The genetic structure of the Turkish population reveals high levels of variation and admixture

M. Ece Kars[a], A. Nazlı Başak[b], O. Emre Onat[a,1], Kaya Bilguvar[c], Jungmin Choi[c,d], Yuval Itan[e,f], Caner Çağlar[g,2], Robin Palvadeau[b], Jean-Laurent Casanova[h,i,j,k,l], David N. Cooper[m], Peter D. Stenson[m], Alper Yavuz[n], Hakan Buluş[o], Murat Günel[c,p], Jeffrey M. Friedman[g,l], and Tayfun Özçelik[a,q,r,3]

[a]Department of Molecular Biology and Genetics, Bilkent University, 06800 Ankara, Turkey; [b]Suna and Inan Kıraç Foundation, Neurodegeneration Research Laboratory, Research Center for Translational Medicine, Koç University School of Medicine, 34450 Istanbul, Turkey; [c]Department of Genetics, Yale Center for Genome Analysis, Yale University School of Medicine, New Haven, CT 06510; [d]Department of Biomedical Sciences, Korea University College of Medicine, 02841 Seoul, Korea; [e]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [f]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [g]Laboratory of Molecular Genetics, Rockefeller University, New York, NY 10065; [h]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, NY 10065; [i]Laboratory of Human Genetics of Infectious Diseases, Necker Branch INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France; [j]Imagine Institute, University of Paris, 75015 Paris, France; [k]Pediatric Immunology-Hematology Unit, Necker Hospital for Sick Children, 75015 Paris, France; [l]HHMI, Rockefeller University, New York, NY 10065; [m]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, United Kingdom; [n]Department of Surgery, Keçiören Training and Research Hospital, 06000 Ankara, Turkey; [o]Keçiören Training and Research Hospital, Department of Surgery, University of Health Sciences, 06000 Ankara, Turkey; [p]Department of Neurosurgery, Yale University School of Medicine, New Haven, CT 06510; [q]Neuroscience Program, Graduate School of Engineering and Science, Bilkent University, 06800 Ankara, Turkey; and [r]Institute of Materials Science and Nanotechnology, National Nanotechnology Research Center, Bilkent University, 06800 Ankara, Turkey.

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved July 13, 2021 (received for review December 18, 2020)

The construction of population-based variomes has contributed substantially to our understanding of the genetic basis of human inherited disease. Here, we investigated the genetic structure of Turkey from 3,362 unrelated subjects whose whole exomes (n = 2,589) or whole genomes (n = 773) were sequenced to generate a Turkish (TR) Variome that should serve to facilitate disease gene discovery in Turkey. Consistent with the history of present-day Turkey as a crossroads between Europe and Asia, we found extensive admixture between Balkan, Caucasus, Middle Eastern, and European populations with a closer genetic relationship of the TR population to Europeans than hitherto appreciated. We determined that 50% of TR individuals had high inbreeding coefficients (≥0.0156) with runs of homozygosity longer than 4 Mb being found exclusively in the TR population when compared to 1000 Genomes Project populations. We also found that 28% of exome and 49% of genome variants in the very rare range (allele frequency < 0.005) are unique to the modern TR population. We annotated these variants based on their functional consequences to establish a TR Variome containing alleles of potential medical relevance, a repository of homozygous loss-of-function variants and a TR reference panel for genotype imputation using high-quality haplotypes, to facilitate genome-wide association studies. In addition to providing information on the genetic structure of the modern TR population, these data provide an invaluable resource for future studies to identify variants that are associated with specific phenotypes as well as establishing the phenotypic consequences of mutations in specific genes.

Turkish Variome | admixture | sequencing | population genetics | variation

Even in the Paleolithic period, Anatolia (or Asia Minor as it was once called) served as a bridge for migrations between Africa, Asia, and Europe. Long before the establishment of nation states, intermixing between human populations occurred in Anatolia. Indeed, Anatolia has been home to many civilizations including Hattians, Hurrians, Assyrians, Hittites, Greeks, Thracians, Phrygians, Urartians, Armenians, and Turks. Gene flow between Anatolian, Caucasus, and northern Levantine populations occurred during the Late Neolithic and Chalcolithic to the Early Bronze age, including long-distance migration from Central Asia to Anatolia (1). The Turkic peoples, a collection of ethnolinguistically related populations originating from Central Asia, were first documented in western Eurasia in the fourth/fifth century BCE and currently live in Central, Eastern, Northern, and Western Asia as well as in parts of Europe and in North Africa. The expansion of Turkic tribes into Western Asia and Eastern Europe occurred between the sixth and 11th centuries, beginning with the Seljuk Turks followed by the Ottomans (2). The sphere of Ottoman influence started to increase greatly, beginning in the 14th century; following the conquest of Constantinople in 1453, the Ottoman Empire controlled a vast region including all of southeastern Europe south of Vienna, parts of Central Europe, Western Asia, the Caucasus, North Africa, and the Horn of Africa. The modern Republic of Turkey was founded in 1923 after the fall of the Ottoman Empire at the end of World War I and is currently home to more than 80 million people. Turkish-speaking people constitute the major ethnolinguistic group in Turkey. There are also more than 70 million

Significance

We delineated the fine-scale genetic structure of the Turkish population by using sequencing data of 3,362 unrelated Turkish individuals from different geographical origins and demonstrated the position of Turkey in terms of human migration and genetic drift. The results show that the genetic structure of present-day Anatolia was shaped by historical and modern-day migrations, high levels of admixture, and inbreeding. We observed that modern-day Turkey has close genetic relationships with the neighboring Balkan and Caucasus populations. We generated a Turkish Variome which defines the extent of variation observed in Turkey, listed homozygous loss-of-function variants and clinically relevant variants in the cohort, and generated an imputation panel for future genome-wide association studies.

Author contributions: M.E.K., J.M.F., and T.Ö. designed research; M.E.K., A.N.B., O.E.O., K.B., J.C., Y.I., C.Ç., R.P., and T.Ö. performed research; A.N.B., K.B., Y.I., J.-L.C., D.N.C., P.D.S., A.Y., H.B., and M.G. contributed new reagents/analytic tools; M.E.K. analyzed data; and M.E.K., J.M.F., and T.Ö. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]Present address: Department of Genome Studies, Institute of Health Sciences, Acıbadem Mehmet Ali Aydınlar University, 34752 Istanbul, Turkey.

[2]Present address: Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, 34820 Istanbul, Turkey.

[3]To whom correspondence may be addressed. Email: tozcelik@bilkent.edu.tr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2026076118/-/DCSupplemental.

Published August 23, 2021.

people who live in the five independent Turkic countries in Central Asia, namely Azerbaijan, Turkmenistan, Uzbekistan, Kazakhstan, and Kyrgyzstan. A study investigating the Y haplogroups of Turkish (TR) males revealed that the proportion of recent paternal gene flow from Central Asia was ~9%, thereby raising the possibility that modern-day Anatolia is an admixture of preexisting Anatolian and Turkic peoples (3).

The practice of consanguineous marriage is frequent in Turkey, especially in the eastern provinces (4). This should, in principle, help to facilitate disease gene discovery, as the increased frequency of homozygosity among members of inbred populations has led to the identification of many disease genes (5–8). The genetic admixture and consanguinity have had a significant effect on the genetic diversity of Middle Eastern populations (9, 10). The characterization of the Greater Middle East (GME) Variome, comprising the most comprehensive genomic database for Middle Eastern populations, has shown that knowledge of the genomic architecture of these populations facilitates disease gene identification in family studies and in genome-wide association studies (GWAS) of populations (11). Until now, the GME has been the largest resource representing the genetic variation in Turkey, albeit with only 140 out of a total of 1,111 samples coming from the TR Peninsula. Thus, based on the larger population of Turkey relative to its immediate neighbors, the TR population is underrepresented in current genomic databases. Furthermore, gnomAD, as one of the most comprehensive genetic variation resources, does not contain TR whole-exome sequencing (WES) or whole-genome sequencing (WGS) data (12). Therefore, a comprehensive database of alleles in the TR population should facilitate disease gene identification in consanguineous families and the assessment of the clinical phenotypes of individuals who are homozygous for mutations in specific genes.

Finally, most GWAS to date have analyzed DNA from European ancestry–derived populations, and it will be important to extend the GWAS of complex traits to underrepresented populations. One of the key steps in GWAS is to "predict" or "impute" the missing genotypes by using a reference haplotype panel. It is becoming increasingly common for researchers to generate such panels for imputation from population-specific WGS data. Population-specific reference panels increase imputation accuracy, especially when they are combined with existing reference panels such as the 1000 Genomes Project (1000GP) (13–15). In this study, we have described the high-resolution genetic structure of the TR population, generated a TR Variome, and imputed a TR reference panel for future genetics studies.

Results

Population Structure of Turkey. WES and WGS data from 3,864 individuals who had participated in genetic studies of amyotrophic lateral sclerosis, ataxia, delayed sleep phase disorder, essential tremor, obesity, Parkinson's disease, polycystic ovarian syndrome, and various assorted neurological and immunological disorders (*SI Appendix*, *Study Samples* and Table S1). WES and WGS samples were processed separately for analyses of the sample quality and familial relationships. A total of 2,589 WES and 773 WGS samples remained after filtration according to quality metrics and relatedness (*SI Appendix*, Table S2 and Figs. S1 and S2). Following a variant filtration process to identify high-quality variants, we obtained 1,123,248 WES and 45,981,721 WGS variants (*SI Appendix*, *Sequencing and Filtering* and Table S3).

The geographical origins of ancestors (birthplaces of maternal and paternal grandparents) of 1,460 TR samples were documented and grouped into six different subregions, namely Balkan (TR-B: 90), West (TR-W: 157), Central (TR-C: 441), North (TR-N: 372), South (TR-S: 116), and East (TR-E: 284) (13). First, we performed a principal component (PC) analysis (PCA) using only TR individuals of known origin (*SI Appendix*, Fig. S3). There were no sharp divisions between TR subregions, yet the position of subregions along PC axes was similar to their geographical location. To evaluate the impact of geography in shaping the genomic variability in Turkey, we tested the correlation between geographic and genetic coordinates by applying a Procrustes analysis. Consistent with the results of the PCA, we did not observe a clear-cut distribution of samples among TR subregions, although we did detect a significant mild positive correlation in our dataset (Fig. 1*A*, correlation in Procrustes rotation, 0.49 $P < 1 \times 10^{-5}$).

We used the populations from Lazaridis et al. and 1000GP to evaluate the genetic differentiation of the TR population on a global scale (*SI Appendix*, Table S4) (16). First, we compared the TR population with eight superpopulations using PCA: Africa (AFR), Europe (EUR), Balkan (BLK), Caucasus (CAU), GME, South Asia (SAS), Central and North Asia (CNA), and East Asia (EAS). EAS, CNA, AFR, and SAS populations were distinguished with PC1, PC2, and PC3, while the other populations displayed an east to west cline in PC4 (Fig. 1*B* and *SI Appendix*, Fig. S4).

We then evaluated the genetic substructure of Turkey using ADMIXTURE, and $k = 4$ was determined as the lowest cross-validation error (*SI Appendix*, Fig. S5*A*) (17). Individuals with unknown ancestral birthplaces (TR-U) exhibited similar ancestral components to those individuals with known ancestral birthplaces. All four ancestries were represented in each geographical region, although in different proportions (*SI Appendix*, Fig. S6). When we employed ADMIXTURE using the global dataset, we noted four major ancestral components, which were predominantly found in EUR, BLK, CAU, and GME populations, formed the genetic substructure of the TR population (Fig. 1 *C* and *D* and *SI Appendix*, Figs. S5*B* and S7). The primary ancestral components of the EUR and BLK populations were remarkably higher in the TR-B and TR-W, whereas the shared ancestry of the TR, CAU, and non-Arab populations of GME increased in the east direction for the TR subregions. The proportion of the ancestral contributions among the TR subregions reflects the importance of geographical location in shaping genetic substructure. Additionally, we calculated the Central Asian contribution to the modern-day TR population as 9.59% based on ADMIXTURE results. This contribution varied for the TR subregions: TR-B, 7.69%; TR-W, 12%; TR-C, 10.1%; TR-N, 10.6%; TR-S, 11.2%; and TR-E, 6.48%.

To further evaluate the genetic relationship in a regional context, we performed a third PCA using a regional dataset that includes the populations closely related to the TR population (*SI Appendix*, *Population Structure Analyses*). Importantly, consistent with a high level of admixture, the degree of variation observed in the TR population was much higher compared to other populations (Fig. 1*E*). As expected, we observed that the genetic connection of European and TR populations was established through BLK (TR-B) and Western Turkey (TR-W), while the links between TR-CAU and TR-GME populations were formed by the other TR subregions, which further emphasize the importance of geography on the genetic variation seen in Turkey (*SI Appendix*, Fig. S8). We tested the correlation between geographic and genetic coordinates of the populations included in the regional dataset by applying a Procrustes analysis and detected a strong positive correlation (*SI Appendix*, Fig. S9, correlation in Procrustes rotation, 0.76 $P < 1 \times 10^{-5}$).

The position of Turkey along historical routes of migration and the effect of genetic drift was assessed using a maximum likelihood phylogenetic tree with the inclusion of the 1000GP and the GME populations (Fig. 2*A*) (18). The clusters of each 1000GP and GME populations were recapitulated with the inferred tree, and Turkey connected the GME and European branches. When the populations were ordered from the root, the ordering corroborated the "out-of-Africa" hypothesis and

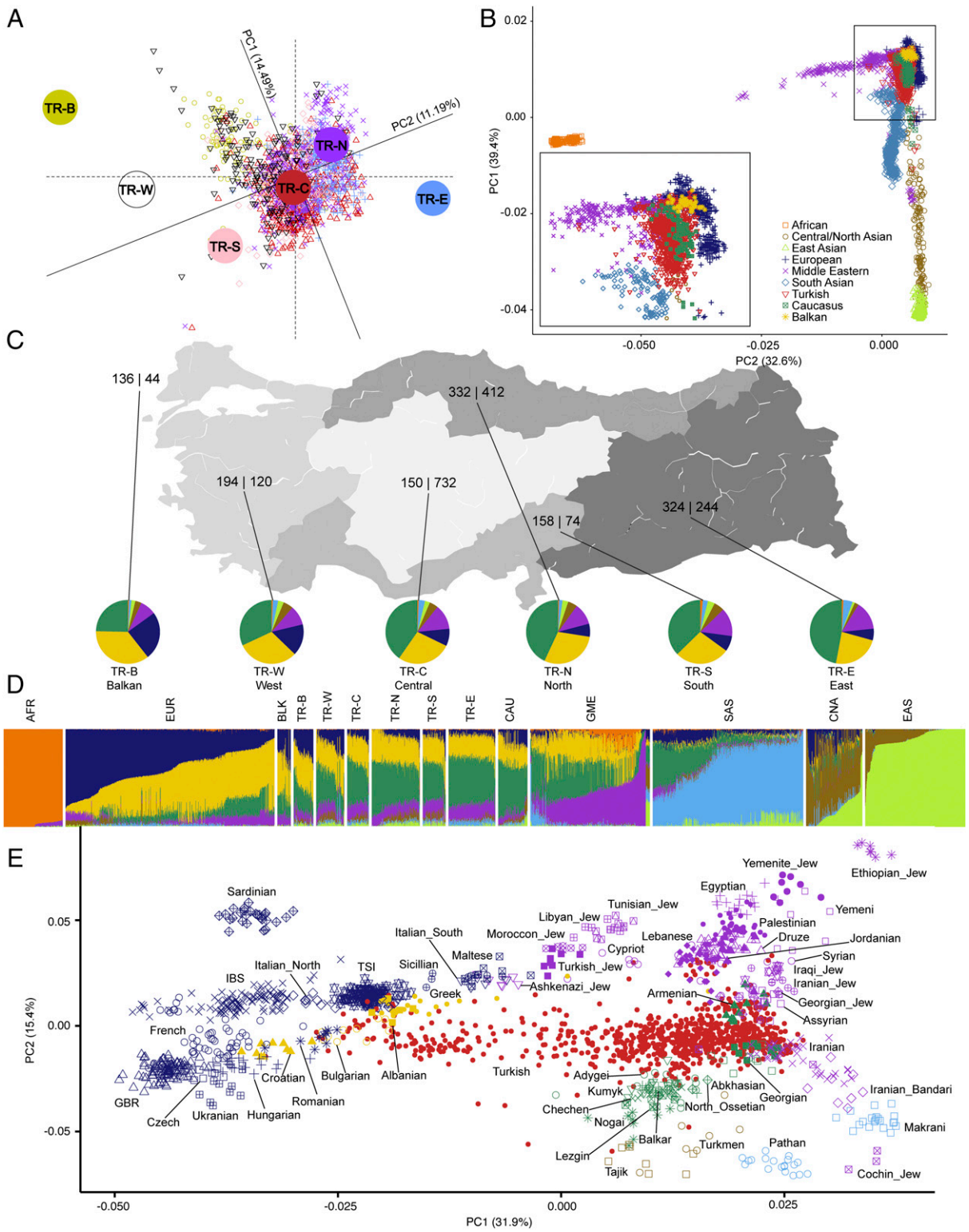

Fig. 1. TR as a hub of extensive human admixture. (*A*) Procrustes analysis based on unprojected coordinates of geographical locations and PC1 and PC2 coordinates of 1,460 TR individuals with known origin. (*B*) PCA for individuals from the TR WGS ($n = 773$), the populations from Lazaridis et al. ($n = 1,430$) (16), and 1000GP populations ($n = 1,299$). Individuals were projected along the PC1 and PC2 axes. (*Inset*) Zoomed view of TR and nearby populations. (*C*) Map of TR showing the number of chromosomes (WGS/WES) and mean admixture proportions of individuals with known birthplaces who originated from present day TR and former Ottoman Empire territories (TR-B, Balkan; TR-W, West; TR-C, Central; TR-N, North; TR-S, South; TR-E, East). (*D*) Admixture results of the TR WGS individuals with known origin ($n = 647$), the populations from Lazaridis et al. (16), and the 1000GP ($k = 8$). (*E*) PCA of TR individuals in a regional context. The populations with the lowest pairwise Wright's F_{ST} values were included.

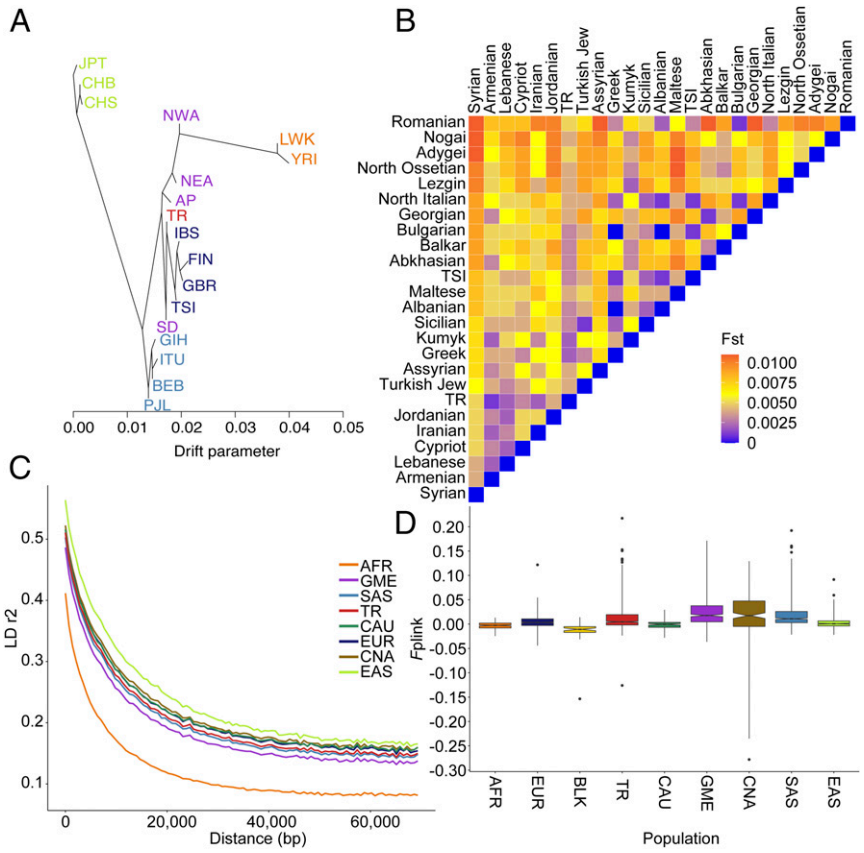

Fig. 2. The Turkish Peninsula as a bridge in the migration trajectories and high inbreeding levels in the TR population. (*A*) TreeMix phylogeny of the TR population (n = 3,362) along with the 1000GP controls (n = 1,299) and the GME populations (n = 696) representing divergence patterns. The length of branches is proportional to the extent of population drift. (*B*) The populations with a pairwise Wright's F_{ST} value < 0.01. The blue color indicates a closer genetic relationship. (*C*) The rate of LD decay in the TR population and in the populations of 1000GP and Lazaridis et al. (16). Mean variant correlations (r^2) are shown for each 700–base pair (bp) bin over 70,000 bp. EUR and BLK samples were combined as EUR because of the relatively low number of samples in the BLK population. (*D*) Distributions of the inbreeding coefficient (*F*plink). Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

supported the west-to-east trajectory of human migration into Asia (19).

The genetic similarity in the regional dataset was further tested with Wright's fixation index (F_{ST}), which revealed that the closest relationship of the TR population, in order of magnitude, is with the eastern and western neighboring populations, followed by Tuscan people in Italy. These results are therefore consistent with high levels of BLK, CAU, EUR, and Middle Eastern admixture (Fig. 2B). Interestingly, the pairwise F_{ST} values of the TR subregions indicated the effect of geographical distance on the genetic structure of the TR population (*SI Appendix*, Table S5).

Researchers investigating founder effects in populations suggest that two ancient population bottlenecks shaped the genetic variation in humans after they migrated out of Africa: the first bottleneck occurred about 50,000 to 60,000 y ago in the Middle East, and the second occurred when people crossed the ancient land bridge separating the Bering Strait from the Americas (20). Therefore, in order to see whether an ancient population bottleneck was shared between the TR and other populations, we calculated the mean rate of linkage disequilibrium (LD) decay (Fig. 2*C*). The LD for the TR population decayed more slowly than for the AFR, GME, and SAS populations and faster than the rest of the populations. The results were consistent with the phylogenetic tree analysis and supported the bottleneck hypothesis (11). The diverse levels of admixture observed in these populations confirm that the results are not due to intermixing and imply the occurrence of a shared ancient bottleneck.

Inbreeding Status and Estimation of Runs of Homozygosity. The consanguineous marriage rate is high in Turkey (22 to 36%), especially when compared to Western Europe and the Americas (<2%), and the majority of consanguineous marriages occur between first cousins (66.3%) (4, 21). High rates of consanguinity have been shown to be associated with an increased rate of recessive Mendelian disease (5, 8, 11). While the median estimated inbreeding coefficient (F_{plink}) for the TR population was similar to that of EUR, AFR, SAS, and EAS populations, it was as high as 0.21 in some of the TR individuals (Fig. 2*D*). These individuals are probably offspring of consanguineous mating given the fact that the inbreeding coefficient of an individual is approximately half the relationship between the parents. Overall, 29.6% of the TR population had an inbreeding coefficient ≥0.0156, which means a kinship greater than that of a second cousin marriage (22). By contrast, the comparable percentages for the control populations for the same threshold were for AFR, 0%; EUR, 14.9%; BLK, 0%; CAU, 1.98%; GME, 53%; SAS, 41.1%; CNA, 51.5%; and EAS 4.91% populations for the same threshold. We also assessed the effect of reported parental relatedness on F_{plink}. The medians for F_{plink} in the offspring of consanguineous or endogamous marriages were significantly higher compared to that of unrelated marriages (*SI Appendix*, Fig. S10*A*). The large negative

F_{plink} values in the dataset possibly reflect the offspring of pairs of unrelated but inbred individuals.

Consanguinity is associated with the increased length and sum total length of runs of homozygosity (ROH), whereas admixture acts to reduce the total number of ROH (23). Extended ROH have been shown to be enriched for rare and deleterious variations (11, 24). Therefore, we assessed the number and lengths of ROH, based on previously published ranges in the 1000GP populations as well as in the TR population, and compared the results (25). We observed the increased sum and number of ROHs in the TR individuals who are offspring of consanguineous marriages, although there was a marked overlap in different degrees of parental relatedness (*SI Appendix*, Fig. S10*B*). Similar to previous publications (25), the smallest median for the sum total length of ROH was observed in sub-Saharan Africans, while it was highest in the TR population. Also, in cases of long ROH (≥1,607 kb), the TR population displayed the highest numbers of individuals, whereas for the short- and medium-length ROH, the TR individuals are comparable to the East Asian, South Asian, and European populations (*SI Appendix*, Fig. S11*A*). The frequency calculations showed that ROH longer than 4 Mb in length were observed exclusively in the TR population (*SI Appendix*, Fig. S11*B*): 385 (49.81%) TR individuals had an ROH longer than 4 Mb in length, whereas none of the 1000GP individuals had an ROH longer than 4 Mb. Moreover, the longest ROH in the TR and 1000GP populations was detected in a TR individual: 41 Mb in length.

We utilized F_{roh} as a measure of autozygosity (*SI Appendix*, *Population Structure Analysis*) and compared it with F_{plink} using long and total classes of ROH. We detected significantly high correlations between F_{roh} and F_{plink} for both classes of ROHs (*SI Appendix*, Fig. S10 *C* and *D*). Similar to what was observed in F_{plink} analysis, we detected increased medians for F_{roh} for both classes in the offspring of consanguineous or endogamous marriages compared to that of unrelated marriages (*SI Appendix*, Fig. S10 *E* and *F*).

The Distribution of Y Chromosome and Mitochondrial DNA Haplotypes. Y chromosome and mitochondrial DNA (mtDNA) haplogroup analyses largely confirmed close genetic connections between the TR and EUR populations as well as with the neighboring populations. The most common Y chromosome haplogroups in TR individuals were from J2a (18.4%), R1b (14.9%), and R1a (12.1%) sublineages, consistent with previous findings (*SI Appendix*, Fig. S12 and Dataset S1) (3). Except for TR-B, in which I2a (20%) was the most prevalent haplogroup followed by R2a (17.1%) and E1b (14.3%), Y chromosome haplogroup distribution was similar with small differences in the TR subregions (*SI Appendix*, Fig. S13). For the mtDNA, the most common haplogroups were from the H sublineage (27.55%) followed by haplogroup U (19.53%) and haplogroup T (10.99%) in the TR population, as would be expected (*SI Appendix*, Fig. S14 and Dataset S1) (26). mtDNA haplogroup distribution showed small variance in the TR subregions, except for TR-B in which the frequency of the T haplogroup was very low (*SI Appendix*, Fig. S15). We also investigated the paternal and maternal gene flow from Central Asia by using the frequency of haplogroups that are restricted to Central Asia (3). Paternal gene flow based on Y chromosome haplogroups C-RPS4Y and O3-M122, which were previously implicated as Central Asian specific, ranged from 8.5 to 15.6%. Maternal gene flow based on mtDNA haplogroups D4c and G2a, which were previously suggested as Central Asian specific, was 8.13% (27) (*SI Appendix*, *Population Structure Analysis*).

The TR Variome. The GME Variome has demonstrated the power of consanguinity to identify causes of recessive disease, which are often the result of population-specific mutations (11). Thus, the comparison of derived allele frequencies (DAFs) of GME populations with that of the Exome Sequencing Project Variant Server revealed a large number of variants unique to the GME populations. We therefore investigated the genetic variation in the TR population at higher resolution by searching for TR DAFs in gnomAD (12) and GME datasets. We observed that ∼28% of the WES and ∼49% of the WGS variants in the very rare derived allele frequency bins (allele frequency [AF] < 0.005) are unique to the TR population (Fig. 3 *A* and *B*). Moreover, ∼79% of the very rare alleles of the TR population were absent from the GME Variome (Fig. 3*C*). The heat maps demonstrating the results of the correlation analyses of the TR and the gnomAD, or the TR and the GME DAFs, revealed that neither is a sufficient estimator for the TR DAFs (Fig. 3 *D*–*F*). These results indicate that the GME Variome is an inadequate representation of the TR population.

Next, we categorized TR variants according to their functional effects into seven main groups: high-confidence or low-confidence predicted loss-of-function variants (HC-pLoFs or LC-pLoFs), missense variants, non-frameshift indels, synonymous variants, noncoding variants, and other effects such as nonessential splice site variants, structural variants, and protein–protein contact variants (*SI Appendix*, *Variome Characterization* and Table S6). The missense variants were classified into two subgroups according to their deleteriousness: deleterious missense or other missense. Variants were also classified according to their allele frequencies in public databases. Overall, we identified 9,999,451 novel variants of which 37,123 were HC-pLoF or deleterious missense. A total of 839,775 variants (2.55%) in the rare and novel categories had an allele frequency higher than 1% in the TR Variome. We also noted that the proportions of HC-pLoFs and deleterious missense variants were higher among the novel and rare categories in the TR Variome, and these results were similar to those of the Iranome study (*SI Appendix*, Fig. S16*A*) (28). We also extracted the private variants (variants which are observed in only one individual either in the heterozygous or the homozygous state) of the TR Variome. We detected 23,403,893 private variants of which 8,898,088 (38%) were not observed in other public databases. A total of 79,947 (0.34%) of the all-private variants were HC-pLoFs or deleterious missense variants, and 32,687 (0.14%) of these variants were specific to the TR Variome.

Homozygous Predicted Loss-of-Function Mutations. Studies performed in populations with a high rate of consanguineous marriage provide researchers with an ideal opportunity to expand the list of naturally occurring human gene knockouts (11, 29, 30). Since common pLoF variants are less likely either to have a functional effect/clinical impact or to be subject to purifying selection (31), we first analyzed the number of high-confidence homozygous pLoF variants with an allele frequency lower than 1% in the TR Variome. We identified 704 rare homozygous HC-pLoF variants in 626 genes (Dataset S2). These homozygous HC-pLoFs were observed in 680 individuals (20.22%) who each had between one and four genes with homozygous HC-pLoFs. We then cross compared our list of homozygous HC-pLoFs and the genes carrying those variants with previously reported homozygous pLoFs lists in Icelanders, Pakistan Risk of Myocardial Infarction Study, Pakistanis living in Britain, and GenomeAsia (29, 30, 32, 33). We also extracted homozygous pLoFs from gnomAD and 1000GP data, thereby identifying a total of 173 genes with homozygous pLoFs specific to the TR Variome.

Homozygosity for pLoF variants with a population frequency higher than 1% might indicate selective advantage or the ameliorating effect of gene redundancy. A list of such variants in gnomAD and ExAC has recently been reported (34). Therefore, we extracted the high-confidence and common homozygous pLoFs of the TR individuals and identified 307 common homozygous HC-pLoF variants in 268 genes (Dataset S3). We then cross compared our list of common homozygous HC-pLoFs and the genes carrying those variants with the list of previously reported homozygous pLoFs from gnomAD and ExAC (34). We

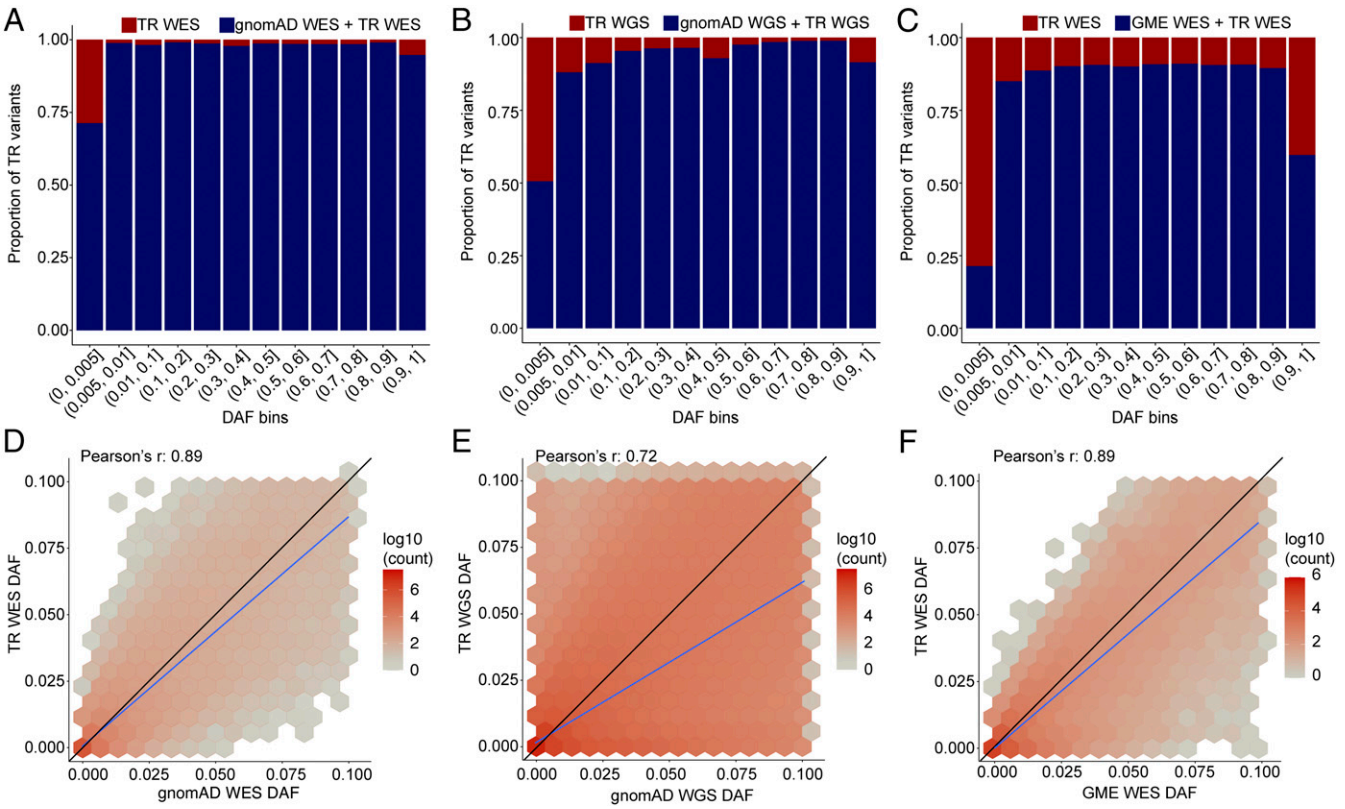

Fig. 3. The TR Variome possesses a significant number of very rare and unique variants that are poorly represented in gnomAD and GME. The proportion of TR variants represented in the TR Variome and other databases. (*A*) TR WES versus gnomAD WES, (*B*) TR WGS versus gnomAD WGS, (*C*) TR WES versus GME WES. The correlation of DAFs of rare TR variants in the (*D*) TR WES versus gnomAD WES, (*E*) TR WGS versus gnomAD WGS, and (*F*) TR WES versus GME WES. Hexagonal bins are shaded by the log-transformed number of variants in each bin.

identified 48 genes (15.64%) with common homozygous HC-pLoFs that were also present in the list of ExAC/gnomAD. We also noted that 259 variants in 227 genes that were listed as rare homozygous pLoFs in previous studies had a population frequency higher than 1% in the TR Variome. Consistent with the high rate of consanguineous marriage in Turkey, we noted that TR individuals carried excess rare homozygous pLoFs over what Hardy–Weinberg equilibrium would predict. We also investigated the effect of single-nucleotide homozygous pLoFs on transcriptional output by using proportion expression across transcripts (pext) values (35). Among 463 rare single-nucleotide homozygous pLoFs, 164 (35.4%) had a high (>0.9), 228 (49.2%) had a medium (0.1 to 0.9), and 71 (15.3%) had a low (<0.1) pext score. Among 105 common single-nucleotide homozygous pLoFs, 12 (11.4%) had a high, 42 (40%) had a medium, and 51 (48.6%) had a low pext score.

Clinically Relevant Variants. To demonstrate the potential of the TR Variome for the identification of disease-relevant variants, we first listed the TR Variome HC-pLoF variants and then searched for them in Online Mendelian Inheritance in Man (OMIM). In the TR Variome, we identified 22,570 HC-pLoF variants in 9,081 unique genes. A total of 76.1% of these variants were located under 6,831 OMIM-listed genes, while 25.37% of them were located under 2,197 OMIM-listed genes with an associated phenotype. We categorized the HC-pLoF variants according to their frequency status in other public databases as either novel, rare, or common. The numbers of novel and rare pLoFs were significantly higher than that of the common HC-pLoFs. However, the proportion of HC-pLoF variants in OMIM-listed genes and OMIM-listed genes with an associated clinical phenotype was comparable between classes (*SI Appendix*, Fig. S16*B*). These findings were similar to those derived from the Iranome database (28).

We then annotated variants that were identified in the TR Variome against the Human Gene Mutation Database (HGMD) (36) and ClinVar (37). A total of 6,537 variants in 2,188 genes from the TR Variome were found to be classified as disease-causing pathological mutations (DMs) in HGMD, and these DMs were observed in 3,362 individuals (100%) who each harbored between 1 and 30 DMs with an average of 12 (0 to 5 in the homozygous state) (*SI Appendix*, Fig. S16*C* and Dataset S4). A total of 1,636 variants in 929 genes were classified as pathogenic or pathogenic/likely pathogenic in ClinVar, and these variants were observed in 3,355 (99.79%) individuals who each had between 0 and 19 pathogenic and/or pathogenic/likely pathogenic variants with an average of 6 (0 to 10 in the homozygous state) (*SI Appendix*, Fig. S16*D* and Dataset S5). Importantly, 1,376 variants in 782 genes were found to be DM in HGMD and pathogenic or pathogenic/likely pathogenic in ClinVar (*SI Appendix*, Fig. S17).

Per-Genome Variant Summary and Imputation Panel. The extent of genetic variation in humans differs between populations. For example, individuals with African ancestry harbor a much higher number of variants in their genomes than Europeans (13). To compare the genetic structure of the TR population with other populations in terms of genome-wide variation, we first cataloged high-quality variants from the WGS dataset of the TR Variome and calculated the number of per-genome variant sites

and singletons and compared them with those of the 1000GP populations (Fig. 4*A* and *SI Appendix, Per-Genome Variant Summary and Imputation Panel*). As with the recently admixed American populations, the TR population displays a high number of per-genome variant sites and contains more variants than the European populations (*SI Appendix*, Fig. S18). Additionally, the average number of variants seen in only one individual—"singletons"—is highest in the TR and Luhya in Webuye, Kenya populations compared to the other 1000GP populations, highlighting the potential of rare variants for making novel discoveries in the TR population (Fig. 4*B*). The numbers of variant sites and singletons could be exacerbated by the high level of admixture in the TR population.

Imputing variants based on shared haplotypes of individuals is widely used for the GWAS of complex traits. Previous studies have shown that the use of population-specific reference panels increases imputation accuracy (13–15). For this reason, we generated a TR haplotype reference panel (*SI Appendix, Per-Genome Variant Summary and Imputation Panel*). When compared with the 1000GP, the TR reference panel alone significantly increased the imputation accuracy, especially for the variants with AF < 5%. The combined panel of the TR and 1000GP haplotypes further improved the imputation accuracy (Fig. 4*C*). Also, the TR reference panel produced higher numbers of high-confidence (expected $R^2 > 0.8$) calls of variants with expected AF < 1% than others, and the combined panel was more beneficial in terms of yielding a higher number of high-confidence variants than both panels for variants with expected AF ≥ 1% (Fig. 4*D*). The TR reference panel added 3,911 high-confidence rare variants (AF < 1%) that were not captured by the 1000GP panel, whereas the combined panel added 20,951 and 3,902 high-confidence variants (AF ≥ %1) that were not detected with the TR and the 1000GP, respectively. We also evaluated the performance on the imputation of genotypes of the individuals from the CAU, BLK, and GME populations of the Simons Genome Diversity Project (*SI Appendix*, Table S4) (38). We detected that the TR reference panel alone provided the highest accuracy in the CAU population, while the combined panel of TR and 1000GP resulted in statistically higher accuracy levels for the BLK and GME populations (*SI Appendix*, Fig. S19).

Discussion

In this report, we delineated the fine-scale genetic structure of the TR population. Consistent with Turkey's location at the crossroads of many historical population migrations, we find a high level of admixture. Studies of ancient DNA suggest that the early farmers of Anatolia in the late Pleistocene period had two significant ancestral contributions from Iran/Caucasus and ancient Levant in addition to the local genetic contribution from Anatolian hunter-gatherers (39). The admixture events in Anatolia extended toward Europe. However, there are also studies which suggest that the early Neolithic central Anatolians were probably descendants of local hunter-gatherers rather than

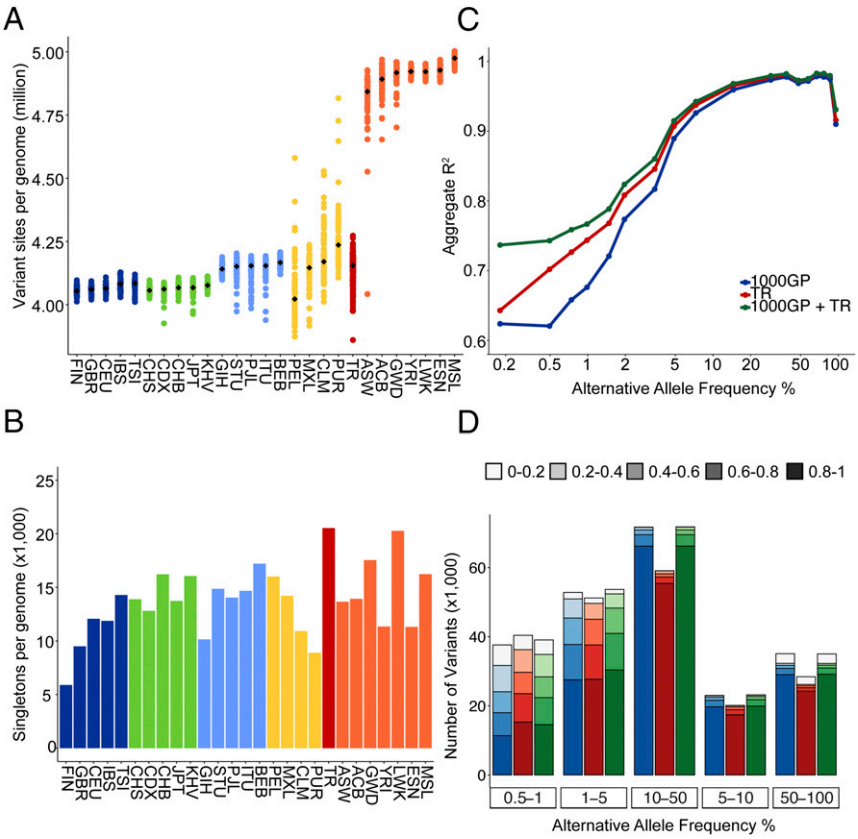

Fig. 4. Per-genome variant summary and imputation. (*A*) The number of variant sites per genome for autosomes. (*B*) The average number of singletons per genome for autosomes. (*C*) Evaluation of imputation performance on chromosome 20. The aggregate squared Pearson correlation coefficient (R^2) was calculated for genotypes called from WGS and imputed genotypes and plotted against alternative allele frequency for the three reference haplotype panels. A two-tailed Wilcoxon rank-sum test was used to assess the significance of the R^2 difference: $P = 0.002$ for the TR (mean ± SD: 0.88 ± 0.12) versus 1000GP (mean ± SD: 0.86 ± 0.14); $P = 0.002$ for the TR + 1000GP (mean ± SD: 0.89 ± 0.09) versus 1000GP. (*D*) The number of imputed variants as a function of expected alternative allele frequency. The density of shading reflects the expected imputation R^2, whereas colors represent the reference panels used for the imputation: 1000GP (blue), TR (red), or combined 1000GP and TR (green).

immigrants from the Levant or Iran (40). The migration of early Neolithic Anatolian farmers to Europe was a particularly important move with a significant impact on the genetic structure of preexisting as well as present-day European populations (16, 41). The most prominent effects of this migration are observed in southern Europe (42, 43). Moreover, an additional migration of later Neolithic Anatolian farmers occurred after the early Neolithic spread (44). Thus, the close genetic relationship of the TR population with the present-day European populations probably reflects these Anatolian migrations to Europe. The mobility of Anatolian and neighboring South Caucasus and North Levantine populations ~8,500 y ago also led to the genetic homogenization of western and eastern Anatolia for the first time (1). We also detected a significant shared ancestry in Turkey, Caucasus, Levantine, and Iranian populations. Mitochondrial DNA studies have suggested that recurrent gene flow between Europe and the Near East took place throughout the past 10,000 y (45). Anatolia has been exposed to many expansions and conquests during classical antiquity and the Middle Ages. The modern-day Anatolian population have traces of admixture events in their genomes from the Middle East, Central Asia, and Siberia (43). The expansion of Turkic tribes into Anatolia in the 11th century is a remarkable event that shaped the genetic structure of Anatolia. The modern-day TR population was previously known to have a Central Asian contribution amounting to between 3 and 30%, which was calculated using *Alu* insertion polymorphisms, mitochondrial, or Y chromosome loci (3, 46, 47). Similarly, we detected ~10% autosomal, 8 to 15% paternal, and ~8% maternal gene flow from Central Asia.

Anatolia was also subject to a high rate of recent external and internal migration events. In recent times, a huge number of permanent internal migrations from the eastern and northern Anatolia to the central, southern, and western provinces have occurred due to economic conditions and urbanization beginning in the late 19th to early 20th centuries (48). Moreover, ~400,000 Balkan refugees settled in western Anatolia during the population exchange with Balkan countries in 1914 (49). By means of admixture and Procrustes analyses, we have demonstrated that the geographical subregions of Turkey have a mild yet significant effect on the genetic structure. These findings revealed the effects of admixture events because of internal migration. Considering there was no clear-cut separation between TR subregions in PCA and Procrustes analysis, the recent migration events might have led to genetic homogenization in Turkey.

Large-scale population-specific genomic databases have the potential to play a pivotal role in enabling precision medicine. Such databases are important for variant prioritization and the identification of causative disease genes. In addition, the generation of high-quality haplotype reference panels for different human populations can be used to improve accuracy by enabling one to impute missing genotypes in large-scale GWAS (14, 15, 33). We therefore further expanded our knowledge of human genetic variation by focusing on the TR population.

Here, we present data derived from high-coverage WES and WGS of 3,362 individuals from TR and identify 9,999,451 novel variants of which 37,123 are deemed likely to have a deleterious effect. Our results also highlight the importance of population-specific reference panels for increasing the accuracy of imputation, especially for rare variation. We also demonstrated that the TR reference panel could also be exploited in the imputation of genotypes from neighboring populations. Genetic variation in the TR Peninsula has previously been investigated by relatively small-scale studies (11, 50); our data have substantially increased the sample size and, more importantly, the representation from all geographical regions and cities in Turkey. These high-resolution WES and WGS data enabled the detection of previously uncaptured rare variants by the GME Variome.

We found that the TR population harbors a considerable proportion of variants that are not yet designated in publicly available databases. Our results show that ~21% of all variants identified in this study were specific to the TR population, and ~38% of the private deleterious variants were not observed in other public databases. The TR Variome also introduces 839,775 novel or previously known rare variants, which have a frequency of higher than 1% in the TR population. Although DAF calculations revealed strong correlations, we observed that neither gnomAD nor GME was sufficient to represent the allele frequencies of a marked number of TR variants. Since allele frequency information is critical for Mendelian disease gene identification studies as well as variant prioritization strategies, the TR Variome will provide valuable data to facilitate the exclusion of low-probability candidates.

The phenotypic consequences of homozygous LoF mutations have long been investigated as a means to define gene function (29). Naturally occurring homozygous LoFs in humans, also termed "human knockouts," provide invaluable information in this context. However, it is not always easy to interpret their phenotypic consequences (if any) due to issues arising during sequence data analysis and differences in the phenotypic impact of knocking out different genes (51). Our list of homozygous pLoFs expanded the previous lists of human knockouts; however, one should also consider that this list is not based on deep phenotyping. Only the phenotypes which brought the family to medical attention were reported. Although these variants are only pLoFs until experimentally verified, sequencing consanguineous populations is one of the most efficient ways to expand the list of homozygous pLoFs (30). Since the TR population, in common with other populations with high consanguinity, contributes substantially to the study of Mendelian phenotypes, we also sought to characterize the extent of inbreeding status by analyzing the length of ROH. We detected several individuals with very high inbreeding coefficients and increased lengths of ROH, which facilitated the discovery of homozygous pLoFs. Moreover, TR individuals carried two to 30 variants classified by the HGMD as DMs and zero to 19 variants classified by the ClinVar as pathogenic or pathogenic/likely pathogenic. These results may have yielded secondary findings, with the potential to provide information on future disease prospects of the individuals concerned. However, such individuals might carry such variants without showing any clinical manifestations for the following reasons: carrying only one copy of the disease allele for a recessive disorder, late-onset disease, variable expression, and reduced penetrance (52, 53). Furthermore, there is the possibility that the TR Variome contains additional clinically relevant variants due to false negatives arising from automated variant filtering (54). Therefore, disease gene/variant identification studies in underrepresented populations are far from complete, and it is crucial to reassess disease-related databases using different population resources (55). Hence, analyses of the TR Variome will help to establish or exclude specific genes in the pathogenesis of a variety of genetic disorders.

In conclusion, we have established the TR Variome as the most comprehensive resource now available reflecting the genetic background of Turkey and suggest that it will provide an invaluable resource for studies of human and medical genetics. The identification of disease causative genes, particularly in the context of recessive disease, could be facilitated once the TR Variome is included alongside other publicly available databases.

Materials and Methods

For a full description of all of the methods and materials, see *SI Appendix, Supplementary Methods and Materials*. We generated combined datasets of 3,072 TR WES and 792 TR WGS samples. After sample-, variant- and genotype-based quality control (QC) filtering, we obtained 3,362 TR samples and 46,739,685 variants. We excluded 206 variants in the genes that were

causally associated with the phenotypes in our cohort (Dataset S6). We produced 38 technical replicates and calculated the concordance rates of these replicates after applying our QC filtering method (*SI Appendix*, Table S7). Using 3,362 unrelated TR individuals, we generated a TR dataset and performed ADMIXTURE (17) PCA and Procrustes analysis. Also, using global (n = 3,502) and regional (n = 1,805) datasets of TR WGS, Near East populations from Lazaridis et al. (16), and 1000GP, we performed all population structure analyses including PCA, ADMIXTURE, Procrustes analysis, phylogenetic tree, Wright's F_{ST}, inbreeding coefficient, ROH, Y chromosome, and mtDNA haplotypes. For the variome characterization analyses, we calculated DAFs of all 3,362 TR individuals and compared them with that of the gnomAD and GME Variome (11, 12). We performed functional annotations to determine the functional impact of the TR variants. We listed homozygous pLoF variants and clinically relevant variants using OMIM, ClinVar and HGMD (36, 37). We generated an imputation panel using the TR WGS dataset and squared Pearson's correlation coefficients (R2) were calculated to evaluate imputation accuracy. All participants gave written informed consent. The Institutional Ethics Committees of Bilkent University and Koç University approved the study.

Data Availability. Turkish Variome and Turkish reference panels for imputation are available for download from Figshare at https://figshare.com/articles/dataset/The_genetic_structure_of_the_Turkish_population_reveals_high_levels_of_variation_and_admixture/15147642. Individual level WES and WGS data are available at the Sequence Read Archive repository BioProject (accession ID: PRJNA670444, PRJNA674530, and PRJNA624188) and dbGAP under accession phs000744.v4.p2. The WES data of immunological disorders cohort is available from J.L.C. upon request. All other data are available in the main text or the supporting information.

ACKNOWLEDGMENTS. A.N.B. expresses her heartfelt gratitude to Suna, Inan, and Ipek Kirac for their vision, devotion, and dedicated mentorship throughout these studies and to Koç University Research Center for Translational Medicine for the inspiring research facilities created. We gratefully acknowledge İclal Büyükdevrim Özçelik and Nezahat Doğan for their insightful communications with the families and Serhan Kars for his help in the computational aspects of the project. We would like to extend our sincere gratitude to Dr. Kristel van Eijk for her help in the coverage analyses of WGS data, Dr. Hamzah Syed for his help in the data deposition, and Prof. Jan Veldink for his always sincere cooperation and assistance in Project MinE-related queries. This study was funded, in part, by Suna and Inan Kirac Foundation and Koç University. The Turkish Academy of Sciences supported this work. M.E.K. is the recipient of fellowship 2211-A National Doctorate Scholarship Program of Scientific and Technological Research Council of Turkey Directorate of Science Fellowships and Grant Programmes. Whole-exome sequencing was performed at the Yale Center for Mendelian Genomics funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute (NIH M#UM1HG006504). Whole-genome sequencing was performed at the University Medical Center Utrecht, Netherlands. The Genome Sequencing Program Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. This work was funded, in part, by the JPB Foundation, National Center for Advancing Translational Sciences, NIH Clinical and Translational Science Award Program (UL1TR001866), NIH (R01AI088364, R01AI127564, R37AI095983, and P01AI61093), the French National Research Agency (ANR) under the "Investments for the future" Program (ANR-10-IAHU-01), Integrative Biology of Emerging Infectious Diseases Laboratoire d'Excellence (ANR-10-LABX-62-IBEID), an Inborn Errors of Immunity to HSV-1 underlying Childhood Herpes Simplex Encephalitis: An Exception or a Rule? Grant (ANR-14-CE14-0008-01), a SEAe-HostFactors Grant (ANR-18-CE15-0020 02), a Childhood Invasive Pneumococcal Disease: Toward the Identification of Novel Primary Immunodeficiencies Project (ANR 14-CE15-0009-01), and a grant from The French National Cancer Institute/Cancéropole Ile-de-France (2013-1-PL BIO-11-INSERM 5-1), the Rockefeller University, INSERM, the HHMI, Paris Descartes University, and the St. Giles Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

1. E. Skourtanioti *et al*., Genomic history of neolithic to bronze age Anatolia, northern Levant, and southern Caucasus. *Cell* **181**, 1158–1175.e28 (2020).
2. P. B. Golden, *An Introduction to the History of the Turkic Peoples. Ethnogenesis and State-Formation in Medieval and Early Modern Eurasia and the Middle East* (Otto Harrassowitz, Wiesbaden, 1992).
3. C. Cinnioğlu *et al*., Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* **114**, 127–148 (2004).
4. S. Akbayram *et al*., The frequency of consanguineous marriage in eastern Turkey. *Genet. Couns.* **20**, 207–214 (2009).
5. A. H. Bittles, M. L. Black, Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. *Proc. Natl. Acad. Sci. U.S.A.* **107** (suppl. 1), 1779–1786 (2010).
6. L. D. Notarangelo, R. Bacchetta, J. L. Casanova, H. C. Su, Human inborn errors of immunity: An expanding universe. *Sci. Immunol.* **5**, eabb1662 (2020).
7. T. Özçelik, Medical genetics and genomic medicine in Turkey: A bright future at a new era in life sciences. *Mol. Genet. Genomic Med.* **5**, 466–472 (2017).
8. T. Özçelik *et al*., Collaborative genomics for human health and cooperation in the Mediterranean region. *Nat. Genet.* **42**, 641–645 (2010).
9. X. Yang *et al*., The influence of admixture and consanguinity on population genetic diversity in Middle East. *J. Hum. Genet.* **59**, 615–622 (2014).
10. Z. Mehrjoo *et al*., Distinct genetic variation and heterogeneity of the Iranian population. *PLoS Genet.* **15**, e1008385–e1008385 (2019).
11. E. M. Scott *et al*.; Greater Middle East Variome Consortium, Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48**, 1071–1076 (2016).
12. K. J. Karczewski *et al*.; Genome Aggregation Database Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
13. A. Auton *et al*.; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. D. Gurdasani *et al*., Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* **179**, 984–1002.e36 (2019).
15. H. Bai *et al*., Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* **50**, 1696–1704 (2018).
16. I. Lazaridis *et al*., Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
17. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
18. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
19. B. M. Henn *et al*., Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8**, e1002397 (2012).
20. W. Amos, J. I. Hoffman, Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci.* **277**, 131–137 (2010).
21. E. Tunçbilek, Genetic services in Turkey. *Eur. J. Hum. Genet.* **5** (suppl. 2), 178–182 (1997).
22. F. C. Ceballos, G. Alvarez, Royal dynasties as human inbreeding laboratories: The Habsburgs. *Heredity* **111**, 114–121 (2013).
23. F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, J. F. Wilson, Runs of homozygosity: Windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).
24. Z. A. Szpiech *et al*., Long runs of homozygosity are enriched for deleterious variation. *Am. J. Hum. Genet.* **93**, 90–102 (2013).
25. T. J. Pemberton *et al*., Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
26. Z. Bánfai *et al*., Revealing the genetic impact of the ottoman occupation on ethnic groups of east-central Europe and on the roma population of the area. *Front. Genet.* **10**, 558 (2019).
27. D. Comas *et al*., Admixture, migrations, and dispersals in Central Asia: Evidence from maternal DNA lineages. *Eur. J. Hum. Genet.* **12**, 495–504 (2004).
28. Z. Fattahi *et al*., Iranome: A catalog of genomic variations in the Iranian population. *Hum. Mutat.* **40**, 1968–1984 (2019).
29. D. Saleheen *et al*., Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
30. V. M. Narasimhan *et al*., Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
31. D. G. MacArthur *et al*.; 1000 Genomes Project Consortium, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
32. P. Sulem *et al*., Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
33. GenomeAsia100K Consortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
34. A. Rausell *et al*., Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13626–13636 (2020).
35. B. B. Cummings *et al*.; Genome Aggregation Database Production Team; Genome Aggregation Database Consortium, Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).
36. P. D. Stenson *et al*., The Human Gene Mutation Database (HGMD®): Optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
37. M. J. Landrum *et al*., ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
38. S. Mallick *et al*., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
39. M. Feldman *et al*., Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. *Nat. Commun.* **10**, 1218 (2019).
40. G. M. Kılınç *et al*., Archaeogenomic analysis of the first steps of Neolithization in Anatolia and the Aegean. *Proc. Biol. Sci.* **284**, 20172064 (2017).
41. I. Mathieson *et al*., Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).

42. A. Raveane *et al*., Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci. Adv.* **5**, eaaw3492 (2019).
43. A. Omrak *et al*., Genomic evidence establishes Anatolia as the source of the European neolithic gene pool. *Curr. Biol.* **26**, 270–275 (2016).
44. G. M. Kılınç *et al*., The demographic development of the first farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
45. M. Richards *et al*., Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).
46. C. C. Berkman, H. Dinc, C. Sekeryapan, I. Togan, Alu insertion polymorphisms and an assessment of the genetic contribution of Central Asia to Anatolia with respect to the Balkans. *Am. J. Phys. Anthropol.* **136**, 11–18 (2008).
47. G. Di Benedetto *et al*., DNA diversity and population admixture in Anatolia. *Am. J. Phys. Anthropol.* **115**, 144–156 (2001).
48. C. C. Clay, Labour migration and economic conditions in nineteenth-century Anatolia. *Middle East. Stud.* **34**, 1–32 (1998).
49. A. Içduygu, Ş. Toktas, B. A. Soner, The politics of population in a nation-building process: Emigration of non-Muslims from Turkey. *Ethn. Racial Stud.* **31**, 358–389 (2008).
50. C. Alkan *et al*., Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics* **15**, 963 (2014).
51. V. M. Narasimhan, Y. Xue, C. Tyler-Smith, Human knockout carriers: Dead, diseased, healthy, or improved? *Trends Mol. Med.* **22**, 341–351 (2016).
52. Y. Xue *et al*.; 1000 Genomes Project Consortium, Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
53. D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, H. Kehrer-Sawatzki, Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
54. C. F. Wright *et al*.; DDD Study, Evaluating variants classified as pathogenic in ClinVar in the DDD Study. *Genet. Med.* **23**, 571–575 (2021).
55. M. Abouelhoda, T. Faquih, M. El-Kalioby, F. S. Alkuraya, Revisiting the morbid genome of Mendelian disorders. *Genome Biol.* **17**, 235–235 (2016).