nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0489-3

The origin and expansion of Pama–Nyungan languages across Australia

Remco R. Bouckaert[1,2], Claire Bowern[3] and Quentin D. Atkinson[2,4*]

It remains a mystery how Pama–Nyungan, the world's largest hunter-gatherer language family, came to dominate the Australian continent. Some argue that social or technological advantages allowed rapid language replacement from the Gulf Plains region during the mid-Holocene. Others have proposed expansions from refugia linked to climatic changes after the last ice age or, more controversially, during the initial colonization of Australia. Here, we combine basic vocabulary data from 306 Pama-Nyungan languages with Bayesian phylogeographic methods to explicitly model the expansion of the family across Australia and test between these origin scenarios. We find strong and robust support for a Pama–Nyungan origin in the Gulf Plains region during the mid-Holocene, implying rapid replacement of non-Pama–Nyungan languages. Concomitant changes in the archaeological record, together with a lack of strong genetic evidence for Holocene population expansion, suggests that Pama–Nyungan languages were carried as part of an expanding package of cultural innovations that probably facilitated the absorption and assimilation of existing hunter-gatherer groups.

Most of human prehistory has played out among hunter-gatherer societies. However, the expansion of agriculture in the past 10,000 years has replaced much of the world's cultural and linguistic diversity[1] and, in so doing, erased evidence of past hunter-gatherer expansions. In Australia, the cultural legacy of one large-scale hunter-gatherer expansion remains uniquely well preserved in the linguistic diversity of the continent. Of the 28 language families of mainland Australia, 27 are restricted to the far north, while one family—Pama–Nyungan—covers the remaining 90% of the continent[2]. It is now well established that Aboriginal Australians have inhabited Australia for more than 50,000 years[1,3,4], but how and why one family came to occupy most of the Australian continent remains a mystery[5].

Proposals for the origin of Pama–Nyungan include time depths ranging from 4 thousand years ago (ka) to more than 40 ka, putative homelands that span the length and breadth of the continent, and a range of mechanisms of expansion[2,6] (Supplementary Table 1). One hypothesis argues for a recent, rapid replacement of non-Pama–Nyungan by Pama–Nyungan languages from an origin around the Gulf of Carpentaria and north-western Queensland 4–6 ka[5,7–9]. Proponents argue that this expansion into already occupied territory was driven by the emergence of putative social and technological advantages during the mid- to late Holocene, including new lithic technologies[2,5,7] and social institutions[7,8], and was possibly associated with the introduction of the dingo[10]. Two further hypotheses hold that the languages spread earlier into uninhabited or sparsely populated areas from refugia, where relict populations were sustained throughout the Last Glacial Maximum[11–15]. One variant links the origin of Pama–Nyungan to evidence of population intensification following climatic amelioration in the early Holocene 7–10 ka[11,15], perhaps triggered by the Last Marine Transgression[14]. A second variant argues for an earlier expansion from a Dividing Range refugium at the end of the Antarctic Cold Reversal in the late Pleistocene 10–13 ka[12]. Finally, a fourth, controversial hypothesis proposes that Pama–Nyungan is in fact much older, reflecting divergence and convergence processes that date back to the initial colonization of the continent ~40–55 ka[16,17].

Current genetic and linguistic evidence for the origin of the family is inconclusive. Recent whole-genome analysis of Pama–Nyungan speakers[3] has revealed an intriguing correspondence between genetic and linguistic divergence[18] and finds evidence for an early northeast–southwest split 10–31 ka. However, a lack of samples from non-Pama–Nyungan languages in the north means this signal cannot be tied to Pama–Nyungan specifically. Simple distance-based metrics using the percentage of homologous words or 'cognates' shared between Pama–Nyungan languages imply an age for the family of ~8 ka[19], but this approach has been shown to produce unreliable estimates[20] and is now largely discredited. In contrast, linguistic arguments for a recent (4–6 ka) expansion of the family from the Gulf of Carpentaria have relied on more impressionistic assessments of the amount of diversity across the continent[9]. An expansion at this time is consistent with evidence from early genetic studies for gene flow into Australia from India ~4–5 ka[21], but more recent work has called these findings into question[3,22–24].

By explicitly modelling the process of language evolution in time and space and incorporating uncertainty in the resulting estimates, new Bayesian phylogeographic methods[25] now make it possible to investigate language family origins in a more principled and transparent way[26,27]. Previous work has successfully used Bayesian inference of phylogeny to quantify support for Pama–Nyungan's internal branching structure[18]. However, phylogeographic methods that map expansion through space and time have yet to be applied to Pama–Nyungan or, indeed, any other large hunter-gatherer language family.

Here, we apply a novel Bayesian phylogeographic approach to analyse newly available data based on the Chirila database[28], recording the presence or absence of 18,238 cognates across 200 vocabulary terms in 306 Pama–Nyungan languages (see Methods). To draw inferences about the earliest branches in the Pama–Nyungan family, our sample spans all 31 major subgroups and includes 7 languages

[1]Center of Computational Evolution, University of Auckland, Auckland, New Zealand. [2]Max Planck Institute for the Science of Human History, Jena, Germany. [3]Department of Linguistics, Yale University, New Haven, CT, USA. [4]School of Psychology, University of Auckland, Auckland, New Zealand. *e-mail: q.atkinson@auckland.ac.nz

from the putative outgroup Tangkic[29–33]. We build on and extend previous work[18,26], modelling language evolution as the birth and death of cognates through time. To more naturally capture the process of language migration, we introduce a phylogeographic 'founder dispersal' model, whereby new territory is colonized at language diversification by one lineage migrating while the other remains. To provide an approximate timescale for the expansion, rates of evolution were calibrated based on archaeological evidence for the colonization of the Western Desert region using a 'relaxed clock'[34] that allows rates to vary across the branches of the tree (see Methods).

Results

The posterior distribution for the location and age of the root of the tree, representing the inferred origin of Pama–Nyungan under our model, is shown in Fig. 1. This reveals clear support for a homeland at the base of the Gulf of Carpentaria, with an estimated age of 4,455–6,966 years (95% highest posterior density; mean = 5,671 years). Both the inferred location and timing of the root fit a mid-Holocene expansion from the Gulf of Carpentaria/north-western Queensland region, as proposed under a rapid-replacement hypothesis. There is very little support for homelands proposed under the rival hypotheses, including the Einasleigh Uplands, Great Dividing Range, Murray–Darling basin and potential colonization routes in the north. One hypothesis (early-Holocene intensification) includes a homeland that partially overlaps this region; however, the inferred timescale is outside the range implied under this hypothesis (Fig. 1b). We quantify the relative strength of prior versus posterior support for the age and homeland combination implied under each origin scenario using Bayes factors (Supplementary Table 2) and find decisive support for the rapid-replacement hypothesis over the three alternative hypotheses.

One factor that makes inferences about Australian prehistory both difficult and consequential is the scarcity of established events that can be used to calibrate rates of change (see Supplementary Note 1). While the fit between the rapid-replacement hypothesis and our inferred location and date of origin is compelling, our time estimates are based on a single age calibration of the Western Desert subgroup and scale approximately linearly with this calibration. We therefore used a broad calibration range that captures uncertainty in the credible age of the subgroup. This calibration would need to be 40–50% older than this range to support the early-Holocene intensification hypothesis, more than 50% older to support an expansion after the Antarctic Cold Reversal and more than 300% older to support an origin with initial colonization. Furthermore, recalculating Bayes factors for our models without incorporating time reveals that support for the rapid-replacement model is not driven simply by our inferred chronology (Supplementary Table 2); the geographic component of our model alone shows substantial support for the origin proposed under the rapid-replacement hypothesis.

Next, we investigated the robustness of our findings to variation in assumptions about the migration process. We inferred an average migration speed for the Pama–Nyungan spread of 0.14 km yr^{-1}, three to four times slower than that observed in the Indo-European agricultural expansion across a comparable range (0.48 km yr^{-1})[26]. While our primary analysis allowed rates of geographic diffusion to vary across branches in the tree, recent Australian genomic and mitogenomic data indicate that gene flow occurred preferentially along the coast and waterways[3,4], suggesting faster movement near water and/or barriers to movement in the arid interior during the Last Glacial Maximum[13]. Conversely, arid or marginal environments are associated with greater range size, larger social networks and increased mobility[35], potentially increasing rates of language spread compared with coastal or riverine areas where greater resource abundance may promote sedentism and diversification over much shorter distances. To accommodate and test between these scenarios, we developed a flexible and computationally efficient graph-based approach to rapidly compare support for models in which rates of movement vary depending on proximity to the coast and major rivers (see Methods and Supplementary Fig. 1). The best-fitting model was one in which rates were two times slower near water (Supplementary Table 3), contradicting proposals that the Pama–Nyungan spread was the result of rapid migration along the coast or waterways, and supporting a link between language spread and ecological factors associated with mobility and range size[35]. Changing the migration model in this way did not substantially affect the inferred origin

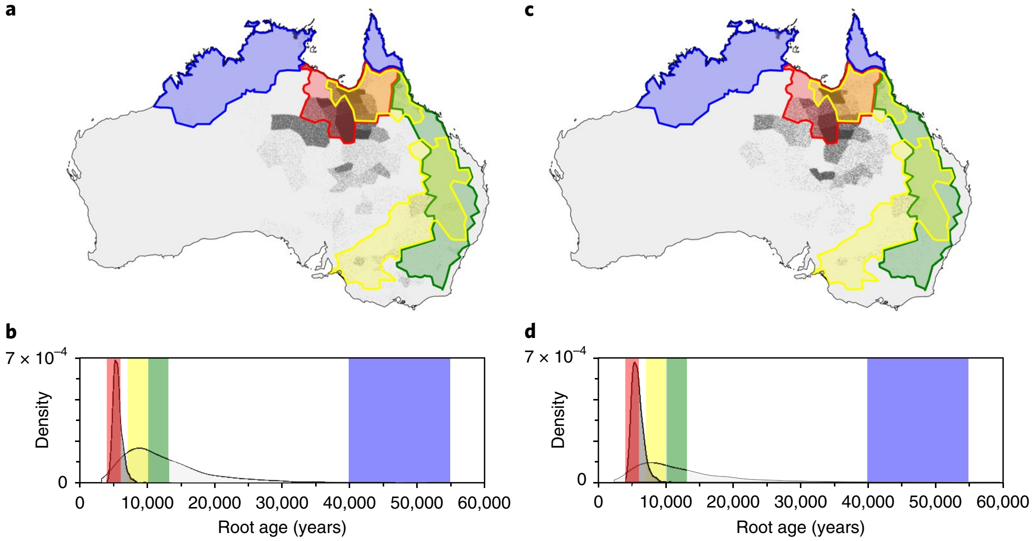

Fig. 1 | Inferred origin of the Pama–Nyungan language family tree. **a**, Map showing the posterior distribution on the root location under the standard founder dispersal model. Darker areas correspond to increased probability mass. Coloured polygons indicate origins implied under the rapid replacement (red), early-Holocene intensification (yellow), post-Antarctic cold reversal (green) and initial colonization (blue) hypotheses. **b**, Histogram showing the prior (light grey) and posterior (dark grey) distributions for the age of the family. Coloured bars indicate hypothesized ages as in **a**. **c**,**d**, Same as **a**,**b**, respectively, for a founder dispersal model with rates two times slower near water; that is, along the coast and adjacent to the Murray–Darling river system (Supplementary Fig. 1). Background map © 2017 ESRI, World Imagery, DigitalGlobe. All rights reserved.

location and age (Fig. 1c,d) or Bayes factor support for the rapid-replacement hypothesis (Supplementary Table 4). In addition to examining variants of the founder dispersal model, we repeated our analyses using a previously published standard Brownian diffusion model of language expansion in which migration rates are drawn from the same distribution in both daughter lineages[25,26] (Supplementary Methods). Again, our findings were robust across this variation in modelling assumptions (Supplementary Fig. 2 and Supplementary Table 5).

Some scholars have argued that languages evolve differently among hunter-gatherer groups and that an assumption of vertical inheritance, rather than horizontal diffusion, is not appropriate[12,17]. Furthermore, the proposal that Pama–Nyungan dates back to the initial colonization of Australia assumes convergence due to widespread areal diffusion[16,17]. In contrast, we infer high branch support values in the posterior distribution of trees consistent with a strong phylogenetic signal in our data. Of the 273 unconstrained internal nodes on our tree, nearly two-thirds (63%) have >90% posterior support. These results provide evidence against widespread word borrowing following initial colonization and are consistent with previous work showing that loans in Pama–Nyungan basic vocabulary are not significantly higher than elsewhere in the world[36] and are well under

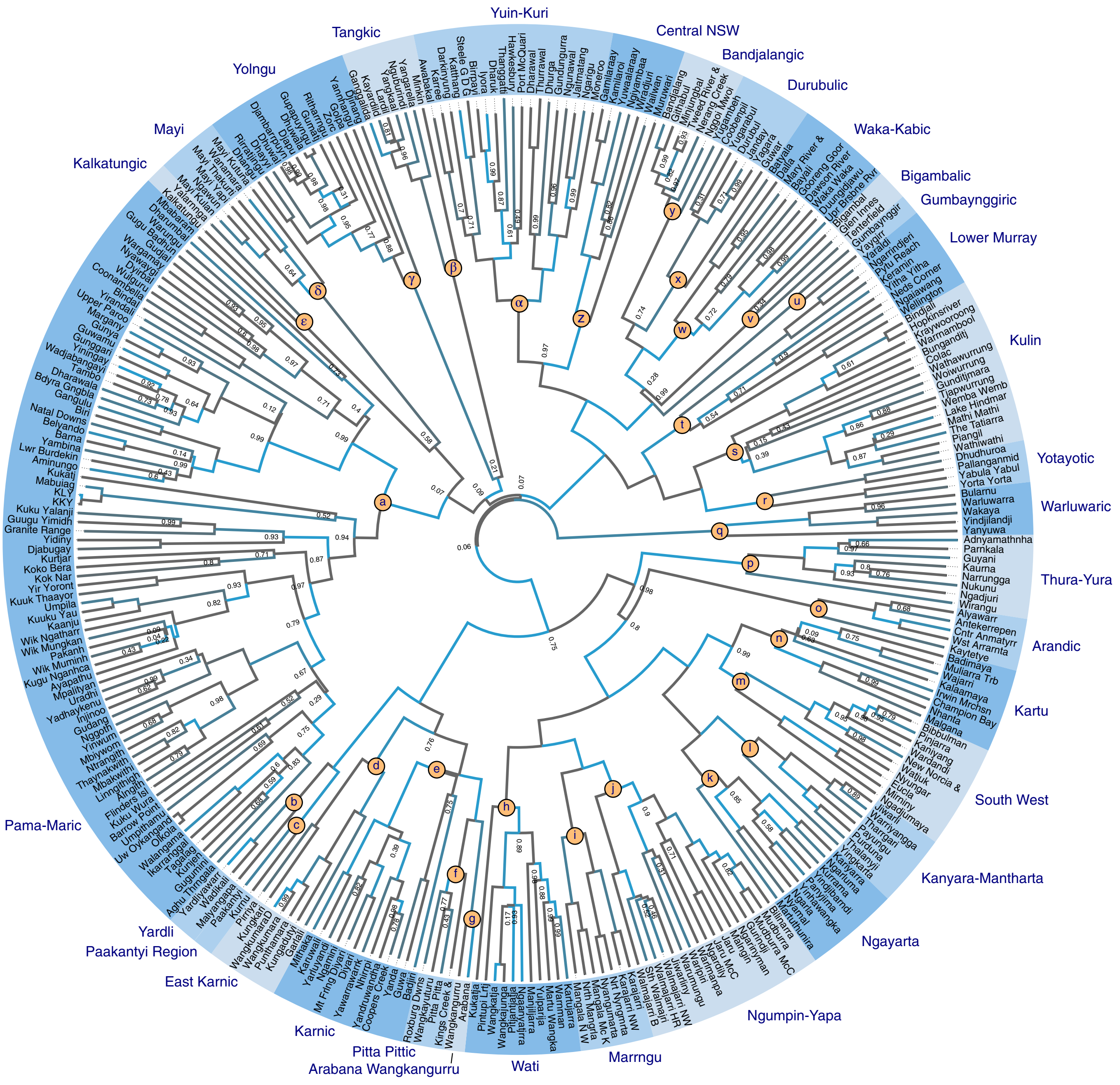

Fig. 2 | Diversification of the Pama–Nyungan language family. Maximum clade credibility tree showing the inferred timing and emergence of the major branches and their subsequent diversification. Values above each branch indicate posterior support for the descendent clade, where support is less than 100%. Letters circled in orange indicate the most recent common ancestor of subgroups as labelled on the tree perimeter. Letter assignment is arbitrary, ordered counterclockwise (a–z then α–ε) from Pama-Maric and corresponds to the letters in Fig. 3. Branches are coloured according to the inferred migration rate from grey (no movement) to light blue (faster rates). Language names are truncated abbreviations, see Supplementary Table 8 for full names and identifiers. The &'s indicate names that comprise more than one location.

a threshold that would impede accurate phylogenetic inference[37]. Our comparison of models of Pama–Nyungan cognate evolution provides further evidence that these languages are not a special case. The best-fitting model, the 'covarion' (see Supplementary Methods and Supplementary Table 6), is also favoured in analyses of language evolution among agriculturalists[26,27,38].

We used two approaches to investigate the robustness of our findings to potential errors in cognate assignments. First, we compared our cognate assignments and tree topology based on the Chirila database[28] with our previous Bayesian phylogenetic classification of 194 Pama–Nyungan languages that used an earlier version of the data[18]. A comparison of cognate assignments across the two datasets provides a count of errors identified after five years of error checking, refinement and consultation with regional language specialists. Among the 38,570 cognate sets represented by languages that occur in both datasets, missing cognates were replaced with attested forms in 1,209 cases, while changes to cognate codes were made for 1,034 cases. While this represents a significant refinement of the data, the implied error rate (changes to cognate codes) was just 2.7% and is unlikely to significantly impact our findings. Consistent with this assessment, the genealogical relationships we infer within and between established groups show broad agreement with the tree topology produced by our earlier analysis (which also used a different model of cognate evolution and did not include a temporal or geographic component), suggesting that our findings are robust to this level of error in the data (see Supplementary Fig. 3 and Supplementary Note 2 for comparison with previous work). As a second robustness check, we quantify the effect of introducing error into our current dataset at rates of 5, 10 and 15% false negative and false positive cognate assignment (Supplementary Methods). In all cases, even for the highest error rates, we continue to find support for a rapid-replacement hypothesis over the alternative hypotheses (Supplementary Table 7).

Beyond the origin of the family, our analyses provide insight into the subsequent breakup and expansion of Pama–Nyungan subgroups (Figs. 2 and 3, Supplementary Fig. 4, and Supplementary Notes 1 and 2). Low branch support values and short branches at the base of the tree support an interpretation of rapid initial diversification. However, among the earliest splits in the tree there are several large and well-supported clades. This includes a Western branch, expanding across South and Western Australia and the Northern Territory to cover an area of 3.75 million square kilometres, and a Southern group, comprising the languages of Victoria, much of New South Wales and the Southeast Queensland coast. Consistent with previous proposals[29–33], Tangkic is among the first groups to separate, although there is also some signal in the data placing the Tangkic branch as a remote sister to the Yolngu languages. The two groups are separated by several non-Pama–Nyungan groups, which makes it less likely that the signal we observe is a result of recent loans. While our migration model does not support faster migration near the coast or rivers, there is some evidence that these areas may be launching points for expansion (Fig. 3). For example, the Eastern, Central and Western Karnic subgroups occur along the Bulloo River, the Diamantina River and Coopers Creek, respectively. Likewise, Gumbaynggirr, Bigambalic, Bandjalangic and Waka–Kabic split sequentially up the coast of Northern New South Wales and Southern Queensland. Another spread, in Western Australia, seems to have occurred along the Gascoyne–Murchison river system. This is consistent with a pattern in which language groups settle along watercourses, undergoing successive group and language fission as they go.

Discussion

Our findings, which integrate over uncertainty in our tree and parameter estimates and hold across a range of models of cognate evolution and migration, support a rapid replacement of

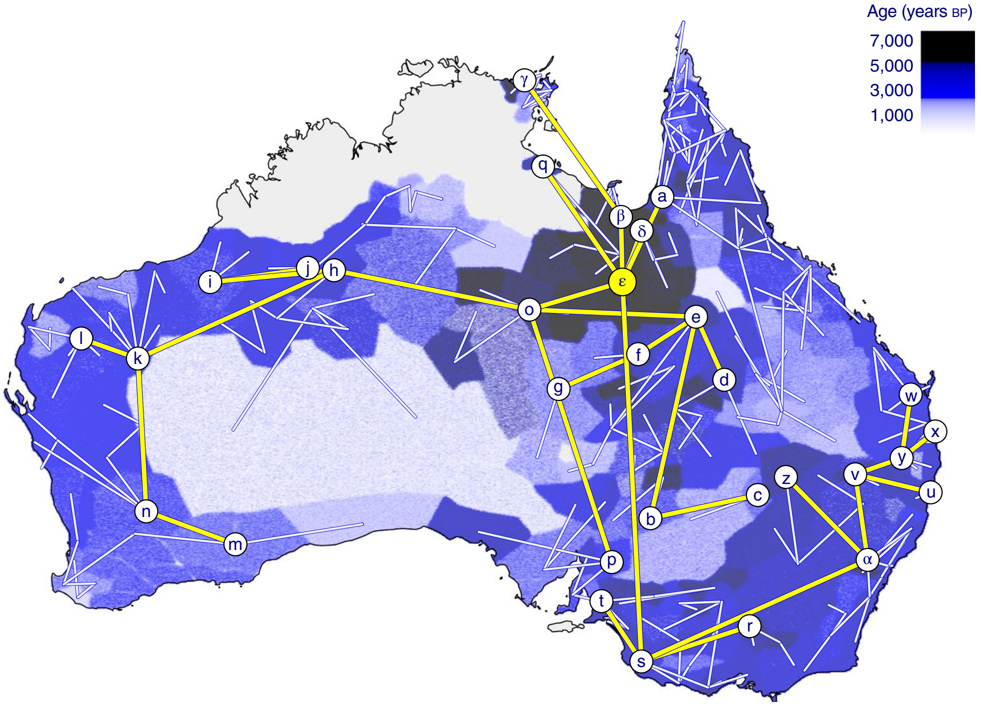

Fig. 3 | Geographical dispersal of Pama–Nyungan languages. Maximum clade credibility tree from Fig. 2 projected onto a map, showing the mean inferred location from the posterior distribution for each of the main subgroups (circles as indicated in Fig. 2). Yellow lines correspond to basal branches linking the main subgroups. Language areas are shaded to indicate the posterior distribution of internal node location estimates through time. Older (darker) nodes are shown on the foreground to clearly depict the temporal diffusion pattern. Our sample represents all known Pama–Nyungan lineages; however, we cannot make inferences about unattested lineages. The chronology represented here therefore offers a minimum age for expansion into a given area. Background map © 2017 ESRI, World Imagery, DigitalGlobe. All rights reserved.

non-Pama–Nyungan by Pama–Nyungan languages from the Gulf Plains region. Such a scenario has been linked to widespread changes in the mid-Holocene archaeological record, including the intensification of land and marine resource use[39–42], new extractive technologies[40,43], rock art[44] and the proliferation of 'backed artefacts' across a range concordant with the distribution of Pama–Nyungan languages[14,45,46]. Archaeological evidence of more recent, localized transitions also accords with the inferred location and timing for the breakup of many of the Pama–Nyungan subgroups (Supplementary Note 1), providing further support for the chronology we infer and highlighting probable connections between the linguistic and archaeological evidence. While the processes that could account for such an expansion among hunter-gatherers remain poorly understood, in the case of Pama–Nyungan, several potential drivers have been proposed. New tools and extractive technologies may have afforded a survival advantage and allowed replacement or recolonization, particularly in the more marginal environments that followed the El Niño–Southern Oscillation climatic shift from 4000–5000 BP[14,40,43,45,46]. It has also been argued that changes to social institutions, including patrilineal kinship, exogamous marriage and multi-group rituals, have played a critical role[7,8].

Whether these changes evident in the linguistic and archaeological data involved demic diffusion and the large-scale movement of people across the continent is unclear. Australian Y-chromosome haplotypes show a mid-Holocene expansion signal, but the sampled populations probably also include non-Pama–Nyungan speakers[47]. Analyses of whole-genome data indicate some support for an expansion from the north-east and find that genetic distances correlate with linguistic distances[3], but the expansion analysis assumes a simple two-population model of Australian genetic diversity and lacks a non-Pama–Nyungan comparison group, and the inferred timing of the split (10–31 ka) does not fit with the archaeological and linguistic chronology. Australian mitogenomic data also support stability and regionalism stretching back well into the Pleistocene and have been used to argue that evidence of more recent expansion is ambiguous[4]. The lack of strong genetic evidence for population expansion after initial colonization supports the proposal that Pama–Nyungan languages spread as part of a suite of technological and social innovations[5,7–9] that probably facilitated assimilation and absorbtion rather than wholesale replacement of the existing hunter-gatherer groups. Our own finding that languages move more slowly near water despite greater gene flow along coasts and waterways[3] also indicates that the movements of genes and languages were not tightly coupled, suggesting that water facilitates the movement of individuals, but makes it less likely that groups (and the languages they speak) will move.

Large-scale language replacement is frequently attributed to demic diffusion[48] related to the spread of agricultural intensification in a more stable Holocene climate[1,49]. Together with archaeological and genetic evidence, the results we report clarify Pama–Nyungan's status as a clear example of hunter-gatherer language replacement and focus attempts to understand the processes at work. More generally, our findings demonstrate how, by locating cultural lineages in time and space, Bayesian phylogeographic methods allow linguistic evidence to be combined with archaeological and genomic data to provide unique insight into the prehistory of hunter-gatherer populations, laying the foundation for further work on the coevolution of genes and culture[50] during this pivotal chapter in the human story.

Methods

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during the experiments and outcome assessment.

Language data. We compiled a binary matrix representing the presence (1) or absence (0) of 18,238 cognate sets across 200 basic vocabulary terms in 306 Greater-Pama–Nyungan languages. This sample spans all the major recognized Pama–Nyungan subgroups and includes seven languages from the well-studied Tangkic group, which is considered to be one of the earliest branches within Greater Pama–Nyungan[29–33]. We did not include languages from the putatively Greater Pama–Nyungan Garrwan group or the more speculative Macro-Pama–Nyungan Gunwinyguan group[51], both of which show few potential cognates with the rest of the family and evidence of extensive regional loans. Words of a given meaning were assigned to a cognate set if they showed shared recurrent sound correspondences thought to indicate common ancestry. For example, the word for boomerang is *boomarring* in Karree, *bumarangga* in Dharuk and *boomerrit* in Birrpayi, which are all cognate, while there are different forms in Duungidjawu (*baran*) and Durubul (*barrakadan*) that are cognate among themselves. The underlying wordlist data were sourced from the Chirila lexical database of Australian languages (http://chirila.yale.edu/)[28]. Cognate judgements were made by C.B. according to the principles of the comparative method[52].

Languages in the Chirila[28] dataset are organized by doculect[53] and holdings for Australian languages are extensive (particularly for Pama–Nyungan) but not complete. Source quality varies extensively across the dataset. Source quality was evaluated in the database on a three-point scale. Where two sources were available for a given language, that of higher quality was used. We also prioritized sources that were explicit about the doculect they referred to; some historical sources combined materials collected from different authors. Full references were given in the Chirila database. We also linked all languages to metadata documenting the time and location at which they were sampled, as well as the subgroup within Pama–Nyungan to which they belong (Supplementary Table 8). The geographic range of sampled languages is shown in Supplementary Fig. 5.

Inferring language trees. We derived a posterior sample of Pama–Nyungan language trees using Bayesian inference and a Markov chain Monte Carlo (MCMC) algorithm as implemented in the BEAST[54] software package and BEAGLE[55] library. MCMC is a stochastic algorithm that performs a random walk through a state space guided by the posterior density on the states it visits. This approach allows us to efficiently sample language trees and model parameters in proportion to their posterior probability, given the data, a model of language evolution and a set of prior assumptions about model parameters. Following previous work[18,26,27,38,56–60], we modelled language change as the birth and death of cognates along the branches of a bifurcating phylogeny, such that the likelihood of all cognates is the product of the likelihoods for each individual cognate[26].

We derived our tree prior using the birth–death skyline model[61], a variant of the widely used pure birth (Yule[62]) tree prior that accounts for the fact that not all of our languages were sampled at the same time (see Supplementary Methods). To improve rate estimates and more efficiently search the set of credible trees, we constrained 31 subfamilies identified by ref. [18] and established on prior grounds within Pama–Nyungan[63–65] as monophyletic (see Supplementary Table 8 for details).

We did not assume a known outgroup. Instead, the appropriate root point on the tree was inferred under the assumption of a relaxed clock (see below). While we found considerable uncertainty in the root point and basal branches of the tree, our estimates for the location and timing of Pama–Nyungan origin were made across the posterior sample of trees, and hence all our inferences integrate over this phylogenetic uncertainty.

Models of cognate evolution. Estimating the tree likelihood requires a model of cognate evolution describing the probability that a given cognate is present or absent at a node on the tree as a function of the state of the parent node and the length of the intervening branch. We evaluated three different model variants that have previously been applied to cognate data: (1) a simple binary continuous-time Markov chain (CTMC) model[26,56]; (2) the covarion model[26,38,66]; and (3) the stochastic Dollo model[67–69].

The CTMC model is analogous to simple nucleotide substitution models that allow uneven transition rates between certain states, such as the widely used Hasegawa–Kishino–Yano model[70]. Applied to binary cognate data, the CTMC model comprises a single parameter that represents the relative rate at which cognates are gained ($0 \rightarrow 1$) and lost ($1 \rightarrow 0$), accounting for the estimated equilibrium frequency of 1s and 0s. We considered the CTMC model both with and without gamma distributed rate heterogeneity across cognates[71]. The Covarion model[66] extends the CTMC model by allowing cognates to switch between slow and fast evolutionary rates across the tree, capturing linguists' intuition that certain words may change more rapidly across part of the tree. Finally, the stochastic Dollo model[38,67–69] is based on the Dollo principle that a feature or cognate is only likely to arise once, but may be lost multiple times. While this fits well with the definition of a cognate, unlike the CTMC and covarion models, the stochastic Dollo model does not allow multiple gains of a cognate on the tree and so cannot accommodate borrowing or other anomalies, such as parallel semantic shift.

For all three models, since the data do not contain entries for latent cognates (not observed in any of the languages in our sample), we used an ascertainment correction for each meaning class and adjusted for missing data according to previously described methods[68,72]. The implementation of each of these models of cognate evolution is explained in more detail in the Supplementary Methods.

We considered two ways in which rates of cognate replacement might vary. First, we compared models under which each meaning class has the same mutation rate with models where each meaning class has its own relative mutation rate. This was motivated by the fact that some meaning classes have been shown to be more stable than others[73], but modelling such variation requires sufficiently informative data to justify the extra 199 parameters. Second, to address the concern that rates of cognate replacement may vary across lineages[20], we compared a strict clock model, in which mean rates are held constant across branches, with an uncorrelated relaxed clock[34] model, in which mean rates are allowed to vary. The relaxed clock lets rates vary across branches according to a log normal distribution. The rate for a cognate on a branch is the product of the relative mutation rate of the meaning class containing the cognate and the branch rate according to the relaxed clock model.

Rate estimates were informed by variation in the sampling times of languages in our sample, together with a calibration on the age of the Wati subgroup based on archaeological evidence for expansion into their current range in the Western Desert region. The current period of occupation of the Western Desert region has been identified to have been from as early as 5 ka until 1.5 ka[13,74], with the greatest activity from 3 ka to 2.5 ka[75]. Although a slightly younger earliest date associated with the shift to an El Niño–Southern Oscillation-dominated climate from 4 ka has been suggested previously[14]. To capture this uncertainty, we constrained the age of the separation of the Wati lineage using a gamma distribution with a 95% highest posterior density interval between 5 ka and 3 ka and a probability mass skewed towards younger ages within this range (modelled as a gamma distribution with a 3,000 year offset, $\alpha=2$ and $\beta=359$). In addition, we constrained the breakup of the Wati subgroup to no later than 1.5 ka, by which time there is evidence of an accelerated population increase and settlement restructuring in the region[14,74,75]. Other calibration points were considered in Queensland, Victoria and the Lake Eyre Basin of South Australia; however, in each case the archaeological evidence could not be linked to a particular node in the Pama–Nyungan tree (see Supplementary Note 1 for further discussion regarding the choice of calibration and correlations between linguistic and archaeological evidence).

Phylogenetic signal and the impact of borrowing. The cognate meaning classes chosen here are resistant (but not immune) to borrowing. Previous work[36] has demonstrated that borrowing across these cognate classes is usually low (around 10% of items in the list). This is well under the threshold identified as distortive to recovering a phylogeny[37] and consistent with what is found in other language families around the world[36]. Posterior branch support values in the maximum clade credibility tree provide further support for clear phylogenetic signal in our data. Of the 273 unconstrained internal nodes on our tree, 63% have >90% posterior support, and more than 93% of clades have >50% posterior support. In comparison, the prior distribution (without cognate data) includes zero unconstrained clades with >50% branch support.

Standard founder dispersal model. We modelled language dispersal as a random walk through continuous space along the branches of a tree[25,26,78]. This approach combines cognate and location data for sampled languages at the tips of the tree to jointly infer ancestral relationships and the location of ancestral nodes. In this way, cognates and geography jointly inform the posterior distribution of trees and inferences about ancestral node locations integrate over uncertainty in the tree topology. We note that a 'random walk' does not mean that each language is moving 'randomly' without influence from social, political or ecological factors, but rather that on average, movement can be approximated by a Brownian diffusion process. This implies that after some finite amount of time, a few languages will have moved far, some will not have moved at all and most will have moved somewhere in between.

Our spatial dispersal model incorporates two innovations. First, standard diffusion-based phylogeographic models[25–27,59,60] assume that following a lineage split, descendent lineages disperse at equal rates. However, this assumption is biased towards a posterior distribution on ancestral nodes (and hence the origin) near the centre of the geographic range of the descendent languages, which is inconsistent with proposed Pama–Nyungan homelands (Fig. 1). We therefore introduced a new and more realistic 'founder-event dispersal' model in which one lineage disperses (the founder) while the other remains. This model naturally captures the expectation that languages arise when a founder population migrates to colonize new territory (analogous to allopatric speciation in biology) and implies a more even prior on the location of the origin across the current range of Pama–Nyungan languages (see Supplementary Fig. 6). Second, the latitudinal range covered by large families like Pama–Nyungan generates distortion when location is simply modelled by treating longitude and latitude as coordinates on a plane. Previous approaches have sought to minimize this effect by translating the coordinate system[26], but cannot eliminate distortion at the extremes of the range. To overcome this problem, here we model diffusion directly on a sphere representing the globe[78].

The founder dispersal model assumes that a population started in some location uniformly chosen from the sample locations on mainland Australia. Each sample location is represented by the centroid of the region over which the language is spoken, or the nearest location within the region in cases where the centroid falls outside the region. For every bifurcation in the tree, we allow a migration event and assume migration occurs only along one descendent branch, while the population on the other branch remains in the same place. The distribution of migration events for the founder group can be described by a diffusion process with a scale defined by a single parameter, called the precision b.

To specify the probability density of tip locations under this migration process, let T be a bifurcating tree over a set of n taxa $x_1, \ldots, x_n$. Internal nodes $x_{n+1}, \ldots x_{2n-1}$ of the tree are numbered $n+1, \ldots, 2n-1$. By convention, the highest numbered node $(2n-1)$ is the root and we use π_i to denote the index of the parent of node 'i'. Each node x_i is associated with a location $\text{pos}_i=(\text{lat}_i, \text{long}_i)$ with latitude lat_i and longitude long_i. Then, the likelihood of observing the set of tip locations for the vector $\text{pos}_{1\ldots n}=\{\text{pos}_1, \ldots, \text{pos}_n\}$ given a tree T, with precision b governing the diffusion process, and other parameters θ (including branch lengths, cognate model parameters and priors) is

$$\begin{aligned} p(\text{pos}_{1\ldots n}|T,b,\theta) \;=\; & \int_{\text{pos}_{n+1}} \cdots \int_{\text{pos}_{2n-1}} \prod_{i=1\ldots 2n-2} \\ & f(x_i=\text{pos}_i|x_{\pi_i}=\text{pos}_{\pi_i}, \theta, b) \\ & f(\text{pos}_{2n-1}|\theta,b)\,d\text{pos}_{n+1}\ldots d\text{pos}_{2n-1} \end{aligned} \tag{1}$$

where the first density $f(x_i=\text{pos}_i|x_{\pi_i}=\text{pos}_{\pi_i}\theta, b)$ represents the migration from parent node x_{π_i} to node x_i and the second density represents the root location prior.

For the founder lineage, $f(x_i=\text{pos}_i|x_{\pi_i}=\text{pos}_{\pi_i}\theta, b)$ is a density representing the probability of migration from the location of x_{π_i} to x_i over the apparent branch length t_i. This length is equal to the length of the branch in time multiplied by the rate associated with that branch (see below). We calculate this density based on a random walk on a sphere[78] using the great circle distance between pos_{π_i} and pos_i (the lineage that remains does not contribute to the likelihood under this model) and use a uniform prior over the root location.

For each internal node x_i, let index h_i be such that x_{h_i} is the child node of x_i that disperses, and g_i is the index of the other child of x_i. Since internal nodes can only be located at tip locations, the state space is greatly reduced for the founder dispersal model, and the integrals in equation (1) can be replaced by sums over all child indices:

$$\begin{aligned} p(\text{pos}_{1\ldots n}|T,b,\theta) \;=\; & \sum_{h_{n+1}} \cdots \sum_{h_{2n-1}} \prod_{i=n+1,\ldots,2n-1} \\ & f(x_{h_i}=\text{pos}_{h_i}|x_i=\text{pos}_i, \theta, b) \\ & p(\text{pos}_{2n-1}|\theta,b) \end{aligned} \tag{2}$$

where pos_i is recursively defined as the position of a leaf (if $i \leq n$) and equal to the position pos_{g_i} otherwise. Although equation (2) can be calculated in polynomial time using dynamic programming resulting in an algorithm similar to Felsenstein's pruning algorithm, the state space increases in size when traversing up the tree. Therefore, it is computationally more efficient to augment the state space by instances of h_i and approximate equation (2) by MCMC, sampling from the augmented distribution:

$$\begin{aligned} & p(\text{pos}_{1\ldots n}, h_{n+1}, \ldots, h_{2n-1}|T,b,\theta)= \\ & \prod_{i=n+1,\ldots,2n-1} f(x_{h_i}=\text{pos}_{h_i}|x_i=\text{pos}_i, \theta, b) \\ & p(\text{pos}_{2n-1}|\theta,b) \end{aligned} \tag{3}$$

To improve the efficiency with which the MCMC algorithm explores the state space under our model of phylogeography, we also developed several novel proposal mechanisms in addition to the default operators in BEAST[54] (Supplementary Methods).

To test for and accommodate variation in migration rates, we fitted two versions of the founder-event dispersal model. The first assumes a random walk governed by a single rate of movement across all branches in the phylogeny. The second uses a relaxed random walk[25,26] analogous to the relaxed clock model of cognate evolution (Supplementary Methods). This model relaxes the assumption of constant rates of movement by allowing rates to vary across branches in the tree according to a log normal distribution. We found strong support for the relaxed random walk over a constant rates model (Bayes factor = 90). The coefficient of correlation was estimated to be 0.83 (with a 95% highest posterior density interval of 0.458–1.1928) for the relaxed clock, so the strict clock geographic model can be rejected. This implies that migration speed varied instead of being constant throughout the tree, although using a strict clock does not change our conclusion (Supplementary Table 2).

Heterogeneous founder dispersal model. The founder dispersal model described above accommodates variation in rates of movement by allowing rates to differ across branches in the tree, but cannot identify and model migration rate variation linked to features of the landscape. Australian human genomic data indicate preferential gene flow along the coasts and waterways of Australia[3,4], suggesting that movement may occur more quickly in these areas. Conversely, coastal and riverine resources may reduce the incentives for groups to migrate near water. To test these predictions and their impact on the inferred origin of Pama–Nyungan, we deployed an alternative landscape-aware migration model.

Rather than a random walk in continuous space, we can approximate continuous space across a set of evenly distributed points with some neighbourhood structure forming a graph, and perform a random walk on the graph. The graph of N nodes overlays the area of interest like a grid, and nodes in the graph are connected to their neighbours. We use indices to indicate node 1, …, N in the graph and define an $N \times N$ rate matrix R as follows:

$$R_{ij} = \begin{cases} f(i,j) & i \text{ and } j \text{ are neighbours} \\ -\sum\limits_{k \neq j} r_{ik} & i = j \\ 0 & \text{otherwise} \end{cases}$$

where $f(i,j)$ is a suitable rate function. Here, $c(i)$/gcd(i,j) where gcd() is the great circle distance between node i and j, and $c(i)$ is a constant determined by the landscape features at location i. Hence, to reflect the assumption that migration along the coast is faster than inland, $c(i)$ can be 1 for nodes in the interior and larger than 1 for nodes near the coast. This rate matrix R defines a CTMC over the nodes in the graph[79], and the probability of arriving at node j after time t when starting at node i can be obtained by matrix exponentiation:

$$f\,(x_{h_i} = \text{pos}_{h_i} | x_i = \text{pos}_i, \theta, b) = P\,(h_i | i, t) = \mathrm{e}^{Rt}\,(h_i, i)$$

where index (h_i, i) indicates entry h_i, i in matrix e^{Rt} (see Supplementary Methods).

We constructed a graph with 1,446 nodes overlaying the Australian mainland and identified all connections adjacent to the coast or Murray–Darling river system (see Supplementary Fig 1). Languages were assigned locations at the node on the graph nearest to the centroid of their range. We assume no migration over the ocean, but movement near the coast or along rivers facilitated by small craft is already captured by our model. We recognize that the coasts and waterways of contemporary Australia have changed significantly over the past 10–50 ky, including in the area of the Gulf of Carpentaria, which could affect the inferred location of the Pama–Nyungan homeland. However, our standard founder dispersal model is not informed by contemporary coastlines and waterways and allows migration over ocean. Hence, to the extent that our findings are consistent across models, our results are likely to be robust to the effects of changing coasts and waterways.

Computation time using the full graph-based CTMC approach is dominated by the geography likelihood calculation and is not feasible for even $N = 1{,}446$. A more efficient approach is to approximate the process by defining distances $c(i)$gcd(i,j) on edges between nodes i and j on the graph and calculate the shortest distance between two nodes in the graph using Dijkstra's algorithm[80]. This captures the approximate distance a random walk has to traverse starting at node i and arriving at node j in time t. When $c(i)$ is constant for all nodes, this model reduces to an approximation of the random walk on a sphere model. To minimize distortion of distances relative to great circle distances between points, we use all eight closest nodes in the grid (north, north-east, east, south-east, and so on), not just the four closest (north, east, south and west). Distances $d(i,j)$ can then be calculated for each pair of nodes i and j in advance, since they are constant throughout the MCMC run.

We compared log marginal likelihoods across a suite of founder dispersal models to evaluate support for differential migration rates along coasts and the Murray–Darling versus inland areas (Supplementary Fig. 1). We considered rates near water that were ten times, five times and two times slower and faster than inland rates, as well as an equal-rates model (equivalent to the standard founder dispersal model). Under each, we also compared the model fit with and without variation in migration rates across branches. The best-fitting model was two times slower near water with constant rates across branches. The next-best-fitting models were the five times and two times slower models with relaxed rates across branches. These three top models fit significantly better (log Bayes factors >10) than any other model, including the equal-rates model (see Supplementary Table 3). Our inferences about the origin of Pama–Nyungan are robust across these landscape-aware models (Supplementary Table 4).

Phylogeographic hypothesis testing. The language evolution and spatial diffusion models were jointly fitted to the data using BEAST 2.4 (ref. [54]). This produced a posterior distribution of Pama–Nyungan language trees with location and age estimates at the root and internal nodes sampled in proportion to their posterior probability. This approach accounts for uncertainty in the phylogeny, age constraints, models of cognate replacement and spatial diffusion process and provides a principled framework for evaluating rival origin hypotheses. Since we model internal node locations as points in space, the inferred posterior location of divergence events represents a combination of the probable range over which ancestral languages were spoken and stochastic uncertainty in the modelled diffusion process. The picture that emerges must be interpreted with the caveat that we model the expansion of language divergence events (not the rapid expansion of a single language), and only between those languages that are in our sample; nodes associated with Pama–Nyungan branches not represented in our sample will not be captured.

We used the inferred location and age of the root of the tree to test between the four theories of Pama–Nyungan origin outlined in Supplementary Table 1, that is, the rapid-replacement hypothesis, early-Holocene intensification hypothesis, post-Antarctic Cold Reversal hypothesis and initial colonization hypothesis. Supplementary Fig. 6 shows the implied geographic range used to test each theory, together with the prior distribution on the location of origin under the standard founder dispersal model and the best-fitting landscape-aware model. The location implied under each hypothesis receives roughly equal support under the prior, with modest bias towards H2 (H1 = 9.6%, H2 = 21.12%, H3 = 11.85% and H4 = 11.9% support a priori), while timing shows modest bias towards H2 and H3 (H1 = 8.35%, H2 = 25.52%, H3 = 20.61% and H4 = 1.01% support a priori). We evaluate the support for these different hypotheses using Bayes factors calculated as $\frac{\text{Posterior(H1) / Prior(H1)}}{\text{Posterior(H2) / Prior(H2)}}$ where the prior represents the probability that the origin will fall within a hypothesized area and/or date range before observing the data.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The authors declare that code supporting the findings of this study, including model parameterization and priors, is documented in XML code in Supplementary Data File 1. BEAST 2 source code is available at https://github.com/CompEvol/beast2.

Data availability. The authors declare that data supporting the findings of this study are available within the paper and its Supplementary Information. Pama–Nyungan vocabulary data are available from the Chirila database (http://chirila.yale.edu/). Cognate presence/absence data are documented in XML code in Supplementary Data File 1.

Received: 7 August 2017; Accepted: 30 January 2018;
Published online: 12 March 2018

References

1. Bellwood, P. *First Migrants: Ancient Migration in Global Perspective* (Wiley Blackwell, Chichester, 2013).
2. McConvell, P. & Bowern, C. The prehistory and internal relationships of Australian languages. *Lang. Linguist. Compass* **5**, 19–32 (2011).
3. Malaspinas, A.-S. et al. A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016).
4. Tobler, R. et al. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature* **544**, 180–184 (2017).
5. Evans, N. & McConvell, P. in *Archaeology and Language II: Archaeological Data and Linguistic Hypotheses* (eds Blench, R. & Spriggs, M.) 174–192 (Routledge, London, 1998).
6. Bowern, C. in *Linguistic Areas* (eds Matras, Y. et al.) 244–265 (Palgrave Macmillan, London, 2006).
7. Evans, N. & Jones, R. in *Archaeology and Linguistics: Aboriginal Australia in Global Perspective* (eds McConvell, P. & Evans, N.) 385–417 (Oxford Univ. Press, Melbourne & New York, 1997).
8. O'Grady, G. N. & Hale, K. L. in *Australian Languages: Classification and the Comparative Method* (eds Bowern, C. & Koch, H.) 69–92 (John Benjamins, Amsterdam, 2004).
9. McConvell, P. Backtracking to Babel: the chronology of Pama–Nyungan expansion in Australia. *Archaeol. Ocean.* **31**, 125–144 (1996).
10. Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M. & Stoneking, M. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc. Natl Acad. Sci. USA* **110**, 1803–1808 (2013).
11. Williams, A. N., Ulm, S., Turney, C. S. M., Rohde, D. & White, G. Holocene demographic changes and the emergence of complex societies in prehistoric Australia. *PLoS ONE* **10**, e0128661 (2015).
12. Clendon, M. Reassessing Australia's linguistic prehistory. *Curr. Anthropol.* **47**, 39–61 (2006).
13. Veth, P. Islands in the interior: a model for the colonization of Australia's arid zone. *Archaeol. Ocean.* **24**, 81–92 (1989).
14. Smith, M. *The Archaeology of Australia's Deserts* (Cambridge Univ. Press, Cambridge, 2013).
15. Williams, A. N., Ulm, S., Cook, A. R., Langley, M. C. & Collard, M. Human refugia in Australia during the last glacial maximum and terminal Pleistocene: a geospatial analysis of the 25–12 ka Australian archaeological record. *J. Archaeol. Sci.* **40**, 4612–4625 (2013).
16. Dixon, R. M. W. *The Rise and Fall of Languages* (Cambridge Univ. Press, Cambridge & New York, 1997).
17. Dixon, R. M. W. in *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics* (eds Aikenvald, A. Y. & Dixon, R. M. W.) 64–104 (Oxford Univ. Press, Oxford & New York, 2001).

18. Bowern, C. & Atkinson, Q. Computational phylogenetics and the internal structure of Pama–Nyungan. *Language* **88**, 817–845 (2012).
19. O'Grady, G. N., Wurm, S. A. & Hale, K. L. *Aboriginal Languages of Australia: (a Preliminary Classification)* (Univ. Victoria, Victoria, BC, 1966).
20. Bergsland, K. & Vogt, H. On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153 (1962).
21. Redd, A. J. et al. Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. *Curr. Biol.* **12**, 673–677 (2002).
22. Hudjashov, G. et al. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl Acad. Sci. USA* **104**, 8726–8730 (2007).
23. McEvoy, B. P. et al. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *Am. J. Hum. Genet.* **87**, 297–305 (2010).
24. Bergström, A. et al. Deep roots for Aboriginal Australian Y chromosomes. *Curr. Biol.* **26**, 809–813 (2016).
25. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
26. Bouckaert, R. et al. Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960 (2012).
27. Grollemund, R. et al. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proc. Natl Acad. Sci. USA* **112**, 13296–13301 (2015).
28. Bowern, C. Chirila: contemporary and historical resources for the Indigenous languages of Australia. *Lang. Doc. Conserv.* **10**, 1–45 (2016).
29. O'Grady, G. N. in *Australian Linguistic Studies* (ed. Wurm, S. A.) 107–139 (Pacific Linguistics, Canberra, 1979).
30. Blake, B. J. in *Aboriginal Linguistics 1* 1–90 (Univ. New England, Armidale, 1988).
31. Evans, N. in *Aboriginal Linguistics 1* 91–110 (Univ. New England, Armidale, 1988).
32. Blake, B. J. in *Language and History: Essays in Honour of Luise A. Hercus* (eds Austin, P. et al.) 49–66 (Australian National Univ., Canberra, 1990).
33. Evans, N. Australian languages reconsidered: a review of Dixon (2002). *Ocean. Linguist.* **44**, 242–286 (2005).
34. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
35. Hill, J. H. in *Language, Archaeology, and History* (ed. Terrell, J.) 257–282 (Bergin and Garvey, Westport, 2001).
36. Bowern, C. et al. Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS ONE* **6**, e25195 (2011).
37. Greenhill, S. J., Currie, T. E. & Gray, R. D. Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306 (2009).
38. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
39. Lourandos, H. Intensification: a late Pleistocene–Holocene archaeological sequence from southwestern Victoria. *Archaeol. Ocean.* **18**, 81–94 (1983).
40. Lourandos, H. *Continent of Hunter-gatherers: New Perspectives in Australian Prehistory* (Cambridge Univ. Press, Cambridge, 1997).
41. Haberle, S. G. & David, B. Climates of change: human dimensions of Holocene environmental change in low latitudes of the PEPII transect. *Quat. Int.* **118**, 165–179 (2004).
42. McNiven, I. J., De Maria, N., Weisler, M. & Lewis, T. Darumbal voyaging: intensifying use of central Queensland's Shoalwater Bay islands over the past 5000 years. *Archaeol. Ocean.* **49**, 2–42 (2014).
43. Smith, M. A. The antiquity of seedgrinding in arid Australia. *Archaeol. Ocean.* **21**, 29–39 (1986).
44. David, B. & Cole, N. Rock art and inter-regional interaction in northeastern Australian prehistory. *Antiquity* **64**, 788–806 (1990).
45. Hiscock, P. Pattern and context in the Holocene proliferation of backed artifacts in Australia. *Archeol. Pap. Am. Anthropol. Assoc.* **12**, 163–177 (2002).
46. Hiscock, P. *The Archaeology of Ancient Australia* (Routledge, London & New York, 2008).
47. Kayser, M. et al. Independent histories of human Y chromosomes from Melanesia and Australia. *Am. J. Human. Genet.* **68**, 173–190 (2001).
48. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science* **259**, 639–646 (1993).
49. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).
50. Richerson, P. J., Boyd, R. & Henrich, J. Gene-culture coevolution in the age of genomics. *Proc. Natl Acad. Sci. USA* **107**, 8985–8992 (2010).
51. Evans, N. *The Non-Pama–Nyungan Languages of Northern Australia: Comparative Studies of the Continent's Most Linguistically Complex Region* (Australian National Univ., Canberra, 2003).
52. Hock, H. H. & Joseph, B. D. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics* (Mouton de Gruyter, Berlin, 1996).
53. Cysouw, M. & Good, J. Languoid, doculect, and glossonym: formalizing the notion 'language'. *Lang. Doc. Conserv.* **7**, 331–359 (2013).
54. Bouckaert, R. R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS. Comput. Biol.* **10**, e1003537 (2014).
55. Ayres, D. L. et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2011).
56. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003).
57. Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C. J. Bayesian phylogenetic analysis of semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. B* **276**, 2703–2710 (2009).
58. Lee, S. & Hasegawa, T. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proc. R. Soc. Lond. B* **278**, 3662–3669 (2011).
59. Lee, S. & Hasegawa, T. Evolution of the Ainu language in space and time. *PLoS ONE* **8**, e62243 (2013).
60. Walker, R. S. & Ribeiro, L. A. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proc. R. Soc. B* **278**, 2562–2567 (2011).
61. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* **110**, 228–233 (2013).
62. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
63. O'Grady, G. N., Voegelin, C. F. & Voegelin, F. M. Languages of the world: Indo-Pacific fascicle six. *Anthropol. Ling.* **8**, 1–197 (1966).
64. Wurm, S. A., Mühlhäusler, P. & Tryon, D. T. *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas: Vol I: Maps. Vol II: Texts* (Walter de Gruyter, Berlin, 1996).
65. Koch, H. & Nordlinger, R. in *The Languages and Linguistics of Australia: A Comprehensive Guide* 23–90 (Walter de Gruyter, Berlin, 2014).
66. Tuffley, C. & Steel, M. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63–91 (1998).
67. Nicholls, G. K. & Gray, R. D. Dated ancestral trees from binary trait data and their application to the diversification of languages. *J. R. Stat. Soc. B Stat. Methodol.* **70**, 545–566 (2008).
68. Alekseyenko, A. V., Lee, C. J. & Suchard, M. A. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.* **57**, 772–784 (2008).
69. Atkinson, Q., Nicholls, G., Welch, D. & Gray, R. From words to dates: water into wine, mathemagic or phylogenetic inference? *Trans. Philol. Soc.* **103**, 193–219 (2005).
70. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
71. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
72. Chang, W., Cathcart, C., Hall, D. & Garrett, A. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**, 194–244 (2015).
73. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720 (2007).
74. Veth, P. Origins of the Western Desert language: convergence in linguistic and archaeological space and time models. *Archaeol. Ocean.* **35**, 11–19 (2000).
75. Thorley, P. & Gunn, R. Archaeological research from the eastern border lands of the Western Desert. In *Paper for the Western Desert Origins Workshop* (Australian Linguistic Institute, Canberra, 1996).
76. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
77. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013).
78. Bouckaert, R. Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ* **4**, e2406 (2016).
79. Stewart, W. J. *Introduction to the Numerical Solution of Markov Chains* Vol. 41 (Princeton Univ. Press, Princeton, 1994).
80. Skiena, S. *Dijkstra's Algorithm Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica* 225–227 (Addison-Wesley, Reading, MA, 1990).

Acknowledgements

We thank P. Hiscock, D. Kühnert, M. Smith, M. Stoneking, P. Veth and A. Williams for helpful advice. R.R.B. and Q.D.A. were supported by a Royal Society of New Zealand Marsden grant (UOA1308). Q.D.A. was supported by a Rutherford Discovery Fellowship (RDF-UOA1101). C.B. is supported by National Science Foundation grants BCS-0844550 and BCS-1423711.

Author contributions

C.B. and Q.D.A. conceived the study. C.B. collected and prepared the data. R.R.B. designed and performed the analyses and prepared the figures and Methods, with input from Q.D.A. and C.B. Q.D.A. wrote the main text with extensive input from C.B. and R.R.B.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0489-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Q.D.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Quentin Atkinson

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We sampled 306 Pama-Nyungan languages (~90% of the family). To the extent that it is possible to increase this sample size, we do not expect this will improve the the accuracy or precision of our date and location estimates, because our sample already represents all the established lineages within the family. Further, the Bayesian approach we deploy means all our inferences are based on a principled quantification of the uncertainty in the parameters of interest.

2. Data exclusions

Describe any data exclusions.

Languages were sampled from the Chirila database (http://www.pamanyungan.net/chirila/). We included Pama-Nyungan languages and 7 languages from the putative Tangkic outgroup. We aimed for as complete geographic and linguistic coverage of the family as possible. We sampled from all subgroups and regions, reaching 90% coverage of the estimated 343 languages of the family. 37 languages could not be included either because no or little data are attested (e.g. Pirlatapa), or because data were not available in Chirila. While we aimed to include only languages which attested more than 60% of the words on the wordlists, in the interests of ensuring complete geographical coverage, some areas included language lists with more sparse resources.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Findings were reproduced under a range of different model assumptions as reported in the manuscript.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

This is not relevant to our study because it does not include an experimental treatment.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

This is not relevant to our study because it does not include an experimental treatment.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☐ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ A statement indicating how many times each experiment was replicated

☐ ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

☐ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

☐ ☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted

☐ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

☐ ☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All analyses were conducted using BEAST 2 software package available here - https://www.beast2.org/. We use the Babel package for BEAST 2, as well as the “break-away” package, which depends on the GEO_SPHERE and BEASTLabs packages, all available via the package manager in BEAUti. The BEAST xml code is included in the manuscript's supplementary material.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants.