PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2011) **278**, 3662–3669
doi:10.1098/rspb.2011.0518
Published online 4 May 2011

Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages

Sean Lee* and Toshikazu Hasegawa

Department of Cognitive and Behavioral Science, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, 153-8902 Tokyo, Japan

Languages, like genes, evolve by a process of descent with modification. This striking similarity between biological and linguistic evolution allows us to apply phylogenetic methods to explore how languages, as well as the people who speak them, are related to one another through evolutionary history. Language phylogenies constructed with lexical data have so far revealed population expansions of Austronesian, Indo-European and Bantu speakers. However, how robustly a phylogenetic approach can chart the history of language evolution and what language phylogenies reveal about human prehistory must be investigated more thoroughly on a global scale. Here we report a phylogeny of 59 Japonic languages and dialects. We used this phylogeny to estimate time depth of its root and compared it with the time suggested by an agricultural expansion scenario for Japanese origin. In agreement with the scenario, our results indicate that Japonic languages descended from a common ancestor approximately 2182 years ago. Together with archaeological and biological evidence, our results suggest that the first farmers of Japan had a profound impact on the origins of both people and languages. On a broader level, our results are consistent with a theory that agricultural expansion is the principal factor for shaping global linguistic diversity.

Keywords: linguistic evolution; phylogenetics; Japonic languages

1. INTRODUCTION

Considerable controversy surrounds prehistoric factors that shaped the patterns of linguistic diversity. On the one hand, farming/language dispersal theory argues that agricultural population expansions since the end of the last Ice Age played a critical role in shaping major patterns of linguistic diversity [1]. This theory posits that geographically uneven opportunities to domesticate wild plants and animals allowed some Holocene societies to acquire agriculture, which subsequently brought about technological advancements based on social stratification, and ultimately led to the outward population expansions of farming societies, thereby shaping both human genetic and linguistic diversity into the areas in which they settled.

On the other hand, diffusion/transformation theory argues that agricultural population expansion and the patterns of linguistic diversity are not closely linked [2]. This theory posits that agricultural social intensification is merely one of the many factors that influence biological and linguistic diversity, and cultural/technological innovations can diffuse between societies [3]. According to this view, it is perfectly possible that diffusion between societies allowed hunter–gatherer societies to adopt cultural innovations from the farmers and gradually transform themselves into modern societies, while maintaining their own genetic and linguistic make-ups.

Until recently, however, it was difficult to investigate which of these two theories has more explanatory power in a systematic manner, owing to the lack of a suitable methodology to study languages. Early attempts to study the patterns of linguistic diversity such as lexicostatistics and glottochronology did not survive scientific scrutiny. This was because these methods not only failed to distinguish shared innovations from shared retentions, but also misconceived that rates of linguistic change both within and between languages were universally constant. Consequently, these methods were often found to produce misleading inferences about divergence times as well as topological relationships among languages [4,5].

Fortunately, recent progresses in phylogenetic methods and their application in studying languages were found to provide adequate solutions for these problems [6]. Accumulating empirical evidence suggests that languages have, astonishingly, gene-like properties in numerous aspects and they also evolve by a process of descent with modification (for review, see [7]). This implies that once the shared innovations among languages are revealed by converting linguistic signals (i.e. presence or absence of homologous words) into discrete binary characters, various stochastic phylogenetic techniques for modelling biological evolution can be used to adequately reconstruct the history of language evolution. During the last decade, therefore, these techniques were quickly adopted to critically examine, and subsequently corroborate, instances of farming/language co-dispersal for Bantu [8], Indo-European [9] and Austronesian speakers [10].

Nevertheless, the debate about the major prehistoric factors that shaped linguistic diversity on a global scale is yet to be resolved. In order to settle this controversy, rigorous and systematic investigations need to be carried out for the remaining language families around the world and the overall coherence of the results must be made available.

*Author for correspondence (seanlee@darwin.c.u-tokyo.ac.jp).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2011.0518 or via http://rspb.royalsocietypublishing.org.

Received 9 March 2011
Accepted 15 April 2011

The evolutionary history of the Japonic language family that consists of mainland Japanese, Ryukyuan and their dialects [11] provides an invaluable opportunity to examine the predictions made by the two competing theories. According to the farming/language dispersal theory, a significant number of farmers from the Korean peninsula expanded into the southwest Japanese island of Kyushu around 1700–2400 years before present (YBP) bearing their pottery styles, agricultural tools and weapons [12]. Also, the evidence from cranial, dental and Y-chromosome analyses suggests that modern Japanese people derive primarily from a hybrid population between dominant farmers and peripheral hunter–gatherers [13–15]. The farming/language dispersal theory thus predicts that time depth for the root of Japonic languages would go back to the time when the first farmers arrived in the Japanese archipelago.

In stark contrast, the diffusion/transformation theory argues that the major settlement of the archipelago occurred sometime between 12 000–30 000 YBP and that the modern Japanese populations are more or less direct descendants of these initial settlers [16]. This theory is supported by genetic data [17] and posits that the apparent transition from hunter–gatherer lifestyle to that of agriculture, as indicated by the archaeological evidence, is a result of cultural diffusion. Accordingly, this theory predicts that the evolutionary history of Japonic languages is not critically linked with the arrival of the farmers.

In this paper, we used a Bayesian phylogenetic method to reconstruct the evolutionary history of 59 Japonic languages. It was hypothesized that if the recent agricultural population were responsible for shaping the diversity of Japonic languages, then the time depth of Japonic origin would be located within 1700–2400 YBP. Conversely, if the older Pleistocene population made the majority of contribution to the linguistic make-up of this region, then the time depth may be found within 12 000–30 000 YBP. In order to estimate the root divergence time, the posterior probability distribution of Japonic language trees was inferred with probabilistic sampling dates for two ancient languages and a probabilistic divergence time calibration for a pair of extant languages. In addition, four different ways of modelling language evolution were evaluated with Bayes factor (BF) tests in order to find an optimal evolutionary model for the current data.

2. MATERIAL AND METHODS

(a) *Lexical data*

Lexical data consist of 59 lists of 210 basic vocabularies (figure 1; a full list is in the electronic supplementary material, figure S1). The basic vocabularies are words for body parts, kinship terms, basic verbs, numbers and pronouns [18]. These words are known to be resistant to change and unlikely to be borrowed between languages [19]. Wordlists were extracted from previously published etymological dictionaries as well as lexicons published in the linguistic literature. More specifically, the wordlist for Old Japanese was extracted from Jōdaigo Jiten Henshū Iinkai [20] and Yasumoto & Honda [21]. The wordlist for Middle Japanese was obtained from Muromachi Jidaigo Jiten Henshū Iinkai [22]. The wordlists for the rest of the Japonic languages and dialects were extracted from lexicons compiled by Hirayama [23,24].

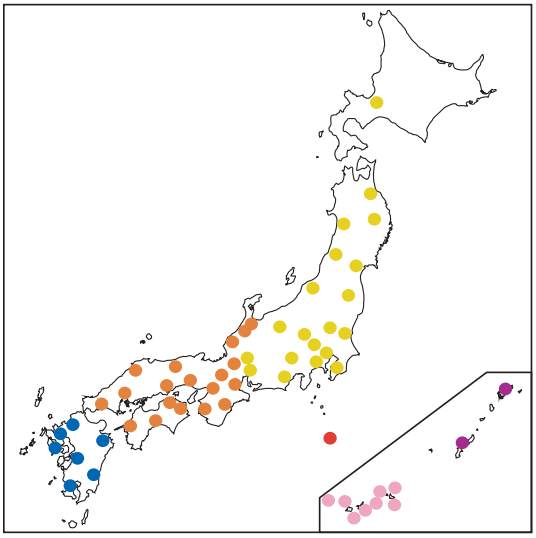

Figure 1. Map of Japonic languages. Subgroups are coded with colour circles: yellow, eastern Japanese; orange, western Japanese; red, Hachijyo; blue, Kyushu; purple, northern Ryukyuan; pink, southern Ryukyuan.

We made cognate judgements by (i) relying on previously identified sound correspondences that were used for reconstructing proto-Japonic ([25]; B. E. Riley 2003, unpublished PhD thesis; J. B. Whitman 1985, unpublished PhD thesis); (ii) working out systematic sound correspondences based on the comparative method [26]; and (iii) consulting previously published cognate judgements in glottochronological studies on Japonic languages [27,28]. For example, there is a known systematic sound correspondence between Tokyo dialect /*j*/ and Yonaguni /*d*/. In addition to this, we identified that the mainland Japanese vowel /*u*/ following affricative consonant /*t*ʃ/ systematically corresponds to the Ryukyuan /*i*/ following the same consonant. With these corresponding patterns, it is clear that Tokyo dialect *jaQt*ʃ*u* (eight), *joQt*ʃ*u* (four), *joru* (night) systematically corresponds to Yonaguni *dat*ʃ*i*, *dut*ʃ*i*, *duru*, respectively. This means that each pair of words is cognate despite their difference at the phonetic level. The cognate sets were encoded into binary states showing presence (‘1’) or absence (‘0’) of a cognate, which resulted in a 59 × 675 matrix (see the electronic supplementary material, tables S2 and S3).

(b) *Phylogenetic analysis*

Phylogenetic analyses were conducted with Bayesian Evolutionary Analysis by Sampling Trees (BEAST), v. 1.6.1 [29]. BEAST uses Bayesian Markov chain Monte Carlo sampling methods to approximate the posterior distribution with sample frequency distribution. This particular application was chosen because it can construct phylogenies without specifying an *a priori* outgroup (which is uncertain for an isolated language family like Japonic) by using a strict clock model or a relaxed clock model. BEAST and its evolutionary clock models have previously been used to make inferences about the origins of Semitic languages [30].

In order to search for the most appropriate model of evolution for our data, we constructed phylogenies with four different ways of modelling language evolution, all of which adopted a single time-reversible substitution rate for cognate gain and loss; accommodating for rate variation within language with gamma correction [31] or covarion model

[32]; and accounting for rate variation between languages with a strict clock or a relaxed clock model [33]. A full description of these models can be found in the electronic supplementary material.

The evolutionary rate and time depth of root were estimated by incorporating calibration priors. This was done with two relaxed sample date calibrations and a probabilistic divergence time calibration. The relaxed sample date calibrations were assigned to Old and Middle Japanese with lognormal priors in which their 95 per cent of the distributions lie between the upper and the lower bounds of each era; 1216–1300 YBP for Old Japanese [20] and 437–674 YBP for Middle Japanese [22]. These time-dated tips were used in a manner similar to how a leaf-dating method is applied to ancient DNA data in order to deal with uncertainty in temporal information [34]. This is deemed to be more appropriate than assigning point calibrations because the wordlists for Old and Middle Japanese are compilations of the lexicons from a range of sources collected over a period of time.

A probabilistic divergence time calibration prior was assigned to Tokyo and Kyoto. From historical records, it is clear that the city of Kyoto has been the political centre of Japan from around 1200 YBP until the Tokugawa military regime took control of the country and moved the government to the city of Edo (present Tokyo) 407 YBP. In addition, both the historical records and linguistic evidence suggest that following the shift of power, the governing elite, merchants and craftsman settled into the new capital, and their languages (ancient western language spoken in the old capital of Kyoto) fused with native dialects spoken by the original inhabitants of Edo to give rise to a distinct dialect, which later evolved into Tokyo dialect [35,36]. Accordingly, we incorporated this prior information into our analyses by assigning a normally distributed prior to the Tokyo–Kyoto pair with the mean of 407 YBP (i.e. the year the Tokugawa regime was established in Edo) and the standard deviation of 135.2 years. This means that 95 per cent of the distribution lies between 142 and 549 YBP, with a 2.5 per cent percentile at the lower bound of the era and a mirroring range towards the 97.5 per cent percentile. The mirroring range was incorporated to introduce some amount of uncertainty into the calibration.

For tree topology, we used a constant coalescence prior informed by Jeffrey's prior (or $1/x$ prior) as this prior makes the simplest assumption and has the least amount of influence on the evolutionary rate.

All four models were run for 30 000 000 steps, with samples taken every 3000 steps. This produced a sample of 9000 trees for each model after discarding the first 1000 trees as burn-in. Post-run inspections using Tracer v. 1.5 [37] indicated that all chains reached convergence after the burn-in and all parameters obtained sufficient effective sample sizes in all four models (greater than 100; chain length divided by the autocorrelation time).

In order to select the most suitable model of evolution for our data, BF was estimated from each pairwise model comparison via importance sampling [38,39]. This method uses the smoothed harmonic mean of sampled likelihood distribution as a means to estimate marginal likelihood, and then produces BF from the difference in the marginal likelihoods between two models in comparison. Significance of BF from each model comparison was evaluated according to a conventional benchmark [40].

3. RESULTS

(a) *Model selection*

A series of BF tests indicated that a model using a relaxed clock with the covarion model (uncorrelated lognormal distribution (UCLD) + Cov) was the best fit for the data. According to the conventional benchmark [40], BF greater than 12 means a model is strongly favoured; BF between 12 and 3 means a model is slightly favoured; and BF between 3 and 1 is considered trivial. The UCLD + Cov model's BF was consistently well beyond 12 when compared with other models (BF ranging from 92 to 538). Thus, all reports of the results hereafter are based on the UCLD + Cov model.

(b) *Date estimation*

It was predicted that if the farming/language dispersal theory were correct, then the estimated time for the root of Japonic language phylogenies would be found within 1700 and 2400 YBP. On the other hand, if the diffusion/transformation theory were correct, then the root would be found anywhere between 12 000 and 30 000 YBP. As it can be seen in figure 2, the median age of the root, 2182 YBP (the mean: 2398 YBP; the standard error: 47.21 years; 95% HPD: 1239–4190 YBP), is clearly in concordance with the age range of the farming/language dispersal theory. All node heights in the tree are scaled to match the posterior median node heights since some of the node height distributions are slightly skewed. Figure 3 is a histogram of the estimated time for the root of Japonic languages. The root time estimates were consistent across all four models and reasonably robust even when the Tokyo–Kyoto calibration was removed.

To investigate the reliability of the result, another set of analyses was carried out with an independent set of Japonic language data compiled by Starostin [41] (available from http://starling.rinet.ru/main.html). We acknowledge that Starostin's data are considered controversial by some scholars, as he made debatable reconstructions of proto-Japonic and used them to argue for genetic relationships to other equally debatable proto-languages (e.g. proto-Tungusic). However, this was not deemed to be a major problem for the current purpose, as those controversial reconstructed lexicons are excluded from our analyses. The data consist of 110 basic vocabulary lists on nine Japonic languages. After the cognate sets were converted to binary codes, a series of model testing was conducted. The maximum clade credibility tree from the best fitting model indicates that the median root divergence time is 1976 YBP (the mean: 2080 YBP; the standard error: 9.13 years; 95% HPD: 1232–3279 YBP), clearly in agreement with the estimation from our data (see the electronic supplementary material, figure S2).

(c) *Phylogenies and networks*

The maximum clade credibility tree from the posterior distribution of 9000 samples is shown in figure 2. A value on each branch of the tree is the posterior probability, which shows the percentage support for the following node (i.e. subgroup). The posterior probability close to 100 per cent implies that there is a strong support for the subsequent node. Any values below 50 per cent

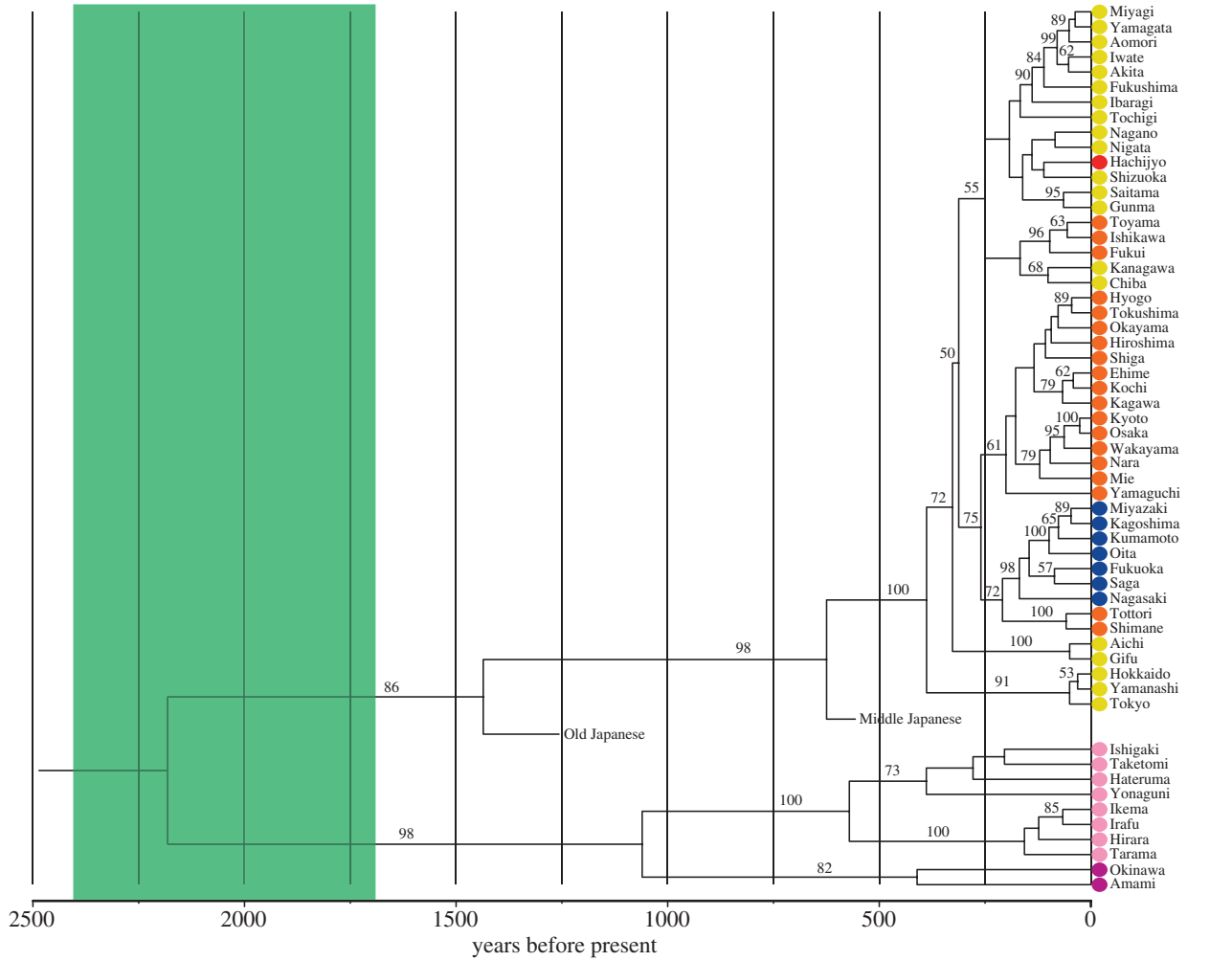

Figure 2. Maximum clade credibility tree of Japonic languages. All node heights in the tree are scaled to match the posterior median node heights. The value on each branch of the tree is the posterior probability, showing the percentage support for a node following a particular branch. Posterior probabilities below 50% are not shown. The green bar represents the age range predicted by the farming/language theory (1700–2400 YBP).

indicate weak support and are thus not shown in the figure.

The tree topology correctly reflects the expected two major subgroups in the Japonic family: the Ryukyuan group (purple and pink circles) and the mainland Japanese group (yellow, orange, blue and red circles) [11,24]. Within the Ryukyuan group, the tree correctly reflects that there are two minor subgroups: northern (purple circles) and southern Ryukyuan (pink circles). Also within the mainland Japanese group, all of the blue circles (Kyushu) and most of the orange circles (western Japanese) cluster with those of the same colours, in addition to these two clusters forming a minor subgroup in accordance with their geographical proximity (figure 1 or the electronic supplementary material, figure S1).

However, our phylogeny fails to recover the entire eastern and western Japanese, and node supports are generally low (below 70%) among mainland Japanese dialects. The cause of this outcome may be explained at two levels: proximate and ultimate. At the proximate level, low node supports among dialects occurred because the isoglosses that separate dialects are small and they do not overlap together; therefore, the algorithm ended up exploring several potential subgrouping patterns with similar probabilities (i.e. dialect chains; [26,42]).

This also means that the relationships among mainland Japanese dialects are non-tree-like, in which its extent can be visualized with NeighborNet analysis [43], as shown in figure 4. It might be expected that under the 'dialect chain formation/break-up' model of lexical evolution [44] in which the intermediate dialect chains would be pruned and produce a tree-like linguistic relationship as a function of time [45], one should see more or less the same amount of reticulations from both mainland Japanese and Ryukyuan dialects, as they are both descendants of a 2200-year-old common ancestor. However, the split graphs in figure 4 seem to suggest that mainland Japanese on the left side has a significantly higher level of conflicting signals than Ryukyuan on the right. If correct, one possible cause for relatively low node supports among mainland Japanese dialects may be found at the ultimate level; the difference in the degree of internal linguistic contact [42] within mainland Japanese and Ryukyuan. An obvious difference between the two groups is that whereas each Ryukyuan dialect is contained within a geographically isolated island, mainland dialects are connected to their neighbours via land routes. Thus, the lack of geographical barriers might have slowed down the pruning process among mainland Japanese dialects (either by allowing horizontal

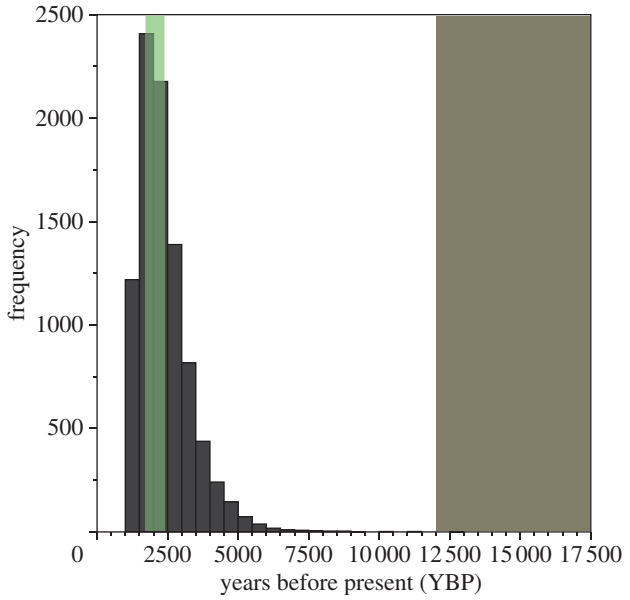

Figure 3. Histogram of the estimated time for the root of Japonic languages. Green bar represents the age range predicted by the farming/language theory and grey bar represents the age range predicted by the diffusion/transformation theory. The median root divergence time is 2182 YBP and the mean is 2398 YBP with the standard error of 47.21 years. The 95% highest probability density is 1239–4190 YBP.

transmission or complex population diffusion), and this could potentially be the cause of low node supports in that part of the tree.

If geographical isolation really did play a role to moderate lexical evolution of Japonic, then the same process may also be observed for other language families; subgroups of the same language family spoken in plains would be slower to achieve tree-likeness than those spoken in mountainous terrains or islands, given that both do not differ significantly in their divergence times.

4. DISCUSSION

The results presented here indicate that the origin of Japonic languages is closely bound with the arrival of the first farmers around 2400 YBP. Together with the archaeological and biological evidence [13–15], these results imply that the first farmers had a profound impact on the origins of people as well as languages in the Japanese archipelago. These observations support the farming/language dispersal theory, which posits that agricultural expansion is the principal factor for shaping the patterns of human genetic as well as linguistic diversity [1].

If our results are correct, one surprising aspect of prehistoric Japan becomes apparent; the hunter–gatherer population, which settled in Japan around 12 000–30 000 YBP, managed to fend off the farmers for thousands of years until being abolished suddenly and dramatically with the arrival of proto-Japonic-speaking farmers around 2400 YBP. To place this in perspective, it should be noted that the hunter–gatherer societies and their languages in Europe began to be abolished by those of the farmers as early as 8500 YBP [9]. Even some of Japan's closest neighbours such as China had started agriculture since 9000 YBP [1], which progressively brought about fully fledged kingdoms equipped with metal tools fighting each other for political unification. During all this transition outside, the hunter–gatherers of Japan continued to prosper by using simple stone tools and without adopting full-scale agriculture, despite knowledge of cultivation of many crops [12]. There are probably two reasons that explain their unusually long survival. First, the population size of the hunter–gatherers may have been too large to be invaded by nearby farmers. The hunter–gatherer of Japan was perhaps one of the most affluent hunter–gatherers known to humankind, endowed with a large range of plants, animals and sea foods [46]. This vast availability of food resources is probably related to the fact that the world's oldest known pottery was made by the hunter–gatherers of Japan [47]. The development of pottery meant that unlike other hunter–gatherers around the world, they had a means to cook and store the foods that were available abundantly in their environment, and such could have triggered a population explosion to the extent that it prevented the farmers asserting any force over the hunter–gatherers for a long time. The second reason behind their long survival could be that it probably took a few thousand years for the farmers to modify rice, one of their main food sources, to grow in cold climate [48]. The archaeological evidence suggest it was not until around 3500 YBP that rice farming of warm southern China spread to the much colder Korean Peninsular [49], which is thought to be the most recent homeland of proto-Japonic-speaking farmers. A combination of these two factors might have contributed to the unusually long occupation of the hunter–gatherers in Japan.

Our Japonic phylogeny seems to imply that the arrival of the farmers did not necessarily lead to a burst of language diversification, or if it were the case, our tree would have a series of short branches following the root leading to the tips. This demands an explanation because (i) soon after their arrival, the proto-Japonic farmers were already divided into several chiefdom-like political units fighting each other to gain access to resources, as indicated by archaeological evidence of defensive moats surrounding settlements, arrowheads and skeletons damaged by sharp objects [50], and (ii) a fully fledged centralized government makes its first appearance 1000 years after the arrival of the farmers; the Nara era that spoke Old Japanese. If we are correct to assume that languages separated by political barriers may take different evolutionary paths [42] and that political power in Japan was not unified for a long time, then there could have been more linguistic diversity in early Japan. Our source for Old Japanese also acknowledges that there could have been some linguistic variations in the Nara era [20]. At present, we do not know the fates of those ancient languages, if there were any. There are two possibilities. One possibility is that the early linguistic diversity could have been wiped out with the emergence of a strong centralized political power in the Nara era, and hence leaving no traces behind. The other possibility is that the early chiefdom-like political units were not able to maintain their states long enough to give rise to any detectable language splits. Further research would be required to clarify this matter.

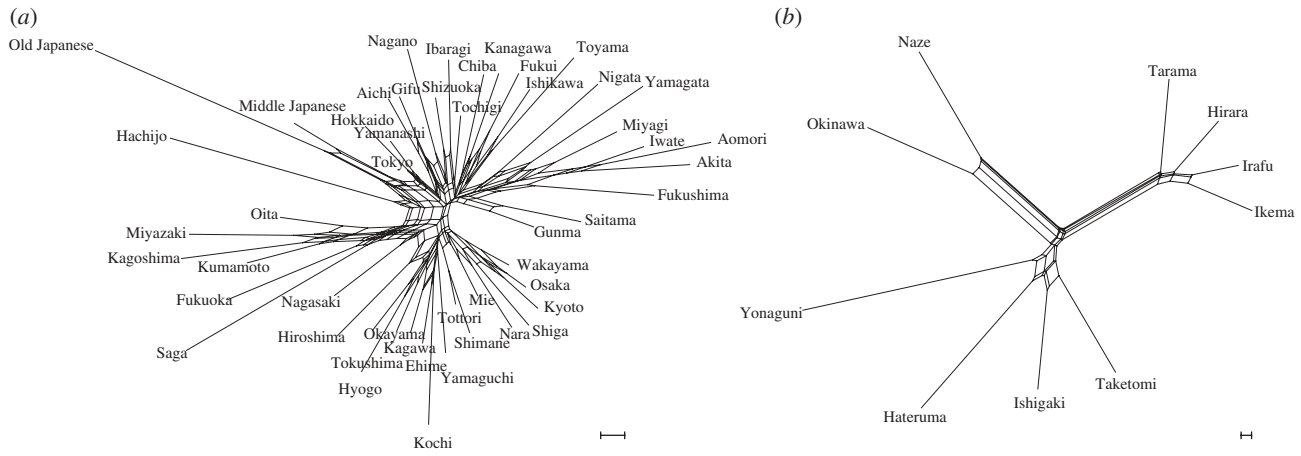

Figure 4. NeighborNet analyses of mainland Japanese and Ryukyuan. Gene-content distances were used and splits were filtered to a threshold of 0.001. Reticulations indicate presence of conflicting signals or dialect chains. Scale bar: 0.01.

A synthesis of evidence from linguistic phylogenetics, biology and archaeology can potentially provide a powerful framework for testing the plausibility of hypotheses regarding possible genetic relationships of Japonic to other language families. For example, if we are correct to think that the agricultural population and their languages dispersed from the Korean Peninsula 2400 YBP, then the hypothetical Japonic-Austronesian family (e.g. [51]) seems implausible because (i) the hypothesis fails to provide neither archaeological nor biological evidence, demonstrating that Austronesian speakers travelled through China and Korea for thousands of years to make their appearance in Japan, and (ii) even if we assume that Austronesian farmers arrived in Japan via the islands of Ryukyu around 2400 YBP and gave rise to Japonic languages, it still cannot explain why the agricultural lifestyle introduced by Austronesian farmers was based on the models found in the Korean Peninsula [12]. Similarly, a hypothesis of linking Japonic with Altaic languages also seems problematic because it contradicts biological evidence that more distant ancestors of Japonic-speaking people are likely to be found in Sino-Tibetan-speaking Southeast Asia, rather than in Altaic-speaking Central Asia [15,52]. The search for genetic links of Japonic to other language families is an area that deserves more attention, and the integrated framework presented here may play an important role in this respect.

A potential problem with any phylogenetic approach to lexical data is a horizontal transmission of words between languages. This is particularly problematic since it is difficult for one to be absolutely sure that every single borrowed word has been detected. However, a recent simulation study suggests that the amount of undetected borrowing needs to be unrealistically high (greater than 20% per 1000 years) to invalidate time estimation or tree topology [53]. We thus argue that our results would withstand some amount of undetected horizontal transmission that may exist in the data. But at the same time, we also acknowledge that high levels of horizontal transmission do occur in some cases [54], and our results would be seriously affected if this turns out to the case. Another potential criticism is that our method cannot accurately distinguish between two theories because the proposed time depth of the diffusion/transformation theory (12 000–30 000 YBP) lies beyond the limit that many linguists consider recoverable from linguistic data (8000–10 000 YBP). However, it should be noted that this limit is estimated under an assumption that all words evolve at the same rate, which is very unrealistic. A recent work suggests that a more realistic model of cognate evolution, which is similar to the models used here, allows linguistic ancestry to be detected even after 20 000 years [55]. Therefore, if our data had any signals indicating deep evolutionary relationships in support of the diffusion/transformation theory, then our method would have accurately reflected such signals to the node heights.

A phylogenetic approach is one of the most notable endeavours to produce insights into the mode and tempo of language evolution. But this approach alone cannot reveal all the details of language evolution and there are other Darwinian approaches that are making great contributions to advance our knowledge of the phenomenon. A few examples are: languages tend to evolve in a punctuational burst-like manner following speciation events [56]; frequency of word-use in everyday speech contributes to evolutionary rate heterogeneity within languages [57]; frequency of use also gives rise to regularity in verb forms [58]; and both the genetic drift-like process [59] and adaptation [60] may regulate vertical transmission of language from a generation to the next. It is needless to say that a complete understanding about why linguistic mutations arise, accumulate and give birth to different languages still lies in the future. Be that as it may, a Darwinian framework holds great promise for further elucidating the intertwined history of physical replicators like our genes and non-physical replicators like our languages.

We thank Tom Currie for helpful comments at the early stage of this work and Teresa Romero for constructive feedback on the manuscript. We also thank two anonymous reviewers for their enlightening comments and Simon Ho for providing useful information on technical issues. S.L. was supported by the University of Tokyo Fellowship.

REFERENCES

1 Diamond, J. & Bellwood, P. 2003 Farmers and their languages: the first expansions. *Science* **300**, 597–603. (doi:10.1126/science.1078208)

2 Campbell, L. 2003 What drives linguistic diversification and language spread? In *Examining the farming/language dispersal hypothesis* (eds P. Bellwood & C. Renfrew), pp. 49–63. Cambridge, UK: McDonald Institute for Archaeological Research.
3 Welsch, R., Terrell, J. & Nadolski, J. 1992 Language and culture on the north coast of New Guinea. *Am. Anthropol.* **94**, 568–600. (doi:10.1525/aa.1992.94.3.02a00030)
4 Bergsland, K. & Vogt, H. 1962 On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153. (doi:10.1086/200264)
5 Blust, R. 2000 Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics* (eds C. Renfrew, A. McMahon & L. Trask), pp. 311–332. Cambridge, UK: McDonald Institute for Archaeological Research.
6 Atkinson, Q. D. & Gray, R. D. 2005 Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* **54**, 513–526. (doi:10.1080/10635150590950317)
7 Pagel, M. 2009 Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10**, 405–415. (doi:10.1038/nrg2560)
8 Holden, C. J., Meade, A. & Pagel, M. 2005 Comparison of maximum parsimony and Bayesian Bantu language trees. In *The evolution of cultural diversity: a phylogenetic approach* (eds R. Mace, C. J. Holden & S. Shennan), pp. 53–65. London, UK: UCL Press.
9 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)
10 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)
11 Lewis, M. P. 2009 *Ethnologue: languages of the world.* Dallas, TX: SIL International.
12 Hudson, M. 1999 *Ruins of identity: ethnogenesis in the Japanese Islands.* Honolulu, HI: University of Hawaii'i Press.
13 Dodo, Y., Doi, N. & Kondo, O. 1998 Ainu and Ryukyuan cranial nonmetric variation: evidence which disputes the Ainu-Ryukyu common origin theory. *Anthropol. Sci.* **106**, 99–120.
14 Matsumura, H. 2001 Differentials of Yayoi immigration to Japan as derived from dental metrics. *HOMO J. Compar. Hum. Biol.* **52**, 135–156. (doi:10.1078/0018-442X-00025)
15 Hammer, M. F., Karafet, T. M., Park, H., Omoto, K., Harihara, S., Stoneking, M. & Horai, S. 2006 Dual origins of the Japanese: common ground for hunter–gatherer and farmer Y chromosomes. *J. Hum. Genet.* **51**, 47–58. (doi:10.1007/s10038-005-0322-0)
16 Suzuki, H. 1981 Racial history of the Japanese. *Rassengeschichte der Menschheit* **8**, 7–69.
17 Nei, M. 1995 The origins of human populations: Genetic, linguistic, and archeological data. In *The origin and past of modern humans as viewed from DNA* (eds S. Brenner & K. Hanihara), pp. 71–91. Singapore, Singapore: World Scientific Publishing.
18 Greenhill, S. J., Blust, R. & Gray, R. D. 2008 The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol. Bioinfo.* **4**, 271–283.
19 Embleton, S. M. 1986 *Statistics in historical linguistics.* Bochum, Germany: Brockmeyer.
20 Jōdaigo Jiten Henshū Iinkai. 1967 *Jidaibetsu kokugo daijiten. Jōdai hen.* Tokyo, Japan: Sanseidō.
21 Yasumoto, B. & Honda, M. 1978 *Nihongo no Tanjyō.* Tokyo, Japan: Taishūkan Shoten.
22 Muromachi Jidaigo Jiten Henshū Iinkai 2001 *Jidaibetsu kokugo daijiten: Muromachi jidai hen.* Tokyo, Japan: Sanseidō.
23 Hirayama, T. 1988 *Minami Ryukyū no Hōgenkisogoi.* Tokyo, Japan: Ohfusya.
24 Hirayama, T. 1992 *Gendai Nihongohōgen Daijiten.* Tokyo, Japan: Meiji Shoin.
25 Frellesvig, B. & Whitman, J. 2008 *Proto-Japanese: issues and prospects.* Amsterdam, The Netherlands: John Benjamins Publishing Co.
26 Crowley, T. & Bowern, C. 2009 *An introduction to historical linguistics.* Oxford, London: Oxford University Press.
27 Hattori, S. 1961 A glottochronological study on three Okinawan dialects. *Int. J. Am. Ling.* **27**, 52–62. (doi:10.1086/464603)
28 Hattori, S. 1978–1979 *Nihon sogo ni tsuite. Gekkan gengo*, vols 7–8.
29 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)
30 Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C. J. 2009 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. (doi:10.1098/rspb.2009.0408)
31 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)
32 Penny, D., McComish, B., Charleston, M. & Hendy, M. 2001 Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* **53**, 711–723. (doi:10.1007/s002390010258)
33 Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
34 Shapiro, B., Ho, S. Y. W., Drummond, A. J., Suchard, M. A., Pybus, O. G. & Rambaut, A. 2011 A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* **28**, 879–887. (doi:10.1093/molbev/msq262)
35 Clarke, H. 1989 The development of Edo language. In *18th century Japan: culture and society* (ed. C. A. Gerstle), pp. 63–72. Richmond, UK: Curzon Press.
36 Twine, N. 1988 Standardizing written Japanese. A factor in modernization. *Monum. Nippon* **43**, 429–454. (doi:10.2307/2384796)
37 Rambaut, A. & Drummond, A. J. 2007. Tracer v. 1. 5. See http://beast.bio.ed.ac.uk/Tracer.
38 Newton, M. A. & Raftery, A. E. 1994 Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. B* **56**, 3–48.
39 Suchard, M., Weiss, R., Dorman, K., Patel, M., McCabe, E. & Sinsheimer, J. 2000 Evolutionary similarity among genes when data are sparse. In *Proceedings of the Section on Bayesian Statistical Science*, pp. 92–97. Alexandria, VA: American Statistical Assocation.
40 Raftery, A. 1996 Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice* (eds W. R. Gilks, S. Richardson & D. J. Speigelhalter), pp. 163–187. New York, NY: Chapman & Hall.
41 Starostin, S. A., Dybo, A. V. & Mudrak, O. A. 2003 *Etymological dictionary of the Altaic languages.* Leiden, The Netherlands: Brill Academic Publishers.
42 Hock, H. H. 1986 *Principles of historical linguistics.* Berlin, Germany: Mouton de Gruyter.
43 Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. (doi:10.1093/molbev/msj030)

44 Gray, R. D., Bryant, D. & Greenhill, S. J. 2010 On the shape and fabric of human history. *Phil. Trans. R. Soc. B* **365**, 3923–3933. (doi:10.1098/rstb.2010.0162)
45 Garrett, A. 2006 Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 139–151. Cambridge, UK: McDonald Institute for Archaeological Research.
46 Koyama, S., Thomas, D. H. & Hakubutsukan, K. M. (eds) 1982 *Affluent foragers: Pacific coasts east and west.* Osaka, Japan: The National Museum of Ethnology.
47 Habu, J. 2004 *Ancient Jomon of Japan.* Cambridge, UK: Cambridge University Press.
48 Diamond, J. M. 1997 *Guns, germs, and steel: the fates of human societies.* New York, NY: W.W. Norton & Company.
49 Nelson, S. M. 1993 *The archaeology of Korea.* Cambridge, UK: Cambridge University Press.
50 Nakahashi, T. 2005 *Nihonjin no kigen.* Tokyo, Japan: Kodansha.
51 Benedict, P. K. 1990 *Japanese/Austro-Tai.* Ann Arbor, MI: Karoma.
52 The HUGO Pan-Asian SNP Consortium 2009 Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545. (doi:10.1126/science.1177074)
53 Greenhill, S. J., Currie, T. E. & Gray, R. D. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)
54 Haspelmath, M. & Tadmor, U. 2009 *Loanwords in the world's languages: a comparative handbook.* Berlin, Germany: Mouton de Gruyter.
55 Pagel, M. 2000 Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. In *Time depth in historical linguistics* (eds C. Renfrew, A. McMahon & L. Trask), pp. 189–207. Cambridge, UK: McDonald Institute for Archaeological Research.
56 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)
57 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)
58 Lieberman, E., Michel, J. B., Jackson, J., Tang, T. & Nowak, M. A. 2007 Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716. (doi:10.1038/nature06137)
59 Reali, F. & Griffiths, T. L. 2010 Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. B* **277**, 429–436. (doi:10.1098/rspb.2009.1513)
60 Kirby, S., Cornish, H. & Smith, K. 2008 Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl Acad. Sci. USA* **105**, 10 681–10 686. (doi:10.1073/pnas.0707835105)