2014

Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins

S. Carmi

K. Y. Hui

F. Grady

S. Guha
Northwell Health

K. Upadhyay

See next page for additional authors

Authors
S. Carmi, K. Y. Hui, F. Grady, S. Guha, K. Upadhyay, D. Ben-Avraham, S. Mukherjee, B. M. Bowen, T. Lencz, I. Pe'er, and +22 additional authors

nature COMMUNICATIONS

ARTICLE

Received 24 Jun 2014 | Accepted 28 Jul 2014 | Published 9 Sep 2014

DOI: 10.1038/ncomms5835 OPEN

Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins

Shai Carmi[1], Ken Y. Hui[2], Ethan Kochav[1], Xinmin Liu[3], James Xue[1], Fillan Grady[1], Saurav Guha[4,5,6], Kinnari Upadhyay[7], Dan Ben-Avraham[7,8], Semanti Mukherjee[4,5], B. Monica Bowen[2], Tinu Thomas[9,10], Joseph Vijai[9,10], Marc Cruts[11], Guy Froyen[12], Diether Lambrechts[13], Stéphane Plaisance[14], Christine Van Broeckhoven[11], Philip Van Damme[13,15], Herwig Van Marck[14], Nir Barzilai[7,8], Ariel Darvasi[16], Kenneth Offit[9,10], Susan Bressman[17], Laurie J. Ozelius[6], Inga Peter[6], Judy H. Cho[2], Harry Ostrer[7,18], Gil Atzmon[7,8], Lorraine N. Clark[3,19], Todd Lencz[4,5,20] & Itsik Pe'er[1,21]

The Ashkenazi Jewish (AJ) population is a genetic isolate close to European and Middle Eastern groups, with genetic diversity patterns conducive to disease mapping. Here we report high-depth sequencing of 128 complete genomes of AJ controls. Compared with European samples, our AJ panel has 47% more novel variants per genome and is eightfold more effective at filtering benign variants out of AJ clinical genomes. Our panel improves imputation accuracy for AJ SNP arrays by 28%, and covers at least one haplotype in ≈67% of any AJ genome with long, identical-by-descent segments. Reconstruction of recent AJ history from such segments confirms a recent bottleneck of merely ≈350 individuals. Modelling of ancient histories for AJ and European populations using their joint allele frequency spectrum determines AJ to be an even admixture of European and likely Middle Eastern origins. We date the split between the two ancestral populations to ≈12–25 Kyr, suggesting a predominantly Near Eastern source for the repopulation of Europe after the Last Glacial Maximum.

[1] Department of Computer Science, Columbia University, 500 W 120th Street, New York, New York 10027, USA. [2] Department of Internal Medicine, Genetics & Pediatrics, Yale School of Medicine, 300 Cedar Street, New Haven, Connecticut 06519, USA. [3] Department of Pathology and Cell Biology, Columbia University Medical Center, 1150 St Nicholas Avenue, New York, New York 10032, USA. [4] Center for Psychiatric Neuroscience, The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York 11030, USA. [5] Department of Psychiatry, Division of Research, The Zucker Hillside Hospital Division of the North Shore-Long Island Jewish Health System, Glen Oaks, New York 11004, USA. [6] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, New York 10029, USA. [7] Department of Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. [8] Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. [9] Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [10] Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [11] VIB Department of Molecular Genetics, University of Antwerp, Universiteitsplein 1, 2610 Antwerpen, Belgium. [12] VIB Center for the Biology of Disease, KU Leuven, Herestraat 49, bus 602, 3000 Leuven, Belgium. [13] VIB Vesalius Research Center, KU Leuven, Herestraat 49, bus 912, 3000 Leuven, Belgium. [14] VIB BioInformatics Training and Services facility, Rijvisschestraat 120, 9052 Gent, Belgium. [15] Neurology Department, University Hospital Leuven, 3000 Leuven, Belgium. [16] Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem 91904, Israel. [17] Department of Neurology, Beth Israel Medical Center, New York, New York 10003, USA. [18] Department of Pathology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. [19] Taub Institute for Research of Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, 1150 St Nicholas Avenue, New York, New York 10032, USA. [20] Departments of Psychiatry and Molecular Medicine, Hofstra University School of Medicine, Hempstead, New York 11550, USA. [21] Center for Computational Biology and Bioinformatics, Columbia University, 1130 St Nicholas Avenue, New York, New York 10032, USA. Correspondence and requests for materials should be addressed to T.L. (email: tlencz@nshs.edu) or to I.P. (email: itsik@cs.columbia.edu).

Ashkenazi Jews (AJ), identified as Jewish individuals of Central- and Eastern European ancestry, form the largest genetic isolate in the United States. AJ demonstrate distinctive genetic characteristics[1,2], including high prevalence of autosomal recessive diseases and relatively high frequency of alleles that confer a strong risk of common diseases, such as Parkinson's disease[3] and breast and ovarian cancer[4]. Several recent studies have employed common polymorphisms[5–13] to characterize AJ as a genetically distinct population, close to other Jewish populations as well as to present-day Middle Eastern and European populations. Previous analyses of recent AJ history highlighted a narrow population bottleneck of only hundreds of individuals in late medieval times, followed by rapid expansion[12,14].

The AJ population is much larger and/or experienced a more severe bottleneck than other founder populations, such as Amish, Hutterites or Icelanders[15], whose demographic histories facilitated a steady stream of genetic discoveries. This suggests the potential for cataloguing nearly all founder variants in a large extant population by sequencing a limited number of samples, who represent the diversity in the founding group (for example, ref. 16). Such a catalogue of variants can make a threefold contribution: First, it will enable clinical interpretation of personal genomes in the sizeable AJ population by distinguishing between background variation and recent, potentially more deleterious mutations. Second, it will improve disease mapping in AJ by increasing the accuracy of imputation. Third, the ability to extensively sample a population with ancient roots in the Levant is expected to provide insights regarding the histories of both Middle Eastern and European populations.

Here we report a catalogue of 128 high coverage, whole-genome AJ sequences. Compared with a European reference panel, the AJ panel has more novel and population-specific variants, and we demonstrate that the AJ panel is necessary for interpretation and imputation of AJ personal genomes. Analysis of long shared segments, which are abundant in AJ, confirms a recent severe bottleneck and potential utility in future sequencing studies. The joint AJ–European allele frequency spectrum suggests that the AJ population is an even mix of European and Middle Eastern ancestral populations and quantifies ancient bottlenecks and population splits. Finally, we report the deleterious mutation load in AJ to be slightly higher than in Europeans.

Results

We sequenced a panel ($n=128$) of controls of self-reported and empirically validated AJ ancestry (Supplementary Note 1; Supplementary Table 1; Supplementary Fig. 1). The high coverage sequence ($>50\times$), generated by Complete Genomics[17], showed multiple quality control (QC) indicators supporting both high quality and completeness of the single-nucleotide variant (SNV) data: 97% coverage of the genome (Supplementary Note 2; Supplementary Table 2), inferred discordance of 0.047% to high quality genotypes in SNP arrays (Supplementary Note 2; Supplementary Data 1), transition/transversion ratio of 2.14, and consistency of QC measures across potential sources of bias (Supplementary Note 2; Supplementary Data 2; Supplementary Figs 1 and 2). The average raw number of non-reference SNVs called per individual was 3.412 M, including 10.5K coding synonymous changes and 9.7K non-synonymous ones (Supplementary Data 2). An additional 538K multinucleotide variants, 4.1K mobile element insertions, and 302 copy number variants (spanning 6.7 Mbp) were observed, on average, in each sample (Supplementary Data 2). However, inspection of novel non-SNVs demonstrated high false-positive rates (Supplementary Note 2), and we thus focused on autosomal, bi-allelic SNVs for all subsequent analyses. We applied strict multisample filters (Supplementary Note 2) to generate a working set of 12,326,197 high quality SNVs, of which 2,891,414 were novel (23.5%; dbSNP135). Quality was gauged by a sequenced duplicate as well as runs-of-homozygosity, which are sufficiently frequent in AJ for this purpose, providing estimates of $\approx$6,000–8,000 false positives genome wide (Supplementary Note 2), in line with previous benchmarks of this technology[17]. Principal component analysis of common variants in the sequenced AJ samples confirmed previous observations[5,6,9,10], namely, that AJ form a distinct cluster with proximity to other Jewish, European and Middle Eastern populations (Supplementary Fig. 1).

Our reference panel is expected to improve the ability to catalogue variants and haplotypes in the Ashkenazi population, beyond what is possible with non-ancestry-matched reference samples. A natural panel for comparison would be the European samples from the 1000 Genomes Project[18]. However, to match the high depth of our data and the sequencing platform used to obtain it, we chose as our primary comparison data set a cohort of Flemish (FL) personal genomes ($n=26$) from Belgium (Supplementary Note 2). We merged our first batch of AJ genomes ($n=57$) with the FL data, applying a QC pipeline attempting to remove all potentially artifactual population-specific variants (Supplementary Note 2). The merged, post-QC data set included 10,499,312 SNVs for comparative analysis.

Comparison of tallies of variants between AJ and FL genomes (Fig. 1a; Supplementary Table 3) suggested that AJ have slightly but significantly more overall variants ($+1.5\%$), mostly as heterozygotes. The increased AJ heterozygosity ($+2.4\%$), in spite of the recent bottleneck, confirms previous observations (Supplementary Note 3)[6,7,10,19]. More pertinently to the utility of a population sequencing endeavour, AJ samples have a much higher fraction ($+47\%$) of novel variants (dbSNP135; Fig. 1a). Clinical AJ genomes will thus be screened more efficiently against the AJ reference panel. For example, an AJ genome has, on average, 36,995 novel variants (160 of which are also non-synonymous). Only 4.0% of them (3.2% for novel and non-synonymous) will be filtered out against the FL panel, whereas an AJ panel of the same size filters out 32.6% of variants (22.4%), 8.2 (7.0) times more. Using the entire AJ panel allows filtering of $\approx$65% of all novel variants (48%). The number of novel and non-synonymous, never-seen variants in an AJ personal genome is therefore only 83.3, making the clinical analysis of such a genome more feasible (Fig. 1b). The number of new variants discovered when sequencing each additional genome is slightly larger in our AJ cohort than in FL (Fig. 1c). However, extrapolation predicts the converse trend already for cohorts larger than $n=49$ samples (Fig. 1c; Supplementary Note 3; Supplementary Fig. 3), suggesting higher efficiency of the AJ cohort in cataloguing population variation.

The effective coverage of variation can also be demonstrated using identical-by-descent (IBD) segments. We detected IBD segments by using the Germline software[20], with additional filtering adapted to sequencing data (Supplementary Note 4; Supplementary Fig. 4). Sharing in AJ was $\geq$7.9-fold more abundant than in FL or between the populations (Fig. 2a). Using the AJ panel, one can cover at least one haplotype in $\approx$67% of the genome of any other AJ individual with long (>3 cM) IBD segments ($\approx$46% using segments >5 cM), compared with much poorer efficiency in Europeans (Fig. 2b; here we used the CEU panel from the 1000 Genomes project; Supplementary Note 4). These results imply that any additional, sparsely genotyped AJ sample can be effectively imputed, at least partially, along haplotypes shared with a small sequenced reference panel. Co-ancestry of copies of IBD segments is expected to be extremely

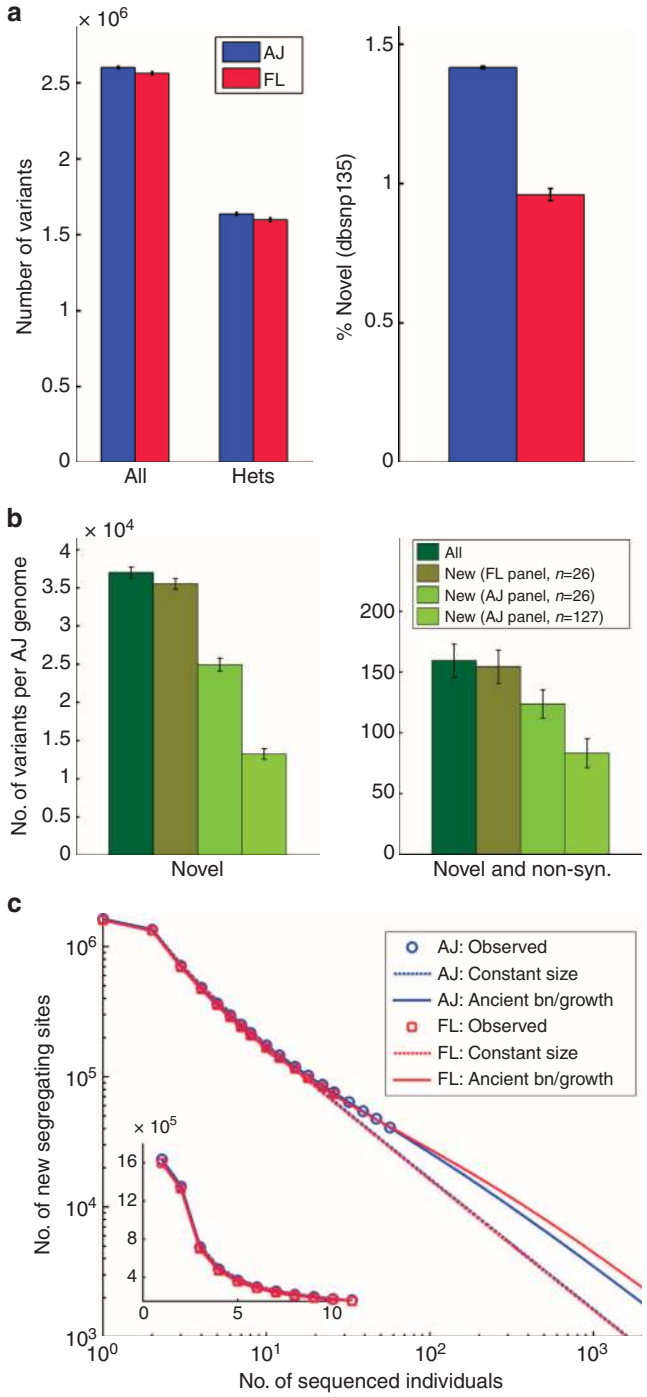

Figure 1 | Novel variants discovered in Ashkenazi Jewish and Flemish genomes. (**a**) Variant counts (all and heterozygous; left) and fraction novel (right) per genome in the Ashkenazi Jewish (AJ) and Flemish (FL) cohorts (corresponding to about ≈80% of the raw variants remaining after QC and cohort merging; Supplementary Note 2; error bars represent s.d.). (**b**) Efficiency of filtering all novel variants detected in an AJ personal genome, measured by counting those that remain new after filtering such a genome against either FL or AJ panels of a matched size ($n=26$) or our complete AJ panel ($n=127$). Left: all novel variants; right: non-synonymous novel variants. Error bars represent s.d. (**c**) The number of newly discovered segregating sites in AJ and FL versus the number of already sequenced individuals in each cohort (markers). Dashed and solid lines are expectations based on either a constant size or a bottleneck and growth model (bn/growth), respectively, fitted to each population separately (Supplementary Note 3). The inset magnifies the region (0, 10).

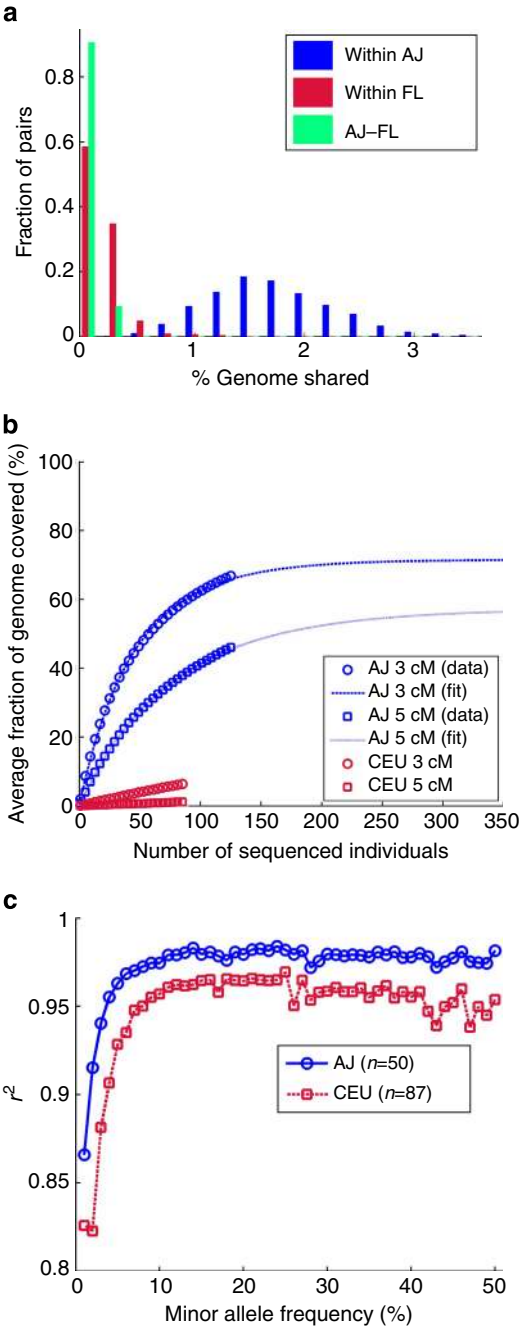

Figure 2 | Utility of the AJ reference panel in IBD-based and traditional imputation. (**a**) The distribution, over all pairs of individuals, of the fraction of the genome shared IBD (segment lengths >3 cM) either within AJ, within FL or between AJ and FL. (**b**) The average fraction of a genome (in AJ and CEU) where at least one haplotype is covered by segments shared with a population-matched panel. Data points (markers) were fit to $c = 1-[1-c_{max}(1-e^{-n/n_0})]^2$ (lines), where c is the average coverage and n is the number of individuals in the panel (Supplementary Note 4). (**c**) The aggregate r^2 (over the AJ study genomes) between the true and the imputed dosages versus the minor allele frequency, when imputing an AJ genome using a reference panel consisting of either AJ or CEU genomes.

recent (typically 30 or fewer generations), thus allowing only very recent mutations to be missed at the imputed genome[21,22]. Whether this strategy will scale for the accurate imputation of the entire genome of an AJ proband will be resolved with the sequencing of additional genomes.

Our sequencing panel is also expected to improve the performance of traditional imputation approaches, which are known to be more accurate when the ancestries of the reference and target populations are matched[23]. To evaluate the quality of imputation, we divided our sequencing cohort into 'reference' and 'study' panels; in the latter, we masked all variants not genotyped on a typical SNP array. We then imputed[24] the 'study' panel using either our 'reference' panel ($n=50$) or the larger ($n=87$) 1000 Genomes CEU panel[18] (Supplementary Note 5; Supplementary Fig. 5). As expected, using an AJ reference panel was more accurate than using a European one, with the number of discordant genotypes 28% lower and the correlation between true and imputed dosages, r^2, increasing from 97.4% to 98.2% (Supplementary Note 5; Supplementary Table 4). Using the AJ panel reduced mostly the number of false negatives (with respect to the reference genome; Supplementary Table 4); it lowered the number of wrongly imputed non-reference variants with minor allele frequency $\leq$1% by 2.7-fold, with the improvement remaining at 1.5–2-fold at higher frequencies (Fig. 2c; Supplementary Fig. 6). This improvement in imputation quality likely reflects both the increased segmental sharing in AJ as well as the large number of AJ-specific alleles. These results motivate using a population-matched, rather than a merely continent-matched, reference panel, even for the closely related AJ and European populations.

Our sequencing data also enables detailed reconstruction of AJ and European population histories. Allele frequency spectra (AFS) are attractive conduits for such an analysis, especially in deeply sequenced cohorts. The AFS of both AJ and FL (Fig. 3a) reject a constant-size population model, which has previously been ruled out across multiple human populations[25]. The two spectra are similar, with AJ showing a slight excess of doubletons. These spectra each fit well to similar models of ancient history, comprising an ancient bottleneck ($\approx$60–86 Kyr) followed by slow exponential growth (Supplementary Note 6; Supplementary Table 5; Supplementary Fig. 7; Supplementary Fig. 8). The joint (AJ–FL) AFS reveals correlated allele counts (Fig. 3b), indicating gene flow between the populations or very recent divergence (Supplementary Note 6). Yet, correlation is not as strong as it would have been had the AJ–FL combined sample been panmictic (Fig. 3b; $F_{ST}=0.016$; Supplementary Note 6). The normalized AFS of population-specific variants (Fig. 3a, inset) is noticeably different between AJ and FL, with higher allele frequencies in AJ. There were overall 14% more population-specific variants in AJ (Supplementary Note 6; Supplementary Figs 9 and 10), pointing to asymmetric gene flow from Europeans into the ancestral population of AJ.

We next turned to inferring an explicit model for the demographic history of AJ and Europeans. Since the allele frequency spectrum, in particular for our sample size, may not be sensitive to recent demographic events, we first reconstructed the very recent AJ history by examining long IBD segments[5,12,14,21], which carry information on recent co-ancestry (last $\approx$50 generations). We used the lengths of shared segments (Fig. 3c) to infer the parameters of a recent AJ bottleneck (effective size 250–420; 25–32 generations ago) followed by rapid exponential expansion (rate per generation 16–53%; Fig. 4, bottom),

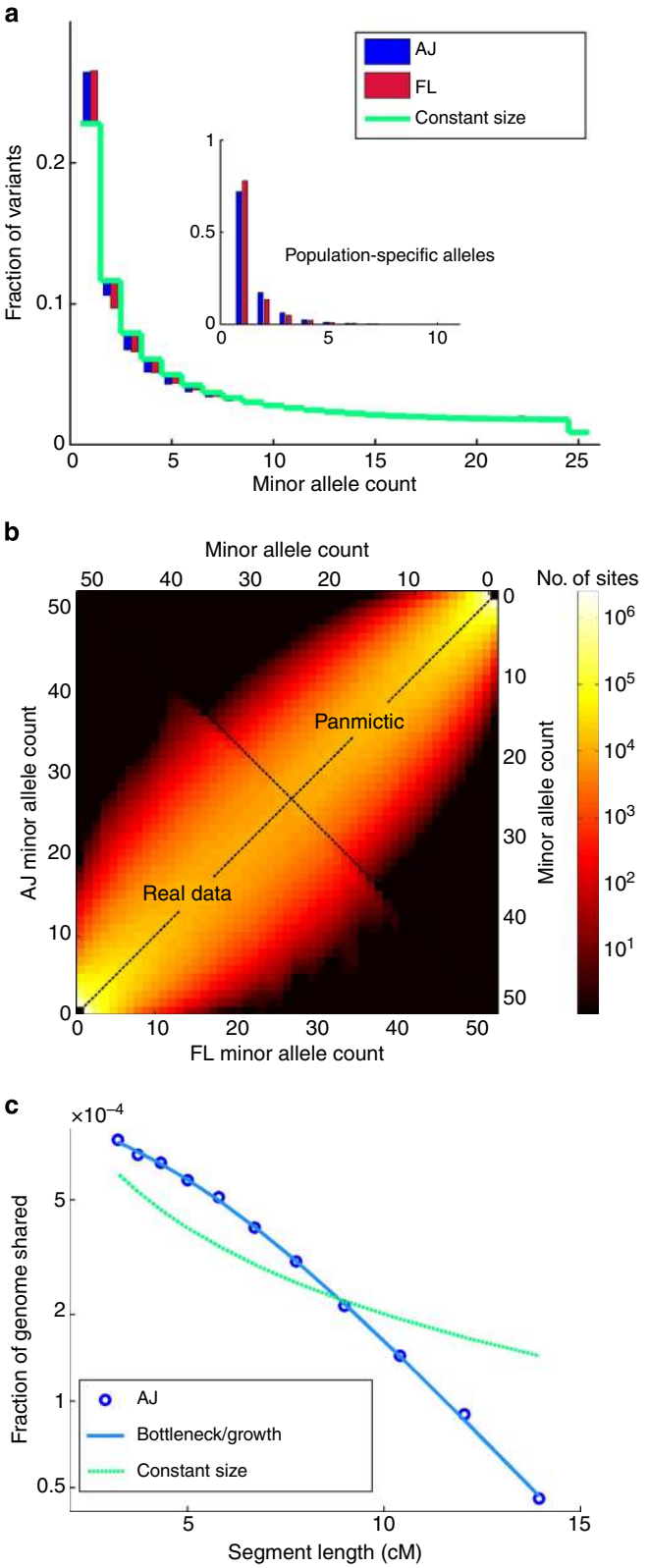

Figure 3 | The AFS and the lengths of shared segments. (**a**) The (normalized) minor allele frequency spectrum in AJ and FL, shown as counts in subsets of $n=25$ genomes in each cohort. The green line corresponds to the expectation in a constant-size population (Wright-Fisher), and bars represent deviations in AJ and FL. The inset shows the spectra of alleles private to each population. (**b**) A heat map of the joint (minor) allele frequency spectrum of AJ and FL (lower left triangle) compared with the expected joint AFS, had population labels been random (upper right triangle)[33]. (**c**) The average fraction of the genome found in shared segments versus the segment length (AJ only; circles), along with the best fit to a recent bottleneck and growth model (solid blue line; Fig. 4) and the expectation in a constant-size population with the same total sharing (dashed green line).

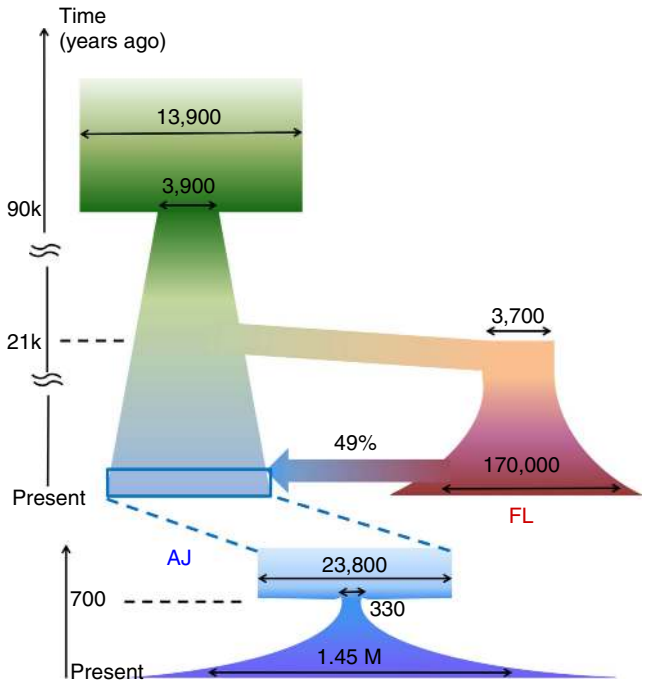

Figure 4 | A reconstruction of the AJ and FL demographic history. The upper part of the diagram shows the reconstruction of the ancient history by fitting the joint AFS (Fig. 3b) using $\partial a \partial i$[26] and using a mutation rate of 1.44×10^{-8} per generation per bp. The lower diagram shows the recent AJ history, reconstructed by fitting the IBD length decay pattern (Fig. 3c). The wide arrow represents an admixture event; all effective population sizes (horizontal arrows) are in number of diploid individuals; all times were computed assuming 25 years per generation. Confidence intervals are provided in Supplementary Tables 6 and 7.

confirming previous analyses conducted on lower throughput data (Supplementary Note 4; Supplementary Table 6; Supplementary Fig. 11)[12,14].

Given the model for the recent AJ history, we inferred the parameters of a model for the ancient history of AJ and FL using an existing method ($\partial a \partial i$[26]) based on the joint frequency spectrum (Supplementary Note 6; Supplementary Data 3). Confidence intervals were computed using parametric bootstrap[26] (Supplementary Note 6), but we did not integrate over the uncertainty in the mutation rate (see the next paragraph). According to the resulting model (Fig. 4, top; Supplementary Table 7; Supplementary Fig. 12), contemporary AJ formed 600–800 years (close to the time of the AJ bottleneck) as the fusion of two ancestral populations. One ancestral population, consistent with being the ancestors of the FL samples, contributed 46–50% of the AJ gene pool. We call that population ancestral European and the other ancestral Middle Eastern. The ancestral European population went through a founding bottleneck (effective size 3,500–3,900) when diverging from ancestral Middle Easterners. We date this event to 20.4–22.1 Kyr, at around the time of the Last Glacial Maximum and preceding the Neolithic revolution ([27]; see Supplementary Note 6 and below for discussion). The ancestors of both populations underwent a bottleneck (3,600–4,100 founders) at 85–94 Kyr, likely corresponding to an Out-of-Africa event[28].

The confidence intervals around our inferred parameters were remarkably small (Supplementary Table 7; coefficient of variation typically $\approx$2–5% and no more than $\approx$8%). However, any sampling noise in our historical reconstruction is negligible compared with possible inaccuracies in the human mutation rate or potentially oversimplified model assumptions. We verified that our main conclusions were robust to variations in the model's fine details (Supplementary Note 6). Conversely, all inferred times and population sizes depend inversely on the mutation rate, μ, and are thus highly sensitive to its precise value. The recent debate over the human mutation rate[28,29] has converged to estimates of μ ranging between $1.0–1.5 \cdot 10^{-8}$ (per generation per bp; obtained using next-generation sequencing of *de novo* mutations), compared with the traditional estimates (using the human–chimpanzee divergence) around $\mu_{phylo} \approx 2.5 \cdot 10^{-8}$. The mutation rate that we used was $\mu = 1.44 \cdot 10^{-8}$, estimated by Gravel *et al.*[30] by matching the relatively well-known time of the population of the Americas with the time of a bottleneck inferred from Native American whole-genome sequences. This estimate is relevant to our evolutionary time scale of interest, and is close to the '*de novo*' estimates[31] (see ref. 32 for a very recent review).

Previous explicit demographic models using genome-wide SNP arrays or low-pass sequencing data time-stamped a European bottleneck at $\approx$40–80 Kyr (recalibrated to the lower mutation rate estimate; Supplementary Note 6), with even the lowest estimates[26,33,34] being higher than our point estimate of $\approx$21 Kyr. However, no previous study has employed deeply sequenced genomes of (partial) Middle Eastern ancestry; in addition, previous studies usually modelled the European founder event simultaneously with the divergence from East Asian populations. As modern humans had colonized Europe already by $\approx$40–45 Kyr[35], our results (across all estimates of the mutation rate) support genetic discontinuity between that (hunter–gatherer) population and contemporary Europeans. A Middle Eastern European divergence time around $\approx$21 Kyr would also suggest (*i*) a near Eastern source for the repopulation of Europe at the end of the Last Glacial Maximum[27,36] and (*ii*) that migration from the Middle East to Europe largely preceded the Neolithic revolution, suggesting that Neolithic population movements were largely within Europe[37–42]. These interpretations, however, strongly depend on the mutation rate: taking into account the uncertainty in the mutation rate, our divergence time estimate is between $\approx$12–25 Kyr, which can be reconciled with Neolithic migrations originating in the Middle East (Supplementary Note 6).

We finally turned to the analysis of the functional elements of the genome. Historically, mapping disease mutations in the AJ population enabled the development of diagnostic panels. Here, our sequencing data allowed us to generate an extensive listing of variants in such genes (Supplementary Data 4, which also demonstrates the detection of carriers for 35 known disease mutations; Supplementary Note 7).

Recently, it was suggested that relaxation of negative selection constraints in bottlenecked populations increases their deleterious mutational burden[43–45]. We therefore looked for patterns of selective constraints at likely functional sites, taking advantage of the availability of non-coding regions as a control. We used again the platform-matched FL samples as a comparison cohort. As expected due to purifying (negative) selection, variants of increasing functional importance appear in lower frequencies in both AJ and FL, but not significantly differently between the populations (Supplementary Note 7; Supplementary Figs 13 and 14). A comparison of the functional mutation load showed slightly increased load in AJ compared with FL (Supplementary Note 7; Supplementary Table 8), consistently with the bottleneck hypothesis. Specifically, the observed number of non-reference, non-synonymous variants in AJ was 0.50% higher than expected based on population differences in neutral variation ($P = 0.006$; Supplementary Note 7; see also Supplementary Fig. 15). We note, however, that the effect is weak and the significance is sensitive to the precise definition of deleterious variation

(Supplementary Note 7). A genome-wide GERP analysis similarly showed that AJ variants overlap with slightly more conserved sites ($P=0.01$; Supplementary Note 7). In conclusion, we observed increased deleterious mutation load in AJ, but the effect is very limited, compared, for example, with French Canadians[43]. Ongoing progress in theory (for example, ref. 46) and data analysis methods is expected to elucidate this difference as well as lead to more decisive results for the AJ load.

Finally, as a number of diseases show higher prevalence in AJ[1], we sought to determine whether there are specific disease categories overabundantly affected by non-synonymous variation[47] (Supplementary Note 7). While a few categories showed higher mutational load than others (Supplementary Table 9), none reached false discovery rate <0.05 (at least in our relatively small sample size).

The AJ population has so far played an important role in human genetics, with notable successes in gene mapping[48,49] as well as prenatal and cancer screening. We have demonstrated that the narrow AJ bottleneck, of just a few hundred individuals, facilitates cost-effective cataloguing of the vast majority of (prebottleneck) AJ variation, even considering the currently large size of this population. It also suggests an increased power to detect rare alleles of large effect that drifted to higher frequencies during the bottleneck (Supplementary Note 8). This is in line with the recent success of detecting such alleles in other isolated populations[16,50,51] and motivates continued studies focusing on such cohorts.

Methods

Sample selection and sequencing. Samples were selected among controls of a longevity study[52] (Albert Einstein College of Medicine; $n=74$) and a Parkinson's study[53,54] (Columbia University Medical Center; $n=54$). The average age was 69 years. Some medically relevant phenotypes are given in Supplementary Table 1. Genotype data were used to validate Ashkenazi ancestry and the absence of cryptic relatedness. Informed consent was obtained in accordance with institutional policies and the study was approved by the corresponding institutional review boards. Sequencing was carried out by Complete Genomics, to average coverage $>50\times$, in three batches (Supplementary Note 1).

QC and processing pipeline. Raw sequencing summary statistics are reported per sample and per batch in Supplementary Data 2. Copy number variants and mobile element insertions were also reported; however, the false-positive rate was high (see below and Supplementary Note 2). All samples were previously genotyped on SNP arrays; concordance was measured using CGA tools and averaged 99.67% over all samples. The discordance was correlated with the array missingness, but not with sequencing metrics; extrapolating to the limit of no array missingness, the discordance approached 0.047% (Supplementary Note 2).

Genotypes calls across individuals were merged using CGA tools and converted to VCF or *Plink*[55] formats. Some of the analyses were carried out on 57 genomes sequenced in the first batch. Otherwise, we used the entire cohort ($n=128$). The merged genotypes were filtered by removing low quality and half-called variants, multiallelic and multinucleotide variants, variants not called as non-reference in any genome, variants with a no-call rate $>10\%$ (6% for the first batch), variants not in Hardy–Weinberg equilibrium ($P<10^{-6}$), and variants outside the autosomes. For some analyses, we excluded a single genome containing an exceptional amount (≈ 200 MB) of runs-of-homozygosity. We validated that monomorphic non-reference variants that we observed were monomorphic (or high frequency) in Complete Genomics' and 1000 Genomes' public sequencing data sets (Supplementary Note 2).

To validate the Ashkenazi ancestry of our samples, we merged the AJ data set with Middle Eastern and European individuals from HGDP[56] and with the Jewish HapMap project[5]. After pruning SNPs in LD (leaving $\approx$48K SNPs), we ran smartPCA[57]. The PCA plot (Supplementary Fig. 1) demonstrates the absence of either outliers or any batch effect (Supplementary Note 2). We also verified the absence of cryptic relatedness (maximum pairwise $\hat{\pi}$ (*Plink*) was $\approx 5.5\%$).

We estimated the false-positive rate using runs-of-homozygosity (inside which almost all heterozygous sites are due to errors), which we detected using *Plink*, after removing low frequency variants and LD pruning. We used high- and low-confidence sets of runs-of-homozygosity to obtain a lower and an upper bound, respectively, for the false-positive rate. After trimming each segment, we estimated the false-positive rate using the number of heterozygote sites along the segment (all variants or SNVs only, and in the original genotype calls or in the cleaned data set). There were overall $\approx$300–600 MB found in autozygous segments, harbouring a few thousands of heterozygous sites. Cleaning reduced the SNV false-positive rate by $\approx$3–4 fold to an extrapolated $\approx$6–8K per genome. The false-positive rate for non-SNVs was ≈ 6 times that of SNVs. We obtained an independent estimate of the error rate using a pair of duplicate genomes, reaching qualitatively similar conclusions (Supplementary Note 2).

The FL samples were mixed controls and cases from VIB in Ghent, Belgium. They were sequenced to coverage $\approx 70\times$ by Complete Genomics, albeit using an earlier pipeline compared with the AJ genomes. PCA validated the FL ancestry (Supplementary Note 2; Supplementary Fig. 1). The FL genomes ($n=26$) were merged and cleaned using a pipeline similar to that of the AJ genomes. We merged the cleaned FL genotypes with the cleaned genotypes of the 57 AJ genomes sequenced in the first batch. We removed any variants that appeared in the cleaned genotypes in one population, but were removed during QC in the other population (Supplementary Note 2), to avoid spurious population-specific variants. We phased the merged data set using *SHAPEIT*[58], with parameters as recommended by the authors, and with the 1000 Genomes reference panel. We used the molecular phasing information (that is, linked heterozygotes calls) to estimate the switch error rate at $\approx 0.95\%$ ($\approx 0.3\%$ for non-singletons). The merged and phased AJ–FL data set was used for most population comparisons.

Annotations. dbSNP annotations were from the UCSC Genome Browser[59]. Functional annotation for Fig. 1b was generated using *ANNOVAR*[60]. In Fig. 1b, the reported counts are means and s.d. over all AJ individuals. For each individual, we randomly selected a set of $n=26$ or $n=127$ other AJ individuals to serve as the reference panel.

Rate of variant discovery. The empirical rate of discovery of segregating sites in Fig. 1c is the average over 50 random orderings of the individuals in each cohort. The theoretical number of segregating sites for the Wright–Fisher model used an estimate of θ based on the average heterozygosity and standard coalescent theory[61]. For variable size populations, we used equations from[62] (Supplementary Note 3). The demographic model we used (for each population separately) is a bottleneck followed by an exponential expansion. The parameters were inferred by fitting the allele frequency spectrum using $\partial a \partial i$[26] (see below and Supplementary Note 6). The higher predicted number of FL sites was significant ($P<0.01$) with respect to parametric bootstrapping of the demographic models (Supplementary Note 3). A picture similar to Fig. 1c was seen when computing the rate of discovery of non-reference variants. There, projection to larger samples was on the basis of the first three entries of the allele frequency spectrum and the method of[33] (Supplementary Note 3; Supplementary Fig. 3).

The joint allele frequency spectrum. Initial inspection of the joint spectrum revealed a few thousands of highly differentiated variants (for example, AJ-specific variants of frequency $>50\%$). We suspected that those variants were due to reference genome mapping discrepancy (hg18/hg19), which we confirmed using Complete Genomics' public genomes resource (Supplementary Note 3). We therefore removed from further analysis $\approx$4,000 population-specific variants with frequency $>25\%$. To facilitate population-genetic comparisons, we downsampled the joint spectrum to 50 AJ and 50 FL haploid genomes analytically using hypergeometric expectations. We folded and marginalized the spectrum using standard definitions (Supplementary Note 3; minor alleles were defined with respect to the combined sample; Fig. 3b). The Wright–Fisher expected spectrum (Fig. 3a) was computed using standard coalescent theory[61]. The panmictic spectrum of Fig. 3b was computed analytically assuming that the appearances of each variant are randomly distributed between AJ and FL (Supplementary Note 3). F_{ST} was computed using $\partial a \partial i$[26].

IBD segment detection. To detect IBD segments, we first assigned genetic map distances using HapMap2 (ref. 63). We then ran Germline[20] using a minimal length cutoff of either 3 cM or 5 cM, and in the 'genotype extension' mode[12], which allows segments to extend as long as double homozygous sites are matching. We followed by filtering segments with particularly short physical length, overlap with sequence gaps or where all matching sites had the major allele. We further filtered segments by computing a score related to the probability of a segment to be truly shared-by-descent, on the basis of the allele frequencies of sites along the segment (Supplementary Note 4). Scores were higher for within-AJ segments than for within-FL or AJ–FL segments (Supplementary Fig. 4). In addition, most non-AJ sharing was concentrated in a handful of peaks (Supplementary Note 4), suggesting that many of the non-AJ detected segments were false positives.

Coverage of the genome by IBD segments. To create Fig. 2b, we considered sharing within-AJ (using all 128 individuals) and within-Europeans (FL or CEU from the 1000 Genomes Project) separately. For each hypothetical reference panel size n, we created a subset of size n of the full panel. For each individual in the subset, we computed the fraction of the genome (in physical distance) shared between that individual and the rest of the subset (which implies sharing of at least one of the haplotypes, but not necessarily both). We then averaged over all individuals in the subset and over 50 random subsets. The coverage curve was fitted to

the expectation from a simple model of a bottleneck lasting a single generation, with the population size being extremely large otherwise (Supplementary Note 4).

Demographic inference using IBD segments. We used the method developed in ref. 14. For each segment length bin, we summed the total length (in cM) of segments having length in the bin and divided by the total genome size and by the total number of (haplotype) pairs. The resulting curve (Fig. 3c) was fitted (by a grid search, minimizing the sum of squared (log-) errors) to a bottleneck and expansion model, with theoretical curves computed as in ref. 14. The constant population size estimator was computed as in ref. 21. The fitting error around the optimal parameters (Supplementary Fig. 11) showed deep minima around the optimal bottleneck time and population size, but less confidence in the values of the ancestral population size and the growth rate. Confidence intervals were obtained using jackknifing (Supplementary Table 6; Supplementary Note 4). Parametric bootstrap gave qualitatively similar results.

Imputation accuracy using the AJ panel. We split the 57 AJ genomes of the first batch (here phased using a variation of *SHAPEIT* that employs molecular phasing information (Supplementary Note 2)) into a reference panel ($n=50$) and a study panel ($n=7$). We reduced the study panel sequences to SNPs typically genotyped on an Illumina Human Omni1-Quad array, and supplemented them with 1000 SNP arrays of AJ controls from a Schizophrenia study[11,48], to emulate a typical imputation scenario. After standard QC procedures (Supplementary Note 5), we phased the entire study panel ($n=1007$) using *SHAPEIT*. We then imputed the study panel, on the basis of the AJ reference panel, using *IMPUTE2* (ref. 64). We also imputed using the CEU reference panel from 1000 Genomes ($n=87$, larger than the AJ panel). We carried out all analyses on chr1 only (Supplementary Note 5).

Imputation accuracy was measured by uncovering the full sequences of the AJ study genomes (Supplementary Table 4). Sites not imputed by the CEU panel were set as homozygous reference, and sites imputed by the CEU panel that were not found in the AJ sequences were (conservatively) discarded (Supplementary Note 5). Monomorphic non-reference sites in the AJ panel were also discarded. The squared correlation coefficient, r^2, was computed between the aggregate of all true genotypes (over all sites and study individuals) and all imputed dosages. Due to our small study panel, we computed the minor allele frequency (plotted in Fig. 2c and Supplementary Fig. 6) in the AJ reference panel ($n=50$). We excluded variants with frequency zero from these plots (leaving finally $\approx$200K variants per individual), since they are necessarily wrongly imputed. They were not removed from the overall accuracy reports (Supplementary Table 4).

Demographic inference using the allele frequency spectrum. We inferred the parameters of demographic models using $\partial a \partial i$[26]. For all models, we used a mutation rate of 1.44×10^{-8} per bp per generation[30] (based on the time of the human settlement in the Americas) and set the genome length to 2.685×10^{9} (autosomal hg19, excluding sequence gaps) times 0.81, which is an estimate of the fraction of variants remaining after cleaning (Supplementary Note 6). We estimated the scaled mutation rate, θ, by matching the number of segregating sites. The generation time we used was 25 years. We inferred single-population models using the individual AJ and FL spectra as well as two-population models using the joint spectrum (downsampled to 50×50 haploid genomes). In each case, the spectrum was fitted, using $\partial a \partial i$, with parameters as recommended by the authors (Supplementary Note 6). For each model, we experimented with different parameter regions until identifying a plausible parameter set. We then initiated the parameters to randomly perturbed values around that set. We repeated optimization with 100 different initial conditions and reported the most likely parameters. We verified that these parameters were not close to the optimization boundaries and not sensitive to the initial perturbation.

Parametric bootstrap was carried out by simulating (using *MaCS*[65], a coalescent simulator) artificial genomes under the demographic model of the most likely parameter set. For each of 100 data sets, the allele frequency spectrum was computed and folded, and $\partial a \partial i$ was used to infer the demographic parameters, exactly as for the real data. The biased-corrected 95% confidence intervals were computed assuming a normal distribution of the inferred parameters (Supplementary Note 6). Note that the confidence intervals account only for sampling noise but not for systematic errors such as sequencing errors or model and mutation rate misspecification.

For the single-population case (Supplementary Note 6, Supplementary Fig. 7 and Supplementary Table 5), we found that a model of a bottleneck followed by exponential growth explains well the spectra of both populations (Supplementary Fig. 8). Our main two-population model is shown in Fig. 4. The parameters of the recent AJ bottleneck were fixed to the values inferred from the IBD analysis (Supplementary Table 6). We verified that the log-likelihood of the optimal model decreased sharply near the values of two key parameters: the fraction of European admixture into AJ and the time of the European–Middle Eastern divergence. Admixture into AJ was shown to be necessary for a reasonable fit (Supplementary Note 6). Most parameters were robust to model specification, specifically, the time of the out-of-Africa bottleneck, the fraction of European admixture into AJ, and to some extent, the European–Middle Eastern divergence time. The time of the European admixture, however, differed considerably between models (Supplementary Note 6). The most promising model refinement included an additional wave of migration from the ancestral Middle Eastern population into Europeans at about $\approx$17 Kyr; experiments with further refinements did not converge to a consistent parameter set (Supplementary Note 6).

The deleterious mutation load. We annotated coding variants in the merged and size-matched AJ–FL data set ($n=26\times2$) using the SeattleSeq Variant Annotation server. We measured the (non-reference) variant load either as unique or total counts, and either for all or low frequency only variants (Supplementary Note 7). We further broke the counts by whether the variants were non-coding, coding synonymous or coding non-synonymous, and by *PolyPhen*'s[66] predicted effect (damaging or benign). To account for the genome wide larger number of variants in AJ, we normalized all counts by the ratio between the number of neutral AJ and FL variants. Significance of AJ–FL differences in any category was evaluated by assuming that all counts were binomial (Supplementary Table 8; Supplementary Note 7). To compare the number of non-synonymous variants per individual (Supplementary Fig. 15), we normalized each count by the number of intergenic variants. The (genome wide) average GERP score over all non-reference variants in each individual[67] was slightly higher (more conserved) in AJ than in FL (Supplementary Note 7).

We also attempted to determine whether there was any disease category with particularly high mutational burden in AJ. We computed the total number (over all individuals in each population) of non-synonymous (non-reference) variants in all genes belonging to each disease category, using the annotation developed in ref. 47 and then by Omicia (assigning 2488 genes into 17 categories; Supplementary Table 9). We then ranked all genes according to the difference between the number of AJ and FL non-synonymous variants, and used GSEA[68] to determine whether any given category had an exceptional number of top ranked genes. Only the aging category reached $P<0.05$, but with false discovery rate >0.05 (Supplementary Note 7).

A catalogue of variants in known disease genes. Our list of AJ disease genes is based on a table from ref. 2. We determined the hg19 coordinates of all disease mutations in that table manually using a number of online resources (Supplementary Note 7). The final list of 73 mutations in 48 genes is reported in Supplementary Data 4, along with some properties of each mutation. We then extracted all variants (including non-SNVs) in these genes from our unfiltered AJ genotypes ($n=128$). We detected carriers of 35 known disease mutations in 29 genes and annotated 953 newly discovered variants (using *ANNOVAR*[60]; also reported in Supplementary Data 4, along with summary statistics per gene; Supplementary Note 7).

References

1. Goodman, R. M. *Genetic Disorders among the Jewish People* (The Johns Hopkins University Press, 1979).
2. Ostrer, H. & Skorecki, K. The population genetics of the Jewish people. *Hum. Genet.* **132,** 119–127 (2013).
3. Ozelius, L. J. *et al.* LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **354,** 424–425 (2006).
4. Struewing, J. P. *et al.* The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N. Engl. J. Med.* **336,** 1401–1408 (1997).
5. 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).
6. Bray, S. M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc. Natl Acad. Sci. USA* **107,** 16222–16227 (2010).
7. 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).
8. Price, A. L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4,** e236 (2008).
9. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).
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. Guha, S. *et al.* Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome. Biol.* **13,** R2 (2012).
12. Gusev, A. *et al.* The architecture of long-range haplotypes shared within and across populations. *Mol. Biol. Evol.* **29,** 473–486 (2012).
13. Olshen, A. B. *et al.* Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. *BMC Genet.* **9,** 14 (2008).
14. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91,** 809–822 (2012).

15. Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J. & Stefansson, K. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37,** 90–95 (2005).
16. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* **497,** 517–520 (2013).
17. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
18. Genomes Project C *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
19. 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).
20. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19,** 318–326 (2009).
21. Carmi, S. *et al.* The variance of identity-by-descent sharing in the wright-fisher model. *Genetics* **193,** 911–928 (2013).
22. Gusev, A. *et al.* Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics* **190,** 679–689 (2012).
23. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84,** 235–250 (2009).
24. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
25. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336,** 740–743 (2012).
26. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5,** e1000695 (2009).
27. Pala, M. *et al.* Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. *Am. J. Hum. Genet.* **90,** 915–924 (2012).
28. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 745–753 (2012).
29. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet* **29,** 575–584 (2013).
30. Gravel, S. *et al.* Reconstructing native american migrations from whole-genome and whole-exome data. *PLoS Genet.* **9,** e1004023 (2013).
31. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
32. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15,** 11–19.24 (2014).
33. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108,** 11983–11988 (2011).
34. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* **39,** 1251–1255 (2007).
35. Higham, T. *et al.* The earliest evidence for anatomically modern humans in northwestern Europe. *Nature* **479,** 521–524 (2011).
36. Haber, M. *et al.* Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet.* **9,** e1003316 (2013).
37. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23,** 388–395 (2013).
38. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (2012).
39. Haak, W. *et al.* Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biol.* **8,** e1000536 (2010).
40. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342,** 257–261 (2013).
41. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans (2013).
42. Sikora, M. *et al.* Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet.* **10,** e1004353 (2014).
43. Casals, F. *et al.* Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet.* **9,** e1003815 (2013).
44. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451,** 994–997 (2008).
45. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
46. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46,** 220–224 (2014).
47. Moore, B. *et al.* Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. *Genet. Med.* **13,** 210–217 (2011).
48. Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat. Commun.* **4,** 2739 (2013).
49. Kenny, E. E. *et al.* A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet.* **8,** e1002559 (2012).
50. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4,** 2872 (2013).
51. Kurki, M. I. *et al.* High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms. *PLoS Genet.* **10,** e1004134 (2014).
52. Huffman, D. M. *et al.* Distinguishing between longevity and buffered-deleterious genotypes for exceptional human longevity: the case of the MTP gene. *J. Gerontol. A. Biol. Sci. Med. Sci.* **67,** 1153–1160 (2012).
53. Marder, K. *et al.* Familial aggregation of early- and late-onset Parkinson's disease. *Ann. Neurol.* **54,** 507–513 (2003).
54. Liu, X. *et al.* Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population. *BMC Med. Genet.* **12,** 104 (2011).
55. 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).
56. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
57. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
58. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods.* **10,** 5–6 (2013).
59. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
60. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
61. Wakeley, J. *Coalescent Theory: An Introduction* (Roberts & Company Publishers, 2009).
62. Zivkovic, D. & Stephan, W. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor. Popul. Biol.* **79,** 184–191 (2011).
63. International HapMap C *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
64. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
65. Chen, G. K., Marjoram, P. & Wall, J. D. Fast and flexible simulation of DNA sequence data. *Genome Res.* **19,** 136–142 (2009).
66. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
67. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).
68. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).

Acknowledgements

We thank Shlomo Hershkop for technical assistance and Barry Moore and Omicia Inc. for providing a disease gene catalogue. We thank Adam Auton and Alon Keinan for commenting on the manuscript. We acknowledge financial support from the Human Frontier Science Program (S.C.); NIH research grants AG042188 (G.A.), DK62429, DK062422, DK092235 (J.H.C.), NS050487, NS060113 (L.N.C.), AG021654, AG027734 (N.B.), MH089964, MH095458, MH084098 (T.L.), GM007205, DK098927 (K.Y.H.), and CA121852 (computational infrastructure, I.Pe'er); NSF research grants 08929882 and 0845677 (I.Pe'er); Rachel and Lewis Rudin Foundation (H.O.); North Shore–LIJ Health System Foundation (T.L.); Brain & Behaviour Foundation (T.L.); US-Israel Binational Science Foundation (T.L., A.D.); New York Crohn's Foundation (I.Peter); Edwin and Caroline Levy and Joseph and Carol Reich (S.B.); the Parkinson's Disease Foundation (L.N.C.); the Sharon Levine Corzine Cancer Research Fund (K.O.); and the Andrew Sabin Family Research Fund (K.O.).

Author contributions

S.C. was the primary analysis and manuscript-writing person. K.Y.H., E.K., X.L., J.X., F.G., S.G., K.U., D.B.-A., S.M., B.M.B., T.T. and J.V. conducted analysis and provided input for the manuscript. M.C., G.F., D.L., S.P., C.V.B, P.V.D., and H.V.M. contributed the Flemish genomes. N.B. contributed Ashkenazi DNA samples. A.D., K.O., S.B., I.Peter, J.H.C., H.O., L.J.O., G.A., L.N.C., T.L., and I.Pe'er initiated and funded the study. I.Peter, J.H.C., H.O., G.A., L.N.C., and T.L. supervised analysis and provided comments on the manuscript. G.A. and L.N.C. conducted lab work. T.L. led the funding of the study. I.Pe'er led the analysis and the writing of the manuscript.

Additional information

Accession codes: Whole-genome sequence data have been deposited at the European Genome-phenome Archive, which is hosted by the EBI, under accession code EGAS00001000664.

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

Competing financial interests: The authors declare no competing financial interests.

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

How to cite this article: Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* 5:4835 doi: 10.1038/ncomms5835 (2014).