PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2009) **276**, 2703–2710
doi:10.1098/rspb.2009.0408
Published online 29 April 2009

Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East

Andrew Kitchen[1,*], Christopher Ehret[2], Shiferaw Assefa[2] and Connie J. Mulligan[1]

[1]*Department of Anthropology, PO Box 103610, University of Florida, Gainesville, FL 32610-3610, USA*
[2]*Department of History, PO Box 951473, University of California—Los Angeles, Los Angeles, CA 90095-1473, USA*

The evolution of languages provides a unique opportunity to study human population history. The origin of Semitic and the nature of dispersals by Semitic-speaking populations are of great importance to our understanding of the ancient history of the Middle East and Horn of Africa. Semitic populations are associated with the oldest written languages and urban civilizations in the region, which gave rise to some of the world's first major religious and literary traditions. In this study, we employ Bayesian computational phylogenetic techniques recently developed in evolutionary biology to analyse Semitic lexical data by modelling language evolution and explicitly testing alternative hypotheses of Semitic history. We implement a relaxed linguistic clock to date language divergences and use epigraphic evidence for the sampling dates of extinct Semitic languages to calibrate the rate of language evolution. Our statistical tests of alternative Semitic histories support an initial divergence of Akkadian from ancestral Semitic over competing hypotheses (e.g. an African origin of Semitic). We estimate an Early Bronze Age origin for Semitic approximately 5750 years ago in the Levant, and further propose that contemporary Ethiosemitic languages of Africa reflect a single introduction of early Ethiosemitic from southern Arabia approximately 2800 years ago.

Keywords: Semitic; language evolution; Middle East; Horn of Africa; Bayesian phylogenetics; population history

1. INTRODUCTION

Semitic languages comprise one of the most studied language families in the world. Semitic is of particular interest due to its association with the earliest civilizations in Mesopotamia (Lloyd 1984), the Levant (Rendsburg 2003) and the Horn of Africa (Connah 2001), which gave rise to several of the world's first major religious traditions (Judaism, Christianity and Islam) and literary works (e.g. the Akkadian poem *The epic of Gilgamesh*). The importance of Semitic dates back at least 4350 years before present (YBP) to ancient Sumer in Mesopotamia, where the Akkadian language replaced Sumerian (Buccellati 1997). From this time forward, archaeological evidence for Semitic among the Hebrews and Phoenicians in the Levant (Diakonoff 1998; Rendsburg 2003) and the Aksumites in the Horn of Africa (Connah 2001) suggests that Semitic-speaking populations and their languages underwent a complex history of geographical expansion, migration and diffusion tied to the emergence of the earliest urban civilizations in these regions (Lloyd 1984; Connah 2001; Richard 2003*b*; Nardo 2007). Uncertainties about key details of this history persist despite extensive archaeological, genetic and linguistic studies of Semitic populations. A more comprehensive understanding of the precise origin and relationship of Semitic populations to each other is necessary to fully appreciate their complex history.

Although multiple genetic studies of extant Semitic-speaking populations have been conducted (Nebel *et al.* 2002; Capelli *et al.* 2006), much is still unknown about the genealogical relationships of these populations. Most previous genetic studies focus on time frames that are either too recent (the origin of Jewish communities in the Middle East and Africa; Hammer *et al.* 2000; Nebel *et al.* 2001; Rosenberg *et al.* 2001) or too ancient (the out-of-Africa migration of modern humans; Passarino *et al.* 1998; Quintana-Murci *et al.* 1999) to provide insight about the origin and dispersal of Semitic languages and Semitic-speaking populations.

Previous historical linguistic studies of Semitic languages have used the comparative method to infer the genealogical relationships of Semitic (for review, see Faber 1997). The comparative method is a technique that uses the pattern of shared, derived changes in language (vocabulary, syntax or grammar), termed innovations, to assess the relative relatedness of languages, although this method cannot date the divergences between languages (Campbell 2000). Cognates, which are words that generally share a common form and meaning through descent from a common ancestor (e.g. the English word 'night' is a cognate with the German word 'Nacht'), serve as the data used most often in comparative analyses.

*Author and present address for correspondence: Department of Biology, Pennsylvania State University, 208 Mueller Laboratory, University Park, PA 16802-5301, USA (aak11@psu.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.0408 or via http://rspb.royalsocietypublishing.org.

Received 9 March 2009
Accepted 8 April 2009

The field of Semitic linguistics has generally coalesced around a model that places the ancient Mesopotamian language Akkadian as the most basal lineage of Semitic (Hetzron 1976; Faber 1997). This standard model divides Semitic into East Semitic, composed of the extinct Akkadian and Eblaite languages, and West Semitic, consisting of all remaining Semitic languages that are distributed from the Levant to the Horn of Africa. West Semitic is in turn divided into South (consisting of Ethiosemitic, Epigraphic South Arabian and Modern South Arabian (MSA)) and Central linguistic groups, but the genealogical relationships of the languages within these two groups are poorly defined (Huehnergard 1990, 1992; Rodgers 1992; Faber 1997). Additionally, no consensus exists for placing Arabic in either the Central or South Semitic group (Hetzron 1976; Blau 1978; Diem 1980; Huehnergard 1990, 1992; Faber 1997), which makes Arabic's genealogical location simultaneously uncertain and interesting, as Central and South Semitic are geographically and genealogically distinct entities.

Dating language divergences has been controversial, especially when linguistic clocks are involved (for discussion, see Renfrew *et al.* 2000). The existence of a linguistic clock is controversial as it assumes that languages evolve at a fixed rate (Ehret 2000), whereas there is evidence for variation in rates of change between words and languages and no reason why languages should evolve at fixed rates (Blust 2000). However, recent studies have shown that much variation in the rates of linguistic change may follow generalized rules that apply across language families (Pagel *et al.* 2007; Atkinson *et al.* 2008). This suggests that variation in the rates of change between words and languages can be modelled by applying techniques used in evolutionary biology (e.g. probabilistic modelling of relative rates of word change with relaxed clock or covarion models of language evolution). Computational phylogenetic methods such as these are consistent with the philosophical underpinnings of the linguistic comparative method (i.e. inferring relationships by the comparison of similar features between languages) and provide an objective statistical framework to accurately estimate language divergences. Furthermore, Bayesian phylogenetic methods offer distinct advantages by allowing for the inclusion of multiple lines of evidence as prior probabilities, incorporating the uncertainty of model parameters in posterior probability estimates, and providing straightforward statistical comparisons of models via Bayes factors (BFs).

In this study, we analyse lexical data from 25 Semitic languages distributed throughout the Middle East and Horn of Africa (figure 1) using a Bayesian phylogenetic method to simultaneously infer genealogical relationships and estimate divergence dates of the Semitic languages investigated here. In order to calibrate a relaxed linguistic clock and increase the accuracy of our divergence date estimates, we use epigraphic data (text inscribed in stone or tablets) from extinct Semitic languages (Akkadian, Aramaic, Ge'ez, ancient Hebrew and Ugaritic) combined with archaeological evidence for the sampling dates of the epigraphic data (the time at which the materials were inscribed). We employ a log BF model-testing technique to statistically assess alternative Semitic histories and investigate different ways of modelling language evolution. Finally, we combine our divergence date estimates with epigraphic and archaeological evidence from all known Semitic languages to create an integrated model of Semitic history.

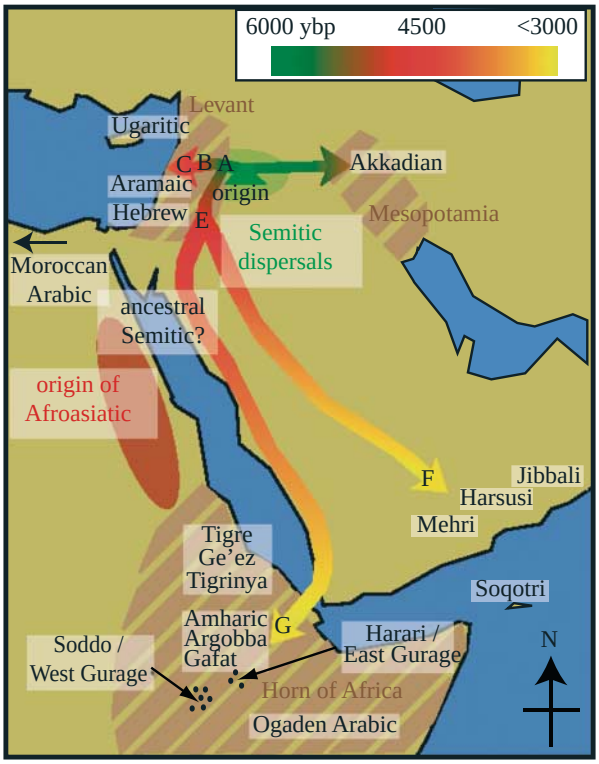

Figure 1. Map of Semitic languages and inferred dispersals. The locations of all languages sampled in this study, both extinct and extant, are depicted on the map. The current distribution of Ethiosemitic languages follows Bender (1971) and distribution of the remaining languages follows Hetzron (1997). The ancient distribution of extinct languages is also indicated (i.e. Akkadian, Biblical Aramaic, Ge'ez, ancient Hebrew and Ugaritic; Bender 1971; Hetzron 1997). The West Gurage (Chaha, Geto, Innemor, Mesmes and Mesqan) and East Gurage (Walani and Zway) Ethiosemitic language groups in central Ethiopia are depicted as two combined groups. The map also presents the dispersal of Semitic languages inferred from our study. An origin of Afroasiatic along the African coast of the Red Sea, supported by comparative analyses (Ehret 1995; Ehret *et al.* 2004), is indicated in red, although other African origins of Afroasiatic have been proposed (e.g. southwest Ethiopia; Blench 2006). The assumed location of the divergence of ancestral Semitic from Afroasiatic between the African coast of the Red Sea and the Near East is indicated in italics. Semitic dispersals are depicted by arrows coloured according to the estimated time of divergence (see coloured time scale at top of figure), and important nodes from the phylogeny (figure 2) are placed on the arrows to indicate where and when these divergences occurred.

2. MATERIAL AND METHODS

(a) *Wordlists and cognate coding*

Wordlists were modified from Swadesh's 100-word list of most conserved words (Swadesh 1955), with the final lists containing 96 words for 25 extant and extinct Semitic languages (fig. S1 in the electronic supplementary material). Wordlists for the Ethiosemitic languages (Amharic, Argobba, Chaha, Gafat, Ge'ez, Geto, Harari, Innemor, Mesmes, Mesqan, Soddo, Tigre, Tigrinya, Walani and Zway) and Ogaden Arabic were drawn from Bender (1971). Wordlists for Moroccan Arabic,

South Arabian languages (Jibbali, Harsusi, Mehri and Soqotri) and extinct non-African Semitic languages (Akkadian, Biblical Aramaic, ancient Hebrew and Ugaritic) were constructed from previously published lexicons (Leslau 1938; Gelb *et al.* 1956; Sobelman & Harrel 1963; Rabin 1975).

Cognate classes were determined for each of the 96 words using a comparative method that emphasizes the similarity of consonant–consonant–consonant roots and known consonant shifts when comparing two words. The cognate data were coded in two ways: (i) as a 25-by-96 multistate character matrix of cognate classes ('A'–'Q') for each of the 96 meanings (fig. S2 in the electronic supplementary material), and (ii) as a 25-by-673 binary matrix coding the presence ('1') or absence ('0') of each of the 673 cognate classes in each language (fig. S3 in the electronic supplementary material). Loanwords were identified using lexical information from distantly related, but geographically close, language families (such as Cushitic), as well as comparisons with lexicons of languages within the Semitic family. Identified loanwords were excluded from all subsequent analyses.

(b) *Phylogenetic analysis and divergence date estimation*

Phylogenies were constructed under a Bayesian framework using BEAST v. 1.4.8 (Drummond & Rambaut 2007). BEAST uses a Markov chain Monte Carlo (MCMC) simulation technique to estimate the posterior distribution of parameters. All Markov chains were run for 20 000 000 generations with samples taken every 1000 generations. The first 4 000 000 generations were discarded as burn-in, and post-run analysis of parameter plots in Tracer v. 1.4 (Rambaut & Drummond 2007) suggested all chains had reached convergence by the end of the burn-in period. MCMC sampling and run conditions, and all prior distributions, were identical for all analyses unless otherwise stated.

An unordered model of cognate class evolution with equal and reversible instantaneous rates of changes between all pairs of cognate classes (i.e. the rates of A-to-B, A-to-C and B-to-A changes were identical) was used to analyse the multistate coded data, while a model with a single reversible rate was used to analyse the binary coded data. Rate heterogeneity across lexical items was modelled by a gamma distribution of item-specific rates. This model accommodated variations in the rate of change across lexical items, such that conserved items (a single cognate class for all languages) were assigned a slower rate than the mean, while highly variable items (few shared cognate classes between languages) were assigned a faster rate than the mean. Priors for the gamma shape parameter were uniform on the interval 0–50.

Divergence times were estimated using an uncorrelated lognormal relaxed-clock model that assumes a single underlying rate for the entire phylogeny, but allows for variations in rates between branches (Drummond *et al.* 2006). In order to calibrate the clock, we used sampling dates for the five extinct languages in our dataset (Akkadian = 2800 YBP, Biblical Aramaic = 1800 YBP, Ge'ez = 1700 YBP, ancient Hebrew = 2600 YBP and Ugaritic = 3400 YBP; Rabin 1975) in a manner similar to how sampling dates are used in the studies of measurably evolving populations, such as fast-evolving viruses or ancient DNA (Drummond *et al.* 2003). These dates come from archaeological and epigraphic evidence associated with the linguistic source material, and thus provide the time at which the wordlists of the extinct languages were sampled (although the languages themselves often continued to exist for some time). Additionally, a set of five constraints taken from a combination of archaeological, epigraphic and historical evidence was placed on interior nodes. Such constraints allow for the inclusion of prior information and uncertainty regarding Semitic divergence times, which are strengths of Bayesian methods and have been successfully used to date the divergences of Indo-European (Gray & Atkinson 2003; Atkinson *et al.* 2005) and Austronesian (Gray *et al.* 2009) languages. These constraints are: (i) the origin of ancient Hebrew 3200–4200 YBP (Steiner 1997), (ii) the origin of Ugaritic 3400–4400 YBP (Pardee 1997), (iii) the origin of Aramaic 2850–3850 YBP (Kaufman 1997) and (iv) the origin of Amharic 700–1700 YBP (Hudson 1997). Each of these constraints spans a 1000-year interval since the earliest epigraphic or historical evidence for the language. An additional constraint (v) was placed on the time of the most recent common ancestor of the included Semitic languages to 4350–8000 YBP (the lower date is based on the earliest known epigraphic evidence of Akkadian; Buccellati 1997). An analysis was also performed without the constraint on the age of the root, which returned an estimate of 4300–7750 YBP for the root, i.e. almost exactly our constraint range. All divergence time constraints are in the form of uniform priors over the indicated interval. A uniform prior of 0.01 to 0.00001 cognate changes per word per year (0.001–1% replacement rate per year) was placed on the mean of the lognormal-distributed clock. The mean rate estimated from analysis of the binary data is 6.1×10^{-5} replacements per cognate per year (95% highest probability density (HPD) $=4.4-7.9\times10^{-5}$).

The robustness of our results was investigated using log BF tests to compare phylogenies that were constrained to model alternative Semitic histories. Specifically, we first compared two versions of the standard model of Semitic history: a model that placed Akkadian (i.e. East Semitic) at the root versus an unconstrained analysis to assess independent support for a non-African Semitic root. We then investigated the position of Arabic in Semitic history by comparing two variations of the standard model, one with Arabic nested within Central Semitic and another with Arabic within South Semitic. We also tested the ability of different models to account for variation in rates of linguistic change between lexical items and languages. In this case, we compared the standard model to two alternatives: (i) no gamma distribution to model variation in the rate of change between lexical items and (ii) no relaxed clock to model variation in the rate of change between languages. All log BF tests of Semitic history incorporated a gamma distribution and relaxed clock since our log BF tests showed support for these models. Marginal likelihoods for each model were estimated using the smoothed harmonic mean of the likelihood distribution (Newton *et al.* 1994; Redelings & Suchard 2005), and all log BF values were calculated by taking the difference in the log of the marginal likelihoods of each model (Kass & Raftery 1995) with log BF values reported in log units.

3. RESULTS

(a) *Genealogy of Semitic languages*

Our phylogenetic analysis of Semitic languages produced the phylogeny shown in figure 2. This phylogeny is based on the binary dataset and incorporates all model features

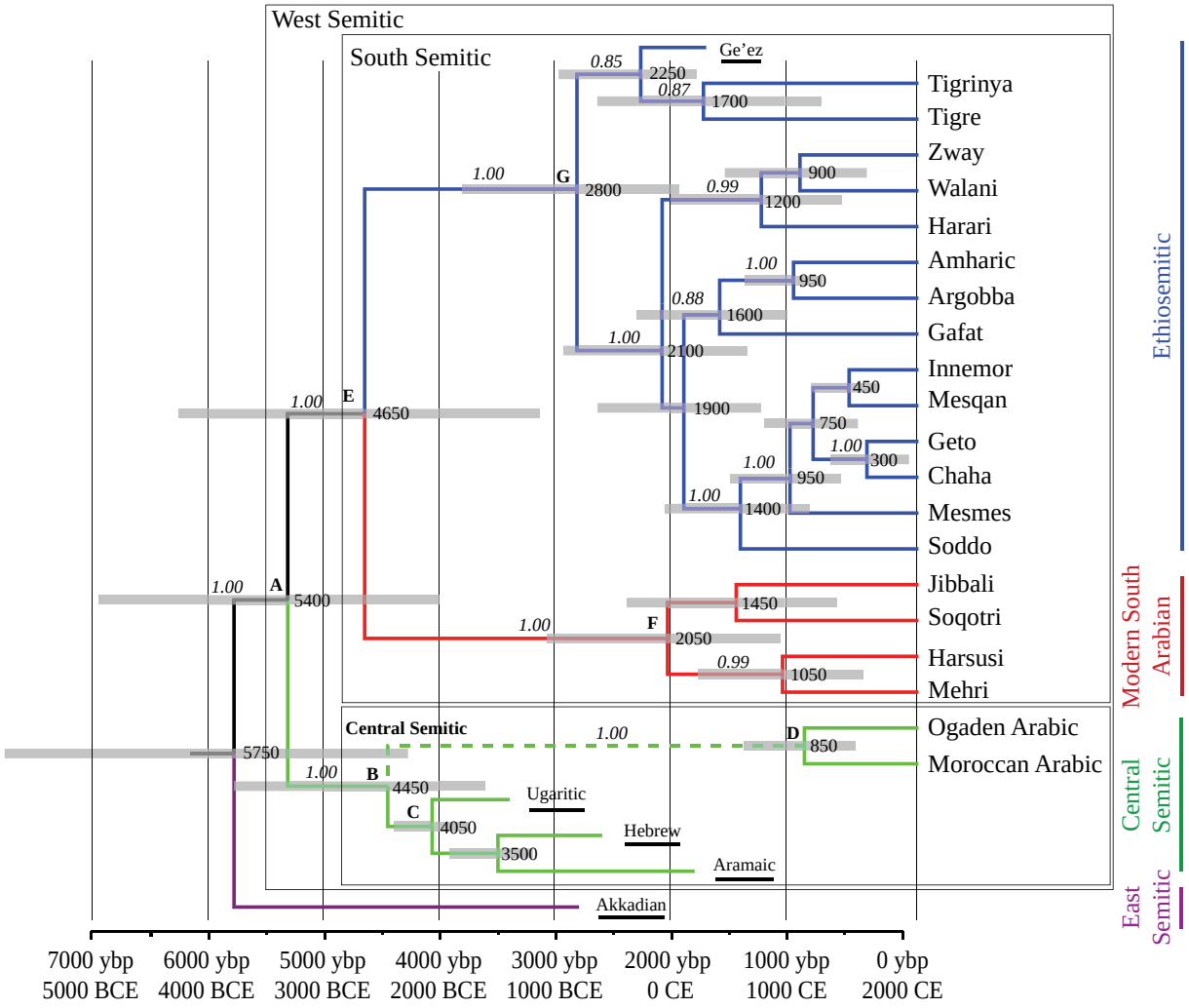

Figure 2. Phylogeny of Semitic languages. Our phylogeny of 25 Semitic languages based on binary encoded data is presented with mean divergence times to the right of each node and 95% HPD intervals indicated by light grey bars. The scale bar along the bottom of the phylogeny presents time in YBP. Posterior probabilities of branches are printed in italics above each branch with >0.75 support. Extinct languages are underlined and all other languages are considered to evolve to the present. Subgroups of Semitic are identified by colour bars to the right of the phylogeny (purple bars, East Semitic; green bars, Central Semitic; red bars, MSA; and blue bars, Ethiosemitic) and by three boxes (West, Central and South Semitic). Important nodes are indicated by letters: A, West Semitic; B, Central Semitic; C, Ugaritic–Hebrew–Aramaic; D, Arabic; E, South Semitic; F, MSA; and G, Ethiosemitic. The dashed line leading to Arabic reflects the fact that log BF tests were equivocal in the placement of Arabic, so we placed Arabic in Central Semitic based on previous linguistic studies (e.g. Hetzron 1976; Faber 1997). The topology is rooted with Akkadian, which is preferred by our log BF analyses, and follows the constraints of the standard model.

that showed significant log BF support (log BF tests were equivocal in the placement of Arabic, so we placed Arabic in Central Semitic based on previous comparative studies; e.g. Hetzron 1976; Faber 1997). A brief summary of the phylogeny highlights include: (i) the greater age of non-African versus African Semitic languages (non-overlapping HPDs of 4150–7400 YBP for East/West Semitic versus 2000–3800 YBP for Ethiosemitic); (ii) the near-simultaneous divergence of East, West, South and Central Semitic languages; (iii) the early divergence of Arabic of approximately 4450 YBP (HPD: 3650–5800 YBP); and (iv) the well-resolved and recent divergences (less than 3800 YBP) of Ethiosemitic languages in a monophyletic (single origin) clade (a group of related languages). It is important to note that each node in the phylogeny represents an ancestral language that is hypothesized to have existed at the time of divergence estimated for that node, whereas branch tips represent actual languages at the time they were sampled. Long branches are representative of long intervals between divergences and the presence of unsampled languages (e.g. the long branch between nodes E and F), whereas short branches indicate rapid language divergence. Posterior probability estimates are shown for each branch and indicate the probability that a group of languages is more closely related to each other than to other languages. Branches with posterior probability estimates ≤ 0.70 were considered to be unresolved (the relative pattern of divergence among the languages could not be ascertained) and were collapsed to reflect this uncertainty.

(b) *Semitic language divergence dates*

In addition to delineating the relationship between different Semitic languages, our phylogenetic analysis provides dates for the divergences of the investigated languages. The mean estimates of all language divergence

times, with associated 95 per cent HPDs, are depicted in years on the phylogeny in figure 2. Our phylogeny indicates the most basal divergence within Semitic occurred at 5750 YBP (HPD: 4400–7400 YBP), suggesting an origin of Semitic during the Early Bronze Age (Ehrich 1992). This result implies that a hypothetical ancestral language was extant during this period and gave rise to all of the Semitic languages investigated in this study. The deepest four branches of the phylogeny indicate the divergences of East (root), West (node A), South (node E) and Central (node B) Semitic; these divergences are nearly coincident with largely overlapping HPDs (3300–7400 YBP), suggesting that Semitic underwent a period of rapid diversification upon its origin.

Central Semitic (node B) initially diverges at approximately 4450 YBP (HPD: 3650–5800 YBP) into Arabic and a group of ancient languages from the Levant (Aramaic, ancient Hebrew and Ugaritic), which in turn diverge (node C) at approximately 4050 YBP (HPD: 3750–4400 YBP). The Arabic languages (node D) have an estimated divergence time of approximately 850 YBP (HPD: 400–1370 YBP).

On the other half of the phylogeny, the South Semitic clade (node E) shows an ancient divergence of Ethiosemitic and MSA languages approximately 4650 YBP (HPD: 3300–6250 YBP), which overlaps with the transition from the Early to Middle Bronze Age. The early divergence between Ethiosemitic and MSA is consistent with previous historical linguistic proposals that MSA is a deep branch of Semitic, linguistically distant even from its closest relatives within the Semitic family (e.g. Murtonen 1967). The hypothetical ancestor of the MSA clade (node F) dates to approximately 2050 YBP (HPD: 1100–3100 YBP), which, coupled with the narrow geographical distribution of MSA along the southern coast of Arabia, suggests that the diversification of MSA occurred in this region.

The single, well-supported (posterior probability = 0.9976) branch leading to modern Ethiosemitic indicates a single origin for Semitic languages in the Horn of Africa with their diversification into North and South clades (node G) occurring at approximately 2850 YBP (HPD: 2000–3800 YBP), during the Iron Age in the Near East and overlapping with the pre-Aksumite and Aksumite periods in the Horn of Africa (Connah 2001). The large number of small internal branches in the Ethiosemitic group indicates a rapid diversification of these languages. The South Ethiosemitic languages separate into three monophyletic clades that correspond to accepted groupings of Ethiosemitic (Bender 1971) and show near-coincident divergences at approximately 1200–1600 YBP.

Our analysis of the multistate-encoded data produced divergence date estimates and 95 per cent HPDs that were consistent with those estimated from the binary encoded data (see fig. S4 in the electronic supplementary material). The mean divergence dates are also altered: the divergences of East versus West Semitic, Central versus South Semitic, MSA versus Ethiosemitic and Ethiosemitic are older in the multistate estimates, and the divergences of Central Semitic and MSA are younger relative to the binary estimates. The topologies are essentially the same with several small changes within the Ethiosemitic languages and a closer clustering of Arabic and Aramaic in the multistate analysis. Importantly, all of the mean divergence date estimates from the binary analysis fall within the HPDs of the multistate analysis. For figure 2, we chose to present the phylogeny based on the binary dataset following conventions of previous linguistic phylogenetic studies (Gray & Atkinson 2003; Atkinson *et al.* 2005; Gray *et al.* 2009).

(c) *Log Bayes factor tests*

We assess the robustness of our analysis by statistically testing alternative Semitic histories. This was done using log BF model tests, which compare the probabilities that various models produced for the observed data (i.e. the lexical list data). Log BF values (all values are in log units) in the intervals 0–0.5, 0.5–1, 1–2 and greater than 2 are considered 'not worth mentioning', 'substantial', 'strong' and 'decisive' support, respectively, for the primary model (Kass & Raftery 1995). We test alternative Semitic histories using two comparisons. The first comparison tests models that root Semitic with Akkadian (i.e. a Near Eastern origin of Semitic) relative to an unconstrained model that allows for Near Eastern (i.e. Akkadian), African (i.e. Ethiosemitic) or Arabian (i.e. MSA) origins for Semitic. This comparison shows substantial support for a model with an Akkadian root ($\log BF = 0.641$), consistent with the consensus of comparative linguistic analyses (Faber 1997). The second comparison concerns the placement of Arabic and compares the standard model, in which Arabic is placed within Central Semitic (e.g. Hetzron 1976; Faber 1997), with a single topological modification that places Arabic within South Semitic. This comparison showed little preference for a model with Arabic within Central Semitic over one with Arabic within South Semitic ($\log BF = -0.438$). Interestingly, the location of Arabic within Semitic is the only discrepancy in topology and divergence date estimates between our binary and multistate analyses (figure 2; fig. S4 in the electronic supplementary material).

We also use log BFs to test the ability of different models to accurately represent variation in the rates of linguistic change between words and languages. Our first comparison was between versions of the standard model that did and did not include a gamma distribution to model variation in the rate of linguistic change between lexical items. This log BF test shows substantial ($\log BF = 0.574$) support for a model that includes a gamma distribution to model rate variation between words. To place this in perspective, our estimate for the shape of the gamma distribution ($\alpha = 24.9$) indicates that there is less variation in the rate of change between lexical items than there is within codon classes in mitochondrial coding genomes of primates (Yang 1996). Our second comparison was between versions of the standard model that used relaxed and strict linguistic clocks to model rate variation between languages. This log BF test shows decisive support ($\log BF = 13.0$) for a model that includes a relaxed clock to model between-language variation. These two results demonstrate that our inclusion of rate variation components in our model of linguistic evolution significantly improves the fit between the data and our model, and that there is substantial variation in the linguistic rate of change between lineages and between lexical items. All log BF tests of Semitic history reported

above incorporated a gamma distribution and relaxed clock since our log BF tests showed support for these models.

4. DISCUSSION

(a) *Semitic origins*

Our analysis of the Semitic language family produced a dated phylogeny that estimates the origin of Semitic at approximately 4400–7400 YBP (figure 2). The phylogeny suggests East Semitic (represented by Akkadian in this study) corresponds to the deepest branch (although the four deepest branches have overlapping HPDs), and our log BF tests indicate that Akkadian is the appropriate root for the Semitic languages analysed here. These results indicate that the ancestor of all Semitic languages in our dataset was being spoken in the Near East no earlier than approximately 7400 YBP, after having diverged from Afroasiatic in Africa (Ehret 1995; Ehret *et al.* 2004; Blench 2006). Lacking closely related non-Semitic languages to serve as out-groups in our phylogeny, we cannot estimate when or where the ancestor of all Semitic languages diverged from Afroasiatic. Furthermore, it is likely that some early Semitic languages became extinct and left no record of their existence. This is especially probable if early Semitic societies were pastoralist in nature (Blench 2006), as pastoralists are less likely to leave epigraphic and archaeological evidence of their languages. The discovery of such early Semitic languages could increase estimates of the age of Semitic, and alter its geographical origin if these early Semitic languages were found in Africa rather than the Middle East.

Our estimate for the origin of Semitic (4400–7400 YBP) predates the first Akkadian inscriptions in the archaeological record of northern Mesopotamia by approximately 100–3000 years (Buccellati 1997). The city-states of Sumer were established and flourishing in Mesopotamia with their own indigenous languages unrelated to Semitic by approximately 5400 YBP (Lloyd 1984), so it is unlikely that Akkadian was spoken in Sumer for the entirety of the possible 3000-year interval between the origin of Semitic and Akkadian's initial appearance in the archaeological record. Furthermore, Eblaite (no Eblaite wordlists were available for our study), the closest relative of Akkadian and the only other member of East Semitic, was spoken in the Levant (specifically the northeast Levant or present-day Syria; Gordon 1997), which is also where some of the oldest West Semitic languages were spoken (Ugaritic, Aramaic and ancient Hebrew). The presence of ancient members of the two oldest Semitic groups (East and West Semitic) in the same region of the Levant, combined with a possible long interval (100–3000 years) between the origin of Semitic and the appearance of Akkadian in Sumer, suggests a Semitic origin in the northeast Levant and a later movement of Akkadian eastward into Mesopotamia and Sumer (see figure 1 for a map of our proposed Semitic dispersals).

(b) *Early Semitic dispersals*

Our Semitic language phylogeny indicates that the initial divergence of ancestral Semitic into East and West Semitic was nearly coincident with the divergence of West Semitic into Central and South Semitic around 5300 YBP (figure 2, node A). The short interval between these two divergences and their overlapping HPDs suggests that both divergences may have occurred in the northeast Levant (see figure 1, node A). The distribution of ancient and modern Central and South Semitic languages is consistent with Central Semitic spreading westward throughout the Levant and South Semitic spreading southward from the Levant, eventually reaching southern Arabia (figure 1, nodes B and E, respectively).

The Central Semitic branch is characterized first by a divergence into Arabic and the Levantine languages (Aramaic, Hebrew and Ugaritic) at least 3650 YBP and possibly shortly after East and West Semitic diverged (figures 1 and 2, node B). The Levantine languages subsequently diverged into separate lineages by approximately 4050 YBP (figures 1 and 2, node C), but possibly as early as approximately 4400 YBP. The expansion of the Levantine languages of Central Semitic approximately 3650–4400 YBP was probably part of the migration process that was definitive of the transition from the Early to the Middle Bronze Age in the Levant (Ehrich 1992; Ilan 2003; Richard 2003*a*). This period in the Levant involved the devolution of many urban societies at the end of the Early Bronze Age (Richard 2003*a*) and their replacement with new urban societies that were culturally and morphologically distinct at the start of the Middle Bronze Age (Ilan 2003). Our analysis suggests that the shift in urban populations during the Early to Middle Bronze Age may be temporally associated with the wider expansion of Central Semitic in the Levant.

Within South Semitic, the early emergence of a South Arabian lineage between approximately 3300 and 6250 YBP (figures 1 and 2, node E) may reflect an Early Bronze Age expansion of Semitic from the Levant southward to the Arabian desert. This lineage was ancestral to the MSA languages, for which the more recent divergence less than 3100 YBP (figures 1 and 2, node F) suggests that early MSA speakers probably inhabited the southern coasts and coastal hinterlands of the Arabian Peninsula (the current distribution of MSA). The recurrent spread of early Semitic peoples and their languages into the steppe and desert lands of the Arabian Peninsula (first South Semitic and later Arabic; see below), combined with Biblical testimony on early Hebrew subsistence, lead us to propose that the earliest West Semitic society may have had a largely pastoralist economy particularly adapted to such conditions.

(c) *Recent Arabic divergence*

The Arabic languages, or dialects, represent the largest group of extant Semitic languages (Gordon 2005). Although our analysis provided inconsistent support for Arabic as a lineage of Central Semitic (i.e. strong support for Arabic within Central Semitic from the multistate analysis, but no support from the binary analysis), most comparative linguistic analyses place Arabic within Central Semitic (for a review, see Faber 1997). Arabic languages originated in northern Arabia and expanded along with Islam in the seventh century to occupy a geographical range that extends from Morocco to Iran in the present day (Kaye & Rosenhouse 1997). Our phylogenetic analysis indicated that the two studied Arabic languages (Moroccan and Ogaden) diverged approximately 400–1350 YBP (node D); that is, after the expansion of Arab populations associated with Islam.

This late divergence suggests that Arabic-speaking populations maintained sufficient contact to preclude the divergence and isolation of their languages for some time, or that, in some regions such as Morocco, it was not until the last millennium that Arabic languages replaced earlier indigenous languages (Berber in this case).

(d) *Origin of Ethiosemitic*

Our Semitic phylogeny indicates that Ethiosemitic had a single, non-African origin; Ethiosemitic forms a well-resolved monophyletic clade nested within non-African Semitic languages, no earlier than approximately 3800 YBP (node G). The simultaneous divergences of many Ethiosemitic subgroups and their current widespread distribution throughout Ethiopia suggest that Ethiosemitic underwent a rapid process of diversification and expansion upon arrival in Africa. Studies have shown that Ethiosemitic-speaking populations are genetically similar to Cushitic-speaking populations within Eritrea and Ethiopia (Lovell *et al.* 2005). Thus, we propose that the current distribution of Ethiosemitic reflects a process of language diffusion through existing African populations with little gene flow from the Arabian Peninsula (i.e. a language shift). Our mean estimate of approximately 2850 YBP for the origin of Ethiosemitic (node G) is contemporaneous with the rise of pre-Aksumite societies in Eritrea and Ethiopia (Connah 2001), although the associated HPD includes the early Aksumite period. This result suggests that the introduction of early Ethiosemitic languages to the Horn of Africa may have been temporally associated with the development of some of the first indigenous complex societies (Ehret 1988), Aksumite or pre-Aksumite, and coincided with a period of South Arabian influence in northern Ethiopia approximately 2400–2700 YBP (Michels 2005).

5. CONCLUSION

We used Bayesian phylogenetic methods to elucidate the relationships and divergence dates of Semitic languages, which we then related to epigraphic and archaeological records to produce a comprehensive hypothesis of Semitic origins and dispersals after the divergence of ancestral Semitic from Afroasiatic in Africa (figure 1). We estimate that: (i) Semitic had an Early Bronze Age origin (approx. 5750 YBP) in the Levant, followed by an expansion of Akkadian into Mesopotamia; (ii) Central and South Semitic diverged earlier than previously thought throughout the Levant during the Early to Middle Bronze Age transition; and (iii) Ethiosemitic arose as the result of a single, possibly pre-Aksumite, introduction of a lineage from southern Arabia to the Horn of Africa approximately 2800 YBP. Furthermore, we employed the first use of log BFs to statistically test competing language histories and provide support for a Near Eastern origin of Semitic. Our inferences shed light on the complex history of Semitic, address key questions about Semitic origins and dispersals, and provide important hypotheses to test with new data and analyses.

We thank Steve Brandt and Johanna Nichols for their enlightening discussions and comments on the manuscript. This research was funded by a grant from the National Science Foundation to C.J.M. (BSR-0518530).

REFERENCES

Atkinson, Q. D., Nicholls, G., Welch, D. & Gray, R. 2005 From words to dates: water into wine, mathemagic or phylogenetic inference? *Trans. Philol. Soc.* **103**, 193–219. (doi:10.1111/j.1467-968X.2005.00151.x)

Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J. & Pagel, M. 2008 Languages evolve in punctuational bursts. *Science* **319**, 588. (doi:10.1126/science.1149683)

Bender, M. L. 1971 Languages of Ethiopia—new lexicostatistic classification and some problems of diffusion. *Anthropol. Linguist.* **13**, 165–288.

Blau, J. 1978 Hebrew and Northwest Semitic: reflections on the classification of the Semitic languages. *Heb. Annu. Rev.* **2**, 21–44.

Blench, R. 2006 *Archaeology, language, and the African past*. Lanham, MD: AltaMira.

Blust, R. 2000 Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics*, vol. 2 (eds C. Renfrew, A. McMahon & L. Trask), pp. 311–331. Cambridge, UK: The McDonald Institute for Archaeological Research.

Buccellati, G. 1997 Akkadian. In *The Semitic languages* (ed. R. Hetzron), pp. 69–99. London, UK: Routledge.

Campbell, L. 2000 Time perspectives in linguistics. In *Time depth in historical linguistics*, vol. 1 (eds C. Renfrew, A. McMahon & L. Trask), pp. 3–31. Cambridge, UK: The McDonald Institute for Archaeological Research.

Capelli, C. *et al.* 2006 Population structure in the Mediterranean basin: a Y chromosome perspective. *Ann. Hum. Genet.* **70**, 207–225. (doi:10.1111/j.1529-8817.2005.00224.x)

Connah, G. 2001 *African civilizations: an archaeological perspective*. Cambridge, UK: Cambridge University Press.

Diakonoff, I. 1998 The earliest Semitic society. *J. Semit. Stud.* **43**, 209–219. (doi:10.1093/jss/43.2.209)

Diem, W. 1980 Die genealogische Stellung des Arabischen in den semitischen Sprachen: ein eingelostes Problem der Semitistik. In *Studien aus Arabistik und Semitistik, A. Spitaler zum 70* (eds W. Diem & S. Wild), pp. 65–85. Wiesbaden, Germany: Harrassowitz.

Drummond, A. J. & Rambaut, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. 2003 Measurably evolving populations. *Trends Ecol. Evol.* **18**, 481–488. (doi:10.1016/S0169-5347(03)00216-7)

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, 699–710. (doi:10.1371/journal.pbio.0040088)

Ehret, C. 1988 Social transformations in the early history of the Horn of Africa: linguistic clues to developments of the period 500 BC to AD 500. In *Proc. Eighth Int. Conf. of Ethiopian Studies*, vol. 1 (ed. T. Bayene) pp. 639–651. Addis Ababa, Ethiopia: Institute of Ethiopian Studies.

Ehret, C. 1995 *Reconstructing proto-Afroasiatic (proto-Afasian): vowels, tone, consonants, and vocabulary*. Berkeley, CA: University of California Press.

Ehret, C. 2000 Testing the expectations of glottochronology against the correlations of language and archaeology in Africa. In *Time depth in historical linguistics*, vol. 2 (eds C. Renfrew, A. McMahon & L. Trask), pp. 373–399. Cambridge, UK: The McDonald Institute for Archaeological Research.

Ehret, C., Keita, S. O. Y. & Newman, P. 2004 The origins of Afroasiatic. *Science* **306**, 1680. (doi:10.1126/science.306.5702.1680c)

Ehrich, R. H. (ed.) 1992 *Chronologies in Old World archaeology*. 3rd edition. Chicago, IL: University of Chicago Press.

Faber, A. 1997 Genetic subgrouping of the Semitic languages. In *The Semitic languages* (ed. R. Hetzron), pp. 3–15. London, UK: Routledge.

Gelb, I. J., Jacobsen, T., Landsberger, B. & Oppenheim, A. L. (eds) 1956 *The Assyrian dictionary of the Oriental Institute of the University of Chicago.* Chicago, IL: Oriental Institute.

Gordon, C. H. 1997 Amorite and Eblaite. In *The Semitic languages* (ed. R. Hetzron), pp. 100–113. London, UK: Routledge.

Gordon, R. G. (ed.) 2005 *Ethnologue: languages of the world.* Dallas, TX: SIL International.

Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)

Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)

Hammer, M. F. *et al.* 2000 Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl Acad. Sci. USA* **97**, 6769–6774. (doi:10.1073/pnas.100115997)

Hetzron, R. 1976 Two principles of genetic reconstruction. *Lingua* **38**, 89–104. (doi:10.1016/0024-3841(76)90074-7)

Hetzron, R. (ed.) 1997 *The Semitic languages.* London, UK: Routledge.

Hudson, G. 1997 Amharic and Argobba. In *The Semitic languages* (ed. R. Hetzron), pp. 457–485. London, UK: Routledge.

Huehnergard, J. 1990 Remarks on the classification of the Northwest Semitic languages. In *Deir 'Alla Symposium* (ed. J. Hoftizjer), pp. 282–293. Leiden, The Netherlands: Brill.

Huehnergard, J. 1992 Languages of the ancient Near East. In *The anchor Bible dictionary*, vol. 4, pp. 155–170. New York, NY: Doubleday.

Ilan, D. 2003 The Middle Bronze Age (circa 2000—1500 B.C.E.). In *Near eastern archaeology: a reader* (ed. S. Richard), pp. 331–342. Winona Lakes, IN: Eisenbrauns.

Kass, R. E. & Raftery, A. E. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.2307/2291091)

Kaufman, S. A. 1997 Aramaic. In *The Semitic languages* (ed. R. Hetzron), pp. 114–130. London, UK: Routledge.

Kaye, A. S. & Rosenhouse, J. 1997 Arabic dialects and Maltese. In *The Semitic languages* (ed. R. Hetzron), pp. 263–311. London, UK: Routledge.

Leslau, W. 1938 *Soqotri avec comparaisons et explications etymologique*. Paris, France: C. Klincksieck.

Lloyd, S. 1984 *The archaeology of Mesopotamia*. The World of Archaeology. New York, NY: Thames and Hudson.

Lovell, A., Moreau, C., Yotova, V., Xiao, F., Bourgeois, S., Gehl, D., Bertranpetit, J., Schurr, E. & Labuda, D. 2005 Ethiopia: between Sub-Saharan Africa and Western Eurasia. *Ann. Hum. Genet.* **69**, 275–287. (doi:10.1046/J.1469-1809.2005.00152.x)

Michels, J. W. 2005 *Changing settlement patterns in Aksum-Yeha region of Ethiopia: 700 BC–AD 850*. Oxford, UK: Archaeopress.

Murtonen, A. 1967 *Early Semitic: a diachronical inquiry into the relationship of Ethiopic to the other so-called South-East Semitic languages*. Leiden, The Netherlands: E. J. Brill.

Nardo, D. 2007 *Ancient Mesopotamia*. Detroit, MI: Greenhaven Press.

Nebel, A., Filon, D., Brinkmann, B., Majumder, P. P., Faerman, M. & Oppenheim, A. 2001 The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am. J. Hum. Genet.* **69**, 1095–1112. (doi:10.1086/324070)

Nebel, A., Landau-Tasseron, E., Filon, D., Oppenheim, A. & Faerman, M. 2002 Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. *Am. J. Hum. Genet.* **70**, 1594–1596. (doi:10.1086/340669)

Newton, M. A. *et al.* 1994 Approximate Bayesian-inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. B: Methodol.* **56**, 3–48.

Pagel, M., Atkinson, Q. D. & Meade, A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)

Pardee, D. 1997 Ugaritic. In *The Semitic languages* (ed. R. Hetzron), pp. 131–144. London, UK: Routledge.

Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M. & Santachiara-Benerecetti, A. S. 1998 Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am. J. Hum. Genet.* **62**, 420–434. (doi:10.1086/301702)

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. 1999 Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat. Genet.* **23**, 437–441. (doi:10.1038/70550)

Rabin, C. 1975 Lexicostatistics and the internal divisions of Semitic. In *Hamito-Semitica* (eds T. Bynon & J. Bynon), pp. 85–102. The Hague, The Netherlands: Mouton.

Rambaut, A. & Drummond, A. 2007 *Tracer* v. 1.4. See http://tree.bio.ed.ac.uk/software/Trace.

Redelings, B. D. & Suchard, M. A. 2005 Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* **54**, 401–418. (doi:10.1080/10635150590947041)

Rendsburg, G. A. 2003 Semitic languages (with special reference to the Levant). In *Near eastern archaeology: a reader* (ed. S. Richard), pp. 71–73. Winona Lakes, IN: Eisenbrauns.

Renfrew, C., McMahon, A. & Trask, L. (eds) 2000 *Time depth in historical linguistics*. Cambridge, UK: The McDonald Institute for Archaeological Research.

Richard, S. 2003*a* The Early Bronze Age in the southern Levant. In *Near eastern archaeology: a reader* (ed. S. Richard), pp. 280–296. Winona Lake, IN: Eisenbrauns.

Richard, S. (ed.) 2003*b* *Near eastern archaeology: a reader.* Winona Lake, IN: Eisenbrauns.

Rodgers, J. 1992 The subgrouping of the South Semitic languages. In *Semitic studies in honor of Wolf Leslau* (ed. A. S. Kaye), pp. 1323–1336. Wiesbaden, Germany: Harrassowitz.

Rosenberg, N. A. *et al.* 2001 Distinctive genetic signatures in the Libyan Jews. *Proc. Natl Acad. Sci. USA* **98**, 858–863. (doi:10.1073/pnas.98.3.858)

Sobelman, H. & Harrel, R. 1963 *A dictionary of Moroccan Arabic: English–Arabic*. Washington, DC: Georgetown University Press.

Steiner, R. C. 1997 Ancient Hebrew. In *The Semitic languages* (ed. R. Hetzron), pp. 145–173. London, UK: Routledge.

Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)

Yang, Z. 1996 Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372. (doi:10.1016/0169-5347(96)10041-0)