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

M. Ece Kars[1], A. Nazlı Başak[2], O. Emre Onat[1], Kaya Bilguvar[3], Jungmin Choi[3], Yuval Itan[4,5], Caner Çağlar[6], Robin Palvadeau[2], Jean-Laurent Casanova[7,8,9,10,11], David N. Cooper[12], Peter D. Stenson[12], Alper Yavuz[13], Hakan Buluş[13], Bülent Yıldız[14], Jeffrey M. Friedman[6,11], Tayfun Özçelik[1,15,16]

[1]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey. [2]Suna and Inan Kıraç Foundation, Neurodegeneration Research Laboratory (NDAL), Research Center for Translational Medicine, Koç University School of Medicine, Istanbul, Turkey. [3]Department of Genetics, Yale Center for Genome Analysis, Yale University School of Medicine, New Haven, Connecticut, USA. [4]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [6]Laboratory of Molecular Genetics, Rockefeller University, New York, New York, USA. [7]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, New York, USA. [8]Laboratory of Human Genetics of Infectious Diseases, Necker Branch INSERM U1163, Necker Hospital for Sick Children, Paris, France. [9]Imagine Institute, University of Paris, Paris, France. [10]Pediatric Immunology-Hematology Unit, Necker Hospital for Sick Children, Paris, France. [11]Howard Hughes Medical Institute (HHMI), Rockefeller University, New York, New York, USA. [12]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. [13]Keçiören Educational and Research Hospital, Department of Surgery, Health Sciences University, Ankara, Turkey. [14]Division of Endocrinology and Metabolism, Department of Internal Medicine, Hacettepe University School of Medicine, Hacettepe, Ankara, Turkey. [15]Neuroscience Program, Graduate School of Engineering and Science, and [16]Institute

of Materials Science and Nanotechnology, National Nanotechnology Research Center (UNAM), Bilkent University, Ankara, Turkey.

**Running Title:** The Turkish population is highly admixed

**Keywords:** Turkish Variome, admixture, exome, whole genome, sequencing

**Abstract**

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 3,599 unrelated subjects whose whole-exomes (*n* = 2,826) or whole-genomes (*n* = 773) were sequenced to generate a Turkish (TR) Variome that should serve to facilitate disease gene discovery in this population. Consistent with the history of present-day Turkey as a crossroads between Europe and Asia, we found extensive admixture between Middle Eastern and European populations with a closer genetic relationship of the TR population to Europeans than hitherto appreciated. Since the TR population, in common with other populations with high consanguinity, contribute substantially to the study of Mendelian phenotypes, we also sought to characterize the extent of inbreeding status by analyzing the length of runs of homozygosity (ROH). We determined that approximately 45% of TR individuals had high inbreeding coefficients (≥ 0.0156) with ROH longer than 4Mb being found exclusively in the TR population [TAYFUN: as compared to which population?]. We also found that approximately 30% of exome and 50% of genome variants in the very rare range (AF< 0.005) are unique to the modern TR population. Finally, we annotated these variants based on their functional consequences to establish a TR Variome containing alleles of potential medical relevance, a repository of human knockouts and a TR reference panel for genotype imputation using high-quality haplotypes, to facilitate genome-wide association studies. In addition to providing new 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. (279 words)

## Introduction

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 Early Bronze age, including long-distance migration from Central Asia to Anatolia (Skourtanioti et al. 2020). The Turkic peoples, a collection of ethnolinguistically related populations originating from Central Asia were first documented in Western Eurasia in the 4th-5th 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 6$^{th}$ and 11$^{th}$ centuries, beginning with the Seljuk Turks followed by the Ottomans (Golden 1992). 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 WW1 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 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 TR males revealed that the proportion of

recent paternal gene flow from Central Asia was ~9% (Cinnioğlu et al. 2004) thereby raising the possibility that modern-day Anatolia is an admixture of pre-existing Anatolian and Turkic peoples.

The practice of consanguineous marriage is frequent in Turkey, especially in the eastern provinces (Akbayram et al. 2009). 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 (Bittles and Black 2010; Özçelik et al. 2010; Özçelik 2017; Notarangelo et al. 2020). The genetic admixture and consanguinity have had a significant effect on the genetic diversity of Middle Eastern populations (Yang et al. 2014; Mehrjoo et al. 2019).

The characterization of the Greater Middle East (GME) Variome, comprising the most comprehensive genomic database for Middle East populations, has shown that knowledge of the genomic architecture of these populations facilitates disease gene identification in family studies and in GWAS studies of populations (Scott et al. 2016). 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 WES or WGS data (Karczewski et al. 2020). Therefore, a comprehensive database of alleles in the TR population should facilitate disease gene identification in consanguineous families and assessment of the clinical phenotypes of individuals who are homozygous for mutations in specific genes.

Finally, most GWAS studies to date have analyzed DNA from European ancestry-derived populations, and it will be important to extend GWAS studies 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) (Auton et al. 2015; Bai et al. 2018; Gurdasani et al. 2019). 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

We analyzed whole-exome (WES) and whole-genome (WGS) sequence data from 4,194 individuals who had participated in genetic studies of obesity, essential tremor, Parkinson's disease, amyotrophic lateral sclerosis, ataxia, delayed sleep phase disorder, polycystic ovarian syndrome, and various assorted neurological and immunological disorders (Table S1). WES and WGS samples were processed separately for analyses of the sample quality and familial relationships using BCFtools. 2,826 WES and 773 WGS samples remained after filtration according to quality metrics and relatedness. The mean target base coverage for the exons of consensus coding sequence (CCDS) build 15 (Farrell et al. 2014) of the WES samples was 68X with 95.33%, 93.83%, and 88.82% coverage at 8X, 10X and 20X or more, respectively. The mean depth of coverage for the WGS samples was 32X with 93%, 91% and 89% coverage at 8X, 10X and 20X or more, respectively. We identified 1,981,939 WES and 72,982,375 WGS variants. Following a variant filtration process to identify high-quality variants, we obtained 1,436,012 WES and 47,322,283 WGS variants (see Methods). Exome data provides accurate results for population structure analyses (Belkadi et al. 2016). Therefore, we combined the exome part of the WGS data with the WES data for further analyses of population structure. The difference in the mean number of novel variants for the samples sequenced with two different technologies can be used to detect potential batch effects (Fakhro et al. 2016). Accordingly, we

evaluated the sample-based quality control measures and observed that the proportions of novel variants (not present in dbSNP build 151) were consistent for the samples sequenced with either WES or WGS using the thresholds selected for minimum allele count and depth (Table S2).

The geographical origins of ancestors (birthplaces of maternal and paternal grandparents) of 1,598 TR samples were documented and grouped into six different subregions, namely Balkan (TR-B:93), West (TR-W:163), Central (TR-C:509), North (TR-N:407), South (TR-S:126) and East (TR-E:300), and then compared with 13 populations from the 1000GP (Table S3) (Auton et al. 2015). First, we performed a principal component analysis (PCA) using only TR individuals of known origin (Fig. S1). 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. 1A, correlation in Procrustes rotation, 0.46 $P = 9.99 \times 10^{-6}$).

We then evaluated the genetic substructure of Turkey using ADMIXTURE (Alexander et al. 2009) and $k = 5$ was determined as the lowest cross-validation error (Fig. S2A). Individuals with unknown ancestral birthplaces (TR-U) exhibited similar ancestral components to those individuals with known ancestral birthplaces. All five ancestries were represented in each geographical region, although in different proportions (Fig. S3). Surprisingly, when we included samples from African, European, South and East Asian populations from 1000GP, $k = 8$ gave the lowest cross-validation error, which revealed that only one ancestry was specific to the TR population (Fig. 1B, C, Fig. S2B, Fig. S4). We observed small differences in the contributions of primary admixture components from Africa, Europe, South Asia and East Asia in the TR population, reflecting the importance of geographical location in shaping genetic substructure. Additionally, we observed a significant TR and European overlap. Consistent with ancient DNA

157 studies, these results firmly establish significant admixture in the genetically homogenous
158 population of present-day Turkey (Lazaridis et al. 2016; Skourtanioti et al. 2020).
159 We employed a two-step evaluation of the relationship between the TR and 1000GP
160 populations using PCA. First, we compared the TR population with three super-populations
161 (Sub-Saharan Africa, East Asia and South Asia) in the 1000GP; Sub-Saharan African, East
162 Asian, and South Asian populations were distinguished with PC1, PC2, and PC3, respectively,
163 while the TR population formed a distinct cluster and also displayed a close relationship with the
164 European populations (Fig. 1D and Fig. S5). To further evaluate this close relationship, we
165 performed a second PCA between TR individuals with known geographical origin of
166 grandparental alleles and European populations. As expected, we observed that the genetic
167 connection of European and TR populations was established through the Balkans (TR-B) and
168 Western Turkey (TR-W), which further emphasizes the importance of geography on the genetic
169 variation seen in Turkey. Importantly, consistent with a high level of admixture, the degree of
170 variation observed in the TR population was much higher than that of distinct European
171 populations (Fig. 1E and Fig. S6).
172 The position of Turkey along historical routes of migration and the effect of genetic drift
173 was assessed using a maximum likelihood phylogenetic tree, with the inclusion of the 1000GP
174 and the GME populations (Fig. 2A) (Pickrell and Pritchard 2012). The clusters of each 1000GP
175 and GME populations were recapitulated with the inferred tree, and Turkey connected the GME
176 and European branches. When the populations were ordered from the root, the ordering
177 corroborated the “out-of-Africa” hypothesis and supported the west-to-east trajectory of human
178 migration into Asia (Henn et al. 2012).
179 The genetic similarity between the TR and the 1000GP populations was further tested
180 with Wright's fixation index ($F_{ST}$), which revealed that the closest relationship, in order of
181 magnitude, is with the Toscani in Italy (TSI), followed by the Iberian population in Spain (IBS),
182 the British in England and Scotland (GBR), the Finnish in Finland (FIN) and the Punjabi from

Lahore, Pakistan (PJL). These results are therefore consistent with high levels of European admixture (Fig. 2B).

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 years ago in the Middle East and the second occurred when people crossed the ancient land bridge separating the Bering Strait from the Americas (Amos and Hoffman 2010). 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. 2C). The LD for the TR population decayed more slowly than for the African populations. Yet, a similar rate was observed in the East Asian, South Asian, European and TR populations, supporting the bottleneck hypothesis (Scott et al. 2016). 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 ROH**

The consanguineous marriage rate is high in Turkey (22-36%), especially when compared to Western Europe and the Americas (<2%), and the majority of consanguineous marriages occur between first cousins (66.3%) (Tunçbilek 1997; Akbayram et al. 2009). High rates of consanguinity have been shown to be associated with an increased rate of recessive Mendelian disease (Bittles and Black 2010; Özçelik et al. 2010; Scott et al. 2016). While the median estimated inbreeding coefficient (*F*) for the TR population was similar to that of European, African, South Asian and East Asian populations, it was as high as 0.27 in some of the TR individuals (Fig. 2D). These individuals are probably offspring of consanguineous matings given the fact that the inbreeding coefficient of an individual is approximately half the relationship between the parents. Overall, 44.6% of the TR population had a kinship coefficient $\geq 0.0156$, which means a kinship greater than that of a second cousin marriage (Ceballos and Alvarez

2013). By contrast, the comparable percentages for AFR, EUR, EAS and SAS populations for the same threshold were 3.9%, 4.7%, 8% and 22.7%, respectively.

Consanguinity is associated with the increased length and sum total length of ROH, whereas admixture acts to reduce the total number of ROH (Ceballos et al. 2018b). Extended ROH have been shown to be enriched for rare and deleterious variation (Szpiech et al. 2013; Scott et al. 2016). 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 (Pemberton et al. 2012). Similar to previous publications, the smallest median for sum total length of ROH was observed in Sub-Saharan Africans (Pemberton et al. 2012), whilst it was highest in the TR population. Also, in cases of long ROH (≥ 1.607) 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 (Fig. 3A). The frequency calculations showed that ROH longer than 4Mb in length were observed exclusively in the TR population (Fig. 3B): 385 (49.81%) TR individuals had ROH longer than 4Mb in length whereas none of the 1000GP individuals had an ROH longer than 4Mb. Moreover, the longest ROH in the TR and 1000GP populations was detected in a TR individual: 41 Mb in length. [TAYFUN: Isn't the length of ROH going to be a function of the SNPs used? The higher the number of SNPs employed, the greater the resolution. The greater the resolution, the more likely are breaks between ROH to be found….which will serve to reduce the length of the ROH. In comparisons between different populations, you are presumably comparing like with like in terms of the numbers of SNPs employed]

**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 (Scott et al. 2016). Thus, the comparison of derived allele frequencies (DAFs) of GME populations with that of the NHBLI GO

Exome Sequencing Project (ESP) 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 (Karczewski et al. 2020) and GME datasets. We observed that approximately 30% of the WES and approximately 50% of the WGS variants in the very rare derived allele frequency bins (AF < 0.005) are unique to the TR population (Fig. 4A, B). Moreover, approximately 80% of the very rare alleles of the TR population were absent from the GME Variome (Fig. 4C). The heatmaps 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. 4 D-F). These results indicate that the GME Variome is an inadequate representation of the TR population.

Next, the WES and WGS variants were annotated by ENSEMBL v.87 using SnpEff (Cingolani et al. 2012; Hunt et al. 2018) and merged (*n* = 48,308,918). The predicted loss of function (pLoF) variants, including frameshifts, stop-gain, stop-loss, start-loss or essential splice site variants, were further annotated using LOFTEE (MacArthur et al. 2012) and classified into high-confidence (HC-pLoFs) or low-confidence pLoFs (LC-pLoFs). We categorized variants according to their functional effects into seven main groups: HC-pLoFs, LC-pLoFs, missense variants, non-frameshift indels, synonymous variants, non-coding variants and other effects such as non-essential splice site variants, structural variants, and protein-protein contact variants (Table S4). The missense variants were annotated using Polyphen-2, SIFT, and Combined Annotation Dependent Depletion (CADD) (Adzhubei et al. 2010; Vaser et al. 2016; Rentzsch et al. 2019) according to their deleteriousness and classified into two subgroups: deleterious missense or other missense. Variants were also classified according to their allele frequencies in public databases. A variant was classified as “Novel”, if it had no record in dbSNP build 151, gnomAD (https://gnomad.broadinstitute.org/), 1000GP (https://www.internationalgenome.org/), and NHBLI GO exome sequencing project (ESP) Exome variant server (https://evs.gs.washington.edu/EVS/ ); it was deemed to be “Common”, if the variant had an AF

≥ 0.01 in any of the above-mentioned databases. If the variant had an AF < 0.01 in all databases, it was classified as “Rare” (Table S4). Overall, we identified 10,116,912 novel variants of which 38,540 were HC-pLoF or deleterious missense. A total of 1,009,435 variants (2.94%) 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 (Fattahi et al. 2019) (Fig. 5A). 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 26,840,965 private variants of which 9,035,278 (33.66%) were not observed in other public databases. A total of 86,779 (0.32%) of the all private variants were HC-pLoFs or deleterious missense variants and 33,771 (0.13%) of these variants were specific to the TR Variome.

**Human knockouts**

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 (Narasimhan et al. 2016a; Scott et al. 2016; Saleheen et al. 2017). Since common pLoF variants are less likely either to have a functional effect/clinical impact or to be subject to purifying selection (MacArthur et al. 2012), we first analyzed the number of high-confidence homozygous pLoF variants with an allele frequency lower than 1% in the TR Variome. We identified 783 rare homozygous pLoF variants in 679 genes. These homozygous pLoFs were observed in 631 individuals (20.31%) who each had between 1 and 4 genes knocked out (Table S5). We then cross-compared our list of homozygous pLoFs and the genes carrying those variants with previously reported human knockout lists in Icelanders (Sulem et al. 2015), PROMIS (Saleheen et al. 2017), Pakistanis living in Britain (Narasimhan et al. 2016a) and GenomeAsia (GenomeAsia100KConsortium 2019). We also extracted homozygous pLoFs from

gnomAD and 1000GP data thereby identifying a total of 221 novel knocked out genes specific to the TR Variome (Table S5). We also noted that 142 variants in 132 genes that were listed as rare knockouts in previous studies had a frequency higher than 1% in the TR Variome.

Homozygosity for pLoF variants with a frequency higher than 1% might indicate selective advantage or the ameliorating effect of gene redundancy [TAYFUN: Have you tried looking at Hardy-Weinburg equilibrium? How does the expected number of homozygotes compare with that expected on the basis of the frequency of the heterozygote allele?]. A list of such variants in gnomAD and ExAC has recently been reported (Rausell et al. 2020). Therefore, we extracted the high-confidence and common homozygous pLoFs of the TR individuals and identified 327 common homozygous HC-pLoF variants in 280 genes (Table S6). We then cross-compared our list of common homozygous HC-pLoFs and the genes carrying those variants with the list of previously reported human knockouts from gnomAD and ExAC (Rausell et al. 2020). We identified 46 genes (16.43%) with common homozygous HC-pLoFs that were also present in the ExAC/gnomAD knockout list. Hence, we have identified 234 new genes harboring homozygous pLoFs with a frequency ≥ 1% in the population [TAYFUN: I assume that the ≥ 1% frequency refers to the frequency of the heterozygous allele not the frequency of the homozygote. You should probably make this clear here and above].

**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 25,804 HC-pLoF variants in 9,634 unique genes. 76.37% of these variants were located under OMIM-listed genes while 24.72% of them were located under OMIM-listed genes with an associated phenotype [TAYFUN: Meaning unclear! You surely cannot mean that 19,000 variants are specifically listed in OMIM. Do you mean the associated genes were listed in OMIM?]. 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 (Fig. 5B). These findings were similar to those derived from the Iranome database (Fattahi et al. 2019).

We then annotated variants that were identified in the TR Variome against the Human Gene Mutation Database (HGMD) (Stenson et al. 2020) and ClinVar (Landrum et al. 2018) (Fig.S8). 6,931 variants in 2,261 genes from the TR Variome were found to be classified as disease-causing pathological mutations (DM) in HGMD, and these DMs were observed in 3,599 individuals (100%) who each harbored between 2 and 30 DMs with an average of 13 (0-5 in the homozygous state) (Fig. 5C, Table S7). 1,778 variants in 982 genes were classified as pathogenic or pathogenic/likely pathogenic in ClinVar and these variants were observed in 3,592 (99.8%) individuals who each had between 0 and 19 pathogenic and/or pathogenic/likely pathogenic variants with an average of 6 (0-10 in the homozygous state) (Fig. 5D, Table S8). Importantly, 1,492 variants in 834 genes were found to be DM in HGMD and pathogenic or pathogenic/likely pathogenic in ClinVar (Fig. S8).

**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 (Auton et al. 2015). To compare the genetic structure of the TR population with other populations in terms of genome-wide variation, we first catalogued high-quality variants from the WGS dataset of the TR Variome with up to 20% missingness and imputed the missing sites by BEAGLE v5.1 (Browning et al. 2018). Then, we calculated the number of per-genome variant sites and singletons from the 773 whole-genome sequenced TR individuals and compared it with

339 those of the 1000GP populations (Fig. 6A). As with the recently admixed American populations,
340 the TR population displays a high number of per-genome variant sites and contains more
341 variants than the European populations (Fig. S7). Additionally, the average number of variants
342 seen in only one individual – 'singletons' - is highest in the TR population compared to the
343 1000GP populations, highlighting the potential of rare variants for making novel discoveries in
344 the TR population (Fig. 6B).
345 Imputing variants based on shared haplotypes of individuals is widely used for the GWAS
346 of complex traits. Previous studies have shown that the use of population-specific reference
347 panels increases imputation accuracy (Auton et al. 2015; Bai et al. 2018; Gurdasani et al. 2019).
348 For this reason, we generated a TR haplotype reference panel and evaluated its performance by
349 comparing it with the existing 1000GP panel. First, using high quality SNPs, we constructed the
350 haplotypes of 773 whole-genome sequenced individuals with BEAGLE v5.1 and re-phased
351 these haplotypes using SHAPEIT v2 (Delaneau et al. 2013) to obtain a reference panel with
352 higher accuracy. Then, we randomly selected 73 whole-genome sequences and created a target
353 panel. Afterwards, we extracted 44,367 SNPs on chromosome 20 from the catalog of Illumina
354 Infinium Omni2.5-8 Kit whereas we marked the remaining sites as missing. The missing
355 positions were imputed with IMPUTE2 (Howie et al. 2009) using the following three reference
356 haplotype panels: the remaining 700 samples of the TR population, the 1000GP individuals,
357 1000GP plus TR individuals. The TR panel comprised only SNPs whereas the 1000GP contains
358 SNPs, short indels plus copy number variations. We assessed the performance of each
359 reference panel by calculating the average aggregate squared correlation ($R^2$) between imputed
360 and sequenced genotypes (Fig. 6C). When compared with the 1000GP, the TR reference panel
361 alone significantly increased the imputation accuracy, especially for the variants with AF < 5%.
362 The combined panel of the TR and 1000GP haplotypes further improved the imputation
363 accuracy (Fig. 6C). We also calculated the number of imputed variants on chromosome 20 in
364 different expected $R^2$ and AF bins by using the summary file generated with IMPUTE2. 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 6D). 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.

## 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 (Feldman et al. 2019). The admixture events in Anatolia extended towards Europe. However, there are also studies, which suggest that the early Neolithic central Anatolians were probably descendants of local hunter–gatherers, rather than immigrants from the Levant or Iran (Kılınç et al. 2017). The migration of early Neolithic Anatolian farmers to Europe was a particularly important move with a significant impact on the genetic structure of pre-existing as well as present-day European populations (Mathieson et al. 2015; Lazaridis et al. 2016). The most prominent effects of this migration are observed in Southern Europe (Omrak et al. 2016; Raveane et al. 2019). Moreover, an additional migration of later Neolithic Anatolian farmers occurred after the early Neolithic spread (Kılınç et al. 2016). 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 approximately 8,500 years ago also led to the genetic homogenization of western and eastern

391 Anatolia for the first time (Skourtanioti et al. 2020). Mitochondrial DNA studies have suggested
392 that recurrent gene flow between Europe and the Near East took place throughout the past
393 10,000 years (Richards et al. 2000). Anatolia has been exposed to many expansions and
394 conquests during classical antiquity and the Middle Ages. The modern-day Anatolian population
395 have traces of admixture events in their genomes from the Middle East, Central Asia, and
396 Siberia (Omrak et al. 2016). The expansion of Turkic tribes into Anatolia in the 11$^{th}$ century is a
397 remarkable event that shaped the genetic structure of Anatolia. The modern-day TR population
398 therefore has a Central Asian contribution amounting to between 3% and 30%, which was
399 calculated using *Alu* insertion polymorphisms, mitochondrial or Y-chromosome loci (Di
400 Benedetto et al. 2001; Cinnioğlu et al. 2004; Berkman et al. 2008).
401 Anatolia was also subject to a high rate of recent external and internal migration events.
402 In recent times, a huge number of permanent internal migrations from the Eastern and Northern
403 Anatolia to the Central, Southern and Western provinces have occurred due to economic
404 conditions and urbanization beginning in the late 19$^{th}$ to early 20$^{th}$ centuries (Clay 1998).
405 Moreover, approximately 400,000 Balkan refugees settled in Western Anatolia during the
406 population exchange with Balkan countries in 1914 (Içduygu et al. 2008). By means of admixture
407 and Procrustes analyses, we have demonstrated that the geographical subregions of Turkey
408 have a mild yet significant effect on the genetic structure. These findings revealed the effects of
409 admixture events due to internal migration. Considering there was no clear-cut separation
410 between TR subregions in PCA and Procrustes analysis, the recent migration events might have
411 led to genetic homogenization in Turkey.
412 Large-scale population-specific genomic databases have the potential to play a pivotal
413 role in enabling precision medicine. Such databases are important for variant prioritization and
414 the identification of causative disease genes. In addition, the generation of high-quality
415 haplotype reference panels for different human populations can be used to improve accuracy by
416 enabling one to impute missing genotypes in large-scale GWAS (Bai et al. 2018; Gurdasani et

417 al. 2019). We therefore further expanded our knowledge of human genetic variation by focusing
418 on the TR population.

419 Here, we present data derived from high-coverage WES and WGS of 3,599 individuals
420 from Turkey and identify 10,116,912 novel variants of which 38,540 are deemed likely to have a
421 deleterious effect. Our results also highlight the importance of population-specific reference
422 panels for increasing the accuracy of imputation, especially for rare variation. Genetic variation
423 in the TR Peninsula has previously been investigated by relatively small scale studies (Alkan et
424 al. 2014; Scott et al. 2016); our data have substantially increased the sample size, and more
425 importantly the representation, from all geographical regions and cities in Turkey. These high-
426 resolution WES and WGS data enabled the detection of previously uncaptured rare variants by
427 the GME Variome.

428 We found that the TR population harbors a considerable proportion of variants that are
429 not yet designated in publicly available databases. Our results show that approximately 21% of
430 all variants identified in this study were specific to the TR population and approximately 39% of
431 the private deleterious variants were not observed in other public databases. The TR Variome
432 also introduces 1,009,435 novel or previously known rare variants, which have a frequency of
433 higher than 1% in the TR population. Although DAF calculations revealed strong correlations, we
434 observed that neither gnomAD nor GME was sufficient to represent the allele frequencies of a
435 marked number of TR variants. Since allele frequency information is critical for Mendelian
436 disease gene identification studies as well as variant prioritization strategies, the TR Variome will
437 provide valuable data to facilitate the exclusion of low-probability candidates.

438 The phenotypic consequences of LoF mutations have long been investigated as a means
439 to define gene function (Saleheen et al. 2017). Naturally occurring homozygous LoFs in humans,
440 also termed ‘human knockouts’, provide invaluable information in this context. However, it is not
441 always easy to interpret their phenotypic consequences (if any) because of issues arising during
442 sequence data analysis and differences in the phenotypic impact of knocking out different genes

(Narasimhan et al. 2016b). Sequencing consanguineous populations is one of the most efficient ways to expand the list of knockouts (Narasimhan et al. 2016a). Consistent with elevated rates of consanguinity in Turkey, we detected several individuals with very high inbreeding coefficients and increased lengths of ROH, which facilitated the discovery of human knockouts. Our list of homozygous pLoFs should contribute to the study of gene function through human knockouts. Moreover, TR individuals carried 2-30 variants classified by the HGMD as DMs and 0-19 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 disease, late-onset disease, variable expression, and reduced penetrance (Xue et al. 2012; Cooper et al. 2013). Further, 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 (Abouelhoda et al. 2016). 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.

## Methods

### Study Samples

Study samples comprised 4,194 TR individuals who either yielded whole exome (WES, *n* = 3,402) or whole genome (WGS, *n* = 792) sequence data, which were collected through different projects related to the molecular bases of human genetic disease (Table S1). We excluded 215 variants in the genes that were causally associated with the phenotypes in our cohort (Table S9). Written informed consent was obtained from all study participants during the sampling process for each study. All informed consents provided the permission to use the DNA samples and basic demographic information for disease gene identification studies and to share the data.

### Sequencing and filtering

WES was performed at the Yale Center for Genome Analysis, TUBITAK or Macrogen using IDT xGen Exome Research 392 Panel v1.0 capture, Roche SeCap EZ Whole Exome V3 or Agilent SureSelect Human All Exon V6 kits according to the manufacturer's protocol. Samples were sequenced on the HiSeq4000 platform with 100-bp paired end-reads. The Illumina processing pipeline was used for base calling, read filtering, and demultiplexing. The read pairs were mapped to the human genome build GRCh37 using Burrows-Wheeler Aligner (BWA) v.0.7.17 (Li and Durbin 2009). Duplicate reads were marked using Mark Duplicates tool in Picard tools. Base quality score recalibration (BQSR) and local realignment around indels were carried out with Genome Analysis Toolkit v.3.7 (GATK) (McKenna et al. 2010). Variant discovery was performed following Best Practices workflows of GATK. HaplotypeCaller was employed to call variants, followed by joint genotyping using GenotypeGVCFs and splitting multiallelic variants with LeftAlignAndTrimVariants. To remove batch effects from the WES data, genotype calling was limited to the intersection of target regions of exome sequencing kits that overlap with CCDS build 15 coding exons (Farrell et al. 2014).

WGS was performed on the Illumina HiSeq 2500 platform using PCR-free library

preparation and 100-bp paired-end sequencing. Reads were aligned to the hg19 human genome build using BWA. The variants were called by the Isaac variant caller (https://github.com/sequencing/isaac_variant_caller). The gVCF files for all WGS samples were jointly genotyped using Illumina gvcfgenotyper (https://github.com/Illumina/gvcfgenotyper). Normalization, realignment around indels and splitting multiallelic variants were performed using BCFtools. The final joint VCF file was lifted over to human genome build GRCh37 with Picard tools using hg19 to b37 chain file, which was downloaded from UCSC website.

Statistical outliers of WES and WGS samples were evaluated separately using BCFtools stats. After the filtration according to the number of singletons, transition/transversion ratio, average depth and total number of variants, 89 WES samples were removed from the dataset because they fell outside five absolute deviations from the median. We did not identify any low-quality samples for WGS batch. Coverage calculations of the WES and WGS samples were performed using VarAFT tool (Desvignes et al. 2018) and BCFtools. For the selection of high-quality variants from the WES and WGS data, we used the following thresholds: a) Variants with Phred-scaled quality score < 30, b) genotypes with depth (DP) < 8, c) genotype quality < 20, and d) a missingness rate higher than 20% across all samples. In addition, variant quality score recalibration (VQSR) was performed for WES samples as implemented in GATK VariantRecalibrator. Variant recalibration was applied by ApplyRecalibration walker of GATK using tranche sensitivity of 99.5% for SNPs and 99.0% for indels. VQSR was used to define low quality variants for downstream processing. To detect the potential batch effects between WES and exome regions of WGS, we calculated the mean number of novel variants (not present in dbSNP build 151) in the two batches and observed that the proportions of novel variants were consistent with the thresholds selected for minimum allele count and depth (Fakhro et al. 2016)(Table S2). Relatedness analysis was performed using KING (Manichaikul et al. 2010) and a kinship coefficient threshold 0.0884 was used to exclude second degree or closer relatives. 506 samples were removed after this step. Finally, 773 WGS and 2,826 WES samples

corresponding to a total of 3,599 individuals and 48,308,918 variants constituted the downstream population structure and variome characterization studies. The GRCh38 positions of the variants were obtained with Picard tools using the hg19 to GRCh38 chain file, which was downloaded from the UCSC website. 48,114,758 (99.59%) variants were successfully lifted over to GRCh38. For population structure analyses, intersection of the target regions of the kits that were used during exome sequencing and CCDS regions were selected from WGS data and combined with WES data.

**Population structure analyses**

13 populations from the 1000GP data were used in comparative analyses: African populations YRI and LWK; European populations GBR, TSI, IBS and FIN; South Asian populations GIH, BEB, PJL and ITU; East Asian populations CHB, CHS and JPT (Table S3). Exome region was selected using the same interval list that was used in TR sequence data and relatedness analysis was performed as previously described. SNPs were extracted from the VCF files of all WES, exome portions of WGS and 1000GP samples using BCFtools. We then combined the SNPs of all three VCF files to demonstrate the population structure of the TR population. After merging the TR and the 1000GP population samples, variants were filtered according to missingness (> 20%), deviation from Hardy-Weinberg equilibrium with a $p$ value of < 0.00005, minor allele frequency (MAF < 0.05) and linkage disequilibrium ($r^2$ = 0.5) using PLINK v.1.9 (Chang et al. 2015). 16,171 variants remained for the analyses of population structure after filtration according to the above-mentioned criteria. All plots were generated with the aid of ggplot2 (Wickham 2009), reshape (Zhang 2016), dplyr (Wickham et al. 2018) and stringr (Wickham 2019) packages of R software.

***Origin of alleles*:** Grand-maternal and grand-paternal birthplace of the 1,598 (44.4%) individuals were obtained from patient records and the numbers of chromosomes from each region were depicted on a map of Turkey.

***Admixture*:** Substructures of the populations were assessed using ADMIXTURE (Alexander et al. 2009). Analysis with *k* from 2 to 12 was run for all TR individuals (*n* = 3,599) in which *k* = 5 resulted in the lowest cross-validation error. Analysis with *k* from 2 to 12 was also run for the origin known TR (*n* = 1,598) and the 1000GP populations (*n* = 1,299) and *k* = 8 was selected as the optimal number since the cross-validation error was lowest when 8 ancestral populations were present.

***Principal components analysis*:** EIGENSOFT SmartPCA (Patterson et al. 2006) tool was used to demonstrate the degree of genetic variation between the populations. Three different PCAs were performed: The first was to explore the variation in Turkey using only the origin-known TR population. The second was to explore the variation in a global context by using all samples included in the study, and the third was to display the close relationship of the TR individuals with known origin and the European populations.

***Procrustes analysis:*** A symmetric Procrustes analysis with 100,000 permutations was performed to evaluate the relationship of geographical distribution and genetic similarity of the TR individuals. The values of the first two PCs of the PCA estimated using the origin-known TR population were employed in the Procrustes analysis. The unprojected geographic coordinates (latitude-longitude) of the TR subregions were determined using geographical midpoints on the map of Turkey.

***Phylogenetic tree:*** Population splitting and genetic drift were evaluated by a maximum likelihood phylogenetic tree using Treemix (Pickrell and Pritchard 2012) software. Greater Middle Eastern populations were included in this analysis by using their allele frequency data.

***Wright's fixation index:*** The degree of differentiation among the populations was evaluated with Fst values produced by Weir and Cockerham estimation, which is included in the EIGENSOFT SmartPCA.

***Linkage disequilibrium decay:*** PLINK --r2 option with 70 kb sliding window and no limit for r2 was used to calculate Pairwise correlations; they were binned by genomic distance between the

SNPs (up to 70 kb), and averages were calculated for each bin.

***Inbreeding coefficient:*** PLINK --het algorithm was used to determine the inbreeding coefficients (*F*) of the individuals. We detected several individuals with negative F values, which could reflect a recent admixture of previously diverse populations or biased variant sampling (Hunter-Zinck et al. 2010).

***Runs of homozygosity:*** Autosomal SNPs of unrelated WGS samples were used to detect ROH. SNPs with minor allele frequencies lower than 0.05 and those that diverted from Hardy-Weinberg equilibrium with $p < 0.00005$ were removed (Ceballos et al. 2018a). The lengths of homozygous regions were calculated using PLINK --homozyg option. With a 50 SNP-containing 50 kb sliding window, ROH longer than 300 kb in length were determined. Three heterozygous calls were allowed during the analysis (Ceballos et al. 2018a).

**Variome characterization**

***Derived allele frequencies:*** Ancestral sequences for *Homo sapiens* (GRCh37), which were generated using the information from Ensembl compara and include the multiple sequence alignment of six primates, were downloaded from the 1000GP FTP site. WES and WGS VCF files were separately annotated with the ancestral alleles using Jvarkit, vcfancestralalleles tool (Lindenbaum 2015). gnomAD WES and WGS VCFs and GME variants were downloaded and annotated using the same ancestral alleles. DAFs were calculated only for variant sites where an ancestral allele is present.

***Functional annotation:*** Variants were annotated by ENSEMBL v.87 (Hunt et al. 2018) using SnpEff v.4.4 (Cingolani et al. 2012) to determine variant functional region and impact on the assigned gene. The HC-pLoFs (frameshift, essential splice site, stop gain, stop loss and start loss) were detected using LOFTEE, which is a VEP plugin designed to identify HC-LoF variants based on their ancestral state, transcript information, and splice prediction (MacArthur et al. 2012). The classification of the missense variants according to their predicted deleteriousness

was performed using PolyPhen-2, SIFT, and CADD (Adzhubei et al. 2010; Vaser et al. 2016). PolyPhen-2 classifies the missense variants as B (benign), P (possibly damaging) or D (probably damaging) whereas SIFT classifies them as T (Tolerated) or D (deleterious). We categorized the missense variants as deleterious if they were listed as "D" in both PolyPhen-2 and SIFT and had a CADD score > 20; it was classified as "other missense" in the rest of the outcomes. Variants were also annotated by the ANNOVAR v.2019Oct24 (Wang et al. 2010) tool using the data from dbnsfp35a, which includes PolyPhen-2 and GERP++ scores (Liu et al. 2016), gnomAD (Karczewski et al. 2020), 1000GP (Auton et al. 2015), the NHBLI GO Exome Sequencing Project (ESP) Exome variant server (https://evs.gs.washington.edu/EVS/ ) and GME (Scott et al. 2016) databases. Annotations were performed separately for "high-quality" WES (*n* = 1,436,012) and WGS (*n* = 47,322,283) variants. Additionally, pLOF variants were annotated with gnomAD_pLI scores. Then, we listed variants detected both in WES and WGS (*n* = 449,378) and re-calculated their allele frequencies. For Fig. 5, variants were classified as "novel", if there was no record in dbSNP build 151, gnomAD, 1000GP or ESP. Variants were classified as common if the variant had an AF ≥ 0.01 in any of the above-mentioned databases. If the variant had AF < 0.01 in all databases, it was classified as rare.

***Human knockouts:*** 1,389 homozygous HC-pLoFs were identified in the TR Variome and 783 of these had an AF < 0.01. Additionally, homozygous pLOFs of gnomAD, and 1000GP were extracted and previously published lists of human-knock-outs including Iceland (Sulem et al. 2015), GME (Scott et al. 2016), PROMIS (Saleheen et al. 2017), British Pakistani (Narasimhan et al. 2016a) and GenomeAsia (GenomeAsia100KConsortium 2019) were downloaded and compared with our list of rare homozygous pLoF variants. The common homozygous pLoFs, which have a frequency in the TR population ≥ 0.01, were listed using HC-pLoF variants. The list was compared to the previously published list of common pLoFs in the ExAC and gnomAD (Rausell et al. 2020).

***Medically relevant variants:*** We annotated the variants that were identified in the TR Variome

against HGMD Professional v.2020.2, ClinVar (Accessed September 9th 2020), and OMIM (Accessed December 10th 2019) (Stenson et al. 2020; Landrum et al. 2018). Only disease-causing pathological mutations (DMs) in HGMD and pathogenic or pathogenic/likely pathogenic variants in ClinVar were used for further analyses. Inheritance types of the phenotypes were extracted from the OMIM database, where applicable.

**Imputation panel**

We used a similar approach to that of previous publications for the generation of the TR reference panel and the evaluation of the imputation performance. Haplotypes of 773 TR individuals were constructed for each autosomal chromosome with BEAGLE v5.1 (Browning et al. 2018) using the high-quality SNPs sequenced with WGS (*n* = 45,698,551). The BEAGLE genotypes re-phased using SHAPEIT v2 (Delaneau et al. 2013) to generate the final TR reference panel. Re-phasing was performed using default parameters except for a window size of 0.5, as it produces more accurate results for sequencing data. To evaluate the performance of the TR reference panel for predicting missing genotypes, we randomly subsampled 73 individuals by extracting their genotypes from unphased WGS data and removed their haplotypes from the TR reference panel. Then, using chromosome 20 variants from the 73 individuals, we generated a Pseudo-GWAS panel, which comprised the 44,367 SNPs represented on Infinium Omni2.5-8 Kit. 1000GP Phase 3 haplotypes were downloaded from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html for comparison with the new TR reference panel. The imputation was performed by IMPUTE2 (Howie et al. 2009) on chromosome 20 split into 5 Mb chunks with 250 kb buffer regions using: 1) 1000GP reference panel, 2) TR reference panel, 3) TR + 1000GP reference panels to predict the “masked” genotypes of the 73 individuals. We used the default parameters of IMPUTE2 except for setting k_hap (Number of reference haplotypes used as templates) to 10,000 since diverse reference panels could contain more useful haplotypes than expected. Squared Pearson’s correlation

coefficients ($R^2$) were calculated between the masked sequence genotypes (0,1,2) and the imputed genotype dosages (0-2), to compare the performance of imputation using each reference haplotype panel. The $R^2$ results were plotted against non-overlapping AF bins. A Wilcoxon rank-sum test was performed to evaluate the statistical significance of the $R^2$ results. The summary file produced by IMPUTE2 was used to show the number of variants with different expected $R^2$ results and expected AF bins for each reference panel.

**Data Access**

**Acknowledgements**

**Disclosure Declaration**

**References**


Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. 2016. Revisiting the morbid genome of Mendelian disorders. *Genome Biology* **17**: 235-235.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248-249.

Akbayram S, Sari N, Akgün C, Doğan M, Tuncer O, Caksen H, Oner AF. 2009. The frequency of consanguineous marriage in eastern Turkey. *Genet Couns* **20**: 207-214.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655-1664.

Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, Dal E, Bugra K, Güngör T, Sahinalp SC et al. 2014. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics* **15**: 963.

Amos W, Hoffman JI. 2010. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society B: Biological Sciences* **277**: 131-137.

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

674 Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, Zhang Y, Bond SR, Pei Z, Zhang D et al. 2018.
675 Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture
676 and gene flow throughout North and East Asia. *Nat Genet* **50**: 1696-1704.

677 Belkadi A, Pedergnana V, Cobat A, Itan Y, Vincent QB, Abhyankar A, Shang L, El Baghdadi J,
678 Bousfiha A, Alcais A et al. 2016. Whole-exome sequencing to analyze population structure,
679 parental inbreeding, and familial linkage. *Proc Natl Acad Sci U S A* **113**: 6713-6718.

680 Berkman CC, Dinc H Fau - Sekeryapan C, Sekeryapan C Fau - Togan I, Togan I. 2008. Alu
681 insertion polymorphisms and an assessment of the genetic contribution of Central Asia to
682 Anatolia with respect to the Balkans. *Am J Phys Anthropol* **136**: 11-18.

683 Bittles AH, Black ML. 2010. Evolution in health and medicine Sackler colloquium: Consanguinity,
684 human evolution, and complex diseases. *Proc Natl Acad Sci U S A* **107 Suppl 1**: 1779-1786.

685 Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next-generation
686 reference panels. *Am J Hum Genet* **103**: 338-348.

687 Ceballos FC, Alvarez G. 2013. Royal dynasties as human inbreeding laboratories: the
688 Habsburgs. *Heredity* **111**: 114-121.

689 Ceballos FC, Hazelhurst S, Ramsay M. 2018a. Assessing runs of homozygosity: a comparison
690 of SNP Array and whole genome sequence low coverage data. *BMC Genomics* **19**: 106.

691 Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018b. Runs of homozygosity:
692 windows into population history and trait architecture. *Nat Rev Genet* **19**: 220-234.

693 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation
694 PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.

Cingolani P, Platts A, Wang lL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

Cinnioğlu C, King R, Kivisild T, Kalfoğlu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K et al. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* **114**: 127-148.

Clay CC. 1998. Labour migration and economic conditions in nineteenth-century Anatolia. *Middle Eastern Studies* **34**: 1-32.

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics* **132**: 1077-1130.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**: 5-6.

Desvignes JP, Bartoli M, Delague V, Krahn M, Miltgen M, Béroud C, Salgado D. 2018. VarAFT: a variant annotation and filtration system for human next generation sequencing data. *Nucleic Acids Res* **46**: W545-W553.

Di Benedetto G, Ergüven A, Stenico M, Castrì L, Bertorelle G, Togan I, Barbujani G. 2001. DNA diversity and population admixture in Anatolia. *Am J Phys Anthropol* **115**: 144-156.

Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, Al-Marri AA, Abi Khalil C, Al-Shakaki A, Chidiac O et al. 2016. The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum Genome Var* **3:** 16016.

717 Farrell CM, O'Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, Diekhans M, Barrell D,
718 Searle SM, Aken B et al. 2014. Current status and new features of the Consensus Coding
719 Sequence database. *Nucleic Acids Res* **42**: D865-872.

720 Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, Amini A, Arzhangi S,
721 Jalalvand K, Jamali P et al. 2019. Iranome: A catalog of genomic variations in the Iranian
722 population. *Hum Mutat* **40**: 1968-1984.

723 Feldman M, Fernández-Domínguez E, Reynolds L, Baird D, Pearson J, Hershkovitz I, May H,
724 Goring-Morris N, Benz M, Gresky J et al. 2019. Late Pleistocene human genome suggests a
725 local origin for the first farmers of central Anatolia. *Nat Commun* **10**: 1218.

726 GenomeAsia100KConsortium. 2019. The GenomeAsia 100K Project enables genetic
727 discoveries across Asia. *Nature* **576**: 106-111.

728 Golden PB. 1992. *An Introduction to the History of the Turkic Peoples. Ethnogenesis and State-*
729 *Formation in Medieval and Early Modern Eurasia and the Middle East.* Otto Harrassowitz,
730 Wiesbaden.

731 Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, Bouman H,
732 Abascal F, Haber M, Tachmazidou I et al. 2019. Uganda Genome Resource enables insights
733 into population history and genomic discovery in Africa. *Cell* **179**: 984-1002.e1036.

734 Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA,
735 Moreno-Estrada A, Bertranpetit J et al. 2012. Genomic ancestry of North Africans supports back-
736 to-Africa migrations. *PLoS Genet* **8**: e1002397.

737 Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for
738 the next generation of genome-wide association studies. *PLoS Genet* **5**: e1000529.

739 Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM,
740 Trevanion SJ, Flicek P et al. 2018. Ensembl variation resources. *Database (Oxford)* **2018**:
741 bay119.

742 Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, Gohar A, Matthews R, Butler MW,
743 Fuller J, Hackett NR et al. 2010. Population genetic structure of the people of Qatar. *Am J Hum*
744 *Genet* **87**: 17-25.

745 Içduygu A, Toktas Ş, Soner BA. 2008. The politics of population in a nation-building process:
746 emigration of non-Muslims from Turkey. *Ethnic and Racial Studies* **31**: 358-389.

747 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia
748 KM, Ganna A, Birnbaum DP et al. 2020. The mutational constraint spectrum quantified from
749 variation in 141,456 humans. *Nature* **581**: 434-443.

750 Kılınç GM, Koptekin D, Atakuman Ç, Sümer AP, Dönertaş HM, Yaka R, Bilgin CC,
751 Büyükkarakaya AM, Baird D, Altınışık E et al. 2017. Archaeogenomic analysis of the first steps
752 of Neolithization in Anatolia and the Aegean. *Proceedings Biological Sciences* **284**: 20172064.

753 Kılınç GM, Omrak A, Özer F, Günther T, Büyükkarakaya AM, Bıçakçı E, Baird D, Dönertaş HM,
754 Ghalichi A, Yaka R et al. 2016. The demographic development of the first farmers in Anatolia.
755 *Curr Biol* **26**: 2659-2666.

756 Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D,
757 Jang W et al. 2018. ClinVar: improving access to variant interpretations and supporting
758 evidence. *Nucleic Acids Res* **46**: D1062-D1067.

759 Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M,
760 Gamarra B, Sirak K et al. 2016. Genomic insights into the origin of farming in the ancient Near
761 East. *Nature* **536**: 419-424.

762 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
763 *Bioinformatics* **25**: 1754-1760.

764 Lindenbaum P. 2015. JVarkit: java-based utilities for Bioinformatics.
765 doi:http://dx.doi.org/10.6084/m9.figshare.1425030.

766 Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional
767 Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**:
768 235-241.

769 MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L,
770 Habegger L, Pickrell JK, Montgomery SB et al. 2012. A systematic survey of loss-of-function
771 variants in human protein-coding genes. *Science* **335**: 823-828.

772 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship
773 inference in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.

774 Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E,
775 Stewardson K, Fernandes D, Novak M et al. 2015. Genome-wide patterns of selection in 230
776 ancient Eurasians. *Nature* **528**: 499-503.

777 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler
778 D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for
779 analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

780 Mehrjoo Z, Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Ardalani F, Jalalvand K, Arzhangi
781 S, Mohammadi Z, Khoshbakht S et al. 2019. Distinct genetic variation and heterogeneity of the
782 Iranian population. *PLoS Genetics* **15**: e1008385-e1008385.

783 Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates
784 C, Bellary S, Bockett NA et al. 2016a. Health and population effects of rare gene knockouts in
785 adult humans with related parents. *Science* **352**: 474-477.

786 Narasimhan VM, Xue Y, Tyler-Smith C. 2016b. Human Knockout Carriers: Dead, Diseased,
787 Healthy, or Improved? *Trends in Molecular Medicine* **22**: 341-351.

788 Notarangelo LD, Bacchetta R, Casanova JL, Su HC. 2020. Human inborn errors of immunity: An
789 expanding universe. *Sci Immunol* **5**: eabb1662.

790 Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, Aylward W,
791 Storå J, Jakobsson M, Götherström A. 2016. Genomic evidence establishes Anatolia as the
792 source of the European Neolithic gene pool. *Curr Biol* **26**: 270-275.

793 Özçelik T. 2017. Medical genetics and genomic medicine in Turkey: a bright future at a new era
794 in life sciences. *Mol Genet Genomic Med* **5**: 466-472.

795 Özçelik T, Kanaan M, Avraham KB, Yannoukakos D, Mégarbané A, Tadmouri GO, Middleton L,
796 Romeo G, King MC, Levy-Lahad E. 2010. Collaborative genomics for human health and
797 cooperation in the Mediterranean region. *Nat Genet* **42**: 641-645.

798 Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**:
799 e190.

800 Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic
801 patterns of homozygosity in worldwide human populations. *Am J Hum Genet* **91**: 275-292.

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

804 Rausell A, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, Stenson PD, Cooper DN,
805 Patin E, Casanova JL et al. 2020. Common homozygosity for predicted loss-of-function variants
806 reveals both redundant and advantageous effects of dispensable human genes. *Proc Natl Acad*
807 *Sci U S A* **117**: 13626-13636.

808 Raveane A, Aneli S, Montinaro F, Athanasiadis G, Barlera S, Birolo G, Boncoraglio G, Di Blasio
809 AM, Di Gaetano C, Pagani L et al. 2019. Population structure of modern-day Italians reveals
810 patterns of ancient and archaic ancestries in Southern Europe. *Science Advances* **5**: eaaw3492.

811 Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the
812 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**: D886-D894.

813 Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F,
814 Kivisild T et al. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J*
815 *Hum Genet* **67**: 1251-1276.

816 Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won HH, Karczewski
817 KJ, O'Donnell-Luria AH, Samocha KE et al. 2017. Human knockouts and phenotypic analysis in
818 a cohort with a high rate of consanguinity. *Nature* **544**: 235-239.

819 Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B,
820 Abel L et al. 2016. Characterization of Greater Middle Eastern genetic variation for enhanced
821 disease gene discovery. *Nat Genet* **48**: 1071-1076.

822 Skourtanioti E, Erdal YS, Frangipane M, Balossi Restelli F, Yener KA, Pinnock F, Matthiae P,
823 Özbal R, Schoop UD, Guliyev F et al. 2020. Genomic history of Neolithic to Bronze Age
824 Anatolia, Northern Levant, and Southern Caucasus. *Cell* **181**: 1158-1175.e1128.

825 Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S,
826 Phillips AD, Cooper DN. 2020. The Human Gene Mutation Database: (HGMD ®): optimizing its
827 use in a clinical diagnostic or research setting. *Hum Genet* **139**: 1197-1207..

828 Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E,
829 Sigurdsson GT, Jonasdottir A, Sigurdsson A et al. 2015. Identification of a large set of rare
830 complete human knockouts. *Nat Genet* **47**: 448-452.

831 Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, Li JZ. 2013. Long runs of
832 homozygosity are enriched for deleterious variation. *Am J Hum Genet* **93**: 90-102.

833 Tunçbilek E. 1997. Genetic services in Turkey. *Eur J Hum Genet* **5 Suppl 2**: 178-182.

834 Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for
835 genomes. *Nat Protoc* **11**: 1-9.

836 Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from
837 high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

838 Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business
839 Media

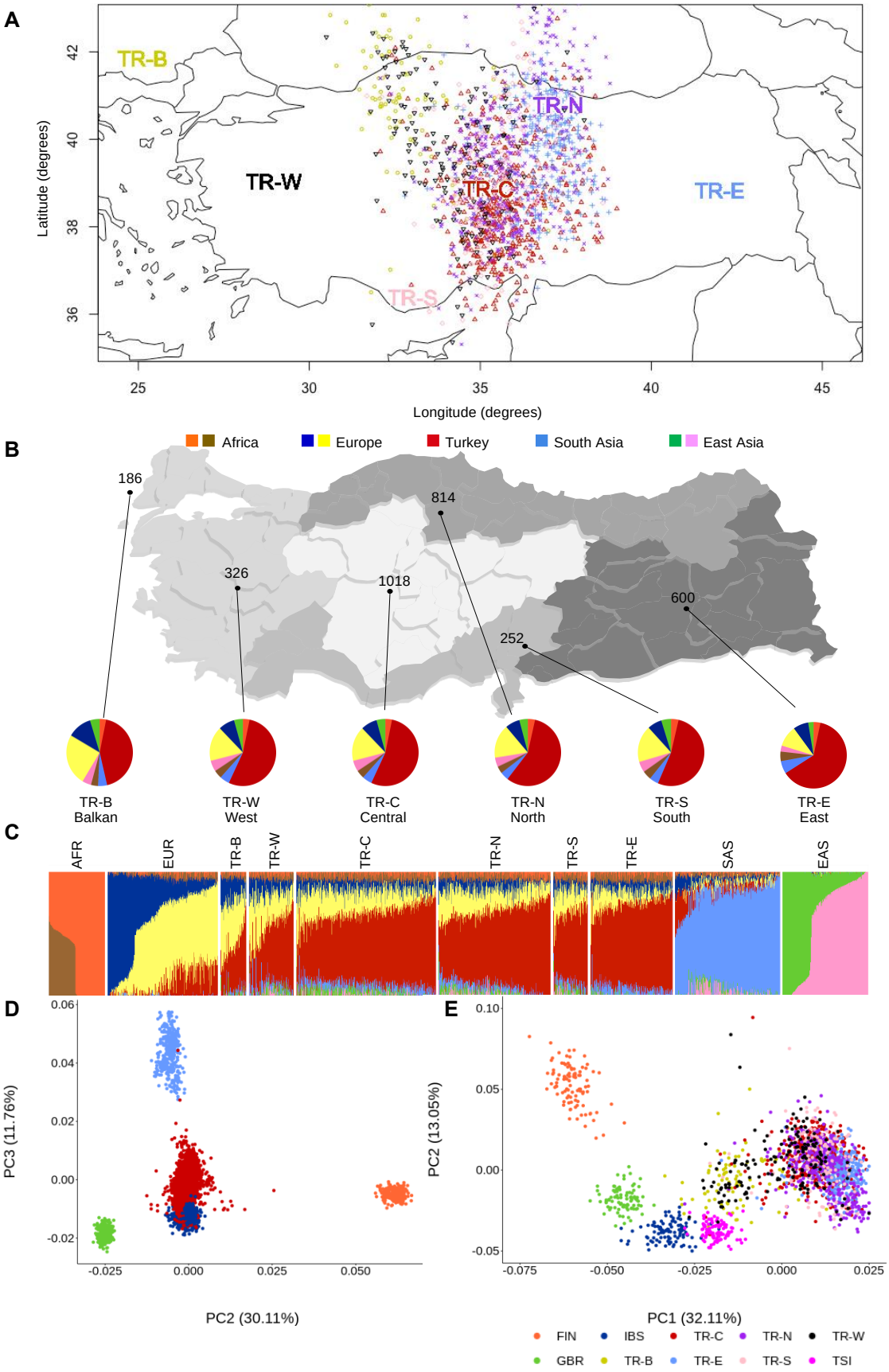
840 Wickham H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations.

841 Wickham H, François R, Henry L, Müller K. 2018. dplyr: A Grammar of Data Manipulation.

842 Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper
843 DN et al. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from
844 current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet*
845 **91**: 1022-1032.

846 Yang X, Al-Bustan S, Feng Q, Guo W, Ma Z, Marafie M, Jacob S, Al-Mulla F, Xu S. 2014. The
847 influence of admixture and consanguinity on population genetic diversity in Middle East. *J Hum*
848 *Genet* **59**: 615-622.

849 Zhang Z. 2016. Reshaping and aggregating data: an introduction to reshape package. *Ann*
850 *Transl Med* **4**: 78.


A
TR-B
TR-N
TR-W
TR-C
TR-E
TR-S
Latitude (degrees)
Longitude (degrees)
B
Africa
Europe
Turkey
South Asia
East Asia
186
814
326
1018
600
252
TR-B Balkan
TR-W West
TR-C Central
TR-N North
TR-S South
TR-E East
C
AFR
EUR
TR-B
TR-W
TR-C
TR-N
TR-S
TR-E
SAS
EAS
D
PC3 (11.76%)
PC2 (30.11%)
E
PC2 (13.05%)
PC1 (32.11%)
FIN
IBS
TR-C
TR-N
TR-W
GBR
TR-B
TR-E
TR-S
TSI

851

**Figure 1. Turkey as a hub of extensive human admixture.** (A) Procrustes analysis based on unprotected coordinates of geographical locations and PC1 and PC2 coordinates of 1,589 TR individuals with known origin. (B) Map of Turkey showing the number of chromosomes and mean admixture proportions of individuals with known birthplaces who originated from present day Turkey and former Ottoman Empire territories (TR-B, Balkan; TR-W, West; TR-C, Central; TR-N, North; TR-S, South; TR-E, East; the map was downloaded from https://www.yourfreetemplates.com/). (C) Admixture results of the TR individuals with known origin and the 1000GP populations. Global ancestry proportions ($k$ = 8) contributed by the 1000GP control populations with seven distinct sources of contribution (orange and camel, sub-Saharan Africa; navy and yellow Europe; light blue, South Asia; green and pink East Asia) (D) PCA for individuals from the TR and 1000GP populations. Individuals were projected along the PC2 and PC3 axes. After separation of African, East Asian and South Asian populations with the first three principal components, the TR population maintained its close relationship with the European populations. (E) PCA of TR individuals with known origin and European populations of 1000GP. The ordering of populations on the PC1 axis was in line with the geographical regions of Turkey, although most of the sub-regions did not form distinct clusters. The variation seen between sub-regions of the TR Peninsula, was significantly higher than that of GBR and TSI.

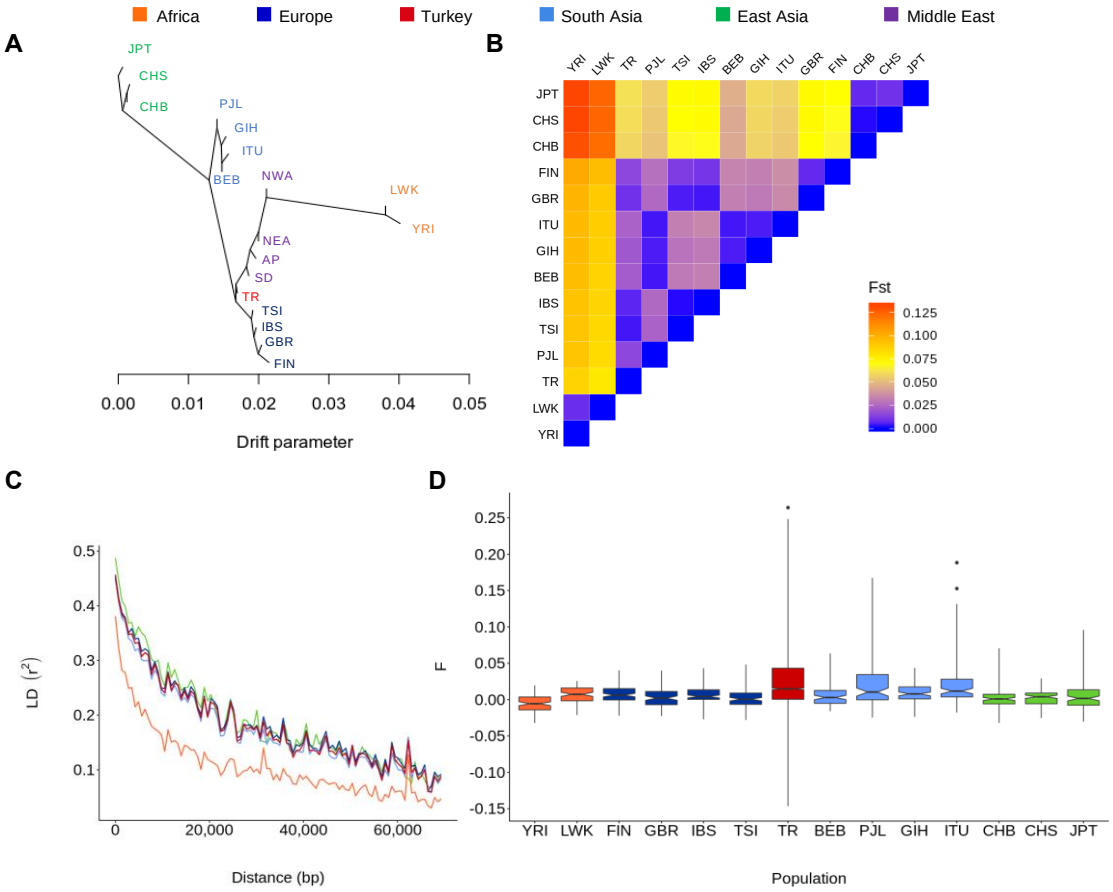



**Figure 2. The Turkish Peninsula as a bridge in the migration trajectories and high inbreeding levels in the TR population.** (A) TreeMix phylogeny of the TR population along with the 1000GP controls and the GME populations representing divergence patterns. The length of branches is proportional to the extent of population drift. The TR population is between the European and GME populations. (B) Wright's fixation index ($F_{ST}$) values for all pairs of the TR and 1000GP populations, showing a smaller distance between the TR and European populations than between the TR and other populations. (C) The TR population showed a similar rate of LD decay to the European (EUR), East Asian (EAS) and South Asian (SAS) populations but a decreased rate in comparison to the African population. Mean variant correlations ($r^2$) are shown for each 700-bp bin over 70,000 bp. (D) Distributions of the inbreeding coefficient (F) for the TR and 1000GP populations. The TR population (red) showed a high number of individuals with elevated F values, consistent with elevated rates of consanguineous marriage. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

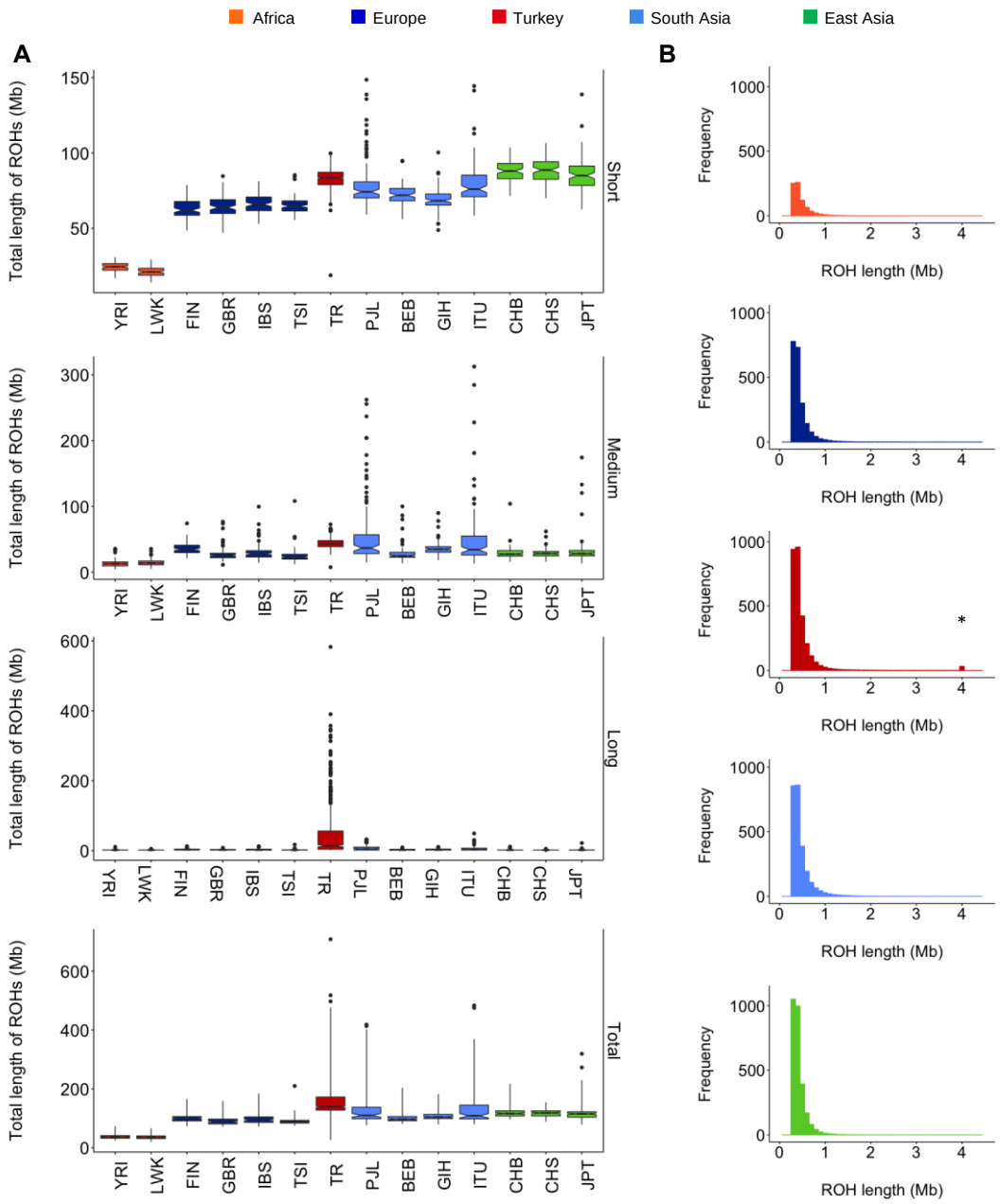



**Figure 3. Distributions of short, medium and long ROH correlate with patterns of bottlenecks and recent consanguinity.** (A) Burden in samples of ROH grouped by length (short, <0.516 Mb; medium 0.516-1,606 Mb; long >1,606 Mb). The TR samples (red) showed a significantly increased number of long ROH in comparison to other populations. (B) Histograms of the frequencies of long ROH in the TR, African, European, South Asian and East Asian populations. Frequencies were calculated by dividing the number of ROH by the population size. ROH > 4 Mb in length are binned together (an asterisk indicates a small peak seen in the TR population).

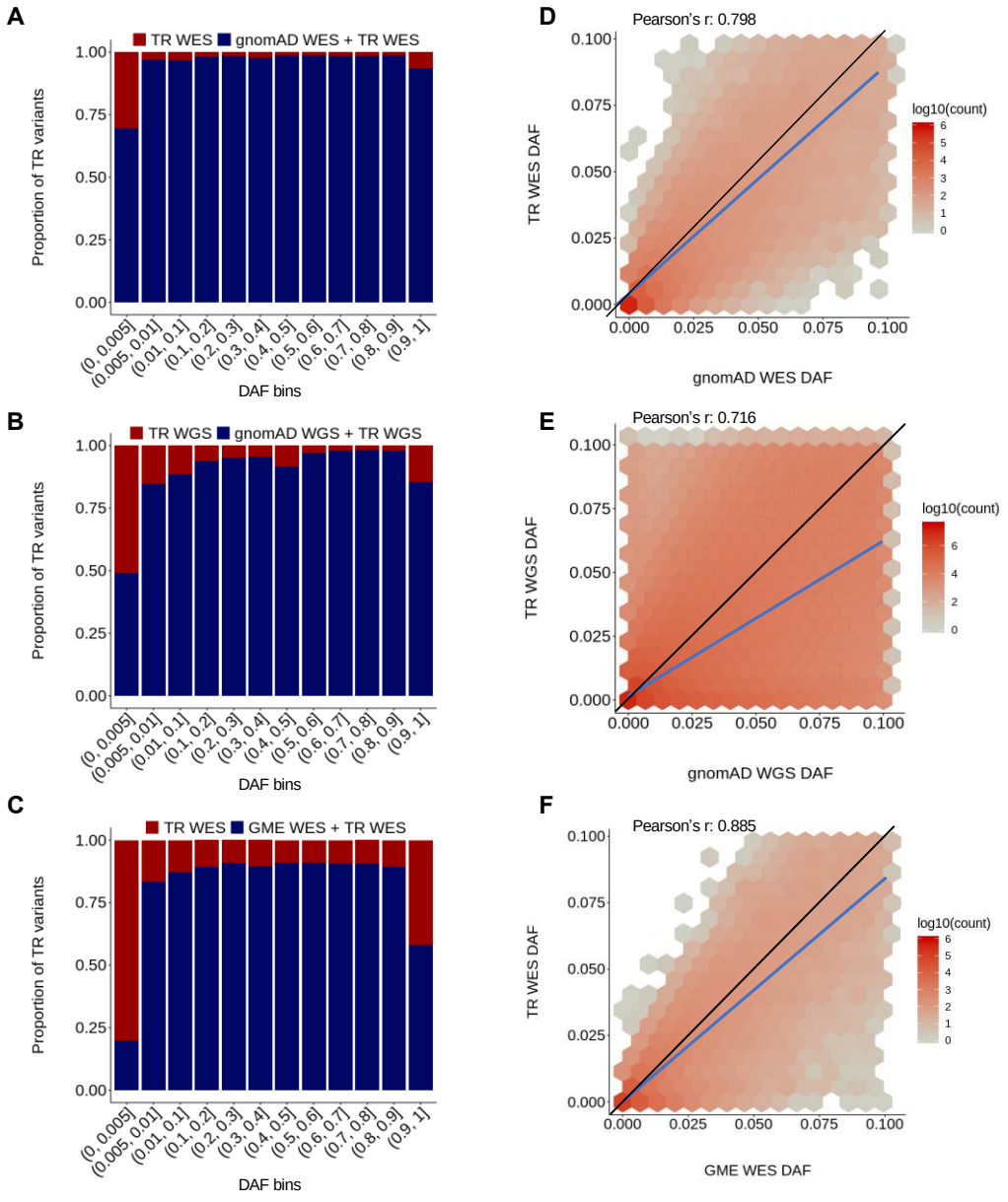



**Figure 4. The TR Variome possesses a significant number of very rare unique variants that are poorly represented in gnomAD and GME.** The proportion of TR variants represented in the TR Variome and other databases. (A) TR WES vs. gnomAD WES, (B) TR WGS vs. gnomAD WGS, (C) TR WES vs. GME WES. The correlation of DAFs of rare TR variants in the (D) TR WES vs. gnomAD WES, (E) TR WGS vs. gnomAD WGS, (F) TR WES vs. GME WES. Hexagonal bins are shaded by the log-transformed number of variants in each bin. Although Pearson's *r* values point to a strong correlation between the TR - gnomAD and TR - GME DAFs, a remarkable number of TR DAFs could not be accurately estimated using gnomAD or GME Variome because of a wide range in allele frequencies between the datasets.

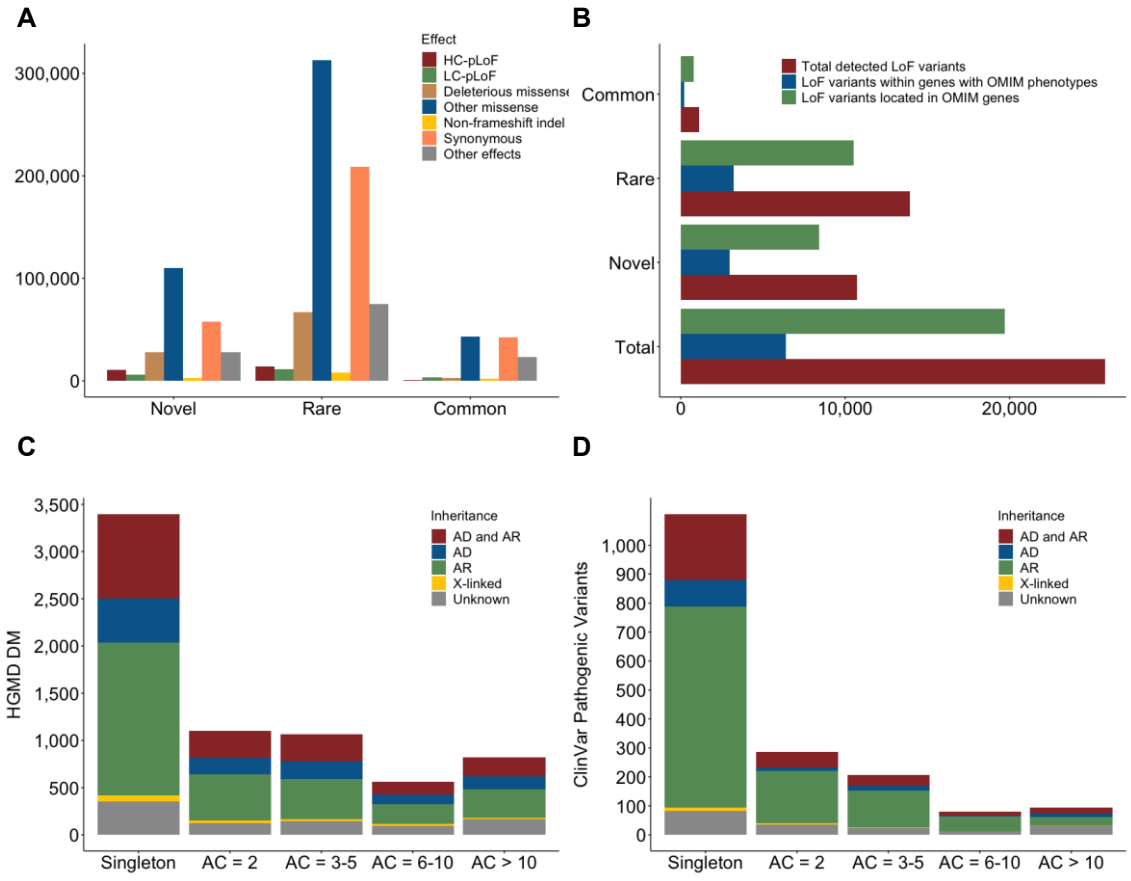



**Figure 5. Distribution of variants based on their frequency, pathogenicity and OMIM phenotypes.** (A) Variants were classified as HC-pLoF, LC-pLoF, missense, non-frameshift indel, synonymous, and other effects as well as according to their frequency. Non-coding variants are not included in the Figure. (B) Proportions of HC-pLoF variants, which were grouped based on their frequency, their location on OMIM genes and the genes with OMIM phenotypes. (C) Distribution of disease-causing pathological mutations from HGMD and (D) pathogenic or pathogenic/likely pathogenic variants from the ClinVar database, categorized based on their frequency and inheritance type as autosomal-dominant (AD) or autosomal-recessive (AR), X-linked or unknown.

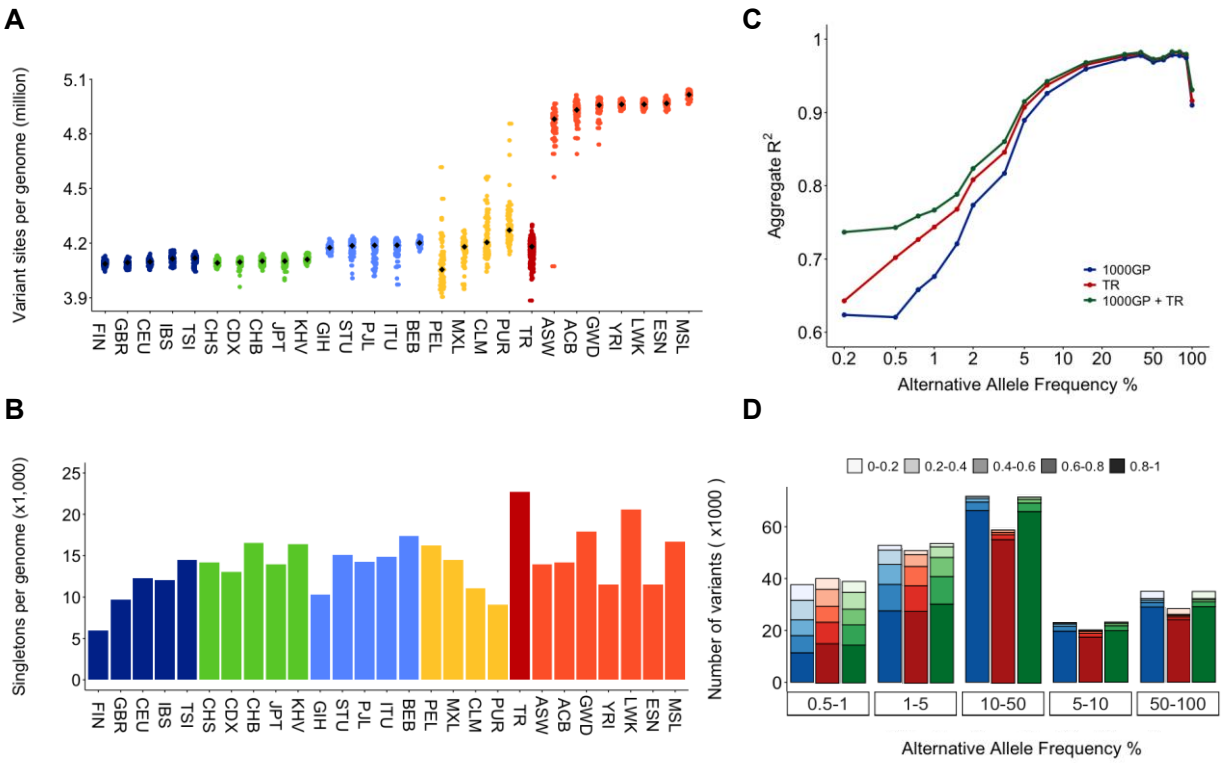



**Figure 6. Per genome variant summary and imputation.** (A) The number of variant sites per genome for autosomes. (B) The average number of singletons per genome for autosomes. (C) Evaluation of imputation performance on chromosome 20. The aggregate squared Pearson correlation coefficient ($R^2$) was 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 ± s.d.: 0.88 ± 0.12) versus 1000GP (mean ± s.d.: 0.86 ± 0.14); P = 0.002 for the TR + 1000GP (mean ± s.d.: 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).

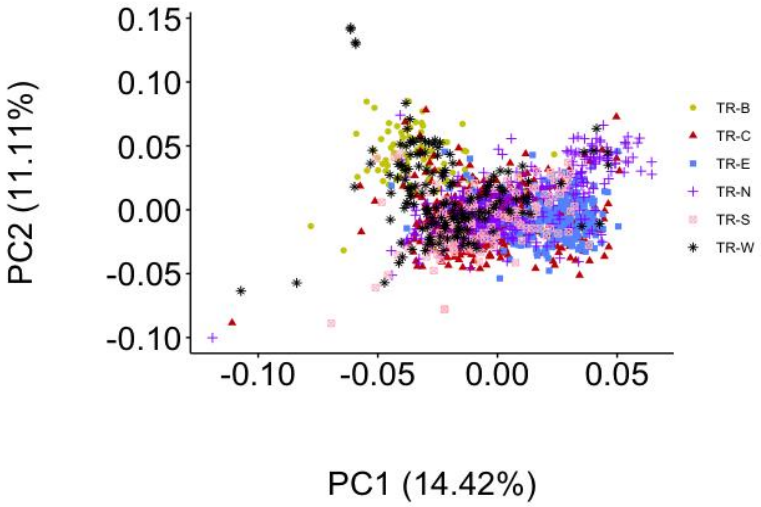
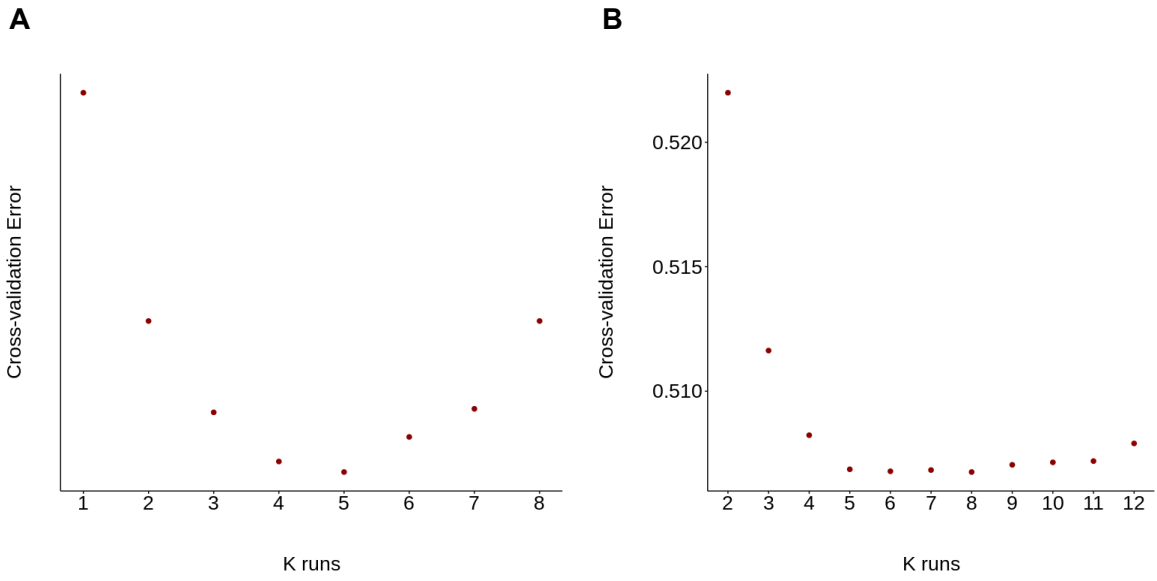
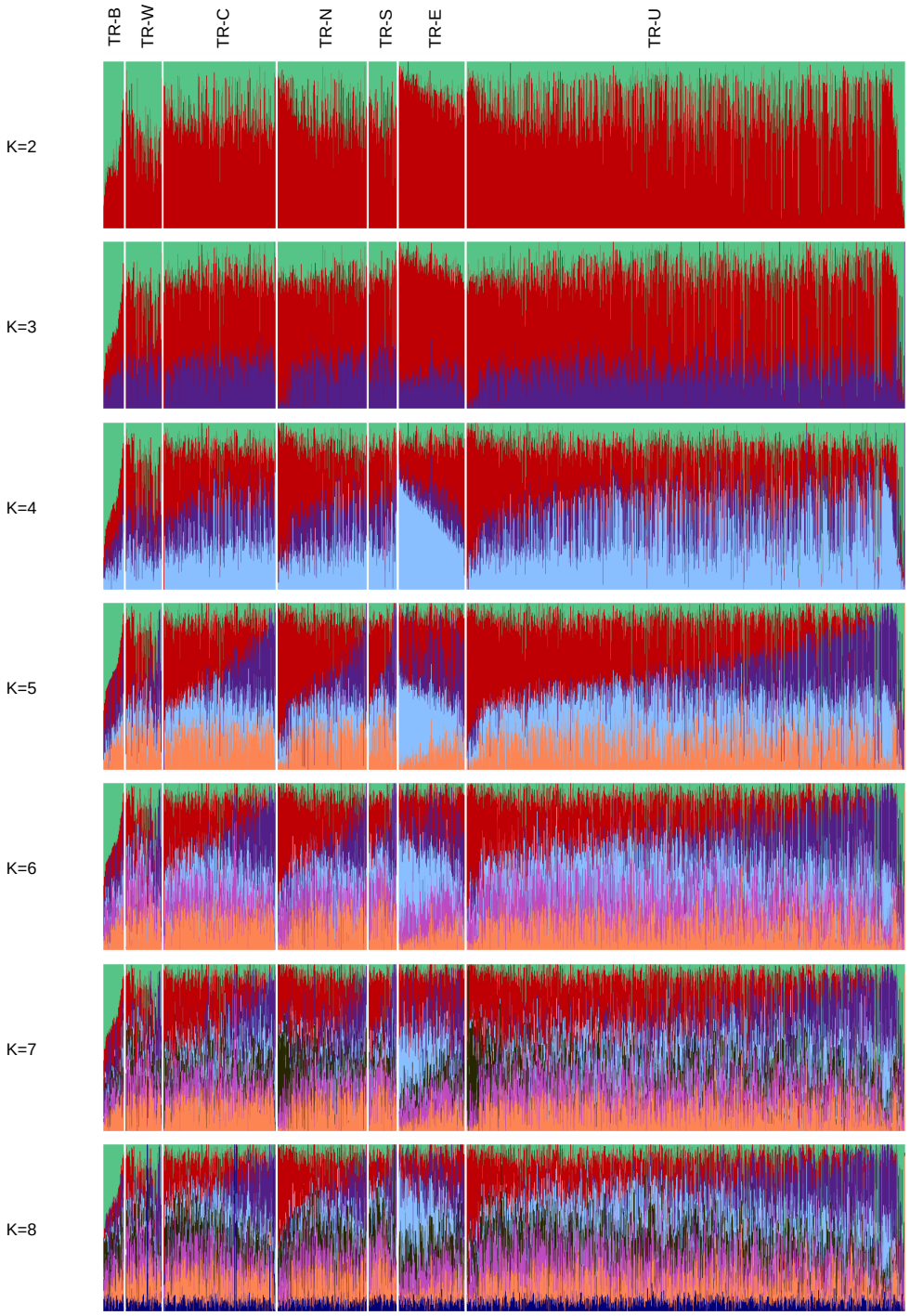



**Figure S1. Principal component analysis on TR individuals with known origin.** Plots for the PC1 and PC2, which explain 14.42% and 11.11% of the total variation seen in Turkey. PC1 distinguishes TR-W, TR-B and TR-E subregions.




924

925 **Figure S2. ADMIXTURE cross-validation.** (A) Cross-validation errors for TR subpopulations according to
926 geographical regions of Turkey. *k* = 5 gave the lowest cross-validation error. (B) Cross-validation errors for
927 the TR and 1000GP samples. Analysis with eight ancestries (*k* = 8) resulted in the lowest cross-validation
928 error.




929

930 **Figure S3. Unsupervised ADMIXTURE analysis of the TR population for clusters *k* = 2 to *k* = 8.**

931 Samples grouped by geographical region and organized from west (left) to east (right). Each column

932 represents an individual. The y axis represents a proportion ranging from 0 to 1. More color suggests

933 multiple ancestral components.

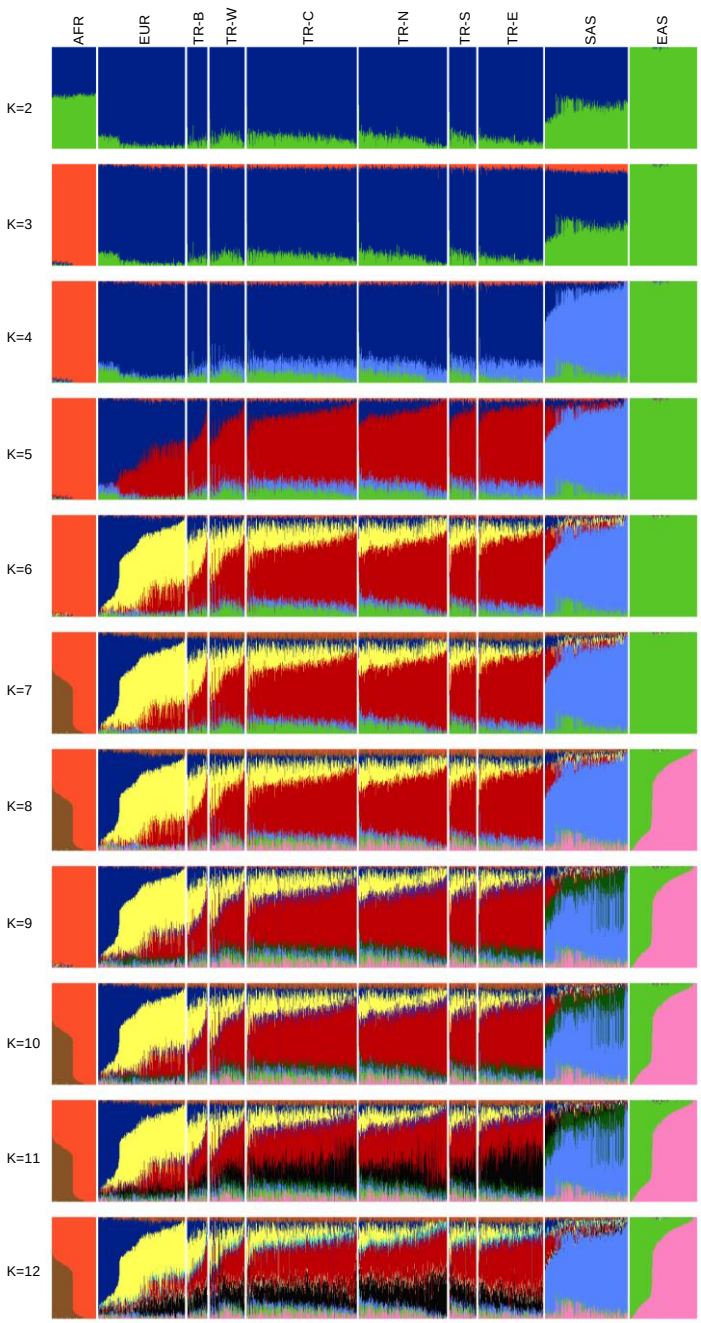



**Figure S4. Unsupervised ADMIXTURE analysis of the TR population in a global context for clusters *k* = 2 to *k* = 12.** Samples from Turkey and 1000GP populations grouped by geographical region and organized from west (left) to east (right), showing trends of overlap. All subregions of the TR Peninsula appear to be represented by multiple ancestral components, including a single ancestry unique to Turkey, two ancestries predominant in European populations, as well as small proportions of Asian and African components.

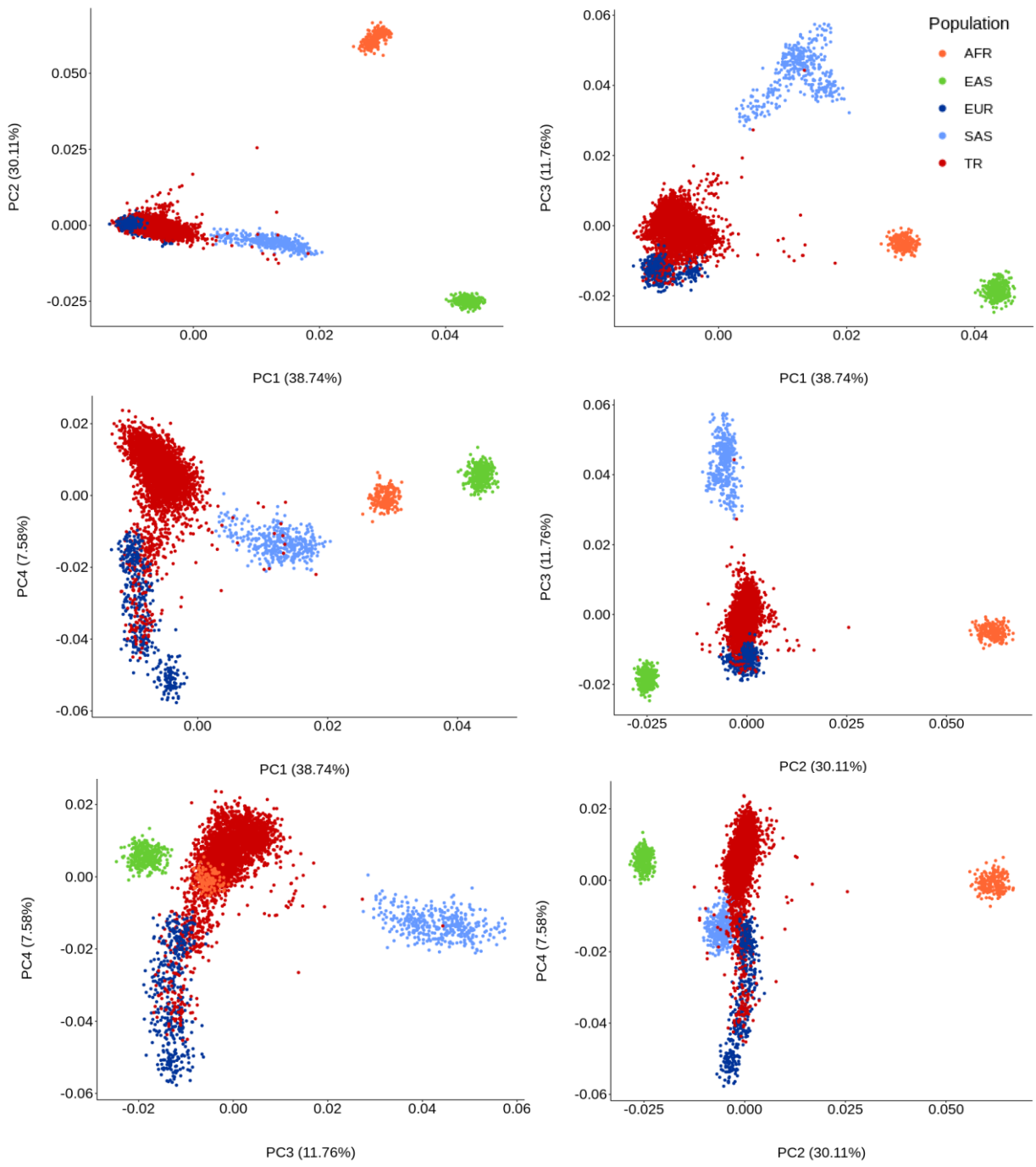
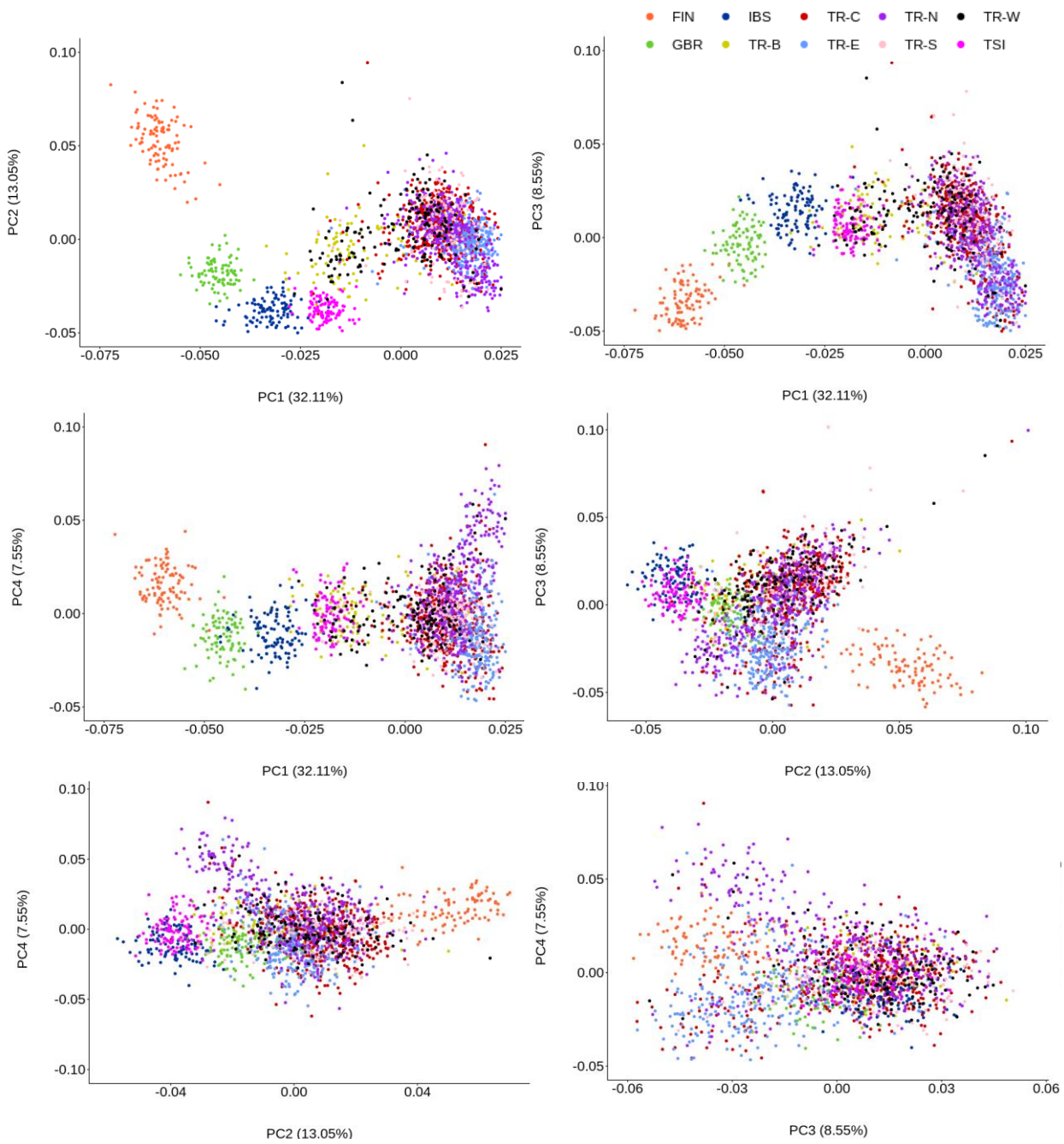



**Figure S5. Principal-component analysis on the TR and 1000GP populations.** Plots for the first four principal components and percentages of variance explained. PC1 (38.74%) and PC2 (30.11%) separate Africans and East Asians respectively from the other populations, while PC3 separates SAS from the TR population. PC4 demonstrated the degree of variance between the TR and European Populations, although it failed to generate distinct clusters.




947

948 **Figure S6. Principal-component analysis on TR individuals with known origin and 1000GP**

949 **European populations.** Plots for the first four principal components and percentages of variance

950 explained. PC1 (32.11%) separated the TR and European populations, although there was a significant

951 overlap between TSI, TR-B and TR-W. PC2 (13.05%) defined the European subpopulations.

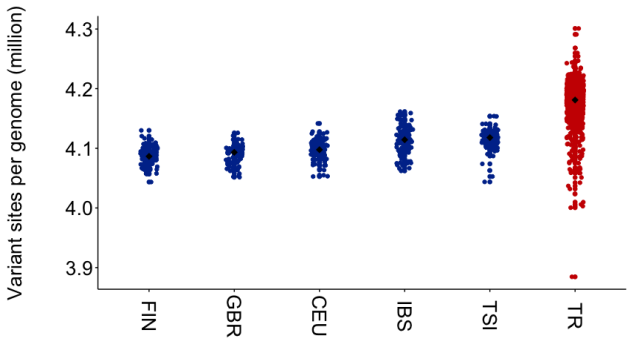



**Figure S7. The number of variant sites per genome for the 1000GP and TR populations.** The average number of variant sites per genome is higher in the TR population than in the European populations.

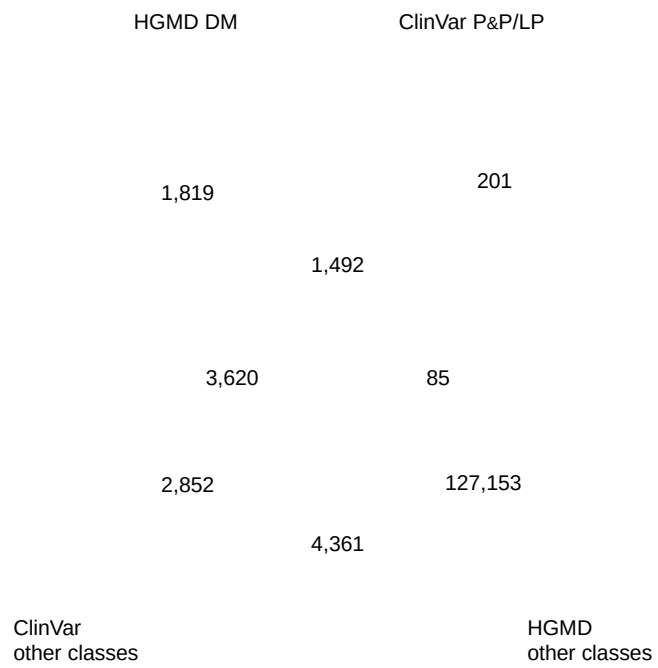



**Figure S8. The Venn Diagram showing the number of the TR variants catalogued in HGMD and/or ClinVar.** HGMD had a higher number of records in the class of disease-causing pathological mutations (DM) when compared to the number of variants that were classified as pathogenic (P) or pathogenic/likely pathogenic (P/LP) in ClinVar.

**Table S1**. **The TR Variome Summary**

| Cohort | *n* | Method |
|---|---|---|
| Amyotrophic lateral sclerosis | 238 | WES |
| Ataxia | 269 | WES |
| Delayed sleep phase disorder | 19 | WES |
| Essential tremors | 154 | WES |
| Obesity | 987 | WES |
| Parkinson's disease | 53 | WES |
| Polycystic ovarian syndrome | 123 | WES |
| Various neurological and immunological disorders | 1,559 | WES |
| Amyotrophic lateral sclerosis | 792 | WGS |
| **Total** | 3,402+792=4,194 | |

**Table S2. Sample based-quality control measures for the integration of the coding regions of WES and WGS data.**

| Sequencing type | Whole genome | | Whole exome | |
|---|---|---|---|---|
| Sample size | 773 | | 2,826 | |
| Minimum depth | 8 | | 8 | |
| Minimum allele count | 1 | | 1 | |
| Mean depth | 25 | | 66 | |
| Novel variants per sample (% not in dbSNP 151) | 0.56% | | 0.55% | |
| | All | Novel | All | Novel |
| Variant alleles | 34,676 | 144 | 24,883 | 102 |
| Heterozygotes | 16,646 | 142 | 11,271 | 98 |
| Variant homozygotes | 9,015 | 1 | 6,806 | 2 |

967 **Table S3. Populations included in the study**

| Abbreviation | Population | Abbreviation for Super population | Super population |
|---|---|---|---|
| TR-B | Turkish with Balkan Ancestry | TR | Turkish |
| TR-W | Western Turkish | TR | Turkish |
| TR-C | Central Turkish | TR | Turkish |
| TR-N | Northern Turkish | TR | Turkish |
| TR-S | Southern Turkish | TR | Turkish |
| TR-E | Eastern Turkish | TR | Turkish |
| TR-U | Turkish with unknown origin | TR | Turkish |
| CHB | Han Chinese in Beijing, China | EAS | East Asian |
| JPT | Japanese in Tokyo, Japan | EAS | East Asian |
| CHS | Southern Han Chinese | EAS | East Asian |
| CDX | Chinese Dai in Xishuangbanna, China | EAS | East Asian |
| KHV | Kinh in Ho Chi Minh City, Vietnam | EAS | East Asian |
| CEU | Utah Residents (CEPH) with Northern and Western European Ancestry | EUR | European |
| TSI | Toscani in Italia | EUR | European |
| FIN | Finnish in Finland | EUR | European |
| GBR | British in England and Scotland | EUR | European |
| IBS | Iberian population in Spain | EUR | European |
| YRI | Yoruba in Ibadan, Nigeria | AFR | African |
| LWK | Luhya in Webuye, Kenya | AFR | African |
| GWD | Gambian in Western Divisions in the Gambia | AFR | African |
| MSL | Mende in Sierra Leone | AFR | African |
| ESN | Esan in Nigeria | AFR | African |
| ASW | Americans of African Ancestry in SW USA | AFR | African |
| ACB | African Caribbeans in Barbados | AFR | African |
| MXL | Mexican Ancestry from Los Angeles, CA, USA | AMR | American |
| PUR | Puerto Ricans from Puerto Rico | AMR | American |
| CLM | Colombians from Medellin, Colombia | AMR | American |
| PEL | Peruvians from Lima, Peru | AMR | American |
| GIH | Gujarati Indian from Houston, Texas, USA | SAS | South Asian |
| PJL | Punjabi from Lahore, Pakistan | SAS | South Asian |
| BEB | Bengali from Bangladesh | SAS | South Asian |
| STU | Sri Lankan Tamil from the UK | SAS | South Asian |
| ITU | Indian Telugu from the UK | SAS | South Asian |

968

969 **Table S4. Functional annotation and allele frequency distribution of TR variants.**

| | | AF in other public databases | | |
|---|---|---|---|---|
| | | Novel | Rare (AF < 0.01) | Common (AF ≥ 0.01) |
| **All variants** | | 10,116,912 | 24,191,524 | 14,000,482 |
| **Functional consequence** | | | | |
| **High-confidence pLoFs** | Frameshift variant | 4,553 | 4,428 | 616 |
| | Splice site variant | 3,147 | 3,905 | 253 |
| | Start loss variant | 1 | | |
| | Stop gain variant | 3,029 | 5,615 | 249 |
| | Stop loss variant | 4 | 2 | 2 |
| **Low-confidence pLoFs** | Frameshift variant | 2,212 | 3,148 | 816 |
| | Splice site variant | 1,867 | 3,430 | 1,877 |
| | Start loss variant | 462 | 1,059 | 162 |
| | Stop gain variant | 1,309 | 3,154 | 363 |
| | Stop loss variant | 342 | 556 | 157 |
| **Missense variants** | Deleterious missense | 27,806 | 67,046 | 2,833 |
| | Other missense | 110,354 | 313,185 | 43,487 |
| **Non-frameshift indels** | | 2,767 | 7,926 | 1,993 |
| **Synonymous variants** | | 57,474 | 208,811 | 42,245 |
| **Other effects** | Protein-protein contact | 154 | 384 | 42 |
| | Exon loss variant | | 2 | |
| | Gene fusion | 12 | 29 | 8 |
| | Structural interaction variant | 3,834 | 11,858 | 1,312 |
| | Bidirectional gene fusion | 16 | 39 | 16 |
| | Transcription Factor Binding Site (TFBS) ablation | 118 | 273 | 102 |
| | Non-essential splice site variant | 23,938 | 62,068 | 21,803 |
| | Initiator codon variant | 35 | 63 | 9 |
| | Stop retained variant | 90 | 194 | 56 |
| **Non-coding variants** | Intergenic region | 3,646,158 | 8,611,593 | 5,376,533 |
| | Intragenic variant | 838 | 1,835 | 995 |
| | Intron variant | 3,562,033 | 8,488,164 | 4,924,545 |
| | Upstream gene variant | 1,403,332 | 3,343,586 | 1,878,325 |
| | TFBS variant | 10,660 | 23,818 | 10,443 |
| | Sequence feature | 71,806 | 173,577 | 94,447 |
| | Downstream gene variant | 995,982 | 2,412,439 | 1,388,865 |
| | Non-coding transcript exon variant | 26,787 | 67,369 | 40,201 |
| | Untranslated region (UTR) variant | 155,792 | 371,968 | 167,727 |

970

971