HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2017 July 07.

Published in final edited form as:

Science. 2013 August 02; 341(6145): 565–569. doi:10.1126/science.1237947.

Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny

Paolo Francalacci[1,*], **Laura Morelli**[1,†], **Andrea Angius**[2,3], **Riccardo Berutti**[3,4], **Frederic Reinier**[3], **Rossano Atzeni**[3], **Rosella Pilu**[2], **Fabio Busonero**[2,5], **Andrea Maschio**[2,5], **Ilenia Zara**[3], **Daria Sanna**[1], **Antonella Useli**[1], **Maria Francesca Urru**[3], **Marco Marcelli**[3], **Roberto Cusano**[3], **Manuela Oppo**[3], **Magdalena Zoledziewska**[2,4], **Maristella Pitzalis**[2,4], **Francesca Deidda**[2,4], **Eleonora Porcu**[2,4,5], **Fausto Poddie**[4], **Hyun Min Kang**[5], **Robert Lyons**[6], **Brendan Tarrier**[6], **Jennifer Bragg Gresham**[6], **Bingshan Li**[7], **Sergio Tofanelli**[8], **Santos Alonso**[9], **Mariano Dei**[2], **Sandra Lai**[2], **Antonella Mulas**[2], **Michael B. Whalen**[2], **Sergio Uzzau**[4,10], **Chris Jones**[3], **David Schlessinger**[11], **Gonçalo R. Abecasis**[5], **Serena Sanna**[2], **Carlo Sidore**[2,4,5], and **Francesco Cucca**[2,4,*]

[1]Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy.

[2]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy.

[3]Center for Advanced Studies, Research and Development in Sardinia (CRS4), Pula, Italy.

[4]Dipartimento di Scienze Biomediche, Università di Sassari, 07100 Sassari, Italy.

[5]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

[6]DNA Sequencing Core, University of Michigan, Ann Arbor, MI 48109, USA.

[7]Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37235, USA.

[8]Dipartimento di Biologia, Università di Pisa, 56126 Pisa, Italy.

[9]Departamento de Genética, Antropología Física y Fisiología Animal, Universidad del País Vasco/Euskal Herriko Unibertsitatea, 48080 Bilbao, Spain.

[10]Porto Conte Ricerche, Località Tramariglio, Alghero, 07041 Sassari, Italy.

[11]Laboratory of Genetics, National Institute on Aging, Baltimore, MD 21224, USA.

Abstract

[*]Corresponding author. pfrancalacci@uniss.it (P.F.); fcucca@uniss.it (F.C.).
[†]Laura Morelli prematurely passed away on 20 February 2013. This work is dedicated to her memory.

Supplementary Materials

www.sciencemag.org/cgi/content/full/341/6145/565/DC1

Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S3
References (*20*–*37*)

Genetic variation within the male-specific portion of the Y chromosome (MSY) can clarify the origins of contemporary populations, but previous studies were hampered by partial genetic information. Population sequencing of 1204 Sardinian males identified 11,763 MSY single-nucleotide polymorphisms, 6751 of which have not previously been observed. We constructed a MSY phylogenetic tree containing all main haplogroups found in Europe, along with many Sardinian-specific lineage clusters within each haplogroup. The tree was calibrated with archaeological data from the initial expansion of the Sardinian population ~7700 years ago. The ages of nodes highlight different genetic strata in Sardinia and reveal the presumptive timing of coalescence with other human populations. We calculate a putative age for coalescence of ~180,000 to 200,000 years ago, which is consistent with previous mitochondrial DNA–based estimates.

New sequencing technologies have provided genomic data sets that can reconstruct past events in human evolution more accurately (1). Sequencing data from the male-specific portion of the Y chromosome (MSY) (2), because of its lack of recombination and low mutation, reversion, and recurrence rates, can be particularly informative for these evolutionary analyses (3, 4). Recently, high-coverage Y chromosome sequencing data from 36males from different worldwide populations (5) assessed 6662 phylogenetically informative variants and estimated the timing of past events, including a putative coalescence time for modern humans of ~101,000 to 115,000 years ago.

MSY sequencing data reported to date still represent a relatively small number of individuals from a few populations. Furthermore, dating estimates are also affected by the calibration of the phylogenetic tree used to establish the rate of molecular change over time. This calibration can either correlate the number of nucleotide substitutions with dates from paleontological/archaeological records (phylogenetic rate) or can use directly observed de novo mutations in present-day families (mutation rate). However, both approaches are complicated by several variables (6, 7).

Some of these problems can be resolved by the analysis of MSY sequencing data from many individuals from a genetically informative population, regarding which archaeological data are available to provide suitable calibration points. This prompted us to use large-scale MSY sequencing data from the island population of Sardinia for phylogenetic analysis. We generated a high-resolution analysis of the MSY from population sequencing of 1204 Sardinian males (8). We used a hierarchical approach and, to be consistent with previous work (5), focused on approximately 8.97 mega–base pairs (Mbp) from the Y chromosome in the X-degenerated region. We inferred 11,763 MSY phylogenetically informative single-nucleotide polymorphisms (SNPs), detected in at least two individuals and unequivocally associated with specific haplogroups and sub-haplogroups; 6751 of these SNPs had not thus far been reported in public databases.

The informative SNPs were used to construct a parsimony-based phylogenetic tree. To root the tree, we used the chimpanzee genome reference as an outgroup and inferred the ancestral status at all SNP sites except for 26 that were discarded in further analysis. The first bifurcation point, and thus the most recent common ancestor, separates samples 1 to 7 from the rest of the samples (samples 8 to 1204) (Table 1). The average number of derived alleles

in the 1204 males is 1002.6 (±21.2 SD) which, consistent with a neutral evolution of these Y polymorphisms, shows a remarkable uniformity of the branch length.

The Sardinian sequences show a very high degree of inter-individual variation. As shown in a schematic tree (Fig. 1), all of the most common Y-chromosome haplogroups previously detected in Europe are present in our sample (Table 1), with the sole exception of the northernmost Uralic haplogroup N. The first bifurcation separates the mostly sub-Saharan haplogroup A (7 individuals, 0.6%in our sample) from the others. Haplogroup E (132, 11.0%) is present with its European clade, characterized by the presence of the M35 marker, together with a small number of individuals belonging to the mainly African clade E1a. The rare haplogroup F (7, 0.6%) is related to haplogroup G (131, 10.9%), which shows a private Sardinian- Corsican clade whose ancient roots have been found in an Eneolithic sample from the Italian Alps (9). Haplogroup I (490, 40.7%) is of special interest because it is mostly represented by the I2a1a clade, identified by theM26 marker, which is at high frequencies in Sardinia (10) but is rare or absent elsewhere (11). Haplogroup J (161, 13.4%) is observed with its main subgroups; and the super-haplogroup K is present with the related L and T branches (36, 3.0%), with a single individual of haplogroup Q (1, 0.08%) and with the more common haplogroup R (239, 19.9%) occurring mostly as the western European M173-M269 branch.

Almost half of the discovered SNPs (4872) make up the skeleton of the phylogenetic tree and constitute the root of the main clades. The skeleton includes lineages that are unbranched for most of their length, with ramifications only in the terminal portion. This indicates an early separation of the clades, followed by new variability generated during subsequent expansion events.

To estimate points of divergence between Sardinian and continental clades, we sequenced two samples from the Basque Country and northern Italy, belonging to haplogroup I, and two, from Tuscany and Corsica, belonging to haplogroup G. We also analyzed the sequence of the so-called Iceman Ötzi (9), together with 133 publicly available European sequences from the 1000 Genomes Project database and those SNPs from the International Society of Genetic Genealogy (ISOGG) database detected outside Sardinia.

The Basque individual separates from the basal position of the I2a1a branch that encompasses 11 Sardinian individuals. The northern Italian sample, instead, most likely reflecting the last step of I2a1 lineages before their arrival in Sardinia, is at the basal point of most of the remaining I2a1a samples (Fig. 2). Considering two other basal lineages encompassing only Sardinian samples, we can infer that when the I2a1a sub-haplogroup entered Sardinia, it had already differentiated into four founder lineages that then accumulated private Sardinian variability. Two other founder clades show similar divergence after entry into the island: one belonging to haplogroup R1b1c (xV35) (whose differentiation is identified contrasting the Sardinian data with the ISOGG and 1000 Genome data), and the other to haplogroup G2a2b-L166 (identified by divergence from a sequenced Corsican sample).

The branch length uniformity observed in our phylogeny is consistent (Fig. 1) with a relatively constant accumulation of SNPs in different lineages over time. Hence, this accumulation can be effectively used as a molecular clock for the dating of branch points. We calibrated the accumulation of Sardinian-specific genetic variation against established Sardinian archaeological records indicating a putative age of initial demographic expansion ~7700 years ago [reviewed in figs. S7 and S8 and the supplementary text (8, 12)] that is also supported by mitochondrial DNA (mtDNA) analyses (13). Comparison of Sardinian genetic variation with that found elsewhere helped us to establish the amount of variability produced during and after this expansion, resulting in sublineages that appear to be unique to the island.

We focused our calibration analyses on the individuals belonging to the I2a1a-δ clade, which is shared by 435 individuals and is best suited to assess the Sardinian specific variability. Taking into account the average variation of all Sardinian individuals in the common I2a1a-δ clade of 37.3 (±7.8) SNPs, a calibration point of 7700 years ago results in a phylogenetic rate of one new mutation every 205 (±50) years. Considering that our analysis focused on approximately 8.97 Mbp of sequence from the Y chromosome X-degenerated region, this rate is equivalent to 0.53×10^{-9} bp^{-1} $year^{-1}$. This phylogenetic rate is consistent with the value of 0.617 (0.439 to 0.707) $\times 10^{-9}$ bp^{-1} $year^{-1}$ from the genomewide mutation rate observed from de novo mutations adjusted for Y chromosome–specific variables (14). Our mutation rate is instead lower than the value of 1.0×10^{-9} bp^{-1} $year^{-1}$ obtained from de novo MSY mutations in a single deep-rooted family (5), which also coincides with that traditionally deduced from the *Homo-Pongo* divergence (15).

Using our phylogenetic rate of 0.53×10^{-9} bp^{-1} $year^{-1}$, we estimated the time to the most recent common ancestor (MRCA) of all samples, whose average variability is 1002.6 (±21.2) SNPs, at ~200,000 years ago. This is older than previously proposed (16) for the Y chromosome but is in agreement with estimates from a de novo mutation rate in an African Y-chromosome lineage (14) and with the revised molecular clock for humans (7) and the TMRCA estimated from analyses of maternally inherited mtDNA (13, 17).

The main non-African super-haplogroup F-R shows an average variation of 534.8 (±28.7) SNPs, corresponding to a MRCA of ~110,000 years ago, in agreement with fossil remains of archaic *Homo sapiens* out of Africa (7, 18) though not with mtDNA, whose M and N super-haplogroups coalesce at a younger age (13). The main European subclades show a differentiation predating the peopling of Sardinia, with an average variation ranging from 70 to 120 SNPs (Table 1), corresponding to a coalescent age between 14,000 and 24,000 years ago, which is compatible with the postglacial peopling of Europe.

However, the inferred phylogenetic rate and dating estimates presented here remain tentative, because the calibration date was deduced from archaeological data, which may be incomplete and typically covers a relatively large temporal interval. In the future, a more precise calibration point might be obtained by sequencing ancient DNAs from prehistoric Sardinian remains dated by radiocarbon methods. Further limitations derive from the scarcity of related samples for rare lineages, coupled with the low-pass sequencing approach we used (8). Low-pass sequencing is expected to detect nearly all common variants

(frequency >1%) but to miss rare variants. Missed variants have competing effects on estimates of ancestral coalescent times: When they lead to missed differences among haplotypes that diverged after the founding of Sardinia, they lower our calibrated estimates of mutation rate and increase coalescent time estimates; when they lead to missed differences among ancestral clades, they lower these time estimates. In fact, despite the overall homogeneity of the length of the branches from the MRCA (Fig. 1), those represented by fewer individuals are generally shorter (8). To estimate the effect of missed variants on the age estimates, we sequenced with deep coverage 7 selected individuals, 4 of them belonging to the I2a1a-δ clade, used for calibration, and 3 to the I2a1a-β, J2b2f and A1b1b2b clades. The deep sequencing of the I2a1a-δ samples yielded an average of 45.7 (±2.2) Sardinian-specific SNPs among these haplotypes [versus 37.3 (±7.8) in low-coverage data], corresponding to a phylogenetic rate of 0.65×10^{-9} bp^{-1} $year^{-1}$. Overall, this reanalysis suggested a slightly more recent MRCA (~8% lower), still in substantial agreement with the antiquity of the main Y-chromosome haplogroups (8).

Hence, despite current limitations, the calibration used from common haplogroups in over 1000 people from this isolated population, including many island-specific SNPs, permits an estimate of main demographic events during the peopling of Sardinia that is concordant with the archaeological/historical record and ancient DNA analysis (8). The initial expansion of the Sardinian population, used for calibration, is marked by six clades belonging to three different haplogroups, with an average variation of around 35 to 40 SNPs, representing the ancient founder core of modern Sardinians.

Our data further suggest a more intricate scenario of Sardinian demographic history. Specifically, clades of E, R, and G that show Sardinian-specific variability of 25 to 30 SNPs are consistent with further expansion in the Late Neolithic (~5500 to 6000 years ago) (Table 1). Additional variation putatively arrived with groups of individuals carrying other haplogroups (namely the I clades different from I2a1, J, and T). Taken together, the genetic data and demographic expansions are consistent with classical archaeological data indicating that Sardinia reached a considerable population size in prehistoric times; the estimated population during the Nuragic Period (~2500 to 3700 years ago) was >300,000 inhabitants (19). Finally, the rare, mostly African A1b-M13 and E1a-M44 clades could have come to Sardinia in more recent times, up to the historic period corresponding to the Roman and Vandalic dominations, suggested by a private Sardinian variability of 7 to 10 SNPs.

Acknowledgments

This research was supported in part by NIH contract NO1-AG-1-2109 from the National Institute of Aging to the IRGB institute; by Sardinian Autonomous Region (L.R. no. 7/2009) grants cRP3-154 to F.C. and cRP2-597 to P.F.; by Fondazione Banco di Sardegna grants to P.F. and L.M.; by Basque Government grant G.I.C. IT-542-10 to S.A.; and by National Human Genome Research Institute grants HG005581, HG005552, HG006513, and HG007022 to G.R.A. We are grateful to all the Sardinian donors for providing blood samples. We also thank the 1000 Genomes Project consortium for making available their sequencing data, in compliance with the Fort Lauderdale principles. We thank H. Skaletsky for detailed information about sequence blocks on the Y chromosome; M. Uda and R. Nagaraja for useful comments; C. Calò, D. Luiselli, and C. de la Rua for providing some non-Sardinian samples; E. Garau, M. Rendeli, A. Moravetti, and B. Wilkens for the archaeological background; and the CRS4 HPC group for their IT support, in particular L. Leoni and C. Podda. Genotype data have been deposited at the European Genome-phenome Archive (EGA, www.ebi.ac.uk/ega/), which is hosted by the European Bioinformatics Institute, under accession number EGAS00001000532.

References

1. 1000 Genomes Project Consortium. Nature. 2012; 491:56–65. [PubMed: 23128226]
2. Skaletsky H, et al. Nature. 2003; 423:825–837. [PubMed: 12815422]
3. Underhill PA, et al. Nat. Genet. 2000; 26:358–361. [PubMed: 11062480]
4. Semino O, et al. Science. 2000; 290:1155–1159. [PubMed: 11073453]
5. Wei W, et al. Genome Res. 2013; 23:388–395. [PubMed: 23038768]
6. Scally A, Durbin R. Nat. Rev. Genet. 2012; 13:745–753. [PubMed: 22965354]
7. Gibbons A. Science. 2012; 338:189–191. [PubMed: 23066056]
8. Materials and methods are available as supplementary materials on *Science* Online.
9. Keller A, et al. Nature Commun. 2012; 3:698. [PubMed: 22426219]
10. Francalacci P, et al. Am. J. Phys. Anthropol. 2003; 121:270–279. [PubMed: 12772214]
11. Rootsi S, et al. Am. J. Hum. Genet. 2004; 75:128–137. [PubMed: 15162323]
12. Tykot, RH. Radiocarbon Dating and Italian Prehistory. Skeates, R., Withehouse, R., editors. Accordia Specialist Studies on Italy; London: 1994. p. 115-145.
13. Olivieri A, et al. Science. 2006; 314:1767–1770. [PubMed: 17170302]
14. Mendez FL, et al. Am. J. Hum. Genet. 2013; 92:454–459. [PubMed: 23453668]
15. Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Genome Res. 2011; 21:349–356. [PubMed: 21270173]
16. Cruciani F, et al. Am. J. Hum. Genet. 2011; 88:814–818. [PubMed: 21601174]
17. Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Nature. 2000; 408:708–713. [PubMed: 11130070]
18. Armitage SJ, et al. Science. 2011; 331:453–456. [PubMed: 21273486]
19. Lilliu, G. La Civiltà Nuragica. Delfino, Sassari; Italy: 1982.

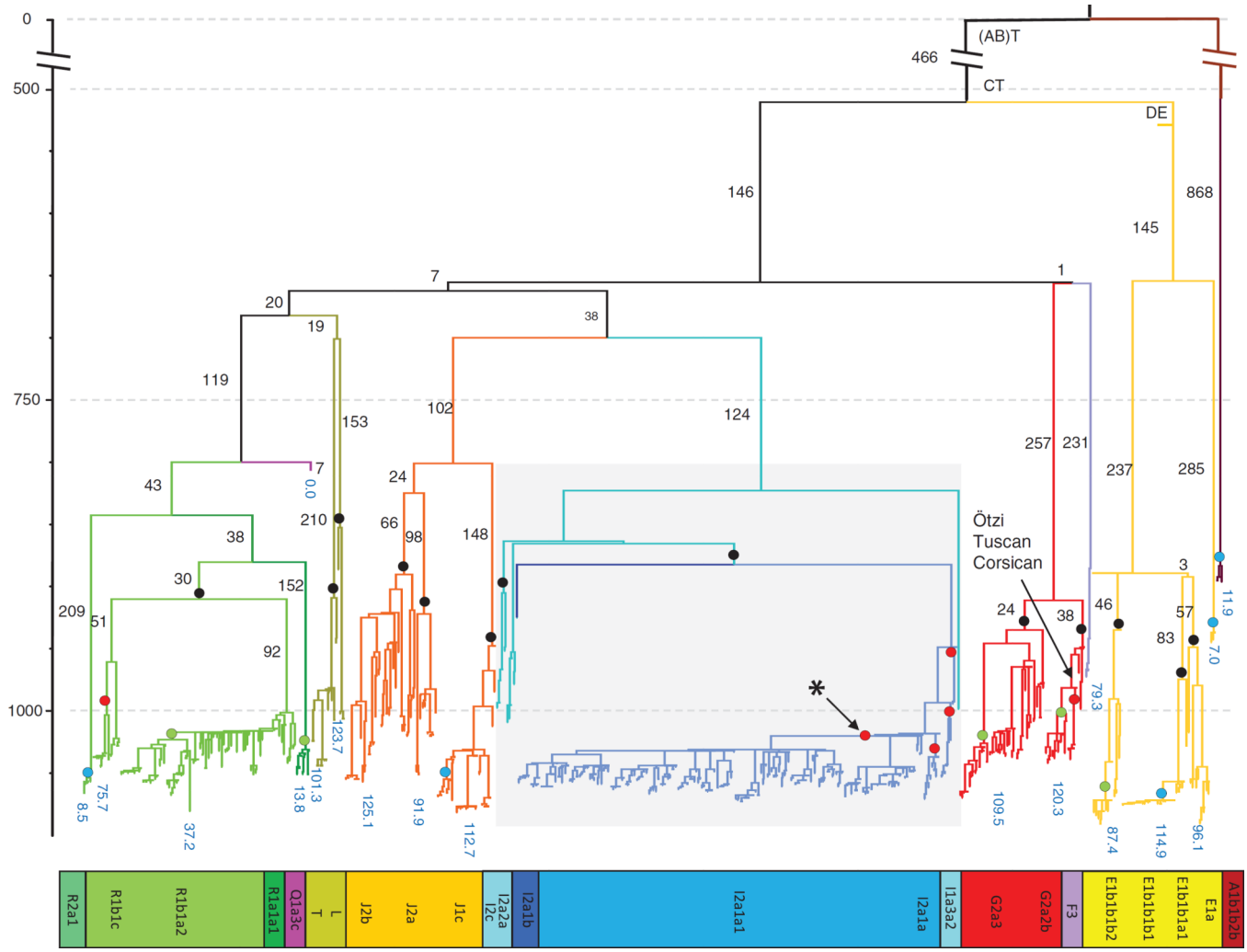

Fig. 1. Phylogenetic tree of the 1209 (1204 Sardinians and 5 non-Sardinians) Y-chromosome sequences

The bifurcations AT, BT, CT, and DE have been inferred because of the absence of individuals belonging to haplogroups B, C, and D in our sample. Colored branches represent different Y-chromosome haplotypes. The number of polymorphisms for the main branches is shown in black; the average number of SNPs of sub-haplogroups is given in blue. The sub-haplogroups are named according to ISOGG nomenclature. The left axis indicates the number of SNPs from the root. The asterisk indicates the calibration point. The colored dots indicate private Sardinian clusters with an average number of SNPs in the range of 35 to 40 in red, 25 to 30 in green, and 7 to 12 in blue. The black dots indicate clusters with an average number of SNPs in the range of 70 to 120. The arrow indicates the position on the tree of the Ötzi, Tuscan, and Corsican samples. The gray box is enlarged in Fig. 2.

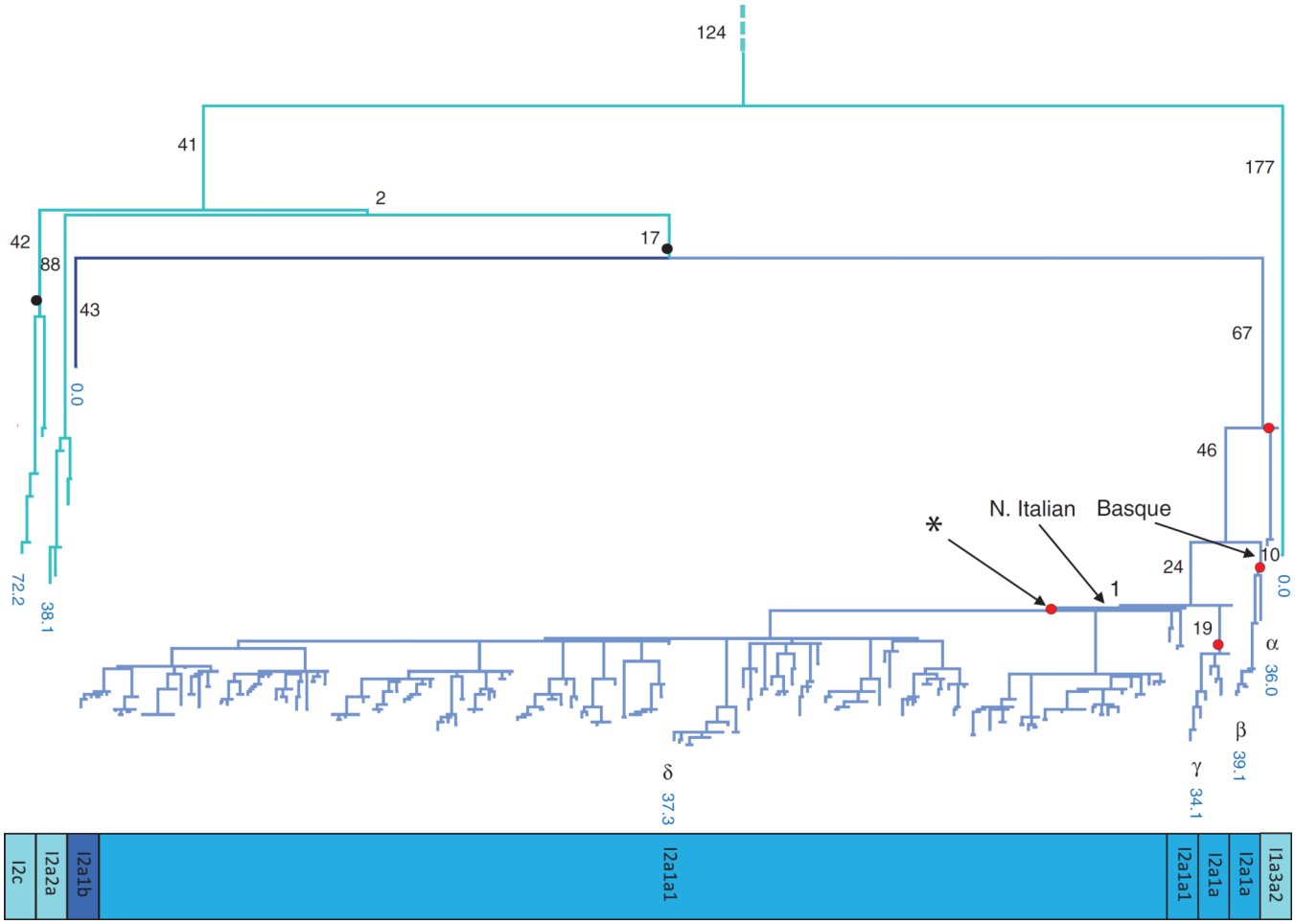

Fig. 2. Phylogenetic tree of the 492 (490 Sardinians and 2 non-Sardinians) Y-chromosome sequences belonging to haplogroup I

The number of polymorphisms for the main branches is shown in black; the average number of SNPs of sub-haplogroups is shown in blue. The sub-haplogroups are named according to ISOGG nomenclature. The red dots indicate Sardinian private clades, labeled in Greek letters as in Table 1. The black dots indicate clusters with an average number of SNPs in the range of 70 to 120. The arrows indicate the position of the northern Italian and Basque samples on the tree. The asterisk indicates the calibration point.

Science. Author manuscript; available in PMC 2017 July 07.

Table 1

Super-haplogroups, haplogroups, sub-haplogroups, and private Sardinian-Corsican clades

Here the average number of SNPs defining each class is shown in our 1209 samples.

Super-haplogroup (individual no.)	Mean SNPs	Haplogroup (individual no.)	Mean SNPs	Sub-haplogroup (individual no.)	Mean SNPs	Private Sardinian clade (individual no.)	Mean SNPs
A-R (1–1204; OTCBI)	1002.6	A (1–7)	879.9 *	A1b1b2b (1–7)	11.9	α (1–7)	11.9
				E1a1 (8–13)	7.0	α (8–13)	7.0
				E1b1b1a1 (14–45)	87.4		
E-R (8–1204; OTCBI)		E (8–139)	541.8 *	E1b1b1b1 (46–115)	96.1	β (49–115)	15.6
				E1b1b1b2 (116–139)	114.9	γ (116–131)	25.8
		F (140–146)	299.0	F3 (140–146)	79.3		
F-R (140–1204; OTCBI)	534.8	G (147–277; OTC)	373.8	G2a2b (147–186; OTC)	109.5	α (C; 155–162)	42.8
						β (163–186)	29.4
				G2a3 (187–277)	120.3	γ (247–277)	25.0
				I1a3a2 (278–279)	0.0		
				I2a1a (280–744; BI)	106.2	α (280–285)	36.0
						β (286–296)	39.1
						γ (297–314)	34.1
		I (278–767; BI)	353.5			δ (315–744)	37.3
I-J (278–928; BI)	387.0			I2a1b (745–746)	0.0		
				I2a2a (747–756)	38.1		
				I2c (757–767)	72.2		
				J1c (768–830)	112.7	α (816–830)	11.0
		J (768–928)	334.3	J2a (831–905)	125.1		
				J2b (906–928)	91.9		
K-R (929–1204)	375.3	K (929–964)	324.9	L (929–936)	123.7		
				T (937–964)	101.3		
		P (965–1204)	359.1	Q1a3c (965)	0.0		
				R1a1a1 (966–980)	13.8		
		R (966–1204)	241.2	R1b1a2 (981–1165)	37.2	α (981–989)	23.0
						β (991–1165)	29.4

Super-haplogroup (individual no.)	Mean SNPs	Haplogroup (individual no.)	Mean SNPs	Sub-haplogroup (individual no.)	Mean SNPs	Private Sardinian clade (individual no.)	Mean SNPs
				R1b1c (1166–1194)	75.7	γ (1177–1194)	36.2
				R2a1 (1195–1204)	8.5	δ (1195–1204)	8.5

The asterisk (*) denotes that the average number of SNPs for haplogroups A and E cannot be determined with precision because of the lack in our sample of individuals belonging to haplogroups B, C, and D. Consequently, the number reported here is an overestimate. The Sardinian samples are progressively numbered from 1 to 1204, and the non-Sardinian samples are labeled as follows: O, Ötzi; T, Tuscan; B, Basque; C, Corsican; I, northern Italian. The clades containing only private Sardinian SNPs are indicated in Greek letters (progressively from α to δ within each haplogroup).