nature COMMUNICATIONS

ARTICLE

Received 28 Jun 2013 | Accepted 13 Nov 2013 | Published 17 Dec 2013

DOI: 10.1038/ncomms3928 OPEN

Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites

Siiri Rootsi[1,*], Doron M. Behar[1,2,*], Mari Järve[1], Alice A. Lin[3], Natalie M. Myres[4], Ben Passarelli[5], G. David Poznik[6], Shay Tzur[2], Hovhannes Sahakyan[1,7], Ajai Kumar Pathak[1], Saharon Rosset[8], Mait Metspalu[1], Viola Grugni[9], Ornella Semino[9,10], Ene Metspalu[1], Carlos D. Bustamante[11], Karl Skorecki[2,12], Richard Villems[1,13], Toomas Kivisild[14] & Peter A. Underhill[11]

Previous Y-chromosome studies have demonstrated that Ashkenazi Levites, members of a paternally inherited Jewish priestly caste, display a distinctive founder event within R1a, the most prevalent Y-chromosome haplogroup in Eastern Europe. Here we report the analysis of 16 whole R1 sequences and show that a set of 19 unique nucleotide substitutions defines the Ashkenazi R1a lineage. While our survey of one of these, M582, in 2,834 R1a samples reveals its absence in 922 Eastern Europeans, we show it is present in all sampled R1a Ashkenazi Levites, as well as in 33.8% of other R1a Ashkenazi Jewish males and 5.9% of 303 R1a Near Eastern males, where it shows considerably higher diversity. Moreover, the M582 lineage also occurs at low frequencies in non-Ashkenazi Jewish populations. In contrast to the previously suggested Eastern European origin for Ashkenazi Levites, the current data are indicative of a geographic source of the Levite founder lineage in the Near East and its likely presence among pre-Diaspora Hebrews.

[1] Estonian Biocentre and Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia. [2] Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa 31096, Israel. [3] Department of Psychiatry, Stanford University, Stanford, California 94305, USA. [4] Ancestry.com DNA, Provo, Utah 84604, USA. [5] Department of Bioengineering, Stanford University, Stanford, California 94305, USA. [6] Program in Biomedical Informatics and Department of Statistics, Stanford University, Stanford, California 94305, USA. [7] Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Yerevan 0014, Armenia. [8] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel. [9] Dipartimento di Biologia e Biotecnologie 'Lazzaro Spallanzani', Università di Pavia, Pavia 27100, Italy. [10] Centro Interdipartimentale 'Studi di Genere', Università di Pavia, Pavia 27100, Italy. [11] Department of Genetics, Stanford University, Stanford, California 94305, USA. [12] Ruth and Bruce Rappaport Faculty of Medicine and Research Institute, Technion-Israel Institute of Technology, Haifa 31096, Israel. [13] Estonian Academy of Sciences, Tallinn 10130, Estonia. [14] Division of Biological Anthropology, University of Cambridge, CB2 3QG Cambridge, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to D.M.B. (email: d_behar@rambam.health.gov.il).

Whole Y-chromosome sequencing holds the promise of stretching the human paternal phylogeny to its maximal resolution. The absence of recombination enables all Y-chromosome sequences to be placed within a single phylogenetic tree. Clearly, a single locus hierarchy may oversimplify the demographic history of a particular individual. However, in contrast to the autosomes, the ordering of the accumulated sequence variants since the most recent common ancestor is preserved. Due to this molecular encapsulation of human male demographic history, the human Y-chromosome phylogeny has become one of the pillars of archaeogenetics and is particularly informative for testing hypotheses based on genealogical information[1,2]. While some Y-chromosome haplogroups show pronounced affiliation to a single continent[3–6], many others, particularly those within the broadly defined Old World (Eurasia), have a wide cross-continental range of distribution[7,8]. Accordingly, paternal population genetic heritage cannot rely solely on the presence of a given list of haplogroups within the populations of interest, but must consider the haplogroup frequencies comprising the respective paternal gene pools[9,10]. At the individual or paternal pedigree level, this means that haplogroups shared between populations do not confirm an unequivocal paternal ancestry to a particular population within the haplogroup's geographic range of distribution. To distinguish between two populations, an informative marker is one that approaches fixation in one group and is absent in another[11]. Such markers should be forthcoming from whole-genome sequences, to be followed by genotyping in larger panels to more precisely delineate patterns of restricted geographic and/or population specificity. Here, we describe the identification of such a Y-chromosome marker from whole Y-chromosome sequencing, which, despite its overall low frequency across Eurasia, coincides with a restricted set of populations within the Near East and among Jews.

The widespread Eurasian distribution zone of haplogroup R1a has been mapped for over a decade[8,10,12,13] since the first reports that the M17 deletion grouped Y-chromosomes into a monophyletic clade that exceeded 30% frequency in some Asian populations[14] and over 50% in Eastern Europe[15]. Several single nucleotide polymorphisms (SNPs) phylogenetically equivalent to M17, such as M198, are now known[16]. Nonetheless, the internal structure of the R1a haplogroup has long resisted fractionation until the recent separation of its post-glacial European and Asian coancestry was partially achieved with the discovery of the M458 SNP[16]. Efforts to clarify the genetic ancestry of East European Jews have been ongoing[17]. Progress towards disentangling the internal structure of haplogroup R1a is integral to our ability to resolve the debate over the geographic origin of the Ashkenazi Jewish Levite founding lineage, which is defined by a short tandem repeat (STR)-based haplotype that is rare elsewhere[18].

The Jewish people can be categorized in a number of ways. For example, one arrangement is into Ashkenazi[19] and non-Ashkenazi[20], which separates Jews recently descended from Central and Eastern Europe from other Jewish Diaspora communities. In addition, Jewish heritage recognizes three paternally inherited castes, Cohen, Levite and Israelite, of which the first two are considered priesthood lineages[18,21–23]. Levites comprise ~4% of the male Jewish population and display genetic evidence for multiple recent origins, with Ashkenazi Levites reported to carry a particularly high frequency (>50%) of a distinctive STR-based lineage nested within haplogroup R1a-M198 (refs 18,24,25). Previous studies have demonstrated that haplogroup R1a is rare in other Jewish castes, other Jewish communities and in non-Jewish groups of Near Eastern origin[18,21]. However, it is found at high frequency in populations of eastern European origin[16,18,26,27]. The greatly elevated frequency of haplogroup R1a-M198 within Ashkenazi Levites, their compact network of STR haplotypes and the recent coalescence time, suggest a founder event specific to the Ashkenazi Levites and a paternal ancestor shared by more than half of contemporary Ashkenazi Levites. The long residence of Ashkenazi Jews in Eastern Europe and the high frequency of haplogroup R1a in the same region suggested that the founder might be of non-Jewish European ancestry, whose descendants were able to assume Levite status. However, because of the paucity of distinctive internal substructure of haplogroup R1a, it was not possible to suggest a particular European source population nor to test the hypothesis of a Turkic-speaking Khazar ancestor, which has been proposed in light of the narrative that members of the Khazar ruling class may have converted to Judaism in the 8th or 9th century[18].

Here we study the phylogenetic origin of the Ashkenazi Levite lineage at the whole Y-chromosome level. First, to allow the appropriate phylogenetic context and depth, we reconstruct the phylogeny of the parental haplogroup R1 using a total of nine R1a and seven R1b whole Y-chromosome sequences. Next, we launch a phylogenetically informative large-scale population-based study in order to resolve the geographic origin of the Ashkenazi Levite founding lineage within the transcontinental distribution range of haplogroup R1a. We find that the set of nucleotide substitutions that characterizes 91.9% of Ashkenazi R1a lineages occurs at low frequencies in non-Ashkenazi Jewish and Near Eastern populations while being absent in Eastern Europe host populations, which were previously considered to be the most likely source for the majority of Ashkenazi Levite Y-chromosome lineages.

Results

Haplogroup R1 whole Y-chromosome phylogeny. We have determined at high coverage the sequence of eight Jewish and five non-Jewish Y-chromosomes (Supplementary Table S1) belonging to haplogroup R1 (Fig. 1a and Supplementary Fig. S1). We reconstructed the phylogenetic relationships of the newly sequenced samples in the context of three other available high coverage sequences from this haplogroup using binary SNP data (Supplementary Data 1) from a total of 8.97 Mbp of sequence from the non-recombining region of nine Y-chromosomes[28]. For both sub-clades of R1, most Jewish sequences cluster separately from the clades that in previous studies have been shown to be common in Europe: in R1b where more than 90% of European lineages are within the L11/L52 clade[29], three of the newly sequenced Jewish Y-chromosomes cluster within its copanion clade defined by the Z2105 mutation; in R1a where the majority of European individuals belong to the S198 (ref. 30) and M458 branches[16], all four newly sequenced Jewish Y-chromosomes fall into the clade defined by Z94 mutation. Furthermore, two Ashkenazi R1a chromosomes were found to share 19 SNPs (Supplementary Fig. S1) of which six were found to be shared with an Iberian individual from the 1,000 Genomes database. One of the mutations defining this clade including Ashkenazi and Iberian samples, which we designate M582 (AKA CTS2253), was subjected to genotyping in 2,834 R1a samples to determine its geographic distribution, diversity and divergence times.

Population-based survey for R1a-M582. Among non-Jewish populations, the overall frequency of R1a-M582 was found to be 0.15% (22/15,138) and among R1a-M198 it was 0.81% (22/2,711) (Table 1). The geographic distribution of the haplogroup appears to be limited to West Eurasia, as we did not observe it in South Asia, Central Asia or Southern Siberia. Haplogroup R1a-M582 was only sporadically observed in Europe, the Diaspora residence of Ashkenazi Jews. Notably, it was not identified among 2,149

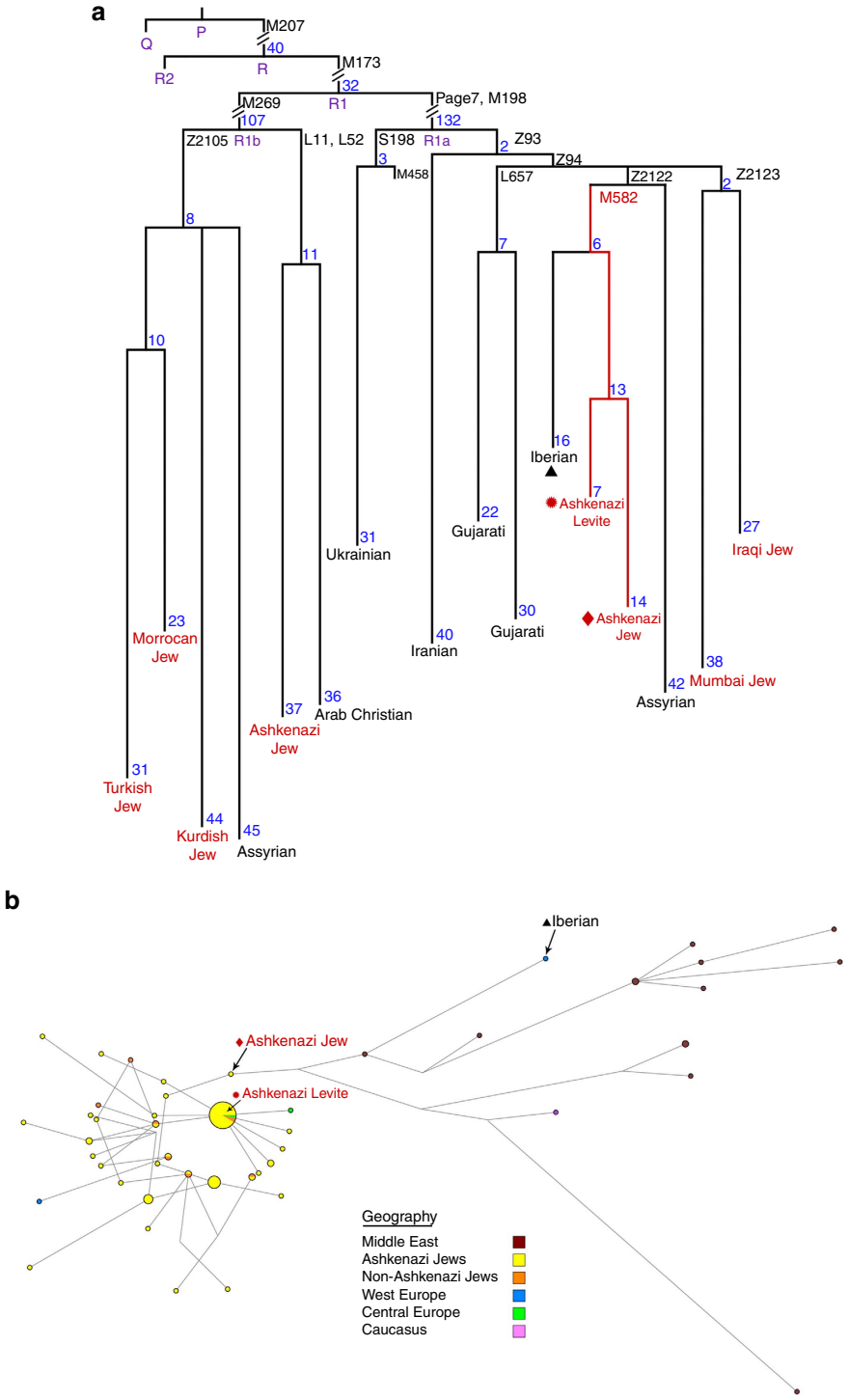

Figure 1 | Y-chromosome haplogroup R1 phylogeny and haplogroup R1a-M582 STR haplotypes diversity. (**a**) Schematic representation of haplogroup R1 phylogeny as reconstructed from 16 whole Y-chromosome sequences. The detailed figure is shown in Supplementary Fig. S1. (**b**) MJN of haplogroup R1a-M582 samples using the 19 Y-chromosome STRs: DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS439, DYS461 = A7.2, DYS385a, DYS385b, DYS437, DYS438, DYS448, DYS456, DYS458, DYS635 and Y GATA H4. Circles represent haplotypes, with area proportional to frequency and coloured according to population, as shown in the legend. The smallest circles represent singletons. The branch lengths are proportional to the number of STRs separating the nodes. The allocations of the Iberian (HG01617), Ashkenazi Levite (16207) and Ashkenazi Jewish (P3) whole-genome samples are marked in both panels of the figure by the triangle, flower and diamond signs, respectively.

samples (including 922 R1a-M198) of non-Jews from East Europe, where the Ashkenazi Jewish community flourished in recent centuries (Table 1). While the frequency of R1a in the paternal gene pool of Eastern European Slavonic and non-Slavonic peoples[16,18,31] reaches as high or higher than 50%, the deeper phylogenetic resolution reported here renders the presumption that similarly high frequencies of R1a among Ashkenazi Levites reflect substantial gene flow or a founder

Table 1 | R1a clade frequencies in different regions and populations.

Country/region or population	N total	M198	%	M582	%	References
Jews	1,421	123	8.66	90	6.33	
Ashkenazi Jews*	600	87	14.50	80	13.33	Updated from[18,39]
Non-Ashkenazi Jews	821	36	4.38	10	1.22	Updated from[18,39]
Ethiopia	5	0	0.00	0	0.00	Updated from[18,39]
India	76	6	7.89	0	0.00	Updated from[18,39]
Italy	11	0	0.00	0	0.00	Updated from[18,39]
Near East	188	7	3.72	2	1.06	Updated from[18,39]
Near East pooled†	180	5	2.78	0	0.00	Updated from[18,39]
Israel	8	2	25.00	2	25.00	Updated from[18,39]
Caucasus	119	7	5.88	0	0.00	Updated from[18,39]
North Africa	173	5	2.89	2	1.16	Updated from[18,39]
North Africa pooled‡	162	3	1.85	0	0.00	Updated from[18,39]
Algeria	11	2	18.18	2	18.18	Updated from[18,39]
Spanish Exile	175	9	5.14	6	3.43	Updated from[18,39]
Spain pooled§	30	0	0.00	0	0.00	Updated from[18,39]
Bulgaria	49	3	6.12	2	4.08	Updated from[18,39]
Slovenia	4	2	50.00	2	50.00	Updated from[18,39]
Turkey	92	4	4.35	2	2.17	Updated from[18,39]
Yemen	74	2	2.70	0	0.00	Updated from[18,39]
Non-Jewish populations	15,138	2,711	17.91	22	0.15	
Western/Northern Europe	1,068	106	9.93	1	0.09	
Western/Northern Europe pooled‖	746	64	8.58	0	0.00	Updated from[16]; this study
Germany	322	42	13.04	1	0.31	Updated from[16]
Eastern Europe¶	2,149	922	42.90	0	0.00	Updated from[16]
Central/Southern Europe	3,756	710	18.90	2	0.05	
Central/Southern Europe pooled#	3,386	570	16.83	0	0.00	Updated from[16,27,49]; this study
Hungary	113	21	18.58	1	0.88	Updated from[16]
Slovakia	257	119	46.30	1	0.39	Updated from[16]
Near East	3,739	303	8.10	18	0.48	
Near East pooled**	2,494	157	6.30	0	0.00	Updated from[16,18,46,48,50,52,58]
Kurds (Turkey and Kazakhstan)	106	7	6.60	3	2.83	This study
Iran	150	21	14.00	2	1.33	Updated from[51]
Iran	188	17	9.04	1	0.53	Updated from[16]
Iran North	336	29	8.63	0	0.00	Updated from[48]
Iran Azeri	360	48	13.33	9	2.50	Updated from[48]; this study
Iran Kerman	105	24	22.86	3	2.86	This study
Caucasus	2,164	211	9.75	1	0.05	
Caucasus pooled††	2,077	200	9.63	0	0.00	Updated from[16]
Nogays	87	11	12.64	1	1.15	Updated from[16]
South Asia‡‡	1,039	189	18.19	0	0.00	Updated from[47,53]; this study
Central/Southern Asia§§	1,223	270	22.08	0	0.00	Updated from[16]; this study

*Ashkenazi Jews from Austrohungary, Belorussia, Czech Rep., Estonia, France, Germany, Latvia, Lithuanians, Moldova, the Netherlands, Poland, Romania, Russia, United Kingdom, Ukraine.
†Afghanistan, Iran, Iraq, Kurds, Uzbekistan.
‡Egypt, Jerba, Libya, Morocco, Tunisia.
§Belmonte, Portugal, Chile, Colombia, France, Greece, Netherlands, Spain, Surinam, Yugoslavia.
‖England, Wales, Ireland, Denmark, Norway, Sweden, Spain.
¶Estonians, Vepsas, Karelians, Russians from four locations (Kostroma, Orjol, Belgorod, Pskov), Belarusians, Ukrainians, Circum-Uralic populations (Maris, Udmurts, Komis, Chuvashes, Tatars).
#Switzerland, Czech Rup., Hungarian Roma, Poland, Italy, Slovenia, Romania, Bulgaria, Croatia, Croatian Roma, Bosnia, Herzegovina, Serbia, Serbian Roma, Macedonia, Greece, Crete.
**Bedouins, Druze, Palestinians, Turkey, Egypt, Qatar, United Arab Emirates, Oman, Iran (3 sets), Assyrians, Iraq.
††Armenians, Georgians, Abkhazes, South and North Osetians, Chamalals, Bagvalals, Andis, Tabassarans, Adyghes, Karachays, Kumyks, Lezgis, Darginians, Abazines, Balkars, Cherkessians, Kabardinians, Karanogays, Ingushes, Avars.
‡‡Pakistan, India, Nepal Tharu, Nepal Hindus.
§§Afghanistan (five populations), Turkmens, Tajiks, Kazakhs, Kyrgyz, Altaians, Tuvinians, Khakassians, Mongolia.

effect from their hosts highly unlikely. Within 1,068 West/North European samples (106 R1a-M198), M582 was observed in just one German sample, and among 3,756 Central/South European samples (710 R1a-M198), it was found only in one Hungarian and one Slovakian sample (Table 1). It is important to note that many of the samples from individuals of European descent from Germany, The Netherlands, England, Ireland and Norway were collected in USA without ethnicity or religious affiliation information, so some of them might represent contemporary Ashkenazi Jews residing in the United States. As such, our results are likely more conservative because fewer of our R1a-M582 samples actually have non-Jewish European ancestry. Among 3,739 Near Eastern samples (303 R1a-M198), R1a-M582 was identified in various populations, with the highest frequency occurring within Iranians collected from the southeastern Kerman population who self-identified as Persians, northwestern Iranian Azeri and in Cilician Anatolian Kurds, at 2.86%, 2.50% and 2.83%, respectively (Table 1). In contrast, among 2,164 samples from the Caucasus (211 R1a-M198), R1a-M582 was found in just one Nogay sample (Table 1). As the overall frequencies obtained for R1a-M582 were low, we applied Fisher's exact test to verify that the different sample sizes did not bias our results (Supplementary Table S2). Out of a total of eight tests for the entire population, all but one was significant at the nominal 0.05 level, and all but three were still significant after a Bonferroni correction. For the R1a-M198 samples only, all eight tests were significant at the 0.05 level and all but one significant after Bonferroni correction.

Among Jews (Table 1), the overall frequency of R1a-M582 was 6.33% (90/1,421), while representing 73.17% (90/123) of the R1a-M198 chromosomes. Two factors must be considered when interpreting these frequencies: the known founder effect previously reported for Ashkenazi Levites[18] and the differing degrees of demographic information available for our 1,421 Jewish samples, compiled from various sampling campaigns[18,21,24]. For part of the sample, only the name of the community and not the paternal caste was documented. Hence, some Ashkenazi samples carrying haplogroup R1a-M582 might be undocumented Levites. Other collections specifically recruited Cohanim and Levites, so their inclusion in this study is likely to increase our overall estimate of R1a-M582 frequency among Ashkenazi Jews. These biases should affect the non-Ashkenazi Jewish data set, comprising 821 samples, to a lesser degree.

Of the 821 non-Ashkenazi Jewish samples, 36 (4.38%) belonged to haplogroup R1a, and ten (1.22%) belonged to R1a-M582 (Table 2). These frequencies are similar to R1a-M582 frequencies observed in non-Jewish Near Eastern populations. Of the 10 non-Ashkenazi R1a-M582 individuals, two were from the North African Algerian Jewish community, six belonged to Spanish expulsion descendent communities of Slovenia, Turkey and Bulgaria, and two were from individuals reporting their last known parental origin as Israel. Significantly, all eight individuals for whom caste information was available self-identified as Levites. Caste information was unavailable for the two Bulgarian individuals (Supplementary Data 2). Moreover, out of nine R1a-M198 non-Ashkenazi Levites, eight were R1a-M582. Among the 600 Ashkenazi Jews, 87 belonged to haplogroup R1a-M198, and of these, 80 (91.95%) were confirmed to belong to haplogroup R1a-M582 (Table 2). The Ashkenazi samples included a total of 279 Israelites, 110 Cohens, 97 Levites and 114 unclassified samples in which the counts and frequencies of haplogroup R1a-M198/R1a-M582 were 12 (4.3%)/7 (2.5%), 1 (0.9%)/1 (0.9%), 63 (64.9%)/63 (64.9%) and 11 (9.6%)/9 (7.9%), respectively (Table 2). Remarkably, all Ashkenazi Levite R1a samples belonged to haplogroup R1a-M582, and this haplogroup is virtually absent among members of the Cohen caste.

STR-based network analysis. Patterns of Y-chromosome STR diversity are useful for identifying additional population substructure among individuals belonging to the same haplogroup[32]. Of the 90 Jewish R1a-M582 samples reported herein, STR haplotypes were obtained for 86 of them (Supplementary Data 2). The Median Joining Network (MJN) demonstrated one clear cluster and a more differentiated set of chromosomes within haplogroup R1a-M582 (Fig. 1b). The first cluster, referred to below as the Ashkenazi cluster, includes Ashkenazi Levites, Ashkenazi non-Levites, non-Ashkenazi Jews and Europeans. This Ashkenazi cluster demonstrates a star-like pattern with Ashkenazi Levites, non-Ashkenazi Jews and Central European samples sharing the modal haplotype. The more differentiated set of chromosomes, referred to below as the Near Eastern set, include populations sampled from an area covering contemporary Iran and eastern Anatolia (Turkey), including southeastern Iranian Kerman, northwestern Iranian Azeri, Kurds from the historic Cilician part of Anatolia and Kazakhstan, one Nogay, and a single Ashkenazi Jew—the index case (P3, Supplementary Table S1). This Near Eastern set demonstrates a non-star-like pattern with multiple mutational events separating the various component haplotypes. These results strongly suggest and are consistent with a split in the R1a-M582 haplogroup, which has not yet been identified. Furthermore, DYS456 had significant discriminatory power within haplogroup R1a-M582 for separating the Ashkenazi Levite cluster from the Near Eastern set of haplotypes (Fig. 1b). Only one STR profile, obtained from the Ashkenazi Jewish R1a-M582 whole-genome sample, exhibited 15 repeats rather than 14 repeats at locus DYS456, and this haplotype was two mutational steps removed from the Levite modal haplotype. While the allocations of the Ashkenazi Levite, Ashkenazi Jewish and Iberian whole-genome samples are marked on the MJN (Fig. 1b), the latter two samples were not part of our population set, and accordingly they were excluded from our population-based analyses.

Table 2 | R1a clade frequencies within Jewish castes.

	Total	R-M198		R-M582	
Ashkenazi	600	87	14.5%	80	13.3%
Israel	279	12	4.3%	7	2.5%
Cohen	110	1	0.9%	1	0.9%
Levite	97	63	64.9%	63	64.9%
Unknown	114	11	9.6%	9	7.9%
Non-Ashkenazi	821	36	4.4%	10	1.2%
Israel	363	16	4.4%	0	0.0%
Cohen	95	3	3.2%	0	0.0%
Levite	51	9	17.6%	8	15.7%
Unknown	312	8	2.6%	2	0.6%

Dating the Ashkenazi Levite lineage. Assuming a mutation rate of 1×10^{-9} per base pair per year in the 8.97 Mbp region[28] of the X-chromosome degenerate non-recombining regions of the Y-chromosome, we estimated the divergence time of the two Ashkenazi M582 whole sequences that differed at 21 positions as 1,200 (SE 300) years. This divergence time represents a lower bound for the age of R1a-M582 because the time estimate was based on just two individual sequences that may share a more recent common ancestor than the rest of the Levites. The Iberian M582 lineage was excluded from age calculations because its low coverage made calling private mutations uncertain. The upper bound for the age of the M582 clade was estimated as the divergence time of the two Ashkenazi M582 lineages and the Assyrian R1a-Z2122(xM582) sequence, yielding an estimate of 4,000 (SE 300) years.

Relying on larger sample sizes, we also estimated the coalescence time of the R1a-M582 lineages from STR data. The coalescence age of the Near Eastern cluster was 11.3 ± 4.1 kya, as compared with a coalescence time of 2.4 ± 1.0 kya for the Ashkenazi Jewish cluster. The coalescence age calculated from Ashkenazi Levites alone was 2.6 ± 1.2 kya, which was similar to the age of 3.1 ± 1.5 kya obtained for non-Ashkenazi Jews. Estimates of the coalescent times must be viewed with caution, in part because of uncertainties involving an appropriate choice of Y-chromosome STR mutation rate, but comparisons between groups are informative[33]. Furthermore, the much lower variance among Jews relative to that observed for the Near Eastern cluster is striking.

Our coalescence analysis should be interpreted cautiously in light of some pertinent caveats. First, our whole Y-chromosome-based analysis relied on results from next generation sequencing. Several factors indicate high sequencing accuracy was achieved, such as the high coverage of the obtained positions, the ability to reconstruct the haplogroup R1 phylogeny, the identification of known SNPs of phylogenetic relevancy, the fact that the studied genetic locus is uniploid, and the ability to obtain the same positions from different samples using different platforms. However, Sanger-type sequencing was not used to validate all of the positions used to reconstruct our phylogeny. Therefore, both false-positive and false-negative calls might render our

estimates provisional. Second, it is also possible that results obtained from using the different genotyping platforms of Complete Genomics and Illumina might have introduced biases into our variant calls. Third, our STR-based coalescence analysis used an evolutionary mutation rate as one of its priors, although considerable controversy exists regarding whether a pedigree-based mutation rate is more appropriate. We opted to use the evolutionary rate because the samples used in the analysis represented very different ancestries and were therefore unlikely to share genealogical ties in recent generations. Last, coalescence analysis is prone to additional biasing factors still much debated in the literature, such as the correct mutation rate and the actual generation time.

Discussion

Our data necessitate revising the current hypotheses regarding the origins of the Ashkenazi Levite founding lineage. Consistent with the previous conclusions[18], the tight cluster of Ashkenazi Levite R1a-M582 haplotypes strongly indicates a recent origin from a single common ancestor who, according to our provisional dating, lived ~1.5–2.5 kya. Importantly, this young STR-based age estimate is now further supported by evidence from Y-chromosome sequences. However, the previously proposed Eastern European origin of this lineage is no longer tenable given that our data suggest haplogroup R1a-M582 actually originated in the Near East[18]. The higher R1a-M582 diversities and frequencies observed among Near Eastern populations indicate R1a-M582 originated in this geographic region. While this conclusion is independent of our findings from Jews, the mere existence of the haplogroup among Jews actually lends further support. However, as the frequencies of R1a-M582 were low overall, we first assessed and demonstrated that the frequencies observed in the different continental/sub-continental regions (Table 1) were statistically significant considering the different sample sizes (Supplementary Table S2), which we now discuss in detail.

Considering the historical records of Ashkenazi Jews, three potential geographic sources should be considered: the Near East, which was the geographic location for the ancient Hebrews; Europe, which was the residence of the Ashkenazi Jewish Diaspora and the region in which they evolved for nearly two millennia; and the region overlapping with the no longer extant mid-11th Century Khazarian Khaganate, whose ruling class has been suggested to have converted to Judaism[18]. Our data render the latter source highly unlikely since the Khazarian Khaganate overlapped with the Northern Pontic-Caspian steppe and the North Caucasus region, in which just one Nogay sample carried the R1a-M582 haplogroup (Table 1). Furthermore, the Nogays, formerly a powerful Kipchak Turkic-speaking nomadic confederation, are relatively recent inhabitants of the Caucasus, and the STR haplotype of the sole R1a-M582 Nogay sample lies outside of the Levite cluster. Had the Caucasus region been the source for the Ashkenazi modal lineage, we likely would have found R1a-M582 Y-chromosomes in some of its 20 local populations examined in our sample of more than 2,000 Y-chromosomes (Table 1). As previously suggested, the European and particularly, the Eastern European paternal gene pool was seen as a natural and highly plausible source for the Ashkenazi Levite lineage as both the Ashkenazi community and haplogroup R1a frequencies peak in this region. But surprisingly, haplogroup R1a-M582 was not detected in non-Jewish Eastern European samples and was found only in singleton samples in various Central and Western European populations (Table 1).

The direction of gene flow involving the Ashkenazi Levite R1a-M582 lineage warrants discussion. The presence of haplogroup R1a-M582 at minute frequencies among Europeans could theoretically suggest introgression into the Ashkenazi Levite community and subsequent expansion among Ashkenazi Levites. Alternatively, the European samples can represent Ashkenazi admixture into the general European gene pool due to assimilation that has occurred in Western Europe throughout the centuries. While it is impossible to refute the former, a few arguments render the latter more likely. The European samples carrying haplogroup R1a-M582 haplotypes add no further diversity to that found among Ashkenazi Jews, despite having been identified at different European locations. These haplotypes are actually identical or one step removed from the Levite modal haplotype. If haplogroup R1a-M582 were of general ancient European ancestry, we would have observed at least some baseline diversity and genetic distance expressed in a few STR-based mutational events between the lineages evolving in the general European gene pool and the one lineage whose evolution is confined to the Ashkenazi samples. While this pattern could not be demonstrated among Europeans, it was clearly evident among Near Eastern samples.

Near Eastern populations are the only populations in which haplogroup R1a-M582 was found at significant frequencies (Table 1). Moreover, the representative samples displayed substantial diversity even within this geographic region (Fig. 1b). Higher frequencies and diversities often suggest lineage autochthony. Hence, we can assess whether or not the origin of haplogroup R1a-M582 is in present-day Iran and eastern Anatolia, or rather the broader region of the Near East. Our data demonstrate the occurrence of R1a-M582 among different Iranian populations, among Kurds from Cilician Anatolia and Kazakhstan, and among Ashkenazi and non-Ashkenazi Jews. These observations, and the STR network delineating an internal R1a-M582 structure, might attest to a broad Near Eastern distribution range of this minor haplogroup that survived to the present day at low frequencies among Iranian Kerman, Iranian Azeri, Kurds and Jews. Haplogroup R1a-M582 was not detected in samples from Iraq or among Bedouins, Druze and Palestinians sampled in Israel.

Additional samples from the region between the Levant and Iraq, such as from Syria, were not available. The identification of an R1a-M582 chromosome in the Iberian sample from the 1,000 Genome Project might represent the legacy of Jews or Moors in Iberia[34]. The STR pattern obtained for this sample is more compatible with the latter (Fig. 1b). The MJN also demonstrates that the STR profile obtained for the whole Ashkenazi Jewish sample is within the range of variation observed in the Near Eastern set of chromosomes (Fig. 1b). While this finding can merely represent a mutational event within a set of fast evolving STRs, it can also suggest the existence of multiple R1a-M582 sub-clades within the presumed ancestral Levantine deme of contemporary Ashkenazi Jews. Notably, the only non-M582 sample we identified within the Z2122 haplogroup (Fig. 1a and Supplementary Fig. S1) corresponds to an individual from the Aramaic speaking population claiming descent from the contemporary ancient Assyrians. Aramaic and Hebrew, the two ancient Western Semitic languages that are still spoken are thought to have diverged ~3,500–4,000 years ago[35], in a time frame coinciding with time estimate suggested above for the age of R1a-Z2122 haplogroup. Taken together, this line of evidence supports a Near Eastern origin for haplogroup R1a-M582, and hence the Ashkenazi Levite lineage as well.

Another notable observation regards the distribution of haplogroup R1a-M582 among Ashkenazi and non-Ashkenazi Levites. While R1a-M582 occurs at 64.9% (63/97) among Ashkenazi Levites, it comprises just 15.7% (8/51) of the non-Ashkenazi Levite paternal gene pool (Table 2). Among non-Ashkenazi Jews, R1a-M582 was observed only in Levites, and the observed

sub-haplogroup shares the same STR signature as that seen in Ashkenazi Levites. A few demographic scenarios can account for this observation. Clearly, a joint Levantine origin before the Diaspora split into the Ashkenazi and non-Ashkenazi Jews must be considered. Under this scenario, this particular R1a-M582 Levite lineage existed among the ancient Hebrews and was carried to the various Jewish Diaspora communities in a manner similar to that of the Cohen Modal Haplotype[21–23]. However, recent studies of genome-wide diversity point to recent shared Levant ancestries and a close genetic proximity among members of Ashkenazi, North African and Spanish Exile Jewish communities[17,36,37]. Similarly, some of the mitochondrial DNA Ashkenazi founding lineages are also found among Spanish Exiles[20]. Therefore, another possible scenario is that of continuous gene flow between Ashkenazi, North African and Spanish Exile Jewish communities. Under this scenario, Ashkenazi Levites must have repeatedly and episodically introgressed into non-Ashkenazi communities, maintaining their Levite status while abandoning their Ashkenazi affiliation.

The haplogroup distribution pattern observed for the Ashkenazi Levite lineage does not seem to be restricted to haplogroup R1a chromosomes. While haplogroup R1a is the most prevalent in East Europe, it is its companion haplogroup, R1b, that is the most frequent among West Europeans[38]. Similar to haplogroup R1a, haplogroup R1b frequencies are also found at low frequencies among non-Europeans, including Jews[39]. Based on distinct STR patterns, it has been suggested that some of the Jewish R1b chromosomes may have a West European origin and that some might be of Near Eastern origin[34]. Our sequence data are compatible with this suggestion as the Kurdish, Moroccan and Turkish Jewish R1b lineages for which whole Y-chromosome data was determined to coincide with a different branch than the one which is common in Europe (Fig. 1a). Intriguingly, we also noticed that like the R1a subclade, the particular subclade within R1b is also shared by Jews and a Y-chromosome, sampled in Aramaic speaking Assyrian descendants. In contrast, one Ashkenazi Jewish and one Arab Christian R1b lineages seem to be nested with the European R1b-L52 haplogroup. Estimating the proportion of haplogroup R1b chromosomes among Jews and in particular among Ashkenazi Jews of Levantine origin remains the scope of further studies.

In summary, we have circumscribed the geography of marker M582 within the broad distribution zone of R1a-M198* lineages. We have shown it to be a minor haplogroup that is primarily shared among Iranian Kerman, Iranian Azeri, Kurds, Ashkenazi Jews and non-Ashkenazi Jews, and that it is virtually absent in the Caucasus region, Europe, South Asia, and southern Siberia. Thus, in contrast to previous suggestions regarding the origin of this Ashkenazi Levite founding lineage, we conclude that haplogroup R1a-M582 was likely carried into Europe by Jewish migrants and that it expanded among Ashkenazi Levites during their subsequent Diaspora period in a region that is incidentally dominated by other R1a parahaplogroups that coalesce with R1a-M582 prior to the establishment of the Jewish people. The existence of the R1a-M582 lineage within non-Ashkenazi Levites from different Jewish communities suggests it to either be a pre-Diaspora Hebrew Levite lineage or that continuous gene flow existed between Jewish communities, presumably, from Ashkenazi to non-Ashkenazi Jews. Our case study of Ashkenazi Levites is a vivid example of the ability to refine haplogroup structure and trace subtle signals of gene flow when sufficiently resolved data are available at the genomic and population levels.

Methods

Preliminary screening of haplogroup R1a markers. Whole-genome sequencing, privately obtained by an individual of Ashkenazi Jewish (Israelite) ancestry (Fig. 1, Supplementary Fig. S1), created the opportunity to address the topic of Ashkenazi Levite paternal heritage. Following Stanford University Human Subjects Internal Revenue Board approval, the list of all Y-chromosome SNP assignments was extracted from the whole-genome data and a saliva sample for DNA extraction was donated for validation and further analysis. This individual's assignment to haplogroup R1a-M198*(xM458, M434) was confirmed using standard PCR-based genotyping assays[16]. The list contained SNPs previously undocumented on the genetic genealogy aggregator ISOGG web site. These putatively private alleles were then intersected with publically available low coverage Y-chromosome sequence data from the 1,000 Genomes Project (http://www.1000genomes.org/data). One of the Ashkenazi Jewish Y-chromosome SNPs appeared to be shared with a single sample from the Iberian population in Spain and hence labelled as potentially non-private. Accordingly, we initially screened a set of haplogroup R1a-M198 samples from the HGDP CEPH panel (Russia, Orkney Island, Caucasus, Israel and five ethnic groups from Pakistan), which was further augmented with Turks, Iranians and Greeks[16]. Since this marker was ascertained in the genome of an individual of Ashkenazi Jewish ancestry carrying an R1a Y-chromosome[18], we also genotyped it in a small set of 13 Ashkenazi Levite known to belong to the same haplogroup[18]. As M582 was detected in all 13 Ashkenazi Levites and in one Iranian Kerman sample, a phylogenetic study of the Ashkenazi Levite lineage at the whole Y-chromosome level and a large-scale genotyping on all samples available to us followed.

Whole Y-chromosome sequencing and phylogeny reconstruction. Supplementary Table S1 details the information for each of the sixteen samples included in this study of which thirteen are first reported herein. Supplementary Data 1 outlines the reference and the alternative alleles for all markers discovered in the sequencing campaign. To reconstruct the phylogeny of haplogroup R1a-M198, we evaluated the two Gujarati R1a individuals sequenced by Complete Genomics[40] and the single Iberian individual sample reported as part of the 1,000 genome project using the Illumina platform[41]. Next, we identified all haplogroup R1a-M198 in all population sample collections available to us and chose seven additional samples for whole-genome sequencing. Samples were chosen to include a wide range of haplogroup R1a-M198 internal variation on the basis of previously available STR haplotypes (Supplementary Data 2), populations and geographic regions relevant to the question of the origin on the Ashkenazi Levite lineage. Accordingly, within Ashkenazi Jews, one sample was from an Ashkenazi Levite carrying the Levite modal haplotype as inferred from STR analysis and one was the Israelite Ashkenazi Jew available from our preliminary work. One Iraqi Jew and one Mumbai Jew were included to allow a general understanding of the R1a-M198 variation within the Jewish world. Similarly, one Assyrian, one Iranian and one Ukrainian samples were chosen to grossly represent the large regions that might be relevant to the understating of the Ashkenazi Levite origin including the Near East and East Europe. In addition, and to accurately confirm our ability to reliably reconstruct the root of haplogroup R1a-M198 within its ancestral R1a, we studied a total of six haplogroup R1b-M269 samples from Near Eastern populations for which the current available public data is scant. It is also worth noting that the quality and source of DNA available to us had a significant role in our ability to include samples in our whole-genome sequencing campaign. In some cases, when samples of sufficient quality were unavailable from our population collections, new blood samples were collected from individuals self-affiliating with these populations to identify individuals carrying R1a-M198 individuals. Accordingly, Supplementary Table S1 details which samples are from our population data set and which are from other non-random collections. All samples reported herein were derived from blood samples that were collected with informed consent according to procedures approved by the Institutional Human Subjects Review Committees in their respective locations.

All sixteen samples were used to reconstruct haplogroup R1 phylogeny (Supplementary Fig. S1). The whole-genome sequences of the eight Jewish and five non-Jewish male individuals belonging to Y-chromosome haplogroup R1 were determined at high coverage using genomic DNA extracted from blood. High-quality variants mapping to Y-chromosome were extracted for downstream analyses. All but one sample were sequenced at Complete Genomics with Y-chromosome coverage $>20\times$ and one Ashkenazi Jewish R1a-M582 sample was sequenced with Illumina HiSeq 2000 at $14\times$ coverage. Alignment and allele calling were performed using Complete Genomics pipelines and CGAtools. The Illumina sequence genomic library was prepared using the Epicentre Nextera Sample Preparation Kit. Paired-end reads were sequenced using four Illumina HiSeq 2000 lanes. Aligned coverage for the Y-chromosome was $14\times$ (ref. 42). Raw variant calls were generated using Broad Institute's GATK v4 best practices pipeline[43,44], including read deduplication, base quality recalibration and local realignment around indels. In phylogenetic analyses we also used publicly available data from two Gujarati R1a individuals sequenced by Complete Genomics[40]. We filtered the data to exclude regions of poor mapping efficiency due to structural variation and focused on the nine X-chromosome degenerate non-recombining regions[28] that altogether encompass 8.97 Mbp of the Y-chromosome sequence. We also excluded from the analyses insertions, deletions and multistate SNPs. The one additional sample from an Iberian individual available from the 1,000 genome project is shown in our phylogeny (Supplementary Fig. S1) but not included in our coalescence analysis because of its low coverage.

We generated maximum-parsimony, neighbor-joining and minimum evolution trees in MEGA[45] for the high coverage Y-chromosome sequences. All trees

generated were identical in topology. The consensus tree was rooted with haplogroup Q sequence from the Complete Genomics public data (http://www.completegenomics.com/public-data/). The Build 37 coordinates of non-recurrent mutations defining branches within haplogroups R1a and R1b are shown in Supplementary Fig. S1.

Population collection and genotyping. We investigated a total of 15,138 non-Jewish individuals from Europe, the Near East, the Caucasus and Asia available to us. This sample set included 2,711 individuals belonging to haplogroup R1a[16,27,46–52]. A total of 1,421 Jewish samples from across the entire range of Diaspora Jewish communities were included[18,21,24,39]. All samples were hierarchically screened for the haplogroup R1a defining markers M198 or M17 (ref. 16) and then to M582, which represents one of the 19 binary SNPs comprising the root of haplogroup R1a-M582 in which the Ashkenazi Levite lineage is nested. M582 is a g.14,236,070T>G transversion according to GRCh37/hg19 assembly version. M582 was genotyped by RFLP assay using Tsp509I enzyme following amplification with primers 5′-3′: F: 5′-GAGGCTGCAGTGAGCTATGAC-3′ and R: 5′-GTCACCTGCTTGGTAAAG ATGAC-3′. Table 1 summarizes the populations used and the frequencies of haplogroup R1a-M198 and its descendant sub-haplogroup R1a-M582 within each population.

To evaluate the significance of differences in prevalence of the R1a-M582 allele between Near Eastern populations and the other non-Jewish populations in our data, we utilized Fisher's exact test for comparisons of prevalence of R1a-M582 within the entire population sample and also within R1a-M198 samples only (Supplementary Table S2).

To assess the genetic distance between the Ashkenazi Levite R1a-M582 haplotypes and all other Jewish or non-Jewish M582 haplotypes from all geographic regions, we used a set of 19 Y-chromosome STRs (Supplementary Data 2). This set of 19 Y-chromosome STRs could not be fully resolved in ten of the Jewish and five on the non-Jewish R1a-M582 samples due to DNA quantity and quality limitations. Previously reported datasets[16,18] were merged to obtain uniform STR haplotype profiles comprising the following 19 markers: DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS439, DYS461 = A7.2, DYS385a, DYS385b, DYS437, DYS438, DYS448, DYS456, DYS458, DYS635, and Y GATA H4.

In addition, the same 19 Y-chromosome STRs were assessed in the Ashkenazi Jewish and the Iberian samples (Coriell #HGO1617) which are not part of our population collection (Supplementary Data 2).

Age estimates using STRs and whole Y-chromosome variation. MJN was generated using the recommended default settings of Network 4.6.1.1 but allowing a higher weight to DYS456, which proved to be particularly informative in our data set. Next, coalescent times (Td)[53,54] of haplogroup R1a-M582 and its sub-clusters were estimated using an evolutionary effective mutation rate of 6.9×10^{-4} per 25 year generation for the following original set of 10 Y-chromosome STRs: DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS439, DYS461 (TAGA counts).

For estimating the divergence times of the whole Y-chromosome sequences using the rho statistic[55,56], we assumed mutation rate of 1×10^{-9} per base pair per year[57]. STR profiles obtained from our whole-genome samples that are not part of our population set were not included in the calculations.

References

1. Underhill, P. A. *et al.* Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41,** 539–564 (2007).
2. Gymrek, M. *et al.* Identifying personal genomes by surname inference. *Science* **339,** 321–324 (2013).
3. Dulik, M. C. *et al.* Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *Am. J. Hum. Genet.* **90,** 229–246 (2012).
4. Mendez, F. L. *et al.* An african american paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am. J. Hum. Genet.* **92,** 454–459 (2013).
5. Semino, O. *et al.* Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* **70,** 265–268 (2002).
6. Zegura, S. L. *et al.* High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol. Biol. Evol.* **21,** 164–175 (2004).
7. Jobling, M. A. *et al.* The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4,** 598–612 (2003).
8. Wells, R. S. *et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl Acad. Sci. USA* **98,** 10244–10249 (2001).
9. Karafet, T. M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* **18,** 830–838 (2008).
10. Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26,** 358–361 (2000).
11. Bamshad, M. J. *et al.* Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* **72,** 578–589 (2003).
12. Rosser, Z. H. *et al.* Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* **67,** 1526–1543 (2000).
13. Quintana-Murci, L. *et al.* Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am. J. Hum. Genet.* **68,** 537–542 (2001).
14. Underhill, P. A. *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7,** 996–1005 (1997).
15. Semino, O. *et al.* The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290,** 1155–1159 (2000).
16. Underhill, P. A. *et al.* Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur. J. Hum. Genet.* **18,** 479–484 (2010).
17. Atzmon, G. *et al.* Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am. J. Hum. Genet.* **86,** 850–859 (2010).
18. Behar, D. M. *et al.* Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. *Am. J. Hum. Genet.* **73,** 768–779 (2003).
19. Behar, D. M. *et al.* The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. *Am. J. Hum. Genet.* **78,** 487–497 (2006).
20. Behar, D. M. *et al.* Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. *PLoS One* **3,** e2062 (2008).
21. Hammer, M. F. *et al.* Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Hum. Genet.* **126,** 707–717 (2009).
22. Skorecki, K. *et al.* Y chromosomes of Jewish priests. *Nature* **385,** 32 (1997).
23. Thomas, M. G. *et al.* Origins of Old Testament priests. *Nature* **394,** 138–140 (1998).
24. Behar, D. M. *et al.* MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. *Eur. J. Hum. Genet.* **12,** 355–364 (2004).
25. Nebel, A. *et al.* Y chromosome evidence for a founder effect in Ashkenazi Jews. *Eur. J. Hum. Genet.* **13,** 388–391 (2005).
26. Battaglia, V. *et al.* Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur. J. Hum. Genet.* **17,** 820–830 (2009).
27. Karachanak, S. *et al.* Y-chromosome diversity in modern bulgarians: new clues about their ancestry. *PLoS One* **8,** e56779 (2013).
28. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23,** 388–395 (2013).
29. Rocca, R. A. *et al.* Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. *PLoS One* **7,** e41634 (2012).
30. Genomes Project, C *et al.* An integrated map of genetic variation from 1,092 human genomes. Nature **491**, 56-65 (2012).
31. Balanovsky, O. *et al.* Two sources of the Russian patrilineal heritage in their Eurasian context. *Am. J. Hum. Genet.* **82,** 236–250 (2008).
32. Shriver, M. D. *et al.* Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.* **5,** 611–618 (2004).
33. Busby, G. B. *et al.* The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc. Biol. Sci.* **279,** 884–892 (2012).
34. Adams, S. M. *et al.* The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am. J. Hum. Genet.* **83,** 725–736 (2008).
35. Kitchen, A. *et al.* Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. Biol. Sci.* **276,** 2703–2710 (2009).
36. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).
37. Gusev, A. *et al.* The architecture of long-range haplotypes shared within and across populations. *Mol. Biol. Evol.* **29,** 473–486 (2012).
38. Myres, N. M. *et al.* A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* **19,** 95–101 (2011).
39. Behar, D. M. *et al.* Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. *Hum. Genet.* **114,** 354–365 (2004).
40. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
41. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. Nature **467**, 1061-1073 (2010).
42. Langmead, B. *et al.* Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9,** 357–359 (2012).
43. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
44. McKenna, A. *et al.* The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

45. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28,** 2731–2739 (2011).
46. Cadenas, A. M. *et al.* Y-chromosome diversity characterizes the Gulf of Oman. *Eur. J. Hum. Genet.* **16,** 374–386 (2008).
47. Fornarino, S. *et al.* Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol. Biol.* **9,** 154 (2009).
48. Grugni, V. *et al.* Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. *PLoS One* **7,** e41252 (2012).
49. King, R. J. *et al.* Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. *Ann. Hum. Genet.* **72,** 205–214 (2008).
50. Luis, J. R. *et al.* The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.* **74,** 532–544 (2004).
51. Regueiro, M. *et al.* Iran: tricontinental nexus for Y-chromosome driven migration. *Hum. Hered.* **61,** 132–143 (2006).
52. Al-Zahery, N. *et al.* In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. *BMC Evol. Biol.* **11,** 288 (2011).
53. Sengupta, S. *et al.* Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* **78,** 202–221 (2006).
54. Zhivotovsky, L. A. *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74,** 50–61 (2004).
55. Forster, P. *et al.* Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59,** 935–945 (1996).
56. Saillard, J. *et al.* mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* **67,** 718–726 (2000).
57. Xue, Y. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19,** 1453–1457 (2009).
58. Di Cristofaro, J. *et al.* Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. *PLoS One* **8,** e76748 (2013).

Acknowledgements

We thank Julie Di Cristofaro for the Afghanistan results, Tuuli Reisberg for expert technical help and Jaime Raijman for preparation of Fig. 1. P.A.U. and A.A.L. thank Kenneth Chahine at Ancestry.com, and P.A.U. also thanks Prof Michael Snyder for support. We thank Levon Yepiskoposyan and Ardeshir Bahmanimehr for providing the Iranian samples. We thank Lars Mouritsen at Sorenson Genomics for technical assistance. We thank the Slava Smolokowski Fund at Rambam Medical Center for support. G.D.P. was supported by NSF graduate research fellowship DGE-1147470. A.K.P. was supported by European Social Fund's Doctoral Studies and Internationalisation Programme DoRa. This work was supported by the European Union European Regional Development Fund through the Centre of Excellence in Genomics, by the Estonian Biocentre and the University of Tartu, by the European Commission grant 205419 ECOGENE to the EBC, by the Estonian Science Foundation grant nr8973, and by the Estonian Basic Research Grant SF 0270177s08.

Author contributions

Si.R., D.M.B., R.V. and P.A.U. conceived and designed the study. Si.R., D.M.B., H.S., A.K.P., V.G. and O.S. provided DNA samples to this study. Si.R. and A.A.L. managed and completed the genotyping campaign. Si.R. and M.J. performed the Y-chromosome STR-based analysis. T.K. performed the whole Y-chromosome analysis. Sa.R. performed the Fisher's exact test. D.M.B., R.V., T.K. and P.A.U. wrote the paper. All of the other authors contributed annotation, analyses, writing or data throughout the project. All authors discussed the results and commented on the manuscript.

Additional information

Accession codes: The whole Y-chromosomes sequences have been deposited to the European Nucleotide Archive under accession code PRJEB4991.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: Peter A. Underhill and Alice A. Lin are partially supported by a grant from Ancestry.com. Ancestry.com currently employs Natalie M. Myres. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Rootsi, S. *et al.* Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites. *Nat. Commun.* 4:2928 doi: 10.1038/ncomms3928 (2013).