Current Biology

Report

Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population

Highlights

- Genetic and linguistic variation are compared for the population of Cape Verde
- Genetic ancestry traces to a combination of Iberian and Senegambian sources
- Measures of African genetic and linguistic admixture are positively correlated
- Patterns accord with vertical cotransmission of genetic and linguistic variation

Authors

Paul Verdu, Ethan M. Jewett, Trevor J. Pemberton, Noah A. Rosenberg, Marlyse Baptista

Correspondence

noahr@stanford.edu

In Brief

Verdu et al. study genetic and linguistic variation among individuals in the genetically admixed creole-speaking population of Cape Verde. They find that genetic and linguistic admixture are correlated, suggesting that genetic and linguistic differences have been vertically cotransmitted from parent to offspring within the population.

Verdu et al., 2017, Current Biology *27*, 2529–2535
August 21, 2017 © 2017 Elsevier Ltd.
http://dx.doi.org/10.1016/j.cub.2017.07.002

CellPress

CellPress

Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population

Paul Verdu,[1,6] Ethan M. Jewett,[2,3,6] Trevor J. Pemberton,[4] Noah A. Rosenberg,[2,7,8,*] and Marlyse Baptista[5,7]
[1]CNRS-MNHN-Université Paris Diderot, UMR7206 Eco-Anthropology and Ethno-Biology, Paris, France
[2]Department of Biology, Stanford University, Stanford, CA 94305, USA
[3]Department of Statistics and Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA
[4]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB R3E0J9, Canada
[5]Department of Linguistics and Department of Afroamerican and African Studies, University of Michigan, Ann Arbor, MI 48109, USA
[6]These authors contributed equally
[7]Senior author
[8]Lead Contact
*Correspondence: noahr@stanford.edu

http://dx.doi.org/10.1016/j.cub.2017.07.002

SUMMARY

Joint analyses of genes and languages, both of which are transmitted in populations by descent with modification—genes vertically by Mendel's laws, language via combinations of vertical, oblique, and horizontal processes [1–4]—provide an informative approach for human evolutionary studies [5–10]. Although gene–language analyses have employed extensive data on individual genetic variation [11–23], their linguistic data have not considered corresponding long-recognized [24] variability in individual speech patterns, or idiolects. Genetically admixed populations that speak creole languages show high genetic and idiolectal variation—genetic variation owing to heterogeneity in ancestry within admixed groups [25, 26] and idiolectal variation owing to recent language formation from differentiated sources [27–31]. To examine cotransmission of genetic and linguistic variation within populations, we collected genetic markers and speech recordings in the admixed creole-speaking population of Cape Verde, whose Kriolu language traces to West African languages and Portuguese [29, 32–35] and whose genetic ancestry has individual variation in European and continental African contributions [36–39]. In parallel with the combined Portuguese and West African origin of Kriolu, we find that genetic admixture in Cape Verde varies on an axis separating Iberian and Senegambian populations. We observe, analogously to vertical genetic transmission, transmission of idiolect from parents to offspring, as idiolect is predicted by parental birthplace, even after controlling for shared parent–child birthplaces. Further, African genetic admixture correlates with an index tabulating idiolectal features with likely African origins. These results suggest that Cape Verdean genetic and linguistic admixture have followed parallel evolutionary trajectories, with cotransmission of genetic and linguistic variation.

RESULTS

We collected genome-wide genotypes and speech samples in the same set of 44 unrelated individuals from Cape Verde (STAR Methods). Compiling the genotypes with similar data from continental African and European groups (Figures S1 and S2), we measured genetic admixture, assigning a genomic proportion for each individual to European ancestry and a proportion to African ancestry. In parallel, from the linguistic data, we computed an African linguistic score that measures the degree to which each individual used a set of 212 lexical items (henceforth "words") that correspond to instances of 29 underlying high-confidence African-origin roots and morphemes. We also tabulated the frequency of each word in the speech of each individual (Figure S3).

Worldwide Context of Genetic Variation in Cape Verde

The Cape Verdean population represents one of the earliest examples of the formation of a modern admixed creole group, originating in the 15th century from contact among Portuguese and West African populations in a previously uninhabited archipelago off the west coast of Africa (Figure S1). To examine the genetic history of the Cape Verdean population in relation to the linguistic history of Kriolu, we considered genetic variation in the Cape Verdean sample together with populations of Western European and continental African origin.

Figure 1A presents a multidimensional scaling (MDS) analysis showing the Cape Verdean population alongside two other populations descended from the transatlantic slave trade. Of 44 Cape Verdeans, 42 lie on a path between Europeans and West Africans represented by the Gambian Mandinka and Senegalese Mandenka, closely related populations that speak different languages from the same Mandé linguistic family (Mandinka ISO

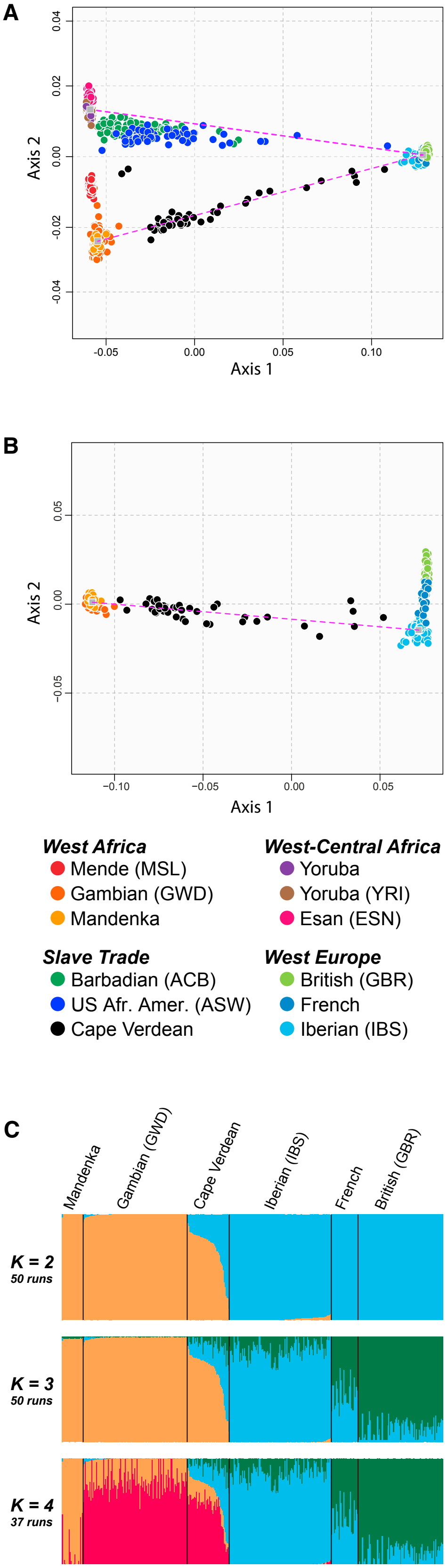

Figure 1. Genetic Variation in African–European Admixed Populations in Relation to Continental African and European Populations

Each point in (A) and (B) represents the projection of an individual in a two-dimensional multidimensional scaling (MDS) analysis of individual-pairwise allele-sharing dissimilarities.

(A) MDS analysis of six continental African populations from West and West-Central Africa, three Western European populations, and three admixed populations that arose during the transatlantic slave trade. The dotted lines connect the centroid of the Europeans with the West-Central African centroid and the centroid of Gambian Mandinka and Senegalese Mandenka. Each centroid is marked with a gray square.

(B) MDS analysis of two West African populations (Gambian Mandinka and Senegalese Mandenka), three European populations, and the Cape Verdean population. The dotted line connects the West African centroid with the Iberian centroid.

(C) Admixture analysis of Cape Verdean individuals together with West African and Western European samples. Each of 399 individuals is represented as a vertical line partitioned into $K = 2$ to $K = 4$ components that correspond to estimated membership in each of the 2 to 4 clusters inferred by Admixture. For each value of K, we ran Admixture 50 times independently, using Clumpp to average replicates with similar clustering. The number of runs averaged to create each plot appears on the left for each value of K.

See Figures S1 and S2 for information about data preparation and Figure S4 for a version of (C) with smaller sample sizes for the populations with large samples.

639-3 [mnk] and Western Maninkakan ISO 639-3 [mlq], respectively; 40). The Gambian Mandinka and Senegalese Mandenka are genetically closer to Cape Verdeans than are the Sierra Leone Mende, who also speak a Mandé language (40, Mende ISO 639-3 [men]). By contrast, Figure 1A shows that West-Central African populations from Nigeria (Esan and two Yoruba samples) represent a more suitable proxy for the continental African origin of the US African American and African Caribbean Barbadian populations.

Excluding African populations other than the Gambian Mandinka and Senegalese Mandenka, the MDS plot in Figure 1B shows that for the European ancestry component, most Cape Verdeans lie on a path connecting Senegambian populations to an Iberian sample, rather than to British or French populations. That Iberians represent the most suitable proxy for the European ancestry of Cape Verdeans is compatible with records of the Portuguese slave trade in Senegambia starting early in the history of Cape Verde [41–43] and accords with an understanding of Cape Verdean Kriolu as a hybrid of Portuguese with languages from the Senegambian region [29, 32]. The genetic relationships are consistent with both the peopling history of the archipelago and with the relationship of Kriolu to other languages.

The MDS results of Figures 1A and 1B are supported and refined in Admixture clustering analyses. Plots with $K = 2$ and $K = 3$ clusters (Figure 1C) suggest that the Cape Verdean sample derives from genetic admixture between African and European populations for which Senegambians and Iberians are close proxies. Admixture at $K = 4$ finds a new cluster that corresponds primarily to Gambian Mandinka individuals and that is distinct from the cluster seen in the Senegalese Mandenka. The clustering component of Cape Verdean individuals at $K = 4$ that reflects similarity to West Africans more closely resembles the patterns exhibited by the Gambian Mandinka.

Genetic and Linguistic Admixture

Analyses of genetic and linguistic admixture can evaluate whether current patterns of genetic and linguistic variation are

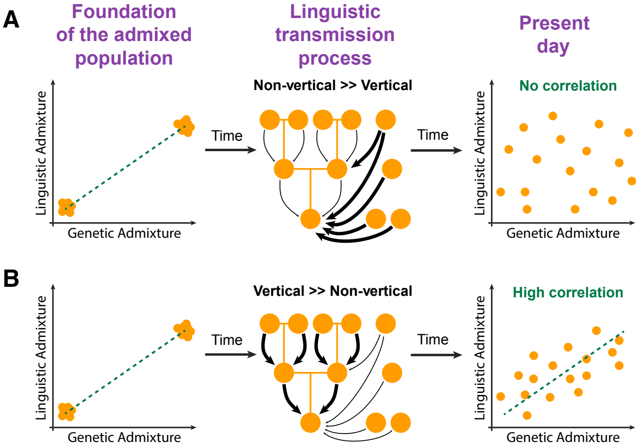

Figure 2. Two Models Predicting the Relationship between Linguistic and Genetic Ancestry

(A) Primarily oblique and horizontal linguistic transmission. Individuals within an admixed population are represented by colored dots. Left: at the foundation of the admixed population, distinct groups of individuals represent the two source populations. Middle: black arrows indicate linguistic transmission among related or unrelated individuals, with stronger weight on oblique transmission from adults to children who are not their own offspring, and horizontal transmission among peers. Right: oblique and horizontal linguistic transmission obscure the signal of vertical transmission (non-vertical $\gg$ vertical), resulting in no correlation between linguistic and genetic admixture among individuals in the admixed population.

(B) Primarily vertical transmission. From the same founding event, largely vertical linguistic transmission (vertical $\gg$ non-vertical) induces a correlation between linguistic and genetic admixture in the admixed population.

consistent with vertical processes of language transmission in a manner analogous to Mendelian genetic inheritance, or instead with predominantly non-vertical linguistic transmission. Figure 2 schematically illustrates the patterns of correlation between genetic and linguistic admixture that might be expected under different mechanisms by which language is transmitted over time among individuals within an admixed creole population. Under vertical transmission of idiolects, the mixture of Portuguese and West African components in the idiolect of a Kriolu speaker is expected to correlate with genetic admixture (Figure 2B), whereas under horizontal or oblique transmission, such a correlation is not expected (Figure 2A). To assess the extent to which an idiolect represents a population mixture similar to the genetic mixture, we compared the genomic fraction inherited from African sources to an African linguistic admixture score (STAR Methods).

African genetic admixture proportions and African linguistic admixture scores for our 212-word set have a significant correlation among Kriolu speakers (Figure 3; Spearman ρ = 0.45, p = 0.0023). Many of the individuals with the lowest African genetic ancestries have among the lowest African linguistic admixture scores, and African linguistic admixture scores are greatest for many individuals with high African genetic ancestries.

To determine whether frequencies of words of African origin are more strongly correlated with African genetic ancestry than are frequencies of randomly chosen words, we sampled 10,000 sets of 212 words chosen uniformly at random without replacement from the 2,474 distinct words appearing in our 44

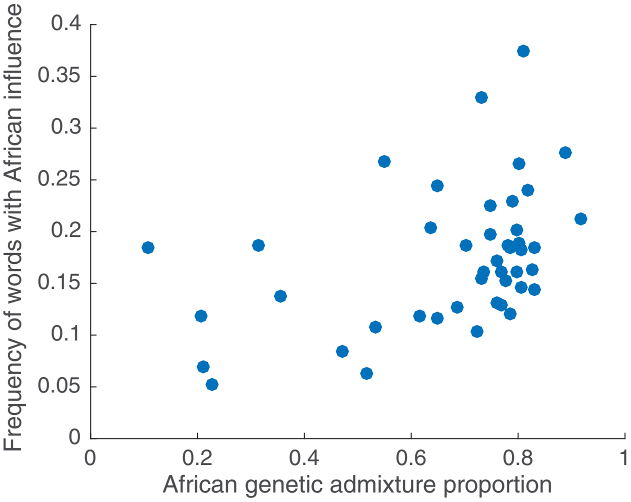

Figure 3. Linguistic and Genetic Admixture in Cape Verde

The 44 points represent African linguistic and genetic admixture scores for individuals sampled in Cape Verde. See Figure S3 and Table S1 for information on the construction of the linguistic admixture score.

speech samples, computing an African linguistic admixture score for each resampled set. The fraction of resampled sets with a correlation coefficient as high as that of the true set of 212 African words was 0.0168 (mean ρ across resamples = -0.005, SD = 0.236). This small value supports the view that the frequency of use of words of African origin is positively correlated with African genetic ancestry.

To further validate the observed correlation between genetic and linguistic admixture scores, we restricted our analysis to a set of 51 words corresponding to instances of 9 distinct roots and morphemes for which confidence in an African origin was particularly high (Table S1). African linguistic admixture scores based on these words had a comparable correlation with African genetic admixture proportions to that seen with the full set of 212 words (ρ = 0.46, p = 0.0017). In 10,000 resampled sets of 51 words chosen uniformly at random without replacement from the full 2,474-word set, the fraction with a correlation coefficient at least as high as that of the 51-word African-origin set was 0.0099 (mean ρ across resamples = -0.027, SD = 0.230). The fact that correlations between genetic and linguistic admixture persist and even slightly strengthen when we restrict attention to the highest-confidence African-origin words provides further evidence that the correlation is driven by shared genetic and linguistic admixture histories.

Intergenerational Transmission of Idiolect

We next investigated further whether the observed correlation between African genetic and linguistic admixture scores could have arisen from vertical transmission of linguistic characters. To examine the transmission of word frequency patterns from parents to offspring, we considered whether individual speech patterns reflected an influence of parental birthplace. In particular, we asked whether parental birthplaces predicted idiolect by performing Mantel correlations between a matrix of pairwise distances between parental birthplaces, averaging maternal and paternal birthplace distances for each individual, and a

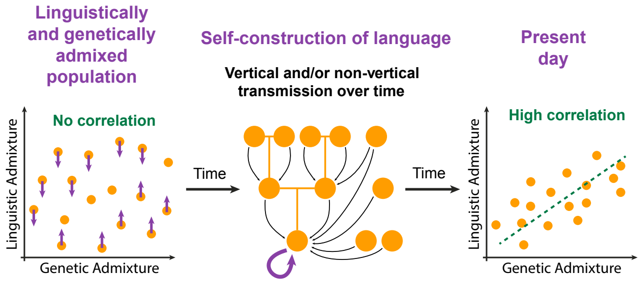
Linguistically and genetically admixed population
Self-construction of language
Present day
Vertical and/or non-vertical transmission over time
No correlation
High correlation
Linguistic Admixture
Genetic Admixture
Time
Time
Linguistic Admixture
Genetic Admixture

Figure 4. Predicting the Relationship between Linguistic and Genetic Ancestry under Social Construction of Linguistic Identity
Individuals within an admixed population are represented by colored dots. Left: when linguistic and genetic admixture processes are not analogous, a population might have no initial correlation between linguistic and genetic admixture. Middle: individuals adopt speech patterns that reflect their identity on a socially constructed spectrum reflecting genetic admixture, adjusting their linguistic admixture upward or downward (arrows in the left panel). For example, individuals who self-identify as more European might develop idiolects that reflect Portuguese influence. Right: social construction of linguistic identity induces a correlation between linguistic and genetic admixture proportions. Correlation between linguistic and genetic admixture can be enhanced at each generation, whether or not vertical transmission occurs.

pairwise distance matrix between individual 2,474-word frequency profiles (STAR Methods).

Consistent with vertical inheritance of word frequency patterns from parents to offspring, parental birthplaces predicted idiolectal variation in their offspring. Within the subset of 34 individuals born in Cape Verde, both of whose parents were also born in Cape Verde, we observed a Mantel correlation of r = 0.47 (p = 0.0006) between pairwise linguistic distances and distance between their parental birthplaces. Note that we would also expect to observe this correlation without vertical inheritance if offspring birthplace correlates with parental birthplace, as idiolect in offspring could simply follow the location-specific dialect of their parents. To test this possibility, we accounted for offspring birthplace by computing a partial Mantel correlation between pairwise linguistic distances and parental birthplace distances, including pairwise distances between individual birthplaces as a covariate. This partial correlation remained significant (r = 0.32, p = 0.0096); thus, vertical transmission of word frequency patterns, rather than shared parent–offspring birthplaces, more likely explains the correlation between offspring linguistic distances and parental birthplace distances.

DISCUSSION

We have examined genetic and linguistic variation in the same set of sampled individuals from the Cape Verde population, a population whose genotypes and language result from an admixture process involving Iberian and West African contributions. We found that the genetic ancestry of Cape Verdeans, unlike that of African Americans and Barbadians, is compatible with contributions from sources best represented here by Iberian and Gambian Mandinka populations (Figure 1); this result accords with the structure of Cape Verdean Kriolu as a Portuguese creole with contributions from Senegambian languages and, in particular, Mandinka. Genetic ancestry is variable among individuals from Cape Verde, and individual genetic admixture from African sources correlates with an index measuring African contributions to individual idiolects (Figure 3). This correlation provides evidence for a cotransmission process of genetic and idiolectal variants.

Our genetic analysis, placing the Cape Verdean population on a trajectory distinct from that of other populations that have European and African ancestry and that are descended from the transatlantic slave trade, highlights the diversity present in populations resulting from intercontinental admixture. The transatlantic slave trade involved both great diversity in the African source populations and systematic pairings of source and destination locations for African slaves [43]; the distinctiveness of the genetic relationships of Cape Verdeans in relation to those of other African–European admixed populations preserves some of the initial heterogeneity in the sources of the African ancestry component. That the genetic signal accords with the character of Cape Verdean Kriolu as a mixture of Portuguese and Senegambian languages supports the use of inferences of genetic ancestry in creole populations as a means of generating hypotheses about linguistic ancestry and provides evidence supporting Mandinka contributions to the Kriolu language. The fact that Gambian Mandinka are genetically more similar to the Cape Verdean sample than are Senegalese Mandenka is of interest for the study of fine-scale origins of Cape Verdean Kriolu. Their languages are from the same language family, unlike more linguistically distant source languages such as Wolof; in assessments of Kriolu origins that take genetics into account, it will be of interest to also consider genetic data on Wolof-speaking populations.

The correlation between genetic and linguistic admixture in Cape Verde might be regarded as unexpected given the multifarious vertical and horizontal processes involved in linguistic transmission (Figure 2), and the complexity of the processes through which languages interact when coming into contact during the emergence of a creole language. Linguistic admixture is affected by complex sociocultural and linguistic pressures [27–32, 44], and creole languages are typically not simple mixtures of source languages [29, 45–48]. Nevertheless, our results support a view in which vertical transmission of linguistic variants in a manner that parallels genetic transmission plays a role in both the development and the subsequent transmission of creole languages [49–51].

Cavalli-Sforza et al. [6, 7] famously superposed genetic trees of population relationships and trees of relationships between languages, finding a striking concordance that suggested cotransmission of genes and languages in the divergence of populations. Our support for cotransmission *within* populations represents the ‘‘microevolutionary’’ analog to the ‘‘macroevolutionary’’ process of genetic and linguistic divergence of whole populations and languages examined in past gene–language studies. It is plausible that simultaneous divergence of genes and languages at the macroevolutionary level of populations and languages occurs by cumulative effects of similar processes occurring at the microevolutionary level of individuals and their distinctive idiolects.

We note that other possible mechanisms of linguistic transmission can be envisioned to explain the gene–language correlations identified here. Figure 4 shows a process in which individuals construct their own way of speaking Kriolu in relation to their genetic admixture composition. Such a process, in which individuals align their speech with constructs of identity arising through genetic admixture, can also produce gene–language correlation.

Alternatively, if a population contains distinct social communities across which an initial correlation between linguistic and genetic admixture exists, the correlation could be maintained over time if transmission of idiolect occurs primarily from the social community of an individual rather than by vertical transmission from parents. Such a process could involve social communities at many levels, including, in Cape Verde, island-level communities, which do show differences in genetic admixture [38]. Social communities of comparable education level, perhaps spanning multiple islands, are also possible; although linguistic distance continued to be correlated with parental birthplace distance after controlling for shared parent–offspring birthplaces, so that offspring idiolect more closely reflected parental idiolect beyond that expected by location, it remains possible that offspring idiolect is transmitted through a non-location-specific social community to which parents belong. Note that in transmission of idiolect via largely hereditary social groups, it is membership in social communities rather than idiolect itself that is vertically transmitted.

Although the high variability in idiolect in Cape Verdean Kriolu makes it well-suited for an analysis of idiolectal variation, our sample is limited in its coverage of Cape Verde. The archipelago has substantial genetic and linguistic variation unlikely to be represented here, including historical distinctions between the two main island groups, the more northerly Barlavento and the more southerly Sotavento (Figure S1). Further analyses considering additional social variables and inter-island differences are required to more definitively resolve the source of the correlation between genetic and linguistic admixture. Nevertheless, our joint analysis of inter-individual genetic and linguistic variation and its contribution to understanding the admixed population of Cape Verde provide an informative new approach for the investigation of genes and languages in human evolution.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Genotyping
  - ○ Genotype Quality Control
  - ○ Relatedness
  - ○ Genetic Data from Worldwide Populations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Allele-Sharing Dissimilarities
  - ○ Multidimensional Scaling
  - ○ Admixture
  - ○ Genetic Admixture Estimates
  - ○ Linguistic Interviews
  - ○ Word Frequency Profiles
  - ○ Identification of Words of African Origin
  - ○ Linguistic Admixture Estimates
  - ○ Geographic Distances
  - ○ Linguistic Distances and Parental Birthplaces
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.07.002.

AUTHOR CONTRIBUTIONS

Conceptualization, M.B., N.A.R., and P.V.; Project Design, P.V., E.M.J., N.A.R., and M.B.; Sampling and Data Collection, M.B. and P.V.; Linguistic Data Preparation, M.B. and E.M.J.; Genetic Data Preparation, P.V. and T.J.P.; Linguistic Data Analysis, E.M.J. and M.B.; Genetic Data Analysis, P.V.; Supervision, N.A.R. and M.B.; Writing, all authors.

ACKNOWLEDGMENTS

We warmly thank all Cape Verdean participants in this study. Support was provided by the France–Stanford Center for Interdisciplinary Studies, the French National Research Agency grant METHIS (ANR-15-CE32-0009-01), and the Stanford Center for Population Research. We thank Angelo Barbosa of the University of Cape Verde for logistical support; Manuel Brito-Semedo and Adalicia Magalhães for help with recruitment; Emanuel de Pina for help conducting interviews; and Diamantino Tavares Freire, Elisangela Spencer Coelho, Emanuel de Pina, Hulda Lopes Costa, Karina Moreira, Ligia Herbert Duarte Lopes, Maria do Céu Santos Baptista, Raimundo Lopes, and Sérgio Costa for transcription assistance. We thank Valentin Thouzeau and Pierre Darlu for useful comments.

Received: April 8, 2017
Revised: May 17, 2017
Accepted: July 3, 2017
Published: August 10, 2017

REFERENCES

1. Cavalli-Sforza, L.L., and Feldman, M.W. (1981). Cultural Transmission and Evolution: A Quantitative Approach (Princeton University Press).
2. Smith, J., Durham, M., and Fortune, L. (2007). "Mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect. Lang. Var. Change *19*, 63–99.
3. Stanford, J.N. (2008). Child dialect acquisition: new perspectives on parent/peer influence. J. Sociolinguist. *12*, 567–596.
4. Schleef, E., Meyerhoff, M., and Clark, L. (2011). Teenagers' acquisition of variation: a comparison of locally-born and migrant teens' realisation of English (ing) in Edinburgh and London. Engl. World-Wide *32*, 206–236.
5. Greenberg, J.H., Turner, C.G., and Zegura, S.L. (1986). The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. Curr. Anthropol. *27*, 477–497.
6. Cavalli-Sforza, L.L., Piazza, A., Menozzi, P., and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA *85*, 6002–6006.
7. Cavalli-Sforza, L.L., Minch, E., and Mountain, J.L. (1992). Coevolution of genes and languages revisited. Proc. Natl. Acad. Sci. USA *89*, 5620–5624.
8. Hewlett, B., De Silvestri, A., and Guglielmino, C.R. (2002). Semes and genes in Africa. Curr. Anthropol. *43*, 313–321.
9. Pagel, M. (2009). Human language as a culturally transmitted replicator. Nat. Rev. Genet. *10*, 405–415.
10. Creanza, N., Ruhlen, M., Pemberton, T.J., Rosenberg, N.A., Feldman, M.W., and Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. Proc. Natl. Acad. Sci. USA *112*, 1265–1272.

11. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.

12. Knight, A., Underhill, P.A., Mortensen, H.M., Zhivotovsky, L.A., Lin, A.A., Henn, B.M., Louis, D., Ruhlen, M., and Mountain, J.L. (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr. Biol. *13*, 464–473.

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

14. Hunley, K.L., Cabana, G.S., Merriwether, D.A., and Long, J.C. (2007). A formal test of linguistic and genetic coevolution in native Central and South America. Am. J. Phys. Anthropol. *132*, 622–631.

15. Lansing, J.S., Cox, M.P., Downey, S.S., Gabler, B.M., Hallmark, B., Karafet, T.M., Norquest, P., Schoenfelder, J.W., Sudoyo, H., Watkins, J.C., and Hammer, M.F. (2007). Coevolution of languages and genes on the island of Sumba, eastern Indonesia. Proc. Natl. Acad. Sci. USA *104*, 16022–16026.

16. Wang, S., Lewis, C.M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M.V., Molina, J.A., Gallo, C., et al. (2007). Genetic variation and population structure in native Americans. PLoS Genet. *3*, e185.

17. Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K., Lea, R.A., Loo, J.-H., Koki, G., Hodgson, J.A., et al. (2008). The genetic structure of Pacific Islanders. PLoS Genet. *4*, e19.

18. Hunley, K., Dunn, M., Lindström, E., Reesink, G., Terrill, A., Healy, M.E., Koki, G., Friedlaender, F.R., and Friedlaender, J.S. (2008). Genetic and linguistic coevolution in Northern Island Melanesia. PLoS Genet. *4*, e1000239.

19. Lewis, C.M., Jr., and Long, J.C. (2008). Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. Mol. Biol. Evol. *25*, 478–486.

20. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.

21. Hunley, K., Bowern, C., and Healy, M. (2012). Rejection of a serial founder effects model of genetic and linguistic coevolution. Proc. Biol. Sci. *279*, 2281–2288.

22. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.

23. Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspinas, A.-S., et al. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. Science *349*, aab3884.

24. Labov, W. (1972). Sociolinguistic Patterns (University of Pennsylvania Press).

25. Verdu, P., and Rosenberg, N.A. (2011). A general mechanistic model for admixture histories of hybrid populations. Genetics *189*, 1413–1426.

26. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.

27. Holm, J. (1989). Pidgins and Creoles, Volume II, Reference Survey (Cambridge University Press).

28. Holm, J. (2000). An Introduction to Pidgins and Creoles (Cambridge University Press).

29. Baptista, M. (2002). The Syntax of Cape Verdean Creole: The Sotavento Varieties (John Benjamins).

30. Stewart, C. (2007). Creolization: History, Ethnography, Theory (Left Coast Press).

31. Baptista, M. (2015). Continuum and variation in Creoles: out of many voices, one language. J. Pidgin Creole Lang. *30*, 225–264.

32. Brüser, M., dos Reis Santos, A., Dengler, E., Blum, A., and Lang, J. (2002). Dicionário do Crioulo da Ilha de Santiago (Cabo Verde) (Gunter Narr).

33. Quint, N. (2008). L'élément Africain dans la Langue Capverdienne (L'Harmattan).

34. Lang, J. (2009). Les Langues des Autres dans la Créolisation (Gunter Narr).

35. Baptista, M. (2014). Echoes from the past: traces of Old Portuguese in Santiaguense TMA system. Papia *24*, 113–132.

36. Parra, E.J., Ribeiro, J.C.T., Caeiro, J.L.B., and Riveiro, A. (1995). Genetic structure of the population of Cabo Verde (west Africa): evidence of substantial European admixture. Am. J. Phys. Anthropol. *97*, 381–389.

37. Gonçalves, R., Fernandes, A.T., and Brehm, A. (2004). Cabo Verde islands: different maternal and paternal heritage testifies the nature of its first settlers. Int. Cong. Ser. *1261*, 372–373.

38. Beleza, S., Campos, J., Lopes, J., Araújo, I.I., Hoppfer Almada, A., Correia e Silva, A., Parra, E.J., and Rocha, J. (2012). The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies. PLoS ONE *7*, e51103.

39. Beleza, S., Johnson, N.A., Candille, S.I., Absher, D.M., Coram, M.A., Lopes, J., Campos, J., Araújo, I.I., Anderson, T.M., Vilhjálmsson, B.J., et al. (2013). Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet. *9*, e1003372.

40. Lewis, M.P., Simons, G.F., and Fennig, C.D., eds. (2016). Ethnologue: Languages of the World, Nineteenth Edition (SIL International).

41. Carreira, A. (1972). Cabo Verde: Formação e Extinção de uma Sociedade Escravocrata (1460-1878) (Centro de Estudos da Guiné Portuguesa).

42. Curtin, R. (1972). The Atlantic Slave Trade: A Census (University of Wisconsin Press).

43. Eltis, D., and Richardson, D. (2010). Atlas of the Transatlantic Slave Trade (Yale University Press).

44. Arends, J., Muysken, P., and Smith, N. (1995). Pidgins and Creoles: An introduction (Benjamins).

45. Schumann, J.H. (1978). The Pidginization Process: A Model for Second Language Acquisition (Newbury House).

46. Muysken, P., ed. (1981). Generative Studies on Creole Languages (Foris).

47. Anderson, R.W., ed. (1983). Pidginization and Creolization as Language Acquisition (Newbury House).

48. Bakker, P. (1997). A Language of Our Own (Oxford University Press).

49. Arends, J. (1989). Syntactic Developments in Sranan: Creolization as a Gradual Process. Thesis (Catholic University of Njimegen).

50. Croft, W. (2000). Explaining Language Change: An Evolutionary Approach (Longman).

51. Mufwene, S.S. (2001). The Ecology of Language Evolution (Cambridge University Press).

52. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

53. Jakobsson, M., and Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics *23*, 1801–1806.

54. Rosenberg, N.A. (2004). DISTRUCT: a program for the graphical display of population structure. Mol. Ecol. Notes *4*, 137–138.

55. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

56. Pemberton, T.J., Wang, C., Li, J.Z., and Rosenberg, N.A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am. J. Hum. Genet. *87*, 457–464.

57. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

58. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza,

L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

59. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., and Cavalli-Sforza, L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. Nature *368*, 455–457.
60. Chafe, W.L., ed. (1980). The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (Ablex).
61. Baptista, M. (2006). When substrates meet superstrate: the case of Cape Verdean Creole. In Cabo Verde: Origens da sua Sociedade e do seu Crioulo, J. Lang, J. Holm, J.-L. Rougé, and M.J. Soares, eds. (Gunter Narr), pp. 91–116.
62. Bartens, A. (2006). A contribuição do substrato africano para a génese dos crioulos caboverdianos: o caso dos ideofones. In Cabo Verde: Origens da sua Sociedade e do seu Crioulo, J. Lang, J. Holm, J.-L. Rougé, and M.J. Soares, eds. (Gunter Narr), pp. 117–131.
63. Brito, A. (1967). Apontamentos para a gramática do criolo que se fala na ilha de Santiago de Cabo Verde. In Estudos linguísticos crioulos: reedição de artigos publicados no Boletim da Sociedade de Geografia de Lisboa, J. Morais-Barbosa, ed. (Academia Internacional da Cultura Portuguesa), pp. 329–404.
64. Lang, J. (2006). L'influence des Wolof et du wolof sur la formation du créole santiagais. In Cabo Verde: Origens da sua Sociedade e do seu Crioulo, J. Lang, J. Holm, J.-L. Rougé, and M.J. Soares, eds. (Gunter Narr), pp. 53–62.
65. Quint, N. (2000). Le Cap Verdien: Origines et Devenir d'une Langue Métisse (L'Harmattan).
66. Quint, N. (2006). Un bref aperçu des racines Africaines de la langue cap-verdienne. In Cabo Verde: Origens da sua Sociedade e do seu Crioulo, J. Lang, J. Holm, J.-L. Rougé, and M.J. Soares, eds. (Gunter Narr), pp. 75–90.
67. Rougé, J.-L. (1988). Petit Dictionnaire Étymologique de Kriol de Guiné Bissau et de Casamance (INEP).
68. Rougé, J.-L. (1994). A propos de la formation des créoles du Cap Vert et de Guinée. Papia *3*, 137–149.
69. Santos, R. (1979). Le Créole des Iles du Cap-Vert: Comparaison avec les Langues Africaines. Thesis (Université de Dakar, Sénégal).
70. Snyder, J.P. (1987). Map Projections: A Working Manual. Technical Report (U.S. Geological Survey).
71. Legendre, P., and Legendre, L. (1998). Numerical Ecology (Elsevier).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Asd	Z.A. Szpiech	https://github.com/szpiech/asd
Admixture	[52]	https://www.genetics.ucla.edu/software/admixture/index.html
Clumpp	[53]	http://rosenberglab.stanford.edu/software.html
Distruct	[54]	http://rosenberglab.stanford.edu/software.html
Plink	[55]	https://www.cog-genomics.org/plink2

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Noah A. Rosenberg (noahr@stanford.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We sampled 50 Kriolu-speaking adult volunteers living on the island of Santiago (Figure S1), 40 in the capital city, Praia, and 10 in the inland town of Picos. We recorded birthplaces of individuals and their parents and grandparents. Birthplaces and parental birthplaces included seven of the nine main inhabited islands of Cape Verde as well as other formerly Portuguese-controlled locations (Angola, Brazil, Guinea-Bissau, Mozambique, Portugal, and São Tomé). The majority of both offspring (38/50) and parents (71/100) were born on Santiago.

Recruitment followed Helsinki guidelines and participants provided informed consent. The study was approved by the Cape Verde Ministério da Saúde, the French Muséum National d'Histoire Naturelle and Ministère de l'Enseignement Supérieur et de la Recherche, Stanford University, and the Universities of Cape Verde, Manitoba, and Michigan. Subjects included 17 males and 33 females, with mean age 52 (standard deviation 19.6, minimum 19, maximum 87).

METHOD DETAILS

Genotyping

DNA samples were collected using DNAGenotek Saliva Sampling Kits, and DNA was extracted following the manufacturer protocol. Next, 48 individuals were typed at 2,379,855 SNPs using the Illumina HumanOmni2.5-8 BeadChip genotyping array. Raw data were generated using an Illumina BeadArray Reader with BeadScan Software, and we called genotypes using the Illumina GenomeStudio (v2011.1) Genotyping Module (v1.9.4). We removed 7,241 SNPs with ambiguous position on the human genome reference sequence and 58,443 sex-chromosomal and mitochondrial SNPs. Thus, we retained 2,314,171 autosomal SNPs in the 48 Cape Verdean individuals for further procedures.

Genotype Quality Control

We prepared a dataset for population-genetic analysis in four stages. We first removed 6,737 SNPs that failed to reach cutoffs for call rate ($\geq$90%) and cluster separation ($\geq$0.2), and one copy of each of 3,206 SNPs included in duplicate on the genotyping array (Figure S2, stage 1). We subjected the remaining 2,304,228 autosomal SNPs to population-genetic quality control [56]. We removed 4 SNPs with high missing data ($\geq$10%), 385,397 SNPs monomorphic in our Cape Verde sample, 115 SNPs on the basis of a Hardy-Weinberg test, and 14 insertion/deletion markers (Figure S2, stage 2). We thus retained 1,918,698 autosomal SNPs in 48 individuals for relatedness inference.

Relatedness

We inferred first- and second-degree genetic relationships among all possible pairs of individuals using identity-by-state allele-sharing and likelihood-based relationship inference [56]. We identified four relative pairs, all compatible with information gathered during interviews. For each pair, we retained the individual with fewer missing genotypes. We then repeated our population-genetic quality control procedures to obtain a final dataset of 44 unrelated Cape Verdeans typed at 2,304,069 autosomal SNPs, including 393,457 monomorphic in our 44-individual set (Figure S2, stage 3).

Genetic Data from Worldwide Populations

We merged our data on 44 unrelated Cape Verdeans with genotypes from The 1000 Genomes Project [57] and Human Genome Diversity Panel [58] (Figure S2, stage 4). Our final worldwide data contained 420,752 SNPs in 3,418 unrelated individuals from 80 populations, including Cape Verdeans (Figure S1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Allele-Sharing Dissimilarities

We computed the allele-sharing dissimilarity (ASD) between pairs of genomes [59], using 3,418 individuals and 420,752 SNPs, employing *asd* v1.0 (https://github.com/szpiech/asd) and considering for each pair only those SNPs for which neither individual had missing genotypes.

Multidimensional Scaling

We performed metric MDS on the pairwise ASD matrices using the *cmdscale* function in R. Each MDS analysis employed a submatrix of the full ASD matrix, considering only individuals from specified populations.

ADMIXTURE

We performed unsupervised model-based clustering using ADMIXTURE [52], considering 44 Cape Verdeans alongside West African (109 Gambian Mandinka, 22 Senegalese Mandenka) and European populations (107 Iberian, 28 French, 89 British). ADMIXTURE was applied to a subset of 72,619 SNPs pruned by linkage disequilibrium (LD), obtained using the PLINK *indep-pairwise* option [55] with a 50-SNP sliding window incremented by 10 SNPs, and an LD threshold of $r^2 = 0.1$.

We performed 50 independent ADMIXTURE runs for each K value from 2 to 4. For each K, we used CLUMPP [53] to identify similar clustering solutions, averaging individual membership proportions across all pairs of runs producing symmetric similarity coefficient SSC > 0.95; for both $K = 2$ and $K = 3$, all pairs had SSC > 0.95. Mean log-likelihood values for runs in the major mode were $-27{,}668{,}563$ for $K = 2$ (SD = 1.18×10^{-6}), $-27{,}617{,}483$ for $K = 3$ (SD = 0.45), and $-27{,}572{,}328$ for $K = 4$ (SD = 7.77), respectively. We plotted membership coefficients using DISTRUCT [54]. Clustering patterns of the major mode shown in Figure 1C are likely not due to uneven population sample sizes; in ADMIXTURE reanalysis of an individual set with the Gambian Mandinka, Iberian, and British samples randomly downsampled population-wise to at most 44 individuals, the number of Cape Verdeans in this study, clustering patterns are similar (Figure S4). This analysis used 71,173 SNPs, selected on the basis of LD pruning in the downsampled set of individuals. Mean log-likelihood values for runs in the major mode were $-15{,}526{,}615$ for $K = 2$ (SD = 6.03×10^{-7}), $-15{,}484{,}779$ for $K = 3$ (SD = 0.04), and $-15{,}441{,}540$ for $K = 4$ (SD = 16.16), respectively.

Genetic Admixture Estimates

We produced estimates of African admixture in the 44 unrelated Cape Verdeans using both ADMIXTURE and MDS. For ADMIXTURE, Cape Verdean membership proportions in the clusters corresponding to Africans (summing the ancestry components in the orange and red clusters) in Figure 1C are similar for $K = 2$, 3, and 4 ($\rho \geq 0.9984$, $p < 2 \times 10^{-16}$, for all three comparisons). For MDS in two dimensions (Figure 1B), we determined the centroid of two West African populations (Gambian Mandinka, Senegalese Mandenka) and the centroid of Iberian individuals and projected each Cape Verdean individual onto the line connecting these centroids (Figure 1B). For each Cape Verdean, we estimated African admixture as one minus the Euclidean distance of the projection to this West African centroid, scaled by the absolute Euclidean distance between the West African and the Iberian centroids. Considering the centroid of all Iberian, British, and French individuals instead of only the Iberians had little effect on Cape Verdean admixture estimates ($\rho = 0.9997$, $p < 2 \times 10^{-16}$).

ADMIXTURE ($K = 2$) and MDS (Figure 1B) estimates of West African admixture in Cape Verde were quite similar ($\rho = 0.9944$, $p < 2 \times 10^{-16}$). Thus, for subsequent analyses, we estimated African admixture proportions in each Cape Verdean as their genotype membership proportions in the ADMIXTURE orange cluster at $K = 2$.

Linguistic Interviews

Each subject separately watched a ~6 min speech-free movie widely used in linguistic research, entitled *The Pear Stories* [60], after which we recorded their narration of the story in Kriolu without interruption or time restriction. The recording for each subject was transcribed into text by one of nine fluent Kriolu speakers.

Word Frequency Profiles

Transcripts for the 50 individuals were parsed to obtain subject-specific word frequency counts. First, punctuation and accents—except for apostrophes and hyphens, which are integral to Kriolu word definition—were removed. Each string of non-whitespace characters between two whitespace characters was considered to be a word. Among the 44 unrelated Cape Verdeans, the number of unique words observed in the combined text from all transcripts was 2,474. These 44 transcripts averaged 517.9 words in length, containing an average of 181.5 unique words.

For each of the 44 transcripts, the number of times each unique word appeared was tabulated in a vector, from which we computed the "word frequency profile" f_i of the individual. Let n_{ij} be the number of times word j appeared in the transcript of individual i, where

$n_{ij} = 0$ if word j was absent for individual i; $f_i = (f_{i1}, f_{i2},\dots, f_{iW})$ is the 1 × W vector in which $f_{ij} = n_{ij}/\sum_{w=1}^{W} n_{iw}$ is the relative frequency of word j in the transcript, and $W = 2{,}474$. The linguistic distance between a pair of word frequency profiles, $d_E(f_i, f_j)$, was computed with the Euclidean metric, $d_E(f_i, f_j) = [\sum_{w=1}^{W} (f_{iw} - f_{jw})^2]^{1/2}$.

Identification of Words of African Origin

From the list of 2,474 Kriolu words in the 44 transcripts, we identified 212 words corresponding to instances of 29 underlying roots and morphemes with African origins (Table S1). Words of African origin were identified on the basis of their form, semantics, or phonology following the scholarship in several sources [32, 34, 61–69].

Each of the items exhibits African substratal influences, in that they are of African origin either in form (morphology), phonology, or semantics. For the latter category, the word is Portuguese but the new meaning, which does not exist in Portuguese, has been traced to a specific African language. These are cases of linguistic reanalysis: the phonetic material is from Portuguese but the grammatical function is African, a widespread and well-documented cognitive process in creolistics.

Linguistic Admixture Estimates

We determined the African admixture score in each word frequency profile as the sum of the frequencies of the 212 words with African origins. Denoting the set of such items B, the African linguistic admixture score Z_i for an individual i was computed as $Z_i = \sum_{w \in B} f_{iw}$.

Geographic Distances

Distances between geographic positions (latitude, longitude) were computed using the haversine formulation of the great circle distance [70], taking 6,371 km for the radius of the Earth. Birthplace coordinates were taken to be the latitude and longitude of birth towns, obtained from Google Earth. For 40 of 44 unrelated subjects, birth town was known; maternal and paternal birth towns were known for 36 and 31 subjects, respectively. When birth town was unknown within an island, the largest city on the island was used. Geographic analyses used 34 subjects for which birth town was known and birthplaces of both parents were in Cape Verde and known to at least island level.

Linguistic Distances and Parental Birthplaces

Mantel and partial Mantel tests [71] were conducted using *mantel* and *mantel.partial* in the R package *vegan*. We evaluated significance by permutation, recording the number among 10,000 replicates in which the correlation in the permuted data exceeded that in the actual data. These tests were applied to submatrices of 44-individual matrices for the 34 individuals born on Cape Verde with parents also born on Cape Verde.

To compare linguistic distances in offspring with geographic distances between parental birthplaces, we determined the Mantel correlation of $d_E(f_i, f_j)$ with $(D_{ij}^m + D_{ij}^p)/2$, where D_{ij}^m is the distance between the birthplaces of the mothers of individuals i and j and D_{ij}^p is the corresponding distance for fathers. We also performed a partial Mantel test of the partial correlation between $d_E(f_i, f_j)$ and $(D_{ij}^m + D_{ij}^p)/2$, conditioning on the distance D_{ij} between the birthplaces of i and j.

DATA AND SOFTWARE AVAILABILITY

Following IRB guidelines, data are available upon request subject to ethical approvals.