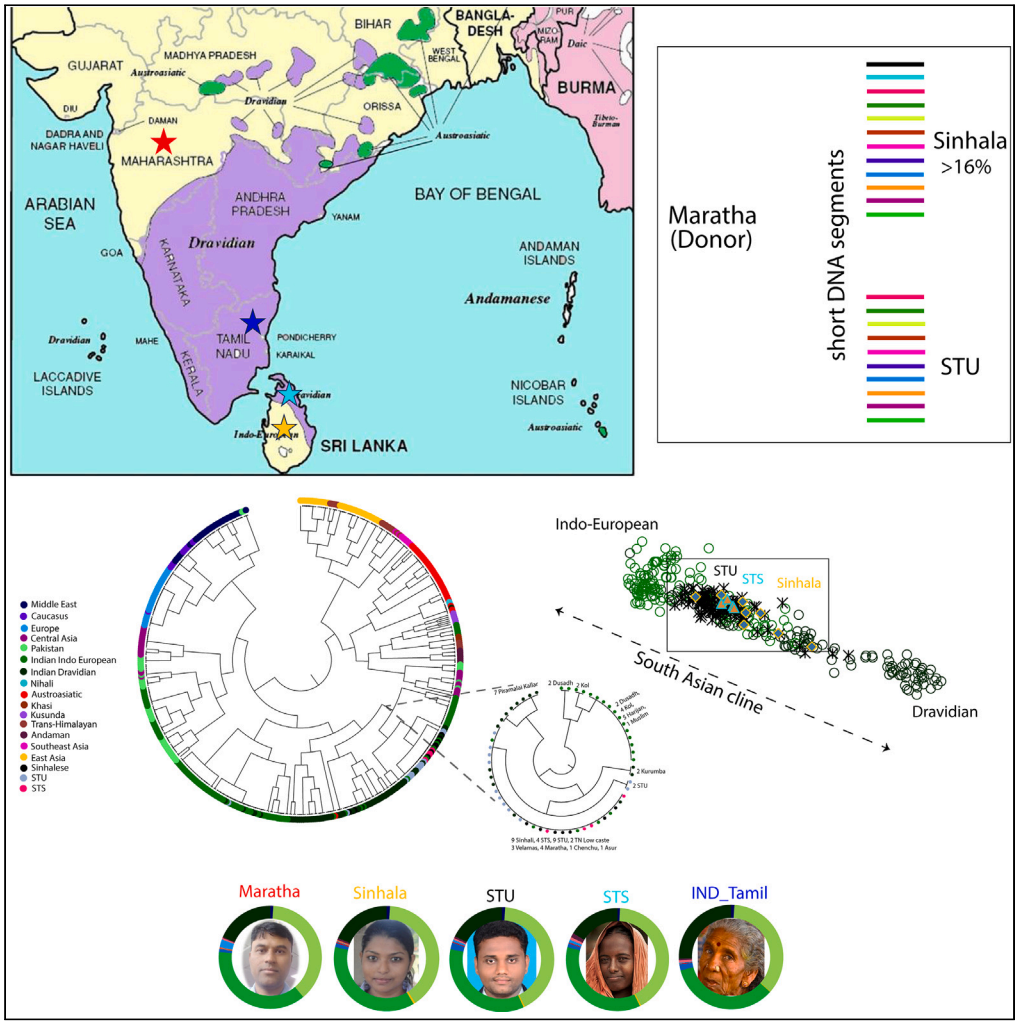
iScience

CellPress
OPEN ACCESS

Article

Reconstructing the population history of the Sinhalese, the major ethnic group in Śrī Laṅkā



Prajjval Pratap Singh, Sachin Kumar, Nagarjuna Pasupuleti, ..., Niraj Rai, Gyaneshwer Chaubey, R. Ranasinghe

nirajrai@bsip.res.in (N.R.)
gyaneshwer.chaubey@bhu.ac.in (G.C.)
ruwa@ibmbb.cmb.ac.lk (R.R.)

Highlights

Higher West Eurasian genetic component in Śrī Laṅkā than South India

A strong gene flow beyond the boundary of ethnicity and language in Śrī Laṅkā

Traces of common roots of Sinhala with Maratha

Singh et al., iScience *26*, 107797
October 20, 2023
https://doi.org/10.1016/j.isci.2023.107797

iScience

CellPress
OPEN ACCESS

Article

Reconstructing the population history of the Sinhalese, the major ethnic group in Śrī Laṅkā

Prajjval Pratap Singh,[1,6] Sachin Kumar,[2,6] Nagarjuna Pasupuleti,[3] P.R. Weerasooriya,[4] George van Driem,[5] Kamani H. Tennekoon,[4] Niraj Rai,[2,*] Gyaneshwer Chaubey,[1,7,*] and R. Ranasinghe[4,*]

SUMMARY

The Sinhalese are the major ethnic group in Śrī Laṅkā, inhabiting nearly the whole length and breadth of the island. They speak an Indo-European language of the Indo-Iranian branch, which is held to originate in northwestern India, going back to at least the fifth century BC. Previous genetic studies on low-resolution markers failed to infer the genomic history of the Sinhalese population. Therefore, we have performed a high-resolution fine-grained genetic study of the Sinhalese population and, in the broader context, we attempted to reconstruct the genetic history of Śrī Laṅkā. Our allele-frequency-based analysis showed a tight cluster of Sinhalese and Tamil populations, suggesting strong gene flow beyond the boundary of ethnicity and language. Interestingly, the haplotype-based analysis preserved a trace of the North Indian affiliation to the Sinhalese population. Overall, in the South Asian context, Śrī Laṅkān ethnic groups are genetically more homogeneous than others.

INTRODUCTION

Śrī Laṅkā is located at 37^0 N and 127^0 30′ E, where the Bay of Bengal meets the Indian Ocean. The surface area comprises 65,610 km^2 of land and water. The ancient Greeks referred to the island Ταπροβάνη Taprobánē, although this toponym may also have been applied by the Greeks to Sumatra. A now-lost indigenous name for the island, ultimately deriving from Sanskrit Siṃhaladvīpa, was transmogrified into Arabic Sarandīb and Persian Sarandīp and so entered European languages as Serendip. The Portuguese name, written Seylan, Ceylan, or Ceylon (cf. modern Portuguese Ceilão), preserves an old native Sinhala name for the island, which derived etymologically from Sanskrit Śrī Laṅkā. The name Ceylon was subsequently adopted by the Dutch, who governed the island from 1640 until 1796, and the British, who ruled Ceylon from 1796 to 1948. The Sanskritised tatsama loanword "Śrī Laṅkā" replaced the European rendition of the original old native Sinhala name in 1972.[1]

The current census estimates Śrī Laṅkā to have 22 million inhabitants, of which the Sinhalese represent the major ethnic group, comprising 74.9% of the population. Other ethnic groups include Śrī Laṅkān Tamils at 11.1%, Muslims or "Moors" at 9.3%, Indian Tamils at 4.1%, and others at 0.6%, i.e., Burgher, Malay, Vedda (Adivasi).[2] It has been conjectured that hunter-gatherers with paleolithic technology settled in Śrī Laṅkā perhaps as early as 125,000 years ago,[3,4] but the earliest anatomically modern human fossil in Śrī Laṅkā dates from 28,500 years ago, found at the Upper Pleistocene site of Batadombalena, evidently inhabited by humans from 36,000 years ago.[5–7] Śrī Laṅkā was inhabited by Mesolithic hunter-gatherers until ca. 800–600 BC when both cattle and agriculture were introduced by the bearers of an Iron Age culture with a Black and Red Ware ceramic culture who practiced megalithic burials. The bearers of this new agricultural civilization are held to have been the Siṃhala, Ceylonese or Sinhalese.[8] The Dīpavaṃsa and Mahāvaṃsa record that Prince Vijaya led the ancestral Siṃhala from Siṃhapura or Sihapura in Lāḷa or Lāṭa in what today is southern Gujarat. Vijaya reigned at the newly established Tambapaṇṇi ca. 468–448 B.C.[9,10]

Wilhelm Ludwig Geiger[11–14] established that Koṅkaṇī, spoken on the Koṅkaṇ coast of India, represented the closest linguistic relative of both Sinhalese, spoken in Śrī Laṅkā, and Divehi, spoken in the Maldives. Geiger inferred that this demonstrable linguistic relationship reflected the ancient maritime migration across the Arabian Sea and Indian Ocean that first brought Divehi- and Sinhalese-speaking populations to their insular habitats in the first millennium BC. Geiger grouped both these languages with Koṅkaṇī, Marāṭhī, and Gujarātī, which in Turner's classification[15] together constitute the Southwestern sub-branch of the Indo-Aryan branch of Indo-European. The Sinhalese chronicles record that for nine months, the newly arrived Sinhala settlers endeavored to exterminate the native populace of the island, whom they called the yakkhas (Skt. yakṣa), which scholars have identified with the Veddas.[1,8,9,16]

[1]Cytogenetics Laboratory, Department of Zoology, Banaras Hindu University, Varanasi 221005, India
[2]Ancient DNA Lab, Birbal Sahni Institute of Palaeosciences, Lucknow 226607, India
[3]Department of Applied Zoology, Mangalore University, Mangalore 574199, India
[4]Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, No. 90, Cumaratunga Munidasa Mawatha, Colombo 03 00300, Śrī Laṅkā
[5]Institut für Sprachwissenschaft, Universität Bern, Länggassstrasse 49, 3012 Bern, Switzerland
[6]These authors contributed equally
[7]Lead contact
*Correspondence: nirajrai@bsip.res.in (N.R.), gyaneshwer.chaubey@bhu.ac.in (G.C.), ruwa@ibmbb.cmb.ac.lk (R.R.)
https://doi.org/10.1016/j.isci.2023.107797

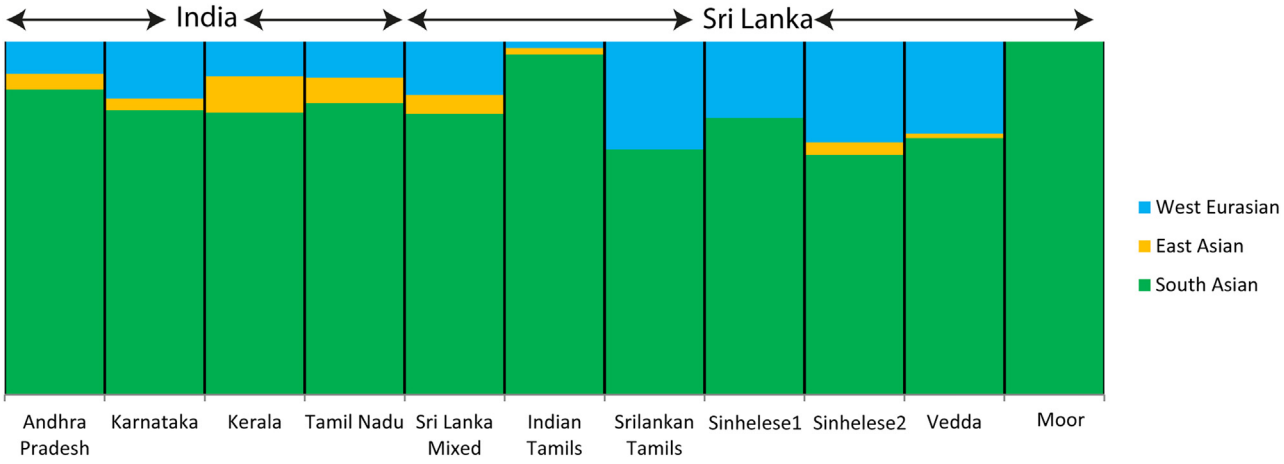


Figure 1. Comparison of maternal ancestry components between Śrī Laṅkā and South India populations

While the Sinhalese are associated with the earliest inscriptions on the island, dating from the time of Aśoka, it has been argued on linguistic grounds that the ancestors of the Tamils crossed the Palk Strait and settled in the North of Śrī Laṅkā at roughly the same time, viz. in the second half of the first millennium BC, during the cultural foment that yielded the dawn of the Cōḻa dynasty on the subcontinent.[1] This linguistic dating is supported by the fact that the thickest bundle of isoglosses runs—as one might expect—between the continental dialects of Tamil and the dialects of Ceylon'.[17]

After Sinhalese and Tamil colonization in the first millennium BC, Śrī Laṅkā's geographical proximity to the Indian subcontinent was enhanced by close cultural ties. This same period saw the dawn of the great maritime Hindu and Buddhist expansion from the subcontinent into mainland and insular Southeast Asia, historically involving both, gene flow and cultural transmission.[8,18–21] Centuries later, Muslim traders arrived from Arabia, Malays from Malaya and, in the British colonial period, Indian Tamils from South India.

Only a few genetic studies, including the mtDNA, Y and X chromosomes have been performed, and these confirmed the Sinhalese connection with mainland India.[22–35] Some studies have shown that the Sinhalese have a distinct origin, while a few of them suggested a connection with South Indian populations.[26,36] Analysis based on classical markers advocated a closer affinity of the Sinhalese population with South and West Indian populations than with the Bengalis.[37,38] The question remains as to how the Sinhalese relate to the other peoples of Śrī Laṅkā in view of ongoing debates on the origin of the Sinhalese and the Śrī Laṅkān Tamils (STU). Therefore, in the present study, we have evaluated various alternatives to establish a molecular genetic perspective on the origin of Sinhalese and preclude possible source populations and genetic admixing.

Śrī Laṅkā also represents an important staging area in any scenario involving the theory of a southern migration route, and so a better understanding of the population genetics of the Sinhalese may offer novel insights into the early peopling of South Asia. Genetic studies on the Śrī Laṅkān population are mainly limited to haploid DNA markers.[22,39] The majority of Śrī Laṅkān individuals studied so far showed an overwhelming presence of South-Asian-specific haplogroups. However, a significant presence of West-Eurasian-specific haplogroups has also been detected. The most common West-Eurasian mtDNA haplogroups are U7 and U1.[22,40] Thus, the West Eurasian connection of Śrī Laṅkā appears likely. The lack of autosomal studies needs to be filled in order to understand the precise nature of peopling of Śrī Laṅkā. Therefore, we have analyzed and evaluated the Śrī Laṅkān Sinhalese and Tamil groups for hundreds of thousands of genetic markers.

RESULTS AND DISCUSSION

To have a detailed understanding on the origin and migration of the Sinhala population, we have first evaluated the maternal gene flow among the Śrī Laṅkān population. We collected data from public sources[22,40] and compared them with the South Indian maternal population composition. The Śrī Laṅkān and South Indian maternal gene pool overwhelmingly showed a South Asian affinity (Figure 1). However, we see a striking difference in the prominence of West Eurasian ancestry. Assuming that the West Eurasian ancestry of Śrī Laṅkā arrived from mainland India, we should expect to see a significantly lower proportion of this ancestry in Śrī Laṅkā than in South Indian populations, but this was not the case. Instead, we observed a significantly higher frequency (two-tailed $p < 0.0001$) of West-Eurasian-specific maternal ancestry in Śrī Laṅkān populations (Figure 1). This high level of West Eurasian ancestry is consistent across all the major Śrī Laṅkān groups except Indian Tamils, who are known to represent a well-documented recent migration during the British colonial period[41] and the Moors, who overwhelmingly exhibit South Asian ancestry. This discrepancy can be explained by independent West Eurasian contribution to Śrī Laṅkā, likely by a sea route and putative migration from Northwest India (Figure 1).

In order to understand more about the West-Eurasian-related ancestry and the population history of the Śrī Laṅkān populations, we used hundreds of thousands of autosomal markers. We extracted a large dataset in addition to Indian samples for comparative autosomal analysis and merged these datasets with our newly generated genome-wide data. First, we performed PCA analysis in order to understand the

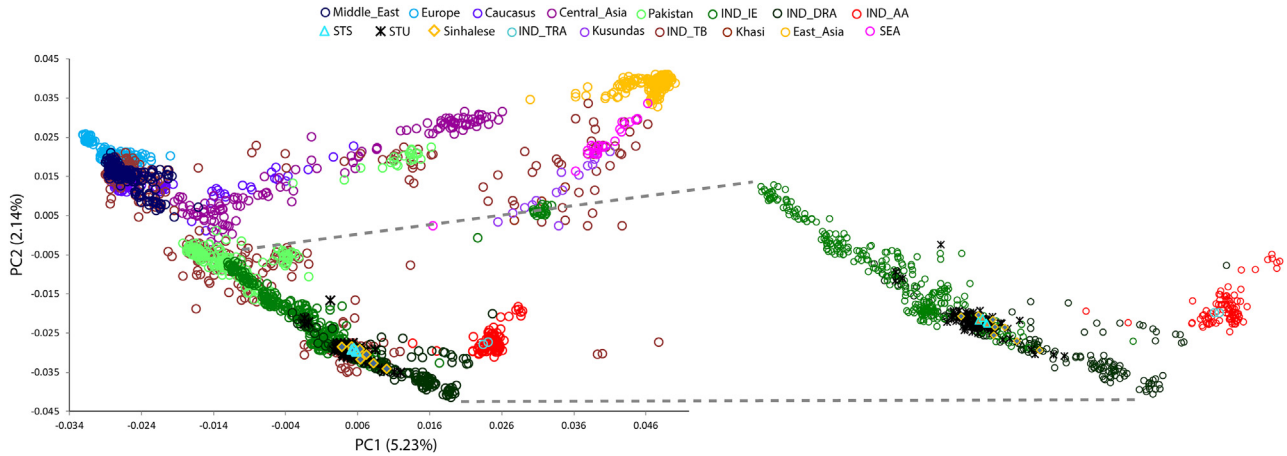


Figure 2. The principal component analysis of studied populations with respect to the Eurasian populations

population affinity. The scatterplot (Figure 2), using the obtained PC1 and PC2 eigenvectors, suggested that the Sinhalese, Śrī Laṅkān Tamils in Śrī Laṅkā (STS), and the Śrī Laṅkān Tamils in the United Kingdom (STU) are close to one another in a large cluster on the South Asian Indo-European to Dravidian cline. This finding suggests a closer genetic affinity of the Sinhalese population with the Śrī Laṅkān Tamil population (Figure 2). In order to investigate ancestral components, ADMIXTURE was performed, which also showed (Figure 3) that the Sinhalese are more similar to Śrī Laṅkān Tamils than to the Indian populations, and both possess a major South-Asian-related ancestral component. The light and dark green color components specific to South Asian populations were nearly equally distributed in Sinhalese and Śrī Laṅkān Tamils (Figure 3).

To ascertain the fine-scale genetic similarity, we performed haplotype-based fine structure analysis. Consistent with the PCA Admixture results, both Śrī Laṅkān populations shared a close genetic affinity (Figure 4) and fell in the same cluster. Both populations also shared a common clade with Indian Indo-European and Dravidian populations. The chunk count comparison suggested that both ethnic groups of Śrī Laṅkā received major chunks from each other and from Indian Indo-European and Dravidian populations.

In order to understand the putative source populations for both ethnic groups, firstly, *f*3-statistics were calculated with the world population, using several sources, such as pop1 and pop2, while Sinhalese and Śrī Laṅkān Tamils (STU) were taken as the target population. The results from *f*3-admix suggested that the Sinhalese and Śrī Laṅkān Tamils are admixed populations of Indian, Indo-European, and Dravidian ancestry (Table S1). Since STU are collected from the UK, we have compared them to see if they have any deviation from the genetic composition of native Sri Lankan Tamils (STS). All the analyses i.e., PCA, ADMIXTURE, outgroup *f*3, and D statistics did not find any significant deviation of STU from STS (Figures 2, 3, and 4; Tables S1 and S2).

To measure the gene flow using the obtained putative source populations and Yoruba as an outgroup population, D-statistics were performed, and the top ten D values for both of the populations suggested that strong gene flow has occurred between the Sinhalese and Śrī Laṅkān Tamils (STU) in the past because they show negative D-values with North Indians (Yoruba; Sinhalese/STS; STU; X) and positive D-values with South Indians. We also calculated D-statistics to infer the direction of gene flow between North vs. South Indian populations models (Yoruba; Sinhalese/STS/STU; X; Y) and obtained results suggesting that higher gene flow occurred between both the populations from the South than the North Indian populations. However, we have found slightly higher gene flow (but non-significant) from some North and Northwest Indian than the South Indian populations (Table S2).

These results are intriguing, considering the distinct linguistic affiliation of Sinhalese and STU/STS. The results indicate a strong gene flow beyond the boundaries of ethnicities' in question, which is usually rare in South Asia.[42,43]

We also evaluated the admixture timing for both ethnic groups using ALDER. Sinhalese and Śrī Laṅkān Tamils were used as target populations, while other world populations were considered as source populations. After several permutations and combinations, we could get a few successful models for STU, while only one model was successful for Sinhalese people. The admixing dates were very recent for both populations, while the low numbers of successful models might be due to high admixing, so the software could not use other populations as a putative source population (Table S3).

Runs of homozygosity (RoH) were calculated to understand the marriage pattern of Sinhalese and Śrī Laṅkān Tamils. The obtained mean values were plotted between the numbers of segments vs. the average numbers length of segments (in Kbs). The STS populations clustered at the base of the scatterplot, followed by Sinhalese, while STU showed a longer and higher number of homozygous segments (Figure 5). Results from the RoH suggest that the effective population size for these populations (Ne) varies. These disparities could be due to the sampling bias where STU were collected from outside South Asia (UK). More Tamil samples from Śrī Laṅkā could help to solve the disparity.

In order to test the linguistic hypothesis that the Sinhalese language shows closer common ancestry with Koṅkaṇī, Marāṭhī, and Gujarātī, we performed identity by descent (IBD) analysis (Figure 6), by comparing larger (2.0 to ∞ cM) and smaller (0–2 cM) chunks of DNA. When two

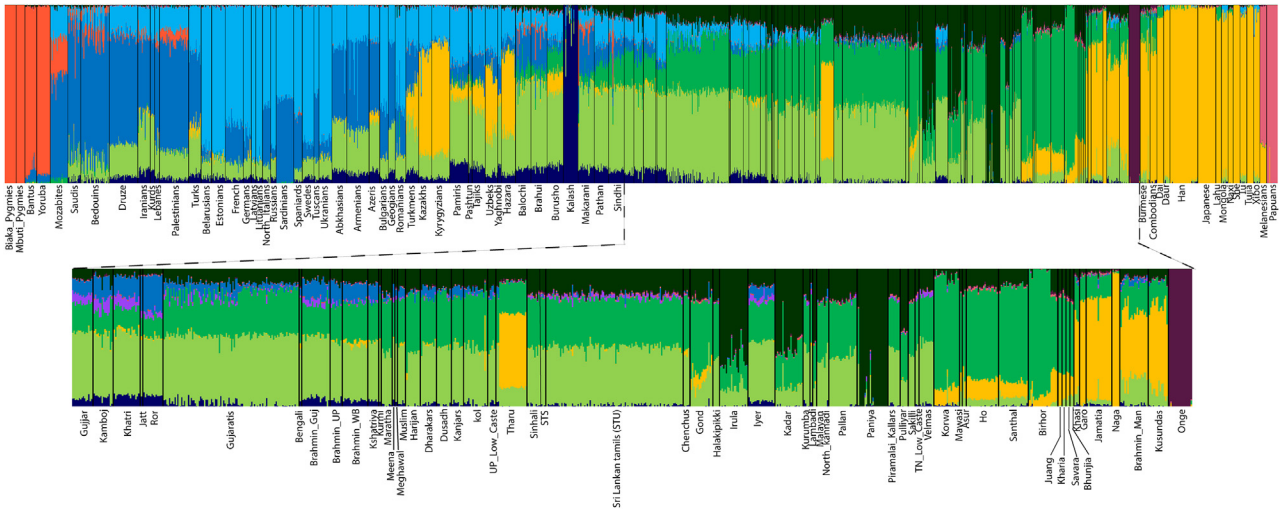


Figure 3. The bar plot of ADMIXTURE analysis showing the ancestral component sharing of studied populations. The Indian and Śrī Laṅkān ethnic groups are projected

population admix, recombination event tend to break the large DNA segments (chunks). With the time, these segment sizes become smaller and smaller. Thus comparing the large and small DNA segments can help us to understand the recent and old admixture processes. Interestingly, we found an unexpected excess of smaller chunks sharing between Marāṭhā and Sinhala (>16%) than between the Marāṭhā and STU, thus supporting the linguistic hypothesis of Geiger, Turner, and van Driem. To confirm the excess sharing, we looked for the population sharing maximum IBD with Sinhala and STU. We observed that South Indian Piramalai Kallar shared the highest IBD with Sinhala and STU, while, both populations showed highest IBD sharing, for short and long DNA segments with Piramalai Kallar. We asked whether Sinhalese or STU shared more DNA segments with Marāṭhā. The Piramalai Kallar shared nearly equally large DNA segments with Sinhalese and STU, respectively (Figure 6), whereas Marāṭhā shared significantly higher (>16%) smaller segments with Sinhalese (two tailed $p < 0.001$).

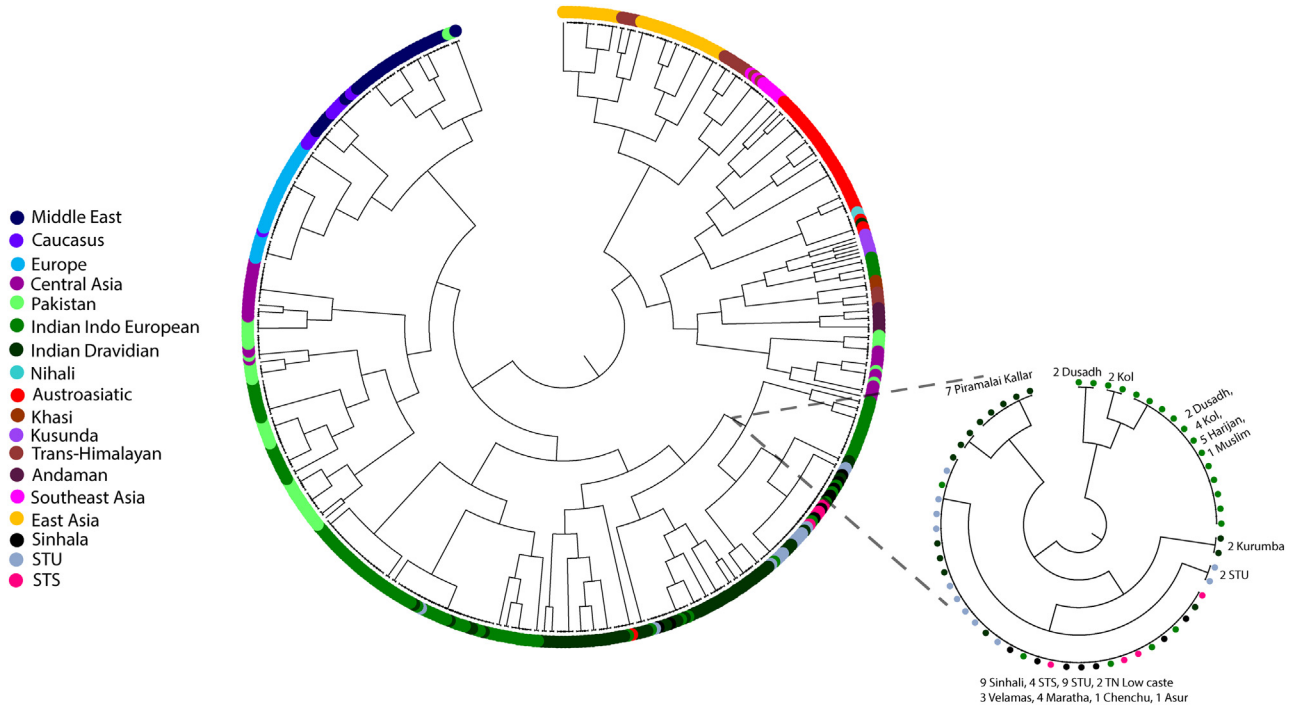


Figure 4. The Maximum Likelihood (ML) tree of Eurasian populations shows the studied populations' genetic affinity. The closest branch of our target populations are zoomed-in and shown in a subset

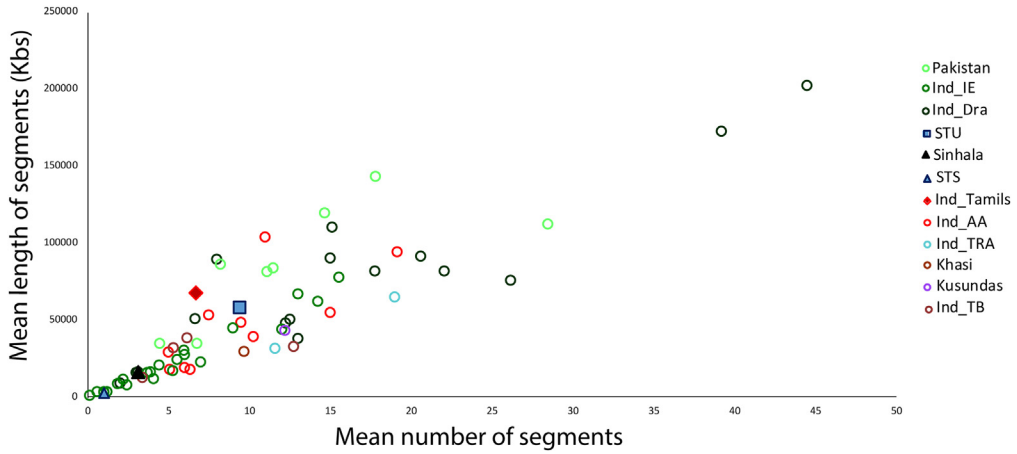


Figure 5. The Runs of Homozygosity (RoH) plot of target populations with respect to the other South Asian ethnic groups

This result is also visible in the D statistics test. However, it was non-significant (Table S2). This excess sharing of smaller segments suggests a closer, deeply rooted common genetic ancestry of the Sinhalese with the Marāṭhā.

In conclusion, this is the first comprehensive analysis with the high-throughput genome-wide autosomal data and comparative analysis of two major linguistically distinct ethnic groups of Śrī Laṁkā with ancient historical settlements. Our findings suggest a close genetic affinity of Sinhalese with STU, irrespective of their linguistic affiliation. This phenomenon is rare in South Asia. The genetic homogeneity of Sinhalese and STU is probably due to long-term close geographic sharing, which facilitated large amounts of gene flow. Furthermore, the traces of common roots of Sinhala with Maratha can also be seen in fine grained genetic analysis. Thus, the genetic analysis of Sinhalese adds another significant chapter to the history of the South Asian genetic landscape.

Limitation of the study

Although we corroborate the linguistic theory, our admixture time analysis was failed to confirm the timeline, likely due to the absence of true putative ancestor. More ancient DNA study and Y chromosomal sequencing would be useful to determine the migration timeline.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

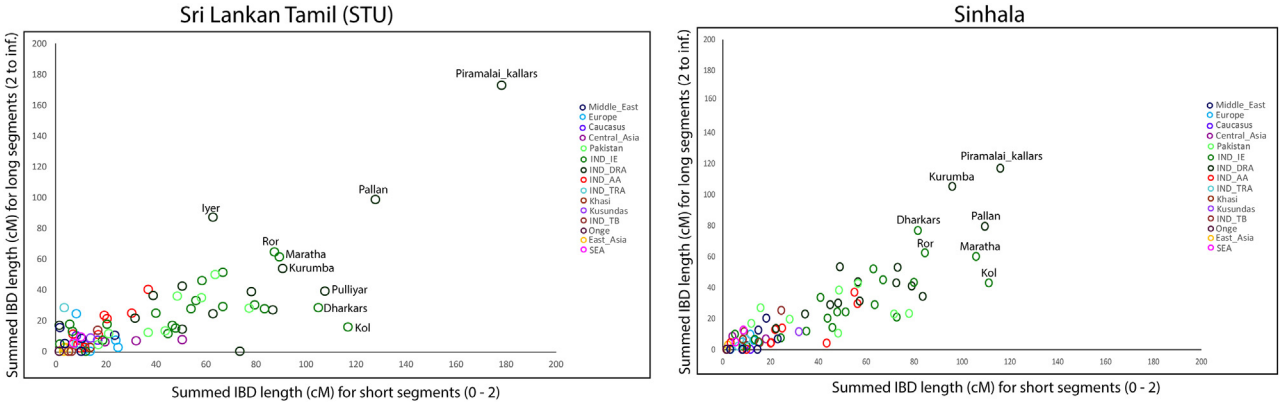


Figure 6. The scatterplot of IBD (Identity by descent) sharing for smaller (x axis) and larger (y axis) IBD segments

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107797.

ACKNOWLEDGMENTS

We are grateful to the volunteers for donating their blood samples. Samples were collected under research supported by National Research Council Sri Lanka, Grant No. 17-042. NR is supported by SERB-CRG/20-21/006762. GC is supported by ICMR ad hoc grants ICMR ad-hoc grants (2021-6389), (2021-11289) and BHU IoE incentive grant BHU (6031). The Open Access Article Processing Charge has been covered by the Institute of Eminence (IoE), Banaras Hindu University, India.

AUTHOR CONTRIBUTIONS

Conceptualization, R.R., N.R., and G.C.; sample collection, PRW. K.H.T., and R.R.; data generation S.K., N.P., and N.R.; formal analysis, P.P.S., S.K., G.C., G.vD.; writing—original draft, P.P.S., S.K., G.vD., and G.C.; writing—review & editing, K.H.T., R.R., and N.R.; supervision, R.R., N.R., and G.C. All authors approved the final draft of the manuscript and take responsibility for its content, including the accuracy of the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 6, 2023
Revised: July 29, 2023
Accepted: August 29, 2023
Published: September 1, 2023

REFERENCES

1. Driem, G.V. (2001). Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region, Containing an Introduction to the Symbiotic Theory of Language (Brill).
2. Government of Śrī Laṅkā (2012). Census of Population and Housing: Population Atlas of Śrī Laṅkā 2012 (Colombo: Department of Census and Statistics, Ministry of Finance and Planning, Government of Śrī Laṅkā).
3. Sarasin, P., and Sarasin, F. (1893). Ergebnisse naturwissenschaftlicher forschungen auf Ceylon, 1 (CW Kreidel).
4. Deraniyagala, S.U. (1992). The Prehistory of Śrī Laṅkā: An Ecological Perspective, Department of Archaeological Survey Memoir, *8* (Department of Archaeological Survey,Government of Śrī Laṅkā), Parts I and II (2 volumes).
5. Kennedy, K.A., Deraniyagala, S.U., Roertgen, W.J., Chiment, J., and Disotell, T. (1987). Upper pleistocene fossil hominids from Sri Lanka. Am. J. Phys. Anthropol. *72*, 441–461. https://doi.org/10.1002/ajpa.1330720405.
6. Deraniyagala, S.U., and Deraniyagala, S.U. (1989). Fossil remains of 28,000-year-old hominids from Sri Lanka. Curr. Anthropol. *30*, 394–399.
7. Perera, N., Kourampas, N., Simpson, I.A., Deraniyagala, S.U., Bulbeck, D., Kamminga, J., Perera, J., Fuller, D.Q., Szabó, K., and Oliveira, N.V. (2011). People of the ancient rainforest: Late Pleistocene foragers at the Batadomba-lena rockshelter, Sri Lanka. J. Hum. Evol. *61*, 254–269. https://doi.org/10.1016/j.jhevol.2011.04.001.
8. van Driem, G.L. (2021). Ethnolinguistic prehistory: The Peopling of The World From the Perspective of Language, Genes and Material Culture*26* (Brill).
9. Geiger, W.L. (1905). Dīpavaṃsa und Mahāvaṃsa und die geschichtliche Überlieferung in Ceylon (A Deichert'sche Verlagsbuchhandlung Nachf. (Georg Böhme)).
10. Geiger, W.L. (1912). The Mahāvamsa: or, The Great Chronicle of Ceylon (Henry Frowde, Oxford University Press).
11. Geiger, W.L. (1900a). Litteratur und Sprache der Singhalesen. Strassburg: Karl J. Trübner.
12. Geiger, W.L. (1900). Máldivische Studien, I.' Sitzungsberichte der Königlichen Bayerischen Akademie der Wissenschaften. Philosophisch-philologische Classe, 641–684.
13. Geiger, W.L. (1901). Máldivische Studien, II.' Sitzungsberichte der Königlichen Bayerischen Akademie der Wissenschaften. Philosophisch-philologische Classe, 371–387.
14. Geiger, W. (1902). Studien III.' Sitzun gsberichte der Königlichen Bayerischen Akademie der Wissenschaften. Philosophisch-philologische Classe, 107–132.
15. Turner, R.L. (1966). A Comparative Dictionary of the Indo-Aryan Languages (Oxford University Press).
16. Parpola, A. (2015). The Roots of Hinduism: The Early Aryans and the Indus Civilization (Oxford University Press).
17. Zvelebil K.V. Tamil. In: Sebeok T.A., editor. Current Trends in Linguistics, Volume 5: Linguistics in South Asia. The Hague; 1969. p. 343–371.
18. Majumdar, R.C. (1927). Ancient Indian Colonies in the Far East, (Vol. I:. Champa) (The Punjab Sanskrit Book Depot).
19. Majumdar, R.C. (1937). Suvarṇadvīpa (Ancient Indian Colonies in the Far East, Vol. II, Part I: Political History) (Asoke Kumar Majumdar).
20. Majumdar, R.C. (1938). Suvarṇadvīpa (Ancient Indian Colonies in the Far East, Vol. II, Part II: Cultural History) (Asoke Kumar Majumdar).
21. Majumdar, R.C. (1944). Hindu Colonies in the Far East (Sures C. Das).
22. Ranaweera, L., Kaewsutthi, S., Win Tun, A., Boonyarit, H., Poolsuwan, S., and Lertrit, P. (2014). Mitochondrial DNA history of Sri Lankan ethnic people: their relations within the island and with the Indian subcontinental populations. J. Hum. Genet. *59*, 28–36. https://doi.org/10.1038/jhg.2013.112.
23. Dissanayake, V.H.W., Weerasekera, L.Y., Gammulla, C.G., and Jayasekara, R.W. (2009). Prevalence of genetic thrombophilic polymorphisms in the Sri Lankan population—implications for association study design and clinical genetic testing services. Exp. Mol. Pathol. *87*, 159–162.

https://doi.org/10.1016/j.yexmp.2009.07.002.
24. Kivisild, T., Rootsi, S., Metspalu, M., Metspalu, E., Parik, J., Kaldma, K., Usanga, E., Mastana, S., Papiha, S.S., and Villems, R. (2003). The genetics of language and farming spread in India. In Examining the Farming/language Dispersal Hypothesis (McDonald Institute Monographs Series, McDonald Institute for Archaeological Research), pp. 215–222.
25. Malavige, G.N., Rostron, T., Seneviratne, S.L., Fernando, S., Sivayogan, S., Wijewickrama, A., and Ogg, G.S. (2007). HLA analysis of Sri Lankan Sinhalese predicts North Indian origin. Int. J. Immunogenet. *34*, 313–315. https://doi.org/10.1111/j.1744-313X.2007.00698.x.
26. Kshatriya, G.K. (1995). Genetic affinities of Sri Lankan populations. Hum. Biol. *67*, 843–866.
27. Papiha, S.S., Mastana, S.S., and Jayasekara, R. (1996). Genetic variation in Sri Lanka. Hum. Biol. *68*, 707–737.
28. Soejima, M., and Koda, Y. (2005). Denaturing high-performance liquid chromatography–based genotyping and genetic variation of FUT2 in Sri Lanka. Transfusion *45*, 1934–1939. https://doi.org/10.1111/j.1537-2995.2005.00651.x.
29. Saha, N. (1988). Blood genetic markers in Sri Lankan populations—reappraisal of the legend of Prince Vijaya. Am. J. Phys. Anthropol. *76*, 217–225. https://doi.org/10.1002/ajpa.1330760210.
30. Illeperuma, R.J., Mohotti, S.N., De Silva, T.M., Fernandopulle, N.D., and Ratnasooriya, W.D. (2009). Genetic profile of 11 autosomal STR loci among the four major ethnic groups in Sri Lanka. Forensic Sci. Int. Genet. *3*, e105–e106. https://doi.org/10.1016/j.fsigen.2008.10.002.
31. Deng, L., Hoh, B.P., Lu, D., Saw, W.Y., Twee-Hee Ong, R., Kasturiratne, A., Janaka de Silva, H., Zilfalil, B.A., Kato, N., Wickremasinghe, A.R., et al. (2015). Dissecting the genetic structure and admixture of four geographical Malay populations. Sci. Rep. *5*, 14375. https://doi.org/10.1038/srep14375.
32. Perera, N., Galhena, G., and Ranawaka, G. (2021). X-chromosomal STR based genetic polymorphisms and demographic history of Sri Lankan ethnicities and their relationship with global populations. Sci. Rep. *11*, 12748. https://doi.org/10.1038/s41598-021-92314-9.
33. Welikala, A.H.J., Ranasinghe, R., Tennekoon, K.H., Kotelawala, J.T., and Weerasooriya, P.R. (2022). Mitochondrial DNA $(CA)_n$ dinucleotide repeat variations in Sinhalese and Vedda populations in Sri Lanka. Genetica *150*, 145–150. https://doi.org/10.1007/s10709-022-00150-0.
34. Ranasinghe, R., Tennekoon, K.H., Karunanayake, E.H., Lembring, M., and Allen, M. (2015). A study of genetic polymorphisms in mitochondrial DNA hypervariable regions I and II of the five major ethnic groups and Vedda population in Sri Lanka. Leg. Med. *17*, 539–546. https://doi.org/10.1016/j.legalmed.2015.05.007.
35. Juneja, R.K., Saha, N., Tay, J.S., Low, P.S., and Gahne, B. (1994). Distribution of plasma alpha-1-B-glycoprotein (A1BG) polymorphism in several populations of the Indian subcontinent. Ann. Hum. Biol. *21*, 443–448. https://doi.org/10.1080/03014469400003462.
36. Kirk, R.L. (1976). The legend of Prince Vijaya—a study of Sinhalese origins. Am. J. Phys. Anthropol. *45*, 91–99.
37. Roychoudhury, A.K., and Nei, M. (1985). Genetic relationships between Indians and their neighboring populations. Hum. Hered. *35*, 201–206. https://doi.org/10.1159/000153545.
38. Mastana, S. (2007). Molecular anthropology: population and forensic genetic applications. Anthropology *3*, 373–383.
39. Fernando, A.S., Wanninayaka, A., Dewage, D., Karunanayake, E.H., Rai, N., Somadeva, R., Tennekoon, K.H., and Ranasinghe, R. (2023). The mitochondrial genomes of two Pre-historic Hunter Gatherers in Sri Lanka. J. Hum. Genet. *68*, 103–105. https://doi.org/10.1038/s10038-022-01099-w.
40. Chaubey, G. (2014). Language isolates and their genetic identity: a commentary on mitochondrial DNA history of Sri Lankan ethnic people: their relations within the island and with the Indian subcontinental populations. J. Hum. Genet. *59*, 61–63. https://doi.org/10.1038/jhg.2013.122.
41. Van Driem, G.L. (2019). The tale of tea: A comprehensive history of tea from prehistoric times to the present day. In The Tale of Tea (Brill).
42. Pathak, A.K., Kadian, A., Kushniarevich, A., Montinaro, F., Mondal, M., Ongaro, L., Singh, M., Kumar, P., Rai, N., Parik, J., et al. (2018). The Genetic Ancestry of Modern Indus Valley Populations from Northwest India. Am. J. Hum. Genet. *103*, 918–929. https://doi.org/10.1016/j.ajhg.2018.10.022.
43. Chaubey, G., Metspalu, M., Choi, Y., Mägi, R., Romero, I.G., Soares, P., van Oven, M., Behar, D.M., Rootsi, S., Hudjashov, G., et al. (2011). Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol. Biol. Evol. *28*, 1013–1024. https://doi.org/10.1093/molbev/msq288.
44. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.
45. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.
46. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.
47. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.
48. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005.
49. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453. https://doi.org/10.1371/journal.pgen.1002453.
50. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096.
51. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. Genetics *193*, 1233–1254. https://doi.org/10.1534/genetics.112.147330.
52. Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics *194*, 459–471. https://doi.org/10.1534/genetics.113.150029.
53. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. *30*, E386–E394. https://doi.org/10.1002/humu.20921.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Sinhala	Field Collection from Śrī Laṅkā	See Figures 2 and 3 for population names
Śrī Laṅkā Tamils (STS)	Field Collection from Śrī Laṅkā	See Figures 2 and 3 for population names
Śrī Laṅkā Tamils (UK)	1000 genomes (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/)	See Figures 2 and 3 for population names
World population data	Pathak et al.[42]	See Figures 2 and 3 for population names
Chemicals, peptides, and recombinant proteins		
Agarose	MERCK	Cat# A9539
Tris EDTA Buffer, Molecular Biology Grade	Fisher Scientific	Cat# AAJ75893AP
Critical commercial assays		
DNA extraction Kit	Qiagen	Cat# 51104
MinElute PCR Purification Kit	Qiagen	Cat# 28006
Illumina-Infinium Global Screening Array	1.0 (GSA-24v1-0)	Cat# 20031669
Deposited data		
The Genotype data of Sinhala	This study	https://doi.org/10.6084/m9.figshare.23975601
The Genotype data of Śrī Laṅkā Tamil	This study	https://doi.org/10.6084/m9.figshare.23975601
Software and algorithms		
PLINK v1.9	Chang et al.[44]	https://www.cog-genomics.org/plink/1.9/
EIGENSOFT v6.1.4	Patterson et al.[45]	https://github.com/DReichLab/EIG
ADMIXTURE	Alexander et al.[46]	https://dalexander.github.io/admixture/
ADMIXTOOL	Patterson et al.[47]	https://github.com/DReichLab/AdmixTools
BEAGLE 5.4	Browning et al.[48]	http://www.gnu.org/licenses/
fineStructure	Lawson et al.[49]	https://people.maths.bris.ac.uk/~madjl/finestructure/fs-2.1.3.tar.gz
MEGA-X	Kumar et al.[50]	https://www.megasoftware.net/show_eua
ChromoPainter	Lawson et al.[49]	https://people.maths.bris.ac.uk/~madjl/finestructure/fs-2.1.3.tar.gz
ALDER	Loh et al.[51]	http://cb.csail.mit.edu/cb/alder/alder_v1.03.tar.gz
runs of homozygosity (RoH)	Chang et al.[44]	https://www.cog-genomics.org/plink/1.9/
merged IBD	Browning & Browning[52]	http://www.gnu.org/licenses/
Refined IBD	Browning & Browning[52]	http://www.gnu.org/licenses
mt-DNA nomenclature	Van Oven & Kayser[53]	Phylotree.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gyaneshwer Chaubey (gyaneshwer.chaubey@bhu.ac.in).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data reported in this paper are publicly available from Figshare repository (https://doi.org/10.6084/m9.figshare.23975601).
- This paper does not report original code.

- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study subjects

The study involve human participation from two major ethnic groups of Śrī Laṅkā. Blood samples (5 mL) were collected in EDTA tubes from thirteen individuals (age 18-60 years), out of which nine belong to Sinhala and four from Śrī Laṅkā Tamil populations of Śrī Laṅkā. The three-generation rule was observed in collecting blood sample; no individuals were each other's blood relatives. Although the sample size is low for the intra-population and genomic selection type studies, they are sufficient for inter-population comparison and understanding the population history. We have used STU code for the Śrī Laṅkā Tamil collected from the UK; STS code for the Śrī Laṅkā Tamil collected from Śrī Laṅkā.

Ethic statement

Ethical approval of the present study was obtained by the Ethics Review Committee, Faculty of Medicine, University of Colombo, Śrī Laṅkā under the approval EC-17-147. Sampling was performed according to the standard guidance given by the ICMR (Indian Council of Medical Research), India and each individual was subjected to interviews and questionnaires, which recorded information such as family, relations and food habitats.

METHOD DETAILS

DNA extraction and genotyping

According to the manufacturer's instructions, DNA was extracted using the Puregene blood kit (Qiagen) at the Birbal Sahni Institute of Palaeosciences (BSIP), Lucknow, India.

Illumina-Infinium Global Screening Array 1.0 (GSA-24v1-0) was performed for all the samples (n=13) collected from Śrī Laṅkā, giving us 618,540 autosomal markers as per Illumina's recommended protocol. Signal intensities detected by the GSA were converted to genotypes using Illumina AutoCall software with a GenCall threshold of 0.15. Primary quality control requirements included per-sample log R standard deviation (SD) less than 0.25 and call rates greater than 98.5% across the array (GSA-24 v1.0), or greater than 99.0% across the autosomes and chromosome X.

Data processing and population genetic analyses

The variant calling factor (vcf) file was converted to binary files with PLINK v1.9 following optimal conditions of quality filtering like, –maf 0.03, geno 0.03 and mind 0.03. The filtered samples were merged with the HGDP Panel. We found 255,063 SNPs common between the samples and panel with a 0.9987 genotyping rate. PLINK v1.9[44] was used for data curation and management for the statistical analyses. The PC1 and PC2 eigenvectors in Principal Component Analysis (PCA) were generated with smartpca[45] (EIGEN v6.1.4), and the plot has been generated with an in-house R script. We used ADMIXTURE[46] to further estimate shared ancestry (K=2 to K=15), and at K=10 the ancestry has been defined with minimum cv error value of 0.5423.

mtDNA analysis

We collected data from published sources for mtDNA comparison and reclassified them in their respective haplogroups manually following the latest nomenclature (Phylotree.org). The regional classification (East Eurasian, South Asian and West Eurasian) of the mtDNA haplogroups was performed manually according to the presence of a particular haplogroup in that region.

QUANTIFICATION AND STATISTICAL ANALYSIS

To understand population relationships, several *f*-statistics were performed in default setting using the Yoruba population as an outgroup population. To know the shared drift and gene flow pattern we used *f*3 and *f*4 statistics, respectively from the ADMIXTOOL package.[47] We have phased the genotypic data with beagle 5.4[48] with default settings. Later the haplotype-based analysis was performed using MCMC algorithm-based software i.e., fineStructure[49] using likelihood modelling approaches to calculate matrices. The obtained output matrix was used for construction of MCMC tree using MEGA-X.[50] ChromoPainter[49] was applied for the estimation of chunk counts donated by reference populations to our targeted population. ALDER[51] was run to understand the admixing time using multiple source populations with default settings. In order to understand the population dynamics, the runs of homozygosity (RoH) was determined for each population using PLINK 1.9[44] The analysis was carried out with the use of the 'homozyg' function and utilised 1000 kb windows for the calculations, allowing one heterozygous call and five missing calls per window, and a minimum of 100 SNPs per window. Every person is successively scanned by the selected window, which estimates the proportion in a homozygous window for each SNP. For understanding the Identity by Descent (IBD) we used refined and merged IBD analysis.[52]