coast constraint on the inferred homelands for major subgroups. A migration model that recognizes waterways (seas and rivers) as the primary method of migration may better orient homelands towards the coasts rather than towards the interior.

Acknowledgements

We want to thank Andrew Meade for his assistance with the BayesTraits software.

References

Adelaar, K. Alexander. 1989. Malay influence on Malagasy: Linguistic and cultural-historical implications. *Oceanic Linguistics* 28. 1–46. https://doi.org/10.2307/3622973

Adelaar, K. Alexander. 1992. *Proto-Malayic: The reconstruction of its phonology and parts of its lexicon and morphology*. Canberra: Pacific Linguistics.

Adelaar, K. Alexander. 2005. Malayo-Sumbawan. *Oceanic Linguistics* 44(2). 357–388. https://doi.org/10.1353/ol.2005.0027

Amigó, Enrique, Julio Gonzalo, Javier Artiles & Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12(4). 461–486. https://doi.org/10.1007/s10791-008-9066-8

Axelsen, Jacob Bock & Susanna Manrubia. 2014. River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences* 281(1784). 20133029.

Bellwood, Peter. 1995. Austronesian prehistory in Southeast Asia: Homeland, expansion and transformation. In Peter Bellwood, James J. Fox & Darrell Tryon (eds.), *The Austronesians: History and comparative perspectives*, 103–118. Canberra: Pacific Linguistics.

Bellwood, Peter. 2007. *Prehistory of the Indo-Malaysian Archipelago: Revised edition*. Canberra: Australian National University E-Press. https://doi.org/10.22459/PIMA.03.2007

Bellwood, Peter & Eusebio Dizon. 2005. The Batanes archaeological project and the "Out of Taiwan" hypothesis for Austronesian dispersal. *Journal of Austronesian Studies* 1(1). 1–32.

Bentz, Christian, Dan Dediu, Annemarie Verkerk & Gerhard Jäger. 2018. The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour* 2. 816–821. https://doi.org/10.1038/s41562-018-0457-6

Blust, Robert. 1974. A Murik vocabulary, with a note on the linguistic position of Murik. In Jerome Rousseau (ed.), *The peoples of Central Borneo* 22(43). 153–189. Special issue of the Sarawak Museum Journal.

Blust, Robert. 1985–1986. The Austronesian homeland: A linguistic perspective. *Asian Perspectives* 26. 45–67.

Blust, Robert. 2005. The linguistic macrohistory of the Philippines: Some speculations. In Hsiu Chuan Liao & Carl R. Galves Rubino (eds.), *Current issues in Philippine linguistics and anthropology: Parangal kay Lawrence A. Reid*, 31–68. Manila: Linguistic Society of the Philippines and SIL International.

Blust, Robert. 2010. The Greater North Borneo hypothesis. *Oceanic Linguistics* 49. 44–118. https://doi.org/10.1353/ol.0.0060

Blust, Robert. 2014. Some recent proposals concerning the classification of the Austronesian languages. *Oceanic Linguistics* 53. 300–391. https://doi.org/10.1353/ol.2014.0025

Blust, Robert. 2019a. The Austronesian homeland and dispersal. *Annual Review of Linguistics* 5. 417–434. https://doi.org/10.1146/annurev-linguistics-011718-012440

Blust, Robert. 2019b. The resurrection of Proto-Philippines. *Oceanic Linguistics* 58(2). 153–256. https://doi.org/10.1353/ol.2019.0008

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960. https://doi.org/10.1126/science.1219669

Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson. 2018. The origin and expansion of Pama-Nyungan languages across Australia. *Nature Ecology & Evolution* 2(4). 741. https://doi.org/10.1038/s41559-018-0489-3

Bowern, Claire & Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 817–845. https://doi.org/10.1353/lan.2012.0081

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. https://doi.org/10.1353/lan.2015.0005

Coedès, G. 1968. *The Indianized states of Southeast Asia*. Honolulu: University of Hawaii Press. Translated by Susan Brown Cowing.

Currie, Thomas E., Andrew Meade, Myrtille Guillon & Ruth Mace. 2013. Cultural phylogeography of the Bantu languages of Sub-Saharan Africa. *Proceedings of the Royal Society B: Biological Sciences* 280(1762). 20130695.

Dahl, Otto Chr. 1951. *Malgache et Maanjan: Une comparaison linguistique* (Studies of the Egede Institute 3). Oslo: Egede Instituttet.

Felsenstein, Joseph. 1992. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution* 46(1). 159–173.

Gavin, Michael C., Carlos A. Botero, Claire Bowern, Robert K. Colwell, Michael Dunn, Robert R. Dunn & Gregor Yanega. 2013. Toward a mechanistic understanding of linguistic diversity. *BioScience* 63(7). 524–535. https://doi.org/10.1525/bio.2013.63.7.6

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gray, Russell D., David Bryant & Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3923–3933. https://doi.org/10.1098/rstb.2010.0162

Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483. https://doi.org/10.1126/science.1166858

Greenhill, Simon J. 2014. Demographic correlates of language diversity. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 557–578. Routledge.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti & Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* 112(43). 13296–13301. https://doi.org/10.1073/pnas.1503793112

Holland, Barbara R., Katharina T. Huber, Andreas Dress & Vincent Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* 19(12). 2051–2059. https://doi.org/10.1093/oxfordjournals.molbev.a004030

Hua, Xia, Simon J. Greenhill, Marcel Cardillo, Hilde Schneemann & Lindell Bromham. 2019. The ecological drivers of variation in global language diversity. *Nature Communications* 10(1). 1–10. https://doi.org/10.1038/s41467-019-09842-2

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267. https://doi.org/10.1093/molbev/msj030

Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2). 245–291. https://doi.org/10.1163/22105832-13030204

Kaboy, Tuton. 1974. The Penan Aput. In Jérôme Rousseau (ed.), *The peoples of Central Borneo* 22(43). 287–293. Sarawak Museum Journal Special Issue.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430). 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lepage, Thomas, David Bryant, Hervé Philippe & Nicolas Lartillot. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* 24(12). 2669–2680. https://doi.org/10.1093/molbev/msm193

List, Johann-Mattis. 2012. SCA: Phonetic alignment based on sound classes. In D. Lassiter & M. Slavkovik (eds.), *New Directions in Logic, Language and Computation*. ESSLLI 2010, ESSLLI 2011. Lecture Notes in Computer Science, vol 7415. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31467-4_3

List, Johann-Mattis, Simon J. Greenhill & Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PloS One* 12(1). e0170046. https://doi.org/10.1371/journal.pone.0170046

Lobel, Jason William. 2016. *North Borneo sourcebook: Vocabularies and functors*, PALI Language Texts. Honolulu: University of Hawaii Press.

Mace, Ruth & Mark Pagel. 1995. A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 261(1360). 117–121. https://doi.org/10.1098/rspb.1995.0125

Nettle, Daniel. 1998. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology* 17(4). 354–374. https://doi.org/10.1006/jaar.1998.0328

Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Nichols, Johanna & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820. https://doi.org/10.1111/j.1749-818X.2008.00082.x

Pigeaud, Theodore G. 1962. *Java in the 14th century: A study in cultural history. The Nāgara-Kĕrtāgama by Rakawi Prapanca of Majapahit, 1365 AD*, vol. 3. The Hague: The Netherlands Institute for International Cultural Relations, 3rd edn.

Rama, Taraka. 2018. Similarity dependent Chinese Restaurant Process for cognate identification in multilingual wordlists. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 271–281. https://doi.org/10.18653/v1/K18-1027

Rama, Taraka, Johann-Mattis List, Johannes Wahle & Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, volume 2 (short papers)*, 393–400. https://doi.org/10.18653/v1/N18-2063

Rambaut, Andrew. 2012. Figtree v1. 4.

Ritchie, Andrew M. & Simon Y W. Ho. 2019. Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. *Journal of Language Evolution* 4(2). 108–123. https://doi.org/10.1093/jole/lzz005. URL https://doi.org/10.1093/jole/lzz005

Ronquist, Fredrik, Maxim Teslenko, Paul Van Der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard & John P. Huelsenbeck. 2012. Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61(3). 539–542. https://doi.org/10.1093/sysbio/sys029

Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill & Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences* 116(21). 10317–10322. https://doi.org/10.1073/pnas.1817972116

Sandin, Benedict. 1994. *Sources of Iban traditional history* (Sarawak Museum Journal Special Monograph 7). Kuching: Sarawak Museum.

Sellato, Bernard. 1994. *Nomads of the Bornean rainforest: The economics, politics, and ideology of settling down*. Honolulu: University of Hawaii Press. Translated by Stephanie Morgan.

Sellato, Bernard. 2001. *Forest, resources and people in Bulungan. Elements for a history of settlement, trade, and social dynamics in Borneo, 1880–2000*. Bogor: Center for International Forestry Research.

Smith, Alexander D. 2015a. On the classification of Kenyah and Kayanic languages. *Oceanic Linguistics* 53(2). 333–357. https://doi.org/10.1353/ol.2015.0016

Smith, Alexander D. 2015b. Sebop, Penan, and Kenyah internal linguistic classification. *Borneo Research Bulletin* 46. 172–193.
Smith, Alexander D. 2017a. *The languages of Borneo: A comprehensive classification*. Ph.D. thesis, University of Hawai'i at Mānoa.
Smith, Alexander D. 2017b. The Western Malayo-Polynesian problem. *Oceanic Linguistics* 56(2). 435–490. https://doi.org/10.1353/ol.2017.0021
Smith, Alexander D. 2018. The Barito linkage hypothesis with a note on the position of Basap. *Journal of the Southeast Asian Linguistic Society* 11. 13–34.
Smith, Alexander D. 2019a. A reconstruction of Proto-Segai-Modang. *Oceanic Linguistics* 58(2). 353–385. https://doi.org/10.1353/ol.2019.0012
Smith, Alexander D. 2019b. A second look at Proto-Land Dayak vowels. *Oceanic Linguistics* 58(1). 110–142. https://doi.org/10.1353/ol.2019.0005
Wichmann, Søren & Taraka Rama. 2020. Testing methods of homeland detection using synthetic data. *bioRxiv* https://doi.org/10.1101/2020.09.03.280826. URL https://www.biorxiv.org/content/early/2020/09/03/2020.09.03.280826
Xie, Wangang, Paul O. Lewis, Yu Fan, Lynn Kuo & Ming-Hui Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* 60(2). 150–160. https://doi.org/10.1093/sysbio/syq085
Yang, Ziheng. 1994. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* 39(1). 105–111. https://doi.org/10.1007/BF00178256
Zhang, Chi, Tanja Stadler, Seraina Klopfstein, Tracy A. Heath & Fredrik Ronquist. 2015. Total-evidence dating under the fossilized birth-death process. *Systematic Biology* 65(2). 228–249. https://doi.org/10.1093/sysbio/syv080
Zhang, Menghan, Shi Yan, Wuyun Pan & Li Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* 569(7754). 112. https://doi.org/10.1038/s41586-019-1153-z

Zusammenfassung

Diese Studie untersucht die Verwandtschaft und Geschichte der austronesischen Sprachen Borneos, der drittgrößten Insel der Welt und Heimat einer bedeutenden sprachlichen Vielfalt. Wir wenden bayessche phylogenetische Datierungsmethoden auf lexikalisch verwandte Daten an, die auf vier historischen Kalibrierungspunkten basieren, um eine datierte Phylogenie von 87 Sprachen abzuleiten. Die abgeleitete Baumtopologie stimmt mit den Vorschlägen zur Untergruppierung auf mittlerer und unterer Ebene überein, die auf der klassischen Vergleichsmethode basieren, sie deutet aber auf eine andere Organisation auf höherer Ebene hin. Das Wurzelal-ter des datierten Baumes ist flacher als die archäologischen Schätzungen, stimmt aber mit der Hypothese eines vergangenen sprachlichen Nivellierungsereignisses überein. Die aus einer bayess-chen phylogeografischen Analyse abgeleiteten Herkunftsgebiete der wichtigsten sprachlichen Un-tergruppen stimmen mit den Heimatvorschlägen der Archäologie und Linguistik überein. Die abgeleitete Heimat für vier der acht Untergruppen unterstützt die Hypothese der Flussheimat, wonach sich die größeren sprachlichen Untergruppen ursprünglich in Gemeinschaften entlang der Hauptflüsse Borneos entwickelten.

Résumé

La présente étude porte sur la relation et l'histoire des langues austronésiennes de Bornéo, qui est la troisième plus grande île du monde, qui abrite une importante diversité linguistique. Nous appliquons des méthodes de datation phylogénétique bayésienne à des données lexicales apparentées basées sur quatre points d'étalonnage historiques pour déduire une phylogénie datée de 87 langues. La topologie arborescente déduite est en accord avec les propositions de sous-groupes de niveaux moyen et inférieur basées sur la méthode comparative classique, mais indiquerait une organisation de niveau supérieur différente. L'âge de la racine de l'arbre daté est moins élevé que les estimations archéologiques, mais il correspond à l'hypothèse d'un nivellement linguistique passé. Les lieux d'origine supposés d'après les principaux sous-groupes linguistiques à partir d'une analyse phylogéographique bayésienne sont en accord avec les propositions sur ce point issues de l'archéologie et de la linguistique. Pour quatre des huit sous-groupes, l'emplacement de leurs foyers d'origine soutient l'hypothèse de la patrie fluviale : les principaux sous-groupes linguistiques se seraient développés initialement dans des communautés situées le long des principaux fleuves de Bornéo.

Address for correspondence

Alexander D. Smith
Department of Linguistics and Modern Languages
The Chinese University of Hong Kong
G17, G/F, Leung Kau Kui Building
Shatin, NT
Hong Kong SAR
alexanderdavidsmith@cuhk.edu.hk
https://orcid.org/0000-0003-2510-8555

Co-author information

Taraka Rama
Department of Linguistics
University of North Texas
taraka.kasicheyanula@unt.edu

Publication history

Date received: 11 May 2020
Date accepted: 26 March 2021
Published online: 19 January 2022