Oceanic barriers promote language diversification in the Japanese Islands

S. LEE* & T. HASEGAWA†

*Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan
†Department of Cognitive and Behavioral Science, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan

Keywords:
allopatric diversification;
cultural FST;
geographic isolation;
language evolution.

Abstract

Good barriers make good languages. Scholars have long speculated that geographical barriers impede linguistic contact between speech communities and promote language diversification in a manner similar to the process of allopatric speciation. This hypothesis, however, has seldom been tested systematically and quantitatively. Here, we adopt methods from evolutionary biology and attempt to quantify the influence of oceanic barriers on the degree of lexical diversity in the Japanese Islands. Measuring the degree of beta diversity from basic vocabularies, we find that geographical proximity and, more importantly, isolation by surrounding ocean, independently explains a significant proportion of lexical variation across Japonic languages. Further analyses indicate that our results are neither a by-product of using a distance matrix derived from a Bayesian language phylogeny nor an epiphenomenon of accelerated evolutionary rates in languages spoken by small communities. Moreover, we find that the effect of oceanic barriers is reproducible with the Ainu languages, indicating that our analytic approach as well as the results can be generalized beyond Japonic language family. The findings we report here are the first quantitative evidence that physical barriers formed by ocean can influence language diversification and points to an intriguing common mechanism between linguistic and biological evolution.

Introduction

The Galápagos Islands, a cluster of extinct volcanoes in the Pacific Ocean, display a spectacular biodiversity that inspired the most important of all biological theories, Charles Darwin’s theory of evolution by natural selection (Darwin, 1882). Finches, iguanas and giant tortoises in these islands appeared unmistakably different not only from mainland South America but also from one island to the next. One hundred and fifty years later, we are beginning to understand that factors giving rise to the biodiversity in these islands are extremely complex (Grant & Grant, 2011), but we know that one simple and the most powerful factor that accounts for many aspects of this biodiversity is geographical isolation imposed by natural barriers among islands (Parent et al., 2008; Losos & Ricklefs, 2009).

The fruits of Darwin’s visit to the Galápagos Islands, including his historical insight that species evolve by a process of descent with modification, have benefited many scientific disciplines ever since (Dennett, 1996). In particular, an area that is flourishing with Darwinian thinking is the study of language change and variation (Atkinson & Gray, 2005; Croft, 2009; Pagel, 2009; Levinson & Gray, 2012); high-resolution phylogenies inferred from a selection of conservative lexicons shed light on the evolutionary history of their speakers (Gray et al., 2009; Bouckaert et al., 2012); words of newly diverged languages evolve in punctuational bursts, resembling DNA of species changing rapidly with speciation events (Pagel et al., 2006; Atkinson et al., 2008); and words that appear more frequently in everyday speech tend to be more conservative in a manner similar to proteins that have a larger impact on fitness tend to be more conservative (Hirsh & Fraser, 2001; Pagel et al., 2007). These parallels between linguistic and bio-

Correspondence: Sean Lee, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, 113-0033 Tokyo, Japan. Tel.: +81 3 5841 0465; fax: +81 3 5841 0824; e-mail: seanlee@darwin.c.u-tokyo.ac.jp

ª 2014 THE AUTHORS. J. EVOL. BIOL. 27 (2014) 1905–1912
JOURNAL OF EVOLUTIONARY BIOLOGY PUBLISHED BY JOHN WILEY & SONS LTD ON BEHALF OF EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY. 1905

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

doi: 10.1111/jeb.12442
logical evolution are striking, but in comparison to biological evolution, our understanding of why linguistic mutations arise, accumulate and give birth to different languages are far from complete.

In this study, we argue that the same factor responsible for much of the biodiversity in the Galápagos Islands is also responsible for the linguistic diversity in the Japanese Islands: the natural oceanic barriers that impede interaction between speech communities. The hypothesis that spatially isolated languages gradually diverge from one another due to reduced linguistic contact has been proposed on theoretical (Sereno, 1991) and anecdotal (Mufwene, 2008) grounds, but the lack of suitable methods and data meant that its validity could not be tested rigorously. A previous investigation on Micronesian languages reported a general trend that distant speech communities tend to speak different languages (Cavalli-Sforza & Wang, 1986), but because it lacked comparable language samples from nonislands, it was impossible to tease apart the influence of geographical isolation from a simple distance decay of linguistic similarity (Nekola & White, 1999; Nettle & Harris, 2003). Another study using more sophisticated methods (Gray et al., 2010) compared tree-likeness scores of Polynesian languages with those from Indo-European and found no evidence for the effect of geographical isolation. This result, however, was difficult to interpret because the Indo-European language family is almost three times older than the Polynesian, and thus the difference between their evolutionary patterns could potentially be attributed to the difference in their time depth.

The Japonic language family provides a convenient testing ground to investigate the influence of geographical isolation on language diversification for two reasons. First, Japonic languages are distributed across islands of different sizes that naturally allow them to be either separated or connected by geography, thereby forming two naturally comparable conditions (Fig. 1). Secondly, as all extant Japonic variants share a recent common ancestor (Lee & Hasegawa, 2011), the time of their origin is reasonably well controlled and it is thus possible for us to interpret the influence of geographical barriers in a relatively straightforward manner. Furthermore, a recent genomewide SNP analysis revealed the structure of Japanese population (Yamaguchi-Kabata et al., 2008) at a resolution high enough to be directly compared with linguistic structure, and as previous studies on cultural diversity have shown (Bell et al., 2009; Rzeszutek et al., 2012; Ross et al., 2013) such a comparison provides an invaluable opportunity to uncover the intertwined history of biological and linguistic evolution.

Materials and methods

For our analyses, we defined linguistic diversity as beta diversity (Anderson et al., 2006) of lexicons, which is expressed as dissimilarity among basic vocabularies of language variants for a given area, measured by patristic or Jaccard distances (Fig. 1; a full list of the variants is in the Fig. S1). Patristic distance is defined as the total branch length connecting two taxa on a tree, and our patristic distances were extracted from a Bayesian inference of Japonic language phylogeny (Lee & Hasegawa, 2011) using the R package ape (Paradis et al., 2004; R Core Team, 2013). The Jaccard distance quantifies the degree of dissimilarity between a pair of variants by estimating the number of dissimilar traits between them, normalized by the total number of their traits. Our Jaccard distances were calculated from binary states, indicating the presence (‘1’) or absence (‘0’) of a cognate (Crowley & Bowern, 2009) among Japonic variants (Lee & Hasegawa, 2011) using the R package vegan (Oksanen et al., 2013). The Jaccard distance is often considered as an appropriate measure of cultural diversity because it disregards shared absence of traits and normalizes the distance for each pair (Rogers & Ehrlich, 2008; Ross et al., 2013). We computed the Jaccard distances to address a potential criticism that our patristic distances are inappropriate estimates derived from an unreliable representation of Japonic language phylogeny.

We used binary states to indicate presence (‘1’) or absence (‘0’) of isolation by ocean between any two variants in a matching matrix. Our coding scheme is conservative in that presence of ocean separating any two variants was coded as 1 regardless of the physical distance between them. We recognize that the scheme
is insensitive to the possibility that many speech communities in the mainland Japan are also effectively isolated from one another by long distance or other geographic features such as rivers and mountains. We expect, however, that our conservative coding scheme based on barriers formed by oceanic barriers is likely to underestimate the effects of geographical isolation rather than overestimate them, because adding more parameters such as rivers and mountains are likely to increase the explanatory power of our statistical tests rather than the other way around. Pairwise geographical proximity among Japonic variants was obtained by calculating great circle distances from their geographic coordinates using GenAIEx v.6.5 (Peakall & Smouse, 2012). The geographical coordinates of Japonic variants were centroids of the locations from which the variants were sampled (Hirayama, 1988, 1992). All data are available in Supporting Information.

The extent of pairwise correlations between geographical proximity, isolation by surrounding water, and patristic/Jaccard distance matrices was determined using the Mantel and partial Mantel tests (Mantel, 1967; Smouse et al., 1986). The Mantel test calculates a correlation between two dissimilarity matrices, and the partial Mantel test calculates a partial correlation between two matrices while controlling for a third matrix. Because the elements of a distance matrix are not independent, statistical significance of the Mantel and partial Mantel tests is determined by permutation testing, and our estimates were obtained from 9999 permutations for each test (Oksanen et al., 2013). In addition to the standard Pearson product-moment correlation coefficient, we also estimated a rank correlation coefficient using Kendall’s tau to examine the robustness of results.

We used the NeighborNet algorithm (Huson & Bryant, 2006) to visualize the relationships among Japonic variants. For this, we used gene-content distances (Gray et al., 2010) and plotted split graphs while filtering out splits below a threshold of 0.001. We then estimated their tree-likeness with the delta (Holland et al., 2002) and Q-residual scores (Gray et al., 2010). Split graphs and their tree-likeness scores allow us to measure the extent of conflicting signal within a dataset, where conflicting signal indicates hybridization, horizontal transfer and convergent evolution.

In order to explore a relationship between genetic and linguistic structures, we used Arlequin v.3.5.1.3 (Excoffier & Lischer, 2010) to calculate \( \Phi_{ST} \) from patristic/Jaccard distance matrices (Supporting Information) and compared them with \( F_{ST} \) obtained from a genomewide SNP analysis of 7003 individuals (Yamaguchi-Kabata et al., 2008). Linguistic subpopulations were defined in the same scheme as the genetic subpopulations [i.e. Hokkaido, Tohoku, Kantō-Koshinetsu, Tokai-Hokuriku, Kinki, Kyushu, and Okinawa (Yamaguchi-Kabata et al., 2008)]. In general, \( \Phi_{ST} \) is considered slightly more informative than \( F_{ST} \) because \( \Phi_{ST} \) takes into account distance differences among variants. In essence, however, they are similar in that both measure the proportion of variation among subpopulations in relation to the total variation, and therefore, we compared \( \Phi_{ST} \) and \( F_{ST} \) directly. We interpreted any negative \( \Phi_{ST} \) value as zero and used permutation testing to assess statistical significance of the relationship between genetic and linguistic structures.

**Results**

Consistent with the hypothesis that geographical barriers promote language diversification in a manner similar to allopatric speciation, the results from simple Mantel tests indicate that pairs of Japonic variants that are separated by ocean tend to be more different from each other than those that are connected by land routes, for both Jaccard (Pearson’s correlation \( r = 0.58, P < 0.001; \) Kendall’s tau \( r_t = 0.49, P < 0.001 \)) and patristic distances \( (r = 0.51, P < 0.001; r_t = 0.42, P < 0.001) \). Also, pairs of variants that are geographically distant from each other tend to be more different than those that are close to each other, for both Jaccard \( (r = 0.78, P < 0.001; r_t = 0.55, P < 0.001) \) and patristic distances \( (r = 0.76, P < 0.001; r_t = 0.56, P < 0.001) \). In general, geographical proximity explains a larger amount of linguistic variability than isolation by surrounding water (Table 1) and this may be related to Honshu (i.e. the largest island) having a linguistic gradient across 1300 km of land without being separated by water. The partial Mantel tests show that the effect of oceanic barriers remains meaningful even after geographical proximity is factored out [Jaccard distances \( (r = 0.30, P < 0.001; r_t = 0.31, P < 0.001) \); patristic distances \( (r = 0.18, P = 0.013; r_t = 0.21, P = 0.002) \)]. We therefore consider that the effect of barriers formed by surrounding water is neither a by-product of geographical proximity nor a statistical artefact derived from an inappropriate language phylogeny (Table 1).

A potential problem with any correlational study is a hidden variable that is linked to the variables of interest (Roberts & Winters, 2013). We thus carried out further analyses to investigate if there is a confounding factor behind the effect of oceanic barriers. On closer inspection of the data, we observed that the majority of signal for the current result comes from small isolated islands (i.e. Hachijō, Amami, Okinawa, Hirara, Ikema, Irafu, Tarama, Taketomi, Ishigaki, Hateruma and Yonaguni; Fig. S1). Considering that smaller communities tend to have higher rates of language evolution as innovations and borrowings spread more easily than in larger communities (Nettle, 1999), one could argue that our results may be an epiphenomenon of accelerated evolutionary rates in small speech communities. It is difficult to directly test for the effect of population size on language diversification within our Mantel test frame-
work because (i) the exact number of speakers for each Japonic variant is unknown and (ii) creating a dissimilarity matrix of population size leads to loss of information about which variant has a larger or smaller population size. Therefore, we took a different approach by extracting mean evolutionary rates for all variants from Japonic language tree (Lee & Hasegawa, 2011) using TreeStat v.1.7.5 (Drummond et al., 2012) and tested if the languages from small islands have higher rates of evolution than the rest (Supporting Information). The Wilcoxon rank-sum test gave no evidence against the null hypothesis of identical distributions for their evolutionary rates ($W = 228$, $P = 0.70$; one-sided), suggesting that accelerated evolutionary rates associated with small speech communities may have little influence on our results.

Split graphs showing the results of NeighborNet analyses on mainland Japanese (left) and Ryukyuan (right). Gene-content distances were used and splits were filtered to a threshold of 0.001. Subgroups of Japonic languages are coded with colour circles: red-Hachijyo; orange-western Japanese; yellow-eastern Japanese; blue-Kyushu; purple-northern Ryukyuan; pink-southern Ryukyuan. Reticulations indicate presence of conflicting signal, which can be interpreted as hybridization, horizontal transfer and convergent evolution. Scale bar: 0.01.

Table 1 Results of the Mantel and partial Mantel tests.

| Measure of lexical diversity | Explained by | While controlling for | Tested by | $r^2$ | P-value |
|-----------------------------|--------------|----------------------|-----------|-------|---------|
| Jaccard Isolation           | Pearson      | 33.86                | <0.001    |
| Jaccard Isolation           | Kendall      | 23.82                | <0.001    |
| Patristic Isolation         | Pearson      | 26.25                | <0.001    |
| Patristic Isolation         | Kendall      | 17.37                | <0.001    |
| Jaccard Proximity           | Pearson      | 60.92                | <0.001    |
| Jaccard Proximity           | Kendall      | 30.56                | <0.001    |
| Patristic Proximity         | Pearson      | 57.90                | <0.001    |
| Patristic Proximity         | Kendall      | 31.46                | <0.001    |
| Jaccard Isolation Proximity | Pearson      | 9.13                 | <0.001    |
| Jaccard Isolation Proximity | Kendall      | 9.78                 | <0.001    |
| Patristic Isolation Proximity| Pearson | 3.35            | 0.013     |
| Patristic Isolation Proximity| Kendall | 4.50              | 0.002     |
| Jaccard Proximity Isolation | Pearson | 46.29              | <0.001    |
| Jaccard Proximity Isolation | Kendall | 17.76              | <0.001    |
| Patristic Proximity Isolation | Pearson | 44.82              | <0.001    |
| Patristic Proximity Isolation | Kendall | 20.78              | <0.001    |
connected by land routes on four large islands (excluding Hachijyo; see Fig. S2 for labels). Clearly, the split graph of mainland Japanese on the left side shows a stronger conflicting signal than that of Ryukyuan on the right. Furthermore, when we quantify the amount of conflicting signal for each group, mainland Japanese shows the average delta score of 0.39 and Q-residual score of 0.02, and Ryukyuan shows the delta score = 0.23 and Q-residual = 0.004. As smaller numbers indicate less conflicting signal, these estimates suggest that, in comparison to Ryukyuan, mainland Japanese carries a stronger signature of hybridization, horizontal transfer and convergent evolution. If we make a crude generalization that these two subgroups roughly represent the presence/absence of isolation by surrounding water, then since (i) Ryukyuan variants and mainland Japanese variants have similar time depth as all Japonic variants are descendants of a 2200-year-old common ancestor, (ii) there is no detectable difference in their evolutionary rates (W = 245, P = 0.84), and (iii) mainland Japanese variants seem to have experienced more intense linguistic contact than Ryukyuan variants, we can infer that the island geography as well as impediment of linguistic contact are highly likely to be the main factors driving linguistic diversity in the Japanese Islands.

In addition, we tested the extent to which our analytic approach is generalizable beyond Japonic language family by replicating the main analyses with the Ainu languages: the languages spoken by an indigenous group that once thrived throughout northern islands of Japan (Asai, 1974; Lee & Hasegawa, 2013). In strong agreement with the results obtained from Japonic language family, the partial Mantel tests show that oceanic barriers explain a significant proportion of the Ainu lexical diversity even after controlling for geographical proximity [Jaccard distances (r = 0.92, P < 0.001; r_1 = 0.57, P < 0.001); patristic distances (r = 0.93, P < 0.001; r_1 = 0.71, P < 0.001)]. A notable difference from the results of Japonic languages is that when controlling for the effect of oceanic barriers, the effect of geographical proximity remains meaningful only when we use Kendall’s tau [Jaccard distances (r_1 = 0.14, P = 0.001); patristic distances (r_1 = 0.13, P = 0.015)]. This seems to be related to Kuril (Fig. S3) being located far from the rest of Ainu variants without any linguistic gradient, because when Kuril is removed from the dataset and the effect of oceanic barriers is controlled for, the relationship between geographical proximity and the Ainu lexical diversity becomes easier to detect [Jaccard distances (r = 0.42, P < 0.001; r_1 = 0.21, P < 0.001); patristic distances (r = 0.41, P < 0.001; r_1 = 0.33, P < 0.001)]. All data used here are available in Supporting Information.

Comparing the structures between genetic and linguistic variation of Japonic speakers reveals that the patterns of their internal population differentiation are strongly correlated [patristic distance (r = 0.79, P = 0.03; r_1 = 0.52, P = 0.04); Jaccard distance (r = 0.75, P = 0.05; r_1 = 0.46, P = 0.05); simple mantel tests with 9999 permutations]. This implies that if genetic variation of a particular subgroup is highly differentiated from the rest, then linguistic variation of the same subgroup is also highly differentiated from the rest, or vice versa. Moreover, it seems unlikely that the similarity between the two structures is a consequence of sharing geographical proximity [patristic distance (r = 0.75, P = 0.03; r_1 = 0.48, P = 0.04); Jaccard distance (r = 0.66, P = 0.08; r_1 = 0.34, P = 0.08); partial mantel tests with 9999 permutations; note that ΦST matrix computed from the Jaccard distances fails to show significance at the 5% level, but because the data points are too small (7 × 7 matrix) to generate a proper null distribution and the P-values are reasonably low, we interpret these estimates to be generally meaningful; linguistic ΦST and physical distance matrices are available in Supporting Information]. Overall, these estimates seem to suggest that the evolution of both systems has experienced similar historical and ecological factors that are relevant to the Japanese Islands, and thus support the idea that human genes and languages often evolve by a shared process of descent with modification. Intriguingly, the range of pairwise linguistic ΦST values (0.0562–0.8903) is orders of magnitude higher than that of genetic ΦST values (0.0002–0.0035). Such a pattern has been argued to be a residual of cultural selection (Bell et al., 2009), and if correct, we can hypothesize that further clues to the forces driving language diversification in the Japanese Islands may be found in culture, rather than in genes, such as political dominance by regional speech communities (Hock, 1986; Renfrew, 1989) or social networks moderated by shared linguistic markers (Nettle & Dunbar, 1997; McElreath et al., 2003).

**Discussion**

Languages grow and diversify across different landscapes. Our preliminary results presented here suggest that geographical isolation imposed by oceanic barriers may have impeded hybridization and/or horizontal transfer among speech communities and promoted language diversification in the Japanese Islands. A series of tests shows that our results are unlikely to be a by-product of (i) using a Bayesian language language phylogeny, (ii) a simple distance decay of similarity and (iii) accelerated language evolution of small speech communities. Based on these observations, we argue that our current understanding of how linguistic diversity arises will be greatly improved if we take into account the same factor that led Darwin to his historical discovery, which is the geographical isolation among island populations (Darwin, 1882; Parent et al., 2008; Losos & Ricklefs, 2009).
At the same time, we acknowledge that the analogy breaks down when we consider that, unlike many species of the Galápagos Islands, the people of the Japanese Islands gradually developed advanced sailing skills which may have enabled them to freely migrate from one island to another in recent times (Hudson, 1999; Smits, 1999). Therefore, although we argue that oceanic barriers among the Japanese Islands played an important role in giving rise to linguistic diversity, we expect that there could also be other factors that helped maintaining the diversity until present.

We suggest that further clues to the process of language diversification in the Japanese Islands might be gained from the comparison between genetic and linguistic population structures. Our results indicate that (i) the degrees of pairwise population differentiation between the two structures are highly correlated, indicating that similar evolutionary forces have shaped both genetic and linguistic diversity, and (ii) linguistic FST values are on average much higher than the corresponding genetic FST values, suggesting that cultural factors may have had more influence on the development of population structure than genetic factors (Bell et al., 2009). If correct, we can formulate two different but related scenarios. The first scenario is a bottom-up process: once sufficient linguistic diversity arose to the point that speech communities could reliably distinguish one variant from another, linguistic dissimilarity may have been further amplified and maintained by being adopted as a marker for detecting as well as signalling one’s membership in reciprocal exchange network (Nettle et al., 2003) which led to the developments of stable social groups shaping genetic and linguistic diversity. Perhaps the use of social markers may have been easier in small communities than in sizeable communities (Boyd & Richerson, 1988) which coincides with our observation that the much of signal for our results comes from smaller islands.

The second scenario is a top-down process: after proto-Japonic speakers arrived in the Japanese Islands around 2500 years ago, they were divided into several small-scale competing groups (Lee & Hasegawa, 2011) and political unification for mainland Japanese was achieved only around 1200 years ago, followed by the unification for isolated islands of Ryukyu around 500 years ago (Hudson, 1999). Therefore, the correlated but linguistically more accentuated population structures could be reflecting the accumulated effects of historical boundaries imposed by regional hereditary clans (Hock, 1986; Renfrew, 1989), meaning that linguistic diversity might have been further amplified and maintained by political barriers that allowed linguistic contact exclusively among genetically close individuals living inside clan borders. The scenarios described here are speculative and should be subjected to further research, but they illustrate how evidence from different lines of enquiry can be synthesized to build a consistent model of human cultural diversity.

Our study makes a contribution to the current state of research on language evolution by demonstrating that there is an alternative way of measuring linguistic diversity, which is beta diversity of lexicons. Previous studies have placed disproportionate emphasis on Greenberg’s diversity index (i.e. the probability of two randomly chosen speakers sharing the same language) or language density over a given area or per population (Gavin et al., 2013). While these are scientifically sound methods, they could potentially suffer from problematic nature of how languages are defined [see Nettle (1998) for conceptual review]. We argue that if (i) language variants are sampled evenly across a region, and (ii) there is a sufficient amount of variation among them, then measuring beta diversity could serve as an excellent complementary strategy for revealing the external factors that shape language diversity (Koleff et al., 2003; Nettle et al., 2007). Further study will determine the precise extent to which these ideas can be utilized beyond the Ainu and Japonic languages.

A clear limitation to our study is the lack of more ecologically sensitive measure for detecting geographical barriers. We focused on separation by ocean as the sole measure for geographical isolation, but it is obvious that numerous mountains of Honshu could have been significant barriers preventing some speech communities from interacting with one another. A previous study that examined the frequencies of 15 genetic markers in Japanese population reported that some of the montane regions of Honshu may have indeed contributed to rapid genetic change (Sokal & Thomson, 1998). Although we did not incorporate this information into our analyses because the identified montane regions were incompatible with the resolution we required, the search for other plausible ecological barriers is a direction that deserves more attention. Additionally, our coding scheme may also be improved by assigning probabilistic weights to different ocean barriers based on seasonal wind change or the direction of water circulation (Moon et al., 2009; Jin et al., 2010) as they would have determined the difficulty of sea travel.

Our findings are mainly correlational and therefore preclude causal interpretation. While we recognize that interpretations from correlations should be made carefully, we believe that the methods and data used in this study are ideally suited for the phenomenon of interest, and our approach opens the possibility for further characterization of this fascinating phenomenon. We still have a long way to go to fully understand the dynamics of language diversification, but the results presented here demonstrate how relatively simple procedures can start revealing linguistic consequences of geographical isolation, and illustrate an intriguing common mechanism between linguistic and biological evolution.
Acknowledgments

We are grateful to the two reviewers for their constructive and insightful comments. This work was supported by a Grant-in-Aid for Japan Society for the Promotion of Science (248973/2604070) awarded to SL.

References

Anderson, M.J., Ellingsen, K.E. & McArdis, B.H. 2006. Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.* **9**: 683–693.

Asai, T. 1974. *Classification of Dialects: Cluster Analysis of Aine Dialects*. Bulletin of Institute for the Study of Northern Eurasian Cultures, Vol. 8, pp. 45–136. Faculty of Letters, Hokkaido University, Sapporo.

Atkinson, Q.D. & Gray, R.D. 2005. Curious parallels and curios connections—phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* **54**: 513–526.

Atkinson, Q.D., Meade, A., Venditti, C., Greenhill, S.J. & Pagel, M. 2008. Languages evolve in punctuational bursts. *Science* **319**: 588.

Bell, A.V., Richerson, P.J. & McElreath, R. 2009. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proc. Natl. Acad. Sci. USA* **106**: 17671–17674.

Bouckaert, R., Lemy, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J. et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* **337**: 957–960.

Boyd, R. & Richerson, P.J. 1988. The evolution of reciprocity in sizable groups. *J. Theor. Biol.* **132**: 337–356.

Cavalli-Sforza, L.L. & Wang, W.S.Y. 1986. Spatial distance and lexical replacement. *Language* **62**: 38–55.

Croft, W. 2009. Evolutionary linguistics. *Annu. Rev. Anthropol.* **37**: 219–234.

Crowley, T. & Bowern, C. 2009. *An Introduction to Historical Linguistics*. Oxford University Press, London.

Darwin, C. 1882. *On the Origin of Species by Means of Natural Selection*. Murray, London.

Dennett, D.C. 1996. *Darwin’s Dangerous Idea*. Simon and Schuster, New York.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**: 1969–1973.

Excoffer, L. & Lischer, H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**: 564–567.

Gavin, M.C., Botero, C.A., Bowern, C., Colwell, R.K., Dunn, M., Dunn, R.R. et al. 2013. Toward a mechanistic understanding of linguistic diversity. *Bioscience* **63**: 524–535.

Grant, P.R. & Grant, B.R. 2011. *How and Why Species Multiply*. Princeton University Press, Princeton.

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

Gray, R.D., Bryant, D. & Greenhill, S.J. 2010. On the shape and fabric of human history. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**: 3923–3933.

Hirayama, T. 1988. *Minami Ryukyu no Hogenkisogoi*. Ohfusya, Tokyo.

Hirayama, T. 1992. *Gendai Nihongo-hogen Daijitoten*. Miji Shoin, Tokyo.

Hirsh, A.E. & Fraser, H.B. 2001. Protein dispensability and rate of evolution. *Nature* **411**: 1046–1049.

Hock, H.H. 1986. *Principles of Historical Linguistics*. Mouton de Gruyter, Berlin.

Holland, B.R., Huber, K.T., Dress, A. & Moulton, V. 2002. δ overlays: a tool for analyzing phylogenetic distance data. *Mol. Biol. Evol.* **19**: 2051–2059.

Hudson, M. 1999. *Ruins of Identity: Ethnicity in the Japanese Islands*. University of Hawai’i Press, Honolulu.

Huson, D.H. & Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**: 254–267.

Jin, B., Wang, G., Liu, Y. & Zhang, R. 2010. Interaction between the East China Sea Kuroshio and the Ryukyu Current as revealed by the self-organizing map. *J. Geophys. Res.* **115**: C12047.

Koleff, P., Gaston, K.J. & Lennon, J.J. 2003. Measuring beta diversity for presence–absence data. *J. Anim. Evol.* **72**: 367–382.

Lee, S. & Hasegawa, T. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proc. Biol. Sci.* **278**: 3662–3669.

Lee, S. & Hasegawa, T. 2013. Evolution of the Aino language in space and time. *PLoS One* **8**: e62243.

Levinson, S.C. & Gray, R.D. 2012. Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn. Sci.* **16**: 167–173.

Losos, J.B. & Ricklefs, R.E. 2009. Adaptation and diversification on islands. *Nature* **457**: 830–836.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**: 209–220.

McElreath, R., Boyd, R. & Richerson, P.J. 2003. Shared norms and the evolution of ethnic markers. *Curr. Anthropol.* **44**: 122–130.

Moon, J.H., Hirose, N., Yoon, J.H. & Pang, I.C. 2009. Effect of the along-strait wind on the volume transport through the Tsushima/Korea Strait in September. *J. Oceanogr.* **65**: 17–29.

Mufwene, S.S. 2008. *Language Evolution: Contact, Competition and Change*. Continuum International Publishing Group, London.

Nekola, J.C. & White, P.S. 1999. The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* **26**: 867–878.

Nettle, D. 1998. Explaining global patterns of language diversity. *J. Anthropol. Archaeol.* **17**: 354–374.

Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* **108**: 119–136.

Nettle, D. & Dunbar, R.I.M. 1997. Social markers and the evolution of reciprocal exchange. *Curr. Anthropol.* **38**: 93–99.

Nettle, D. & Harriss, L. 2003. Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Hum. Biol.* **75**: 331–344.

Nettle, D., Grace, J.F., Choisy, M., Cornell, H.V., Guégan, J.F. & Hochberg, M.E. 2007. Cultural diversity, economic development and societal instability. *PLoS One* **2**: e929.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O’Hara, R.B., Simpson, G.L. et al. 2013. Vegan: community ecology package. 2013. R package version 2.0-7.

Pagel, M. 2009. Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10**: 405–415.

Pagel, M., Venditti, C. & Meade, A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**: 119–121.
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.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Parent, C.E., Caccone, A. & Petren, K. 2008. Colonization and diversification of Galapagos terrestrial fauna: a phylogenetic and biogeographical synthesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**: 3347–3361.

Peakall, R. & Smouse, P.E. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**: 2537–2539.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renfrew, C. 1989. Models of change in language and archaeology. *Trans. Philol. Soc.* **87**: 103–155.

Roberts, S. & Winters, J. 2013. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS One* **8**: e70902.

Rogers, D.S. & Ehrlich, P.R. 2008. Natural selection and cultural rates of change. *Proc. Natl. Acad. Sci.* **105**: 3416–3420.

Ross, R.M., Greenhill, S.J. & Atkinson, Q.D. 2013. Population structure and cultural geography of a folktale in Europe. *Proc. Biol. Sci.* **280**: 20123065.

Rzeszutek, T., Savage, P.E. & Brown, S. 2012. The structure of cross-cultural musical diversity. *Proc. Biol. Sci.* **279**: 1606–1612.

Sereno, M.I. 1991. Four analogies between biological and cultural/linguistic evolution. *J. Theor. Biol.* **151**: 467–507.

Smits, G. 1999. *Visions of Ryukyu: Identity and Ideology in Early-Modern Thought and Politics*. University of Hawai‘i Press, Honolulu.

Smouse, P.E., Long, J.C. & Sokal, R.R. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* **35**: 627–632.

Sokal, R.R. & Thomson, B.A. 1998. Spatial genetic structure of human populations in Japan. *Hum. Biol.* **70**: 1–22.

Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M. *et al.* 2008. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83**: 445–456.

**Supporting information**

Additional Supporting Information may be found in the online version of this article:

**Figure S1** Map of 57 Japonic languages. Subgroups of Japonic languages are coded with colour circles: yellow-eastern Japanese; orange-western Japanese; red-Hachijyo; blue-Kyushu; purple-northern Ryukyuan; pink-southern Ryukyuan.

**Figure S2** Labelled split graphs showing the results of NeighborNet analyses on mainland Japanese (left) and Ryukyuan (right).

**Figure S3** Map of 21 Ainu languages.

**Figure S4** Bayesian phylogeny of 21 Ainu languages.

**Data S1** Supporting data.

Received 1 April 2014; accepted 30 May 2014