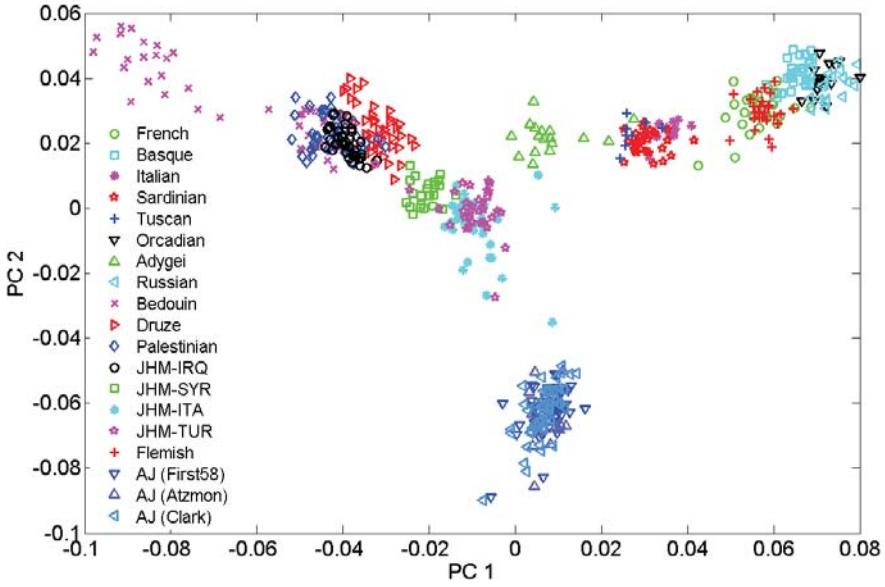
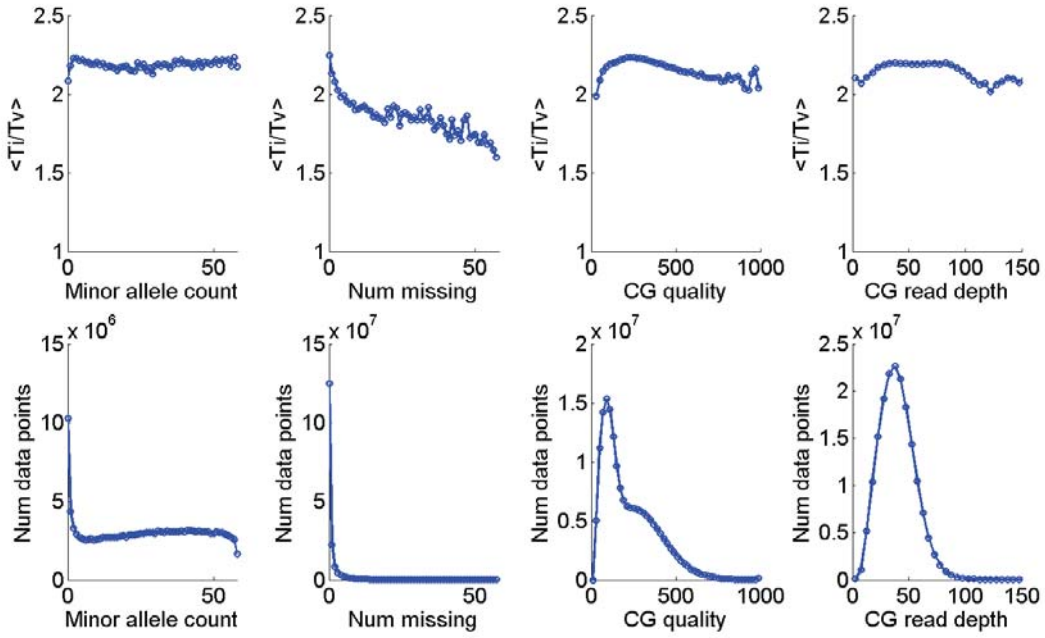
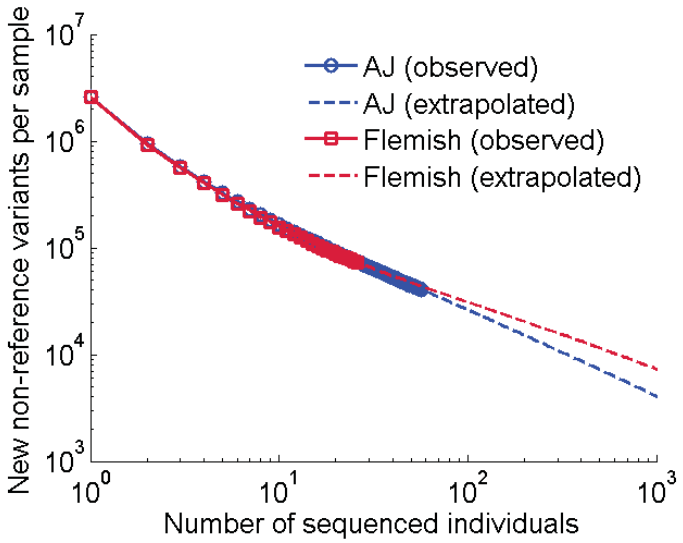
Supplementary Figures



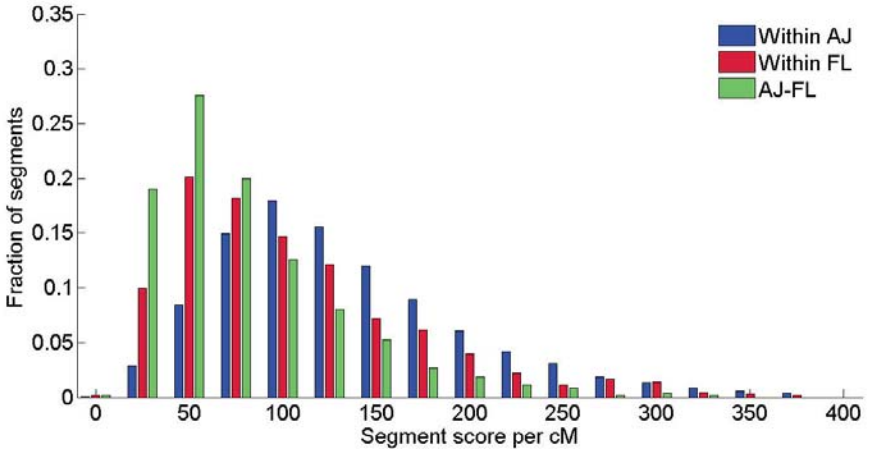
Supplementary Figure 1. *Principal Component Analysis (PCA) of common variants in AJ, non-AJ Jewish, European, and Middle-Eastern populations.* The samples designated AJ and Flemish are the ones reported in this study; non-AJ Jewish populations are from the Jewish HapMap project (JHM) [1]; other European and Middle-Eastern populations are from the Human Genome Diversity Project (HGDP) [2]). Genomes from different AJ sequencing batches are shown in different symbols and different shades of blue.



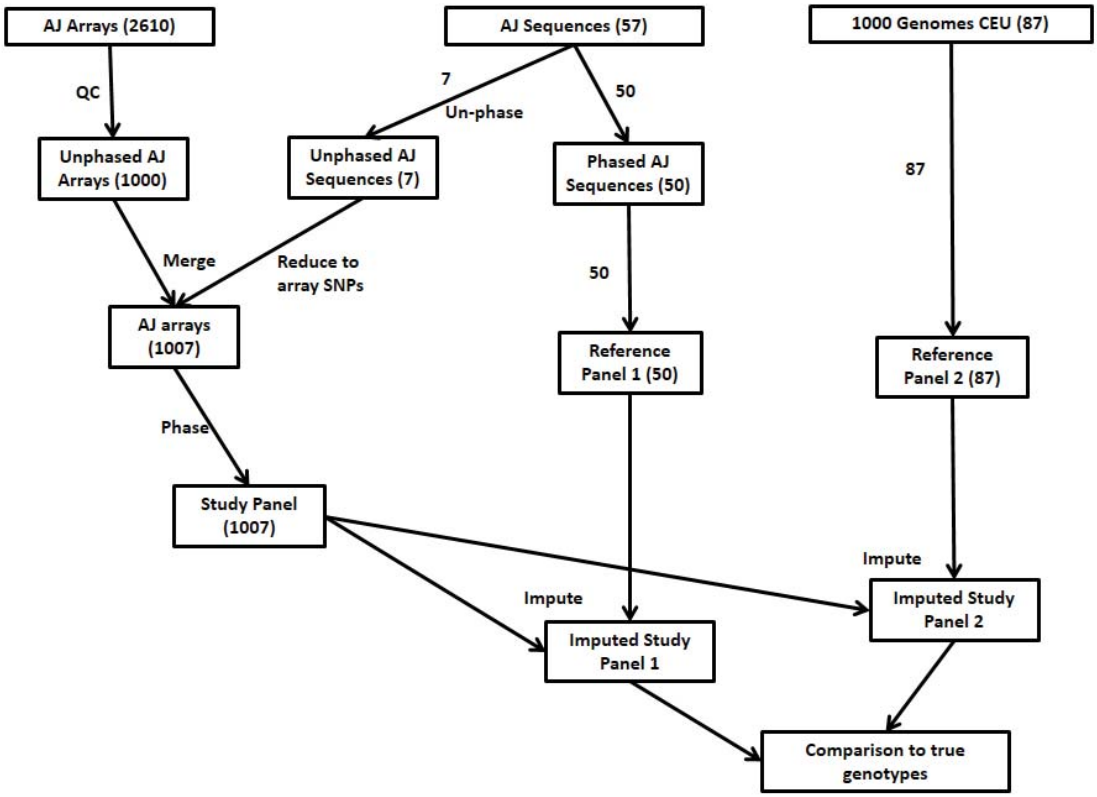
Supplementary Figure 2. *Transition/transversion (Ti/Tv) ratio.* The top row shows the Ti/Tv ratio (averaged over the 58 AJ individuals sequenced in our first batch) vs. properties of the variants. In the bottom row, we plot, for each property, the total number of variants that were used to compute each Ti/Tv data point in the corresponding panel in the top row. CG: Complete Genomics.



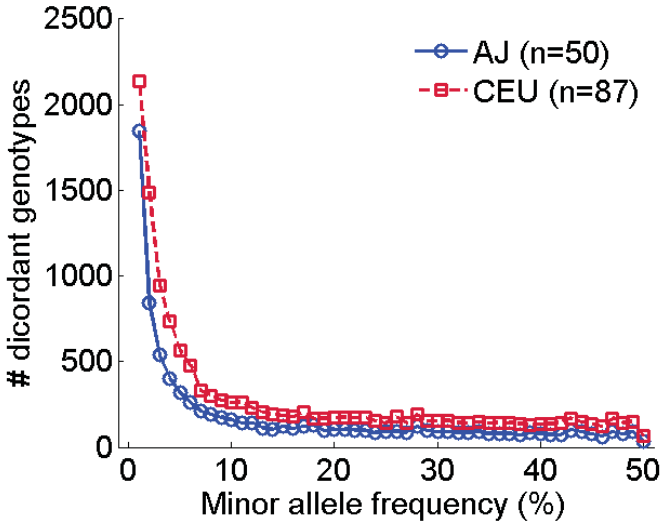
Supplementary Figure 3. *Rate of non-reference variant discovery.* Symbols joined by solid lines represent the empirical average number of new non-reference variants discovered vs. the number of already sequenced individuals. Dashed lines are the projections to larger sample sizes, obtained using the estimator developed by Gravel et al. (2011) [3].



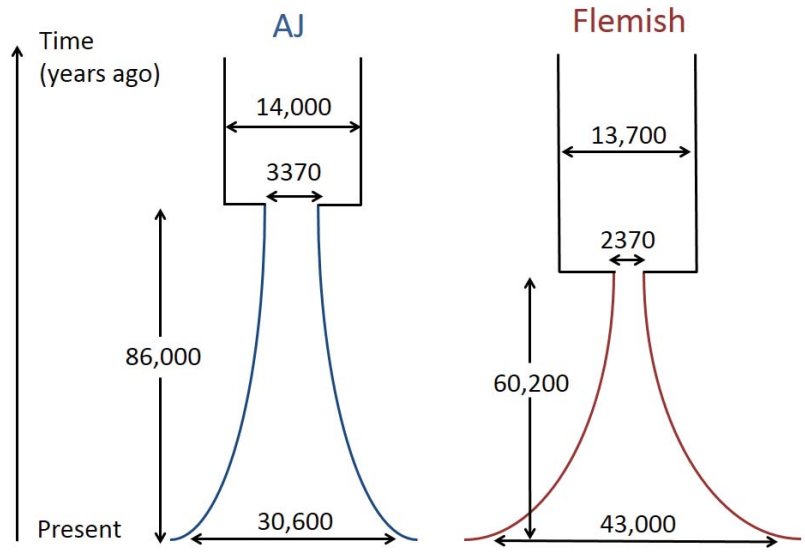
Supplementary Figure 4. The *distribution of the IBD segment scores.* We plot the distribution of the segment scores ($\boldsymbol{D}$; Supplementary Eq. (11)) per cM for segments shared within AJ (blue), within Flemish (red), and between AJ and Flemish (green). Within-AJ segment scores are significantly higher than within-Flemish or AJ-Flemish scores.



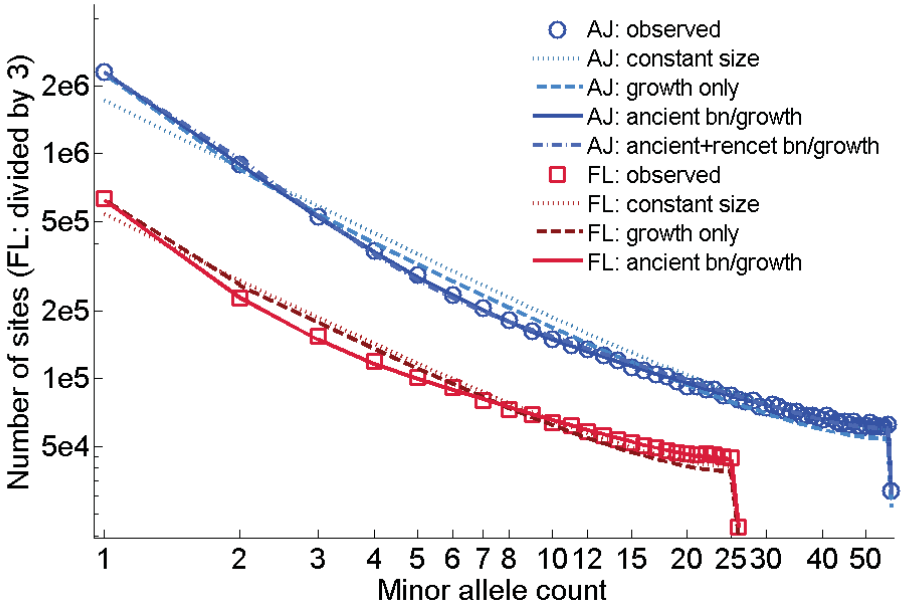
Supplementary Figure 5. *Our imputation study design.*



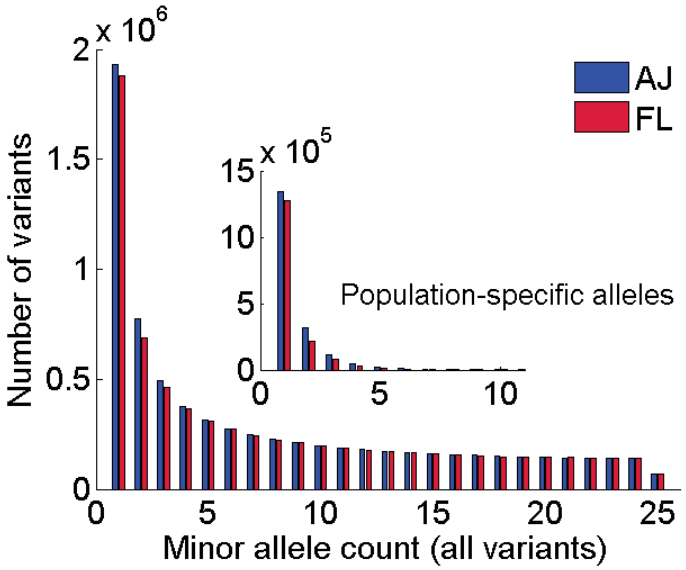
Supplementary Figure 6. *Accuracy of the imputation results vs. the minor allele frequency.* For each reference panel and for each minor allele frequency (see details in Supplementary Note 5), the average number of discordant genotypes (over the seven AJ study sequences) is shown.



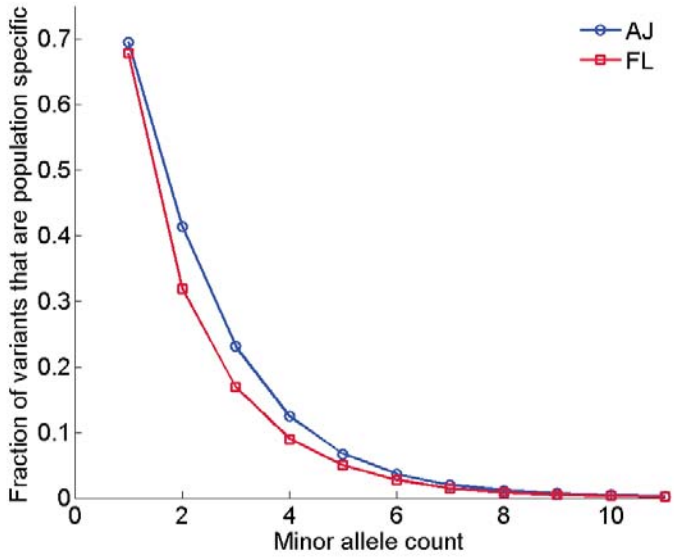
Supplementary Figure 7. *The single-population demographic models along with the inferred parameters.* For both AJ and Flemish, we assumed a history of an ancient bottleneck followed by slow exponential growth. The demographic parameters were inferred, using *∂a∂i* [4], based on the allele frequency spectrum of each population and then parametric bootstrap (Supplementary Note 6, Supplementary Table 5). The reported parameters correspond to the bias-corrected bootstrap means. Population sizes (horizontal arrows) are in diploid individuals, and times (vertical arrows) are in years, assuming 25 years per generation.



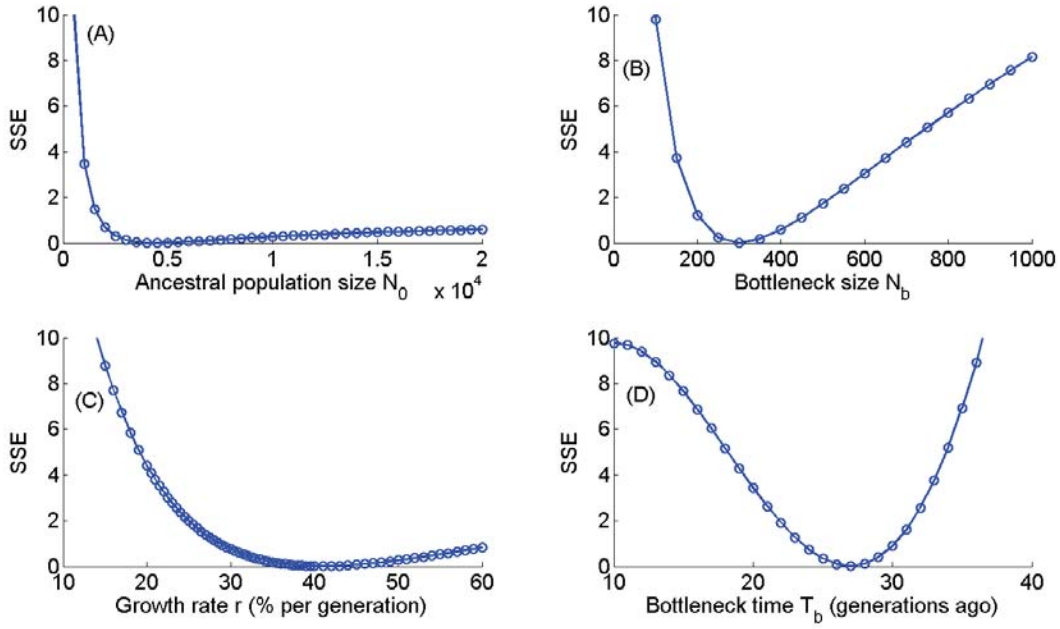
Supplementary Figure 8. *The observed and fitted single-population allele frequency spectrum.* For each population (AJ, blue shades; Flemish, red shades), the minor allele frequency spectrum is plotted (symbols) along with the maximum likelihood spectrum (lines) for a number of demographic models. The models are defined in Supplementary Note 6 and the parameter values are reported in Supplementary Table 5. For visibility, the number of sites in all Flemish spectra was divided by 3. bn/growth: a bottleneck followed by exponential growth.



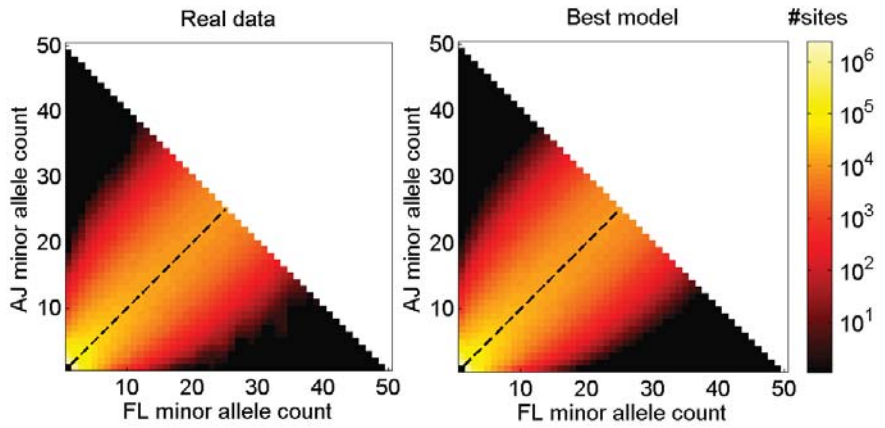
Supplementary Figure 9. *The (non-normalized) single-population frequency spectra of AJ and Flemish.* The total number of variants is shown vs. the minor allele count, after each population has been reduced to 50 haploid genomes. Inset: The total number of population-specific variants vs. the minor allele count in the population where each variant exists.



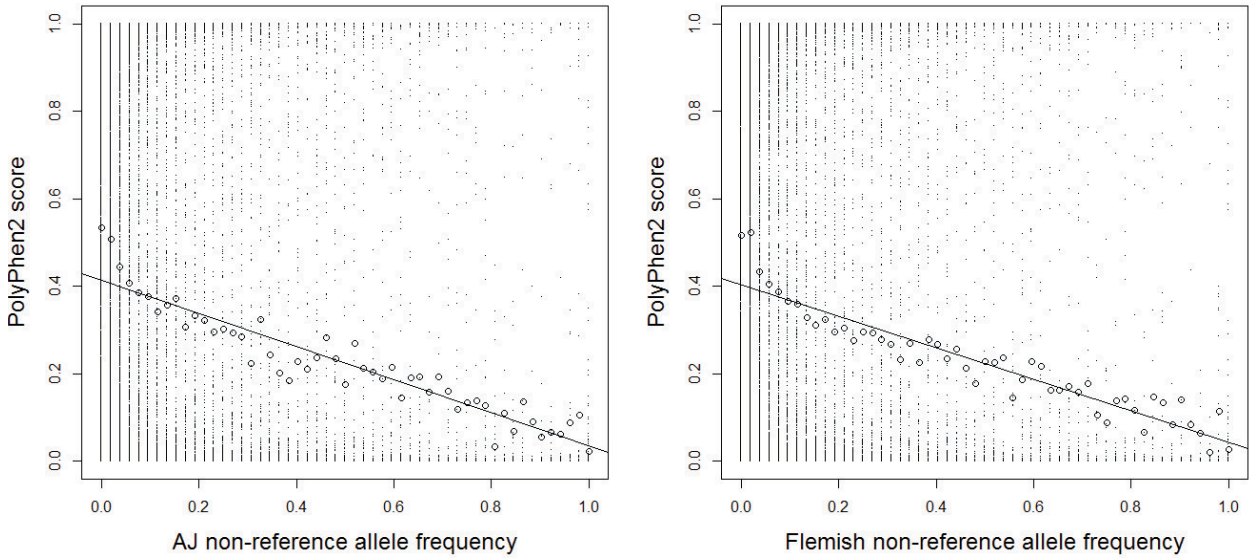
Supplementary Figure 10. *The fraction of variants that are population specific in AJ and Flemish.* The fraction is plotted vs. the minor allele count (in the population where each variant exists), after each population has been reduced to 50 haploid genomes.



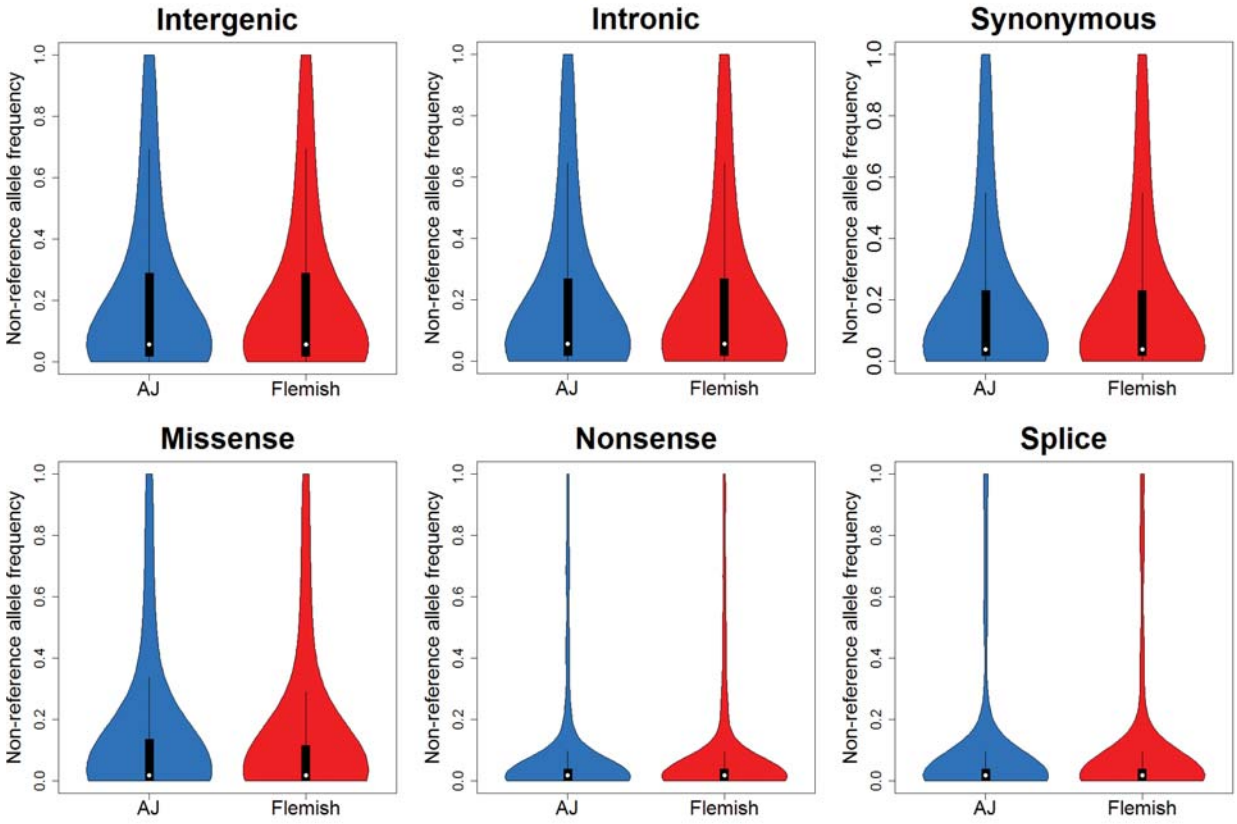
Supplementary Figure 11. *The sum of squared errors (SSE) around the recent history parameters inferred using IBD segment lengths.* The demographic model we inferred in Supplementary Note 4 has four parameters. In each panel, we fixed three of the parameters to their optimal values and varied the fourth. We then computed the SSE between $\boldsymbol{p}_{\mathrm{real}}(\boldsymbol{\ell})$ (plotted in Figure 2 of the main text) and $\boldsymbol{p}_{\mathrm{model}}(\boldsymbol{\ell})$ according to Supplementary Eq. (13). (A) The ancestral population size, $\boldsymbol{N_0}$. (B) The bottleneck size, $\boldsymbol{N_b}$. (C) The growth rate, $\boldsymbol{r}$. (D) The bottleneck time, $\boldsymbol{T_b}$. Note that the same y-axis scaling was used for all panels.



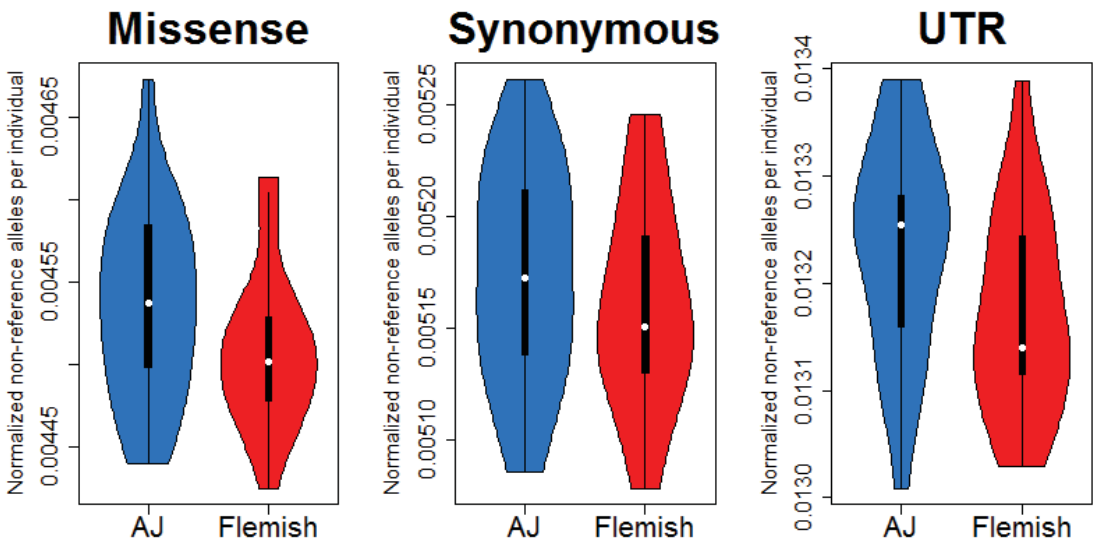
Supplementary Figure 12. *The real joint AJ-Flemish allele frequency spectrum and the best fit (maximum likelihood) model spectrum.* In each panel, the total number of variants is shown vs. the AJ and Flemish minor allele counts (according to the color bar), after each population has been down-sampled to 50 haploid genomes. The dashed line corresponds to equal frequencies in AJ and Flemish. Left: the real joint spectrum, reproducing Figure 3B of the main text. Right: the best fitting model spectrum, corresponding to the demographic model of the top part of Figure 4 of the main text (Supplementary Note 6, section 6.2.3.2) with the maximum likelihood parameters reported in Supplementary Table 7.



Supplementary Figure 13. *PolyPhen2 score vs. non-reference allele frequency in AJ (left) and Flemish (right).* Dots represent individual variants, circles represent average PolyPhen2 score within each allele frequency bin, and lines represent linear regression models fitted to the average scores (not significantly different between AJ and Flemish; Supplementary Note 7).



Supplementary Figure 14. *Violin plots of non-reference allele frequency spectra in AJ and Flemish, by variant functional class.*



Supplementary Figure 15. *Violin plots of the non-reference allele counts per individual in AJ and Flemish, by functional class*. The number of variants in each category was normalized by the number of intergenic variants, to account for difference in the total number of variants between the populations due to their different demographic histories.

Supplementary Tables

Trait (Atzmon's lab; Einstein)	Mean ± Standard Deviation	Trait (Clark's lab; Columbia)	Mean ± Standard Deviation
All (n)	74	**All (n)**	54
Female (n)	45	**Female (n)**	33
Male (n)	29	**Male (n)**	21
Age (years)	68.8±7.7 (range 49-85)	**Age (years)**	68.7±10.4 (range 39-88)
Cholesterol (mg/dL)	200±42.1	**Intellectual impairment (n)**	1
Triglycerides (mg/dL)	132±71.4	**Thought disorder (n)**	1
HDL (mg/dL)	65.4±17.2	**Depression (n)**	2
LDL (mg/dL)	108±34.4	**Family history of PD in first degree relatives (conservative) (n)**	2
Glucose (mg/dL)	81.9±14.6	**Family history of AD in first degree relatives (conservative) (n)**	3
Waist circumference (inch)	35.1±6.9	**Total mMMS Score**	56.2±1.4
Body Mass Index (kg/m^2)	26.4±5.1	**Total UPDRS part II score**	0.25±0.85
Systolic Blood pressure (mm Hg)	139±20.6	**Total UPDRS part III score**	1.81±3.20
Diastolic Blood pressure (mm Hg)	79.6±11.2		

Supplementary Table 1. *Demographic and medical characteristics of the AJ samples.* Means and standard deviations are shown. For a description of the cohorts (Einstein and Columbia), see Supplementary Note 1. Except the gender and the mMMS score, all traits in the Columbia cohort were computed over 53 samples. PD: Parkinson's disease. AD: Alzheimer's disease. The mMMS score was calculated from a modification of the modified Mini-Mental State Examination, with a maximum score of 57 (computed for 15 samples). The Unified Parkinson's Disease Rating Scale (UPDRS) parts II and III contain 44 questions, each measured on a 5-point scale (0-4).

	Fully called genome fraction	Fully called exome fraction	SNPs total count	SNPs novel fraction	SNPs het/hom ratio
Average	96.66%	98.10%	$3.412 \cdot 10^{6}$	3.84%	1.647
Coefficient of variation (CV)	$3 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$2.1 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$
	SNPs Ti/Tv ratio	**Insertions**	**Deletions**	**Multi-nucleotide substitutions**	**Synonymous SNPs**
Average	2.142	$220 \cdot 10^{3}$	$235 \cdot 10^{3}$	$83 \cdot 10^{3}$	10,536
CV	$2 \cdot 10^{-3}$	$3.3 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$9 \cdot 10^{-3}$
	Non-synonymous SNPs	**Non-sense SNPs**	**CNV segments**	**Structural variants**	**Mobile element insertions**
Average	9706	72	302	1480	4090
CV	$1.0 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	0.503	$4.8 \cdot 10^{-2}$	0.161

Supplementary Table 2. *Selected quality control and variant count statistics for 128 AJ genome sequences.* Statistics were reported for each individual by Complete Genomics. Means and coefficients of variation (standard deviation/mean) are shown.

	AJ	Flemish	P-value
Non-reference variants ($\times 10^3$)	2602 ± 7.8	2563 ± 11.2	$3.9 \cdot 10^{-13}$
Heterozygous variants ($\times 10^3$)	1637 ± 10.2	1599 ± 12.7	$1.3 \cdot 10^{-6}$
Homozygous variants ($\times 10^3$)	966 ± 5.7	964 ± 5.2	0.30
Het/hom ratio	1.695 ± 0.019	1.658 ± 0.019	$8.2 \cdot 10^{-10}$
Fraction novel in dbSNP132 (%)	3.14 ± 0.01	2.52 ± 0.02	$3.9 \cdot 10^{-13}$
Fraction novel in dbSNP135 (%)	1.42 ± 0.004	0.96 ± 0.02	$3.9 \cdot 10^{-13}$

Supplementary Table 3. *A comparison of variant statistics between AJ and Flemish.* The quantities reported are the mean and standard deviation over 57 AJ and 26 Flemish individuals, respectively, after cleaning and merging. The het/hom ratio and the dbSNP novelty were computed with respect to the non-reference variants in each individual. The P-values were computed using the rank-sum test.

Ref panel	Discordant genotypes	False positives	False negatives	
AJ (n=50)	12,181	4615	7566	
CEU (n=87)	16,901	4769	12,132	
Ref panel	**Fraction of non-ref variants wrongly imputed**	**Fraction of non-ref variants with minor allele freq ≤1% wrongly imputed**	r^2	***IMPUTE2*'s self-estimated discordance**
AJ (n=50)	4.08%	13.01%	98.24%	2.79%
CEU (n=87)	6.53%	34.67%	97.37%	3.62%

Supplementary Table 4. *Summary of the imputation results.* The numbers of discordant genotypes, false positives, and false negatives, as well as the fractions of wrongly imputed non-reference genotypes are the averages over the seven AJ study sequences and were computed using the most likely imputed genotypes. r^2 is the aggregate (squared-) correlation between the true genotypes and the imputed dosages over all study individuals and sites. False negatives occur when at least one non-reference allele was missed; false positives occur when *IMPUTE2* wrongly suggests at least one non-reference allele. Sites that were monomorphic non-reference in the AJ panel were excluded, and the minor allele frequency was computed in the AJ reference panel. *IMPUTE2*'s estimate of the discordance is the average over the 1000 array genotypes.

Wright-Fisher (section 6.2.2.1)	N_0			
AJ	13,609			
Flemish	12,664			
Growth-only (section 6.2.2.2)	N_0	T_g	N_f	
AJ	12,298	11,847	57,988	
Flemish	12,006	5636	38,239,322	
Bottleneck/Growth (section 6.2.2.3)	N_0	N_b	T_b	N_f
AJ	13,987 13,968±42 [13,885, 14,050]	3373 3502±30 [3443 , 3561]	86,083 86,270±955 [84,399 , 88,142]	30,604 30,731±219 [30,301 , 31,162]
Flemish	13,658 13,643±40 [13,564 , 13,722]	2370 2451±31 [2389 , 2512]	60,219 60049±1146 [57,803 , 62,295]	43,020 45,074±926 [43,258 , 46,890]
Bottleneck/Growth + known recent B/G (section 6.2.2.4)	N_0	$N_{b,a}$	$T_{b,a}$	$N_{f,a}$
AJ	13,660 13,651±31 [13,591 , 13,711]	2375 2336±19 [2299 , 2374]	58,556 57,816±594 [56,651 , 58,981]	79,347 83,547±1266 [81,065 , 86,028]

Supplementary Table 5. *The inferred parameters for our single-population demographic models.* See Supplementary Note 6, section 6.2 for definitions. For the Wright-Fisher and the growth-only models, only the maximum likelihood parameter values are reported. For the two bottleneck/growth models, the parametric bootstrap results (Supplementary Note 6) are also reported: the bias-corrected means and the standard deviations (second line of each cell) and the 95% confidence intervals (third line). All population sizes are reported in number of diploid individuals; times are reported in years (assuming 25 years per generation).

Parameter	Mean	Standard deviation	95% confidence interval
Ancestral size N_0	4755	562	[3654,5856]
Bottleneck size N_b	334	43	[249,419]
Growth rate r (Final population size)	34% ($N_f = 1.450 \cdot 10^6$)	10%	[16%,53%]
Bottleneck time T_b	28	2	[25,32]

Supplementary Table 6. *The demographic parameters inferred using IBD sharing.* The demographic model (a bottleneck followed by exponential growth) is schematically plotted in Supplementary Note 4, section 4.3.1. The parameters were inferred by fitting the observed decay of IBD sharing at increasing genetic distances (section 4.3.1), followed by jackknife resampling (section 4.3.4). The means, standard deviations, and confidence intervals were computed over 100 resampling iterations. The final (current) population size was computed using the means of the other parameters. The population sizes are given in number of diploid individuals, the time in generations, and the growth rate in percent per generation.

Parameter	Maximum likelihood	Bias-corrected mean±SD	95% confidence interval
N_0	13,945	13,940±34	[13,872 , 14,007]
$N_{b,OOA}$	3874	3843±115	[3618 , 4069]
$T_{b,OOA}$	89,785	89,342±2459	[84,523 , 94,161]
$N_{f,AJ}$	23,784	24,184± 1198	[21,837 , 26,531]
$N_{b,EU}$	3692	3695± 110	[3479 , 3911]
$T_{b,EU}$	21,016	21,264± 430	[20,421 , 22,108]
$N_{f,EU}$	170,465	173,771± 13,715	[146,889 , 200,653]
T_a	681	673± 24	[626 , 721]
f_a	49%	48%±1%	[46% , 50%]

Supplementary Table 7. *The inferred parameters for the joint AJ-Flemish demographic model.* The model is defined in Supplementary Note 6, section 6.2.3.2. The maximum likelihood parameters were computed using *∂a∂i* [4] (section 6.3.1); confidence intervals were obtained using parametric bootstrap (section 6.3.2): we report the bias-corrected means, the standard deviations (SD), and the 95% confidence intervals. Population sizes are given in number of diploid individuals and times in years, assuming 25 years per generation.

AJ variant count	Non-syn.	Baseline: non-coding+syn.		Baseline: synonymous		Coding damaging Baseline: benign		
	Observed	Expected	P-value	Expected	P-value	Observed	Expected	P-value
Unique	27,219	26,595	$6\cdot10^{-5}$	26,964	0.06	12,190	12,099	0.20
All	251,788	250,527	$6\cdot10^{-3}$	251,109	0.09	61,393	60,454	$7\cdot10^{-5}$
Unique freq.<10%	17,179	16,259	$3\cdot10^{-13}$	16,484	$3\cdot10^{-8}$	9357	9124	$7\cdot10^{-3}$
All freq.<10%	31,179	28,428	$4\cdot10^{-60}$	28,494	$3\cdot10^{-57}$	16,065	15,566	$1\cdot10^{-35}$

Supplementary Table 8. *A comparison of the functional mutation burden between AJ and Flemish*. We considered a reduced set of genotypes for 26 individuals in each of the AJ and Flemish populations. The observed number of either non-synonymous (non-syn.) or damaging (annotated using PolyPhen2) [5] non-reference variants is compared to the expected number based on the Flemish genomes, with the non-functional variation as a baseline (either non-coding+synonymous (syn.) or synonymous only for non-syn. variation, and benign coding variants for the damaging variation). Counts are also reported when considering only variants having (non-reference) allele frequency (in the combined AJ-Flemish dataset) of <10%. The P-value is approximate, assuming standard normal distribution of the scaled difference $(\text{observed} - \text{expected})/\sqrt{\text{expected}}$.

Disease category	#genes	#AJ non-syn. variants	#FL non-syn. variants	AJ/FL ratio
Aging	106	1177	1105	1.07
Infectious	70	1100	1065	1.03
Neonatal	956	13387	13123	1.02
Gastrointestinal	254	5577	5479	1.02
Dental	86	1564	1543	1.01
Immunological	474	7325	7241	1.01
Hemic	202	2791	2759	1.01
Cardiovascular	502	7714	7626	1.01
Endocrinological	750	10466	10374	1.01
Oncological	471	7965	7898	1.01
Women's	39	409	408	1.00
Drug	82	1661	1667	1.00
Neurological	980	12139	12198	1.00
Nutrition	29	256	258	0.99
Respiratory	187	3517	3570	0.99
Kidney	285	3740	3877	0.96
Psychiatric	21	271	291	0.93

Supplementary Table 9. *The non-synonymous mutational burden for different disease categories.* Gene annotation was provided by Omicia Inc. The number of non-synonymous (non-syn.), non-reference variants was computed for each category and for each population (AJ and Flemish) in the reduced set of genotypes for 26 individuals in each population. Categories were ordered according to their ratio of AJ/FL non-syn. burden.

1. Supplementary Note 1: Sample selection and sequencing

Our 128 DNA samples were collected from the following sources.

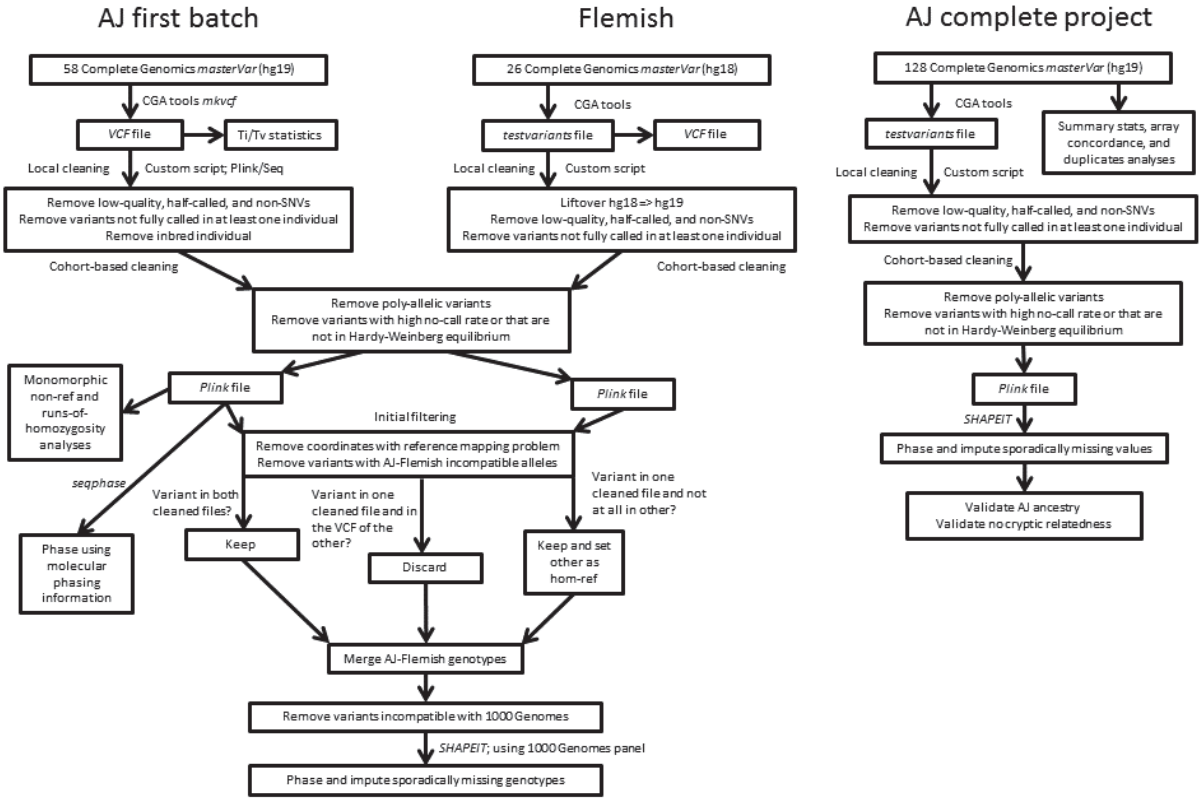
Atzmon's lab, Albert Einstein School of Medicine (n=74). The sequenced individuals were controls in a longevity study. Subjects were recruited by word of mouth and through advertisement in Jewish aging centers and homes. All subjects were disease-free and were verified by PCA to have four grandparents of Ashkenazi origin (see also section 2.9). Additionally, both parents of all subjects died before the age of 85 and were without longevity history in the family. Cryptic relatedness has been excluded based on identity-by-descent analysis using Affy 6.0 data (see also section 2.10). Summary statistics for medically-relevant phenotypes are given in Supplementary Table 1. Written informed consent was obtained in accordance with the policy of the Committee on Clinical Investigation of the Albert Einstein College of Medicine. A single nurse practitioner visited all participants to conduct a physical examination and obtain a medical history report, including review of the questionnaire [6]. The majority of the samples from this source (n=58) were sequenced in summer 2012. Those 58 genomes (less one; see section 2.7.1) were used for most of the population genetic comparisons against European genomes reported in the paper. The remaining 16 samples were sequenced separately in winter 2012-2013.

Clark's lab, Columbia University Medical Center (n=54). The sequenced individuals were controls in two studies: (*i*) The Genetic Epidemiology of Parkinson's Disease study at Columbia University and (*ii*) The New York Ashkenazi Jewish study at Columbia University. Information on Jewish origin in each of the grandparents was obtained during an interview. Ashkenazi ancestry was not specifically inquired; however, ≈90% of Jews in the United States are Ashkenazi and this was verified by PCA (section 2.9). Ascertainment and a description of the study participants is provided in detail in Marder et al. (2003) [7] and Liu et al. (2011) [8]. All control probands were evaluated with a medical history, modified Mini-Mental State Examination (mMMSE), Unified Parkinson's Disease Rating Scale (UPDRS), and when possible, a videotaped assessment that included items from the UPDRS rating scale. Summary statistics for medically-relevant phenotypes are given in Supplementary Table 1. The study was approved by the Institutional Review Board at Columbia University Medical Center. Each study participant signed a written informed consent approved by the University Human Ethics Committee. Genomes from this source were sequenced in winter 2012-2013.

In all samples, DNA was isolated from blood. Sequencing was carried out by Complete Genomics (CG) [9, 10]. The average raw sequencing depth was 56x (Supplementary Data 2). The first 58 genomes were called using CG pipeline 2.0.2.26. All other genomes were called using pipeline 2.0.4.14. Both pipelines mapped variants to reference genome version hg19 [11].

2. Supplementary Note 2: Quality control and processing pipeline

2.1. A diagram demonstrating our processing and quality-control pipeline



Supplementary Note Figure 1. *A diagram of the various steps in our processing and quality control pipeline.* All items are described in detail throughout section 2.

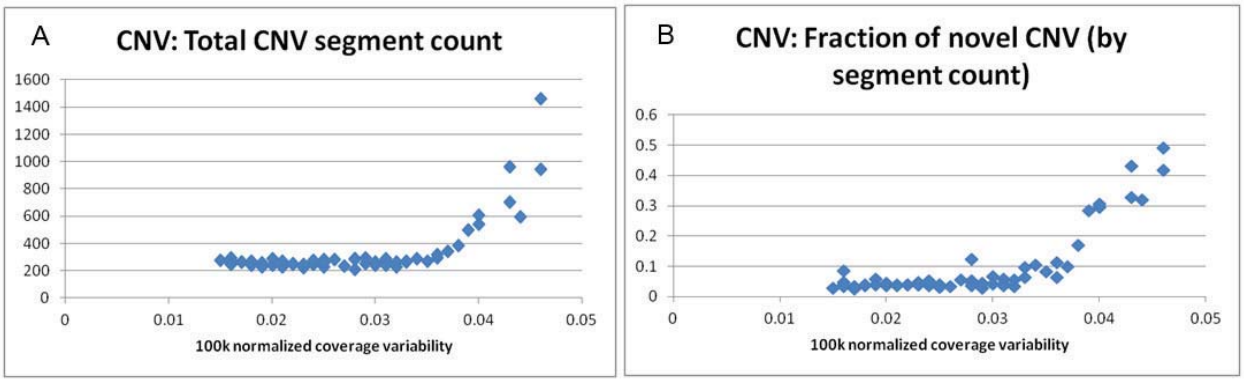
2.2. Quality control and variant count statistics

Several quality control and variant count statistics were reported (per genome) by Complete Genomics. Selected statistics (averages and coefficients of variation) are shown in Supplementary Table 2. Note the very low level of variation in single-nucleotide and short variants, as opposed to longer variants, particularly CNVs (suggesting lower quality of those variants; see also sections 2.3 and 2.4). The complete list of statistics for all individuals is given in Supplementary Data 2. Note that these statistics refer to the original data before filtering (section 2.7).

2.3. Copy Number Variants (CNVs)

As with other next-generation sequencing platforms, the Complete Genomics (CG) pipeline calls copy number variants (CNVs) based on localized variations in read depth/coverage. Coverage is assessed in sliding windows at 2kb intervals, and normalized according to the GC content of the window (http://media.completegenomics.com/documents/CNV+Methods.pdf). Quality metrics and summary statistics were assessed in the first batch of genomes delivered (n=58), and compared to publicly available European (CEU) genomes from CG (n=22), as described below.

In general, 200-400 CNVs were called in each genome. However, a subset (n=9) demonstrated a notable excess in called CNV segments. As depicted in Supplementary Note Figure 2A, these samples were also marked by outlier values in a quality metric provided by CG, which represents the overall variability across windows, normalized as a function of GC content. High values on this metric may indicate artifacts emerging during library construction, with the excess CNVs representing false positive calls. As demonstrated in Supplementary Note Figure 2B, the excess calls represent novel CNVs (not present in Complete Genomics public genomes or the Database of Genomic Variants (DGV)), which are not likely to be present in such volume in control individuals such as the participants in our study. Consequently, nine subjects with 100k normalized coverage variability ≥0.038, who demonstrated the nine highest CNV call rates, were excluded from the comparison to European genomes that we describe next. Notably, none of our subsequent genomes (n=70), drawn from other extraction batches, demonstrated coverage variability or CNV segment counts above these levels.



Supplementary Note Figure 2. *Properties of Copy Number Variants (CNVs) detected in our sequencing cohort.* The CNV count (A) and fraction novel (B), as well as the coverage variability, are as reported by Complete Genomics for our first sequencing batch (n=58). See also Supplementary Data 2.

We then compared the n=49 remaining genomes from our first batch to the n=22 European (CEU) genomes made publicly available by Complete Genomics. The overall number of called CNV segments (deletions and duplications only) did not significantly differ between AJ (mean=245.7±24.9) and CEU (mean=243.3±8.1) groups, nor was there any significant difference in average segment size (22527±30594 vs. 22848±32484). However, when compared to the Database of Genomic Variants (DGV), the rate of novel CNV calls in our AJ samples (3.86%) was approximately double that of CEU (1.8%).

2.4. Mobile Elements Insertions (MEIs)

2.4.1. MEI detection

Mobile Elements (ME) are repetitive genomic sequences comprising a large proportion of the human genome [12, 13]. The major ME families that are active in human and primate genomes are the long interspersed element-1 (LINE-1 or L1), Alu, and SVA (SINE-R/VNTR/Alu). MEI events have been reported in the literature as a cause of single-gene disease (reviewed in [14, 15]). Common MEIs may also contribute to genetic variation and gene expression in the human genome [15].

The Complete Genomics pipeline generated a genome wide map of MEIs in our 128 Ashkenazi Jewish control samples. Insertion sites were identified by searching for mate-paired reads that mapped uniquely to the reference with one arm and to repetitive sequences with the other arm. The location, type, orientation, and score of the inserted elements were reported, as well as the number of reads supporting either the insertion or the reference. MEIs that were previously observed in the 1000 Genomes Project [13] were indicated as such; we refer to those MEIs as *known* and to the rest as *novel*. Since the reported positions of the insertions were imprecise, we did not attempt to identify recurring MEIs across individuals, except for the known variants (since those were mapped by CG to specific 1000 Genomes MEIs).

2.4.2. MEI results

We detected 4090±662 MEIs per individual (mean±standard deviation), or 523,572 overall. However, we noticed a strong batch effect between the first batch (58 genomes), where the number of MEIs per individual was 3475±339, and the other genomes, where it was 4600±362 ($P = 2 \cdot 10^{-21}$, rank-sum). The batch effect was mostly due to the novel MEIs (the number of known MEIs was very similar between the batches).

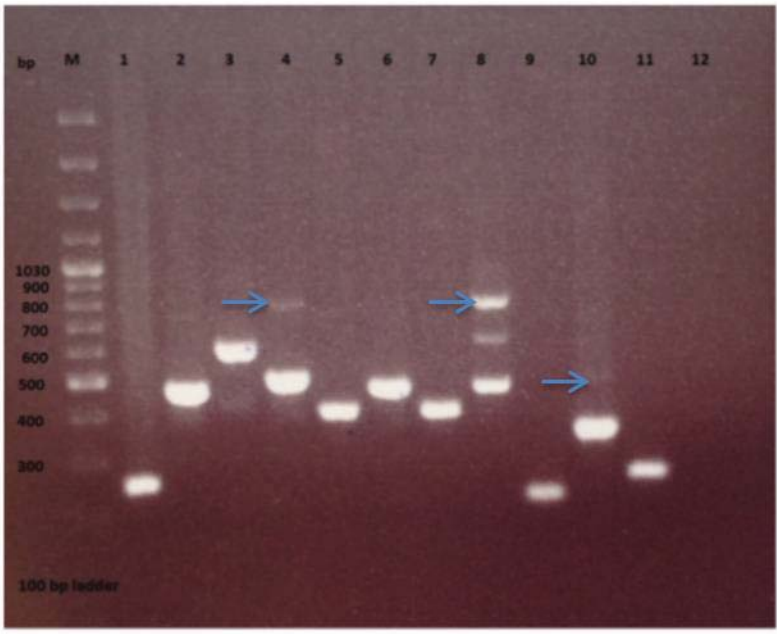
2.4.3. Experimental validation of novel MEIs

We attempted to validate a number of novel MEIs (n=11) that mapped to the introns of known Ashkenazi Jewish (AJ) disease genes (LOXHD1, ABCC8, and MAK; section 7.4). PCR primers (Supplementary Note Table 1) were designed to localize at least 50 bp upstream and 100 bp downstream of the insertion regions. PCR amplifications were performed in 25 µl reactions using Eppendorf Mastercycler. Each reaction contained 100 ng of template DNA in Roche FastStart Taq DNA polymerase system (Cat No 12032 953 001) with GC-RICH solution (200 nM of each oligonucleotide primer, 1x GC-RICH solution, 2mM MgCl2, 1× PCR buffer, 0.2 mM dNTPs, and 1U Taq DNA polymerase). PCR conditions were 95°C for 15 min, followed by 40 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 3 min, with a final cycle of 72°C for 10 min. PCR products were analyzed on a 1.7% agarose gel stained with Crystalgen Dye and a 100 bp ladder (Cat No 65-0321). Images were taken and saved using a BioRad ChemiDoc XRS imaging system (Hercules, CA). Out of 11 tested MEIs, only three were validated (Supplementary Note Figure 3.; Supplementary Note Table 1; false discovery rate of 8/11=73%), indicating low confidence in our detected novel MEIs. We therefore focus next on the known MEIs.

Lane	Sample ID	Gene	ME type	PCR region	Size of PCR product without insertion	Size of PCR product with insertion	Validated
1	14990	LOXHD1	AluYb9	chr18:44125811-44126018	207	247	N
	Primers:	TCTCGATCTCCTGACCTCGT			TAAGCCAGAGGCAGAGGACT		
2	14993	LOXHD1	AluYd2	chr18:44126100-44126586	486	630	N
	Primers:	GGGAGGAGTTTTAGGGATGC			AACTGAATTGGGATGATTGGA		
3	15048	LOXHD1	L1HS	chr18:44126007-	613	868	N

				44126808			
	Primers:	GGGAGGAGTTTTAGGGATGC			AAAGGGGGCATAGTCTCACA		
4	16091	LOXHD1	AluYa1	chr18:44126074-44126584	510	751	Y
	Primers:	GAGCCCCATTCTGACTCCTC			ACTGAATTGGGATGATTGGA		
5	16098	LOXHD1	AluYd2	chr18:44126165-44126584	419	658	N
	Primers:	TCTTTCCCTTGACAAAAATGC			ACTGAATTGGGATGATTGGA		
6	16302	LOXHD1	AluYa8	chr18:44126100-44126586	486	565	N
	Primers:	GGGAGGAGTTTTAGGGATGC			AACTGAATTGGGATGATTGGA		
7	16304	LOXHD1	AluYc1	chr18:44126165-44126584	419	552	N
	Primers:	TCTTTCCCTTGACAAAAATGC			ACTGAATTGGGATGATTGGA		
8	16304	PCDH15	AluY	chr10:55969867-55970354	487	791	Y
	Primers:	TTTTTGACGCAGTCATAAGTAGC			AGAAGACATTTGCCCTCGAA		
9	15044	ABCC8	L1PREC2	chr11:17436288-17436757	235	274	N
	Primers:	CCCTGCAGTCTGTTGTTCCT			TCTTCAAAAACCACATCACTCAA		
10	15043	MAK	L1PA3	chr6:10806196-10806830	355	450	Y
	Primers:	TCCTGAGAGAGTGGGTTGCT			AGCTTGCAGTGAGCGAAGAT		
11	15044	MAK	L1PA7	chr6:10805875-10806345	289	329	N
	Primers:	TGACGAATATTTTTACAAGCTTTATTG			TGCGAATGTGACCTTATTTTG		
12	Control						

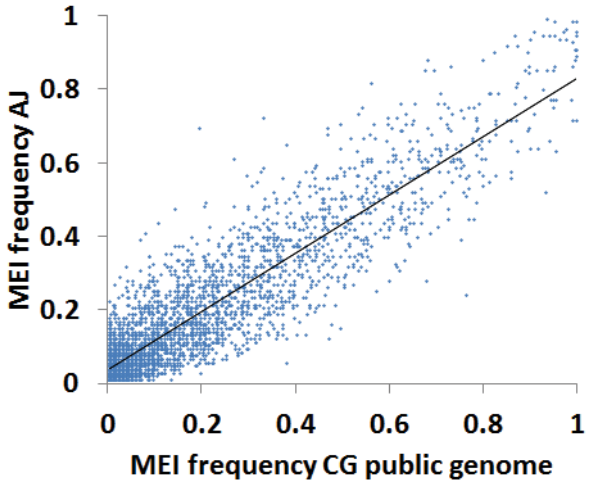
Supplementary Note Table 1. *Experimental details on MEI validation.*



Supplementary Note Figure 3. *Experimental validation of novel MEIs.* See text and Supplementary Note Table 1 for experimental details. Each lane corresponds to a single insertion event. The validated MEIs are indicated with blue arrows.

2.4.4. MEI allele frequencies and ME types

Comparing the frequencies of known MEIs (i.e., those appearing in the 1000 Genomes Project) to the frequencies of the same MEIs in CG's public genomes (http://www.completegenomics.com/public-data/69-Genomes/) revealed high correlation (Supplementary Note Figure 4; $r = 0.90$), despite the diversity of populations sampled by CG. This may be consistent with MEI events representing relatively ancient coalescence events (due to their rarity) and hence being less sensitive to recent demographic history [16]. The fraction of insertions coming from each ME family (Alu, L1, and SVA) is shown in Supplementary Note Table 2. The distribution of known MEIs among the ME families is similar to that of the 1000 Genomes MEIs. For novel MEIs, there is an excess of L1 and SVA insertions.



Supplementary Note Figure 4. *MEI frequencies in AJ and Complete Genomics (CG) data.* The CG data is for 54 unrelated genomes in a diversity panel of worldwide populations. The solid line is a linear fit.

	% of known AJ MEIs	% of novel AJ MEIs	% of 1000 Genomes MEI
Alu	89.5%	54.4%	87.9%
L1	9.0%	38.9%	9.3%
SVA	1.6%	5.1%	2.8%

Supplementary Note Table 2. *The fraction of Mobile Element Insertions (MEIs) per genome, broken by element type and novelty.* Novelty was determined with respect to the 1000 Genomes Project. The fraction of 1000 Genomes MEIs from each family is also shown. Note that the sum of each column is not necessarily 100% due to additional MEIs of other rare families (evident for the novel MEIs).

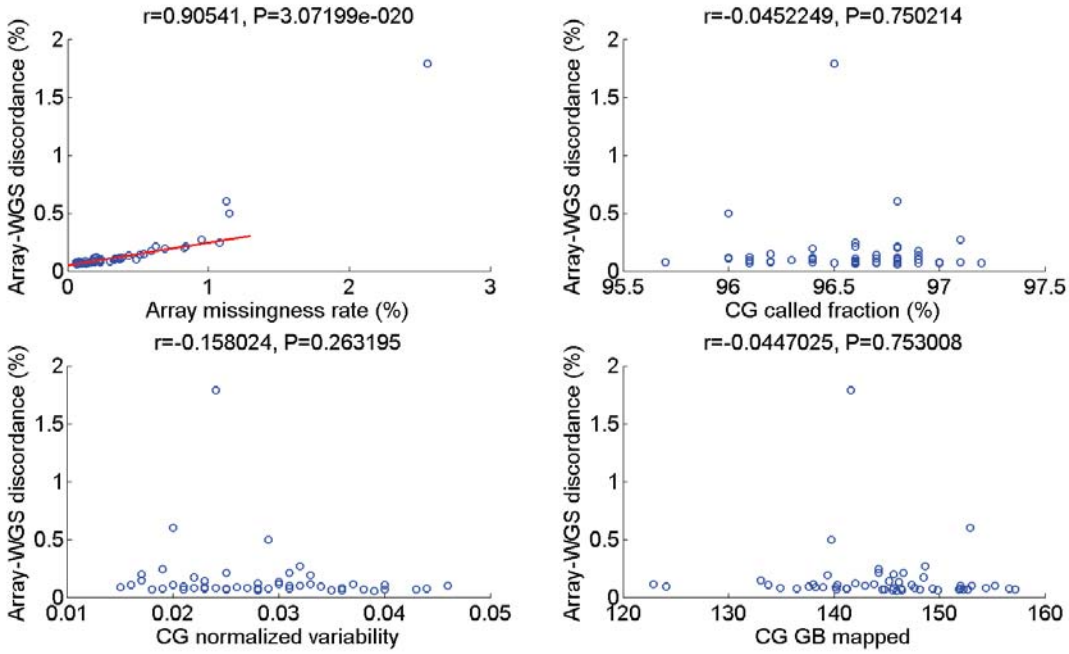
2.5. Concordance with SNP arrays

Genome-wide SNP arrays were available for all samples. Samples from Atzmon's lab were genotyped on Affy 6.0 and called using Birdsuite [17]. Fifty two out of the first 58 and 10 of the 16 in the other batch were jointly processed and cleaned by Kenny et al. (2012) [18]. For the remaining 12 samples, allele codes were determined using Affymetrix's annotation (sites lacking annotation were discarded) and no further cleaning was carried out. Samples from Clark's lab were genotyped on Illumina Human 610k- or 660k-quad bead arrays and processed by Liu et al. (2011) [8].

In all samples (except the 12 Atzmon samples that were directly processed using the Affymetrix annotation), array coordinates were lifted over from hg18 to hg19 using UCSC Genome Browser tools [19], and variants that could not be mapped were discarded. The strand of each allele was determined according to the reference allele, except when it was a no-call or an A/T or C/G polymorphism, in which case it was discarded. Concordance with the sequencing data (in the form of *masterVar* files) was then computed, for each individual separately, using CGA tools (*snpdiff*).

The results are summarized in Supplementary Data 1. The average concordance was 99.85% in the first 52/58 genomes and 99.67% in the entire study (128 genomes). As expected, the concordance was highest for sites genotyped as homozygous-reference (99.94% and 99.82%, respectively), then homozygous non-reference (99.83% and 99.61%), and finally heterozygous (99.72% and 99.46%). The

average fraction of array genotypes not called in the sequencing data was 0.65% in the 52 samples and 0.67% in the entire study (128 individuals). Due to the large degree of heterogeneity in genotyping platforms, processing, and quality, the reported concordance should not be taken as a direct measure of sequencing quality. To demonstrate that, we show in Supplementary Note Figure 5 that the discordance is positively correlated with the array missingness rate (a proxy of the array quality) but not with sequencing metrics such as the fraction of the genome called or the depth of coverage. This result suggests that most discordances are due to genotyping errors; at the limit of no array missingness, linear extrapolation (Supplementary Note Figure 5) gives discordance of 0.047%.



Supplementary Note Figure 5. *Discordance between our whole-genome sequencing data (WGS) and SNP array genotypes.* The discordance is plotted vs. quality characteristics of either the array (top-left panel) or the sequencing data (all other panels; as reported by Complete Genomics (CG)). Data is shown for the 52 of the 58 samples in the first batch that were jointly processed (see text). The Pearson correlation coefficient and the corresponding P-value are indicated at the top of each panel. In the top-left panel we also plot the linear fit of the discordance vs. the array missingness rate, computed using all data points except the three most discordant.

2.6. Ti/Tv analysis

The Ti/Tv (transition/transversion) ratio is known to be about 2.1 in humans and is a useful measure of sequencing quality (e.g., [20, 21, 22, 23, 24]). The "raw" genome-wide Ti/Tv ratio, as reported by Complete Genomics, was between 2.13 and 2.15 for all 128 samples. To determine how the Ti/Tv ratio varied between variant classes, we considered the non-reference single-nucleotide variants (SNVs) in the VCF file of the first 58 genomes, after sex chromosomes and multi-allelic and half-called variants were excluded (see details in section 2.7.1 on VCF generation). The Ti/Tv ratio, averaged over all individuals, was 2.18 for known variants (dbSNP 132; section 2.15) and 2.16 for novel variants. In Supplementary Figure 2 we plot the Ti/Tv ratio (averaged over all individuals) vs. the minor allele count, the number of individuals not-called at the site, the genotype quality, and the read depth (the latter two as reported by

CG). The Ti/Tv was >2, with little fluctuations, with respect to the frequency and quality/depth (within the range where most variants were concentrated). With respect to the missingness rate, however, the Ti/Tv decreased sharply, implying that high no-call rate is a strong indication of lower quality.

2.7. Merging and filtering pipeline

2.7.1. The first batch (58 genomes)

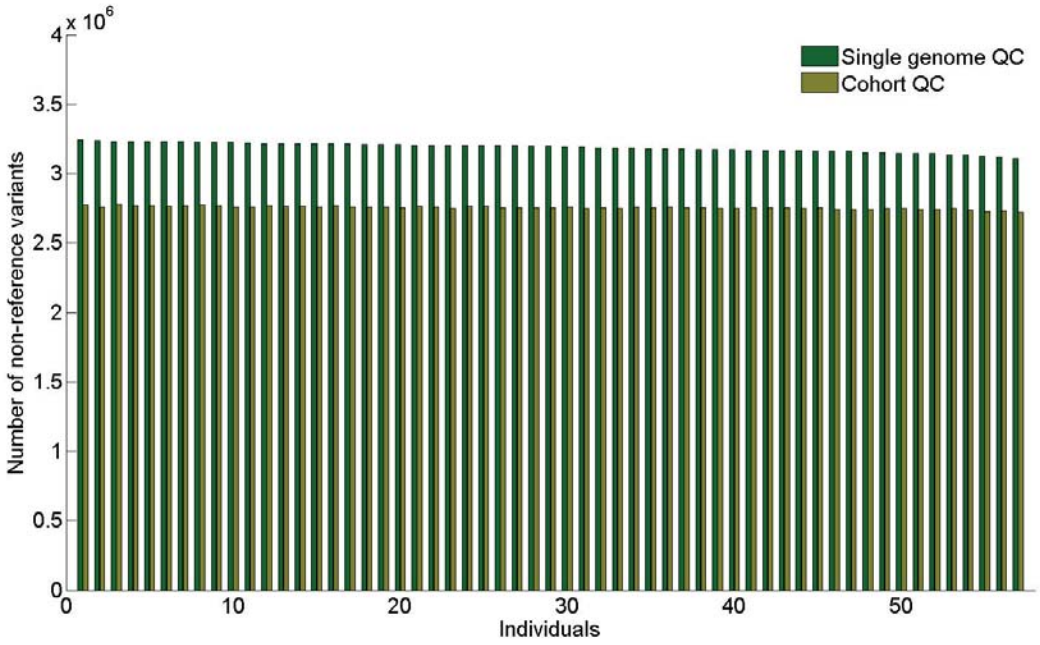
Genotype calls were provided by CG in the form of one *masterVar* file per individual. To merge the individual genomes, we used the CGA tools *mkvcf* command with default parameters and generated a VCF file listing sites where at least one individual was non-reference. We observed that *mkvcf* occasionally generated multi-nucleotide variants that differed from the reference by just a single nucleotide (due to the way in which multi-nucleotide half-calls were processed). We applied a custom script that transformed those variants back into simple single-nucleotide variants.

To obtain high quality genotypes, we first set low quality calls as missing and loaded the VCF file into Plink/Seq (http://atgu.mgh.harvard.edu/plinkseq/). Using Plink/Seq, we discarded variants in the sex chromosome, variants that were not bi-allelic, and multi-nucleotide variants (indels and substitutions). We then removed variants not called in more than three genomes (≈6% no-call rate) and variants not in Hardy-Weinberg Equilibrium (HWE; $P<10^{-6}$). Finally, we set half-calls as no-calls and removed any variants that became monomorphic reference (monomorphic non-reference variants were retained). The remaining genotypes were recorded in Plink format [25]. This stringent quality control procedure, which derives from standard GWAS practices [26] and is similar to pipelines used in comparable sequencing studies [27, 28], was designed to generate genotypes that are most suitable for population genetic analyses. Specifically, indels (or multi-nucleotide substitutions) were excluded due to their high false positive rate (see sections 2.11 and 7.4.2).

Inspection of variant statistics reported by CG indicated that one female individual (GS000010967-ASM) had about 10k more homozygous and 20k less heterozygous variants compared to the rest of the cohort. We later found (see more in section 2.11) that GS000010967-ASM had an exceptional number of runs-of-homozygosity. We concluded that she likely the daughter of cousins, and removed her variants from the cleaned dataset. Genotypes including this individual were used for the purpose of the runs-of-homozygosity analysis described in section 2.11, but not for other population genetic analyses (which were carried out on the filtered set of 57 genomes).

The total number of variants in the VCF file (i.e., the original genotypes) was 19,612,060, of which 11,128,604 were high-quality, bi-allelic SNPs. Further 1,501,127 variants were removed because of high no-call rate, deviation from Hardy-Weinberg equilibrium, being only half-called as non-reference, or presence only in the individual with the consanguineous parents. Of those removed variants, 583,074 were novel (dbSNP132) and 450,373 were singleton with respect to the non-reference allele. The total number of variants (SNVs only) remaining after cleaning was 9,627,477, out of them 26.3% were novel with respect to dbSNP132 (15.2% with dbSNP135), and 24.7% were singletons (in 98.98% of which it was the reference allele).

Per individual, the number of non-reference alleles was ≈$3.188 \cdot 10^6$ (standard deviation (SD) ≈$33 \cdot 10^3$) when applying "local" cleaning only, that is, filtering each individual separately (removing non-autosomal, low-quality, half-calls, and non-SNVs). After "cohort" cleaning, which included filtering multi-allelic variants, variants with high no-call rate, and variants out of HWE, the number of non-reference variants per individual was ≈$2.755 \cdot 10^6$ (SD ≈$12 \cdot 10^3$), a reduction of 13.6%. The average reduction in novel variants (dbsnp135) was higher, as expected, at 28.7% ($P = 3 \cdot 10^{-20}$, rank-sum test). Interestingly, the variance of the number of variants per individual was much lower after cohort-based cleaning compared to the local cleaning (Supplementary Note Figure 6; coefficient of variation (SD/mean) $4.2 \cdot 10^{-3}$ vs. $1.0 \cdot 10^{-2}$; $P = 1.1 \cdot 10^{-7}$, Levene's test), demonstrating the utility of our cleaning pipeline.



Supplementary Note Figure 6. *The number of non-reference variants per individual.* Single-genome QC refers to the variant filters that work on one genome at a time (e.g., removing half-calls or non-SNVs), while cohort QC refers to the complete cleaning process, which includes filters applicable to multiple genomes (e.g., no-call rate or HW equilibrium). The individuals were sorted in decreasing order of their number of variants after single-genome QC.

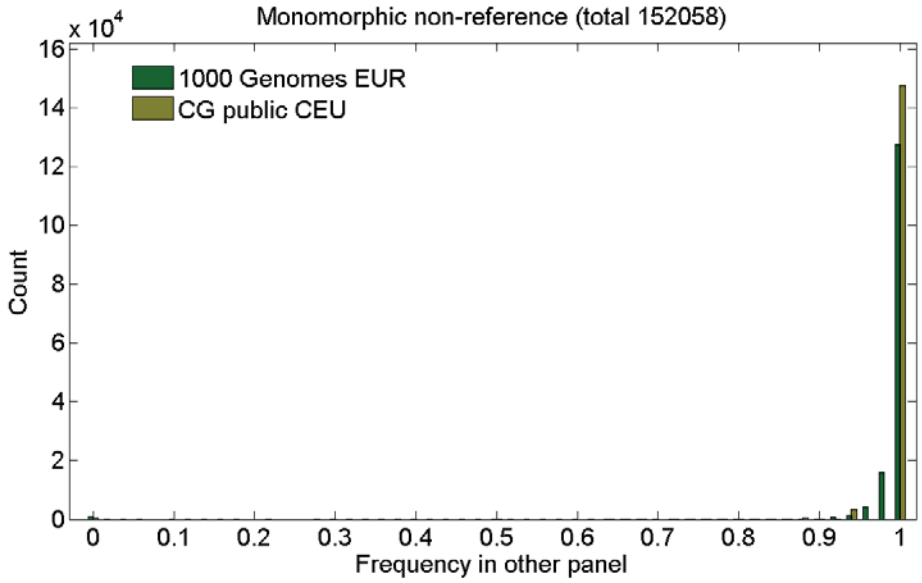
2.7.2. The complete project (128 genomes)

We first removed a number of genomes that were derived from either non-control individuals (n=9), individuals having unclear ancestry (using PCA; section 2.9, n=4), or duplicates (n=1), leaving 128 genomes (including the individual with the consanguineous parents, GS000010967-ASM). We again merged all genomes using CGA tools, but here using the *listvariants* and *testvariants* commands. We then used a custom script to generate a Plink file directly from the *testvariants* output. Our script removed variants just as in the smaller dataset: we retained only autosomal, bi-allelic, single-nucleotide variants that were fully called as non-reference in at least one genome. We further used Plink to filter out variants with >10% no-call rate or not in HWE ($P<10^{-6}$). Most variants removed in the last step were

due to missingness; out of 1,441,960 removed variants (of unfiltered 13,768,157; 10.5%) only 62,220 violated HWE (0.45%). The final number of remaining variants was 12,326,197.

2.8. Monomorphic non-reference variants

Our final cleaned genotypes (for the first batch) contained 152,058 monomorphic non-reference variants. To determine whether those are due to a platform- or a sample-specific error, we inspected their frequencies in other sequencing datasets. Specifically, we extracted the non-reference allele frequencies from the 1000 Genomes project (EUR; ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/) and from CG's public genomes (CEU; ftp://ftp2.completegenomics.com/Diversity/ASM_Build37_2.0.0/). We assigned frequency zero to Ashkenazi variants that were not found in those datasets. The frequency spectra are plotted in Supplementary Note Figure 7, showing that the monomorphic variants tend to be of a very high frequency in the other datasets as well. Specifically, 83.9% and 96.9% of the variants were also monomorphic in the 1000 Genomes and the CG datasets, respectively. We therefore conclude that the vast majority of our monomorphic non-reference variants are not due to a platform- or a sample-specific error.



Supplementary Note Figure 7. *Monomorphic non-reference variants.* The frequency of each monomorphic non-reference variant detected in the first batch of our AJ genomes was extracted from the 1000 Genomes project (Europeans) and CG's public genomes (CEU). The histograms of the frequencies are plotted.

2.9. Ashkenazi Jewish ancestry

To verify that all of our sequenced individuals have Ashkenazi Jewish genetic ancestry, we ran Principal Component Analysis (PCA) on a merged dataset that included our whole genome samples ($n = 128$), the Flemish whole genomes (see section 2.12; $n = 26$), the Human Genome Diversity Project (HGDP) genotypes [2] (Illumina 650k; $n = 939$) and the Jewish HapMap [1] genotypes (Affymetrix 6.0; $n = 237$).

HGDP Genotypes were downloaded from the HGDP website and converted to Plink format using a script from www.harappadna.org. A subset of unrelated individuals was selected according to Rosenberg, 2006 [29]. Following Leutenegger et al., (2011) [30], individual HGDP01097 was also removed. We further filtered out SNPs with >5% no-call rate, monomorphic SNPs, and the sex chromosomes. All coordinates were lifted over from hg18 to hg19 and SNPs that could not be lifted over were removed. The strand was determined according to the reference allele, and negative strand alleles were flipped using Plink. Jewish HapMap (JHM) genotypes were available from Atzmon et al. (2010) [1]. A/T and C/G polymorphisms, non-autosomal sites, and monomorphic sites were removed, and lift over and strand identification were carried out as above.

To create the final dataset for PCA, the Ashkenazi Jewish (AJ) and Flemish Plink files (after filtering) were used. Only SNPs that existed in all four datasets (AJ, Flemish, JHM, HGDP) and that had the same two alleles in all datasets were retained. Merging was carried out using Plink, and the merged dataset was pruned using Plink's --indep-pairwise with parameters 50, 10, and 0.1 (leaving eventually 47,713 SNPs). Finally, we removed all HGDP individuals that were neither European nor Middle-Eastern, as well as the outlier Bedouin individual HGDP00621 (leaving $n = 290$). PCA was then performed using smartPCA [31] with default parameters. The results (first two PCs) are plotted in Supplementary Figure 1, where for clarity, we did not show data points for the JHM Ashkenazi samples (which completely overlapped our Ashkenazi cluster), the JHM samples from Iran and Greece, and the outlier JHM individuals JHM70 (TUR) JHM457 (IRQ). Additionally, each of our AJ sequencing batches (section 1) was given a different symbol.

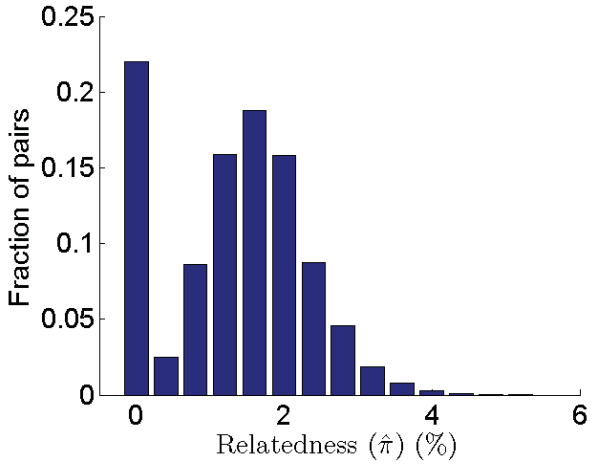
The PCA results largely recapitulate previous Jewish genetics studies [1, 32, 33, 34, 35], showing that the AJ samples cluster tightly between European and Middle-Eastern populations. For our matter here, we note the absence of outliers: all of our samples cluster together, indicating common Ashkenazi Jewish ancestry (except perhaps for GS000010961-ASM and GS000014999-ASM, which are slightly outside the cluster, but nevertheless have neither European nor Middle-Eastern or Jewish non-Ashkenazi ancestry).

Supplementary Figure 1 also demonstrates that samples from different batches cannot be distinguished, and therefore (at least for common SNPs), a batch effect does not exist. We verified that the picture is similar for all first 10 principal components as well as when carrying out PCA of the AJ samples alone. Note that the two samples that are somewhat distant from the cluster come from two different batches, further justifying their inclusion in the study.

To formally test for the absence of substructure in the AJ population, we used the full dataset of our AJ cohort (128 individuals), removed all variants with minor allele frequency >5% and the individual with the consanguineous parents (GS000010967-ASM) and pruned SNPs in LD, as above, leaving finally 165,306 SNPs. We ran PCA as above and computed the Tracy-Widom statistic to test for population substructure [31]. The P-value for the first PC was 0.18, indicating no significant substructure [31].

2.10. Cryptic relatedness

To verify that there is no cryptic relatedness in the cohort (all 128 samples), we ran *Plink*'s --genome command. The average $\hat{\pi}$ was 1.48% and the maximum was 5.48%, indicating no close relatives in our dataset. The distribution of $\hat{\pi}$ values is plotted in Supplementary Note Figure 8.



Supplementary Note Figure 8. *The distribution of the relatedness $\hat{\pi}$ among individuals in our full cohort (128 individuals).*

2.11. Estimation of the false positive rate

2.11.1. Using runs-of-homozygosity

In theory, any two long genomic segments that descend from a recent common ancestor should be separated by just 1-2 mutations, independently of the time to the ancestor or the segment length (see, e.g, [36]). In practice, sequencing errors generate a much larger number of mismatches, and therefore, if identical-by-descent (IBD) segments can be accurately identified, the error rate can be calibrated. While segments shared between individuals can be detected only with moderate confidence (see section 4.1), long segments shared within individuals, or runs-of-homozygosity (or autozygosity), can be relatively accurately detected even in the presence of errors. Note that using this error rate corresponds mostly to false positives flipping a homozygous reference genotype into a heterozygous call.

To detect runs-of-homozygosity (rohs) in our full cohort (128 individuals, which, importantly, included the individual with the consanguineous parents, GS000010967-ASM), we used the cleaned genotypes. We removed variants with minor allele frequency <20% or missingness >5%, then variants that were in LD (using the Plink command --indep-pairwise with parameters 50 10 0.5; leaving 243,880 variants), and then used Plink's --homozyg command to detect rohs. We noticed, however, that even long rohs can be occasionally misidentified, particularly around their borders, which can affect the estimated error rate. This is because considering similar but non-autozygous segments as rohs will count true heterozygous sites as sequencing errors, whereas considering only nearly identical segments will bias the estimate towards the rate at the most error-free regions. Therefore, instead of attempting to obtain a point-estimate of the error rate, we looked for lower and upper bounds by using either high-quality roh segments only or all roh segments, respectively. To detect high quality segment, we used Plink's --homozyg command with parameters --homozyg-window-snp 50 --homozyg-window-het 0 --homozyg-window-threshold 0.25 --homozyg-density 100 and minimum segment length of 7.5MB. To detect all segments, we used the parameter set --homozyg-window-snp 25 --homozyg-window-het 1 --homozyg-window-threshold 0.1 --homozyg-density 50. We then also removed segments around the HLA region (chr6:20-35M).

Using the strict parameters, we discovered nine segments in the individual with the consanguineous parents, GS000010967-ASM, of total length 184MB. In all other individuals, there were 18 segments (in 17 different individuals) of total length 183MB. For the loose parameters, there were 13 segments in GS000010967-ASM (of total length 232MB) and 81 other segments (59 individuals) of total length 598MB. None of the roh segments had an overlap of more than 5% with a gap in the reference genome (e.g., centromeres, telomeres, etc.; according to the UCSC genome browser). The detected segments enabled us to determine the error rate both in the genotypes originally reported by Complete Genomics (SNVs and other variants) as well as in our cleaned genotypes.

In the original genotypes, we removed 1MB from each side of each segment and considered first all heterozygous high-quality SNVs. Using the strict parameters, there were overall 2445 hets in 313MB of sequence, corresponding to an error rate of $7.81 \cdot 10^{-6}$ hets/bp or ≈21,000 errors genome-wide (using the autosomal hg19 sequence length, excluding Ns). Using the loose parameters, there were 8235 hets in 641MB, corresponding to an error rate of $1.28 \cdot 10^{-5}$ hets/bp or ≈34,500 errors genome-wide. When considering all high-quality hets (not only SNVs) using the strict parameters, there were 5152 such variants, corresponding to an error rate of $1.65 \cdot 10^{-5}$ hets/bp or ≈44,200 errors genome-wide. Using the loose parameters, there were 14,477 hets, corresponding to an error rate of $2.26 \cdot 10^{-5}$ hets/bp or ≈60,600 errors genome-wide. While the overall number of SNV and multi-nucleotide errors came out similar, there are, genome-wide, ≈6-fold more SNVs than multi-nucleotide variants, and therefore, the fraction of false non-SNVs is ≈6 times higher than in SNVs, justifying our decision to eliminate non-SNVs from our population genetic analyses. Manual inspection of the errors in GS000010967-ASM revealed, as expected, either high missingness rate or an obvious violation of Hardy-Weinberg equilibrium (e.g., almost all genotypes were hets).

In the cleaned genotypes, we first removed 5000 SNPs (≈1MB) from each side of each segment, and then counted all heterozygous calls along the segment. Using the strict parameters, there were overall 706 hets over 306MB, corresponding to an error rate of $2.31 \cdot 10^{-6}$ hets/bp or ≈6200 errors genome-wide. Using the loose parameters, there were 1821 hets over 614MB, corresponding to an error rate of $2.97 \cdot 10^{-6}$ hets/bp or ≈8000 errors genome-wide. The significant reduction (3-4 fold) in the error rate demonstrates the efficacy of our cleaning pipeline. The average number of hets per segment (using the strict parameters and after cleaning) was 26, justifying our assumption that recent mutations are negligible. Qualitatively similar results were observed for the Flemish genomes (not shown).

2.11.2. Using a duplicate sample

One of the samples we sequenced (not included in the final 128) was a duplicate of another sample (GS000010774-ASM). Comparing the genotypes of those samples gave another, independent estimate of the sequencing error rate. To compare the genomes, we used the CGA tools *mkvcf* command to generate a VCF file containing only the two duplicates. After removing low-quality calls and no-calls, we remained with ≈24,000 SNV differences and ≈47,000 total differences. Assuming symmetry, this amounts to ≈12,000 SNV differences and ≈23,500 total differences per sample. This is somewhat lower than the estimates obtained using runs-of-homozygosity (Section 2.11.1), perhaps since the duplicate samples were derived from the same DNA extraction, such that the actual error rate is somewhat higher than observed by comparing the sequences. Using the CGA tools *testvariants* command, we obtained

similar figures of ≈25,000 and ≈51,000 SNV and total differences, respectively. Our cleaning pipeline, had it operated on the entire dataset including the duplicates, would have decreased the number of (SNV, autosomal) differences to just ≈12,000 (or ≈6,000 per individual).

To summarize, in the absence of ground truth, we employed two independent methods to estimate the sequencing false positive rate. Estimates vary, for SNVs, between 12,000 and 35,000 errors per genome before cleaning (24,000 and 61,000 for all variants) and between 6,000 to 8,000 errors after cleaning.

2.12. Flemish genomes

2.12.1. Samples

The Flemish genomes reported in this paper are from VIB, a life science research institute based in Ghent, Flanders, Belgium. Of the 26 samples, 13 are the parents in a study of seven trios of healthy volunteers (one sample was dropped because it was related to one of the other samples). These samples were recruited under the condition of Flemish ancestry (up to grandparents). Another 10 samples are blood samples from Flemish cancer patients. The remaining three samples consist of two normal control samples, and one sample of an Amyotrophic Lateral Sclerosis (ALS) affected individual, all of Flemish ancestry. Two more samples were initially included in the analysis but then removed during QC (see below). We verified (using the actual sequences) that all remaining 26 individuals are indeed unrelated and of Flemish ancestry (see Supplementary Figure 1). The Flemish samples were sequenced by Complete Genomics, as our AJ samples, but using earlier computational pipelines (1.8, 1.10, and 1.11 vs. 2.0 for AJ) and earlier reference genome version (hg18 vs. hg19). The average raw sequencing depth was 70x and average fraction of the genome called was 95.9%.

2.12.2. Processing pipeline

To merge the Flemish genomes, we ran CGA tools *listvariants* and *testvariants* commands (*mkvcf* is not compatible with the 1.x pipeline). We then lifted over the resulting *testvariants* file from hg18 to hg19. We removed variants that could not have been lifted-over or for which the reference allele has changed. We created a VCF file (to be used later when merging with the AJ genomes; see section 2.13) using the *testvariants2VCF-v2* Perl script available at the Complete Genomics tool repository. Using a custom script, we generated a Plink file from the *testvariants* output. Two genomes were then removed due to unclear ancestry and abundance of runs-of-homozygosity. The rest of the cleaning pipeline was identical to the one used for the AJ samples. Specifically, we removed the non-autosomal variants, multi-allelic and half-called variants, non-SNVs, variants not fully called in any individual, monomorphic reference variants, and variants with >10% call rate or not in HWE ($P<10^{-6}$). The number of remaining SNVs was 7,613,082.

2.13. Merging the AJ and Flemish genotypes

While both the AJ and Flemish samples were sequenced to high coverage by Complete Genomics, a direct comparison of the cohorts was complicated by the use of two different reference genome versions (hg19 for AJ; hg18 for Flemish) and assembly pipelines. The genome comparison methods provided by CGA tools accept only genomes assembled using the same reference, and CG's assembly tools are unavailable to the public. Remapping and reassembling all genomes using the raw reads and

publicly available software, while perhaps being a principled solution, is strongly advised against by CG (due to the peculiarities of their raw read structure; see, e.g., http://www.completegenomics.com/FAQs/Data-Results/#q5) and was also logistically prohibiting. To minimize the heterogeneity due to the differences between the cohorts, we merged the AJ and Flemish genotypes using the following pipeline.

We considered as input the processed genotypes for the 57 AJ genomes of the first batch and the 26 Flemish genomes. Then, we first removed variants that might have exhibited discrepancy between the different reference genomes used. We defined those variants as having hg19 coordinates that either could not be mapped to hg18 (variants that could not be mapped in the opposite direction were removed when processing the Flemish genomes; see section 2.12.2), that mapped to hg18 non-autosomal chromosomes, or that after mapping to hg18 and back to hg19 did not map to the original coordinate. We also removed SNVs with a different non-reference allele in AJ and in Flemish.

Next, a fundamental problem in merging two sets of filtered genotypes from whole-genome sequences is how to treat missing variants. Had sequencing been error-free, a missing variant in one set would indicate that all samples in that set are homozygous-reference. However, suppose a variant exists in one set but not the other; it might be that the variant was observed in the other set but was filtered out. Interpreting the missing variant as homozygous-reference, therefore, would falsely create the impression that the variant is specific to the first set. In our case, we have genotypes available from both before and after cleaning. We therefore approached the merging problem as follows. If a variant was found in the cleaned genotypes of both sets, it was retained. If it was found in the cleaned genotypes of the first set but never in the second set, it was retained and the second set was assigned the homozygous-reference genotype. However, if the variant (found in the cleaned genotypes of the first set) was found in the original, but not in the cleaned, genotypes of the second set, it was removed from both sets. This strategy has the advantage of enabling set-specific processing, without falsely creating set-specific variants (although at the expense of missing some true variants). In our case, the original genotypes were extracted from our VCF files and the cleaned genotypes from our filtered Plink files (see sections 2.7.1 and 2.12.2). All downstream analyses involving AJ and Flemish genome comparisons were carried out on the above-described merged dataset, even when single-population parameters were compared (in which case we considered only individuals and variants specific to the population under consideration, e.g., as in section 3.1). The number of SNVs in the merged dataset was 10,499,312.

2.14. Phasing and imputation of sporadically missing genotypes

2.14.1. Using molecular phasing information

Complete Genomics provides partial molecular phasing information along with its reported genotypes. This is implemented by assigning a "HapLink ID" to heterozygous alleles that were sequenced in the same read or mate-paired reads. Recently, a *SHAPEIT*-based [37] phasing tool called *seqphase* was developed to take advantage of this information [38]. To create the molecular phasing input for *seqphase*, we considered each of our *masterVar* files separately. From each file, we extracted all autosomal, high-quality, heterozygous SNPs that had HapLink ID for both alleles. We further filtered out sites that were absent or no-call in the cleaned Plink file. We then searched for chains of alleles with identical HapLink

ID, and for each pair detected, we created an entry in *seqphase* format that had the coordinates of the linked sites as well as the linked alleles. To run *seqphase*, we used the genetic maps of HapMap2 [39] and the parameters -burn 10 -prune 20 -main 50. *seqphase* was unpublished at the time of our data analysis; we used a developer version kindly provided by Fouad Zakharia of Stanford University. Using *seqphase*, we were able to phase most (but not all) chromosomes of the cleaned genotypes of the 57 AJ individuals (the first batch), before running out of computational resources. We therefore limited the use of those phased genotypes to the imputation analysis, where only chr1 was used (section 5). We verified that all alleles for which we provided *seqphase* with molecular phasing information were correctly phased; in the absence of ground truth phase, we did not further benchmark the phasing quality.

2.14.2. Using *SHAPEIT*

To phase and impute (for sporadic missingness) the merged 83 (57+26) AJ-Flemish genomes, we used *SHAPEIT* version 2 [40], without employing any molecular phasing information. As recommended by *SHAPEIT*'s authors, for samples of size <100, a reference panel needs to be provided. We used the 1000 Genomes Project reference panel (all populations), available from *SHAPEIT*'s website (www.shapeit.fr). Alleles that were in strand conflict with the 1000 Genomes data were removed (leaving 10,473,620 SNVs), and SNVs appearing in the merged AJ-Flemish genomes but missing in the 1000 Genomes data were added to the 1000 Genomes as homozygous-reference. *SHAPEIT* was run with the 1000 Genomes genetic map and default parameters, except for a window size of 0.5 (as recommended for sequencing data). Unless otherwise mentioned, all population genetic analyses reported below used the phased and imputed dataset.

We also used *SHAPEIT* to phase and impute the genotypes for the complete AJ dataset (128 genomes). As the number of samples was >100, we did not use an external reference panel. All other parameters were as above.

2.14.3. Assessment of phasing quality

When phasing using *SHAPEIT*, the molecular phasing information was not utilized. We could therefore use it as ground truth to evaluate the phasing quality, as measured by the switch error rate. The molecular phasing dataset described in section 2.14.1 was used, and we stratified the switch error rate by whether or not one of the sites was a singleton. For the AJ-Flemish merged data, the switch error rate was 0.965% (0.960% for AJ and 0.98% for Flemish) for all variants and 0.329% for non-singletons. For the complete AJ dataset (128 genomes), the switch error rate was 0.892% for all variants and 0.268% for non-singletons.

2.15. dbSNP comparisons

Coordinates of variants in dbSNP (version 135) were obtained from the UCSC Table Browser [41] group: Variation and Repeats, track: All SNPs (135), table: snp135, and filtered according to molType="genomics"; class="single"; locType="exact"; weight=1; exceptions=does not match everything except "SingleClassTriAllelic" or "SingleClassQuadAllelic". dbSNP132 coordinates were similarly extracted.

3. Supplementary Note 3: Comparison of variant statistics between AJ and Flemish

3.1. Variant counts, heterozygosity, and novelty

Basic variant statistics were computed for each population using the merged and imputed AJ-Flemish genotypes and are presented in Supplementary Table 3 and Figure 1 of the main text. The reported quantities are the mean and standard deviation (SD) over 57 AJ and 26 Flemish individuals. All P-values were computed using the non-parametric ranksum (Mann-Whitney U) test. The results did not qualitatively change when we normalized the first three properties (variant counts) by the fraction of the genome called in each individual.

The results suggest that AJ have a slightly but significantly larger number of variants (1.5%) compared to Flemish, mostly due to increased heterozygosity (2.4%; the number of homozygous variants was only 0.13% larger in AJ (P=0.30)). The fact that AJ show higher genetic diversity than Flemish is somewhat surprising. Recently, purportedly unrelated AJ were found to share, by-descent, a large fraction of their genomes (see also section 4), which was inferred to be due to a recent narrow bottleneck [1, 36, 42, 43]. European populations do not exhibit such a large degree of IBD sharing, suggesting no bottlenecks in their recent history [42]. Additionally, mutations for a number of genetic diseases show elevated frequencies in AJ, again, consistently with strong genetic drift [44, 45, 46]. We would therefore expect to see *less* genetic diversity in AJ compared to Flemish. There are several possible resolutions, which we explore in section 6.1. For now, we note that the larger AJ genetic diversity has been in fact already observed using SNP arrays [34, 47, 48], microsatellites [33], and even Y-chromosome markers [49] (see, in particular, the discussion in Bray et al. (2010) [34] and Behar et al. (2004) [49]).

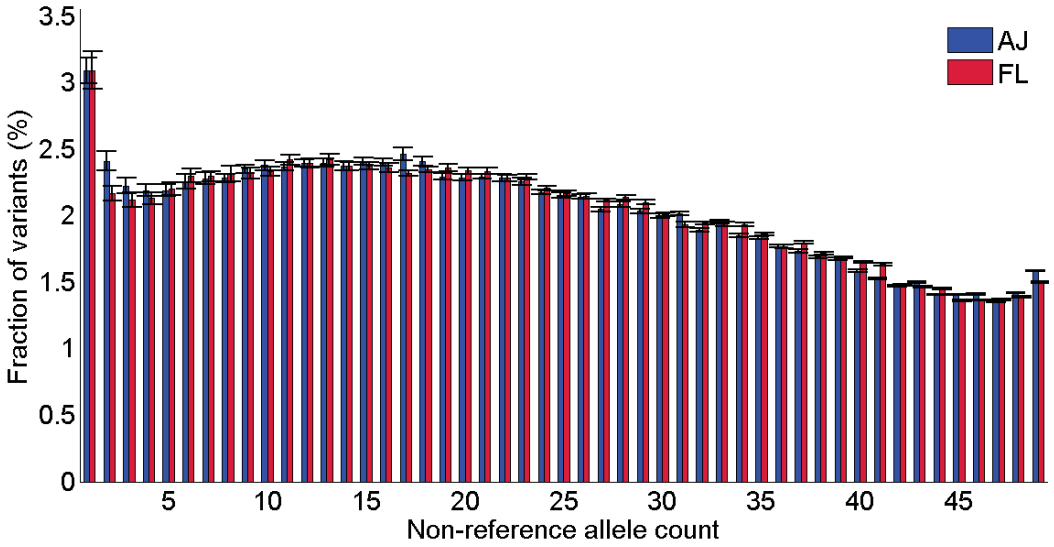
Finally, it is interesting to note that AJ have lower or equal SD in all categories. However, this was significant only for the dbSNP novelty (either dbSNP version; Levene's test)) and may represent a larger heterogeneity in the Flemish sample selection and sequencing quality rather than a true population difference.

3.2. The frequency spectrum of the variants within each individual

When considering all variants in a cohort, most variants are rare, and the frequency spectrum decreases sharply with increasing frequencies. When concentrating only on the variants carried by a single individual, the frequency spectrum (where the frequency is still defined with respect to the cohort) changes, with most variants being common (e.g., [50, 51]). To compute the per-individual frequency spectrum, we first down-sampled each population to n=25 individuals. Then, for each individual, we determined the counts of its non-reference alleles (in each population separately) and generated a histogram. Finally, we removed monomorphic variants and averaged the normalized spectrum over all individuals in each population.

The average fraction of singleton non-reference variants was the same in AJ and Flemish and equal to 3.09%. The average fraction of doubletons was 2.41% in AJ and 2.17% in Flemish. The complete per-individual spectra are plotted in Supplementary Note Figure 9, and indeed, there is only little excess of rare variants: for any given individual, most variants are common (cf. the population frequency

spectrum in Figure 3 of the main text). AJ have a slightly larger fraction of rare variants than Flemish (variants with frequency $\leq 10\%$ vs. all others; $P < 10^{-16}$; χ^2-test). There does not seem to be any consistent trend for higher frequencies. We defer a population-genetic interpretation to later sections (specifically 3.5 and 6).



Supplementary Note Figure 9. *The per-individual frequency spectrum.* After sampling 25 individuals (50 chromosomes) from each population, we calculated, for each individual, the frequency spectrum of its non-reference alleles. We plot the average and standard deviation over all individuals in each population.

3.3. Utility of the sequencing panel for clinical genetics

In Figure 1 of the main text, we plot the fraction of variants in an AJ individual that are found in a panel of either other AJ or Flemish. For that analysis, we used the merged genotypes (57 AJ, 26 Flemish) after phasing and imputation of sporadically missing genotypes. Variant novelty was determined based on dbSNP135 (see section 2.15). To find variants that are novel *and* non-synonymous (non-syn.), we used *ANNOVAR* [52]. Specifically, we used the --geneanno command and considered as non-synonymous each exonic variant that was not annotated as "synonymous" or "unknown" as well as splicing variants. For each AJ individual, we selected 26 other, random AJ individuals (to match the Flemish panel size) out of the remaining 56. For each novel (or novel and non-syn.) non-reference variant found in the given individual, we determined whether it appears in any of the (26) Flemish individuals or the selected 26 AJ individuals. The reported counts are the average and standard deviation over all 57 AJ individuals. To compute the number of variants left after filtering with a panel of 127 individuals, we used the complete dataset of 128 AJ individuals. For each of the 128 AJ individuals, we counted how many of their novel (or novel and non-syn.) non-reference variants appear in any of the other 127. We then reported the average and standard deviation over all 128 individuals. Comparing the number of variants in each category, all differences were significant ($P < 10^{-10}$; rank-sum test).

3.4. Rate of variant discovery

3.4.1. Non-reference variants

To compute the number of non-reference variants discovered with the sequencing of each additional genome, we used the AJ-Flemish merged genotypes after phasing and imputation. For each population, and for each of 50 iterations, we ordered the individuals randomly and counted the number of non-reference variants (regardless of zygosity) in the first n individuals, where $n = 1, \ldots, 57$ for AJ and $n = 1, \ldots, 26$ for Flemish. We then averaged the number of discovered variants over all iterations. To predict the number of discovered variants in a sample size larger than ours, we used the jackknife estimator of Gravel et al. (2011) [3], which is "based on sampling theory and inspired by an analogy with capture-recapture approaches to estimating animal population sizes". Specifically, the estimator uses the number of non-reference variants that appeared once, twice, or three times in a sample to predict the total number of variants that will be discovered in a larger sample. The results are shown in Supplementary Figure 3, demonstrating the same trend as Figure 1 of the main text: while in our sample the number of discovered variants was greater in AJ than in Flemish, an opposite trend is predicted for larger samples.

3.4.2. Segregating sites

The variant discovery rate can also be predicted based on demographic historical models and population genetics theory, which predicts the number of segregating sites in a sample. To calculate the empirical number of segregating sites in the AJ and Flemish samples, we used the same dataset and approach as in section 3.4.1, except that for a given subset of individuals, a site is segregating only if both alleles have been seen (i.e., seeing the non-reference allele alone did not designate the site as segregating). The number of segregating sites for $n = 1$ is, of course, just the average heterozygosity.

To calculate the theoretical expected number of segregating sites in a model of constant-size population (Wright-Fisher; WF), we first estimated the scale mutation rate θ as the number of sites discovered after sequencing one individual (i.e., the average heterozygosity). The average number of segregating sites observed after sequencing n (diploid) individuals, $S(n)$, is [53]

$$(1)\ S(n) = \theta \sum_{i=1}^{2n-1} (1/i).$$

For the more complex models with variable historical population size, the average number of segregating sites was calculated by Živković and Stephan (2011) [54] based on the diffusion equation for the evolution of the allele frequency spectrum. Specifically, we used a slightly simplified version of their Eq. (37),

$$(2)\ S(n;t) = \theta \sum_{k=1}^{n} \frac{(4k-1)\binom{2n}{2k}}{\binom{2n+2k-1}{2k}} \int_{-\infty}^{t} \exp\left(-\int_{s}^{t} \frac{\binom{2k}{2}}{\rho(u)} du \right) ds,$$

where $\rho(t) = N(t)/N_0$. In Živković and Stephan's (2011) demographic model, $t = 0$ is some time in the past before which the population size has always been N_0 (diploids) and $N(t)$ is the population size $2N_0 t$ generations later. Finally, the scaled mutation rate is defined as $\theta = 4N_0\mu$, where μ is the mutation rate per generation per site. Here too, we estimated θ as the average heterozygosity. The

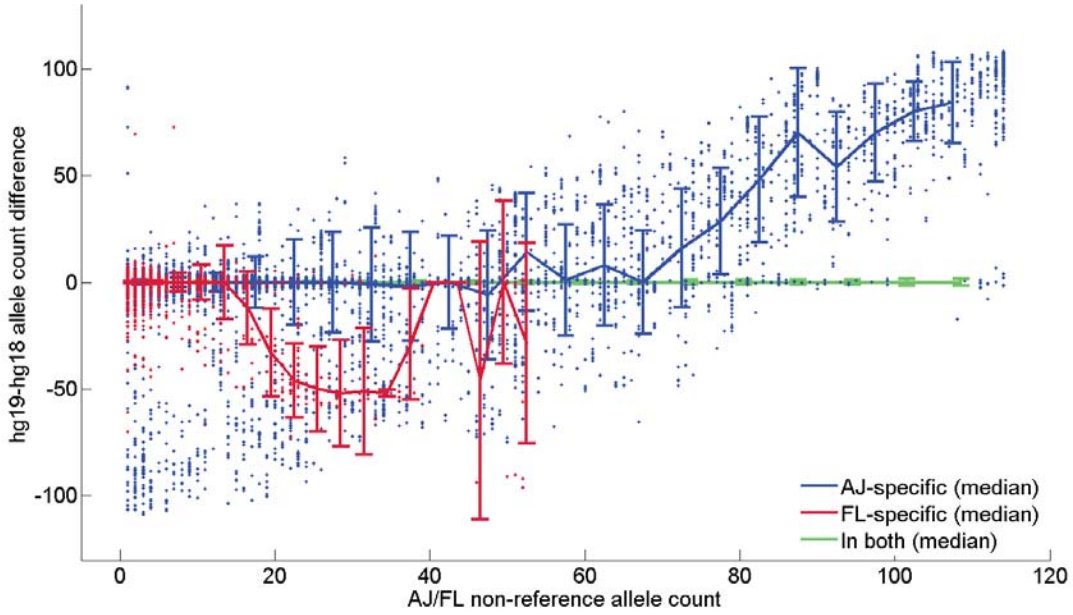
population size at each generation was computed based on demographic models of an ancient bottleneck followed by slow exponential growth that we inferred using the allele frequency spectra of AJ and Flemish (section 6.3.3). For AJ, we also used a model of an ancient bottleneck/growth with an additional recent one (corresponding to the recent AJ expansion and inferred using IBD sharing; section 4.3.2), giving similar results (not shown). We also obtained qualitatively similar results also when considering the joint AJ-Flemish demographic model (section 6.3.3). To evaluate the integrals in Eq. (2), we used *Matlab*'s *quadgk*. We computed the ratio of the binomial coefficients in the prefactor by taking the logarithm of the ratio and then exponentiating; to compute $\ln m!$, we used the exact factorial up to $m < 20$ and otherwise the Stirling approximation, $\ln m! \approx m \ln m - m + \ln(2\pi m)/2$. The empirical and theoretical numbers of new segregating sites ($S(n) - S(n-1)$) are shown, for both populations, in Figure 1 of the main text.

To attach significance to the observed trend (larger number of Flemish variants for large samples), we sampled demographic models for each population using a parametric bootstrap approach, as explained in section 6.3.2. For each cohort that we simulated with the maximum-likelihood parameters, we used the bias-corrected inferred demographic history to compute the expected number of segregating sites. For all ($n = 99$) bootstrapped demographic reconstructions of the two populations, the number of Flemish sites exceeded the number of AJ sites when sequencing $n > 212$ individuals (and on average when sequencing $n > 178$).

3.5. The allele frequency spectrum

3.5.1. Likely artifacts in high-frequency, population-specific variants

Initial inspection of the joint AJ-Flemish allele frequency spectrum revealed a relatively large number of high-frequency, population-specific non-reference variants. In AJ, there were 2000 non-reference variants with frequency ≥50%; in Flemish, there were 159 such variants. While these numbers are virtually negligible relative to the total number of variants (≈10 million in the two populations), we also noted that our demographic (neutral) models predicted zero population-specific variants at those frequencies (see, e.g., section 6.3.5). We therefore suspected that those frequency differences are, for the most part, artifacts of the different reference genome used for variant calling (hg18 for Flemish, hg19 for AJ). To test this hypothesis, we used CG's 54 public genomes (ftp://ftp2.completegenomics.com/Multigenome_summaries/), called either using hg18 or hg19. For each variant found in either AJ or Flemish and that was also found in the public genomes, we recorded the non-reference allele count difference between the hg19 and hg18 versions the public genomes. In Supplementary Note Figure 10, we plot the hg19-hg18 differences vs. the non-reference allele count for variants that are either AJ-specific, Flemish-specific, or appear in both populations. As expected, for variants that are observed in both populations, the allele counts in hg19 and in hg18 are usually the same. However, AJ-specific variants tend to have higher counts in hg19, and Flemish-specific variants tend to have higher counts in hg18. Hence we conclude that even though we removed, prior to merging, all variants with potential mapping problem (i.e., coordinate that did not map from hg18 to hg19 or backwards; or did not remap to the original coordinate), some poorly mapped variants escaped this filtering. We therefore removed from the allele frequency analysis (including the demographic inference of section 6.3) all population-specific variants of frequency of 25% or higher (4048 variants).



Supplementary Note Figure 10. *Likely artifacts in high-frequency population-specific variants.* For each variant found in either AJ-only (blue), Flemish-only (red), or both (green), and that is found in the Complete Genomics public genomes, we plot the difference between the allele count (in the public genomes) when called either using hg19 or using hg18. Each variant is plotted as a dot (only for the population-specific variants; the allele count difference is shifted by a random $\mathcal{N}(\mathbf{0},\mathbf{1})$ to improve visibility), the medians over each allele count bin are connected by lines, and the error bars show standard deviations. While variants found in both populations show no bias, high frequency variants in AJ only tend to have higher allele counts in hg19 (and vice versa), indicating a likely mapping artifact.

3.5.2. Computing the spectrum, projecting to equal sample sizes, and folding

The joint non-reference allele frequency spectrum was computed using the merged and imputed AJ-Flemish genotypes. High-frequency population specific variants were removed as explained in section 3.5.1. As there were 57 AJ and 26 Flemish, the resulting spectrum was of size 115x53. To perform population comparisons, we reduced the spectrum to equal population sizes of 50 haploids each. To avoid sampling artifacts, we computed the expected down-sampled variant counts analytically. Denote the original AJ population size as $N_1 = 114$, the original Flemish population size as $N_2 = 52$, the new population size (for both) as $n = 50$, and denote the original number of sites with i_1 (non-reference) alleles in AJ and i_2 (non-reference) alleles in Flemish as $C[i_1, i_2];\ i_1 = 0, \dots, N_1, i_2 = 0, \dots, N_2$. The expected counts after down-sampling, $S[i_1, i_2];\ i_1 = 0, \dots, n, i_2 = 0, \dots, n$, are given by

$$(3)\ S[i_1, i_2] = \sum_{j_1=i_1}^{N_1} \sum_{j_2=i_2}^{N_2} C[j_1, j_2] \times HG(i_1, n, j_1, N_1) \times HG(i_2, n, j_2, N_2),$$

where $HG(i, n, j, N)$ is the hypergeometric probability of ending up with i labeled items when choosing n items out of total N, j of which are labeled $\left(HG(i, n, j, N) = \binom{j}{i}\binom{N-j}{n-i}/\binom{N}{n}\right)$. After applying Eq. (3), $S[0,0]$ was set to zero (variants that disappeared due to down-sampling).

To obtain the joint minor allele frequency spectrum, S_{min}, the spectrum was "folded" across the diagonal. We set $S_{\min}[i_1, i_2] = S[i_1, i_2] + S[n - i_1, n - i_2]$ and $S_{min}[0,0] = 0$ for $i_1 + i_2 < n$ and then $S_{\min}[i_1, i_2] = 0$ for $i_1 + i_2 > n$. For sites on the diagonal, we set $S_{\min}[i, n - i] = S_{\min}[i, n - i] = (S[i, n - i] + S[n - i, i])/2$ for $i = 0,1, \dots, (n/2 - 1)$ and $S_{\min}[n/2, n/2] = S[n/2, n/2]$. To marginalize the (non-reference allele) spectrum over one of the populations, say, over the Flemish, we used $S_{\text{AJ}}[i] = \sum_{j=0}^{n} S[i, j]\,; i = 1, \dots, n$. To compute the spectrum for, say, AJ-specific variants, we used $S_{\text{AJ-spe}}[i] = S[i, 0]\,; i = 1, \dots, n$. To fold a single-population spectrum, we used $S_{\min}[i] = S[i] + S[n - i]; i = 1, \dots, (n/2 - 1)$ and $S_{\min}[n/2] = S[n/2]$.

3.5.3. The single-population spectrum

The total number of variants (after down-sampling) was 7,316,494 in AJ and 7,083,447 in Flemish. The number of population-specific variants was 1,866,963 in AJ and 1,633,916 in Flemish, leaving 5,449,531 variants polymorphic in both populations. In Figure 3 of the main text, we plot the single-population normalized frequency spectra for AJ and Flemish. We also plot the theoretical expected normalized spectrum for a constant size population, or the Wright-Fisher (WF) model [53],

$$(4)\ S_{\text{WF}}[i] = \frac{1/i + 1/(n-i)}{\sum_{j=1}^{n-1} 1/i}\,, i = 1, \dots, (n/2 - 1)\,;\ S_{\text{WF}}[n/2] = \frac{2/n}{\sum_{j=1}^{n-1} 1/i}.$$

The fraction of singletons was about the same in both populations. The fraction of doubletons was larger in AJ compared to Flemish, while the fraction of variants of frequencies >10% was slightly smaller. Note, however, that the total number of variants is larger in AJ across all frequency bins (Supplementary Figure 9). As expected, for both populations, the Wright-Fisher model underestimated the number of singletons. However, the discrepancy is not as large as observed in other studies (e.g., [51, 55, 56, 57]), likely due to our small sample size (see also section 3.5.5). In Figure 3 of the main text, we also plot the normalized frequency spectra for AJ- and Flemish- specific variants, showing a depletion of singletons in AJ but an excess of variants across all higher frequencies. Here too, the total number of variants is larger in AJ for all frequency bins (Supplementary Figure 9, inset). We validated that all the results for the frequency spectra were qualitatively identical when using the non-reference allele frequency (instead of the minor allele frequency; not shown).

In Supplementary Figure 10, we plot the fraction of variants that are population specific for each minor allele frequency. The fraction of variants that are population specific reaches ≈0.7 for singletons, and is larger in AJ for all frequencies (particularly for intermediate ones).

3.5.4. The joint AJ-Flemish spectrum

The joint AJ-Flemish (minor) allele frequency spectrum is plotted in Figure 3 of the main text. The results were qualitatively similar when using the non-reference allele count. The spectrum shows correlation of allele frequencies between the two populations ($r = 0.88$). To determine, qualitatively, whether the AJ and Flemish populations are distinct or are, alternatively, two samples from a single population, we compared the real spectrum to the expected spectrum assuming random mating. Under random mating (or a panmictic population), the total number of copies of each variant (i.e., AJ+Flemish) can be distributed randomly between the two populations. Mathematically, if $S[i, j]$ is the minor allele frequency spectrum (after down-sampling to n haploids in each population), define $T[k]$ as the number

of copies of the minor allele in the combined population: $T[k] = \sum_{i+j=k} S[i,j]$, $k = 1, \ldots, n$ (where in the sum, $i = 0, \ldots, n$ and $j = 0, \ldots, n$). Then the panmictic spectrum, $S_p[i,j]$, is

$$(5)\ S_p[i,j] = T[i+j] \cdot HG(i, n, i+j, 2n),$$

where $HG(i, n, j, N)$ is the hypergeometric probability of ending up with i labeled items when choosing n items out of total N, j of which are labeled. The last equation follows because to observe i AJ and j Flemish minor alleles when randomly choosing the n AJ alleles, we need i alleles to be the minor (out of total $i+j$). Eq. (5) is equivalent to Eq. (3) of Gravel et al. (2011) [3]. The panmictic spectrum is plotted in Figure 3 of the main text, demonstrating that the two populations are distinct.

3.5.5. Population genetic parameters: θ, Tajima's D, F_{ST}, and f_2 variant sharing

A number of population genetic indices can be computed directly from the allele frequency spectrum. Denote the (single-population) spectrum as $S[i], i = 1, \ldots, n/2$, where $S[i]$ is the number of sites with i copies of the minor allele. The average number of nucleotide differences, π (also equal to the average heterozygosity) is $\pi = \sum_{i=1}^{n/2} \frac{i(n-i)}{n(n-1)/2} S[i]$, and the total number of sites, S, is simply $S = \sum_{i=1}^{n/2} S[i]$. According to the theory [53], for a constant-size population (the Wright-Fisher model), the scaled mutation rate, $\theta = 4\mu N_e$, (where μ is the genome-wide mutation rate per generation and N_e is the effective populations size) can be estimated either as $\hat{\theta}_\pi = \pi$ or as $\hat{\theta}_S = S/\left(\sum_{i=1}^{n-1} 1/i\right)$ (Watterson's estimator). Assuming the mutation rate is $1.44 \cdot 10^{-8}$ per bp per generation [58] and using the (autosomal non-N) genome size, $2.685 \cdot 10^9$ bp (for further correction due to false negatives, see section 6.2.1), we can immediately obtain an estimate of N_e. For AJ, we have $\hat{\theta}_\pi = 1.634 \cdot 10^6$ and $\hat{\theta}_S = 1.633 \cdot 10^6$, giving population size estimates of $\hat{N}_{e,\pi} = 13{,}048$ and $\hat{N}_{e,S} = 13{,}041$, respectively. For Flemish, $\hat{\theta}_\pi = 1.603 \cdot 10^6$ and $\hat{\theta}_S = 1.581 \cdot 10^6$, giving $\hat{N}_{e,\pi} = 12{,}797$ and $\hat{N}_{e,S} = 12{,}626$, respectively. Therefore, the AJ effective population size (as we already mentioned in section 3.1) is slightly larger than that of the Flemish. This is of course not to say that the AJ population size has always been larger (or smaller; in fact, it has likely been highly variable (e.g., sections 4.3 and 6.3)). It is to say that when summarizing the genetic diversity into a single statistic, the AJ population seems more diverse. Indeed, the picture changes when using more complex demographic models (section 5.3), as we have seen in section 3.4. In section 6.1, we investigate possible reasons for the AJ's increased heterozygosity.

Tajima's D is a statistic based on the difference between the two estimates of θ [59]. For both populations, the (genome-wide) Tajima's D was positive (0.0018 for AJ, 0.0497 for Flemish). We expect negative values for larger samples, where an excess of rare variants (due to the recent growth) is expected to reduce Tajima's D sharply (see, e.g., Figure S9 in Tennessen et al. (2012) [51]). The fixation index, F_{ST}, a measure of population differentiation, was calculated using *∂a∂i* [4] (based on [60]), and came out as 1.56%. This value of the F_{ST} is of the same order of magnitude as previously found between AJ and European (non-Jewish) populations [1, 34, 35, 48]. Variants that appear twice in the entire sample are known as f_2 variants [23]. In our dataset, 39.9% of f_2 variants were AJ-specific, 27.2% were Flemish-specific, and 32.9% were shared.

4. Supplementary Note 4: Identical-by-descent (IBD) shared segments

4.1. Detecting IBD segments

4.1.1. Initial detection

Identical-by-descent (IBD) shared segments were detected based on the genetic map distance between sites. We used the HapMap2 genetic maps [39], linearly interpolated at sites not in the map. We then used *Germline* [42, 61], a window-based IBD detection tool that works by finding and extending seeds of exact matches. We used the default parameters except for a window size of 100 (-bits), one allowed homozygote mismatch per window (-err_hom), and one allowed heterozygote mismatch per window (-err_het). We ran *Germline* in the "genotype extension" mode [42], in which a matching segment is extended either if one of the haplotypes is matching or if sites that are homozygous in both individuals are matching. While the second criterion is rather liberal (and is expected to lead to false positives), we preferred to avoid over-dependence on phasing quality and employed an extensive series of post-*Germline* filtering steps (sections 4.1.2 and 4.1.3), which, as can be seen in section 4.2, seem to reduce the number of false positives considerably. Finally, we used a minimal segment length of either 3cM or 5cM (-min_m).

4.1.2. Initial filtering

In the initial filtering step, we removed segments whose length in MB was (numerically) <0.4 of their cM length (the average ratio is ≈0.8). We further removed segments with >10% overlap with any of the gaps in the reference genome (UCSC Table Browser [41]). We then retrieved, for each shared segment, the original genotypes of the two individuals along with their allele frequencies. We removed segments where all (double homozygous) matching genotypes were of the major allele.

4.1.3. Additional filtering using segment scores

We further filtered our segments based on a score related to the probability of a segment to be truly shared by-descent. We considered only sites homozygous in both individuals. For each segment, we computed a score as a product of approximate likelihoods over all double-homozygous sites, either under the hypothesis that the segment is truly IBD, or assuming it is a random segment (see below). Finally, we filtered out all segments with a score ratio less than an arbitrary cutoff.

For each (double homozygous) site i, denote by $M(i)$ the indicator that the two individuals are matching and by p the probability of having a sequencing error or a recent mutation. For a pair of truly IBD haplotypes, the probability of a site to be matching (neglecting the possibility of a double error) is $1-p$, and the probability of the site to be a mismatch is p. We used $p = 0.001$, which is the order of magnitude of the estimated error rate (sections 2.5 and 2.11). Taken together, for a segment containing sites $i = 1, \ldots, n$, the approximate likelihood is

$$(6)\ P_{\mathrm{IBD}} = \prod_{i=1}^{n} (1-p)^{M(i)} p^{1-M(i)}.$$

For a random, non-IBD segment, denote the minor allele frequency at site i as f_i and assume Hardy-Weinberg equilibrium at all sites. We also assume first that it is given that at least one of the individuals

is homozygous to the minor allele. The probability of a minor allele match is therefore (neglecting sequencing errors)

$$(7)\ p_{\text{minor match}} = \frac{f^4}{f^4+2f^2(1-f)^2} = \frac{f^2}{f^2+2(1-f)^2},$$

where, e.g., f^4 is the probability of both individuals to be homozygous to the minor allele, etc., and the probability of a mismatch is

$$(8)\ p_{\text{mismatch}} = 1 - p_{\text{minor match}} = \frac{2(1-f)^2}{f^2+2(1-f)^2}.$$

For the probability of a major allele match, we assume that it is given that at least one individual is homozygous to the major allele, and then

$$(9)\ p_{\text{major match}} = \frac{(1-f)^4}{(1-f)^4+2f^2(1-f)^2} = \frac{(1-f)^2}{(1-f)^2+2f^2}.$$

Since the probabilities in Eqs. (7), (8), and (9) do not sum to 1, they should be thought of as scores that aim to capture the degree of surprise in the observed match or mismatch. The distinction between major and minor matches is needed, because otherwise, the probability of a match would be $p = \frac{f^4+(1-f)^4}{f^4+2f^2(1-f)^2+(1-f)^4}$, which would be high even when the match is of the minor allele (which is ought to be surprising).

Denote by ρ_i the frequency of the shared allele (for a match). The score of a segment is given by the product of the probabilities over all sites,

$$(10)\qquad P_{\text{random}} = \prod_{i=1}^{n}\left[\mathrm{M(i)}\frac{{\rho_i}^2}{{\rho_i}^2+2(1-\rho_i)^2} + \left(1-M(i)\right)\frac{2(1-f_i)^2}{{f_i}^2+2(1-f_i)^2}\right].$$

Note that as the frequencies in neighboring sites are not independent, this is again only an approximation. Our final quality score for the segment is the log-ratio of the probabilities under the IBD and the random segment hypotheses,

$$(11)\qquad \mathrm{D} = \log(P_{\text{IBD}}) - \log(P_{\text{random}}).$$

We removed segments with D smaller than an arbitrary cutoff of 100.

4.2. IBD analysis

4.2.1. Sharing within and between populations

We used the AJ-Flemish merged and phased dataset (57 AJ, 26 Flemish; section 2.14), ran *Germline* with a minimal segment length of $m = 3$cM, and filtered the results as explained in section 4.1. For each pair of individuals, we computed the total shared genetic map distance (cM) and the fraction of genome shared (using a total autosomal map distance of 3546cM [39] and for sharing between any of the two haplotypes of each individual). In Figure 2 of the main text, we plot the distribution of the fraction of the genome shared, broken by population: within AJ, within Flemish, or between AJ and Flemish. The

average fraction of the genome shared was 1.85% within AJ, 0.23% within Flemish, and 0.10% between AJ and Flemish, a trend consistent with previous findings [1, 34, 42]. All AJ pairs shared at least one segment, compared to 86.1% for Flemish and 63.1% for AJ-Flemish pairs. We also repeated the analysis with a minimal segment length of $m = 5$cM, and the results came out qualitatively the same, but with an even larger ratio between the AJ and the non-AJ sharing (mean sharing 0.84% in AJ, compared to 0.03% in Flemish and 0.005% between AJ-Flemish; fraction not sharing 1.6% in AJ, compared to 87.7% in Flemish and 97.2% in AJ-Flemish). These results suggest that within-Flemish and AJ-Flemish sharing is mostly detection noise, and may vanish completely with better IBD detection methods (see also sections 4.2.2 and 4.2.3).

The amount of IBD sharing we detected is close to what is expected based on our runs-of-homozygosity (roh) analysis. Searching for roh in the complete project data, excluding the individual with the consanguineous parents (GS000010967-ASM; 127 remaining), and using the same approach as in section 2.11.1 and with a minimal length of 4MB (corresponding roughly to 5cM) yields a total of 935MB shared. Per autosomal genome ($2.881 \cdot 10^9$bp) and per 127 haplotype pairs, this is ≈0.26% of the genome. Since there are four haplotype pairs between any pair of diploid individuals, for IBD with $m = 5$cM the fraction of the genome shared per haplotype pair is ≈0.21%, close to the roh fraction. In the rest of the section, unless otherwise mentioned, we use $m = 3$cM.

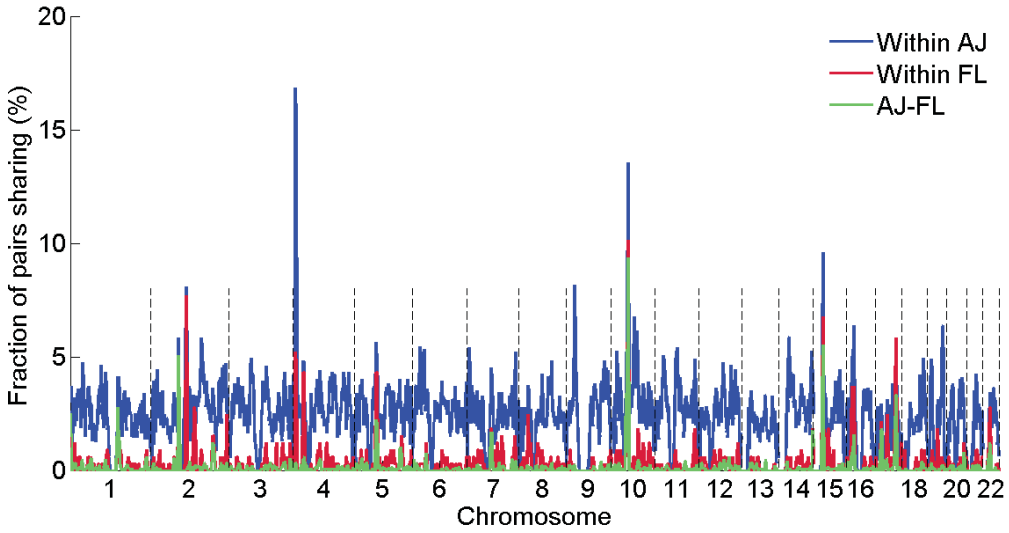
4.2.2. Segment quality scores

To evaluate a possible quality difference between the segments detected in each population, we studied the distribution of the segment scores (D in Eq. (11)) for segments shared either within AJ, within Flemish, or between AJ and Flemish. However, as the segment score increases roughly linearly with the segment length ($r = 0.69$), comparing the raw scores between populations would reflect the differences in overall IBD sharing. We therefore show, in Supplementary Figure 4, the distribution of segment scores per cM. The scores of the intra-AJ segments have the highest quality (mean score per cM 148.8), followed by the intra-Flemish segments (118.7) and the AJ-Flemish segments (92.0). The P-value for the difference between the distribution of the AJ segment scores and the Flemish scores is $1.2 \cdot 10^{-38}$ (rank-sum test); for the difference between AJ scores and AJ-Flemish scores the P-value is $1.2 \cdot 10^{-253}$. These results suggest again that the non-AJ segments are of lower quality and are largely due to noise.

4.2.3. Number of pairs sharing at each locus

In Supplementary Note Figure 11, we plot the "sharing intensity" along the genome. We divided each chromosome into bins of 1MB each and used BEDTools [62] to determine the fraction of pairs sharing at each bin, broken by population. Sharing within-AJ is consistently higher than non-AJ sharing along the entire genome, as expected. The majority of sharing within-Flemish and between AJ and Flemish is concentrated in a handful of peaks; sharing in the rest of the genome is sporadic. In the absence of a gold standard for IBD detection, it is hard to evaluate the importance of those peaks; we note, however, that except for the AJ-specific peaks on chrs 9 and 19, all other peaks disappear when using $m = 5$cM (not shown). This observation further supports our identification of most of the non-AJ shared segments as noise.

We also note that there was no enrichment of sharing in the HLA region (≈chr6:25-35MB), as previously observed in both Jewish and non-Jewish populations (e.g., [42, 63]), even when directly examining the unfiltered *Germline* output and even when using more liberal detection parameters. However, increasing dramatically the window size (up to 2000 or even 5000 SNPs) did show ≈5-fold enrichment in the HLA region. The HLA region has about double the number of SNPs (in the merged AJ-Flemish dataset) and about half the recombination rate compared to the rest of the chromosome. Therefore, even very short IBD segments in this region can potentially comprise a large number of SNPs and hence be detected even with very strict parameters. While this would suggest that the previously observed HLA enrichment was an artifact, improved detection methods and additional datasets will be required to reach a definite conclusion.



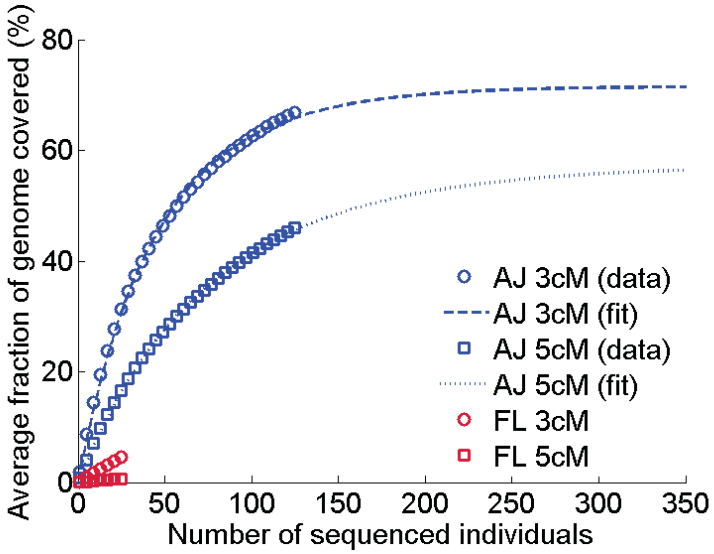
Supplementary Note Figure 11. *IBD sharing along the genome.* The fraction of pairs sharing (at least one haplotype) at each 1MB bin along the genome is plotted for segments shared within AJ (blue), within Flemish (red), or between AJ and Flemish (green). Sharing within AJ is consistently higher than in the other groups; most of the within-Flemish and AJ-Flemish sharing is concentrated in a small number of peaks.

4.2.4. Coverage of the genome by IBD segments

A potential application of our AJ sequencing dataset is imputation of partly genotyped samples (e.g., SNP arrays). In section 5, we study the performance of an "off-the-shelf" imputation algorithm using an AJ reference panel. Most standard methods, however, do not take into account the long-range information offered by the presence of IBD shared segments. Here, we quantify the potential benefit of using such information. In the AJ population, where IBD is abundant , one can impute a partly genotyped segment, shared IBD with a fully sequenced individual, either by simply copying the shared segment (if phase data is available) or by more sophisticated approaches [36, 64, 65, 66, 67, 68, 69]. As shared segments differ, on average, by only ≈1-2 recent mutations [36], this approach is expected to be highly accurate. The key question is therefore, what is the fraction of the genome, in an average AJ individual, that is "covered" by such shared segments.

In this analysis, we considered sharing within-AJ and within-Europeans separately. For AJ, we used the complete dataset of 128 genomes (which was processed exactly as the AJ-Flemish dataset, section 4.1). For Europeans, we used either the Flemish IBD data (section 4.1) or data from the larger CEU cohort in the 1000 Genomes project. Phased haplotypes for CEU were downloaded from http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_interim.html. Non-CEU haplotypes were removed, as well as sites that became monomorphic reference. The haplotypes were then converted to Plink format (including adding genetic map distances) and processed using the pipeline detailed in section 4.1.

To quantify the rate of coverage gain with increasing panel size, we used subsets of sequenced individuals of size n and varied n between 2 and 128 for AJ (2 to 26 for Flemish, 2 to 87 for CEU). For each n and for each individual in the subset, we calculated the fraction of the (autosomal; using physical distance) genome found in IBD segments shared with others in the subset. Note that this guarantees sharing of at least one of the haplotypes but not necessarily both; we did not attempt to resolve the shared haplotypes or determine whether both are covered. We then averaged the coverage over all individuals in the subset and over 50 random orderings of the individuals. The results are shown in Figure 2 of the main text for CEU and in Supplementary Note Figure 12 here for Flemish (for both $m = 3$cM and $m = 5$cM), demonstrating that as expected, the coverage in AJ is much higher than in Europeans.



Supplementary Note Figure 12. *A comparison of coverage by IBD in AJ and Flemish.* This figure is exactly as Figure 2B in the main text, except that the European comparison cohort is the Flemish.

To predict the coverage for sample sizes larger than ours, we considered a toy model where a population is assumed to undergo a bottleneck G generations ago, lasting a single generation. The population size is assumed to be extremely large otherwise. Our mathematical analysis (not shown) demonstrates that under these assumptions, a given study haplotype is shared with at least one haplotype from the panel with probability $\approx c_{\max}\left(1 - e^{-n/n_0}\right)$ (see also [36]). The prefactor $c_{\max}$, the upper bound on the probability of haplotype coverage, is a product of two terms. The first is $(1 +$

$mG/100)e^{-mG/100}$, the probability that the given haplotype cannot be shared due to excessive recombination along its lineage. For a bottleneck $G = 28$ generations ago (the value inferred for AJ in section 4.3) and $m = 3$cM, this equals to ≈0.8 (≈0.6 for 5cM). The second term is the fraction of the population not admixed, because admixed haplotypes are assumed to have a recent co-ancestry different from those in the panel and therefore not to be shared. Next, $\left[1 - c_{\max}\left(1 - e^{-n/n_0}\right)\right]^2$ is the probability that both haplotypes of the given individual are *not* covered. Therefore, the average coverage of a given individual's diploid genotype by IBD segments is

$$\langle c \rangle = 1 - \left[1 - c_{\max}\left(1 - e^{-n/n_0}\right)\right]^2. \quad (12)$$

We obtained the best-fit parameters using *Matlab*'s *nlinfit* (for AJ only). The asymptotic coverage, $c_{\max}$, came out as 46.7% and n_0 was 56.4 (34.7% and 88.7, respectively, for $m = 5$cM). Even when taking into account the theoretical limit on sharing due to recombination, the fraction of admixed ancestry was surprisingly high at about ≈41%. This is nevertheless of the same order of magnitude as our admixture fraction estimate using the allele frequency spectrum (Supplementary Table 7). On the other hand, the admixture event was dated either just at the bottleneck or much earlier (depending on the precise model inferred; Supplementary Table 7 and Supplementary Note Table 3) and additionally, some of our unpublished results suggest that using a much larger sample of over 2600 AJ arrays (from [35, 70]), the asymptotic coverage is close to 100%. The coverage results from the arrays were consistent with the sequencing results for small cohorts, but deviated from Eq. (12) for larger cohorts. Note also that the true fraction covered (in this study) is potentially somewhat higher: for ≈8% of the genome, there was not even a single segment shared, usually due to sequence gaps (centromeres, etc.), and additionally, the asymptotic coverage, $c_{\max}$, increased up to ≈55% when using minimal segment lengths smaller than 3cM (not shown). Our estimate of ≈70% diploid coverage should therefore be considered as conservative in the context of the potential imputation power.

4.3. Demographic inference

4.3.1. Method

IBD sharing is due to strong genetic drift in the recent history of the population. Therefore, it can be used to infer demographic parameters of the recent history. We used the inference method of Palamara et al. (2012) [43], which works by matching the decay of the amount of IBD sharing vs. the segment length to the theoretical expectation for a given demographic model. To create the decay curve, we first recorded the lengths of segments shared between all pairs of individuals (AJ only). We then divided the space of lengths ($m = 3$cM to an arbitrary cutoff of 15cM) into 11 intervals, such that the length of each interval is a constant factor times the length of the previous interval. For each interval, we summed the total length (in cM) of segments shared having length in the interval and divided by the total genome size (3546cM [39]) and by the total number of (haplotype) pairs. The resulting curve, which we denote $p_{\text{real}}(\ell)$, is shown in Figure 3 of the main text. We then searched for a demographic model whose decay curve, $p_{\text{model}}(\ell)$, is closest to the empirical curve.

We examined a model of a recent bottleneck followed by an exponential expansion (Supplementary Note Figure 13). The population size is N_0 (diploids) until T_b generations ago, when it is reduced to N_b.

From T_b generations ago until the present, the population size increases at rate r percent per generation (the final population size, N_f, is therefore $N_f = N_b(1 + r/100)^{T_b}$). The theoretical $p_{\text{model}}(\ell)$ was calculated using Eq. (6) in Palamara et al. (2012) [43]. The (haploid) population size at generation g in the past, needed in that equation, was set to $N(g > T_b) = 2N_0$ and $N(g \le T_b) = 2N_b(1 + r/100)^{T_b - g}$. The approximate summation in Appendix A of Palamara et al. (2012) was replaced by the exact solution

$$\sum_{g=G+1}^{\infty}(1-1/N_0)^{g-G-1}\int\left(\frac{g}{50}\right)^2 xe^{-gx/50}dx = \frac{N_0 e^{-Gx/50}\left[(N_0-1)(50+Gx)-N_0 e^{x/50}(50+x+Gx)\right]}{50\left[1+N_0\left(e^{x/50}-1\right)\right]^2}.$$

To identify the model with the best fit to the empirical $p_{\text{real}}(\ell)$, we used a simple grid search centered approximately around the bottleneck parameters inferred by Palamara et al. (2012). The parameter N_0 was varied between 500 and 20,000 (increments of 500), N_b between 50 and 1000 (increments of 50), r between 1% to 60% (increments of 0.5% or 1%), and T_b between 10 and 60 (increments of 1). For each configuration, we computed $p_{\text{model}}(\ell)$ analytically, as explained above, and then the sum-of-squared-error (SSE) between the model and the real curves,

$$\text{SSE} = \sum_{i\in\text{intervals}}\left[\log\left(p_{\text{real}}(\ell_i)\right) - \log\left(p_{\text{model}}(\ell_i)\right)\right]^2. \quad (13)$$

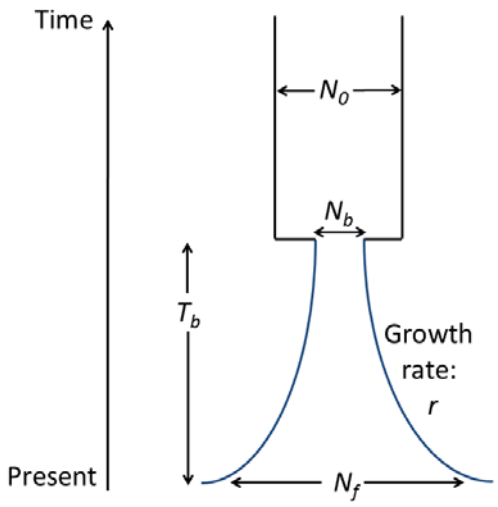
The inferred demographic model had the minimal SSE.

For comparison with a constant-size population (Wright-Fisher model), we inferred the best fitting (haploid) population size as [36, 43]

$$\widehat{N} = \frac{100}{m\langle f\rangle} - \frac{75}{m}, \quad (14)$$

Where $\langle f\rangle$ is the average (haploid) fraction of the genome shared IBD. Using $\widehat{N}$ (haploids), and for segment lengths in the range $[\ell_1, \ell_2]$, the fraction of the genome shared is given by [43]

$$p(\ell_1,\ell_2) = \frac{100\widehat{N}^2(\ell_2-\ell_1)[25(\ell_1+\ell_2)+\ell_1\ell_2\widehat{N}]}{(50+\ell_1\widehat{N})^2(50+\ell_2\widehat{N})^2}. \quad (15)$$



Supplementary Note Figure 13. *A diagram of the demographic model inferred using IBD information.*

4.3.2. Results

Inferring the bottleneck and growth parameters as explained in section 4.3.1 gave the following best fit model: $N_0 = 4500$ (diploids), $N_b = 300$ (diploids), $T_b = 27$ (generations), and $r = 41\%$ (growth rate per generation; corresponding to $N_f = 3.2 \cdot 10^6$). The decay curve, $p_{\text{model}}(\ell)$, is plotted in Figure 3 of the main text, showing good agreement with $p_{\text{real}}(\ell)$. Using $m = 5$cM gave similar results ($N_0 = 2500$, $N_b = 400$, $T_b = 25$, and $r = 45\%$ ($N_f = 4.3 \cdot 10^6$)). The curve corresponding to a constant-size population (with the same overall total sharing; $\widehat{N} \approx 3620$ diploids) shows, as expected, less sharing than the real data for short segments (since the bottleneck is not modeled) and more sharing for long segments (since the recent explosive expansion is not modeled).

Palamara et al. (2012) [43], who developed the IBD-based inference method , used a bottleneck/expansion model identical to ours (Supplementary Note Figure 13) to infer the history of AJ using SNP arrays for a different set of AJ samples. They demonstrated that the bottleneck/expansion model fitted well for segments of length >2cM, but for shorter segments, a double-bottleneck/expansion model fitted better. Our inferred parameters are in general compatible with those of Palamara et al. (2012) [43], except for N_0, which is here ≈10-fold smaller than in their single bottleneck/expansion model (but is in agreement with their ancestral population size for the double-bottleneck/expansion model). The current population size that we inferred, ≈3 million, is reasonable given the current census size, ≈10 million.

When fitting the demographic model, we assumed that the population is isolated; that is, each pair of lineages must coalesce within the population. However, our results in section 4.2.4 show that the AJ population may be admixed, with up to ≈40% of ancestry not traceable to the bottleneck (see also section 6.3.5). This implies that the actual bottleneck has likely been even more severe (or prolonged) [36]. More accurate inference of the bottleneck parameters will require identifying the admixed segments, perhaps through local ancestry inference.

4.3.3. Sum of Squared Errors (SSE) plots

In section 4.3.2 we provided a point estimate of the demographic parameters. Before turning to confidence intervals, we examined the SSE surface around the best-fit parameters. To facilitate visualization, we plotted (Supplementary Figure 11) the one-dimensional projection of the SSE surface (Eq. (13)) when varying each parameter at a time, fixing all others at their optimal value. Note that the same y-axis scaling was used in all panels. The parameters with the narrowest SSE minima were the bottleneck parameters: the time T_b and the size N_b. The SSE surface is flatter around the ancestral population size N_0 and the growth rate r. The lower precision for the ancestral population size is expected, as IBD is less informative on the ancient history. The growth rate (or the current population size) is also less precise, with growth rates between ≈35% and ≈45% differing little in their SSE (corresponding to a current population size between ≈1 and ≈10 million).

4.3.4. Jackknife resampling

To estimate the variance of the inferred demographic parameters, we used a simple delete-n jackknife resampling. In each iteration, we randomly selected a subset of 50 AJ individuals (out of overall 57), and repeated the demographic inference exactly as in section 4.3.1. For each parameter θ, the 95% confidence interval was computed, assuming normal distribution of the estimated parameter $\hat{\theta}$, as $\left[\langle\hat{\theta}\rangle - 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right), \langle\hat{\theta}\rangle + 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right)\right]$, where $\langle\hat{\theta}\rangle$ and $\mathrm{SD}\left(\hat{\theta}\right)$ are the mean and standard deviation of $\hat{\theta}$, respectively, over 100 iterations [4]. The results (for $m = 3\mathrm{cM}$) are presented in Supplementary Table 6.

The final mean values of N_b, N_f, and T_b given in Supplementary Table 6 were later used for the inference of the more ancient history, as explained in sections 6.2.2.4 and 6.2.3.2. The results for $m = 5\mathrm{cM}$ were qualitatively similar, with, as expected, a slightly larger variance (not shown).

4.3.5. Parametric bootstrap

To obtain an alternative estimate of the confidence intervals for the demographic parameters, we carried out parametric bootstrap resampling, as in Gutenkunst et al. (2009) [4]. Specifically, we generated simulated populations having a demographic history equal to that we inferred for the real data (section 4.3.4). Then, the simulated IBD decay curves were fitted to the demographic model and the model parameters were inferred. The confidence intervals were then computed based on the distribution of inferred parameters.

To generate simulated data, we used *MaCS* (version 0.4f) [71]. In each run, we generated 114 artificial genomes (corresponding to the 57 AJ individuals), each of which consisted of 22 chromosomes with lengths equal to the corresponding hg19 autosomal chromosomes. We set the mutation rate to $2.35 \cdot 10^{-8}$ per bp per generation (see discussion in section 6.4.16.4.3; however, here this should have only a minor effect) and the recombination rate to $1.23 \cdot 10^{-8}$ per bp per generation (corresponding to the total hg19 genetic map distance [39]). The "recombination history" parameter was set to 100 bp, and all other parameters were left at their default values. We simulated the demographic model of Supplementary Note Figure 13 with the parameters given in section 4.3.4 or Supplementary Table 6 ($N_0 = 4755$, $N_b = 334$, $r = 34\%$, and $T_b = 28$).

To detect IBD segments, we first converted each synthetic dataset into the *Plink* format (maintaining haplotypes phase). We then carried out IBD detection and filtering exactly as for the real data, as described in section 4.1 (except, of course, that no segments were removed due to overlap with gaps or a too short physical length; we did not introduce sequencing or phasing errors). Inference of the demographic parameters was carried out as in section 4.3.1. For each parameter θ, the biased-corrected 95% confidence interval was computed as $\left[\theta^* - \left(\langle\hat{\theta}\rangle - \theta^*\right) - 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right), \theta^* - \left(\langle\hat{\theta}\rangle - \theta^*\right) + 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right)\right]$, where $\langle\hat{\theta}\rangle$ and $\mathrm{SD}\left(\hat{\theta}\right)$ are the mean and standard deviation of $\hat{\theta}$, respectively, over 100 simulated datasets, and θ^* is the value inferred from the real data [4]. The final results (for $m = 3\mathrm{cM}$) were:

$$(16) \qquad N_0 = 6090 \pm 398\ [5310{,}6870],\ N_b = 246 \pm 75\ [99{,}392],\ r = 44\% \pm 14\%\ [17\%, 72\%]$$

(Or, for the mean rate, $N_f = 4.464 \cdot 10^6$), $T_b = 27 \pm 2\ [22{,}31]$,

where for each parameter, the mean and SD are given, followed by the biased-corrected 95% confidence interval. While the results of Eq. (16) indicate a bias of up to ≈30% compared to Supplementary Table 6 (although remarkably, not for the bottleneck time T_b), our approach confirms the existence of a very recent and very narrow bottleneck in the AJ population. The results for $m = 5\mathrm{cM}$ were qualitatively similar, with, as expected, a slightly larger variance (not shown).

5. Supplementary Note 5: Utility of the AJ genomes as an imputation reference panel

5.1. Study design

We set out to determine whether we could gain accuracy in imputation of AJ array genotypes by using AJ sequences as a reference panel instead of the 1000 Genomes Project CEU panel (Northern and Western European ancestry) [22, 23, 72, 73]. We used the 57 AJ genomes of first batch, processed according to the pipeline described in section 2.7.1. In the absence of AJ sequences from an independent source, we used 50 AJ sequences as our reference panel and the remaining seven as our study panel.

In a typical imputation pipeline, the array genotypes to be imputed consist of ≈0.5-1M common SNPs and are unphased. To simulate a realistic study panel using our seven AJ sequences, we masked all genotypes but a small subset (typical of a modern commercial array; see below) and discarded any phasing information. Since our study panel was too small for effective phasing and imputation, we supplemented our seven AJ (reduced) sequences with SNP arrays for 1000 additional AJ individuals. We then jointly phased and imputed the combined study panel of 1007 individuals. By uncovering, after imputation, the true genotypes of the seven sequences, we could compare the relative accuracies of using the AJ panel vs. using the CEU panel. A schematic of our study design is shown in Supplementary Figure 5. Details are provided in the following sections.

5.2. Preparation of the datasets

5.2.1. The 1000 Genomes Project CEU panel

The 1000 Genomes dataset was downloaded from http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1interim_jun2011_impute.tgz.
We extracted the 87 CEU individuals and removed all sites that were monormorphic reference. Note that the CEU panel was larger than the AJ panel (see also section 5.3). We ran all of our imputation experiments on chr1 only. The total number of CEU variants (on chr1) was 880,219.

5.2.2. SNP arrays for 1000 AJ individuals

We started with SNP arrays for 2610 AJ, 938 of which were schizophrenia cases, genotyped on Illumina HumanOmni1-Quad arrays in the Long Island Jewish Medical Center (LIJMC) and previously reported in Guha et al. (2012) and Lencz et al. (2013) [35, 70]. After removing all cases, we removed individuals with cryptic relatedness (Plink's $\hat{\pi} > 0.15$) or high inbreeding coefficient (Plink's $F > 0.05$), SNPs and individuals with missingness rate >1%, and SNPs not in Hardy-Weinberg Equilibrium ($P < 0.01$). We then used PCA to compare our genotypes to those of HapMap Europeans [39] and removed all individuals with full or partial non-AJ ancestry. We finally retained the 1000 individuals with the lowest missingness rate, genotyped on 726,252 SNPs each. To match the AJ sequences, we first performed an hg18=>hg19 lift-over and then strand flipping of array alleles that were given in the negative strand. To determine the strand, we usually used the hg19 reference allele. For A/T and C/G polymorphisms, we determined the strand by comparing the allele frequencies to the 57 AJ sequences (all of which are known to be in the positive strand), except for any SNPs whose minor allele frequency was in the range [0.35,0.5], which were discarded. We finally considered chr1 only (60,476 SNPs).

5.2.3. Splitting the AJ sequences between a reference panel and a study panel

We randomly selected seven of the 57 AJ samples to become our test (or study) individuals; the remaining 50 would serve as a reference panel for imputation. The sequences of the reference panel were previously phased using molecular phasing information, as explained in section 2.14.1. For the seven individuals serving as our study panel, we used the original un-phased data, as they would later be re-phased, once merged with the 1000 AJ arrays. Only chr1 was used for the imputation experiments, with genotypes available for 757,752 variants.

5.2.4. Merging the study panel genotypes and phasing

To reduce the seven AJ sequences in the study panel to array genotypes, we removed all variants not in the arrays. Then, we removed from the arrays all SNPs not in the sequences and merged the two datasets (1000 LIJMC arrays + 7 reduced sequences). The total number of remaining SNPs, on chr1, was 57,036. As recently recommended [74], we "pre-phased" our merged genotypes prior to imputation, using *SHAPEIT* (version 2) with default parameters and the 1000 Genomes genetic map.

5.3. Imputing

We imputed our merged and pre-phased study panel using *IMPUTE2* (version 2.3.0) [75] with default parameters (except for the -allow_large_regions flag). Imputation was carried out in ≈5MB blocks, chosen to have an approximately similar number of SNPs. We used the two reference panels in two

separate imputation experiments; one of the panels comprised the 50 AJ sequences and the other the 87 CEU individuals. While using a larger CEU panel would supposedly give it an unfair advantage, our results (section 5.4) nevertheless show that the AJ panel was superior. We also performed an imputation experiment using the two reference panels together, a new feature in *IMPUTE2* (using the -merge-ref-panels flag). The results from this experiment were not reported in the main analysis due to potential technical problems (see below).

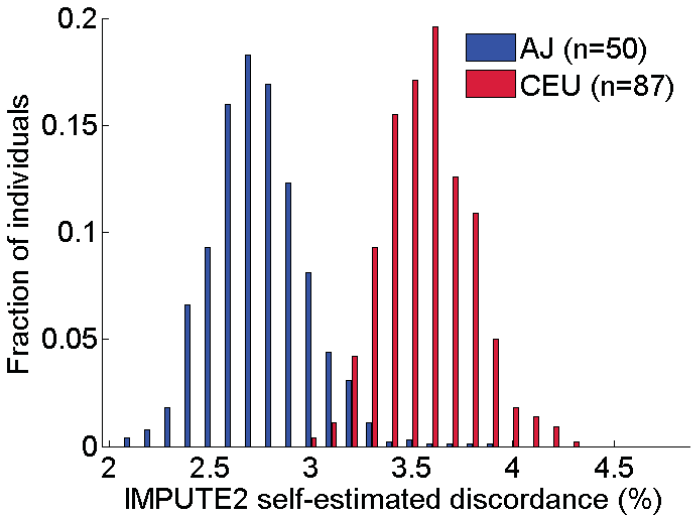
5.4. Analysis of the imputation accuracy

5.4.1. Genome-wide discordance and r^2

After imputation completed, we uncovered the true genotypes of the seven AJ sequences (at sites in the original genotypes of the 57 AJ individuals) to evaluate the imputation accuracy. The true genotypes were compared to the most likely genotypes and dosages returned by *IMPUTE2*. For sites not imputed by the CEU panel, we set the imputed genotypes as homozygous reference. Sites imputed by the CEU panel that were not found in the AJ sequences were discarded. Our reasoning was that counting those sites as discordant would discriminate against the CEU panel, because those variants might have actually been present in the AJ sequences but removed during cleaning. Note that this is conservative: it reduces the discordance when using the CEU panel but not when using the AJ panel, whereas our goal is to demonstrate the higher accuracy of the AJ panel. We also discarded sites that were monomorphic non-reference in the AJ panel, again, conservatively, since some of those sites may have been missing from the CEU panel just because they were monomorphic world-wide (see section 2.8). That left ≈200,000 non-reference variants per individual. The average number of discordant genotypes (over the study individuals) as well as other statistics are presented, for each reference panel, in Supplementary Table 4, showing that the AJ panel achieves higher accuracy than the CEU panel. Supplementary Table 4 also shows r^2, which was computed between the aggregate of all true genotypes (over all sites and study individuals) and all imputed dosages.

In Supplementary Table 4, we break the discordant genotypes based on whether they were *false negatives* (non-reference alleles missed) or *false positives* (non-reference alleles wrongly suggested). The greatest gain using the AJ panel came from reducing the number of false negatives. This is expected, since AJ-specific variants cannot be imputed using a European panel. The number of false positives was only slightly smaller using the AJ panel. We also ran an imputation experiment using the combined AJ-CEU panel. However, while the number of false negatives was 13% lower than when using the AJ panel alone, the number of false positives was 31% higher, and the overall imputation accuracy was slightly lower. This is at odds with our expectation that combining panels should eliminate many false positives [75], and, we believe, might reflect a side-effect of *IMPUTE2*'s yet unpublished new approach (at the time of writing) for merging reference panels. In their implementation, each panel is imputed using the other panel prior to imputing the study genotypes. Therefore, the frequency of rare variants that are specific to one of the panels may be artificially inflated, which could lead to false positives. Note that a poor performance of the combination panel was also recently reported by Duan et al. (2013) [73]. One could, alternatively, merge the panels by setting as homozygous reference all sites missing in one panel and existing in the other (see also the discussion in section 2.13), which improved accuracy in preliminary results from another study [76]. However, we did not further pursue this direction here.

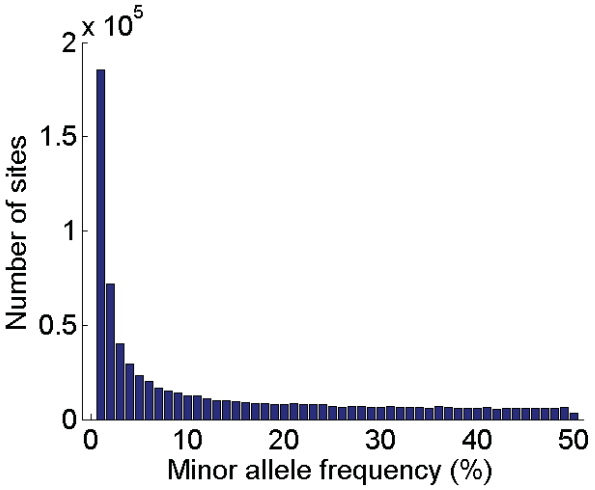
IMPUTE2 also provides, as a diagnostic tool, an estimated discordance between the real and imputed genotypes at the known array SNPs, based on a leave-one-out approach. The average estimated discordance over the 1000 array genotypes is presented in Supplementary Table 4. The distribution of the discordances over the 1000 arrays is plotted in Supplementary Note Figure 14.



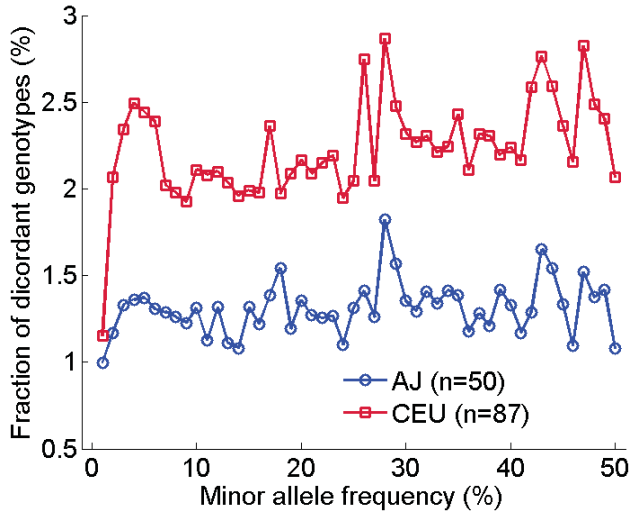
Supplementary Note Figure 14. *IMPUTE2*'s *self-estimated discordance.* The figure compares the distribution of the discordances (over 1000 arrays genotypes) when using different reference panels: AJ or CEU. The distributions are distinct with P-value $< \mathbf{10^{-100}}$ (signed rank test).

5.4.2. The results stratified by minor allele frequency

It is of interest to evaluate the imputation accuracy at different minor allele frequencies [77]. In Figure 2 of the main text and Supplementary Figure 6, we plot r^2 and the number of discordant genotypes, respectively, vs. the minor allele frequency for the two reference panels. In Supplementary Note Figure 15, we plot the number of variants in each frequency bin. The minor allele frequency was determined using the 50 AJ individuals in the AJ panel. According to this definition, variants with frequency 0% are monomorphic reference in the AJ panel and are thus necessarily wrongly imputed whenever a study individual has a non-reference allele. We therefore excluded them from the plots, although not from the numbers reported in Supplementary Table 4. Figure 2 of the main text and Supplementary Figure 6 agree with the results of section 5.4.1, showing that accuracy is improved by using the AJ panel instead of the CEU panel. The number of discordant genotypes is higher, using both panels, for low frequency variants (Supplementary Figure 6). However, this mostly reflects the total number of variants at each frequency bin (Supplementary Note Figure 15 and Supplementary Note Figure 16).



Supplementary Note Figure 15. *The total number of variants at each minor allele frequency*. Minor allele frequencies were computed using the 50 AJ sequences that served as the reference panel. The figure demonstrates that the higher number of discordant genotypes at low frequencies (Supplementary Figure 6) is largely due to a larger number of overall variants.



Supplementary Note Figure 16. The *fraction of discordant genotypes vs. the minor allele frequency.* The average fraction (over the seven AJ study sequences) of genotypes that are discordant is plotted (using the original genotypes of the 57 individuals; excluding sites monomorphic non-reference at the AJ panel).

We finally note that it is easy to see several ways by which our study could be extended. First, we could expand our working dataset from the 57 genomes of the first batch to the full project data (128 sequences), and to the entire genome instead of just chr1. We could also use the entire 1000 Genomes Project European samples, or even the entire cosmopolitan panel, and test other approaches for merging panels. Finally, it would also be interesting to determine whether the AJ reference panel can improve imputation in populations other than AJ (say, in populations from the Near-East). Such extensions are left for future studies.

6. Supplementary Note 6: Demographic models based on the allele frequency spectrum

6.1. Reasons for increased heterozygosity

Our initial comparison of AJ and Flemish genomes showed higher heterozygosity in AJ (section 3.1), which we considered as somewhat counterintuitive, given the larger amount of IBD sharing in AJ (section 4.2). However, we notice that for an average pair of AJ individuals, long IBD segments cover only about ≈1-2% of their genomes (Figure 2 of the main text). While for the shared loci, coalescence times are extremely recent (≈20-30 generations), for the remaining loci, coalescence times are much more ancient and likely longer in AJ compared to Europeans. We consider three demographic processes that could have led to increased coalescence times (and hence genetic diversity).

1. A larger AJ population size in ancient history, either due to a larger ancestral population size or due to an additional bottleneck in Europeans. An additional bottleneck would be consistent with the (partly) Middle-Eastern origin of AJ and the "serial founder model" describing the human expansion out of Africa [2, 78, 79, 80, 81], which asserts that the farther the population is, by land, from East or South Africa, the more bottlenecks its founders underwent and the smaller its contemporary genetic diversity is. Additionally, AJ genomes were shown to have ≈3% West-African ancestry [82]. As heterozygosity is ≈1.35-fold larger in Africans than in AJ or Europeans (e.g., [28]), this could explain ≈40% of the increased heterozygosity in AJ (+2.4% compared to Flemish).
2. "Explosive", ≈1000-fold growth of AJ population size in the recent millennium, which might have not been paralleled in the rest of Europe [42, 43, 51, 57]. To test this hypothesis, we used *∂a∂i*, a demographic inference tool [4], to generate synthetic allele frequency spectra (and thereby the heterozygosity) for a number of representative models. Our results (not shown) suggest that explosive growth would lead to very little increase in heterozygosity, if at all. Intuitively, this is because even though recent growth introduces a very large number of rare variants, those are usually in a homozygous (ancestral) state and therefore contribute very little to the heterozygosity.
3. The AJ ancestry has been shown to be a mixture of Middle-Eastern and European ancestries. Recent European admixture might have introduced many new variants into AJ and thereby increased the AJ heterozygosity [34]. Our numerical results using *∂a∂i* showed that admixture is a plausible cause of increased heterozygosity.

The goal of this subsection was to develop an intuitive understanding of the increased heterozygosity in AJ. In the following subsections, we infer specific demographic models using the complete frequency spectrum.

6.2. Specification of demographic models

6.2.1. General

To infer demographic models from the allele frequency spectrum, we use *∂a∂i*, which determines the most likely demographic parameters by fitting the observed allele frequency spectrum (AFS) to the theoretical AFS, computed using the diffusion approach. Some general settings of the demographic models are listed below. Specific models are described in the following subsections.

To compute the theoretical AFS, we set the mutation rate to $1.44 \cdot 10^{-8}$ per bp per generation, following Gravel et al. (2013) [58] and based on the time of the human settlement in the Americas (see discussion in section 6.4.3). The total genome length was set to $2.685 \cdot 10^{9}$ bp (the autosomal hg19 less the number of N's), further multiplied by 0.81, which is the fraction of variants that remained after our quality-control filters (≈2.6M out of raw 3.2M variants; see section 2.7.1 and Supplementary Table 3). In total, the mutation rate per autosomal genome was $\mu_0 = 1.44 \cdot 10^{-8} \times 0.81 \times 2.685 \cdot 10^{9} = 31.3$. Some of our initial analysis was carried out using a higher mutation rate of $2.35 \cdot 10^{-8}$, following [4] (based on the human-chimp divergence time) as well as the full genome length $2.881 \cdot 10^{9}$, yielding $\mu_0 = 67.7$. As we elaborate in section 6.4.3, this change has the effect of rescaling all parameters (except admixture fractions). The results of any analyses that used the higher mutation rate were then converted to their appropriate value under the lower mutation rate.

In all demographic models, the ancestral population size was assumed to be N_0 diploids up to the point when specific population dynamics begins (see sections 6.2.2 and 6.2.3). The scaled mutation rate is then defined as $\theta = 4N_0\mu_0$. As explained in Gutenkunst et al. (2009) [4], since the AFS is proportional to θ, then for *any* demographic model, θ can be estimated directly from the total number of segregating sites in the real data. That is, each theoretical spectrum was generated using $\theta = 1$, and then θ was computed by dividing the total number of segregating sites in the real data by the number of sites in the theoretical spectrum. Using θ and μ_0, we computed the ancestral population size as $N_0 = \theta/(4\mu_0)$.

In the rest of the section, the units used are as follows. Population sizes (N) are given in number of diploid individuals. Times (T) are usually given in number of generations. When given in years, 25 years per generation were assumed. Migration rates (m) are given, for each population, as the fraction of individuals migrating from the other population in each generation. Admixture fractions (f) are given, for each population, as the fraction of individuals originating from the other population at the admixture event.

6.2.2. Single population models

We used a number of single population models to fit the history of the AJ and Flemish populations.

6.2.2.1 Wright-Fisher model (constant-size)

The simplest demographic model is the Wright-Fisher (constant-size): the population size is N_0 at all times. The AFS is given by Eq. (4) in section 3.5.3 [53]. Inference of the population size N_0 for the Wright-Fisher model was described also in section 3.5.5. Here (section 6.3.3), we infer N_0 using the total number of segregating sites (as explained in section 6.2.1) and using the full dataset (without down-sampling).

6.2.2.2 Growth-only

The model has three parameters: N_0, T_g, and N_f. The population size is N_0 until T_g generations ago, when exponential growth begins. The growth rate is such that the current (final) population size is N_f.

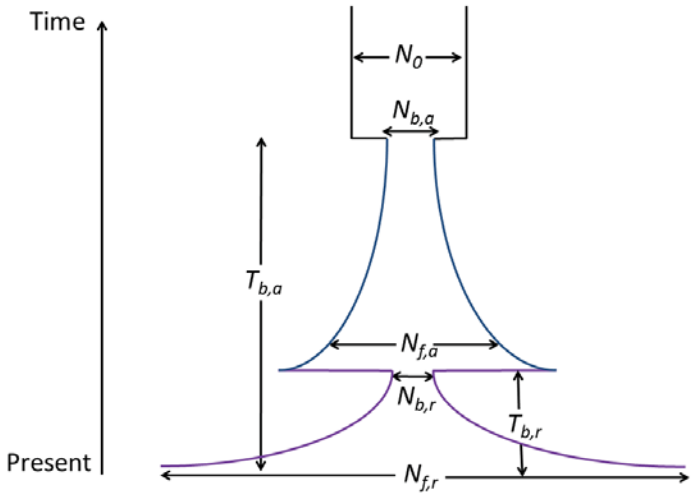
6.2.2.3 Bottleneck and growth

The model has four parameters: N_0, N_b, T_b, and N_f. The population size is N_0 until T_b generations ago, when it is instantaneously reduced to N_b. The population then begins an exponential expansion until

reaching a final size N_f. The model is schematically depicted in Supplementary Note Figure 13. When the inferred bottleneck is ancient, say, of the order of 2000-4000 generations ago (≈50-100 kya), this likely corresponds to the Out-of-Africa (OoA) founder event (see more in section 6.4.2).

6.2.2.4 Ancient bottleneck/growth with additional, known, recent bottleneck/growth

The model has two bottleneck and growth episodes: ancient and recent. There are six parameters beyond N_0: $N_{b,a}$, $T_{b,a}$, and $N_{f,a}$, the ancient bottleneck/growth parameters, and $N_{b,r}$, $T_{b,r}$, and $N_{f,r}$, the recent bottleneck/growth parameters (Supplementary Note Figure 17).



Supplementary Note Figure 17. *A diagram of a demographic model with an ancient and then a recent bottleneck and growth episodes.*

When using this model to infer the history of the Ashkenazi Jewish population, we assumed that the recent bottleneck/growth episode describes the founder effect and explosive growth of the AJ population in the past millennium. For such recent events, the allele frequency spectrum has little power, in particular with our relatively small sample size [43]. Indeed, our attempts to fit the full seven-parameter model did not converge to sensible results (not shown). However, the parameters of the recent bottleneck and growth were successfully inferred using information on IBD sharing (section 4.3). Therefore, we fixed the recent demographic parameters, $N_{b,r}$, $T_{b,r}$, and $N_{f,r}$, to their corresponding values obtained via the IBD analysis (section 4.3.4, Supplementary Table 6): $N_{b,r} = 334$, $T_{b,r} = 28$, and $N_{f,r} = 1.450 \cdot 10^6$. [Note that the value of N_0 inferred using IBD was not used, because the simple ancestral constant size history assumed in section 4.3 is replaced here by a more elaborate bottleneck/growth model]. We then used the allele frequency spectrum and *∂a∂i* to infer the values of the remaining four ancient-history parameters (N_0, $N_{b,a}$, $T_{b,a}$, and $N_{f,a}$).

6.2.3. Two population models

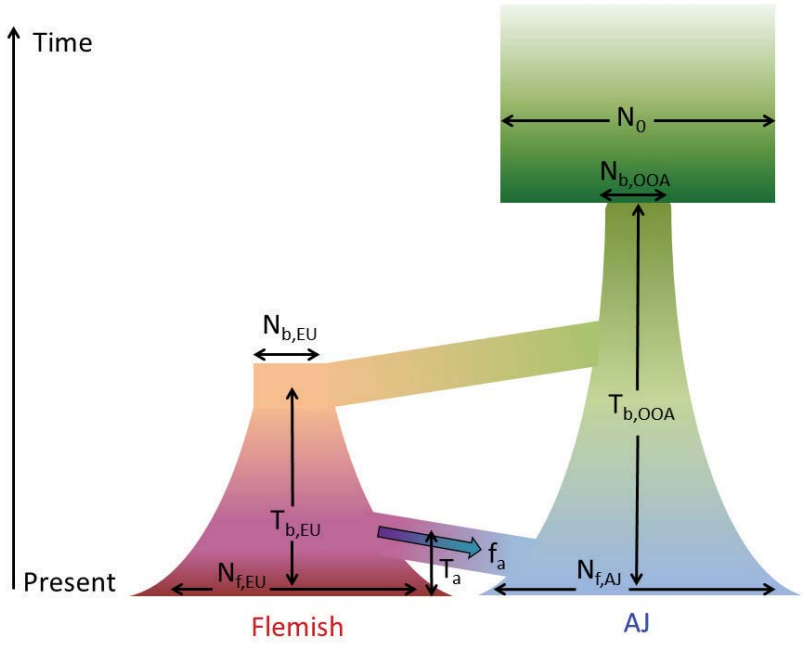
While some of the single-population demographic models fitted quite well to the observed AFS (section 6.3.5), those inferred parameters are “projections” of the real history, had the populations been

evolving in isolation. In reality, gene flow between population has likely played a major role in human evolution in general (e.g., [83]) and in Ashkenazi Jews in particular [34, 36]. We therefore attempted to fit a two-population model to the AJ-Flemish joint spectrum.

6.2.3.1 The basic model

In our model, both AJ and Flemish populations originated from a single ancestral population of size N_0 (presumably living in Africa). At $T_{b,OOA}$ generations ago, the ancestral population underwent a bottleneck and its size was reduced to $N_{b,OOA}$ (corresponding to the Out-of-Africa event and the colonization of the Middle East). It then began growing exponentially at a rate that would bring its final size to the AJ current population size, $N_{f,AJ}$. At $T_{b,EU}$ generations ago, a second population of size $N_{b,EU}$ split from the ancestral one, corresponding to the founder event at the population of Europe. The second population then began growing exponentially at a rate that would bring its final size to the current European population size, $N_{f,EU}$. At T_a generations ago, a fraction f_a of the first population (AJ) was assumed to be replaced by migrants from the second (European), corresponding to a unidirectional admixture event of Europeans into AJ. The model is illustrated in Supplementary Note Figure 18. In summary, there are nine parameters: N_0, $N_{b,OOA}$, $T_{b,OOA}$, $N_{f,AJ}$, $N_{b,EU}$, $T_{b,EU}$, $N_{f,EU}$, T_a, and f_a.

Further interpretation and justification of the demographic model is discussed in section 6.4.4. Alternative models are discussed in section 6.4.7.

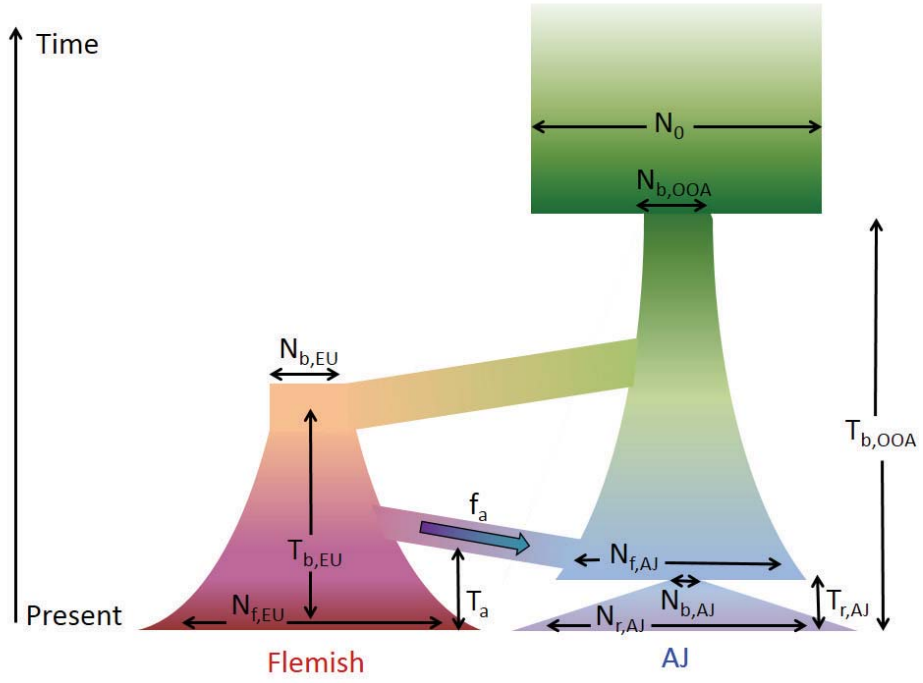


Supplementary Note Figure 18. *A diagram of our two-population demographic model.*

6.2.3.2 A model with recent, known, AJ bottleneck/growth

We also consider a variant of the model of Supplementary Note Figure 18, where, as in section 6.2.2.4, we assume that the AJ population underwent an additional bottleneck/growth episode. In this model (Supplementary Note Figure 19), $N_{f,AJ}$ is the AJ population size *before* the recent bottleneck, $N_{b,AJ}$ is the

population size at the bottleneck, $T_{r,AJ}$ is the time of the bottleneck, and $N_{r,AJ}$ is the current AJ population size. Due to the limited power of the frequency spectrum to resolve the parameters of recent events, here too we fix the recent bottleneck/growth parameters to the values inferred from the IBD analysis (section 4.3.4, Supplementary Table 6): $N_{b,AJ} = 334$, $T_{r,AJ} = 28$, and $N_{r,AJ} = 1.450 \cdot 10^6$.



Supplementary Note Figure 19. *A diagram of our two-population demographic model with a recent AJ bottleneck growth episode.* The model is exactly as in Supplementary Note Figure 18, except that the AJ population is assumed to undergo an additional bottleneck and expansion epoch (based on the IBD results, section 4.3).

6.3. Inference

6.3.1. Method

The observed allele frequency spectrum (AFS) was computed, as in section 3.5.2, using the merged and imputed genotypes and with respect to the minor allele (in the entire sample) and without down-sampling (57 AJ, 26 Flemish). High frequency population-specific alleles, which are likely mapping artifacts, were removed as explained in section 3.5.1. The effect of the removal on the final inferred demographic parameters was negligible (not shown). While we initially computed the derived allele frequency based on the genomes of chimpanzee or other primates [84], we suspected that a large number of alleles remained misclassified (a jump was observed in the AFS at frequencies close to 100%; not shown), similar to observations in previous studies [3, 56]. We therefore opted to use the unambiguous minor allele frequency.

Demographic inference was carried out using *∂a∂i* version 1.63 [4] and the *optimize_log* function. The maximal number of iterations was set to 50 and the fold change perturbation of the initial parameters to 2^1. The integration time-scale factor varied between 10^{-5} for models with a known recent expansion

(where according to the *∂a∂i* manual, a relatively short time-step is necessary to maintain accuracy) to 10^{-3} for the other models, which did not exhibit any sharp recent changes in population size. For the single-population models, spectra were computed by extrapolating over three grids with $(140,160,180)$ points for AJ (114 chromosomes) and $(70,85,100)$ points for Flemish (52 chromosomes). For the joint AJ-Flemish spectrum, we projected the original spectrum to $50x50$ chromosomes, and did not use extrapolation when integrating (a warning appeared that extrapolation was not accurate) but rather used a single grid of 400 points.

For each population and demographic model, we manually experimented with different parameter regions until we identified the most plausible one. We then set the initial parameters to some arbitrary values in that region and randomly perturbed them before launching the optimization procedure. The lower and upper bounds on the inferred parameters were set as the (unperturbed) initial parameters divided and multiplied by a number between 100 and 1000, respectively. In all cases, the final parameters were not close to the boundary and were not sensitive to the magnitude of the initial perturbation. The only exception was the Flemish population with the growth-only model (section 6.2.2.2), where the final population size was always as large as the boundary, with negligible change in the model likelihood. This, however, is in line with our observation that the model provides poor fit (section 6.3.5). For each model, we repeated the inference process 100 times (10 for the synthetic spectra used for the bootstrap, section 6.3.2) with 100 different initial values and reported (section 6.3.3) the configuration that yielded the maximal likelihood. We discarded runs that did not converge.

6.3.2. Parametric bootstrap

Parametric bootstrap sampling was carried out essentially as in section 4.3.5. Using *MaCS* (version 0.4f) [71], we generated artificial genomes consisting of 22 chromosomes with the lengths of the hg19 autosomal chromosomes. We generated $57 \times 2 = 114$ genomes for inference on AJ alone, $26 \times 2 = 52$ genomes for Flemish alone, and $25 \times 2 = 50$ of each population for the joint inference. We set the mutation rate to $1.44 \cdot 10^{-8}$ per bp per generation and the recombination rate to $1.23 \cdot 10^{-8}$ per bp per generation (corresponding to the total hg19 genetic map distance [39]). The "recombination history" parameter was set to 100 bp, and all other parameters were left at their default values. We did not introduce artificial sequencing errors. For each demographic model, we generated 100 synthetic allele frequency spectra, which we then folded and converted to the *∂a∂i*'s format. We then ran *∂a∂i* on each spectrum exactly as in the real data (sections 6.2.1 and 6.3.1). For each parameter θ, the biased-corrected 95% confidence intervals were computed as $\left[\theta^* - \left(\langle\hat{\theta}\rangle - \theta^*\right) - 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right), \theta^* - \left(\langle\hat{\theta}\rangle - \theta^*\right) + 1.96 \cdot \mathrm{SD}\left(\hat{\theta}\right)\right]$, where $\langle\hat{\theta}\rangle$ and $\mathrm{SD}\left(\hat{\theta}\right)$ are the mean and standard deviation (SD) of $\hat{\theta}$ over the simulated datasets, and θ^* is the value inferred from the real data [4]. 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 (see sections 6.4.3 and 6.4.7).

6.3.3. Results

In Supplementary Table 5, Supplementary Table 7, and Supplementary Note Table 3, we present the inferred parameters for all models and populations. Whenever parametric bootstrap sampling was carried out (section 6.3.2), we report the bias-corrected means, the standard deviations (SD), and the

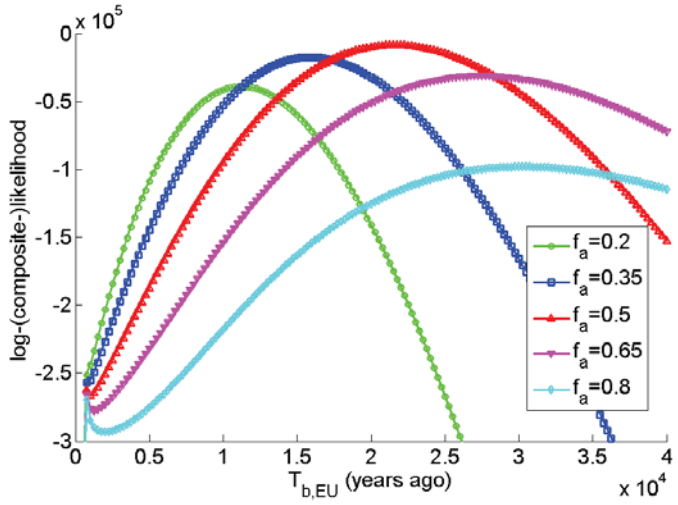
confidence intervals. Otherwise, we report the maximum-likelihood parameters. A diagram with the inferred parameters for the single-population bottleneck/growth models for AJ and Flemish appears in Supplementary Figure 7; a diagram for the two-dimensional model (with a recent AJ bottleneck) is presented in Figure 4 of the main text. The results are discussed in section 6.4.

Parameter	Maximum likelihood	Bias-corrected mean±SD	95% confidence interval
N_0	14,148	14,105±54	[13,999 , 14,210]
$N_{b,OOA}$	4879	5004±101	[4806 , 5202]
$T_{b,OOA}$	114,733	112448±3375	[105,832 , 119,064]
$N_{f,AJ}$	16,502	16,203±417	[15,387 , 17,020]
$N_{b,EU}$	3365	3897±60	[3779 , 4014]
$T_{b,EU}$	22,952	23,426±406	[22,631 , 24,222]
$N_{f,EU}$	122,204	125,357±6296	[113,017 , 137,698]
T_a	4196	3646±211	[3232 , 4060]
f_a	55%	55%±1%	[53% , 57%]

Supplementary Note Table 3. *The inferred parameters for the joint AJ-Flemish demographic model without recent AJ bottleneck.* The model is defined in section 6.2.3.26.2.3.1 and illustrated in Supplementary Note Figure 18. The maximum likelihood parameters were computed using *∂a∂i* [4] (section 6.3.1); confidence intervals were obtained using parametric bootstrap (section 6.3.2): we report the bias-corrected means, the standard deviations (SD), and the 95% confidence intervals. Population sizes are given in number of diploid individuals and times in years.

6.3.4. Likelihood surfaces

Two interesting results of the two-population model (see discussion in section 6.4.4) are the large fraction of European ancestry in AJ ($f_a \approx 50\%$) and the relatively recent date of the European founder event ($T_{b,EU} \approx 21{,}000$ years ago). Both parameters have standard deviation of just ≈2% of the mean (Supplementary Table 7), indicating high confidence. To examine the curvature of the likelihood surface around these parameters, we plot, in Supplementary Note Figure 20, the log-likelihood when fixing all other parameters to their maximum likelihood estimate and varying f_a and $T_{b,EU}$. For both f_a and $T_{b,EU}$, the likelihood is clearly maximized at the inferred values, lending support to an admixture fraction around 50% and to a European founder event taking place around 21,000 years ago.

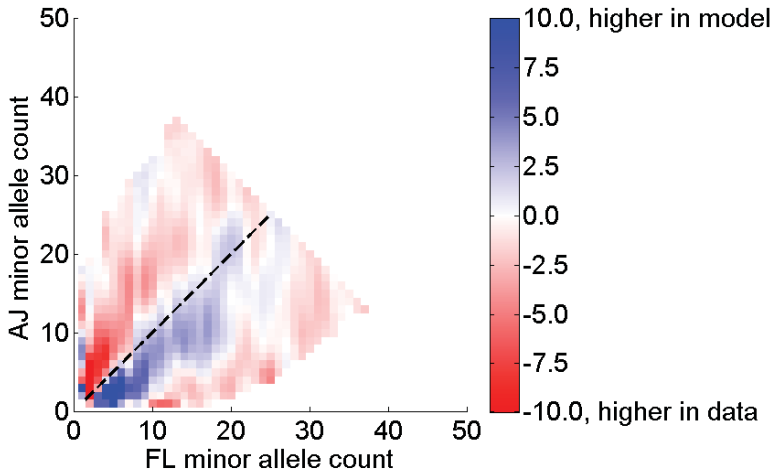


Supplementary Note Figure 20. *Likelihood curves for the admixture fraction (f_a) and the time of the European founder ($T_{b,EU}$).* We used the two-population model of section 6.2.3.2 (Supplementary Note Figure 19). We fixed the values of all parameters to their maximum likelihood estimates, and varied f_a and $T_{b,EU}$. The log-(composite-) likelihood was computed, for each parameter set, using *∂a∂i*.

6.3.5. The best fitting spectra

In Supplementary Figure 8, we compare the single-population real spectra to the maximum likelihood spectra of the different models. To generate the model spectra, we used the same *∂a∂i* configuration as used for inference, with an integration time-scale factor of 10^{-4} (or 10^{-5}, whenever the model included a recent expansion). The constant-size and growth-only models fit the real spectrum poorly for both populations. The ancient bottleneck and growth models fit the data well, with or without a recent bottleneck and growth episode for AJ.

In Supplementary Figure 12, we plot the joint two-population real and maximum likelihood spectra. When generating the inferred spectrum using *∂a∂i*, we set the spectrum to zero at any frequencies appearing non-zero in the "mask" field, indicating that the spectrum could not have been reliably computed. To visualize the difference between the real and maximum likelihood joint spectra, we used the Anscombe residuals, as recommended by Gutenkunst et al. (2009) [4] (Eq. (2) in the *∂a∂i* manual; Supplementary Note Figure 21). We set to zero any residual that was infinite or undefined, as well as residuals at frequencies where the number of sites was <5 for both the real data and the model. For visibility purposes, all residuals larger, in absolute value, than an arbitrary cutoff of 10 were set to the cutoff. The maximum likelihood spectrum qualitatively reproduces the features of the real spectrum, but with room for improvement: Supplementary Note Figure 21 reveals regions in the joint frequency space where the model is biased one way or another. However, we verified the correlation between the residuals at nearby frequencies (the "splotchy" pattern) is due to the down-sampling, as previously observed (Figure S8 in Gutenkunst et al. (2009) [4]).



Supplementary Note Figure 21. *The normalized differences (residuals) between the real joint AJ-Flemish spectrum and the best fitting spectrum.* The model inferred is from Supplementary Note Figure **19**, which includes the recent AJ bottleneck. The color of each square corresponds to the Anscombe residuals at the given allele count, which were computed according to the recommendation of Gutenkunst et al. (2009) [4]. Blue colored squares correspond to a larger number of sites in the model (see the color bar on the right); red colors correspond to more sites in the real data. Residuals larger in absolute value than 10 were truncated. Squares appear white where there were too few sites having the given allele count or when the model could not have been reliably computed. The dashed line corresponds to equal frequencies in AJ and Flemish.

6.4. Discussion

6.4.1. Consistency of the ancestral population size across models and populations

Remarkably, the estimated ancestral population size N_0 is highly similar among all models and populations (≈14,000). The ancestral size was estimated using the total number of polymorphic sites (section 6.2.1). Using Eq. (2) of section 3.4.2, the number of sites is $S(n) = \theta \sum_{k=1}^{n} \frac{(4k-1)\binom{2n}{2k}}{\binom{2n+2k-1}{2k}} \int_{-\infty}^{t} \exp\left(-\int_{s}^{t} \frac{\binom{2k}{2}}{\rho(u)} du\right) ds$, where $\rho(t)$ is the scaled population size at time t and n is the sample size. We noticed that for all bottleneck/growth models reported in Supplementary Table 5, $S(n)$ differs by no more than ≈10% from its value under the Wright-Fisher model ($\theta \sum_{i=1}^{2n-1} 1/i$). Therefore, the effect of the relatively recent demography (≈100,000 years) on the number of sites, and hence the estimated N_0, is small.

6.4.2. Interpretation of the single-population models

Our results showed that a bottleneck/growth model fits the empirical allele frequency spectrum very well for both AJ and Flemish (Supplementary Figure 8), whereas simpler models (constant-size or growth-only) do not. Therefore, for our sample size, the four-parameter bottleneck/growth model is "necessary and sufficient" to describe the AFS (but larger samples will likely require more elaborate models). The inferred parameters suggest some very general trends, such as a more recent and severe bottleneck at the founder event of Europeans compared to AJ. We do not attempt to attach a more precise historical interpretation to the inferred values, which are averages (and single-population projections) of the more complex actual demographic histories. We provide some historical

interpretations for the more detailed two-population model below (section 6.4.4). Our four-parameter set (ancestral population size, bottleneck time and size, and final population size) can also be thought of as a generalization of the single effective population size N_0 (or equivalently, the scaled mutation rate θ), which is traditionally used to describe genetic data. This is an important extension; for example, when considering the expected number of variants in large samples (section 3.4.2, Figure 1 of the main text), using the simple constant-size model predicted more variants in AJ than in Flemish, whereas the more elaborate bottleneck/growth model predicted the opposite trend.

6.4.3. The effect of the mutation rate

As mentioned in section 6.2.1, the ancestral population size N_0, and consequently, all inferred times and population sizes, are inversely proportional to the mutation rate. Therefore, uncertainty in the mutation rate has chief effect on the inferred parameters. For example, if the true mutation rate is two-fold smaller, all population sizes and times will double; this magnitude of uncertainty is at least 10-fold the uncertainty associated with sampling noise, as captured by our confidence intervals (see, e.g., Supplementary Table 5 and Supplementary Table 7). Gutenkunst et al. (2009) [4], who introduced *∂a∂i*, as well as others (e.g., [3, 85, 86, 87]), used a mutation rate calibrated using the human-chimp divergence, or the *phylogenetic* rate estimate (in [4], $\mu = 2.35 \cdot 10^{-8}$ per bp per generation). We carried out our initial analysis using that rate. However, the phylogenetic mutation rate depends crucially on rare fossils to estimate the human-chimp divergence time. Recently, the mutation rate was estimated more directly by looking at differences between individuals in a pedigree (*de novo* rate estimate) and was found to be ≈1-1.5$\cdot 10^{-8}$ per bp per generation, that is, about half of the phylogenetic estimate (see, e.g., the recent reviews at Refs. [14, 88, 89, 90] and references therein). Very recently, Gravel et al. (2013) [58] estimated the mutation rate as $1.44 \cdot 10^{-8}$, at the upper range of the *de novo* estimates, by analyzing Native American whole-genomes. To obtain that rate, Gravel et al. (2013) inferred the parameters of a demographic model for the settlement of the Americas using Native Americans tracts in whole genomes from Colombia, Mexico, and Puerto-Rico. Then, they equated the time of a narrow bottleneck in the early history of the region with the accepted time of the population of the Americas (16,000 years). The results we present here are based on this rate.

We note that concerns over the mutation rate are not specific to our study and will affect any attempt of demographic inference based on allele counts/frequencies. In our case, using whole genomes reduced the (sampling) confidence intervals so dramatically that uncertainties due to the mutation rate (or the model specification) became prominent. As the topic remains debatable (e.g., [91]), we occasionally suggest alternative historical interpretations based on the higher mutation rate. Additionally, when referring to the key parameter of the Middle-Eastern-European divergence time, we often cite a conservative interval of ≈12-25 kya (using the point estimate of ≈21 kya and mutation rate between $1.2 \cdot 10^{-8}$ to $2.5 \cdot 10^{-8}$), to fairly represent the uncertainty in the mutation rate.

6.4.4. Historical interpretation of the two-population model

We constructed two-population demographic models (section 6.2.3; Supplementary Note Figure 18 and Supplementary Note Figure 19) to capture the main events in the AJ and European histories. Multiple lines of evidence from archaeology, linguistics, and human and microbial genetics suggest that all anatomically and behaviorally modern humans descend from an ancestral population living in Africa

≈150-200 kya. Non-African populations are thought to have arisen as a result of a single, or at most a handful of, migration events into the Near East, with time estimates varying between ≈50 to ≈100 kya. More distant regions were populated in a series of bottlenecks and subsequent expansions, known as the "serial founder effect", and leading to gradual loss of genetic diversity with increasing distance from Africa [2, 78, 79, 80, 81, 92, 93, 94]. In our model, the population on the right side of Supplementary Note Figure 18 (or Supplementary Note Figure 19) is assumed to initially represent the ancestral African population, and then to undergo a bottleneck when the Middle-Eastern population was formed. The timing of the bottleneck, corresponding to the Out-of-Africa event, was estimated as ≈90,000 years ago, within the range previously suggested (in particular since we used a lower mutation rate; see also section 6.4.5). We do not model the African population after the exit from Africa. Following the Out-of-Africa bottleneck, the Middle-Eastern population is assumed to begin a slow exponential expansion.

In line with the serial founder effect, the European population (represented by the Flemish, left side of Supplementary Note Figure 18 or Supplementary Note Figure 19) is then assumed to depart from the Middle-Eastern population and undergo a bottleneck. Fossils of anatomically modern humans from as early as ≈40-45 kya [95, 96, 97] have been found in Europe. However, the point estimate of the divergence time between the European and Middle-Eastern populations was more recent at ≈21 kya. A possible explanation is that contemporary Europeans descend, for the most part, from a near-Eastern population that repopulated Europe at the end of the Last Glacial Maximum (LGM) between ≈19-26 kya [98, 99, 100]. These results have consequences to European origins, suggesting genetic discontinuity between modern Europeans and the original hunter-gatherer inhabitants of the continent, and that the major dispersal from the Near-East to Europe preceded the Neolithic revolution. However, the implications to Neolithic migrations are highly sensitive to the mutation rate and the model specification, as we further discuss in section 6.4.6. At any rate, this is a coarse grained picture; finer details being resolved by examination of relevant ancient genomes [101, 102, 103, 104, 105, 106].

The inferred exponential population growth in Europe from ≈3700 to ≈170,000 is in line with other studies observing a dramatic recent population expansion [51, 55, 57, 107, 108]. The slower growth rate in AJ could be due to a more stable ancient population size (post-Out-of-Africa) [78]. The inferred time of European admixture into AJ coincides with the time we inferred (using IBD, section 4.3) for the recent AJ bottleneck, at ≈700 years ago. While the proposed time is plausible (the AJ population has resided in Europe at that time), we also cannot rule out numerical issues due to the very fast changes in population size after the recent bottleneck. When not fixing a recent bottleneck, the admixture time came out as ≈4000 years ago. We are not aware of a possible interpretation for such an early date, but a somewhat earlier date of ≈3000 years ago may be consistent with gene flow from Philistines [109]. An even more recent date of ≈2000 years ago (using a higher mutation rate) may correspond to the exile of Jews from Palestine, partly into Rome [110]. The fraction of European ancestry, around 50%, is a little higher than was previously thought [1], but still within the range of previous estimates [34], and consistent with the substantial European maternal ancestry recently suggested by Costa et al. (2013) [111].

6.4.5. Comparison of our inferred dates to previous estimates

There is vast literature on (genetic) estimation of the Out-of-Africa and the European bottleneck parameters, including some very recent publications (e.g., [3, 4, 85, 86, 87, 93, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121,

[122, 123, 124, 125, 126]). The type, magnitude, and population of origin of the genetic data vary widely between studies, as well as the summary statistics used for inference, the proposed model, and the inference method. To complicate things further, almost all previous studies used the higher phylogenetic mutation rate estimate. Therefore, a complete review of all previously published estimates is beyond the scope of this study. Li and Durbin (2011) [85] summarized the main features and inferred parameters of papers that appeared at the time of their publication, most of which relied on small-scale data. We summarize below a number of recent estimates for the timing of the Out-of-Africa and European foundation events.

Li and Durbin (2011) [85], followed by Sheehan et al. (2013) [121] (using the de novo estimate), developed a coalescent-based Hidden Markov Model for the inference of the population size history using whole genomes. However, the method considers one population at a time and has lower resolution for relatively recent events. Broadly, both studies inferred divergence of Africans and Europeans ≈50-100 kya followed by a reduction in the population size. Gronau et al. (2011) [115] used six whole genomes from six diverse populations and a Bayesian, coalescent-based approach to infer European-African divergence time of ≈50 kya and European-East-Asian divergence time of ≈30-40 kya. Gutenkunst et al. (2009) [4] and Lukić and Hey (2012) [86] used 5MB of re-sequencing of ≈200 non-coding regions and the joint allele frequency spectrum (computed using different numerical approaches) to infer European-African divergence time of ≈140 kya or ≈52 kya and European-East-Asian divergence time of ≈21 kya or ≈30 kya (Gutenkunst et al. and Lukić and Hey, respectively). The studies also disagreed on the magnitude of inter-continental migration. Laval et al. (2010) [87] used data of similar nature and allele frequency and haplotype summary statistics combined with Approximate Bayesian Computation to estimate the out-of-Africa event at ≈60 kya and the European-Asian divergence at ≈22 kya. Lohmueller et al. (2009) [117] used haplotype diversity statistics and genome-wide SNP arrays for European individuals to infer a European bottleneck time of ≈37 kya. Theunert et al. (2012) [122] used the average haplotype length around mutations of given frequencies, Approximate Bayesian Computation, and European genome-wide array data to infer a bottleneck time of ≈40 kya. Harris and Nielsen (2013) [125] used the coalescent-based distribution of identity-by-state (IBS) tract lengths and African and European pairs of whole genomes to infer a divergence time of ≈55 kya. Finally, a particularly relevant study is that of Gravel et al. (2011) [3], since we used the same inference method (based on the allele frequency spectrum and *∂a∂i* [4]) and similar type and magnitude of data (few tens of whole genome sequences; albeit our data is high coverage). They used genomes of Africans, Europeans, and East-Asians to infer the Out-of-Africa time at ≈51 kya and the European-Asian divergence at ≈23 kya.

Taken together, with the phylogenetic mutation rate estimate, previous studies have estimated the time of the Out-of-Africa dispersal at ≈50-80 kya, and the time of a European bottleneck, sometimes bundled with the divergence from Asians, at ≈20-50 kya. Due to the extreme diversity of data types and modeling approaches, a direct comparison to our results is impractical. However, in general, had we also used the phylogenetic estimate, the time we would have inferred for the Out-of-Africa event (≈50-60 kya) is within the range of previous estimates (and particularly close to [3, 86, 113, 115]). The time we inferred for the European bottleneck (≈10-15 kya under the phylogenetic estimate) was more recent than all previous studies, with [3, 4, 116] being the closest at ≈20 kya. The ratio between the Out-of-Africa time and the

European divergence time that we inferred (≈4-5), which is independent of the mutation rate, is higher than most studies, again except [4, 116] (which are also based on the allele frequency spectrum), but who also inferred a less realistic Out-of-Africa time of >100 kya (>200 kya under the de novo rate estimate). We propose that our more recent time estimate is (*i*) due to our novel data on and explicit modeling of genomes with (partly) Middle-Eastern origin, and (*ii*) because we do not force the European founder event to take place simultaneously with the European-Asian divergence. Indeed, the recent study of Haber et al. (2013) [99], who used SNP array data of Lebanese and European populations, inferred the divergence time between Levantines and Europeans (but without explicit demographic modeling) at ≈9-16 kya, close to our estimate.

6.4.6. The debate over European origins

Our results for the European-Middle-Eastern divergence time have potential implications regarding open questions on European origins and the Neolithic revolution. The fundamental point of contention is whether the transition to farming was due to cultural exchange ("cultural diffusion") or was also accompanied by human migration, replacing all or most of the existing hunter-gatherer population ("demic diffusion") [98, 100, 101, 102, 103, 104, 105, 106, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145]. Studies have so far employed a wide variety of methods and datasets, with most studies focusing on modern Europeans and unipaternal markers. With advancement in ancient DNA technology, several studies have incorporated the sequencing of ancient early farmers and their contemporary hunter-gatherers. Inferences from previous studies, however, did not converge to a single conclusion, with some studies supporting cultural diffusion and others demic diffusion. Many studies adopted a middle ground, assuming that farming technology was introduced by migrants, but that those early farmers admixed with existing hunter-gatherers, hence giving rise to only a fraction of the modern European ancestry. Estimates of that ancestry fraction, however, also diverged widely between studies and ranged between 20% and close to 100%. Recent studies of ancient European mitochondrial and nuclear DNA suggested multiple waves of migration as well as recovery of some of the hunter-gatherer lineages. To our knowledge, no modeling of European origins so far has been carried out with genome-wide sequencing data for modern European and Middle-Eastern populations.

Our estimated European-Middle-Eastern divergence time of ≈12-25 kya suggests that the major dispersal from the Middle-East into Europe took place long before the invention of agriculture [98, 100]. Consequently, the spread of agriculture within Europe has been facilitated by either cultural exchange or migrations within the continent.

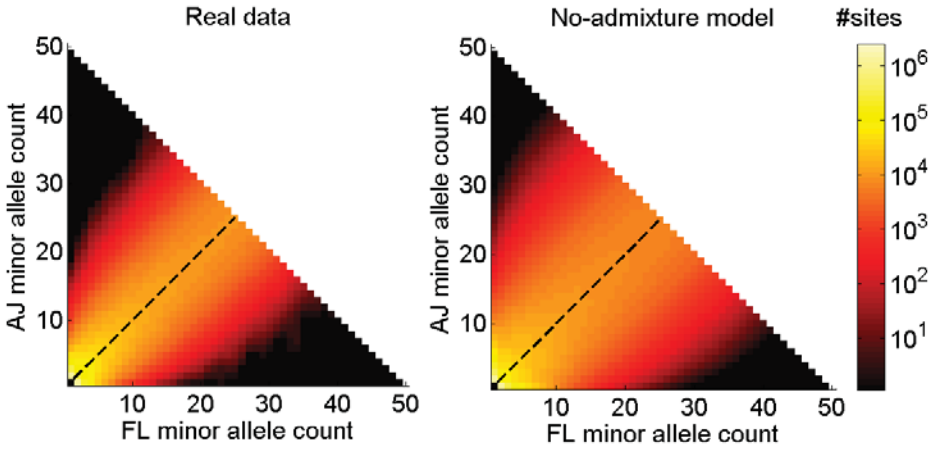
However, the picture is complicated by the uncertainty in estimating the mutation rate and, to lesser extent, by our inability to infer significantly more complex models (for our sample size). If the phylogenetic mutation rate estimate is true, then the divergence time would be dated to ≈10-15 kya. This is still earlier than the time of the Neolithic revolution in Europe (≈5-8 kya). However, this smaller gap can be reconciled with demic diffusion, since Neolithic migrants may have not taken over the population entirely. In this view, which is also supported by recent ancient DNA studies [101, 102], modern Europeans descend from both the original hunter-gatherers and recent Neolithic migrants (≈5-8 kya), and our inferred time reflects a weighted average of the two. Another potential reason why even an early divergence time of ≈10-15 kya could support demic diffusion is an ancient Middle-Eastern

substructure. The debate is expected to last at least until the uncertainty in the mutation rate is significantly reduced.

6.4.7. Robustness to model specification

6.4.7.1 The necessity of admixture into AJ

One of the components of the two-dimensional model is a recent admixture pulse from Europeans into AJ. The admixture pulse is necessary to guarantee that the AJ-Flemish allele frequencies are correlated. To see this, we generated the joint AFS for a model that has exactly the same parameters as the inferred AJ-Flemish model (without the recent bottleneck; Supplementary Note Table 3), except with no admixture. The synthetic AFS is compared to the real AFS in Supplementary Note Figure 22, clearly showing that in the absence of recent admixture (even with a relatively recent EU-Middle Eastern divergence time ≈21 kya), the allele frequencies are considerably less correlated than in the real data. When re-inferring the parameters of a model similar to that of Supplementary Note Figure 18 except without recent admixture, the correlation is similar to that of the real data (not shown), but with a much lower overall likelihood (log-likelihood -14,901 vs. -8079 for the full model) and with the inferred current Flemish population size being unrealistic at $1.7 \cdot 10^9$. To test formally for the necessity of admixture, we generated 100 synthetic spectra for the inferred model without admixture (using the same simulation method of section 6.3.2). Then, we fitted each spectrum to the demographic model either with or without admixture. Since the full model has two additional parameters (the time and fraction of admixture) it is expected to fit better, even for data generated without admixture. However, the improvement in the log-likelihood when using the full model was on average only 32.7 log-likelihood units (maximum 117.1), compared to 6822 for the real data. We therefore conclude that including admixture in the full model (Supplementary Note Figure 18) significantly improves the model's fit to the AJ-Flemish joint spectrum.



Supplementary Note Figure 22. *The real AJ-Flemish spectrum compared to the spectrum of a model without admixture.* The real spectrum (left) is as in Figure 3B of the main text. The parameters of the model spectrum are the ones inferred from the real data and using the full model (without the recent bottleneck; Supplementary Note Table 3), except with no recent pulse admixture from the European population into AJ. The allele frequencies in the no-admixture model are less correlated than in the real data ($r = 0.79$ vs. 0.88 in the real data).

6.4.7.2 European population derives from African, not from the Middle-East

One of the assumptions of our model (Supplementary Note Figure 18) is that the European population is derived from the Middle-Eastern population. Alternatively, one may suggest a model where *both* European and Middle-Eastern populations descend from the ancestral (supposedly) African population. We therefore created a new model, in which an ancestral population of size N_0 gives rise, via a bottleneck, to both populations at different times. The two populations are then assumed to grow exponentially. We allowed either population to be the first to diverge, and we did not follow the ancestral population after the second split. We then assumed recent admixture of Europeans into the Middle-Eastern population (corresponding to AJ), as in section 6.2.3. However, fitting the model gave very poor results compared to our original model. We therefore conclude that a Middle-Eastern origin of the European population is more likely.

6.4.7.3 Two-way, continuous migration between Europeans and AJ

Another assumption of our model (Supplementary Note Figure 18) is that gene flow between Europeans into AJ takes the form of instantaneous , directional admixture from Europeans into AJ (where the directionality is justified, to some extent, by the larger number of AJ-specific variants). To relax this assumption, we allowed bi-directional, asymmetric migration for a variable period of time since the beginning of the admixture epoch. The results for the maximum likelihood fit are given in Supplementary Note Table 4. The best fitting model is generally similar to that inferred in Supplementary Note Table 3. The Out-of-Africa event was dated to ≈108 kya, and the European bottleneck to ≈25 kya. Gene flow, however, is dated to ≈11 kya, likely capturing a non-Jewish-specific event. The extent of migration into AJ, ≈0.0026 per generation for ≈300 generations, corresponds to ≈ $1-(1-0.0026)^{300} \approx 54\%$ European admixture, as inferred in Supplementary Note Table 3. The time when admixture ended, ≈4000 years ago, is in agreement with the admixture time inferred without the migration period. We note, however, that our experiments (not shown) demonstrated limited power of *∂a∂i* to infer migration parameters even in simple models, in particular for asymmetric migration. The above results should therefore be interpreted with caution.

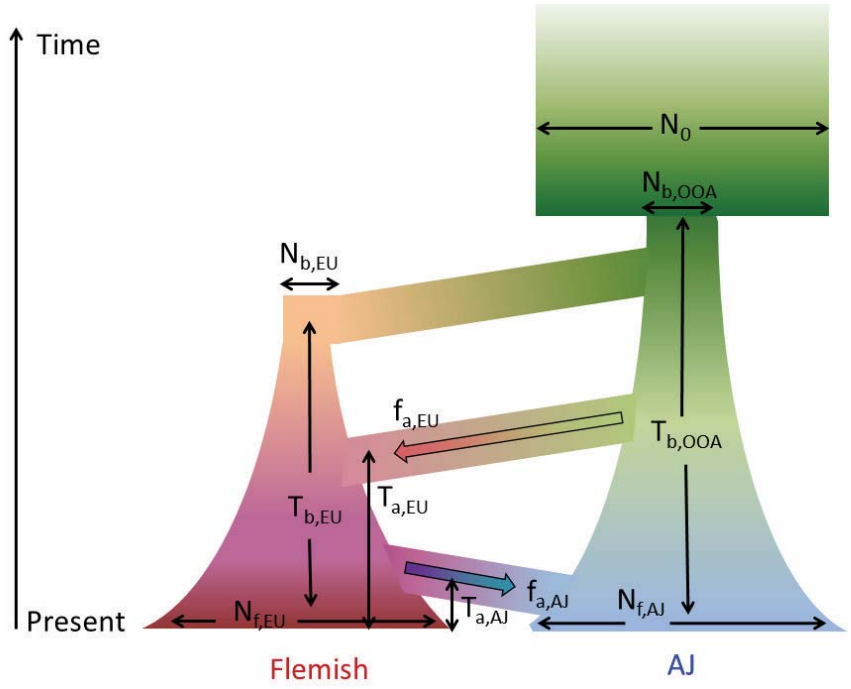
Two-population model with known, recent AJ bottleneck	Maximum likelihood values
N_0	14,113
$N_{b,OOA}$	4335
$T_{b,OOA}$	107,569
$N_{f,AJ}$	20,811
$N_{b,EU}$	2238
$T_{b,EU}$	24,498
$N_{f,EU}$	51,021
$T_{a,start}$	11,450
$m_{EU\to AJ}$	$2.61\cdot 10^{-3}$
$m_{AJ\to EU}$	$0.84\cdot 10^{-3}$
$T_{a,end}$	3925

Supplementary Note Table 4. *The inferred demographic parameters for a two-population model with recent, bi-directional, continuous migration.* The parameter $T_{a,start}$ stands for the time when admixture started. Then,

$m_{EU \to AJ}$ is the migration rate from the European population into the AJ (or ancestral Middle-Eastern) population in units of the fraction of the AJ population per generation replaced by migrants. $m_{AJ \to EU}$ is the migration rate in the opposite direction. Migration ended $T_{a,end}$ years ago, after which the populations were isolated. All other parameters are as in Supplementary Note Figure 18. Population sizes are given in number of diploid individuals and times in years.

6.4.7.4 An additional migration wave from the Middle-East into Europe

In section 6.4.6, we suggested that the ancestry of modern Europeans could be traced partly to Neolithic migrants and partly to pre-existing hunter-gatherers. To investigate this possibility, we inferred the parameters of the model shown in Supplementary Note Figure 23, which is exactly as in section 6.2.3 (Supplementary Note Figure 18), except that we added an admixture event from the Middle-Eastern population into the European population. We hypothesized that in this model, the divergence of Europeans would correspond to the initial peopling of Europe (≈40-45 kya), while the more recent admixture event would correspond to incoming Neolithic migrants. The maximum likelihood parameters (Supplementary Note Table 5) show, surprisingly, that even with the additional admixture event, the divergence time of Europeans from Middle-Easterners is as recent as ≈27 kya. This further supports genetic discontinuity of modern Europeans with the original settlers. According to the inferred model, the admixture event from the Middle-East into Europe occurred ≈17 kya and replaced 63% of the European population, which could correspond to the recovery from the LGM or to Neolithic migration into Europe, in case of a higher mutation rate. Other parameters with notable difference from the simpler model (Supplementary Note Table 3) are a very narrow European bottleneck size (≈400 diploids) and a more ancient admixture of Europeans into AJ (≈6000 ya). We note that this model achieved a much higher log-likelihood than the simpler model of Supplementary Note Figure 18 (-5170 vs. -8078), but we did not compare the models formally using simulations. The confidence intervals (generated as usual using parametric bootstrap) were of similar magnitude to those reported in Supplementary Note Table 3 (not shown). Inferring a model as presented in Supplementary Note Figure 23 but with a (fixed) recent AJ bottleneck did not converge to a consistent parameter set.



Supplementary Note Figure 23. *A diagram of a two-population demographic model with an admixture event from the Middle-East into Europe.* The model is exactly as in Supplementary Note Figure **18**, except that $T_{a,EU}$ generations ago, there was an admixture event from the Middle-Eastern population (right) into the European population (left), replacing a fraction $f_{a,EU}$ of Europeans. The time and fraction of the more recent admixture event from Europeans into AJ are denoted $T_{a,AJ}$ and $f_{a,AJ}$, respectively.

Two-population model with two waves of migration into Europe	Maximum likelihood values
N_0	14,093
$N_{b,OOA}$	4402
$T_{b,OOA}$	106,044
$N_{f,AJ}$	18,305
$N_{b,EU}$	436
$T_{b,EU}$	26,980
$N_{f,EU}$	185,868
$T_{a,EU}$	17,259
$f_{a,EU}$	63%
$T_{a,AJ}$	6087
$f_{a,AJ}$	54%

Supplementary Note Table 5. *The inferred demographic parameters for a two-population model with an additional admixture event from the Middle-East into Europe.* Population sizes are given in number of diploid individuals and times in years.

6.4.7.5 More complex models

We also attempted to fit a number of more complex demographic models. Those were based on the model of Supplementary Note Figure 23 with additional parameters, such as different growth rates at

each epoch (specifically, attempting to capture the different growth rates before and after the Neolithic revolution). However, no model more complex than Supplementary Note Figure 23 showed consistent convergence into a single parameter set— that is, optimizing from different initial conditions gave markedly different final parameters with a comparable likelihood (not shown). We therefore conclude that more elaborate models would require either more samples or improved inference methods. Another potential future refinement of the model is the inclusion of an African population, either as a "ghost" population that is eventually integrated out or using real data.

7. Supplementary Note 7: Functional variants

7.1. The allele frequency spectrum of coding variants

Recently, several genome-wide studies have demonstrated that deleterious alleles have lower frequencies than benign ones, as expected under purifying (negative) selection [23, 51, 56, 57, 146, 147, 148, 149]. To explore the relationship between allele function and frequency in our sequencing data, we used the merged and imputed AJ-Flemish genotypes (section 2.14.2) and further retained random 26 AJ individuals to match the Flemish sample size (variants appearing only in the removed individuals were discarded). We then annotated all variants using the SeattleSeq Variant Annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation137/), which provides relevant information, including PolyPhen2 score for coding variants [5]. In Supplementary Figure 13, we plot the average PolyPhen2 score vs. the (non-reference) allele frequency for the coding variants in AJ and Flemish. Indeed, the score (which corresponds to the posterior probability that the variant is damaging) decreased with increasing allele frequency, with no significant difference between AJ and Flemish ($P = 0.982$ for combined linear regression with ancestry as an additional covariate). Supplementary Figure 14 shows violin plots for the distribution of (non-reference) allele frequencies for AJ and Flemish and for several functional categories (intergenic, intronic, coding synonymous, missense, non-sense, and splice sites). Here too, allele frequencies decrease with increasing functional significance, with no visual difference between the AJ and Flemish spectra.

7.2. A comparison of the deleterious allele burden between Ashkenazi Jews and Flemish

The comparative load of deleterious mutations has been a subject of recent interest and debate [51, 107, 149, 150, 151, 152, 153, 154, 155, 156, 157]. It has been conjectured that "bottlenecked" populations harbor (proportionally) more deleterious mutations, because of the increased genetic drift and thus the weakening of natural selection during the bottleneck. In AJ, such a mechanism was proposed to explain the accumulation of several AJ-specific genetic disorders [46, 158]. To determine whether such an effect is observed when comparing the AJ population and the Flemish, we used the merged (and reduced to equal sample sizes) dataset described in section 7.1. We recorded the number of (non-reference) variants in each population according to four definitions (Supplementary Table 8, rows): (*i*) the total number of unique variants; (*ii*) the total number of appearances of the variants; that is, weighing each variant according to its frequency; and (*iii*) and (*iv*) same as (*i*) and (*ii*), but only for variants of (non-reference allele) frequency <10% in the combined AJ-Flemish dataset. The reason for considering only

variants with low frequency is that more common variants are less likely to be deleterious. With each of the above definitions, we counted variants that were either (*i*) non-coding; (*ii*) coding and synonymous; (*iii*) coding and non-synonymous; (*iv*) coding and benign (according to the PolyPhen2 annotation); and (*v*) coding and damaging ('possibly-damaging' or 'probably-damaging' according to PolyPhen2) (Supplementary Table 8, columns).

To determine whether there are proportionally more deleterious variants in AJ, one must account for the genome-wide larger number of variants in AJ (section 3.1). This was carried out by using the background neutral variation to compute the expected number of deleterious variants, had no enrichment of deleterious variants existed:

$$\#\text{expected\_deleterious\_AJ} = \frac{\#\text{neutral\_AJ}}{\#\text{neutral\_FL}} \times \#\text{deleterious\_FL}, \quad (17)$$

where, e.g., $\#\text{neutral\_AJ}$, is the number of neutral variants in AJ, etc. In other words, we assumed that in the absence of enrichment, the ratio between the number of deleterious variants in AJ and in Flemish should be the same as the ratio for neutral variation. When computing the expected number of deleterious variants of low frequency (see above), we used all variants to compute the expected neutral ratio. We finally considered three comparisons: (*i*) the number of non-synonymous AJ variants compared to the expected number based on both non-coding and synonymous variation; (*ii*) the number of non-synonymous AJ variants compared to the expectation based on synonymous variation only; and (*iii*) the number of coding damaging AJ variants compared to the expectation based on coding benign variation. To obtain an approximate P-value, we first computed the Poisson residual:

$$R = \frac{\#\text{observed\_deleterious\_AJ} - \#\text{expected\_deleterious\_AJ}}{\sqrt{\#\text{expected\_deleterious\_AJ}}}. \quad (18)$$

To transform the residual R into a P-value, we assumed that R is distributed like a standard normal variable. We did not correct for multiple comparisons.

The results for all (3x4=12) comparisons are shown in Supplementary Table 8. In all comparisons, the number of deleterious AJ variants was larger than expected, and in most cases, the difference was significant ($P < 0.05$). Specifically, the P-value was lowest for comparisons involving the total number of appearances of variants with low frequency, which is expected to be the most informative on deleterious variant load. However, care must be taken when interpreting the results. First, the enrichment is not as high when comparing non-syn. to syn. (or damaging vs. benign) variation as it is when comparing non-syn. to non-coding variation. This could be due to differences in exome sequencing quality between the AJ and Flemish genomes. Indeed, for AJ, the average fractions of the genome and exome called were 96.5% and 98%, respectively, while for Flemish (for the genomes where those statistics were available), the fractions were 97% and 94.6%, respectively. On the other hand our AJ-Flemish merging pipeline has specifically removed variants systematically not called in one of the populations. The next concern is with respect to the definition of a deleterious variant. We assumed that the non-reference allele is the deleterious, but other choices are possible, such as using the derived allele (but see section 6.3.1).

Another concern is whether high frequency or even fixed alleles (which are only weakly selected, if at all) should be considered deleterious. We showed results when considering either all variants or only variants of low frequency (<10%). However, our frequency cutoff is arbitrary, and there might have been some non-obvious artifacts due to removing variants found above the cutoff in one population and below in the other (but with an overall high frequency). When considering singletons only, there is no enrichment (although this is reasonable, if the effect of the genetic drift was indeed to increase frequencies of deleterious variants). Finally, a recent paper [151] has demonstrated technical problems in previous comparative studies of the mutation burden and suggested, based on simulations and theory, that recent bottlenecks are expected to have only a negligible effect on the total load.

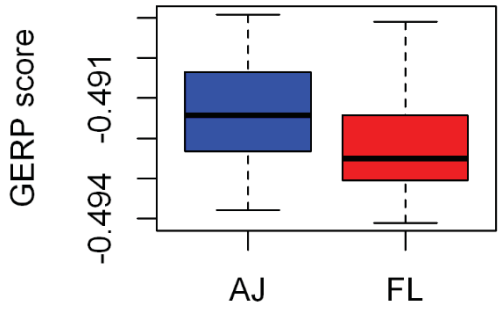
7.2.1. Additional analyses

7.2.1.1 The number of functional variants per individual

To address the question from additional angles, we show in Supplementary Figure 15 violin plots for the distribution of the number of coding variants per individual (among 26 AJ and 26 Flemish) for synonymous, UTR, and missense variants (the number of nonsense and splice variants was <100 per genome). As above, to account for the effect of neutral evolution, we normalized the number of variants in each category by the corresponding number of intergenic variants. There were ≈0.3-1% more coding variants in AJ, across all categories. However, the differences between AJ and Flemish were not or barely significant (P=0.321 for synonymous, 0.021 for UTR, and 0.074 for missense).

7.2.1.2 Genome-wide GERP analysis

Finally, we examined variation outside coding variants by computing GERP scores for variants along the genome. GERP scores provide indication of the evolutionary constraints at each position in the genome based on alignment of several mammalian genomes [159]. We computed the average GERP score for all non-reference variants in each of 26 AJ and 26 Flemish genomes. A box plot for the distribution of the average scores within AJ and within Flemish is presented in Supplementary Note Figure 24, demonstrating slightly higher scores (i.e., more conserved sites) for AJ (P=0.01; t-test). This result is consistent with (small-scale) relaxation of natural selection in the AJ population.



Supplementary Note Figure 24. *The distribution of GERP scores in AJ and Flemish*. The average GERP score was computed over all variants in each individual. The distributions of the average scores over 26 AJ individuals and 26 Flemish individuals are compared.

In summary, our results support a slight (and only weakly significant) enrichment in the deleterious mutation load in AJ compared to Flemish. We expect more definitive conclusions regarding the significance of this effect to be reached with future improvements in sequencing quality, annotation tools, and population genetics theory. The results, however, exclude a large effect such as the one found in French Canadians [155].

7.3. A comparison of the deleterious variant load between AJ and Flemish in different disease categories

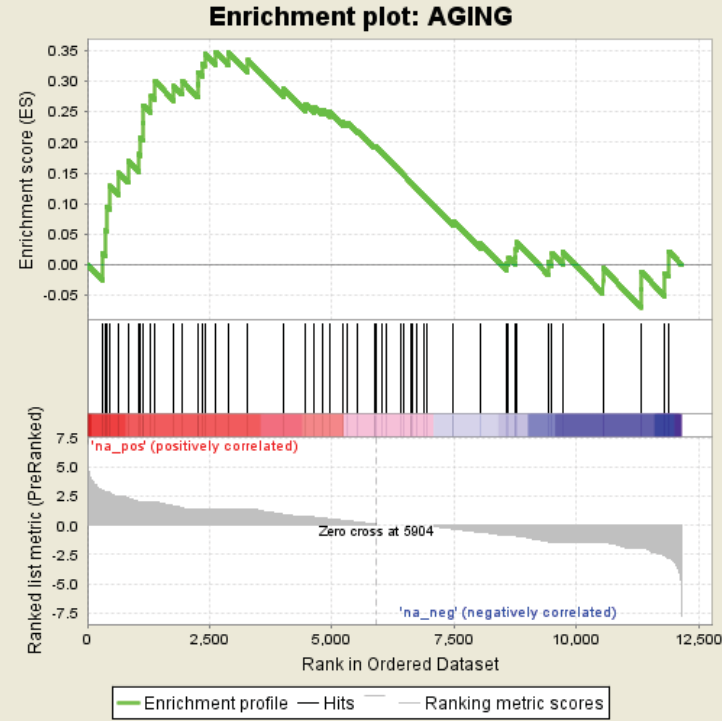
Ever since the past century, AJ have been shown to have a higher prevalence of a number of diseases (compared to non-Jewish Europeans), including several Mendelian disorders (e.g., Tay-Sachs disease, Gaucher disease, etc.) [45], cancers (e.g., hereditary breast cancer and colorectal cancer [160, 161]), inflammatory bowel diseases [162], diabetes [163], and some psychiatric diseases [163, 164]. Many of the suggested health disparities were later refuted [163, 165, 166]. We sought to determine, using our sequencing data, whether there is any disease category with particularly high deleterious mutational burden in AJ. To this end, we used the merged (and reduced to equal sample sizes) AJ-Flemish dataset described in section 7.1, and computed, for each gene, the total number of non-synonymous (non-reference) variants appearing in each population. To determine the disease categories associated with each gene, we used the gene annotation dataset developed by Moore et al. (2011) [21] and later expanded and kindly provided to us by Omicia Inc. (http://www.omicia.com/). The Omicia catalog has 5494 annotations of 2488 genes into 17 disease categories (the rows of Supplementary Table 9). For each category, we counted the total number of non-synonymous variants appearing in all genes belonging to the category, either in AJ or in Flemish.

The results (Supplementary Table 9) show an overall slight excess of non-syn. variation in AJ, as expected from the results of section 7.2. We also observe variability between different categories, with aging genes, in one extreme, demonstrating an excess of 6.5% in AJ, compared to psychiatric genes, in the other extreme, with an excess of 7.4% in Flemish. To determine whether such variability is expected by chance or, alternatively, there is a disease category with an unexpected excess of non-syn. variation in one of the populations, we used the following approach. First, we computed, for each annotated gene, a Poisson residual, R, according to

$$R = \frac{\#\text{non\_syn\_AJ} - \#\text{non\_syn\_FL}}{\sqrt{\max(\#\text{non\_syn\_AJ}, \#\text{non\_syn\_FL})}}. \quad (19)$$

This is similar to Eq. (18), except that due to the often very small number of overall (non-syn.) variants, the denominator is the maximum of the AJ and Flemish counts. We then ranked all annotated genes according to their value of R. Finally, we used Gene Set Enrichment Analysis (GSEA) [167] to determine whether any of the categories is enriched with particularly top (or bottom) ranked genes. GSEA works by computing, for each gene set, the maximal enrichment score over each possible definition of how many genes constitute the "top" of the list, and then computing empirical (permutation-based) P-value and false discovery rate (FDR). The analysis showed no gene set had reached FDR<0.05. Moreover, except for the aging-related genes, no gene set was even nominally significant (P<0.05). The GSEA report for the aging category is displayed in Supplementary Note Figure 25; the top-ranked aging genes are listed

in Supplementary Note Table 6. In conclusion, our results suggest that (at least using our data) no disease category can be associated with high mutational burden in AJ.



Supplementary Note Figure 25. *A screen shot of the Gene Set Enrichment Analysis (GSEA) applied to genes in the aging disease category.* The aging category attained the highest "maximal enrichment score" (green) among all disease categories. The ranks of the genes in the set are shown as black vertical bars.

Number	Aging gene	Rank among all annotated genes
1	IL2RB	302
2	FBN2	362
3	TGFBR2	391
4	MMP3	468
5	MMP12	646
6	GPX1	832
7	TYMS	1048
8	SAFB	1081
9	CSF3R	1139
10	NUP88	1145
11	MCL1	1300
12	MYT1	1388
13	PIK3CA	1771
14	TNFRSF11B	1937
15	MADD	2257
16	CSE1L	2265

17	DRAP1	2351
18	DAD1	2398
19	HPGD	2625

Supplementary Note Table 6. *Aging genes top ranked for higher non-synonymous variant load in AJ.* All genes with annotated variants (12,157) were ranked according to Eq. (19). The first 19 aging genes were ranked higher than expected by chance (P=0.04; Fisher's exact test).

7.4. Mutations in known AJ disease genes

7.4.1. Creating a list of known disease genes and mutations

To create a catalog of mutations in known AJ disease genes, we started with the list of genes in the Supplementary Table of Ostrer and Skorecki (2013) [44]. We did not consider non-Ashkenazi diseases. Gene names were occasionally corrected and updated to comply with standard gene symbols. For each disease, the coordinates of the indicated mutations (with respect to hg19) were manually determined by combining a number of online databases and tools. We usually began with the HGVS symbol (http://www.hgvs.org/) and attempted to convert it to an absolute coordinate using Mutalyzer (https://mutalyzer.nl/positionConverter) [168] or Variation Reporter (http://www.ncbi.nlm.nih.gov/variation/tools/reporter). Occasionally, we located the disease mutation by examining the sequence surrounding it, as reported in the original publication. We cross-validated the resulting coordinates against OMIM (http://www.omim.org/), UCSC Genome Browser (http://genome.ucsc.edu/), dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), 23andMe (https://www.23andme.com/health/), and occasionally, the Israeli Ministry of Health catalog of genetic diseases in Jews (http://www.health.gov.il/Subjects/Genetics/Documents/book_jews.pdf) [169]. Specifically, we validated that all sources agreed on the gene, coordinate, reference allele, alternate allele, and wherever available, the dbSNP "rsID", the RefSeq ID, and the amino acid change (or position with respect to the exon, for intronic mutations). In Supplementary Data 4, we report, for each mutation, the disease name, the gene symbol and OMIM ID, the mutation in HGVS format (with coordinates following the current RefSeq annotation, even if that changed the "classic" name for the mutation), the dbSNP ID (whenever exists), the chromosome and coordinate, the reference allele, and the alternate allele. For the latter three fields, we followed the convention of the VCF format (http://www.1000genomes.org/node/101). Our final list included 73 mutations in 48 genes (Supplementary Data 4). For the few long, structural variants listed in Ostrer and Skorecki (2013), we only report the corresponding gene start and end coordinates.

7.4.2. Detecting disease genes variants in the sequencing data

To detect variants in the known AJ disease genes, we used the complete project data (128 individuals) and the CGA tools *mkvcf* command to generate a combined sample VCF. We did not carry out any filtering, even of low-quality variants. Using the UCSC Table Browser, we extracted the coordinates of the coding exons (plus two nucleotides upstream and downstream) of all UCSC genes that correspond to our list of disease genes. The resulting intervals were merged using BEDTools [62]. We then used Tabix [170] to extract all variants in either the disease mutation coordinates or in the entire coding regions of the disease genes. We used VCFtools [171] to compute the alternate allele frequency, as well as extract disease genes variants that were previously unknown (with respect to our list).

We detected carriers of 35 known disease mutations in 29 genes. For each known disease mutation, we report the alternate allele count and frequency in our 128 AJ genomes (Supplementary Data 4). We verified that for all mutations, the alternate allele corresponded to the known mutation allele. The absence of an alternate allele count in the table indicates that the locus was homozygous reference forn all individuals. We do not report counts for the structural variants.

We then annotated, using *ANNOVAR* [52], the previously unknown variants for their effect on the protein product. We considered exonic variants as well as splice variants, and treated each allele separately in case of multi-allelic variants. The list of our 953 "discovered" variants along with their annotation and allele counts is reported in Supplementary Data 4. In the future, we plan to integrate our list with the "catalog of risk alleles for Ashkenazi Jewish genetic disorders" maintained by the Erlich lab at the Whitehead Institute (http://erlichlab.wix.com/riskcatalog). We note that theoretically, with a reference panel of 128 diploid individuals, variants with allele frequency >1% should be detected with probability at least $1 - (1 - 0.01)^{256} \approx 0.92$.

For further analysis, we retained only non-synonymous variants with non-reference allele frequency <10% (to eliminate likely benign alleles) and no-call rate <10%. We classified the remaining 533 variants according to the following categories: missense, non-frameshift (for multi-nucleotide variants), nonstop, nonsense, splicing, and frameshift. For each gene and each category, we computed the number of unique variants, the number of singletons, the number of doubletons, and the total number of non-synonymous variants appearing in all individuals. We also computed similar counts for variants not in dbSNP135 (see section 2.15 for definitions). A table summarizing the results per gene as well as the totals for each functional category also appears in Supplementary Data 4.

The results show, as expected, that most non-synonymous variants (in particular in the loss-of-function categories) are rare— mostly singletons or doubletons. The total number of non-synonymous variants per gene is mostly a function of the gene coding length ($r = 0.92$). The number of multi-nucleotide variants was surprisingly high (although note that for the non-frameshift variants, 11 were in fact single-nucleotide variants co-localizing with a multi-nucleotide variant at multi-allelic loci). Manual inspection of a number of frameshift variants indicated that they generally occur on the main RefSeq transcript, and thus cannot be assumed to affect only rare splicing isoforms. The initial data from the Erlich lab for a comparable number of individuals (n=96) showed a significantly lower number of multi-nucleotide variants (about a dozen, compared to ≈200 here). Many of our multi-nucleotide variants are likely false positives (see section 2.11; using the same method, we observe that without filtering of low-quality variants, the false positive rate for multi-nucleotide variants can be as high as ≈18%). However, this still does not explain the magnitude of the difference, and more work is required to reconcile the two estimates.

8. Supplementary Note 8: A simple association study model

In this section, we analyze a simple model for disease architecture and association study in order to demonstrate the increased power to detect causal variants in founder populations (see also [172]). Suppose the genetic component of a disease is due to N_a alleles, each having frequency f_0. Each allele is

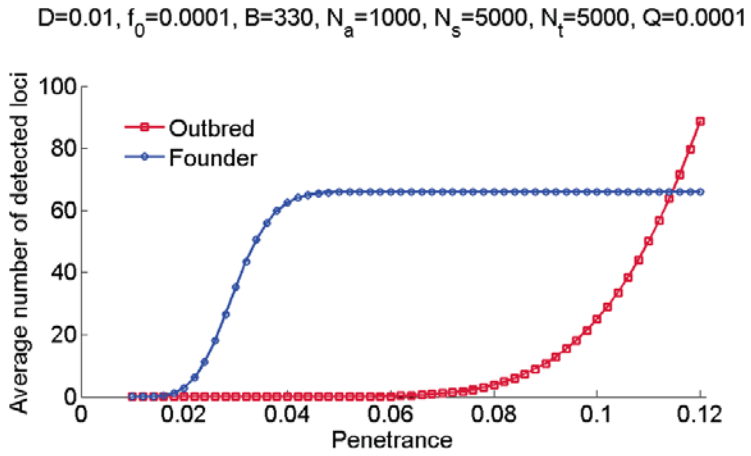
assumed to have penetrance p and a dominant effect, independently of all other alleles. Denote the disease prevalence as D. Assuming that $N_a f_0 \ll 1$, each (diploid) individual has probability $2N_a f_0$ to be a carrier, or probability $2N_a f_0 p\ (< D)$ to be infected. The population is assumed to maintain size $\gg {f_0}^{-1}$ in its recent history, implying that that the disease allele frequency is approximately constant during that period. Suppose a case-control study is carried out with N_s cases and N_t controls. For a given locus, the frequency of carriers in the controls, f_t, is given by Bayes' theorem,

$$(20) \qquad f_t = P(\text{carrier}|\text{control}) = \frac{P(\text{control}|\text{carrier})\times P(\text{carrier})}{P(\text{control})} = \frac{(1-p)\cdot 2f_0}{(1-D)}.$$

Similarly, the frequency in cases, f_s, is

$$(21) \qquad f_s = P(\text{carrier}|\text{case}) = \frac{P(\text{case}|\text{carrier})\times P(\text{carrier})}{P(\text{case})} = \frac{p\cdot 2f_0}{D}.$$

The number of carriers among the controls (cases) is therefore binomial with parameters N_t (N_s) and f_t (f_s). Suppose a χ^2 test (with 1 degree of freedom) is used to compare the number of carriers in the cases and the controls, and that P-values lower than Q are considered significant. The power to detect an association, Π, can then be easily computed using the binomial probabilities. The average number of detected loci, $\Pi \cdot N_a$, is plotted for typical parameter values in Supplementary Note Figure 26.



Supplementary Note Figure 26. *Power to detect an association for outbred and founder populations.* In the outbred population, the frequency of the disease allele is assumed to be $\boldsymbol{f_0 = 10^{-4}}$, while in the founder population, we assume a bottleneck of (diploid) size 330, thus enforcing a minimal disease allele frequency (for the surviving alleles) of $\boldsymbol{f_0 = 1/660}$. The disease prevalence, the number of total alleles, the number of cases and controls, and the P-value threshold are assumed to be $\boldsymbol{D = 0.01}$, $\boldsymbol{N_a = 1000}$, $\boldsymbol{N_s = N_s = 5000}$, and $\boldsymbol{Q = 10^{-4}}$, respectively, typical values for exome-based association studies in complex diseases. The increase in allele frequencies due to the bottleneck makes detection of disease alleles feasible for a wide range of penetrance values.

In a founder population, we assume a similar model, except that the population underwent a recent bottleneck of diploid size B (as we describe for AJ). At the bottleneck, most disease alleles are lost. However, the $\approx 2Bf_0N_a$ alleles that survive increase in frequency to $\approx 1/2B$. We assume that the bottleneck was so recent that we can neglect the effects of genetic drift and natural selection in the period between the bottleneck and the present. The disease allele frequency is therefore assumed to

equal $f_0 = 1/2B$. Eqs. (20) and (21) are then still valid with the new value of f_0, and the power can be computed assuming a χ^2 test as above. The average number of detected loci for an AJ-like bottleneck ($B = 330$) is also plotted in Supplementary Note Figure 26, demonstrating an increased success rate compared to the outbred population (at least as long as the penetrance is not very high).

9. Supplementary References

1. Atzmon G*, et al.* Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *American journal of human genetics* **86**, 850-859 (2010).

2. Li JZ*, et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).

3. Gravel S*, et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**, 11983-11988 (2011).

4. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (2009).

5. Adzhubei IA*, et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).

6. Huffman DM*, et al.* Distinguishing between longevity and buffered-deleterious genotypes for exceptional human longevity: the case of the MTP gene. *The journals of gerontology Series A, Biological sciences and medical sciences* **67**, 1153-1160 (2012).

7. Marder K*, et al.* Familial aggregation of early- and late-onset Parkinson's disease. *Annals of neurology* **54**, 507-513 (2003).

8. Liu X*, et al.* Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population. *BMC medical genetics* **12**, 104 (2011).

9. Drmanac R*, et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

10. Roach JC*, et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639 (2010).

11. Carnevali P*, et al.* Computational techniques for human genome resequencing using mated gapped reads. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 279-292 (2012).

12. Lander ES*, et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

13. Stewart C*, et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236 (2011).

14. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. *Trends in genetics : TIG*, (2013).

15. Hancks DC, Kazazian HH, Jr. Active human retrotransposons: variation and disease. *Current opinion in genetics & development* **22**, 191-203 (2012).

16. Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB. Mobile elements reveal small population size in the ancient ancestors of Homo sapiens. *Proc Natl Acad Sci U S A* **107**, 2147-2152 (2010).

17. Korn JM*, et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* **40**, 1253-1260 (2008).

18. Kenny EE*, et al.* A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet* **8**, e1002559 (2012).

19. Kent WJ*, et al.* The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).

20. Pelak K*, et al.* The characterization of twenty sequenced human genomes. *PLoS Genet* **6**, (2010).

21. Moore B, Hu H, Singleton M, De La Vega FM, Reese MG, Yandell M. Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. *Genetics in medicine : official journal of the American College of Medical Genetics* **13**, 210-217 (2011).

22. Wong LP*, et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *American journal of human genetics* **92**, 52-66 (2013).

23. Genomes Project C*, et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

24. DePristo MA*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

25. Purcell S*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).

26. Turner S*, et al.* Quality control procedures for genome-wide association studies. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 1**, Unit1 19 (2011).

27. Lachance J*, et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457-469 (2012).

28. Kidd JM*, et al.* Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *American journal of human genetics* **91**, 660-671 (2012).

29. Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of human genetics* **70**, 841-847 (2006).

30. Leutenegger AL, Sahbatou M, Gazal S, Cann H, Genin E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *European journal of human genetics : EJHG* **19**, 583-587 (2011).

31. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

32. Behar DM*, et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238-242 (2010).

33. Kopelman NM*, et al.* Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. *BMC genetics* **10**, 80 (2009).

34. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc Natl Acad Sci U S A* **107**, 16222-16227 (2010).

35. Guha S*, et al.* Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome biology* **13**, R2 (2012).

36. Carmi S, Palamara PF, Vacic V, Lencz T, Darvasi A, Pe'er I. The Variance of Identity-by-Descent Sharing in the Wright-Fisher Model. *Genetics* **193**, 911-928 (2013).

37. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-181 (2012).

38. Zakharia F, Bustamante CD. Improved haplotyping of rare variants using next-generation sequence data. (2012).

39. International HapMap C*, et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

40. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5-6 (2013).

41. Meyer LR*, et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research* **41**, D64-69 (2013).

42. Gusev A*, et al.* The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution* **29**, 473-486 (2012).

43. Palamara PF, Lencz T, Darvasi A, Pe'er I. Length distributions of identity by descent reveal fine-scale demographic history. *American journal of human genetics* **91**, 809-822 (2012).

44. Ostrer H, Skorecki K. The population genetics of the Jewish people. *Human genetics* **132**, 119-127 (2013).

45. Charrow J. Ashkenazi Jewish genetic disorders. *Familial cancer* **3**, 201-206 (2004).

46. Risch N, Tang H, Katzenstein H, Ekstein J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *American journal of human genetics* **72**, 812-822 (2003).

47. Need AC, Kasperaviciute D, Cirulli ET, Goldstein DB. 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 biology* **10**, R7 (2009).

48. Olshen AB*, et al.* Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. *BMC genetics* **9**, 14 (2008).

49. Behar DM*, et al.* Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. *Human genetics* **114**, 354-365 (2004).

50. International HapMap C. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

51. Tennessen JA*, et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69 (2012).

52. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

53. Wakeley J. *Coalescent Theory: An Introduction*. Roberts & Company Publishers (2009).

54. Zivkovic D, Stephan W. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theoretical population biology* **79**, 184-191 (2011).

55. Coventry A*, et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010).

56. Marth GT*, et al.* The functional spectrum of low-frequency coding variation. *Genome biology* **12**, R84 (2011).

57. Nelson MR*, et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-104 (2012).

58. Gravel S*, et al.* Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genet* **9**, e1004023 (2013).

59. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

60. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370 (1984).

61. Gusev A*, et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome research* **19**, 318-326 (2009).

62. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

63. Albrechtsen A, Moltke I, Nielsen R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**, 295-308 (2010).

64. 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).

65. Uricchio LH, Chong JX, Ross KD, Ober C, Nicolae DL. Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genetic epidemiology* **36**, 312-319 (2012).

66. Glodzik D*, et al.* Inference of identity by descent in population isolates and optimal sequencing studies. *European journal of human genetics : EJHG* **21**, 1140-1145 (2013).

67. Palin K, Campbell H, Wright AF, Wilson JF, Durbin R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology* **35**, 853-860 (2011).

68. Genovese G, Leibon G, Pollak MR, Rockmore DN. Improved IBD detection using incomplete haplotype information. *BMC genetics* **11**, 58 (2010).

69. Kong A*, et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics* **40**, 1068-1075 (2008).

70. Lencz T*, et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun* **4**, 2739 (2013).

71. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. *Genome research* **19**, 136-142 (2009).

72. Boomsma DI*, et al.* The Genome of the Netherlands: design, and project goals. *European journal of human genetics : EJHG*, (2013).

73. Duan Q*, et al.* Imputation of Coding Variants in African Americans: Better Performance using Data from the Exome Sequencing Project. *Bioinformatics*, (2013).

74. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-959 (2012).

75. Howie BN, 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).

76. Menelaou A, Pulit SL, Francioli LC, Marchini J, de Bakker PIW, consortium GotN. Construction of an accurate haplotype reference panel that incorporates multi-allelic variants from sequencing data. In: *ASHG 2013* (ed^(eds) (2013).

77. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nature reviews Genetics* **11**, 499-511 (2010).

78. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. *Proc Natl Acad Sci U S A* **109**, 17758-17764 (2012).

79. Henn BM*, et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci U S A* **108**, 5154-5162 (2011).

80. Jakobsson M*, et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003 (2008).

81. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**, 15942-15947 (2005).

82. Moorjani P*, et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* **7**, e1001373 (2011).

83. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).

84. Genomes Project C*, et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).

85. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).

86. Lukic S, Hey J. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192**, 619-639 (2012).

87. Laval G, Patin E, Barreiro LB, Quintana-Murci L. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PloS one* **5**, e10284 (2010).

88. Keightley PD. Rates and fitness consequences of new mutations in humans. *Genetics* **190**, 295-304 (2012).

89. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews Genetics* **13**, 745-753 (2012).

90. Ségurel L, Wyman MJ, Przeworski M. Determinants of Mutation Rate Variation in the Human Germline. *Annu Rev Genomics Hum Genet* **15**, 19.11-19.24 (2014).

91. Langergraber KE*, et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A* **109**, 15716-15721 (2012).

92. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nature genetics* **33 Suppl**, 266-275 (2003).

93. Liu H, Prugnolle F, Manica A, Balloux F. A geographically explicit genetic model of worldwide human-settlement history. *American journal of human genetics* **79**, 230-237 (2006).

94. Ray N, Currat M, Berthier P, Excoffier L. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome research* **15**, 1161-1167 (2005).

95. Benazzi S*, et al.* Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479**, 525-528 (2011).

96. Higham T*, et al.* The earliest evidence for anatomically modern humans in northwestern Europe. *Nature* **479**, 521-524 (2011).

97. Mellars P. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* **439**, 931-935 (2006).

98. Olivieri A*, et al.* Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within Europe. *PloS one* **8**, e70492 (2013).

99. Haber M*, et al.* Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet* **9**, e1003316 (2013).

100. Pala M*, et al.* Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. *American journal of human genetics* **90**, 915-924 (2012).

101. Lazaridis I, Patterson N, Mittnik A, etal. Ancient human genomes suggest three ancestral populations for present-day Europeans. (2013).

102. Brandt G*, et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342**, 257-261 (2013).

103. Skoglund P*, et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466-469 (2012).

104. 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).

105. Fernandez E*, et al.* Ancient DNA Analysis of 8000 B.C. Near Eastern Farmers Supports an Early Neolithic Pioneer Maritime Colonization of Mainland Europe through Cyprus and the Aegean Islands. *PLoS Genet* **10**, e1004401 (2014).

106. Skoglund P*, et al.* Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747-750 (2014).

107. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-743 (2012).

108. Gazave E*, et al.* Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci U S A*, (2013).

109. Meiri M*, et al.* Ancient DNA and population turnover in southern levantine pigs--signature of the sea peoples migration? *Scientific reports* **3**, 3035 (2013).

110. *A History of the Jewish People*. Harvard University Press (1976).

111. Costa MD*, et al.* A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nat Commun* **4**, 2543 (2013).

112. Adams AM, Hudson RR. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699-1712 (2004).

113. Fagundes NJ*, et al.* Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**, 17614-17619 (2007).

114. Garrigan D*, et al.* Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics* **177**, 2195-2207 (2007).

115. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* **43**, 1031-1034 (2011).

116. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature genetics* **39**, 1251-1255 (2007).

117. Lohmueller KE, Bustamante CD, Clark AG. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182**, 217-231 (2009).

118. Marth G*, et al.* Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci U S A* **100**, 376-381 (2003).

119. Reich DE*, et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).

120. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome research* **15**, 1576-1583 (2005).

121. Sheehan S, Harris K, Song YS. Estimating Variable Effective Population Sizes From Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **194**, 647–662 (2013).

122. Theunert C, Tang K, Lachmann M, Hu S, Stoneking M. Inferring the history of population size change from genome-wide SNP data. *Molecular biology and evolution* **29**, 3653-3667 (2012).

123. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* **102**, 18508-18513 (2005).

124. Wall JD, Lohmueller KE, Plagnol V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular biology and evolution* **26**, 1823-1827 (2009).

125. Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet* **9**, e1003521 (2013).

126. Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351-372 (2004).

127. Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proc Natl Acad Sci U S A* **95**, 9053-9058 (1998).

128. Richards M*, et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *American journal of human genetics* **67**, 1251-1276 (2000).

129. Rosser ZH*, et al.* Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American journal of human genetics* **67**, 1526-1543 (2000).

130. Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. Estimating the impact of prehistoric admixture on the genome of Europeans. *Molecular biology and evolution* **21**, 1361-1372 (2004).

131. Semino O*, et al.* Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *American journal of human genetics* **74**, 1023-1034 (2004).

132. Haak W*, et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**, 1016-1018 (2005).

133. Belle EM, Landry PA, Barbujani G. Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proceedings Biological sciences / The Royal Society* **273**, 1595-1602 (2006).

134. Battaglia V*, et al.* Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *European journal of human genetics : EJHG* **17**, 820-830 (2009).

135. Bramanti B*, et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**, 137-140 (2009).

136. Balaresque P*, et al.* A predominantly neolithic origin for European paternal lineages. *PLoS biology* **8**, e1000285 (2010).

137. Haak W*, et al.* Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS biology* **8**, e1000536 (2010).

138. Lacan M*, et al.* Ancient DNA suggests the leading role played by men in the Neolithic dissemination. *Proc Natl Acad Sci U S A* **108**, 18255-18259 (2011).

139. Busby GB*, et al.* The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proceedings Biological sciences / The Royal Society* **279**, 884-892 (2012).

140. Fu Q, Rudan P, Paabo S, Krause J. Complete mitochondrial genomes reveal neolithic expansion into Europe. *PloS one* **7**, e32473 (2012).

141. Hervella M*, et al.* Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PloS one* **7**, e34417 (2012).

142. Patterson N*, et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).

143. Rasteiro R, Chikhi L. Female and male perspectives on the neolithic transition in Europe: clues from ancient and modern genetic data. *PloS one* **8**, e60944 (2013).

144. Wei W*, et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome research* **23**, 388-395 (2013).

145. Semino O*, et al.* The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155-1159 (2000).

146. Zhu Q*, et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American journal of human genetics* **88**, 458-468 (2011).

147. MacArthur DG*, et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).

148. Li Y*, et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics* **42**, 969-972 (2010).

149. Hodgkinson A, Casals F, Idaghdour Y, Grenier JC, Hernandez RD, Awadalla P. Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC genomics* **14**, 495 (2013).

150. Lohmueller KE*, et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997 (2008).

151. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. *Nature genetics* **46**, 220-224 (2014).

152. Fu W*, et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).

153. Torkamani A*, et al.* Clinical implications of human population differences in genome-wide rates of functional genotypes. *Frontiers in genetics* **3**, 211 (2012).

154. Gazave E, Chang D, Clark AG, Keinan A. Population Growth Inflates the Per-Individual Number of Deleterious Mutations and Reduces Their Mean Effect. *Genetics*, (2013).

155. 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).

156. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T. Finnish founding bottleneck leads to excess of damaging loss-of-function variants with medically relevant associations. In: *ASHG 2013* (ed^(eds) (2013).

157. Boyko AR*, et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083 (2008).

158. Slatkin M. A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. *American journal of human genetics* **75**, 282-293 (2004).

159. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology* **6**, e1001025 (2010).

160. Lynch HT, Rubinstein WS, Locker GY. Cancer in Jews: introduction and overview. *Familial cancer* **3**, 177-192 (2004).

161. Rubinstein WS. Hereditary breast cancer in Jews. *Familial cancer* **3**, 249-257 (2004).

162. Lynch HT, Brand RE, Locker GY. Inflammatory bowel disease in Ashkenazi Jews: implications for familial colorectal cancer. *Familial cancer* **3**, 229-232 (2004).

163. Goodman RM. *Genetic Disorders among the Jewish People*. The Johns Hopkins University Press (1979).

164. Dohrenwend BP*, et al.* Socioeconomic status and psychiatric disorders: the causation-selection issue. *Science* **255**, 946-952 (1992).

165. Fallin MD*, et al.* Genomewide linkage scan for schizophrenia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 10q22. *American journal of human genetics* **73**, 601-611 (2003).

166. Fallin MD*, et al.* Genomewide linkage scan for bipolar-disorder susceptibility loci among Ashkenazi Jewish families. *American journal of human genetics* **75**, 204-219 (2004).

167. Subramanian A*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

168. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human mutation* **29**, 6-13 (2008).

169. Zlotogora J, van Baal S, Patrinos GP. The Israeli National Genetic Database. *The Israel Medical Association journal : IMAJ* **11**, 373-375 (2009).

170. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718-719 (2011).

171. Danecek P*, et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

172. Zuk O*, et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-464 (2014).