Human Biology

Volume 85
Issue 1 *Special Issue on Revisiting the "Negrito" Hypothesis*

Article 17

2013

Time and Place in the Prehistory of the Aslian Languages

Michael Dunn
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, michael.dunn@mpi.nl

Nicole Kruspe
Lund University, Lund, Sweden

Niclas Burenhult
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Time and Place in the Prehistory of the Aslian Languages

Abstract
The Aslian language family, located in the Malay Peninsula and southern Thai Isthmus, consists of four distinct branches comprising some 18 languages. These languages predate the now dominant Malay and Thai. The speakers of Aslian languages exhibit some of the highest degree of phylogenetic and societal diversity present in Mainland Southeast Asia today, among them a foraging tradition particularly associated with locally ancient, Pleistocene genetic lineages. Little advance has been made in our understanding of the linguistic prehistory of this region or how such complexity arose. In this article we present a Bayesian phylogeographic analysis of a large sample of Aslian languages. An explicit geographic model of diffusion is combined with a cognate birth-word death model of lexical evolution to infer the location of the major events of Aslian cladogenesis. The resultant phylogenetic trees are calibrated against dates in the historical and archaeological record to infer a detailed picture of Aslian language history, addressing a number of outstanding questions, including (1) whether the root ancestor of Aslian was spoken in the Malay Peninsula, or whether the family had already divided before entry, and (2) the dynamics of the movement of Aslian languages across the peninsula, with a particular focus on its spread to the indigenous foragers.

Keywords
Austroasiatic Languages, Aslian Languages, Phylogeography, Historical Linguistics

Time and Place in the Prehistory of the Aslian Languages

MICHAEL DUNN,[1]* NICOLE KRUSPE,[2] AND NICLAS BURENHULT[1,2]

Abstract The Aslian language family, located in the Malay Peninsula and southern Thai Isthmus, consists of four distinct branches comprising some 18 languages. These languages predate the now dominant Malay and Thai. The speakers of Aslian languages exhibit some of the highest degree of phylogenetic and societal diversity present in Mainland Southeast Asia today, among them a foraging tradition particularly associated with locally ancient, Pleistocene genetic lineages. Little advance has been made in our understanding of the linguistic prehistory of this region or how such complexity arose. In this article we present a Bayesian phylogeographic analysis of a large sample of Aslian languages. An explicit geographic model of diffusion is combined with a cognate birth-word death model of lexical evolution to infer the location of the major events of Aslian cladogenesis. The resultant phylogenetic trees are calibrated against dates in the historical and archaeological record to infer a detailed picture of Aslian language history, addressing a number of outstanding questions, including (1) whether the root ancestor of Aslian was spoken in the Malay Peninsula, or whether the family had already divided before entry, and (2) the dynamics of the movement of Aslian languages across the peninsula, with a particular focus on its spread to the indigenous foragers.

The Aslian branch of Austroasiatic is recognized as the oldest recoverable language family in the Malay Peninsula, predating the now dominant Austronesian languages. In this article we address the dynamics of the prehistoric spread of Aslian languages across the peninsula, including the languages spoken by Semang foragers, traditionally associated with the "negrito" phenotype.

In this article the received view of an early and uniform tripartite breakup of Proto-Aslian in the Early Neolithic period, and subsequent differentiation driven by

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.
[2]Lund University, Lund, Sweden.
*Correspondence to: Michael Dunn, Max Planck Institute for Psycholinguistics, 6500AH Nijmegen, The Netherlands. E-mail: michael.dunn@mpi.nl.

Human Biology, February–June 2013, v. 85, no. 1–3, pp. 383–400.

KEY WORDS: AUSTROASIATIC LANGUAGES, ASLIAN LANGUAGES, PHYLOGEOGRAPHY, HISTORICAL LINGUISTICS.

societal modes, is challenged. We present a Bayesian phylogeographic analysis of our data set of vocabulary from 28 Aslian varieties. An explicit geographic model of diffusion is combined with a cognate birth-word death model of lexical evolution to infer the location of the major events of Aslian cladogenesis. The resultant phylogenetic trees are calibrated against dates in the historical and archaeological record to extrapolate a detailed picture of Aslian language history.

We conclude that a binary split between Southern Aslian and the rest of Aslian took place in the Early Neolithic (~4,000 BP). This was followed much later in the Late Neolithic (2,000–3,000 BP) by a tripartite branching into Central Aslian, Jah Hut, and Northern Aslian. Subsequent internal divisions within these subclades took place in the early Metal phase (post-2,000 BP). Significantly, a split in Northern Aslian between Ceq Wong and the languages of the Semang was a late development and is proposed here to coincide with the adoption of Aslian by the Semang foragers.

Given the difficulties involved in associating archaeologically recorded activities with linguistic events, as well as the lack of historical sources, our results remain preliminary. However, they provide sufficient evidence to prompt a rethinking of previous models of both clado- and ethnogenesis within the Malay Peninsula.

Background

The Aslian languages fall into four subgroups (Diffloth and Zide 1992; Dunn et al. 2011), and the speakers of the various Aslian languages are traditionally divided into three ethnographically defined societal types (Benjamin 1985). The Southern Aslian subbranch is associated with the collector/trader group referred to as "Aboriginal Malay"; the Central Aslian languages are mostly associated with so-called Senoi cultures of semisedentary swidden horticulturalists; and the Northern Aslian languages are spoken predominantly by "Semang" groups, comprising nomadic rainforest foragers. The Semang are distributed across the northern part of the Malay Peninsula and the neighboring southern Thai Isthmus (see also Benjamin this issue; Lye this issue). The Maniq speakers of South Thailand are also classified as Semang (in Thai literature the Maniq are often called Sakai, a derogatory term in Malaysia). Apart from their societal mode, the Semang are also typically associated with a phenotype that distinguishes them from coexisting populations and includes dark skin, short stature, and tightly curled hair. These traits led to them being labeled "negrito" by early researchers. There are, however, mismatches between the mapping of linguistic, biological and social-economic groupings. For example, the Northern Aslian–speaking Ceq Wong do not belong to the Semang societal type, while some Semang groups, including the Lanoh, speak Central Aslian languages. Furthermore some Southern Aslian–speaking Semaq Beri engage in sympatric Semang-type foraging with their Semang neighbors (see Lye this issue). In the absence of genetic samples for these three groups, it is impossible to comment further on their biological profile. While all three groups have traces of ancient lineages, samples from core Semang

groups exhibit a significant genetic link to first migrations of modern humans into the region (50 kya), (see below; see also Jinam et al. this issue). While the Aslian languages present as a robust linguistic grouping, the attested biological and cultural diversity has persisted as an intriguing and challenging backdrop. The aforementioned mismatch between the ethnographic and linguistic classifications of the various Aslian groups based on our same sample of 28 Aslian varieties has been discussed in previous publications (Dunn et al. 2011; Burenhult et al. 2011). Here we move away from the societal focus and introduce phylogeographic models to explore the Aslian languages from a new perspective. We test (and confirm) the generally held position that Proto-Aslian developed and then later split up already in situ in the Malay Peninsula.

The Neolithic period in the Malay Peninsula appeared approximately 4,000 years BP with the appearance of cord-marked pottery, stone axes and bark-cloth beaters and extended burials associated with the Ban Kao assemblages from south-central Thailand. This technological innovation is usually associated with the first migration of Austroasiatic speakers into the area and has been interpreted as an example of the demic diffusion model of language expansion driven by agriculturalists (Bellwood 1992). The extant archaeological record suggests that this migration initially took place down the west coast. The suggested homeland for Proto-Aslian has been situated west of the Main Range, in the Malay Peninsula (Benjamin 1985). Soon after this in the Early Neolithic phase Proto-Aslian underwent a tripartite split into the new clades—Proto-Northern, -Central, and -Southern Aslian—and they subsequently spread out from the original homeland across the Peninsula (Benjamin 1997; Bulbeck 2004). Following the developments in the Early Neolithic, the next major event of significance is the early Metal phase (~2,000 BP), which saw the intensification of state formation in the lowlands. The impact of intrusive external societies has been associated by Benjamin (1985) with the development of the different societal modes among the Aslians, and in particular the emergence of an intensive focus on foraging among the Semang.

The dates inferred for the origins of Aslian in this article are consistent with Bulbeck's and Benjamin's dates (see "Results"). It is important to note that the migration of Aslian speakers into the peninsula may not have taken place as an overwhelming wave of replacement agriculturalist settlement at the expense of indigenous foragers. Modeling of genetic drift by Fix (2011) shows that a long-term trickle of Aslian-speaking migrants into small communities could have produced the same effect.

The scenario of the entry of Proto-Aslian into the Peninsula followed by an expansion based on uniform cladogenesis in the Early Neolithic presented above is problematic on a number of counts, including that (1) the irregular rates of lexical change within the three major subbranches do not reflect a purported parallel splits at the subbranch level, and (2) it does not seriously entertain the possibility of language shift, and in particular relatively recent language shift, as a factor in accounting for the biological and cultural diversity represented within Aslian.

Materials and Methods

Theory. In this study, as in previous work by the same authors (Burenhult et al. 2011; Dunn et al. 2011), the relationships between the Aslian languages are inferred using methods from a family of statistical techniques known as Bayesian phylogenetic inference. In these methods, an algorithm is used to infer a set of parameters to a mathematic model of evolutionary change that best explain a set of observations. The parameters encoded by the evolutionary model minimally include the topology of a phylogenetic tree and some expression of the rates of change of the characters being modeled. More complex—and more realistic—models can also be implemented. The function for maximizing the likelihood value of the model is however computationally costly. To render this calculation tractable, Bayesian phylogenetic inference uses a Monte Carlo Markov chain (MCMC) process to search the space of possible parameters to the evolutionary model for the zones of highest likelihood. The results of the MCMC search are a set of parameter values sampled in proportion to their likelihood, in effect giving a sample of more or less equally high-likelihood possibilities for explaining the phylogenetic relationships between the observed taxa. Methods for interpreting and summarizing this sample are discussed below.

Not surprisingly, different evolutionary models produce somewhat different results. The performance of an evolutionary model can be evaluated by two criteria: likelihood and convergence. Each sample from the Markov chain parameter search has a likelihood score. Initial samples in the chain are generally poor fits to the data, but ideally, as the search progresses, the likelihood score increases as regions of the parameter space with more plausible tree topologies and parameter values are found. The convergence of a model is simply whether the search settles on a region of the search space with uniformly high likelihood. The part of the search prior to convergence is called the "burn-in" and is discarded. The mean likelihood of the post-burn-in sample can be compared between different models: a special kind of likelihood ratios test called the Bayes factor allows a conventional interpretation of the degree to which one model can or cannot be preferred over another.

The models of evolution used in this analysis are made up of two parts, the clock model and the substitution model. The clock model expresses how rates of change are allowed to vary: in a strict clock model, the rate in each rate category is fixed over the entire tree; in a relaxed model, rates vary over branches. The substitution model expresses the patterns of change that cognates undergo: in a simple model, reflexes of cognate sets can be gained or lost; in a gamma model, there is a statistical distribution of different rate classes of change. Under the stochastic Dollo model, cognate sets rarely come into being, but reflexes can always be lost, and in the covarion model branches of the tree are allowed to vary between faster and slower rates of change. Furthermore, most of these parameters can be combined to make more complex models of evolution. These models each have

their advantages and disadvantages. The covarion model is good at accounting for varying rates of change, while only requiring a simple "switch-rate" parameter. Gamma models allow for different rates of change of different characters, but each character is assumed to have a fixed rate over the tree. In language trees the overall shifts in rate of evolution captured by the covarion model often overwhelm the effects of the differences in rates of individual characters captured by the gamma model. Stochastic Dollo models are designed for linguistic realism, best approximating the known processes of cognate substitution. Dollo models are, however, sensitive to undetected borrowings and incorrect cognate coding (e.g., based on chance similarity) in the data. Model selection in phylogenetic inference is—as in modeling generally—a matter of trading off realism against predictiveness by minimizing the number of parameters and avoiding over-fitting of the data.

The raw inferences from a Bayesian phylogenetic analysis are in the form of trees with branch length indicating "substitutions," the amount of evolutionary change that has occurred on a branch. These substitution trees are transformed into trees scaled to time by use of an algorithm that infers the branch-specific variation in rates. In a sample of trees sampled contemporaneously, the terminal nodes (the nodes representing the observed taxa) should all be aligned if the branches are scaled to time (i.e., the sum of branch lengths from tip to root should be the same for any tip). In a tree scaled by substitutions the terminal nodes are ragged, since some terminal nodes have been subject to more evolutionary change than others. The rate smoothing algorithm infers the branch-specific rate variation necessary to morph the substitution tree into a time scaled tree, given that the temporal position of the terminal nodes and some of the internal nodes are fixed (or at least probabilistically fixed to a particular distribution of likely values). The inferred values for rates of change along branches can subsequently be used to calculate the dates of other, previously unspecified nodes of the tree. Gray et al. (2011) discuss how these methods are applied specifically to the dating of language trees.

We adopt an approach to inferring spatial diffusions that simultaneously reconstructs evolutionary history in space and time (Lemey et al. 2010). These kinds of methods have been applied to language data from the Arawakan (Walker and Ribeiro 2011) and Indo-European (Bouckaert et al. 2012) families. Spatial data are treated as a distinct partition, subject to different evolutionary models to the lexical data. Languages are treated as inhabiting a continuous landscape, and spatial diffusion is modeled as a Brownian process. Two different models are tested: a simple continuous model—the standard random walk model, and a "relaxed random walk" (RRW) model, which allows for branch-specific variation in rate of dispersal (in analogy to the relaxed random clocks). While more complex spatial models are certainly possible, Brownian models give us a simple, and well-tested, baseline. In addition, there are in fact good reasons to begin from the assumption that incremental spatial change accounts for the true diffusional process. In particular, linguistic data demonstrably shows spatial autocorrelation (Dunn 2009)—spatial dispersion tracks language genealogy—as would be expected under a Brownian process.

Methods. This analysis makes use of phylogeographic methods, a form of Bayesian phylogeographic inference that incorporates lexical and geographic partitions of data and is implemented in the software package BEAST (Lemey et al. 2010; Drummond et al. 2012). The lexical partition is analyzed according to a cognate birth-word death model of lexical change, and the geographic data are analyzed according to a Brownian model of spatial diffusion. In addition, certain nodes of the inferred trees are calibrated according to known historical/archaeological events, and a rate-smoothing algorithm is used to transform the tree into one where branch length indicates calendar years (i.e., rather than "substitutions," or amount of evolutionary change). The phylogenetic hypotheses sampled in the MCMC tend to have low likelihood until the search has run long enough to identify the high-likelihood regions of parameter space. A trace of the likelihood over the chain can be plotted, and the experimenter can identify whether the likelihood trace has converged on a stable level. The leading section of the trace, prior to convergence, is discarded.

The lexical data used in this analysis takes the form of cognate candidates taken from wordlists collected in situ during periods of fieldwork in the period 1990–2009 (sources listed Dunn et al. 2011: 299, Table 1). Words were sought corresponding to 146 basic meanings using a regionally adapted Swadesh-type list. Malay loanwords were excluded (112 instances of borrowing from 55 distinct Malay source words); chance resemblances were excluded where they could be identified. The database contained 984 distinct Aslian cognate sets. The cognate candidates were identified according to explicit criteria, devised according to language-family specific knowledge of general patterns of language change in the Aslian family. The primary criterion is place of articulation of onset and coda for the final syllable; a fuller account of these principles and their exceptions is given in Dunn et al. (2011: 300).

Languages are classified geographically according to their known historical center. Where we know about modern removals and resettlements we have classified languages in their historical location. The geographic centers of the languages in our sample are shown in Figure 1.

The rate variation in the Aslian phylogenetic tree is calculated by inference from known calibration points—internal nodes of the tree that can be associated, on nonlinguistic grounds, with a datable archaeological or historic event. These calibrations are expressed in terms of probability distributions. The two kinds of distributions we make use of are the normal distribution, for when we believe in a particular most likely date with an estimated error factor distributed evenly to either side, and the log-normal distribution, for when there is a skewed uncertainty, for example, for a date that could well be earlier than our best estimate but that is very unlikely to be later.

Bulbeck (2004) identifies the following eras in Aslian prehistory: Early Neolithic (~4,000 BP), Late Neolithic (~3,000 BP), early Metal phase (~2,000 BP), and late Metal phase (~1,000 BP). Some of the splitting events discussed in Bulbeck are not supported by our reconstructions and so are not usable as calibrations.

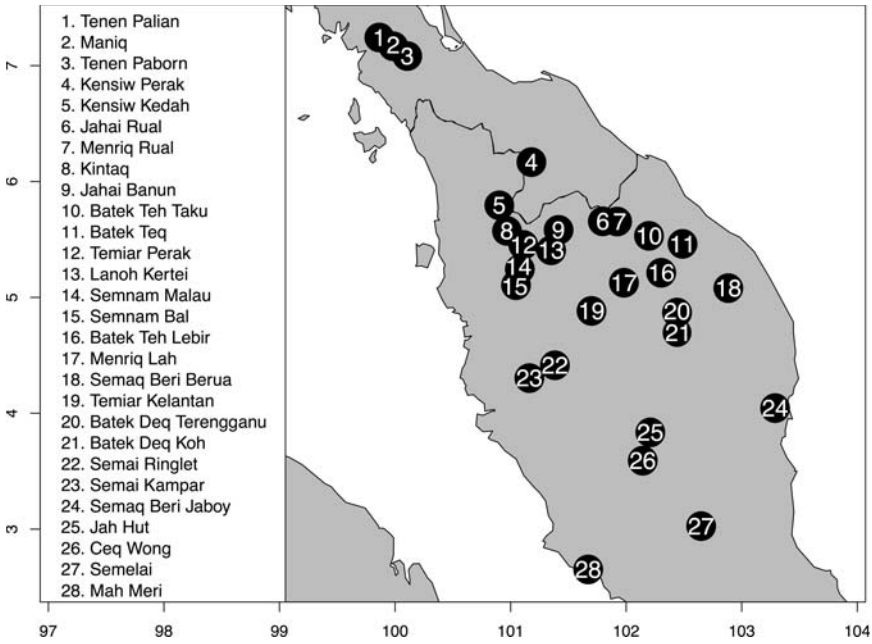

Figure 1. The geographic centers of the 28 Aslian varieties analyzed.

Identifying calibration points is fraught with peril in the absence of written historical evidence. The present analysis concerns a nonliterate environment that has lasted from prehistory to the present. Thus, we rely entirely on archaeologically datable events whose connection to linguistic events can be regarded as particularly well motivated. We motivate our calibration points as follows:

- One calibration point should be present somewhere within each major subbranch of Aslian (Northern, Central, and Southern); one calibration point should involve the Proto-Aslian starting point.
- Our choice of archaeologically dated events rely predominantly on Bulbeck (2004); Bulbeck proposes three alternative scenarios of Aslian spread, but these differ mostly in the order of more recent changes. We have selected dates that are explicitly linked to protostages that remain the same across all his different scenarios: Proto-Mah Meri in the Southern Aslian clade and Proto-Temiar-Lanoh in the Northern Aslian clade.
- The calibration point within Northern Aslian—Proto-Maniq-Menraq/Bate—is not from Bulbeck 2004 but a hypothesis put forward by Benjamin (1985: 244, 261) that Semang cultural development is linked to the rise of Sathing Phra, a first millennium AD trading civilization on the southern Isthmus of Kra (Andaya 2008); its rise coincides with the beginning of the early Metal period.

Table 1. Calibration Points Used to Estimate the Dating of the Aslian Tree

CALIBRATION POINT	DATE	DESCRIPTION
Root of Aslian	Normal distribution: 4,000 BP ± 500	Early Neolithic. Note that this calibration was *not* used in the Aslian root date inference in Figure 2.
Proto-Mah Meri	Normal distribution: 2,000 BP ± 200	All three scenarios in Bulbeck 2004 connect Proto-Mah Meri (or "Proto-Besisi," in his terminology) with Early Metal Phase coastal finds at 2000 BP.
Proto-Temiar-Lanoh	Lognormal distribution: around 2,000 BP or earlier	Proto-Temiar-Lanoh; all three scenarios in Bulbeck (2004) connect Proto-Temiar-Lanoh with the appearance of bronze in the Lenggong Valley by 2,000 BP.
Proto-Maniq-Menraq/ Batek	Normal distribution: 1,750 BP ± 250	Growth of Sathing Phra civilization in South Thailand, starting around 2,000 BP and expanding significantly around 1,500 BP; Benjamin connects development of Semang culture with this growth; start of Sathing Phra coincides with the start of the early Metal phase; most likely timing of Proto-MMB split at the beginning of this period? (based on Benjamin 1985; Andaya 2008)

Near-contemporary language samples were taken between 3 and 40 years BP but were uniformly treated as originating from the present.

The calibration points identified for our study are listed in Table 1.

Results

Model Choice. The phylogenetic models of lexical evolution run in BEAST consist of a clock model (strict or relaxed) and a substitution model (simple, gamma-distributed, Dollo, and covarion, as well as combinations, e.g., covarion + gamma). The spatial diffusion analysis adds to this a partition with a diffusion model (standard or relaxed); where there is branch-specific variation in rate of diffusion, it is additionally necessary to select the distribution of rates (Cauchy, exponential, lognormal, or gamma).

The covarion model with a relaxed clock outperforms (strongly) all other models except for relaxed clock + covarion + gamma, but even in this case the relaxed clock plus covarion is weakly preferred (see Table 2). Since the parameters of the weakly preferred model form a simpler subset of the other model, there are additional good reasons to prefer it: all else being equal, a simpler model is generally more predictive than a more complex model, and all the more so in this case where addition of the gamma parameter actually makes the model fit worse.

Table 2. Bayes Factor Comparisons of Models of Lexical Evolution

	MODEL 2								
MODEL 1	STRICT + SIMPLE	STRICT + GAMMA	STRICT + DOLLO	RELAXED + SIMPLE	RELAXED + GAMMA	RELAXED + COVARION	RELAXED + DOLLO	STRICT + COVARION + GAMMA	RELAXED + COVARION + GAMMA
Strict + simple	—	–10.62	86.76	–28.78	–38.59	–58.92	73.19	–25.62	–55.36
Strict + gamma	10.62	—	97.39	–18.162	–27.97	–48.30	83.82	–14.99	–44.74
Strict + Dollo	–86.76	–97.39	—	–115.55	–125.36	–145.69	–13.57	–112.39	–142.13
Relaxed + simple	28.78	18.16	115.55	—	–9.80	–30.14	101.98	3.16	–26.58
Relaxed + gamma	38.59	27.97	125.36	9.80	—	–20.33	111.79	12.97	–16.77
Relaxed + covarion	**58.92**	**48.30**	**145.69**	**30.14**	**20.33**	—	**132.12**	**33.30**	**3.56**
Relaxed + Dollo	–73.19	–83.82	13.57	–101.98	–111.79	–132.12	—	–98.82	–128.56
Strict + covarion + gamma	25.62	14.99	112.39	–3.16	–12.97	–33.30	98.82	—	–29.74
Relaxed + covarion + gamma	55.36	44.74	142.13	26.58	16.77	–3.56	128.56	29.74	—

A positive value means that model 1 (rows) is preferred over model 2 (columns). The relaxed covariation model (boldface) is preferred overall.

Holding the lexical evolution model constant with the relaxed clock and covarion, we then added the spatial data partition and tested the different spatial diffusion models. RRW diffusion models fit the data better than standard diffusion models, with a $\log_{10}$ Bayes factor of 5.2 (a value that is conventionally treated as "strong" evidence to prefer this model). This tells us that the data are better accounted for in a spatial diffusion model that allows branch-specific variation in rates. The phylogenetic trees and parameter estimates emerging from these two models are for the most part indistinguishable. The only major difference between the results of the two models is that the estimated age of the Temiar-Lanoh clade is considerably lower under the RRW model than the standard diffusion model (discussed below; see Figure 4). Figure 2 shows the maximum clade credibility phylogenetic tree of Aslian using the highest-likelihood phylogeographic model—a covarion model of lexical evolution using a relaxed clock with an RRW model of spatial diffusion. This analysis successfully recovers the known structure of the Aslian family: the three major subgroups, Northern, Central, and Southern, are clearly distinguished in clades with 100% posterior probability. Jah Hut is not strongly associated with any other group, consistent with the claim that Jah Hut forms a single member sister group to the other three subgroups (Dunn et al. 2011; Burenhult et al. 2011).

Proto-Aslian. In order to estimate the age of Proto-Aslian, we removed the constraints on root age of the tree (see Table 1) and reran the analysis retaining only the calibration points from the more recent events in the history. The highest-likelihood models are in close agreement: the covarion model (highest likelihood) gives a distribution of dates for Proto-Aslian with a maximum probability density at 4,305 BP and a 95% highest probability density range of 2,943–6,570 BP. This is consistent with an origin in the early Neolithic, although slightly earlier than the dates used in the calibration (mean ± 4,000 BP ± 500).

Benjamin proposes a starting point or homeland for Proto-Aslian to the west of the Main Range (Benjamin 1985, 1997), with a later secondary dispersal east of the range on the southern slopes of Gunung Benom. Bulbeck (2004: 377) suggests an area in the south-central west that straddles the main range. His homeland hypothesis is motivated as follows:

> A zone containing the maximum diversity of a group of related languages is parsimoniously treated as the homeland for those related languages. It is less presumptive to assume that these diverse representatives are the relatively stationary relics of an earlier diversification than to assume that they have all by chance converged on the same location from some other homeland.

The position of the homeland in our analysis (Figure 3) coincides with Bulbeck's eastern extent, and Benjamin's location of secondary dispersal. Uncontroversially, the phylogeographic model infers a most likely origin near the center of phylogenetic diversity. This is considerably to the south of the geometric center of the attested languages.

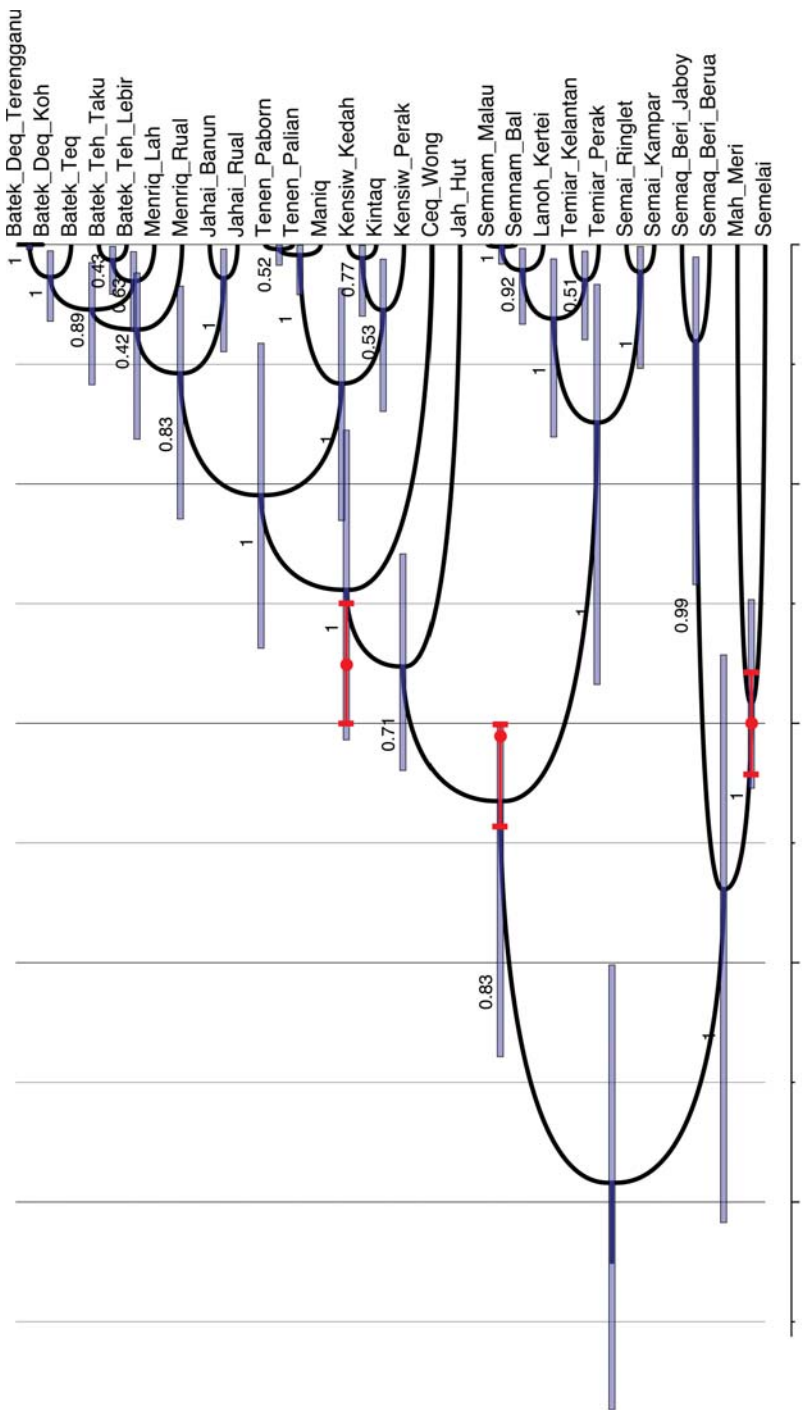

Figure 2. A phylogenetic tree of Aslian using the highest-likelihood phylogeographic model: a covarion model of lexical evolution using a relaxed clock, with a relaxed random walk model of spatial diffusion. The tree is calibrated to time before present; the blue bars show the 95% confidence intervals of the node dates, and the numbers on the branches show the posterior probability of the following node.

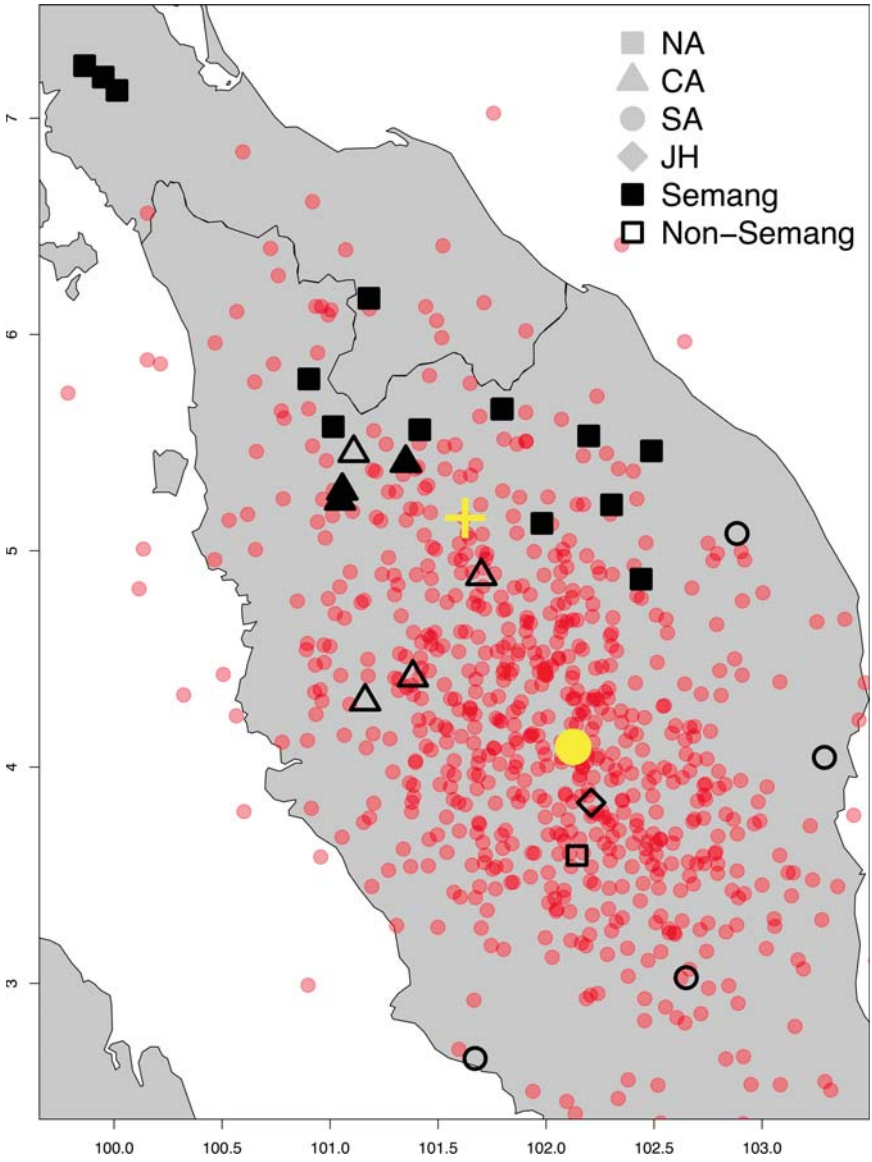

Figure 3. The distribution of inferred origin points (red) for the root node of Proto-Aslian. The yellow dot shows the maximum probability density value, which represents the most likely homeland inferred by the analysis. By way of contrast, the cross shows the geometric center of the attested languages. The locations of present-day languages (showing linguistic affiliation and societal type) are shown for orientation purposes. Abbreviations: NA, Northern Aslian; CA, Central Aslian; SA, Southern Aslian; JH, Jah Hut.

As discussed earlier, the inferred dates for Proto-Temiar-Lanoh are considerably more recent under the RRW model than the standard model (see Figure 4). While the 95% confidence interval includes almost any time up to nearly 2,000 years BP, the median value is only 699 years before present (early fourteenth century CE). The RRW was strongly preferred over the standard model, with a Bayes factor of 5.222. This is not unexpected, since the relaxed model of diffusion better fits what we know about real-world diffusion of languages and peoples—that rates of spatial diffusion are highly heterogeneous. But it is interesting that with this one notable exception, the inferred data ranges for most clades were not so different under each of these two models. This suggests that Proto-Temiar-Lanoh is the product spatial process involving a period of much faster diffusion than is typical in the history of the other Aslian languages.

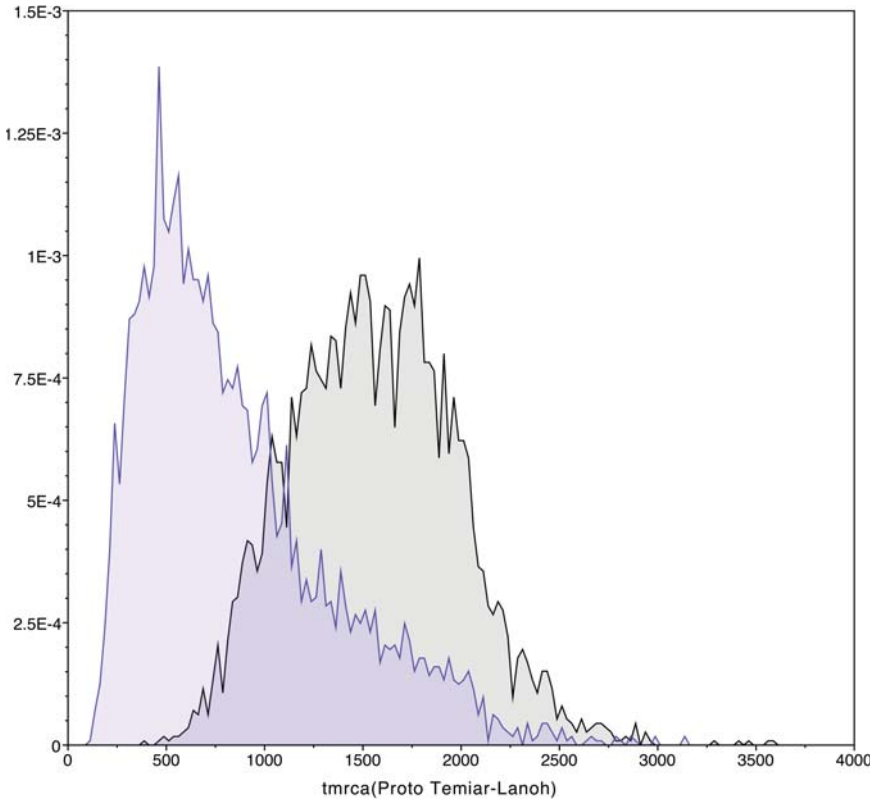

Figure 4. The inferred date range of the time to most recent common ancestor (tmrca) of Temiar-Lanoh under the standard spatial diffusion model (gray) and the relaxed random walk model (blue).

Discussion and Conclusions

This analysis represents a first investigation of Aslian prehistory by means of phylogeographic techniques. Although largely exploratory and clearly subject to future refinements, our study provides some findings of core concern to peninsular prehistory and to our understanding of biological, cultural and linguistic dynamics in the region:

1. Our analysis firmly supports the long-held ideas that Austroasiatic languages arrived in the peninsula from the outside with the Neolithic and that this arrival coincides with Proto-Aslian. The most likely homeland of Proto-Aslian is in the very center of the Malay Peninsula, on the slopes of Gunung Benom, just east of the Main Range (the most prominent mountain group on the Malay Peninsula). This area may represent an initial settlement of Austroasiatic colonizers from outside or, perhaps more likely, a secondary geographic bottleneck in the very earliest history of Austroasiatic in the peninsula. The location is in concord with Early Neolithic dates at sites like Gua Kecil (Bulbeck 2004: 371) but does not quite coincide with previously suggested homelands, which are typically placed west of the Main Range (Benjamin 1997; Bulbeck 2004).
2. Like our previous studies of Aslian lexicon (Dunn et al. 2011; Burenhult et al. 2011), this one suggests that the first Aslian split was binary (Southern Aslian vs. the rest) and occurred during the Early Neolithic, shortly after the arrival of Austroasiatic. A secondary split occurred in the non-Southern Aslian branch during the Late Neolithic, some 1,500 years later (2,500 BP), separating Northern Aslian, Central Aslian, and Jah Hut. This departs from previous reconstructions, which involve a tripartite split between Southern, Central, and Northern Aslian in the Early Neolithic.
3. Subsequent branching within Northern-Central-Jah Hut occurred during the course of the Metal Age. Our analysis presumes that, within Northern Aslian, the Maniq-Menraq/Batek (MMB) and Ceq Wong branches split much later than previously assumed (compare the Neolithic date proposed by Bulbeck 2004). This is significant in the context of Semang forager relationships with Aslian (see below). Furthermore, Maniq is not an early offshoot from MMB, as suggested by Bulbeck 2004, but well contained within a Maniq/Kensiw/Kentaq subclade. This is firmly supported throughout our lexical analyses. Within Central Aslian, our analysis suggests very late splits. Conversely, within Southern Aslian, the analysis supports an unexpectedly early (Late Neolithic) first branching.

Our results point to hitherto unrecognized irregularities in the chronology of Aslian branching. In particular, the conventional three subclades of Aslian (Northern, Central, Southern) cannot lay claim to equal continuity from the Early Neolithic. Southern Aslian did split off in the early history of Aslian, but the splitting up of

Northern, Central, and Jah Hut was propelled by dynamics occurring during the Late Neolithic and early Metal age, some 1,500 years later. Turning specifically to Northern Aslian, we observe that its history as a distinct clade goes no further back than around 1,500–2,000 BP and that soon after its crystallization it split up into two subclades, Ceq Wong and MMB. The latter subclade has since experienced a number of splits.

This chronologically skewed model of Aslian genealogy is congruent with the very unequal rates of lexical diversification observed in our previous analyses (Burenhult et al. 2011; Dunn et al. 2011). These reveal a cline in rates of lexical change, where Southern Aslian languages display the lowest rates of change, and the MMB clade of Northern Aslian display the highest (but shows low internal diversity). Present-day Semang society enforces a mobile lifestyle and social structures that encourage dispersal and flux in personal relations and space. While the present does not provide a window to the past, such a societal mode may have been a contributing factor to the high rates of lexical change in MMB. What is striking in the context of the present analysis is the relationship between MMB and its sister branch Ceq Wong, which together make up Northern Aslian. Ceq Wong has not experienced the very high rate of lexical change displayed by MMB as a whole. Thus, the recent date of the Northern Aslian branch transpiring from the present study, and the rapid subsequent split of the clade into MMB and Ceq Wong, suggest that the branching off of MMB was a significant, sudden and perhaps dramatic event in the history of Aslian. As noted, the MMB clade coincides with the foraging Semang societal mode, and we have argued elsewhere that the MMB split represents the first entry of Aslian into this forager niche (Burenhult et al. 2011). The results of the present study lend support to this interpretation.

The new perspective on Aslian genealogy presented here provides an interesting framework for discussing language in relation to genetics. Many previous accounts of Aslian prehistory have emphasized the shared biological, cultural and linguistic roots of Aslian speakers (see especially Rambo 1988; Fix 1995). However, recent mitochondrial analyses reveal that the Aslian-speaking populations carry genetic lineages of diverse geographical origins and varying antiquity in the peninsula (Hill et al. 2006, 2007). In particular, locally ancient lineages (suggested to have persisted in the peninsula since the Pleistocene and as far back as 50–60 kya) are present in all sampled indigenous peninsular populations, but to varying degrees. Northern Aslian (specifically MMB) speakers, who coincide with the foraging Semang societal mode, show a particularly close connection with these ancient lineages, whereas Semelai (Southern Aslian) speakers show only traces of them and a high proportion of lineages originating outside the peninsula. Temiar (Central Aslian) speakers display an equal mix of the locally ancient and external lineages.

Consequently, it is clear that the well-established and genealogically and geographically well-contained Aslian language group, introduced to the peninsula from the outside some 4,300 years ago, has evolved, spread and diversified in a genetically heterogeneous setting. The significant indigenous component in

this course of events makes it clear that we have to take into consideration the phenomenon of *language shift*, that is, the process through which a population abandons a language in favor of another (Sasse 1992). To the extent that it has been considered at all, a language shift to Austroasiatic in the peninsula has in previous work been addressed only in terms of a "paleo-sociolinguistic problem" restricted to the Semang and, as such, an obstacle to models proposing a distinct and locally ancient origin (Benjamin 2002: 35). But it is becoming increasingly clear that the history of most and perhaps all of Aslian should be considered in the context of varying degrees of genetic intermixing between local and foreign lineages and successive situations of language expansion and shift. The skewed chronology of Aslian subbranching distinguished by the present study, as well as the pan-Aslian cline in rates of lexical change, is congruent with such a successive and drawn out spread of Aslian across diverse populations. The MMB-speaking Semang foragers only represent an extreme end of this spectrum, with a shallow Aslian identity, a retained forager mode of subsistence, and a high proportion of locally ancient and pre-Neolithic genetic lineages. Arguably they are the most recent Aslian conquest in a chain of shift events. The Southern Aslian-speaking groups sit at the other end of the spectrum, with distinct Aslian continuity since the Early Neolithic, lexical conservatism, and a much smaller proportion of locally ancient genetic lineages.

As noted, our results support the idea that Austroasiatic arrived in the peninsula with the Early Neolithic. Agriculture and its associated demographic characteristics must have had some significant role to play in the prelude to this process. But we observe that most of the major branchings in Aslian occurred in response to a range of later dynamics, and the advent of metal around 2 kya gives the impression of being the single most significant factor for the successful spread and current appearance of the family.

Acknowledgments Dunn's research was supported by the Max Planck Society and the Max Planck Institute for Psycholinguistics, Nijmegen; Kruspe's research was supported by the Max Planck Institute for Evolutionary Anthropology, the Research Center for Linguistic Typology at LaTrobe University, the Hans Rausing Endangered Languages Program, and the Bank of Sweden Tercentenary Foundation; Burenhult's research was supported by the Swedish Research Council (421-2007-1281), the Volkswagen Foundation (DOBES), and the European Research Council (the European Union's Seventh Framework Programme, Grant agreement n° 263512).

Received 9 October 2012; revision accepted for publication 22 April 2013.

Literature Cited

Andaya, L. Y. 2008. *Leaves of the Same Tree: Trade and Ethnicity in the Straits of Melaka*. Honolulu: University of Hawai'i Press.

Bellwood, P. 1992. Southeast Asia before history. In *The Cambridge History of Southeast Asia,* Vol. 1, *From Early Times to c. 1800*, N. Tarling, ed. Cambridge: Cambridge University Press, 55–136.

Benjamin, G. 1985. In the long term: Three themes in Malayan cultural ecology. In *Cultural Values and Human Ecology in Southeast Asia*, K. L. Hutterer, A. T. Rambo, and G. Lovelace, eds. University of Michigan: Center for South and Southeast Asian Studies, 219–278.

Benjamin, G. 1997. Issues in the ethnohistory of Pahang. In *Pembangunan Arkeologi Pelancongan Negeri Pahang*, Nik Hassan Shuhaimi bin Nik Abdul Rahman, Mokhtar Abu Bakar, Ahmad Hakimi Khairuddin, and Jazamuddin Baharuddin, eds. Pekan: Muzium Pahang, 82–121.

Benjamin, G. 2002. On being tribal in a Malay world. In *Tribal Communities in the Malay World: Historical, Cultural and Social Perspectives*, G. Benjamin and C. Chou, eds. Singapore: Institute of Southeast Asian Studies, 7–76.

Benjamin, G. 2013. Why have the peninsular "negritos" remained distinct? *Hum. Biol.* 85:445–484.

Bouckaert, R., P. Lemey, M. Dunn et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337:957–960.

Bulbeck, D. F. 2004. An integrated perspective on Orang Asli ethnogenesis. In *Southeast Asian Archaeology*, V. Paz, ed. Dilimon, Quezon City: University of the Philippines Press, 366–399.

Burenhult, N., N. Kruspe, and M. Dunn. 2011. Language history and culture groups among Austroasiatic-speaking foragers of the Malay Peninsula. In *Dynamics of Human Diversity: The Case of Mainland Southeast Asia*, N. J. Enfield, ed. Canberra: Pacific Linguistics, 257–275.

Diffloth, G., and N. Zide. 1992. Austro-Asiatic languages. In *International Encyclopedia of Linguistics,* Vol. 1, William Bright, ed. New York: Oxford University Press, 137–142.

Drummond, A. J., M. A. Suchard, D. Xie et al. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29:1,969–1,973.

Dunn, M. 2009. Contact and phylogeny in island Melanesia. *Lingua* 119:1,664–1,678.

Dunn, M., N. Burenhult, N. Kruspe et al. 2011. Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronica* 28:291–323.

Fix, A. G. 1995. Malayan paleosociology: Implications for patterns of genetic variation among the Orang Asli. *Am. Anthropol. (n.s.)* 97:313–323.

Fix, A. G. 2011. Origin of genetic diversity among Malaysian Orang Asli: An alternative to the demic diffusion model. In *Dynamics of Human Diversity,* N. J. Enfield, ed. Canberra: Pacific Linguistics, 277–294.

Gray, R. D., Q. D. Atkinson, and S. J. Greenhill. 2011. Language evolution and human history: What a difference a date makes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366:1,090–1,100.

Hill, C., P. Soares, M. Mormina et al. 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol. Biol. Evol.* 23:2,480–2,491.

Hill, C., P. Soares, M. Mormina et al. 2007. A mitochondrial stratigraphy for Island Southeast Asia. *Am. J. Hum. Genet.* 80:29–43.

Jinam, T. A., M. E. Phipps, N. Saitou, and The Hugo Pan-Asian SNP Consortium. 2013. Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data. *Hum. Biol.* 85:173–188.

Lemey, P., A. Rambaut, J. J. Welch, and M. A. Suchard. 2010. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* 27:1,877–1,885.

Lye Tuck-Po. 2013. Making friends in the rainforest: "Negrito" adaptation to risk and uncertainty. *Hum. Biol.* 85:417–444.

Rambo, A. T. 1988. Why are the Semang? Ecology and ethnogenesis of aboriginal groups in Peninsular Malaysia. In *Ethnic Diversity and the Control of Natural Resources in Southeast Asia*, A. T. Rambo, K. Gillogly, and K. L. Hutterer, eds. Ann Arbor: University of Michigan Press, 19–35.

Sasse, H.-J. 1992. Theory of language death. In *Language Death: Factual and Theoretical Explorations with Special Reference to East Africa*, M. Brenzinger, ed. Berlin: Mouton de Gruyter, 7–30.

Walker, R. S., and L. A. Ribeiro. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proc. Biol. Sci.* 278:2,562–2,567.