The genome-wide structure of the Jewish people

Doron M. Behar[1,2]*, Bayazit Yunusbayev[2,3]*, Mait Metspalu[2]*, Ene Metspalu[2], Saharon Rosset[4], Jüri Parik[2], Siiri Rootsi[2], Gyaneshwer Chaubey[2], Ildus Kutuev[2,3], Guennady Yudkovsky[1,5], Elza K. Khusnutdinova[3], Oleg Balanovsky[6], Ornella Semino[7], Luisa Pereira[8,9], David Comas[10], David Gurwitz[11], Batsheva Bonne-Tamir[11], Tudor Parfitt[12], Michael F. Hammer[13], Karl Skorecki[1,5] & Richard Villems[2]

Contemporary Jews comprise an aggregate of ethno-religious communities whose worldwide members identify with each other through various shared religious, historical and cultural traditions[1,2]. Historical evidence suggests common origins in the Middle East, followed by migrations leading to the establishment of communities of Jews in Europe, Africa and Asia, in what is termed the Jewish Diaspora[3–5]. This complex demographic history imposes special challenges in attempting to address the genetic structure of the Jewish people[6]. Although many genetic studies have shed light on Jewish origins and on diseases prevalent among Jewish communities, including studies focusing on uniparentally and biparentally inherited markers[7–16], genome-wide patterns of variation across the vast geographic span of Jewish Diaspora communities and their respective neighbours have yet to be addressed. Here we use high-density bead arrays to genotype individuals from 14 Jewish Diaspora communities and compare these patterns of genome-wide diversity with those from 69 Old World non-Jewish populations, of which 25 have not previously been reported. These samples were carefully chosen to provide comprehensive comparisons between Jewish and non-Jewish populations in the Diaspora, as well as with non-Jewish populations from the Middle East and north Africa. Principal component and structure-like analyses identify previously unrecognized genetic substructure within the Middle East. Most Jewish samples form a remarkably tight subcluster that overlies Druze and Cypriot samples but not samples from other Levantine populations or paired Diaspora host populations. In contrast, Ethiopian Jews (Beta Israel) and Indian Jews (Bene Israel and Cochini) cluster with neighbouring autochthonous populations in Ethiopia and western India, respectively, despite a clear paternal link between the Bene Israel and the Levant. These results cast light on the variegated genetic architecture of the Middle East, and trace the origins of most Jewish Diaspora communities to the Levant.

Recently, the capacity to obtain whole-genome genotypes with the use of array technology has provided a robust tool for elucidating fine-scale population structure and aspects of demographic history[17–23]. This approach, initially used to account for population stratification in genome-wide association studies, identified genome-wide patterns of variation that distinguished between Ashkenazi Jews and non-Jews of European descent[7,11,12,14–16]. Similarly, a large-scale survey of autosomal microsatellites found that samples from four Jewish communities clustered close to each other and intermediate between non-Jewish Middle Eastern and European populations[10].

Illumina 610K and 660K bead arrays were used to genotype 121 samples from 14 Jewish communities. The results were compared with 1,166 individuals from 69 non-Jewish populations (Supplementary Note 1 and Supplementary Table 1), with particular attention to neighbouring or 'host' populations in corresponding geographic regions. These results were also integrated with analyses of genotype data from about 8,000 Y chromosomes and 14,000 mitochondrial DNA (mtDNA) samples (Supplementary Note 6 and Supplementary Tables 4 and 5). Several questions were then addressed: What are the locations of the various Jewish communities in a global genetic variation context? What are the features of the Middle Eastern (Supplementary Fig. 1) population genetic substructure? What are the genetic distances between contemporary Jewish communities, their Diaspora neighbours and Middle Eastern populations? Can the genetic origin of Jews be pinpointed within the Middle East?

The EIGENSOFT package[24] was used to identify the principal components (PCs) of autosomal variation in our Old World sample set (Fig. 1 and Supplementary Fig. 2a). This analysis places the studied samples along two well-established geographic axes of global genetic variation[18,19,22]: PC1 (sub-Saharan Africa versus the rest of the Old World) and PC2 (east versus west Eurasia). Focusing on the Middle Eastern populations in the PC1–PC2 plot (Fig. 1b) reveals more geographically refined groupings. Populations of the Caucasus, flanked by Cypriots, form an almost uninterrupted rim that separates the bulk of Europeans from Middle Eastern populations. Bedouins, Jordanians, Palestinians and Saudi Arabians are located in close proximity to each other, which is consistent with a common origin in the Arabian Peninsula[25], whereas the Egyptian, Moroccan, Mozabite Berber, and Yemenite samples are located closer to sub-Saharan populations (Fig. 1a and Supplementary Fig. 2a).

Most Jewish samples, other than those from Ethiopia and India, overlie non-Jewish samples from the Levant (Fig. 1b). The tight cluster comprising the Ashkenazi, Caucasus (Azerbaijani and Georgian), Middle Eastern (Iranian and Iraqi), north African (Moroccan) and Sephardi (Bulgarian and Turkish) Jewish communities, as well as Samaritans, strongly overlaps Israeli Druze and is centrally located on the principal component analysis (PCA) plot when compared with Middle Eastern, European Mediterranean, Anatolian and Caucasus non-Jewish populations (Fig. 1). This Jewish cluster consists of

[1]Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa 31096, Israel. [2]Estonian Biocentre and Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia. [3]Institute of Biochemistry and Genetics, Ufa Research Center, Russian Academy of Sciences, Ufa 450054, Russia. [4]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. [5]Rappaport Faculty of Medicine and Research Institute, Technion – Israel Institute of Technology, Haifa 31096, Israel. [6]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow 115478, Russia. [7]Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia 27100, Italy. [8]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto 4200-465, Portugal. [9]Faculdade de Medicina, Universidade do Porto, Porto 4200-319, Portugal. [10]Institute of Evolutionary Biology (CSIC-UPF), CEXS-UPF-PRBB and CIBER de Epidemiología y Salud Pública, Barcelona 08003, Spain. [11]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel. [12]Department of the Languages and Cultures of the Near and Middle East, Faculty of Languages and Cultures, School of Oriental and African Studies (SOAS), University of London, London WC1H 0XG, UK. [13]ARL Division of Biotechnology, University of Arizona, Tucson, Arizona 85721, USA.
*These authors contributed equally to this work.

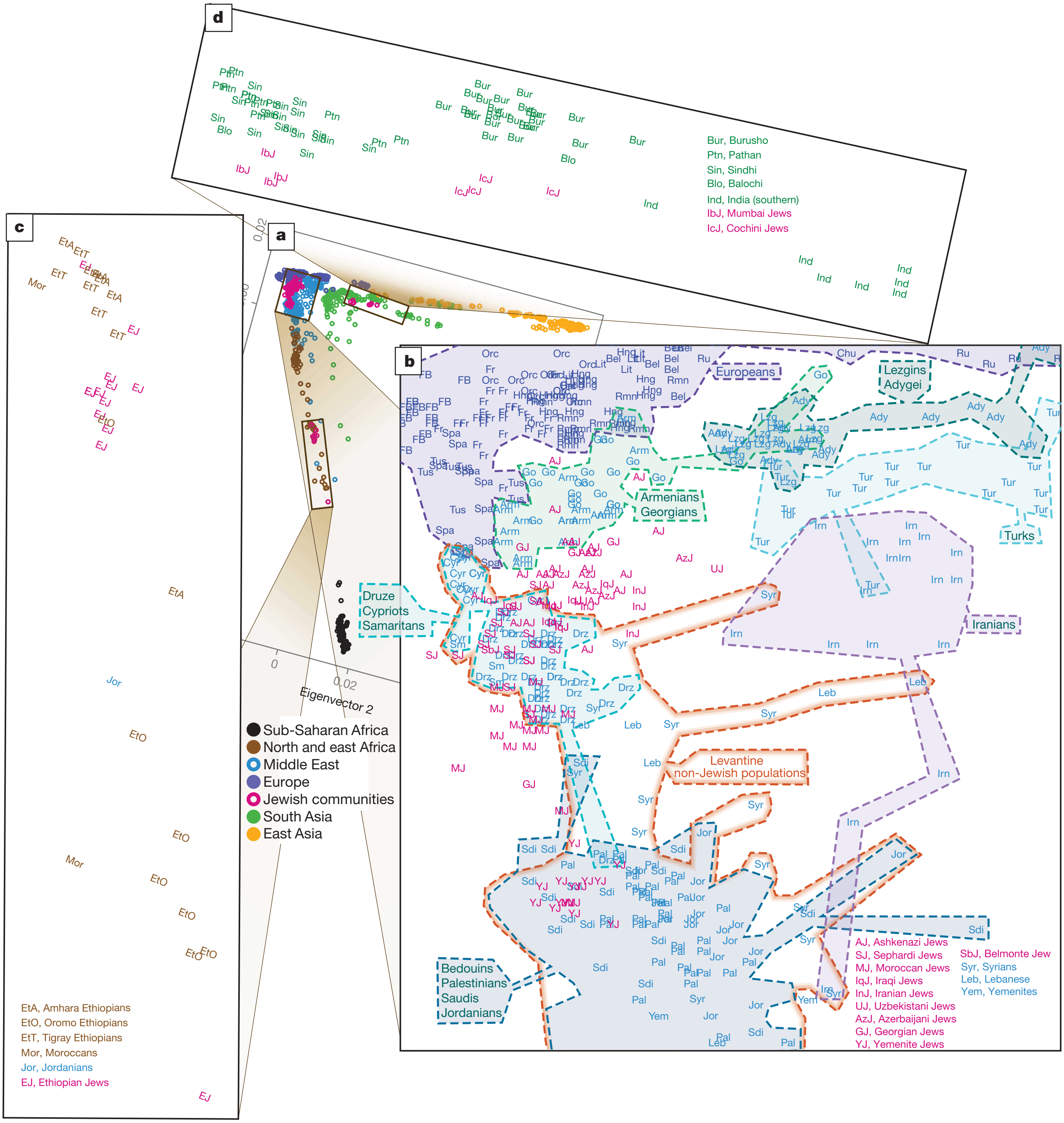

Figure 1 | PCA of high-density array data. **a**, Scatter plot of Old World individuals, showing the first two principal components. Each ring corresponds to one individual and the colour indicates the region of origin (for the full figure see Supplementary Fig. 2). **b**–**d**, A series of magnifications showing samples from Europe and the Middle East (**b**), Ethiopia (**c**) and south Asia (**d**). Each letter code (Supplementary Table 1) corresponds to one individual, and the colour indicates the geographic region of origin. In **b**, a polygon surrounding all of the individual samples belonging to a group designation highlights several population groups.

samples from most Jewish communities studied here, which together cover more than 90% of the current world Jewish population[5]; this is consistent with an ancestral Levantine contribution to much of contemporary Jewry. A compact cluster of Yemenite Jews, which is also located within an assemblage of Levantine samples, overlaps primarily with Bedouins but also with Saudi individuals (Fig. 1b). In contrast, Ethiopian and Indian Jews are located close to those from neighbouring host populations (Fig. 1c, d). Ethiopian Jews clustered with Semitic-speaking rather than Cushitic-speaking Ethiopians. See Supplementary Note 2 for a discussion of the assignment of samples representing the Belmonte and Uzbek (Bukharan) Jewish communities.

To glean further details of Levantine genetic structure, we repeated PCA on a restricted set of samples from west Eurasia (Fig. 2, Supplementary Fig. 3 and Supplementary Note 2) and by inspecting lower-ranked PCs in the Old World context (Supplementary Fig. 2b, c; PC1 versus PC3 and PC4). These analyses reveal three

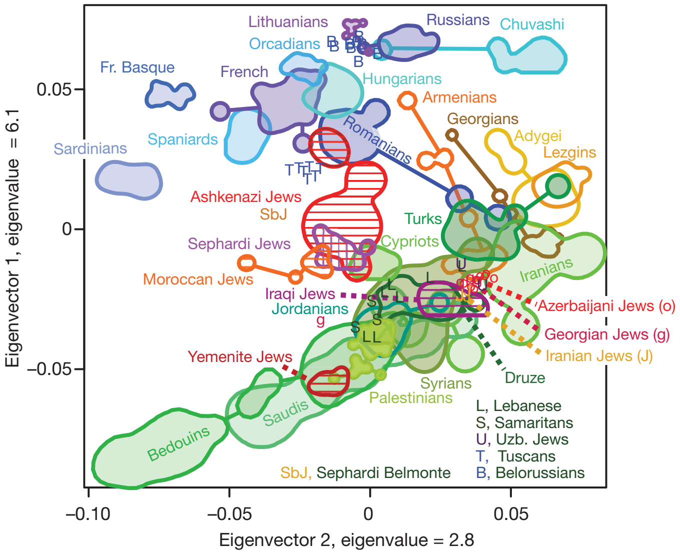

Figure 2 | PCA of west Eurasian high-density array data. Plot of kernel densities (Supplementary Note 2) for each population sample ($n > 10$) was estimated on the basis of PC1 and PC2 coordinates in Supplementary Fig. 3. Individuals from these samples were plotted by using PC1 and PC2 coordinates and were overlaid with the plot of kernel density.

distinct Near Eastern Jewish subclusters: the first group is located between Middle Eastern and European populations and consists of Ashkenazi, Moroccan and Sephardi Jews. The second group, comprising the Middle Eastern and Caucasus Jewish communities, is positioned within the large conglomerate of non-Jewish populations of the region. The third group contains only a tight cluster of Yemenite Jews.

After elucidation of these groupings by PCA, we turned to structure-like analysis[26] with the algorithm ADMIXTURE[27] to assign individuals proportionally to hypothetical ancestral populations (Supplementary Note 3). Initially, all Jewish samples were analysed jointly with 25 novel reference populations (Supplementary Note 1) in combination with the Human Genome Diversity Panel[18] samples representing Africa, the Middle East, Europe, and central, south and east Asia (Fig. 3 and Supplementary Fig. 4). This analysis significantly refines and reinforces the previously proposed partitioning of Old World population samples into continental groupings[18,19] (Supplementary Fig. 4 and Supplementary Note 4). We note that membership of a sample in a component that is predominant in, but not restricted to, a specific geographic region is not sufficient to infer its genetic origins. Membership in several genetic components can imply either a shared genetic ancestry or a recent admixture of sampled individuals[18,28]. An illustrative example at $K = 8$ (Fig. 3 and Supplementary Note 3) is the pattern of membership of Ashkenazi, Caucasus (Azerbaijani and Georgian), Middle Eastern (Iranian and Iraqi), north African (Moroccan), Sephardi (Bulgarian and Turkish) and Yemenite Jewish communities in the light-green and light-blue genetic components, which is similar to that observed for Middle Eastern non-Jewish populations, suggesting a shared regional origin of these Jewish communities. This inference is consistent with historical records describing the dispersion of the people of ancient Israel throughout the Old World[1–4]. Our conclusion favouring common ancestry over recent admixture is further supported by the fact that our sample contains individuals that are known not to be admixed in the most recent one or two generations. It is also evident that among the Ashkenazi, Moroccan and Sephardi Jewish communities the dark-blue component dominating European populations is more substantial than the corresponding proportion of this component among the Middle Eastern Jewish communities (Fig. 3). For the Indian and Ethiopian Jewish communities the dark-green and light-brown genetic components are consistent with corresponding membership of their respective host populations (Fig. 3). ADMIXTURE was also run on the west Eurasian subset of the Old World sample, which highlights differentiation between the Middle East and Europe (Supplementary Fig. 4b). Here, comparison between the ADMIXTURE-derived component patterns for Sephardi and Ashkenazi Jews shows that the former have only slightly greater similarity to the pattern observed for Middle Eastern populations than do the latter.

Genetic relationships between our population samples were then explored with the measure of allele sharing distances (ASDs)[29]. Table 1 provides genetic distances between each Jewish community and its corresponding host population, all Jewish communities, west Eurasian Jewish communities, their respective Jewish group inferred from the PCA, and non-Jewish Levantine populations. The Ashkenazi, Sephardi, Moroccan, Iranian, Iraqi, Azerbaijani and Uzbekistani Jewish communities have the lowest ASD values when compared with their PCA-based inferred Jewish sub-cluster (Fig. 3 and Supplementary Figs 2c and 3). In all except the Sephardi Jewish community, this ASD difference is statistically significant ($P < 0.01$, bootstrap t-test). ASD values between Ashkenazi, Sephardi and Caucasus Jewish populations and their respective hosts are lower than those between each Jewish population and non-Jewish populations from the Levant. This might be the result of a bias inherent in our calculations as a result of the genetically more diverse non-Jewish populations of the Levant. The Ethiopian and Indian Jewish communities show the lowest ASD values when compared with their host population (Supplementary Tables 2 and 3 and Supplementary Note 5).

Although uniparental markers[8,9] (Supplementary Note 6) are limited in their capacity to uncover genetic substructure within the Middle East, they do provide important insights into sex-specific processes that are not unambiguously evident from the autosomal data alone. For example, Y-chromosome data point to a unique paternal genetic link between the Bene Israel community and the Levant, whereas the absence of sub-Saharan African maternal lineages in Yemenite and Moroccan Jews (in contrast to their hosts) suggests limited maternal gene flow.

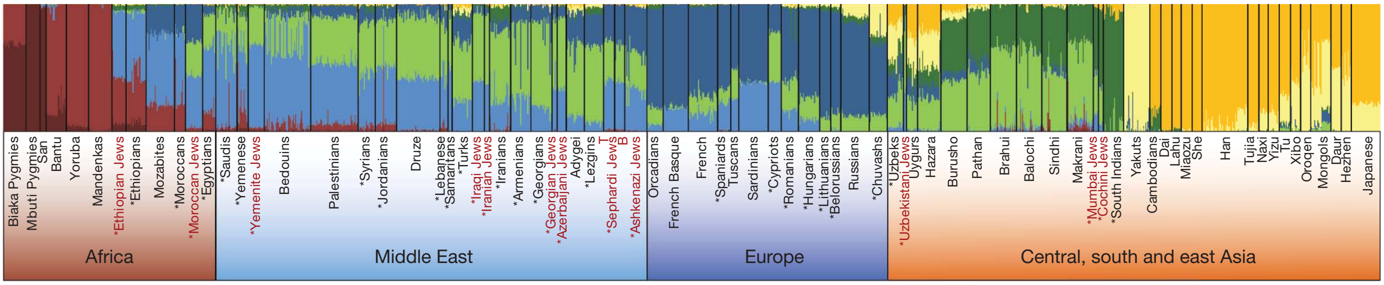

Figure 3 | Population structure inferred by ADMIXTURE analysis. Each individual is represented by a vertical (100%) stacked column of genetic components proportions shown in colour for $K = 8$. The Jewish communities are labelled in colour and bold. T and B further specify Sephardi Jews from Turkey and Bulgaria, respectively. Populations introduced for the first time in this study and analysed together with the Human Genome Diversity Panel[18] data are marked with an asterisk.

Table 1 | Genetic distances (ASD) between Jewish, Levantine and Diaspora host populations

Jewish community	Host population	Hosts	Levant*	All Jews	West Eurasian Jews†	Jewish cluster‡
Ashkenazi	Europe§	0.236	*0.239*‖	*0.240*	0.236	**0.235**
Sephardi	Spain	0.236	*0.238*	*0.239*	0.236	0.235
Moroccan	Morocco	0.246	**0.239**	**0.240**	**0.237**	**0.236**
Georgian	Georgia	0.234	*0.238*	*0.239*	*0.236*	*0.236*
Azerbaijani	Lezgin	0.238	*0.240*	*0.241*	0.238	**0.237**
Iranian	Iran	0.239	0.239	0.240	**0.237**	**0.236**
Iraqi	Syria, Iran	0.238	0.238	*0.239*	**0.236**	**0.236**
Uzbekistani	Uzbekistan	0.243	**0.238**	**0.239**	**0.236**	**0.235**
Bene Israel	India (Mumbai)	0.240	*0.245*	*0.245*	*0.243*	0.241
Cochini	India (Kerala)	0.238	*0.247*	*0.247*	*0.245*	*0.241*
Ethiopian	Ethiopia¶	0.245	*0.253*	*0.255*	*0.254*	
Yemenite	Yemen	0.243	0.238	0.240	*0.237*	

* Levant populations included Bedouin, Cypriots, Druze, Jordanians, Lebanese, Palestinians, Samaritans and Syrians.
† All Jewish populations excluding Ethiopian and Indian Jews.
‡ Jewish communities in the same cluster as obtained from the PCA analysis (Supplementary Fig. 3) are indicated by bold, italic or underlined type under the heading Jewish community.
§ Russians, Romanians, Hungarians, Belorussians, French and Lithuanians.
‖ Significance throughout the table: italic entries are significantly bigger than ASD from hosts (that is, further away), bold entries are significantly smaller than ASD from hosts; see Supplementary Table 3 for details.
¶ Amhara, Oromo and Tigray.

Our PCA, ADMIXTURE and ASD analyses, which are based on genome-wide data from a large sample of Jewish communities, their non-Jewish host populations, and novel samples from the Middle East, are concordant in revealing a close relationship between most contemporary Jews and non-Jewish populations from the Levant. The most parsimonious explanation for these observations is a common genetic origin, which is consistent with an historical formulation of the Jewish people as descending from ancient Hebrew and Israelite residents of the Levant. This inference underscores the significant genetic continuity that exists among most Jewish communities and contemporary non-Jewish Levantine populations, despite their long-term residence in diverse regions remote from the Levant and isolation from one another. This study further uncovers genetic structure that partitions most Jewish samples into Ashkenazi–north African–Sephardi, Caucasus–Middle Eastern, and Yemenite subclusters (Fig. 2). There are several mutually compatible explanations for the observed pattern: a splintering of Jewish populations in the early Diaspora period, an underappreciated level of contact between members of each of these subclusters, and low levels of admixture with Diaspora host populations. Equally interesting are the inferences that can be gleaned from more distant Diaspora communities, such as the Ethiopian and Indian Jewish communities. Strong similarities to their neighbouring host populations may have resulted from one or more of the following: large-scale introgression, asymmetrical sex-biased gene flow, or religious and cultural diffusion during the process of becoming one of the many and varied Jewish communities.

METHODS SUMMARY

Blood or buccal samples were collected with informed consent from unrelated volunteers who self-identified as members of one of the Jewish communities or non-Jewish populations studied here (Supplementary Note 1). The term 'Old World' refers to populations of the Eastern Hemisphere, specifically Europe, Asia and Africa. Whenever the term Jewish is not part of the population designation, this refers to a non-Jewish population. DNA samples chosen for the biparental analysis were genotyped on Illumina 610K or 660K bead arrays and showed a genotyping success rate of more than 97%. Data management and quality control were aided by PLINK 1.05 (ref. 30). For comparison, the relevant populations from the Illumina 650K-based data set of the Human Genome Diversity Panel, excluding relatives[18], were included in our analysis. After identification of the intersection of genotypes from the various Bead-Arrays, quality control (QC) and linkage disequilibrium (LD) pruning, a total of 226,839 autosomal single nucleotide polymorphisms (SNPs) remained for further analysis. PCA of autosomal variation using the smartpca of the EIGENSOFT package[24] was performed (Supplementary Note 2). Samples were modelled as comprising a mixture of major genetic components using the structure-like ADMIXTURE program[27], and the inferred genetic membership of each individual from this analysis was studied (Supplementary Notes 3 and 4). ASD[29] between groups was assessed, and a bootstrap procedure to determine the significance of differences in ASD between pairs of populations was adapted (Supplementary Note 5). Our uniparental data was merged with previously reported data sets for Y-chromosome and mtDNA analysis (Supplementary Note 6). A matrix of Y-chromosome and mtDNA haplogroup frequencies was constructed, and PCA was performed in the R environment (using the function princomp).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 December 2009; accepted 21 April 2010.
Published online 9 June 2010.

1. Ben-Sasson, H. H. *A History of the Jewish People* (Harvard Univ. Press, 1976).
2. De Lange, N. *Atlas of the Jewish World* (Phaidon Press, 1984).
3. Mahler, R. *A History of Modern Jewry* (Schocken, 1971).
4. Stillman, N. A. *Jews of Arab Lands: A History and Source Book* (Jewish Publication Society of America, 1979).
5. Della Pergola, S. in *Papers in Jewish Demography 1997* (eds Della Pergola, S. & Even, J.) 11–33 (The Hebrew University of Jerusalem, 1997).
6. Cavalli-Sforza, L. L., Menozzi, A. & Piazza, A. in *The History and Geography of Human Genes* 4 (Princeton Univ. Press, 1994).
7. Bauchet, M. *et al.* Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* **80,** 948–956 (2007).
8. Behar, D. M. *et al.* Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. *PLoS ONE* **3,** e2062 (2008).
9. Hammer, M. F. *et al.* Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl Acad. Sci. USA* **97,** 6769–6774 (2000).
10. Kopelman, N. M. *et al.* Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. *BMC Genet.* **10,** 80 (2009).
11. Need, A. C., Kasperaviciute, D., Cirulli, E. T. & Goldstein, D. B. A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. *Genome Biol.* **10,** R7 (2009).
12. Olshen, A. B. *et al.* Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. *BMC Genet.* **9,** 14 (2008).
13. Ostrer, H. A genetic profile of contemporary Jewish populations. *Nature Rev. Genet.* **2,** 891–898 (2001).
14. Price, A. L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4,** e236 (2008).
15. Seldin, M. F. *et al.* European population substructure: clustering of northern and southern populations. *PLoS Genet.* **2,** e143 (2006).
16. Tian, C. *et al.* Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* **4,** e4 (2008).
17. Abdulla, M. A. *et al.* Mapping human genetic diversity in Asia. *Science* **326,** 1541–1545 (2009).
18. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
19. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451,** 998–1003 (2008).
20. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
21. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
22. Biswas, S., Scheinfeldt, L. B. & Akey, J. M. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* **84,** 641–650 (2009).
23. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–1044 (2009).

24. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
25. Hourani, A. *A History of the Arab Peoples* (Faber & Faber, 1991).
26. Weiss, K. M. & Long, J. C. Non-Darwinian estimation: my ancestors, my genes' ancestors. *Genome Res.* **19,** 703–710 (2009).
27. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
28. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
29. Gao, X. & Martin, E. R. Using allele sharing distance for detecting human population stratification. *Hum. Hered.* **68,** 182–191 (2009).
30. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the individuals who provided DNA samples for this study, including the National Laboratory for the Genetics of Israeli Populations; Mari Nelis, Georgi Hudjashov and Viljo Soo for conducting the autosomal genotyping; Lauri Anton for computational help. R.V. and D.M.B. thank the European Commission, Directorate-General for Research for FP7 Ecogene grant 205419. R.V. thanks the European Union, Regional Development Fund through a Centre of Excellence in Genomics grant and the Swedish Collegium for Advanced Studies for support during the initial stage of this study. E.M. and Si.R. thank the Estonian Science Foundation for grants 7858 and 7445, respectively. K.S. thanks the Arthur and Rosalinde Gilbert Foundation fund of the American Technion Society. Sa.R. thanks the European Union for Marie Curie International Reintegration grant CT-2007-208019, and the Israeli Science Foundation for grant 1227/09. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partly supported by Fundação para a Ciência ea Tecnologia, the Portuguese Foundation for Science and Technology.

Author Contributions D.M.B. and R.V. conceived and designed the study. B.B.T., D.C., D.G., D.M.B., E.K.K., G.C., I.K., L.P., M.F.H., O.B., O.S., T.P. and R.V. provided DNA samples to this study. E.M., J.P. and G.Y. screened and prepared the samples for the autosomal genotyping. D.M.B., E.M., G.C., M.F.H. and Si.R. generated and summarized the database for the uniparental analysis. B.Y., M.M. and Sa.R. designed and applied the modelling methodology and statistical analysis. T.P. provided expert input regarding the relevant historical aspects. B.Y., D.M.B., K.S., M.F.H., M.M., R.V. and Sa.R. wrote the paper. B.Y., D.M.B. and M.M. contributed equally to the paper. All authors discussed the results and commented on the manuscript.

Author Information The array data described in this paper are deposited in the Gene Expression Omnibus under accession number GSE21478. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.M.B. (behardm@usernet.com), K.S. (skorecki@tx.technion.ac.il) or R.V. (rvillems@ebc.ee).

METHODS

Sample collection. All samples reported here were derived from a buccal swab or blood cells collected with informed consent in accordance with protocols approved by the National Human Subjects Review Committee in Israel and Institutional Review boards of the participating research centres. Participants were recruited during scheduled archaeogenetics lectures addressing the general public, genealogical societies, heritage centres and the scientific community. Each volunteer reported ancestry by providing information on the origin of all four grandparents. Samples were also obtained from the National Laboratory for the Genetics of Israeli Populations (http://www.tau.ac.il/medicine/NLGIP/). Comparative data sets for the uniparental and biparental analysis were assembled from the literature as summarized in Supplementary Note 1 and Supplementary Tables 1 and 4 and 5.

Genotyping autosomal markers. Illumina 610K or 660K bead arrays were used for genotyping with standard protocols, and Bead Studio software was used to assign genotypes. PLINK 1.05 (ref. 30) was used to perform data management and QC operations. Samples and SNPs with success rates of less than 97% were excluded. A total of 475 novel samples were analysed, 121 of which were from 14 Jewish communities representing most of the known geographic range of Jews during the past 100 years. The other 354 samples were chosen from 27 non-Jewish populations to enable paired analysis with the Jewish sample set. For comparison, relevant populations were further included (Supplementary Table 1) from the Illumina 650K-based data set of the Human Genome Diversity Panel after excluding relatives as in ref. 18. Because background LD can distort both PCA[24] and structure-like analysis[27] results, one member of any pair of SNPs in strong LD ($r^2 > 0.4$) in windows of 200 SNPs (sliding the window by 25 SNPs at a time) was removed using indep-pairwise in PLINK. After identifying the intersection of genotypes from the two types of bead array (Illumina 610K and 660K), QC and LD pruning, a total of 226,839 autosomal SNPs were chosen for all autosomal analyses.

Principal component analysis. PC analysis was performed with the smartpca program of the EIGENSOFT package[24]. To express the relative importance of the top two eigenvectors in the resulting PC plot, two axes were scaled by a factor equal to the square root of the corresponding eigenvalue (Supplementary Note 2). Our analysis was repeated for the entire set of populations and for the subset of west Eurasian populations (Supplementary Table 1). The R environment was used to perform PCA (using the function princomp) and plot the results for all analyses of uniparental data.

Structure-like analysis. The recently introduced structure-like approach was applied as assembled in the program ADMIXTURE[27] (Supplementary Notes 3 and 4). ADMIXTURE was run on our global and west Eurasian data sets 100 times in parallel at $K = 2$ to $K = 10$ (using random seeds). Convergence between independent runs at the same K was monitored by comparing the resulting log-likelihood scores (LLs). The minimal variation in LLs (less than 1 LL unit) within a fraction (10%) of runs with the highest LLs was assumed to be a reasonable proxy for inferring convergence[28]. In the global data set, convergence was observed in the case of all explored K values ($K = 2$ to $K = 10$). Results from runs at all values of K are shown rather than restricting the reader to one chosen K (Supplementary Note 3). To focus on population structure in the relevant regions of the Middle East and Europe we performed analyses on a data set restricted to west Eurasian samples. In this analysis, convergence was reached at $K = 2$ to $K = 5$; $K = 7$ and $K = 8$. Only $K = 4$ was highlighted in Supplementary Fig. 5 because components appearing at higher values of K were predominantly restricted to a single population and were therefore less informative for our purposes. Judging from the distribution of LLs of the converged K values, the maximum-likelihood solutions with LLs very close to the highest LLs were also the most frequent solutions (except for $K = 6$ of the global data set). One run from the top LLs fraction of each converged K (from global and west Eurasian data set) was plotted with Excel (Supplementary Fig. 4a, b).

Allele sharing distances. ASD was used for measuring genetic distances between populations. ASD is less sensitive to small sample size than the Fixation Index (F_{ST}) and other measures[29], and more appropriate for our goal of measuring genetic distances between groups regardless of their internal diversity. Standard errors of ASD values were calculated with a bootstrap approach, accounting for variance resulting from both sample selection and site selection. ASDs between individual Jewish populations and population groups representing a geographic region or ethnic group were calculated. In each case, the population under consideration was removed from all groupings with which it was compared. To test significance of differences in pairs of ASD values in each row in Table 1, a bootstrap approach was used (Supplementary Note 5 and Supplementary Tables 2 and 3).

Genotyping uniparental markers. Our data from the Y chromosome and mtDNA were combined with previously published data sets from populations of interest (Supplementary Note 6). Markers were chosen to match the phylogenetic level of resolution achieved in previously reported data sets. A total of 8,210 samples were assembled for Y-chromosome analysis (Supplementary Table 4). Genotypes for these sites were determined by using multiple techniques, such as allele-specific PCR, TaqMan, Kaspar and direct sequencing. A total of 13,919 samples were assembled for mtDNA analysis (Supplementary Table 5).