Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery

Eric M Scott1–3, Anason Halees4, Yuval Itan5, Emily G Spencer1–3, Yupeng He1–3, Mostafa Abdellateef Azab1–3, Stacey B Gabriel6, Aziz Belkadi7,8, Bertrand Boisson5,7,8, Laurent Abel5,7,8, Andrew G Clark9, Greater Middle East Variome Consortium10, Fowzan S Alkuraya11,12, Jean-Laurent Casanova1,5,7,8,13 & Joseph G Gleeson1–3

The Greater Middle East (GME) has been a central hub of human migration and population admixture. The tradition of consanguinity, variably practiced in the Persian Gulf region, North Africa, and Central Asia1–3, has resulted in an elevated burden of recessive disease4. Here we generated a whole-exome GME variome from 1,111 unrelated subjects. We detected substantial diversity and admixture in continental and subregional populations, corresponding to several ancient founder populations with little evidence of bottlenecks. Measured consanguinity rates were an order of magnitude above those in other sampled populations, and the GME population exhibited an increased burden of runs of homozygosity (ROHs) but showed no evidence for reduced burden of deleterious variation due to classically theorized ‘genetic purging’. Applying this database to unsolved recessive conditions in the GME population reduced the number of potential disease-causing variants by four- to sevenfold. These results show variegated genetic architecture in GME populations and support future human genetic discoveries in Mendelian and population genetics.

The GME, loosely defined as a large swath of Arab and non-Arab countries from Morocco in the west to as far east as Pakistan5, is home to approximately 10% of the world’s population. Despite the invaluable contribution of GME populations to the understanding of genetic causes of inherited conditions, especially recessive conditions, and the role of the region as a critical hub, serving as a crossroad between early civilizations, genetic architecture and the extent of rare genetic variation remain poorly defined in GME populations6–8.

To address this shortcoming, the GME Variome Consortium collected whole-exome data on 1,794 self-reported nationals from GME regions participating in ongoing genetics studies. To minimize selection bias and over-representation of disease-associated alleles, we selected primarily healthy individuals from families and, wherever possible, removed from data sets the allele that brought the family to medical attention. Samples were jointly processed and filtered for quality and familial relation, leaving 1,111 high-quality samples from unrelated individuals.

We grouped the 1,111 GME exomes into six different GME subregions, northwest Africa (NWA; 85 samples), northeast Africa (NEA; 423 samples), Turkish Peninsula (TP; 140 samples), Syrian Desert (SD; 81 samples), Arabian Peninsula (AP; 214 samples), and Persia and Pakistan (PP; 168 samples) (Supplementary Fig. 1 and Supplementary Table 1), which represent historical groupings, and then compared our data with exomic data for nine established continental populations from the 1000 Genomes Project5. Unbiased identity-by-state (IBS) clustering showed that samples largely grouped according to the location of ascertainment, validating the grouping criteria (Supplementary Fig. 2).

To evaluate GME genetic substructure, we ran the unsupervised algorithm ADMIXTURE10, where k = 6 clusters minimized cross-validation error (Supplementary Fig. 3). We found some overlap with the primary admixture components from Africa, Europe, and East Asia at the edges of the geographical region but also a large proportion of admixture not found in previous reference samples (Fig. 1a and Supplementary Fig. 4). The admixture results also aligned with publications reporting common variation11–13.

The least admixed samples were found in the NWA, AP, and PP subregions, suggesting that populations in these regions are derived...
Figure 1: The Greater Middle East Variome as a hub of human genetics. (a) Map of the GME subregions. Lines between regions define the borders used in analyses of admixture from East Asia, Europe, sub-Saharan Africa, and the new GME components (NWA, NEA, TP, SD, AP, and PP). Pie charts show admixture proportions from 1000 Genomes Project continental populations in analysis with k = 6 clusters. The map was downloaded from https://www.presentationmagazin.com/. (b) Global ancestry proportions (k = 6) contributed by 1000 Genomes Project control populations with three distinct sources of contribution (red, Africa; green, Europe; blue, East Asia) and GME populations (orange, NWA; orange, AP; yellow, PP). Plotted using the hclust algorithm in R. (c) TreeMix phylogeny of GME populations along with 1000 Genomes Project controls representing population divergence patterns. The length of branches is proportional to population drift. GME populations were grouped around the African branch but showed substantial divergence. YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; FIN, Finnish in Finland; GBR, British in England and Scotland; TSI, Toscani in Italia; CHS, Southern Han Chinese; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan. (d) Wright’s fixation index (FST) values for all pairs of GME and 1000 Genomes Project European populations, showing a smaller distance between GME and European populations than between GME and sub-Saharan African populations. The greatest FST value for any two GME populations was 0.026 (a quarter of the distance between the FIN and JPT populations).

From founder populations, but there was evidence of inter-regional variation in GME-specific components, suggesting the occurrence of local admixture (Fig. 1b) and potentially supporting historical events. The NWA component was found in regions from west to east across North Africa, likely representing the Berber genetic background. The AP component likely represents ancestral Arab populations and was observed in nearly all regions, possibly as a result of the Arab conquests of the seventh century coincident with the expansion of the Arabic language now spoken over much of the GME. Similarly, the Persian expansion into the TP and SD regions and parts of NEA in the fifth century was the most likely contributor of the PP signal.

Additional sources of human heterogeneity derive from ancient introgression. We found similar patterns of Neanderthal introgression across all GME populations with the exception of NWA individuals, who clustered closer to sub-Saharan Africans (Supplementary Fig. 5). These data support the reduced extent of Neanderthal introgression observed in native African populations.

Patterns of human migration and drift were recapitulated using TreeMix for GME subregions, on the basis of 1000 Genomes Project control populations (Fig. 1c). The inferred tree with no migration showed the largest inferred drift parameter, supporting a west-to-east trajectory of human migrations.

Assessment of Wright’s fixation index (FST) demonstrated that the GME populations grouped with European populations, agreeing with the TreeMix results. This analysis resulted in three distinct clusters with a low degree of differentiation (Fig. 1d and Supplementary Fig. 6). The PP and NWA populations represented the extremes of the identified subregions and showed the highest degree of differentiation (FST = 0.026; corresponding to a genetic distance more than two times that between the Finnish in Finland and Toscani in Italy). GME populations but smaller than the distances from intercontinental comparisons). FST measurements from comparisons of GME populations to four 1000 Genomes Project European populations showed that GME populations were closest to the TSI population, especially SD and TP, consistent with higher levels of European admixture in these populations. Despite the contribution of admixture, these
values suggest extended periods of isolation relative to 1000 Genomes Project populations within each subregion.

We tested the relationships among populations from the subregions using principal-component analysis (PCA). As expected, the first two principal components separated samples along well-established geographical axes: PC1 separated sub-Saharan Africans, and PC2 separated Eurasian populations (Supplementary Fig. 7). Populations from the GME subregions fell between the 1000 Genomes Project African, East Asian, and European populations, supporting the occurrence of recent admixture. PP and TP were closer to the East Asian populations, whereas NEA, NW A, and SD were closer to the sub-Saharan African populations. PC3 and PC4 separated samples along topographically north–south and east–west gradients, respectively, resulting in largely distinct but overlapping groups with a high degree of inter-regional diversity (Fig. 2a).

To test whether these populations were subject to bottlenecks, we calculated the mean rate of linkage disequilibrium (LD) decay for each, as LD should decay more slowly over haplotypes, as a function of size, in the presence of population bottlenecks (Fig. 2b). LD for each GME population decayed faster than it did in the European and East Asian populations but slower than in African populations. LD decayed faster in NW A and NEA than in GME populations from other regions, in agreement with our TreeMix results. Diverse patterns of admixture across these regions suggest that these trends are not predominantly due to intermixing but instead argue for a historical shared ancient bottleneck.

From 20–50% of all marriages in the GME are consanguineous (as compared with ∼0.2% in the Americas and Western Europe)\(^1\)–\(^3\), with the majority between first cousins. This roughly 100-fold higher rate of consanguinity has correlated with roughly a doubling of the rate of recessive Mendelian disease\(^19\)–\(^20\), European, African, and East Asian 1000 Genomes Project populations all had medians for the estimated inbreeding coefficient (\(F\)) of ~0.005, whereas GME \(F\) values ranged from 0.059 to 0.098, with high variance within each population (Fig. 2c). Thus, measured \(F\) values were approximately 10- to 20-fold higher in GME populations, reflecting the shared genomic blocks common to all human populations. \(F\) values were dominated by structure from the immediate family rather than historical or population-wide data trends (Supplementary Fig. 8). Examination of the larger set of 1,794 exomes that included many parent–child trios also showed an overwhelming influence of structure from the immediate family, with offspring from first-cousin marriages displaying higher \(F\) values than those from non-consanguineous marriages (Fig. 2d).

We expected that higher \(F\) values would correlate with an increased burden and length of ROHs, defined as homozygous haplotypes as a function of length\(^1\). The 1000 Genomes Project sub-Saharan African populations displayed the smallest total lengths for ROHs, as expected\(^22\), whereas ROH lengths in the two other populations assessed by the 1000 Genomes Project were relatively similar to those in the GME populations (Fig. 3a and Supplementary Fig. 9), probably reflecting similar numbers of short (<0.515 Mb) and medium-length (0.516–1.606 Mb) ROHs. Most striking was the increase in the burden of long ROHs (>1.607 Mb) found nearly exclusively in GME samples, especially when considering regions over 4 Mb in length (Fig. 3b). In the GME populations, there was an enrichment of rare and very rare variants (allele frequency <0.05 and <0.01, respectively) in longer ROHs and of common variants (allele frequency ≥0.05) in shorter ROHs (Fig. 3c), suggesting that the longer ROHs result from recent consanguinity\(^23\).

The increased length of ROHs provided an opportunity to identify homozygous loss-of-function variants in healthy humans. Although these variants are only putatively loss of function until experimentally verified\(^24\), they exhibit the strongest signs of selective pressure and are the first checked as disease candidates\(^25\). Recently, in 2,636 sequenced and 101,584 chip-imputed Icelanders, 1,171 genes were predicted to be inactivated\(^26\). In our 354 exomes from adults verified...
Figure 3  Distributions of short and long runs of homozygosity correlate with patterns of bottlenecks and recent consanguinity.  (a) Burden in samples of ROHs grouped by length (short, <0.155 Mb; medium length, 0.156–1.606 Mb; long, >1.607 Mb). GME samples (purple) showed a unique contribution from long ROHs in comparison to other populations (indicated by an asterisk), with fewer ROHs in the short and medium-length bins as compared to European and East Asian populations. Total ROH lengths in populations from the GME subregions overlapped with those in European and East Asian populations, likely because of greater bottlenecks in the latter populations.  (b) Histograms of the frequencies of long ROHs in the GME, African, European, and East Asian populations. GME samples more frequently harbored runs of >4 Mb than other populations. ROHs >15 Mb in length are binned together (an asterisk indicates a peak unique to the GME populations).  (c) Longer ROH spans in GME populations were enriched for rare variation, whereas shorter runs were enriched for more common variation. Probabilities are shown for variants binned by allele frequency as a function of ROH length (binned in 0.5-Mb intervals).  A probability density function was calculated for each allele frequency class. Note that allele frequencies for common alleles declined whereas allele frequencies for rare and very rare alleles rose as ROHs increased in length (common, frequency >0.05; rare, frequency = 0.01–0.05; very rare, frequency <0.01).

to be healthy, we found 301 genes with rare homozygous putative loss-of-function variants (Supplementary Tables 2 and 3), with only 50 genes overlapping with the Icelandic gene list. Similarly, the Exome Aggregation Consortium (ExAC) data set of 60,706 sequenced individuals identified 2,068 genes that were inactivated, of which only 94 genes overlapped with our 301 genes. These findings suggest that the set of non-clinically relevant loss-of-function variants is far from being complete. The GME population represents an optimal population in which to identify homozygous variants because of the elevated rates of consanguinity.  Darwin observed that rare self-fertilizing orchid strains exhibited fitness levels surprisingly higher than those of founder strains, and he termed these ‘hero strains’ (ref. 27). This observation led to the concept of purging of recessive alleles by Haldane28, referring to increased loss of deleterious alleles due to increased selective pressure in inbred populations. Purging was hypothesized to influence the genome in GME populations through higher rates of birth defects being incompatible with future reproduction29, but this has yet to be documented in humans. We compared the distributions of derived allele frequencies (DAFs) in GME and 1000 Genomes Project populations30. Variants were divided into seven functional deleterious classes. We calculated mean DAFs using chimpanzee (PanTro2) as the common ancestor (Supplementary Figs. 10 and 11)31. Neither autosomal nor X-linked variants showed substantial differences in frequency between the populations (Supplementary Fig. 12), arguing against a measurable effect of consanguinity on overall variant burden.

Numerous studies have relied on the increased power of consanguineous families from GME populations in identifying causes of recessive disease, but the lack of an accessible variome has hindered progress. Efforts like the National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project (ESP) produced variomes for European-American and African-American populations, but poor correlation of DAFs in pairwise comparisons determined that neither is a good estimator for DAFs in GME populations (Pearson’s r = 0.7979 for GME individuals versus European Americans, 0.385 for GME individuals versus African Americans, and 0.1447 for European Americans versus African Americans; Fig. 4a and Supplementary Fig. 13). Moreover, we found much of the variation in GME populations to be poorly represented outside the GME (Fig. 4b), with a majority of the rarest variants (lowest DAF bin) found only in the GME.

To assess how well the GME Variome captured extant exome variation, we subsampled the cohort over 100 iterations, selecting 5 to 700 individuals and considering variants from eight variant classes (Fig. 4c and Online Methods). There was decay in the number of unique variants and accumulation of rare variants as sample size increased, owing to the scaled ability to estimate prevalence. When close to 1,000 individuals were sampled, the change in mean numbers for these variants was negligible as new samples were added. Thus, the GME Variome should allow accurate determination of population-level DAFs for all but the rarest alleles.

Current methods for selecting candidate variants require a combination of function and deleteriousness prediction and allele frequency filtering32,33. To investigate the potential of the GME Variome to expedite the discovery of new disease-associated genes, we compared the causal variant sets among GME families displaying recessive hereditary spastic paraplegia (HSP), for which we recently established 17 new genetic forms of disease34. For a disease like recessive HSP with a prevalence of 3–10 cases per 100,000 (ref. 35) and where there are more than 40 genetic forms and hundreds of individual genetic mutations known, the expected allele frequency for any causative mutation should be <1:1,000 (Online Methods). Select individuals from 20
Variome to future sequencing projects for subjects originating from the GME, shaped by prehistoric as well as historic migrations, conquests, and Phenotypes (dbGaP) under accession phs000288.v1.p1. Methods and any associated references are available in the online version of the paper.

Accession codes. Allele frequencies for the GME Variome can be found in the online repository (see URLs). Exome data from the Gleeson laboratory have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession phs000288.v1.p1.
The authors declare no competing financial interests.

ACKNOWLEDGMENTS

The authors thank S. Sunyaev and D. Reich for help with PolyPhen-2 and DAF corrections. M. Turchin for help with purging analysis, J. Pickrell for help with TreeMix, and V. Bafna, N. Schork, and S. Bonnison for suggestions. Work was supported by grants from the US National Institutes of Health (P01HD070494 and R01NS048543), the Qatar National Research Foundation ( NPRP6-1-163), the Simons Foundation Autism Research Initiative (105525 to J.-L.C.), the World Health Organization (to J.G.G.), the Yale Center for Mendelian Disorders (U54HG006504), the Broad Institute (U54HG003067), the Rockefeller University CTSU (SULRIR024134-03), the Howard Hughes Medical Institute (to J.G.G. and J.-L.C.), INSERM, the St. Giles Foundation, and the CANDISID Association and by grants R01AI088364, R37AI099593, P01AI061903, U01AI109697 (to J.-L.C.), U01AI088685 (to J.-L.C. and L.A.), R21AI107508 (to E. Jouanguy), the DHFMR Collaborative Research Grant, and KACST 13-BIO1113-20 (to F.S.A.).

Greater Middle Eastern Variome Consortium. Sohail Abdel Rahim, Sawsan Abdel-Hadi, Ghada Abdel-Salam, Ekram Abdel-Salam, Mohammed Abdou, Aivinis Ahbytankar, Parisa Adimi, Jamal Ahmad, Mustafa Afkacika, Guskide Aksu, Sami Al Hajjar, Suliman Al Izzaaah, Saleh Al Muhsen, Nouriya Al Sannaa, Avinash Abhytankar, Parisa Adimi, Jamil Ahmad, Mustafa Akcakus, Guside Aksu, Asma Al-Tawari, Abdullah Alangari, Alexandre Alcais, Tariq S AlFawaz, Lihadh Al-Gazali, Amal Al-Hashem, Waleed Al-Herz, Deema Al-Jeaid, Zeynep Peker Koc, Isabelle Pellier, Capucine Picard, Anne Puel, Tarek Omar, Azza Oraby, Ayse Pandaluz, Nima Parvaneh, Turkan Patiroglu, Reem A Elfeky, Gehad Elghazali, Eissa Faqeih, Elif Fenerci, Claire Fieschi, Catherinot, Fatma C Celik, Michael Ciancanelli, Funda E Cipe, Gary Clark, Aurélie

CORRECTIONS Flow of text from North Africa contributes to differential human genetic diversity in southern Europe. Proc. Natl. Acad. Sci. USA 110, 11791–11796 (2013).

Li, J.Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100–1104 (2008).

Henn, B.M. et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 8, e1002397 (2012).

Gérard, N. et al. Can the impact of human genetic variations be predicted? J. Exp. Med. 208, 710–720 (2011).

Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507, 354–357 (2014).

SIGMA Type 2 Diabetes Consortium. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 506, 97–101 (2014).

Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from whole exome sequencing data: state of the art and perspectives for clinical, population and epidemiological studies. Hum. Hered. 77, 63–72 (2014).

Pemberton, T.J. et al. Genomic patterns of homopropy in worldwide human populations. Am. J. Hum. Genet. 91, 275–292 (2013).

Spiecz, Z.A. et al. Runs of homopropy are enriched for deleterious variation. Am. J. Hum. Genet. 93, 102–103 (2013).

Iran, Y. & Casanova, J.L. Can the impact of human genetic variations be predicted? Proc. Natl. Acad. Sci. USA 112, 11426–11427 (2015).

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

Sulem, P. et al. Identification of a large set of rare complete human knockouts. Nat. Genet. 47, 448–452 (2015).

Jones, S. The Darwin Archipelago. (Yale University Press, 2011).

Haldane, J.B.S. The effect of variation of fitness. Proc. Natl. Acad. Sci. USA 112, 11426–11427 (2015).
**ONLINE METHODS**

**Definition of the Greater Middle East.** The term Greater Middle East has been used to refer to a large swath of Arab and non-Arab countries, stretching from Morocco in the west to as far east as Pakistan in southeast Asia. However, no precise listing of designated countries has yet emerged (see URLs).

**Exome resequencing. Study sample.** The 2,497 individuals used in the analysis were selected from samples ascertained across three laboratories and recruited with the help of the clinicians who constitute the GME Consortium. Although these individuals were not a random sample, they were ascertained in the context of a wide variety of distinct phenotypes, such that cohort-specific effects were not expected to bias patterns of variation. All study participants in each of the component studies provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing. Institutional certification was obtained for each sample to allow deposition of genotype data in dbGaP and for other purposes. Institutional review board approval was obtained from the Rockefeller University, King Faisal Specialist Hospital, and Necker Hospital for Sick Children.

**Exome resequencing, variant calling, and filtering.** Blood DNA was extracted using Qiagen reagents, subjected to exome capture with the Agilent SureSelect Human All Exome 50Mb kit, and sequenced on an Illumina HiSeq 2000 instrument, resulting in ~94% target coverage at >30x depth36,38. FASTQ files were reprocessed and jointly called to minimize batch effects and ensure consistent variant calling, using the Genome Analysis Toolkit (GATK) pipeline (version 3.1-1) adhering to best practices39, eliminating duplicate reads. Paired-end reads were aligned to human reference genome NCBI Build 37 using BWA (version 0.7.5)40. PCA was run on the resultant set of variants to identify potential batch effects between laboratories, sequencing centers, or identified statistical outliers. The metrics included the total number of variants were observed. The PCA-based outlier analysis considered samples grouped by geographical region independently. Samples and the number of singletons. Because of possible reference distance bias, we identified no additional kinship. The final continental sample counts after unrelated individuals. Remaining samples were rerun through KING, which were removed to reduce data set relatedness, leaving a final cohort of 1,516 samples. Of the remaining 2,183 samples after outlier filtering, 667 samples were removed to reduce data set relatedness, leaving a final cohort of 1,516 unrelated individuals. Remaining samples were rerun through KING, which identified no additional kinship. The final continental sample counts after filtering were as follows: sub-Saharan Africa, 19; Americas, 33; Europe, 378; Oceania, 1; and GME, 1,111.

Coverage statistics were generated across all internal exome data sets using BEDTools, to calculate the average coverage across each exon43. Exons were filtered from the analysis if greater than 5% of samples had less than 10x average coverage. Of the initial 192,056 exons targeted by the Agilent SureSelect II capture kit, 170,032 exons were well covered in at least 95% of samples. Variants were filtered out if identified outside of these genomic regions, leaving 32,967,859 bases under consideration (~1% of the human genome) within 17,800 genes.

Standard filters were applied to include variants that were called with posterior probability >99% (gldMultiples SNP quality >20), were at least 5 bp away from an indel detected in the 1000 Genomes Project Pilot Project, and were targeted in at least 95% of individuals. Variant positions were filtered on the basis of population statistics, including a missingness rate (referring to the percentage of samples where information was missing) of less than 5% and a P value for deviation from Hardy–Weinberg equilibrium of <0.00005 (ref. 44).

We generated a subset of variants in minimal LD by pruning variants exhibiting pairwise LD (r^2)>0.5. Variants were filtered to exclude SNPs with minor allele frequency (MAF) <5% and all indels. Remaining SNPs were pruned, adhering to a maximum threshold of 1,000, using the –indep-pairwise command in PLINK44. Of the initial 578,231 variants, 182,967 SNPs passed the filters. This LD-pruned data set was used for the characterization of population structure, including PCA, Wright’s fixation index (FST) measurement, admixture analysis, KING relationship testing, and estimation of the inbreeding coefficient (F).

**Geographical region assignment.** Samples were recruited from 20 countries and territories across the GME and grouped into a set of six geographical regions: NWA (85 samples), NEA (423 samples), AP (214 samples), SD (81 samples), TP (140 samples), and PP (168 samples). Country boundaries were not used to group samples for two reasons: (i) inconsistent sampling left several countries with too few samples to accurately represent the diversity of the population and (ii) current country borders frequently fail to accurately separate ancestry groups.

Information on self-identified ancestry was available for some samples, but the incompleteness of this annotation and the great diversity in populations affiliating as Arab prompted the use of geography for grouping. As much as possible, we assigned location to the current residence, rather than ancestral residence or location where samples were drawn. Although some reference GME ancestry groups exist in public resources, such as the Human Genome Diversity Project (HGDP)45, we found both the breadth of ascertained ancestry groups and sample size insufficient to imbibe ancestry groups where absent. To ensure consistency in our geographical designations, we performed linkage clustering based on pairwise distances between samples using the –distance-matrix command in PLINK44. We performed hierarchical clustering on all samples using Ward’s hierarchical clustering method (ward.D2 option for the hclust algorithm in R)46.

**Population structure of the Greater Middle East. Data integration.** Population structure was analyzed in the context of continental populations from the 1000 Genomes Project Phase I data set47. As 1000 Genomes Project samples were generated from a combination of whole-genome and exome sequencing, variants falling outside of ReSeq exonic regions by ≥30 bp were filtered out using BEDTools and remaining variants were merged with those in the GME cohort47,48. Nine populations from 1000 Genomes Project data were used in comparative analyses: African populations YRI and LWK; East Asian populations CHB, CHS, and JPT; and European populations GBR, TSI, IBS, and FIN. Related 1000 Genomes Project samples were filtered out using KING analysis as previously described. A total of 1,821 samples remained after filtering, representing 15 geographical regions—6 from the GME and 9 from the 1000 Genomes Project.

**Substructure analysis.** To investigate the influence of admixture on the GME samples, we used the block relaxation algorithm implemented in ADMIXTURE to estimate individual ancestry proportions given k ancestral populations49. Unsupervised ADMIXTURE analysis was run using default settings (k = 5) on merged variants from GME and 1000 Genomes Project samples with iterations of k from 2 to 14. Minimum squared error values calculated from the cross-validation procedure in ADMIXTURE to evaluate the fit of different values of k determined that k = 6 was optimum for only GME samples and k = 7 was optimum when 1000 Genomes Project control data were included.

**Principal-components analysis and Wright’s fixation index.** PCA was used to investigate the affinities in human populations and the relationships between them. We performed PCA on GME and 1000 Genomes Project samples using the SmartPCA tool from the EIGENSOFT software library, and the first four principal components were compared graphically47,50. Wright’s fixation index (FST) was used to explore the degree of differentiation between populations. FST values and standard errors for all population pairs were calculated using the estimator of Weir and Cockerham, also included in the EIGENSOFT software library. All plots were generated using ggplot2 (ref. 51).

**Linkage disequilibrium decay.** The LD between pairs of SNPs is an indicator of the past history of recombination and genetic drift. To calculate LD, we tallied pairwise r^2 values for SNP pairs in all GME and control populations using the r2 option in PLINK44. Correlations among all SNPs in each 70-kb
sliding window were calculated with no lower limit on $r^2$ values. Pairwise correlations were binned by genomic distance between the SNPs (up to 70 kb), and averages were calculated for each bin. Control samples followed expected patterns of LD decay.

**Estimation of inbreeding.** The inbreeding coefficient of an individual ($F$) was used to represent the probability that two randomly chosen alleles at a homologous locus in an individual were identical by descent (IBD) with respect to a base reference population in which all alleles were independent. Although the true inbreeding coefficient of an individual is often unknown, several estimation methods have been shown to give a reasonable estimate.

Estimates of $F$ value were calculated using the het algorithm in PLINK on LD-pruned variants following the authors’ guidelines. We compared the results to those from the HMM algorithm Festim and found that the two estimates were very similar (Pareion’s $r = 0.874$), but the Festim algorithm frequently failed to return results for samples with missing data. Negative $F$ values were most likely the result of either biased variant sampling, or recent intermixing of previously disparate populations.

**Estimation of runs of homozygosity.** To infer estimates of autozygosity and relative recent population size, we estimated ROHs using the HMM algorithm H3M2. We compared the results to those from the HMM algorithm Festim and found that the two estimates were very similar (Pareion’s $r = 0.874$), but the Festim algorithm frequently failed to return results for samples with missing data. Negative $F$ values were most likely the result of either biased variant sampling, or recent intermixing of previously disparate populations.

**Variant annotation and classification.** Variant annotation. Functional annotation was performed for analyses of genetic purging and loss of function. Variants were annotated using the ANNOVAR suite of scripts (version 2014Nov12). ANNOVAR classified variants into eight coding-region values in defining the classes. Functional designations for PolyPhen-2 (ref. 55). The functional designations for PolyPhen-2 location within 2 bp of a splice junction, either on the intronic or exonic side. Variants causing splicing defects were identified by their classified as ‘unknown’. We compared these annotations to those generated by SnpEff.

Ancestral allele identification. We used the chimpanzee genome as the closest assembled outgroup genome. Ancestral allele estimates were obtained by pairwise alignments in the UCSC Genome Browser between the hg19 human reference sequence and the PanTro2 and PanTro4 chimpanzee reference sequences. Systematic lookups for all GME and 1000 Genomes Project variants were performed using UCSC Genome Browser tools and custom scripts to identify associated chimpanzee alleles. We compared PanTro2 and PanTro4 to assess their difference in correcting apparent reference bias but found that both worked equally well.

Estimated ancestral alleles were used as the reference allele to calculate DAFs. DAFs were not calculated for variants where the ancestral allele was not present in the human germ line.

Identity-by-state distance to the reference. To interrogate potential biases that might result from reference selection, we calculated the IBS distance between samples and multiple different references, including hg19 and chimpanzee sequences. The distance represents the proportion of positions that diverge from the reference and was calculated for all pairs of samples and references.

The IBS distance $d$ represents the number of alleles differing between the two samples divided by the total number of alleles compared. More formally, $d$ for two n-length vectors $p$ and $q$ (in our case, $p$ is the reference sample and $q$ is the sample being compared) in a vector space $v$, where $v = [0,1,2]$, corresponding to the homozygote for the human reference allele, the heterozygote, and the homozygote for the alternate allele, respectively.

For any two samples, we calculate $d$ as

$$d(p,q) = \sum_{i=1}^{n} \left| p_i - q_i \right| / n$$

where $(p, q)$ are vectors such that $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$. Each vector represented all genotype calls between the two samples, excluding filtered sites or missing positions.

The IBS distance was calculated for all GME and 1000 Genomes Project samples against the hg19 and chimpanzee reference genomes. All genotypes from the merged VCF file were coded on the basis of comparison to the hg19 reference sequence. Variant positions were filtered to remove indels, because of the possibility of alignment errors, and non-biallelic sites. When comparing to hg19, vector $p$ was represented by a vector of zeros.

**Hereditary spastic paraplegia candidate variant variant analysis.** Samples from 20 consanguineous families displaying an autosomal recessive inheritance pattern for HSP were selected from a previously analyzed cohort of 55 families because there was a single genetic cause identified in these 20 families. Families were analyzed according to published methods.

Homozygous variants were filtered out on the basis of family structure to ensure that variants segregated with the disease phenotype. We performed deleterious filtering using functional classes and GERP++ scores. All candidate variants were potentially loss of function (frameshift, stop gain or loss, or perturbing splicing) or a coding variant with a GERP score >4.

The maximum allele frequency for candidate variants was based on established rates of disease prevalence, estimated at 1:10,000 for clinical presentations classified as HSP. Approximately 50% of HSP is autosomal dominant, and about 50% of the remaining incidence is explained by mutations in $SPG11$ (ref. 59), leaving only 1 in 40,000 with recessive HSP caused by mutations in other genes. Mutation of at least 35 other genes is reported to cause recessive HSP. Thus, the contribution to HSP disease prevalence for any given gene is unlikely to be more than 1:1,000,000. Although the prevalence of HSP-associated mutations is not expected to be uniform, we expect the maximum carrier frequency for any new causal variant to be no more than 1:1,000, assuming full penetrance and classic recessive inheritance; in actuality, the frequency is likely to be much lower given allelic diversity.

With roughly 1,100 individuals in our cohort, we calculated that variants with DAFs <1:1,000 should not be commonly observed in our data set, allele frequencies <1:500 should not be observed in more than one individual, allele frequencies <1:333 should not be observed in more than two individuals, and allele frequencies <1:250 should not be observed in more than three individuals. Variants passing deleterious and allele frequency thresholds were treated as candidates to calculate the usefulness of the GME Variome in limiting the number of deleterious variants considered as candidates.

**Testing for the influence of genetic purging.** Consanguinity has been practiced in the GME for at least several centuries. Simulations of GME-like populations have found that sufficient time has passed for purging to have been effective in reducing genetic load. Clinical studies aimed at comparing clinical rates of birth defects, premature births, and miscarriages between communities that practice consanguinity and largely outbreeding populations have found that all metrics fall within the range for the rate expected for the immediate form of consanguinity. More recent genetic studies investigating differential selective pressure across human populations focused on the role of population bottlenecks, neglecting the potential influence of consanguinity, and lacked representation from the GME.

To approach the question of variable selective pressure across human populations, we implemented a variation of the DAF comparison method. We assumed that any change in the efficacy of natural selection should be evident across populations in the mean DAF for each variant class.

For all variants described across the GME and 1000 Genomes Project populations, we filtered for high-quality calls, identified ancestral alleles, annotated predicted function and PolyPhen-2 classes using ANNOVAR, and downsampled to achieve an equivalent number of chromosomes across populations; we then calculated DAFs for all positions. Variants were grouped by class, and mean DAFs were calculated for each
population. Standard errors were calculated by bootstrapping mean DAFs for 1,000 iterations.

Recent studies using PolyPhen-2 demonstrated a deflation in deleteriousness scores for derived variants found in the hg19 reference genome1,61. Bias correction was implemented by grouping variants by DAF bin and calculating the proportion of variants in each PolyPhen-2 class per bin for ancestral reference positions; all derived reference positions were randomly reassigned to a new PolyPhen-2 class on the basis of a hypergeometric distribution within each DAF bin62.

**Neanderthal and Denisovan introgression analysis.** Neanderthal-derived variants are often subject to strong negative selection, thereby making exome analysis inadequate in estimating the age of introgression. Thus, we calculated Neanderthal and Denisovan introgression analysis. Neanderthal-derived reference genomes. Alignments of Neanderthal and Denisovan genomes to reference positions; all derived reference positions were randomly reassigned to a new PolyPhen-2 class on the basis of a hypergeometric distribution within each DAF bin62.

To estimate introgression, we identified aligned consensus calls for all human variant positions from the chimpanzee, Neanderthal, and Denisovan reference genomes. Alignments of Neanderthal and Denisovan genomes to 1000 Genomes Project variant positions were downloaded from the Max Planck Institute for Evolutionary Anthropology13,64. Chimpanzee alleles were in the ‘ancestral allele’ section of these methods.

We projected GME and 1000 Genomes Project control populations on the principal components calculated using representative samples from Neanderthal, Denisovan, and chimpanzee16,66,67 and aligned the human samples to these ancestral populations. Principal components were computed using the prcomp function in R (see URLs), and projected vectors were calculated for all 1000 Genomes Project and GME samples. Distance from the readjusted origin to each species reflects the proportion of introgression calculated for all 1000 Genomes Project and GME samples. Distance from the readjusted origin to each species reflects the proportion of introgression observed in each sample. Similar to previous work, European, East Asian, and GME populations overlapped and demonstrated larger proportions of Neanderthal introgression than African populations66–68.

---

36. Dixon-Salazar, T.J. et al. Exome sequencing can improve diagnosis and alter patient management. *Sci. Transl. Med.* 4, 138ra78 (2012).
37. Okada, S. et al. Impairment of immunity to *Candida* and *Mycobacterium* in humans with bi-allelic *RORC* mutations. *Science* 349, 606–613 (2015).
38. Alsalem, A.B., Halees, A.S., Anazi, S., Alshamekh, S., Alkuraya, F.S. Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. *PLoS Genet.* 9, e1004030 (2013).
39. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498 (2011).
40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595 (2010).
41. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* 2, e193 (2006).
42. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873 (2010).
43. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).
44. Purrini, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
45. Cann, H.M. et al. A human genome diversity cell line panel. *Science* 296, 261–262 (2002).
46. Behar, D.M. et al. The genome-wide structure of the Jewish people. *Nature* 466, 2199–2202 (2010).
47. Danzeck, P. et al. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158 (2011).
48. Pruitt, K.D. et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 42, D756-D763 (2014).
49. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664 (2009).
50. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909 (2006).
51. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).
52. Polasek, O. et al. Comparative assessment of methods for estimating individual genome-wide homozgyosity-by-descent from human genomic data. *BMC Genomics* 11, 139 (2010).
53. Magi, A. et al. H3M2: detection of runs of homozgyosity from whole-exome sequencing data. *Bioinformatics* 30, 2852–2859 (2014).
54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164 (2010).
55. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249 (2010).
56. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92 (2012).
57. Davydov, E.V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6, e1001025 (2010).
58. Erichsen, A.K., Kohl, J., Stray-Pedersen, A., Abdelnoor, M. & Tallaksen, C.A. Prevalence of hereditary ataxia and spastic paraplegia in southeast Norway: a population-based study. *Brain* 132, 1577–1588 (2009).
59. Stevanin, G. et al. *Mutations in SPG4 are frequent in autosomal recessive spastic paraplegia with thin corpus callosum, cognitive decline and lower motor neuron degeneration.* *Brain* 131, 772–784 (2008).
60. Vardi-Saliternik, R., Friedlander, Y. & Cohen, T. Consanguinity in a population sample of Israeli Muslims. *Arabian* 29, 422–431 (2002).
61. Shami, S.A., Qasim, R. & Bittles, A.H. Consanguinity and adult morbidity in Pakistan. *Lancet* 338, 954 (1991).
62. Stoltenberg, C., Magnus, P., Lie, R.T., Daltveit, A.K. & Irgens, L.M. Birth defects and parental consanguinity in Norway. *Am. J. Epidemiol.* 145, 439–448 (1997).
63. Do, R. et al. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* 47, 126–131 (2015).
64. SIGMA Type 2 Diabetes Consortium. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population. *J. Am. Med. Assoc.* 311, 2305–2314 (2014).
65. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338, 222–226 (2012).
66. Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197 (2014).
67. Wang, S., Lachance, J., Tishkoff, S.A., Hey, J. & Xing, J. Apparent variation in the readjusted origin to each species reflects the proportion of introgression calculated for all 1000 Genomes Project and GME samples. Distance from the readjusted origin to each species reflects the proportion of introgression observed in each sample. Similar to previous work, European, East Asian, and GME populations overlapped and demonstrated larger proportions of Neanderthal introgression than African populations66–68.