

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/144097/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Kars, M. Ece, Basak, A. Nazli, Onat, O. Emre, Bilguvar, Kaya, Choi, Jungmin, Itan, Yuval, Çalar, Caner, Palvadeau, Robin, Casanova, Jean-Laurent, Cooper, David N. ORCID: <https://orcid.org/0000-0002-8943-8484>, Stenson, Peter D., Yavuz, Alper, Bulus, Hakan, Günel, Murat, Friedman, Jeffrey M. and Özçelik, Tayfun 2021. The genetic structure of the Turkish population reveals high levels of variation and admixture. *Proceedings of the National Academy of Sciences* 118 (36) , e2026076118. 10.1073/pnas.2026076118 file

Publishers page: <http://dx.doi.org/10.1073/pnas.2026076118>
<<http://dx.doi.org/10.1073/pnas.2026076118>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies.

See

<http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



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

3
4 M. Ece Kars¹, A. Nazlı Başak², O. Emre Onat¹, Kaya Bilguvar³, Jungmin Choi³, Yuval Itan^{4,5},
5 Caner Çağlar⁶, Robin Palvadeau², Jean-Laurent Casanova^{7,8,9,10,11}, David N. Cooper¹², Peter D.
6 Stenson¹², Alper Yavuz¹³, Hakan Buluş¹³, Bülent Yıldız¹⁴, Jeffrey M. Friedman^{6,11}, Tayfun
7 Özçelik^{1,15,16}

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

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

29

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

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

32

33 **Abstract**

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

53 provide an invaluable resource for future studies to identify variants that are associated with
54 specific phenotypes as well as establishing the phenotypic consequences of mutations in
55 specific genes. (279 words)

56

57 **Introduction**

58 Even in the Paleolithic period, Anatolia (or Asia Minor as it was once called), served as a bridge
59 for migrations between Africa, Asia and Europe. Long before the establishment of nation states,
60 intermixing between human populations occurred in Anatolia. Indeed, Anatolia has been home
61 to many civilizations including Hattians, Hurrians, Assyrians, Hittites, Greeks, Thracians,
62 Phrygians, Urartians, Armenians and Turks. Gene flow between Anatolian, Caucasus and
63 Northern Levantine populations occurred during the Late Neolithic and Chalcolithic to Early
64 Bronze age, including long-distance migration from Central Asia to Anatolia (Skourtanioti et al.
65 2020). The Turkic peoples, a collection of ethnolinguistically related populations originating from
66 Central Asia were first documented in Western Eurasia in the 4th-5th century BCE, and currently
67 live in Central, Eastern, Northern and Western Asia as well as in parts of Europe and in North
68 Africa. The expansion of Turkic tribes into Western Asia and Eastern Europe occurred between
69 the 6th and 11th centuries, beginning with the Seljuk Turks followed by the Ottomans (Golden
70 1992). The sphere of Ottoman influence started to increase greatly, beginning in the 14th
71 Century; following the conquest of Constantinople in 1453, the Ottoman Empire controlled a vast
72 region including all of Southeastern Europe south of Vienna, parts of Central Europe, Western
73 Asia, the Caucasus, North Africa and the Horn of Africa. The modern Republic of Turkey was
74 founded in 1923 after the fall of the Ottoman empire at the end of WW1 and is currently home to
75 more than 80 million people. Turkish-speaking people constitute the major ethnolinguistic group
76 in Turkey. There are also more than 70 million people who live in the five independent Turkic
77 countries in Central Asia, namely Azerbaijan, Turkmenistan, Uzbekistan, Kazakhstan and
78 Kyrgyzstan. A study investigating the Y haplogroups of TR males revealed that the proportion of

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

82 The practice of consanguineous marriage is frequent in Turkey, especially in the eastern
83 provinces (Akbayram et al. 2009). This should in principle help to facilitate disease gene
84 discovery as the increased frequency of homozygosity among members of inbred populations
85 has led to the identification of many disease genes (Bittles and Black 2010; Özçelik et al. 2010;
86 Özçelik 2017; Notarangelo et al. 2020). **The genetic admixture and consanguinity have had a**
87 **significant effect on the genetic diversity of Middle Eastern populations (Yang et al. 2014;**
88 **Mehrjoo et al. 2019).**

89 The characterization of the Greater Middle East (GME) Variome, comprising the most
90 comprehensive genomic database for Middle East populations, has shown that knowledge of the
91 genomic architecture of these populations facilitates disease gene identification in family studies
92 and in GWAS studies of populations (Scott et al. 2016). Until now, the GME has been the largest
93 resource representing the genetic variation in Turkey, albeit with only 140 out of a total of 1,111
94 samples coming from the TR Peninsula. Thus, based on the larger population of Turkey relative
95 to its immediate neighbors, the TR population is underrepresented in current genomic
96 databases. Furthermore, gnomAD, as one of the most comprehensive genetic variation
97 resources, does not contain TR WES or WGS data (Karczewski et al. 2020). Therefore, a
98 comprehensive database of alleles in the **TR** population should facilitate disease gene
99 identification in consanguineous families and assessment of the clinical phenotypes of
100 individuals who are homozygous for mutations in specific genes.

101 Finally, most GWAS studies to date have analyzed DNA from European ancestry-derived
102 populations, and it will be important to extend GWAS studies of complex traits to
103 underrepresented populations. One of the key steps in GWAS is to 'predict' or 'impute' the
104 missing genotypes by using a reference haplotype panel. It is becoming increasingly common

105 for researchers to generate such panels for imputation from population-specific WGS data.
106 Population-specific reference panels increase imputation accuracy, especially when they are
107 combined with existing reference panels such as the 1000 Genomes Project (1000GP) (Auton et
108 al. 2015; Bai et al. 2018; Gurdasani et al. 2019). In this study, we have described the high-
109 resolution genetic structure of the TR population, generated a TR Variome and imputed a TR
110 reference panel for future genetics studies.

111

112 **Results**

113 **Population structure of Turkey**

114 We analyzed whole-exome (WES) and whole-genome (WGS) sequence data from 4,194
115 individuals who had participated in genetic studies of obesity, essential tremor, Parkinson's
116 disease, amyotrophic lateral sclerosis, ataxia, delayed sleep phase disorder, polycystic ovarian
117 syndrome, and various assorted neurological and immunological disorders (Table S1). WES and
118 WGS samples were processed separately for analyses of the sample quality and familial
119 relationships using BCFtools. 2,826 WES and 773 WGS samples remained after filtration
120 according to quality metrics and relatedness. The mean target base coverage for the exons of
121 consensus coding sequence (CCDS) build 15 (Farrell et al. 2014) of the WES samples was 68X
122 with 95.33%, 93.83%, and 88.82% coverage at 8X, 10X and 20X or more, respectively. The
123 mean depth of coverage for the WGS samples was 32X with 93%, 91% and 89% coverage at
124 8X, 10X and 20X or more, respectively. We identified 1,981,939 WES and 72,982,375 WGS
125 variants. Following a variant filtration process to identify high-quality variants, we obtained
126 1,436,012 WES and 47,322,283 WGS variants (see Methods). Exome data provides accurate
127 results for population structure analyses (Belkadi et al. 2016). Therefore, we combined the
128 exome part of the WGS data with the WES data for further analyses of population structure. The
129 difference in the mean number of novel variants for the samples sequenced with two different
130 technologies can be used to detect potential batch effects (Fakhro et al. 2016). Accordingly, we

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

134 The geographical origins of ancestors (birthplaces of maternal and paternal
135 grandparents) of 1,598 TR samples were documented and grouped into six different subregions,
136 namely Balkan (TR-B:93), West (TR-W:163), Central (TR-C:509), North (TR-N:407), South (TR-
137 S:126) and East (TR-E:300), and then compared with 13 populations from the 1000GP (Table
138 S3) (Auton et al. 2015). First, we performed a principal component analysis (PCA) using only TR
139 individuals of known origin (Fig. S1). There were no sharp divisions between TR subregions, yet
140 the position of subregions along PC axes was similar to their geographical location. To evaluate
141 the impact of geography in shaping the genomic variability in Turkey, we tested the correlation
142 between geographic and genetic coordinates by applying a Procrustes analysis. Consistent with
143 the results of the PCA, we did not observe a clear-cut distribution of samples among TR
144 subregions, although we did detect a significant mild positive correlation in our dataset (Fig. 1A,
145 correlation in Procrustes rotation, 0.46 $P = 9.99 \times 10^{-6}$).

146 We then evaluated the genetic substructure of Turkey using ADMIXTURE (Alexander et
147 al. 2009) and $k = 5$ was determined as the lowest cross-validation error (Fig. S2A). Individuals
148 with unknown ancestral birthplaces (TR-U) exhibited similar ancestral components to those
149 individuals with known ancestral birthplaces. All five ancestries were represented in each
150 geographical region, although in different proportions (Fig. S3). Surprisingly, when we included
151 samples from African, European, South and East Asian populations from 1000GP, $k = 8$ gave
152 the lowest cross-validation error, which revealed that only one ancestry was specific to the TR
153 population (Fig. 1B, C, Fig. S2B, Fig. S4). We observed small differences in the contributions of
154 primary admixture components from Africa, Europe, South Asia and East Asia in the TR
155 population, reflecting the importance of geographical location in shaping genetic substructure.
156 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

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

185 Researchers investigating founder effects in populations suggest that two ancient
186 population bottlenecks shaped the genetic variation in humans after they migrated out of Africa:
187 the first bottleneck occurred about 50,000 to 60,000 years ago in the Middle East and the
188 second occurred when people crossed the ancient land bridge separating the Bering Strait from
189 the Americas (Amos and Hoffman 2010). Therefore, in order to see whether an ancient
190 population bottleneck was shared between the TR and other populations, we calculated the
191 mean rate of linkage disequilibrium (LD) decay (Fig. 2C). The LD for the TR population decayed
192 more slowly than for the African populations. Yet, a similar rate was observed in the East Asian,
193 South Asian, European and TR populations, supporting the bottleneck hypothesis (Scott et al.
194 2016). The diverse levels of admixture observed in these populations confirm that the results are
195 not due to intermixing, and imply the occurrence of a shared ancient bottleneck.

196

197 **Inbreeding status and estimation of ROH**

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

209 2013). By contrast, the comparable percentages for AFR, EUR, EAS and SAS populations for
210 the same threshold were 3.9%, 4.7%, 8% and 22.7%, respectively.

211 Consanguinity is associated with the increased length and sum total length of ROH,
212 whereas admixture acts to reduce the total number of ROH (Ceballos et al. 2018b). Extended
213 ROH have been shown to be enriched for rare and deleterious variation (Szpiech et al. 2013;
214 Scott et al. 2016). Therefore, we assessed the number and lengths of ROH, based on previously
215 published ranges in the 1000GP populations as well as in the TR population, and compared the
216 results (Pemberton et al. 2012). Similar to previous publications, the smallest median for sum
217 total length of ROH was observed in Sub-Saharan Africans (Pemberton et al. 2012), whilst it was
218 highest in the TR population. Also, in cases of long ROH (≥ 1.607) the TR population displayed
219 the highest numbers of individuals, whereas for the short and medium-length ROH the TR
220 individuals are comparable to the East Asian, South Asian and European populations (Fig. 3A).
221 The frequency calculations showed that ROH longer than 4Mb in length were observed
222 exclusively in the TR population (Fig. 3B): 385 (49.81%) TR individuals had ROH longer than
223 4Mb in length whereas none of the 1000GP individuals had an ROH longer than 4Mb. Moreover,
224 the longest ROH in the TR and 1000GP populations was detected in a TR individual: 41 Mb in
225 length. [TAYFUN: Isn't the length of ROH going to be a function of the SNPs used? The higher
226 the number of SNPs employed, the greater the resolution. The greater the resolution, the more
227 likely are breaks between ROH to be found....which will serve to reduce the length of the ROH.
228 In comparisons between different populations, you are presumably comparing like with like in
229 terms of the numbers of SNPs employed]

230

231 **The TR Variome**

232 The GME Variome has demonstrated the power of consanguinity to identify causes of recessive
233 disease, which are often the result of population-specific mutations (Scott et al. 2016). Thus, the
234 comparison of derived allele frequencies (DAFs) of GME populations with that of the NHBLI GO

235 Exome Sequencing Project (ESP) revealed a large number of variants unique to the GME
236 populations. We therefore investigated the genetic variation in the TR population at higher
237 resolution by searching for TR DAFs in gnomAD (Karczewski et al. 2020) and GME datasets.
238 We observed that approximately 30% of the WES and approximately 50% of the WGS variants
239 in the very rare derived allele frequency bins ($AF < 0.005$) are unique to the TR population (Fig.
240 4A, B). Moreover, approximately 80% of the very rare alleles of the TR population were absent
241 from the GME Variome (Fig. 4C). The heatmaps demonstrating the results of the correlation
242 analyses of the TR and the gnomAD, or the TR and the GME DAFs, revealed that neither is a
243 sufficient estimator for the TR DAFs (Fig. 4 D-F). These results indicate that the GME Variome is
244 an inadequate representation of the TR population.

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

261 ≥ 0.01 in any of the above-mentioned databases. If the variant had an AF < 0.01 in all
262 databases, it was classified as “Rare” (Table S4). Overall, we identified 10,116,912 novel
263 variants of which 38,540 were HC-pLoF or deleterious missense. A total of 1,009,435 variants
264 (2.94%) in the rare and novel categories had an allele frequency higher than 1% in the TR
265 Variome. We also noted that the proportions of HC-pLoFs and deleterious missense variants
266 were higher among the novel and rare categories in the TR Variome, and these results were
267 similar to those of the Iranome study (Fattahi et al. 2019) (Fig. 5A). We also extracted the private
268 variants (variants which are observed in only one individual either in the heterozygous or the
269 homozygous state) of the TR Variome. We detected 26,840,965 private variants of which
270 9,035,278 (33.66%) were not observed in other public databases. A total of 86,779 (0.32%) of
271 the all private variants were HC-pLoFs or deleterious missense variants and 33,771 (0.13%) of
272 these variants were specific to the TR Variome.

273

274 **Human knockouts**

275 Studies performed in populations with a high rate of consanguineous marriage provide
276 researchers with an ideal opportunity to expand the list of naturally occurring human gene
277 knockouts (Narasimhan et al. 2016a; Scott et al. 2016; Saleheen et al. 2017). Since common
278 pLoF variants are less likely either to have a functional effect/clinical impact or to be subject to
279 purifying selection (MacArthur et al. 2012), we first analyzed the number of high-confidence
280 homozygous pLoF variants with an allele frequency lower than 1% in the TR Variome. We
281 identified 783 rare homozygous pLoF variants in 679 genes. These homozygous pLoFs were
282 observed in 631 individuals (20.31%) who each had between 1 and 4 genes knocked out (Table
283 S5). We then cross-compared our list of homozygous pLoFs and the genes carrying those
284 variants with previously reported human knockout lists in Icelanders (Sulem et al. 2015),
285 PROMIS (Saleheen et al. 2017), Pakistanis living in Britain (Narasimhan et al. 2016a) and
286 GenomeAsia (GenomeAsia100KConsortium 2019). We also extracted homozygous pLoFs from

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

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

304

305 **Clinically relevant variants**

306 To demonstrate the potential of the TR Variome for the identification of disease-relevant
307 variants, we first listed the TR Variome HC-pLoF variants and then searched for them in Online
308 Mendelian Inheritance in Man (OMIM). In the TR Variome, we identified 25,804 HC-pLoF
309 variants in 9,634 unique genes. 76.37% of these variants were located under OMIM-listed genes
310 while 24.72% of them were located under OMIM-listed genes with an associated phenotype
311 [TAYFUN: Meaning unclear! You surely cannot mean that 19,000 variants are specifically listed
312 in OMIM. Do you mean the associated genes were listed in OMIM?]. We categorized the HC-

313 pLoF variants, according to their frequency status in other public databases, as either novel, rare
314 or common. The numbers of novel and rare pLoFs were significantly higher than that of the
315 common HC-pLoFs. However, the proportion of HC-pLoF variants in OMIM-listed genes and
316 OMIM-listed genes with an associated clinical phenotype was comparable between classes (Fig.
317 5B). These findings were similar to those derived from the Iranome database (Fattahi et al.
318 2019).

319 We then annotated variants that were identified in the TR Variome against the Human
320 Gene Mutation Database (HGMD) (Stenson et al. 2020) and ClinVar (Landrum et al. 2018)
321 (Fig.S8). 6,931 variants in 2,261 genes from the TR Variome were found to be classified as
322 disease-causing pathological mutations (DM) in HGMD, and these DMs were observed in 3,599
323 individuals (100%) who each harbored between 2 and 30 DMs with an average of 13 (0-5 in the
324 homozygous state) (Fig. 5C, Table S7). 1,778 variants in 982 genes were classified as
325 pathogenic or pathogenic/likely pathogenic in ClinVar and these variants were observed in 3,592
326 (99.8%) individuals who each had between 0 and 19 pathogenic and/or pathogenic/likely
327 pathogenic variants with an average of 6 (0-10 in the homozygous state) (Fig. 5D, Table S8).
328 Importantly, 1,492 variants in 834 genes were found to be DM in HGMD and pathogenic or
329 pathogenic/likely pathogenic in ClinVar (Fig. S8).

330

331 **Per-genome variant summary and imputation panel**

332 The extent of genetic variation in humans differs between populations. For example, individuals
333 with African ancestry harbor a much higher number of variants in their genomes than Europeans
334 (Auton et al. 2015). To compare the genetic structure of the TR population with other populations
335 in terms of genome-wide variation, we first catalogued high-quality variants from the WGS
336 dataset of the TR Variome with up to 20% missingness and imputed the missing sites by
337 BEAGLE v5.1 (Browning et al. 2018). Then, we calculated the number of per-genome variant
338 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

365 reference panel produced higher numbers of high-confidence (expected $R^2 > 0.8$) calls of
366 variants with expected AF < 1% than others, and the combined panel was more beneficial in
367 terms of yielding a higher number of high-confidence variants than both panels for variants with
368 expected AF $\geq 1\%$ (Fig 6D). The TR reference panel added 3,911 high-confidence rare variants
369 (AF < 1%) that were not captured by the 1000GP panel whereas the combined panel added
370 20,951 and 3,902 high-confidence variants (AF $\geq 1\%$) that were not detected with the TR and
371 the 1000GP, respectively.

372

373 Discussion

374 In this report, we delineated the fine-scale genetic structure of the TR population. Consistent
375 with Turkey's location at the crossroads of many historical population migrations, we find a high
376 level of admixture. Studies of ancient DNA suggest that the early farmers of Anatolia in the late
377 Pleistocene period had two significant ancestral contributions from Iran/Caucasus and ancient
378 Levant in addition to the local genetic contribution from Anatolian hunter-gatherers (Feldman et
379 al. 2019). The admixture events in Anatolia extended towards Europe. However, there are also
380 studies, which suggest that the early Neolithic central Anatolians were probably descendants of
381 local hunter-gatherers, rather than immigrants from the Levant or Iran (Kilinc et al. 2017). The
382 migration of early Neolithic Anatolian farmers to Europe was a particularly important move with a
383 significant impact on the genetic structure of pre-existing as well as present-day European
384 populations (Mathieson et al. 2015; Lazaridis et al. 2016). The most prominent effects of this
385 migration are observed in Southern Europe (Omrak et al. 2016; Raveane et al. 2019). Moreover,
386 an additional migration of later Neolithic Anatolian farmers occurred after the early Neolithic
387 spread (Kilinc et al. 2016). Thus, the close genetic relationship of the TR population with the
388 present-day European populations probably reflects these Anatolian migrations to Europe. The
389 mobility of Anatolian and neighboring South Caucasus and North Levantine populations
390 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 11th 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 19th to early 20th 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

443 (Narasimhan et al. 2016b). Sequencing consanguineous populations is one of the most efficient
444 ways to expand the list of knockouts (Narasimhan et al. 2016a). Consistent with elevated rates
445 of consanguinity in Turkey, we detected several individuals with very high inbreeding coefficients
446 and increased lengths of ROH, which facilitated the discovery of human knockouts. Our list of
447 homozygous pLoFs should contribute to the study of gene function through human knockouts.
448 Moreover, TR individuals carried 2-30 variants classified by the HGMD as DMs and 0-19
449 variants classified by the ClinVar as pathogenic or pathogenic/likely pathogenic. These results
450 may have yielded secondary findings, with the potential to provide information on future disease
451 prospects of the individuals concerned. However, such individuals might carry such variants
452 without showing any clinical manifestations for the following reasons: carrying only one copy of
453 the disease allele for a recessive disease, late-onset disease, variable expression, and reduced
454 penetrance (Xue et al. 2012; Cooper et al. 2013). Further, disease gene/variant identification
455 studies in underrepresented populations are far from complete and it is crucial to reassess
456 disease-related databases using different population resources (Abouelhoda et al. 2016).
457 Hence, analyses of the TR Variome will help to establish or exclude specific genes in the
458 pathogenesis of a variety of genetic disorders.

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

464 **Methods**

465 **Study Samples**

466 Study samples comprised 4,194 TR individuals who either yielded whole exome (WES, $n =$
467 3,402) or whole genome (WGS, $n = 792$) sequence data, which were collected through different
468 projects related to the molecular bases of human genetic disease (Table S1). We excluded 215
469 variants in the genes that were causally associated with the phenotypes in our cohort (Table
470 S9). Written informed consent was obtained from all study participants during the sampling
471 process for each study. All informed consents provided the permission to use the DNA samples
472 and basic demographic information for disease gene identification studies and to share the data.

473

474 **Sequencing and filtering**

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

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

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

497 Statistical outliers of WES and WGS samples were evaluated separately using BCFtools
498 stats. After the filtration according to the number of singletons, transition/transversion ratio,
499 average depth and total number of variants, 89 WES samples were removed from the dataset
500 because they fell outside five absolute deviations from the median. We did not identify any low-
501 quality samples for WGS batch. Coverage calculations of the WES and WGS samples were
502 performed using **VarAFT tool (Desvignes et al. 2018) and BCFtools**. For the selection of high-
503 quality variants from the WES and WGS data, we used the following thresholds: a) Variants with
504 Phred-scaled quality score < 30, b) genotypes with depth (DP) < 8, c) genotype quality < 20, and
505 d) a missingness rate higher than 20% across all samples. In addition, variant quality score
506 recalibration (VQSR) was performed for WES samples as implemented in GATK
507 VariantRecalibrator. Variant recalibration was applied by ApplyRecalibration walker of GATK
508 using tranche sensitivity of 99.5% for SNPs and 99.0% for indels. VQSR was used to define low
509 quality variants for downstream processing. To detect the potential batch effects between WES
510 and exome regions of WGS, we calculated the mean number of novel variants (not present in
511 dbSNP build 151) in the two batches and observed that the proportions of novel variants were
512 consistent with the thresholds selected for minimum allele count and depth (Fakhro et al.
513 2016) (**Table S2**). Relatedness analysis was performed using KING (Manichaikul et al. 2010) and
514 a kinship coefficient threshold 0.0884 was used to exclude second degree or closer relatives.
515 506 samples were removed after this step. Finally, 773 WGS and 2,826 WES samples

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

523

524 **Population structure analyses**

525 13 populations from the 1000GP data were used in comparative analyses: African populations
526 YRI and LWK; European populations GBR, TSI, IBS and FIN; South Asian populations GIH,
527 BEB, PJI and ITU; East Asian populations CHB, CHS and JPT (Table S3). Exome region was
528 selected using the same interval list that was used in TR sequence data and relatedness
529 analysis was performed as previously described. SNPs were extracted from the VCF files of all
530 WES, exome portions of WGS and 1000GP samples using BCFtools. We then combined the
531 SNPs of all three VCF files to demonstrate the population structure of the TR population. After
532 merging the TR and the 1000GP population samples, variants were filtered according to
533 missingness (> 20%), deviation from Hardy-Weinberg equilibrium with a p value of < 0.00005,
534 minor allele frequency (MAF < 0.05) and linkage disequilibrium ($r^2 = 0.5$) using PLINK v.1.9
535 (Chang et al. 2015). 16,171 variants remained for the analyses of population structure after
536 filtration according to the above-mentioned criteria. All plots were generated with the aid of
537 ggplot2 (Wickham 2009), reshape (Zhang 2016), dplyr (Wickham et al. 2018) and stringr
538 (Wickham 2019) packages of R software.

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

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

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

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

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

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

566 **Linkage disequilibrium decay:** PLINK --r2 option with 70 kb sliding window and no limit for r^2
567 was used to calculate Pairwise correlations; they were binned by genomic distance between the

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

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

573 **Runs of homozygosity:** Autosomal SNPs of unrelated WGS samples were used to detect
574 ROH. SNPs with minor allele frequencies lower than 0.05 and those that diverted from Hardy-
575 Weinberg equilibrium with $p < 0.00005$ were removed (Ceballos et al. 2018a). The lengths of
576 homozygous regions were calculated using PLINK --homozyg option. With a 50 SNP-containing
577 50 kb sliding window, ROH longer than 300 kb in length were determined. Three heterozygous
578 calls were allowed during the analysis (Ceballos et al. 2018a).

579

580 **Variome characterization**

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

588 **Functional annotation:** Variants were annotated by ENSEMBL v.87 (Hunt et al. 2018) using
589 SnpEff v.4.4 (Cingolani et al. 2012) to determine variant functional region and impact on the
590 assigned gene. The HC-pLoFs (frameshift, essential splice site, stop gain, stop loss and start
591 loss) were detected using LOFTEE, which is a VEP plugin designed to identify HC-LoF variants
592 based on their ancestral state, transcript information, and splice prediction (MacArthur et al.
593 2012). The classification of the missense variants according to their predicted deleteriousness

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

610 **Human knockouts:** 1,389 homozygous HC-pLoFs were identified in the TR Variome and 783 of
611 these had an AF < 0.01. Additionally, homozygous pLoFs of gnomAD, and 1000GP were
612 extracted and previously published lists of human-knock-outs including Iceland (Sulem et al.
613 2015), GME (Scott et al. 2016), PROMIS (Saleheen et al. 2017), British Pakistani (Narasimhan
614 et al. 2016a) and GenomeAsia (GenomeAsia100KConsortium 2019) were downloaded and
615 compared with our list of rare homozygous pLoF variants. The common homozygous pLoFs,
616 which have a frequency in the TR population ≥ 0.01 , were listed using HC-pLoF variants. The list
617 was compared to the previously published list of common pLoFs in the ExAC and gnomAD
618 (Rausell et al. 2020).

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

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

625

626 **Imputation panel**

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

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

652 **Data Access**

653 **Acknowledgements**

654 **Disclosure Declaration**

655 **References**

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

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

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

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

665 Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, Dal E, Bugra K, Güngör T,
666 Sahinalp SC et al. 2014. Whole genome sequencing of Turkish genomes reveals functional
667 private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics*
668 **15**: 963.

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

671 Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S,
672 McVean GA, Abecasis GR et al. 2015. A global reference for human genetic variation. *Nature*
673 **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.

695 Cingolani P, Platts A, Wang IL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A
696 program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:
697 SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-
698 92.

699 Cinnioğlu C, King R, Kivisild T, Kalfoğlu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin
700 AA, Prince K et al. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet*
701 **114**: 127-148.

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

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

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

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

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

714 Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, Al-Marri AA, Abi Khalil C,
715 Al-Shakaki A, Chidiac O et al. 2016. The Qatar genome: a population-specific tool for precision
716 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, Arzhanghi 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, Fadhlouli-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 İçduygu 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, Altınışik 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, Kernytzky 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, Arzhang
781 S, Mohammadi Z, Khoshtakht 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

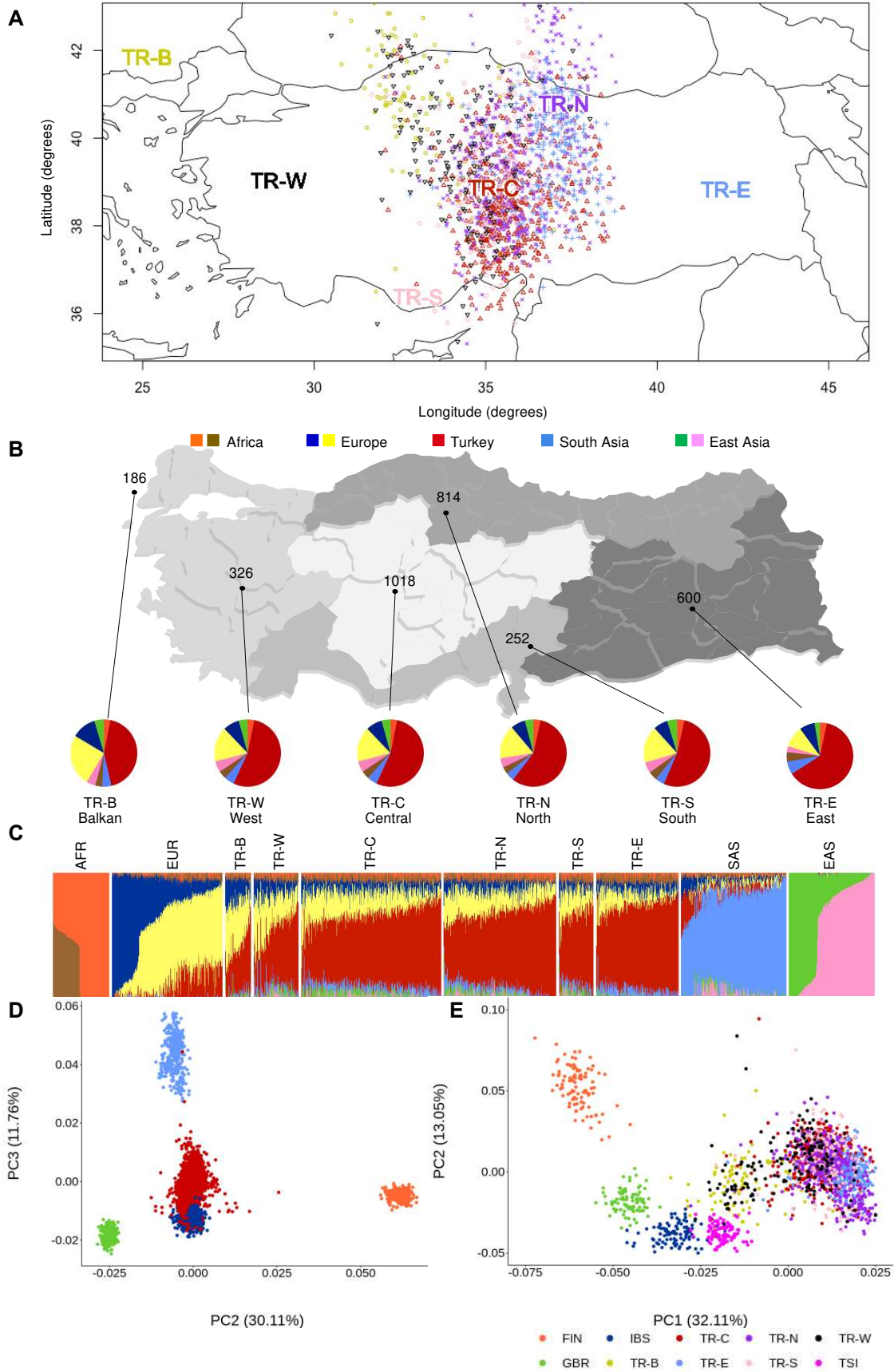
840 Wickham H. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations*.

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

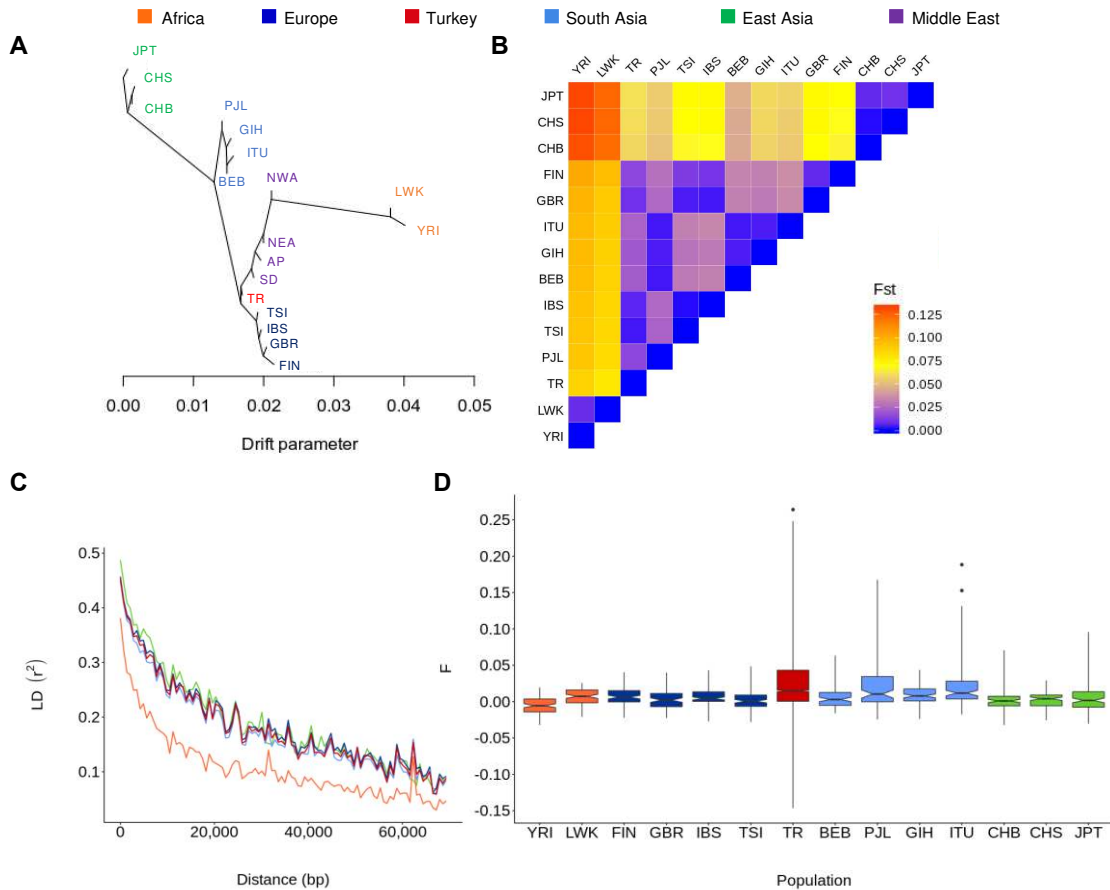
842 Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper
843 DN et al. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from
844 current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet*
845 **91**: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

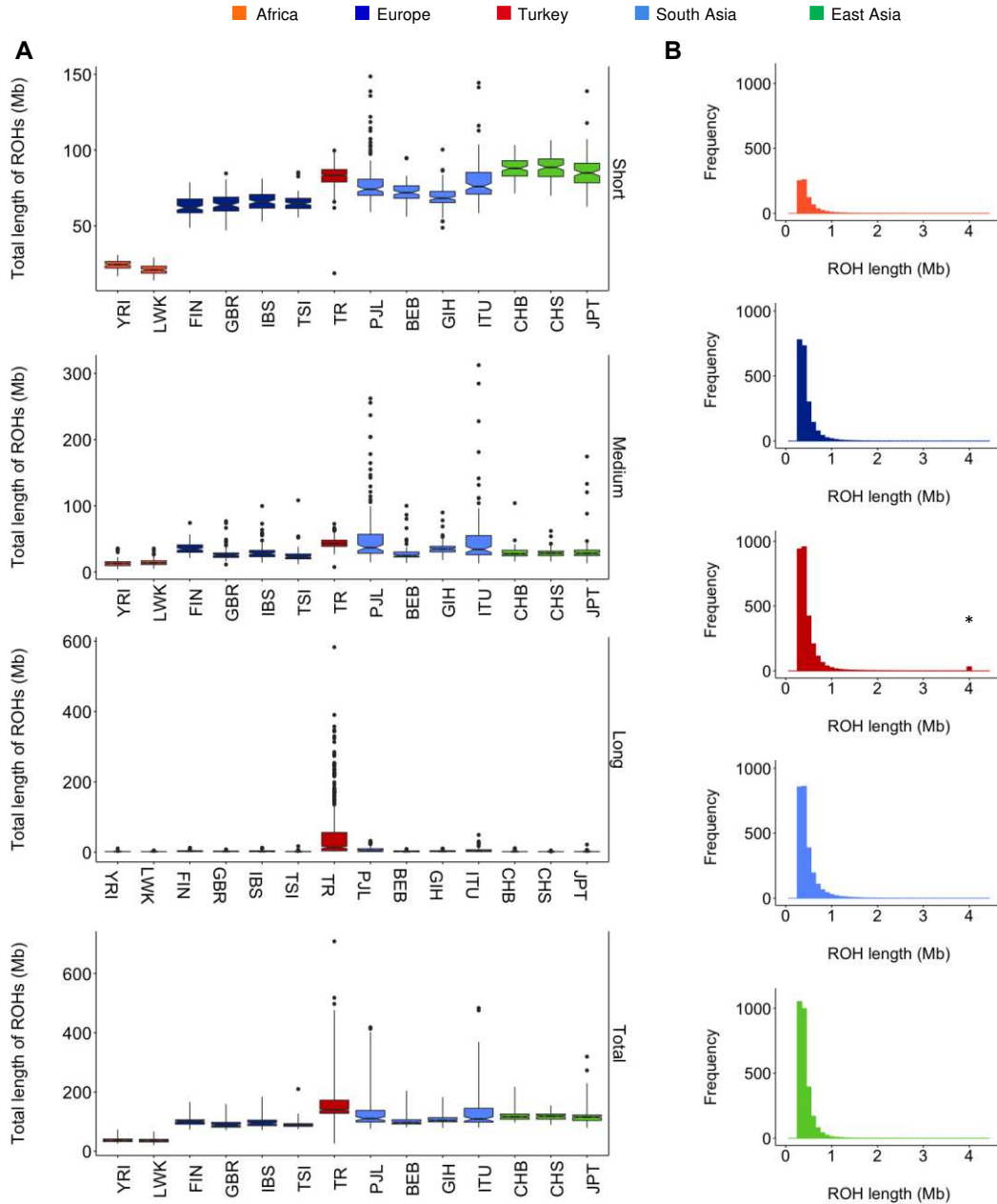
849 Zhang Z. 2016. Reshaping and aggregating data: an introduction to reshape package. *Ann*
850 *Transl Med* **4**: 78.



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



868
 869 **Figure 2. The Turkish Peninsula as a bridge in the migration trajectories and high inbreeding levels**
 870 **in the TR population.** (A) TreeMix phylogeny of the TR population along with the 1000GP controls and
 871 the GME populations representing divergence patterns. The length of branches is proportional to the
 872 extent of population drift. The TR population is between the European and GME populations. (B) Wright's
 873 fixation index (F_{ST}) values for all pairs of the TR and 1000GP populations, showing a smaller distance
 874 between the TR and European populations than between the TR and other populations. (C) The TR
 875 population showed a similar rate of LD decay to the European (EUR), East Asian (EAS) and South Asian
 876 (SAS) populations but a decreased rate in comparison to the African population. Mean variant correlations
 877 (r^2) are shown for each 700-bp bin over 70,000 bp. (D) Distributions of the inbreeding coefficient (F) for the
 878 TR and 1000GP populations. The TR population (red) showed a high number of individuals with elevated
 879 F values, consistent with elevated rates of consanguineous marriage. Box plots show the median
 880 (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum
 881 observations (whiskers).



882

883 **Figure 3. Distributions of short, medium and long ROH correlate with patterns of bottlenecks and**

884 **recent consanguinity.** (A) Burden in samples of ROH grouped by length (short, <0.516 Mb; medium

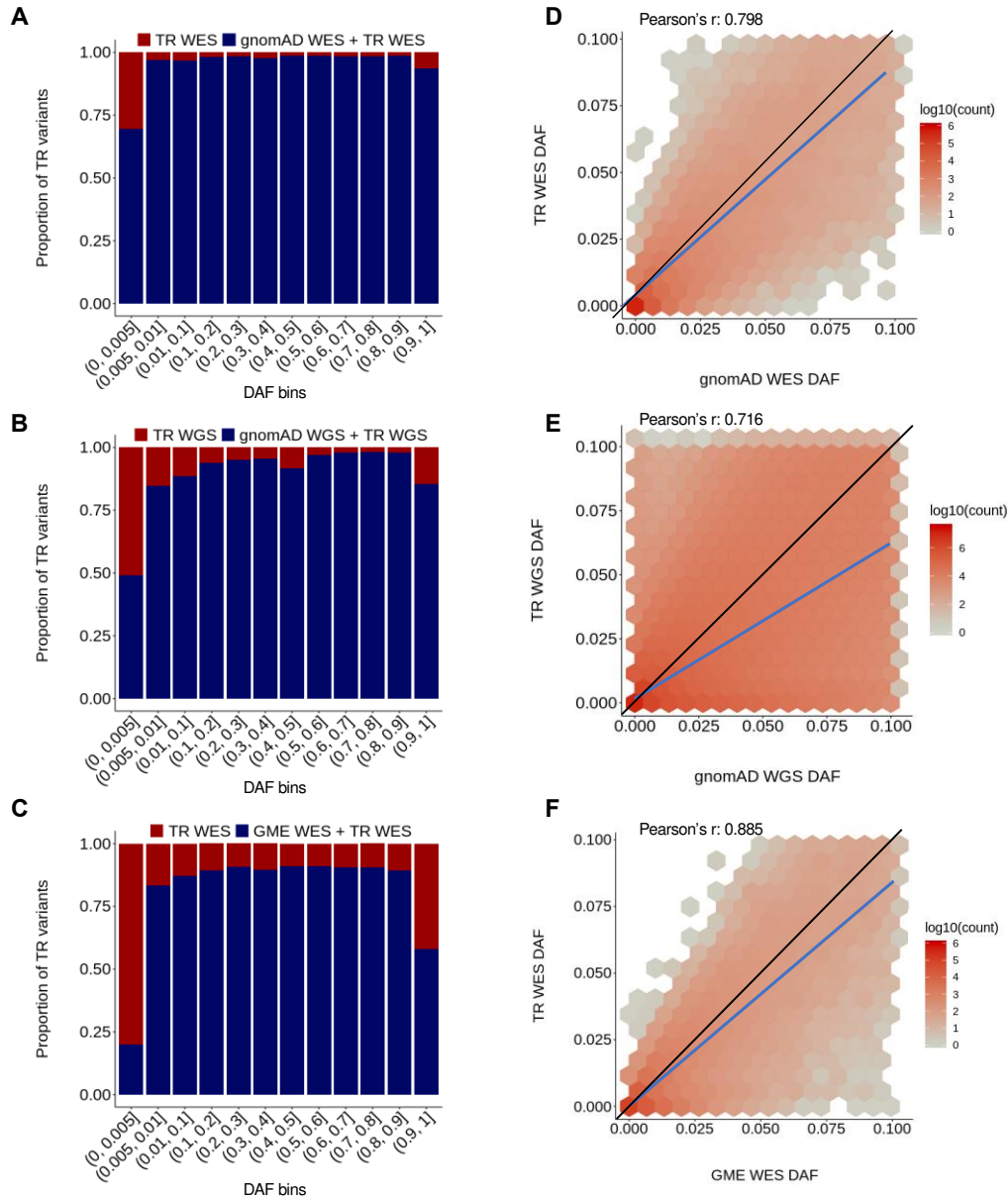
885 0.516-1,606 Mb; long >1,606 Mb). The TR samples (red) showed a significantly increased number of long

886 ROH in comparison to other populations. (B) Histograms of the frequencies of long ROH in the TR,

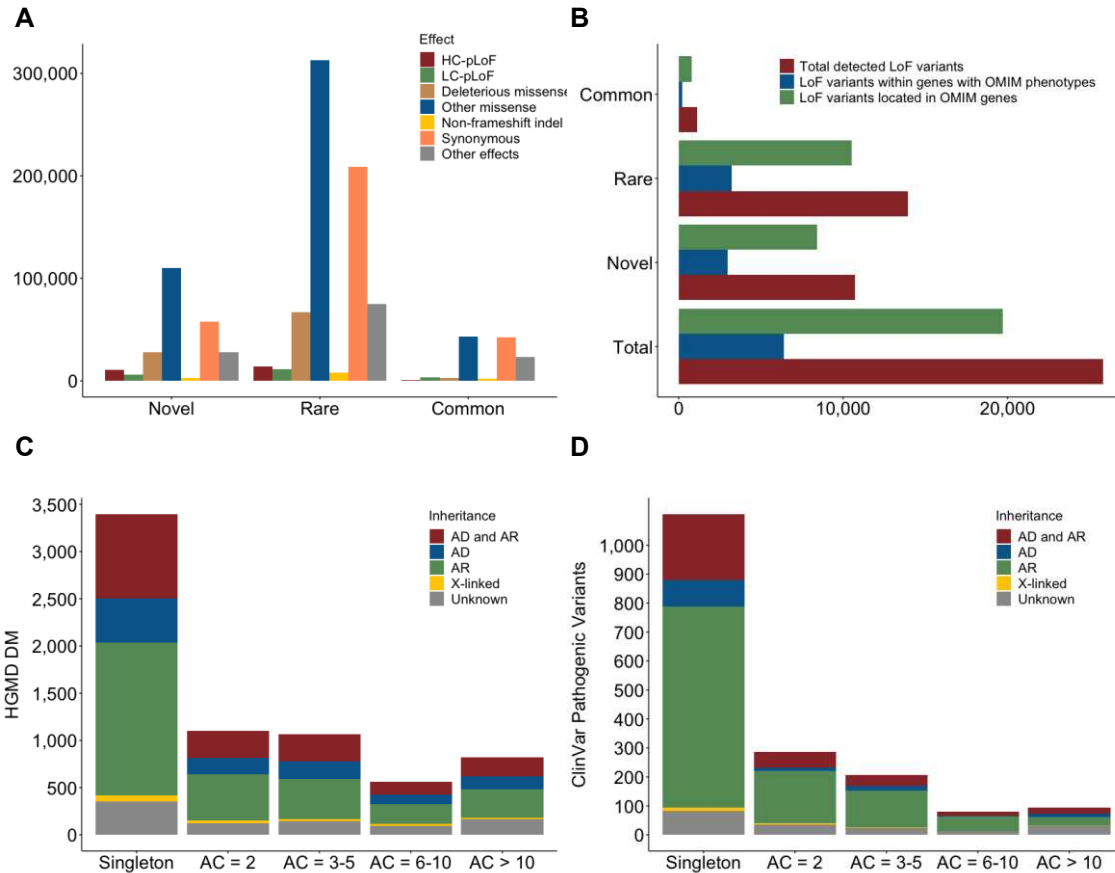
887 African, European, South Asian and East Asian populations. Frequencies were calculated by dividing the

888 number of ROH by the population size. ROH > 4 Mb in length are binned together (an asterisk indicates a

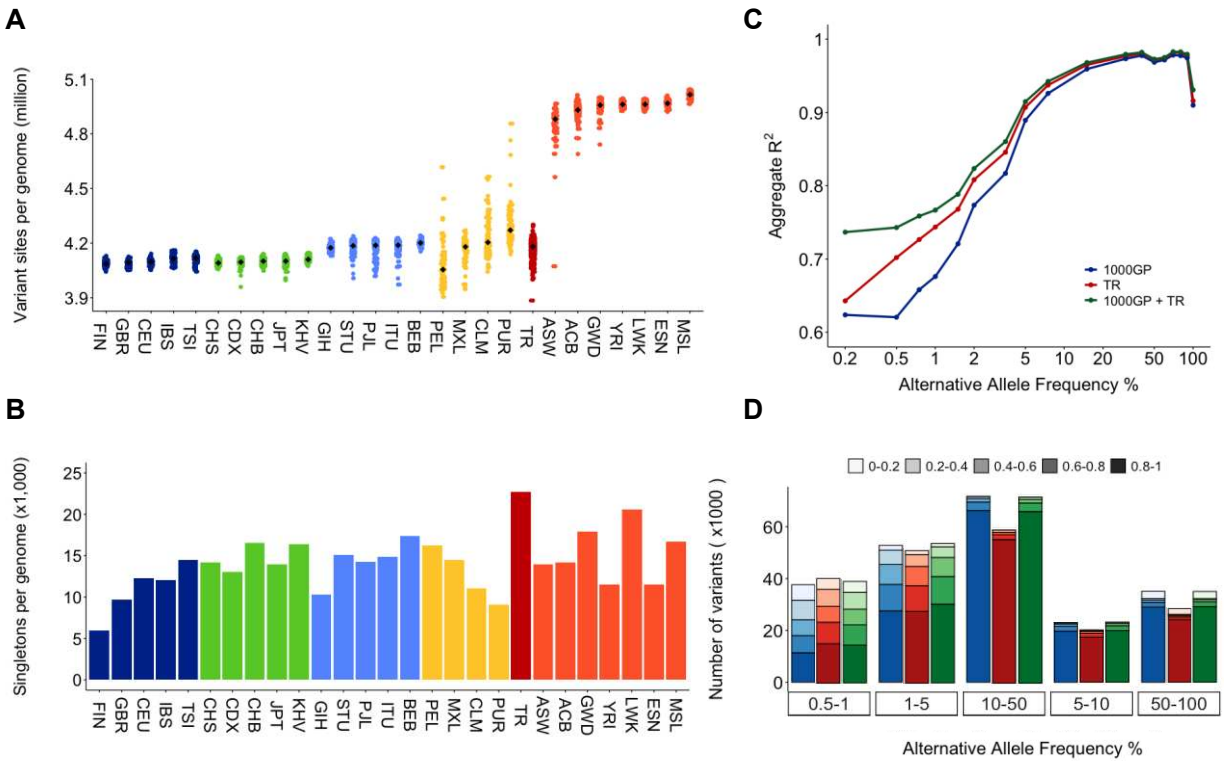
889 small peak seen in the TR population).



890
 891 **Figure 4. The TR Variome possesses a significant number of very rare unique variants that are**
 892 **poorly represented in gnomAD and GME.** The proportion of TR variants represented in the TR Variome
 893 and other databases. (A) TR WES vs. gnomAD WES, (B) TR WGS vs. gnomAD WGS, (C) TR WES vs.
 894 GME WES. The correlation of DAFs of rare TR variants in the (D) TR WES vs. gnomAD WES, (E) TR
 895 WGS vs. gnomAD WGS, (F) TR WES vs. GME WES. Hexagonal bins are shaded by the log-transformed
 896 number of variants in each bin. Although Pearson's r values point to a strong correlation between the TR -
 897 gnomAD and TR - GME DAFs, a remarkable number of TR DAFs could not be accurately estimated using
 898 gnomAD or GME Variome because of a wide range in allele frequencies between the datasets.



899
 900 **Figure 5. Distribution of variants based on their frequency, pathogenicity and OMIM phenotypes.**
 901 (A) Variants were classified as HC-pLoF, LC-pLoF, missense, non-frameshift indel, synonymous, and
 902 other effects as well as according to their frequency. Non-coding variants are not included in the Figure.
 903 (B) Proportions of HC-pLoF variants, which were grouped based on their frequency, their location on
 904 OMIM genes and the genes with OMIM phenotypes. (C) Distribution of disease-causing pathological
 905 mutations from HGMD and (D) pathogenic or pathogenic/likely pathogenic variants from the ClinVar
 906 database, categorized based on their frequency and inheritance type as autosomal-dominant (AD) or
 907 autosomal-recessive (AR), X-linked or unknown.



908

909 **Figure 6. Per genome variant summary and imputation.** (A) The number of variant sites per genome

910 for autosomes. (B) The average number of singletons per genome for autosomes. (C) Evaluation of

911 imputation performance on chromosome 20. The aggregate squared Pearson correlation coefficient (R^2)

912 was calculated for genotypes called from WGS and imputed genotypes and plotted against alternative

913 allele frequency for the three reference haplotype panels. A two-tailed Wilcoxon rank-sum test was used

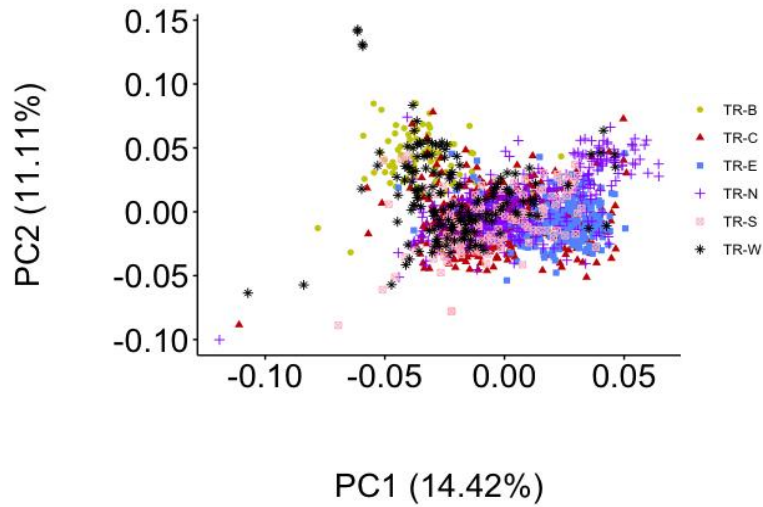
914 to assess the significance of the R^2 difference: $P = 0.002$ for the TR (mean \pm s.d.: 0.88 ± 0.12) versus

915 1000GP (mean \pm s.d.: 0.86 ± 0.14); $P = 0.002$ for the TR + 1000GP (mean \pm s.d.: 0.89 ± 0.09) versus

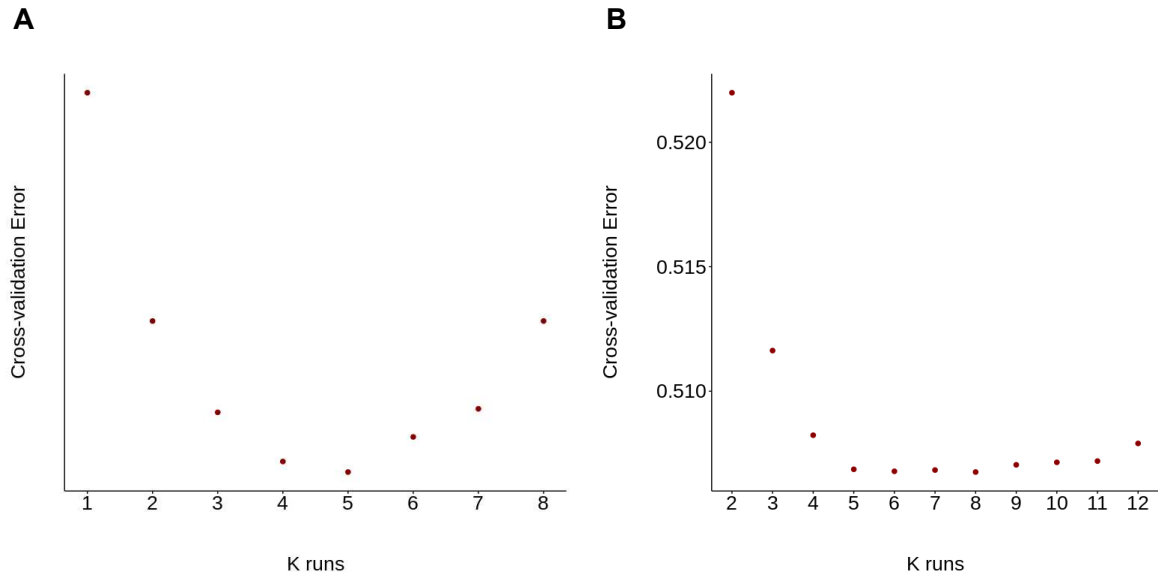
916 1000GP. (D) The number of imputed variants as a function of expected alternative allele frequency. The

917 density of shading reflects the expected imputation R^2 , whereas colors represent the reference panels

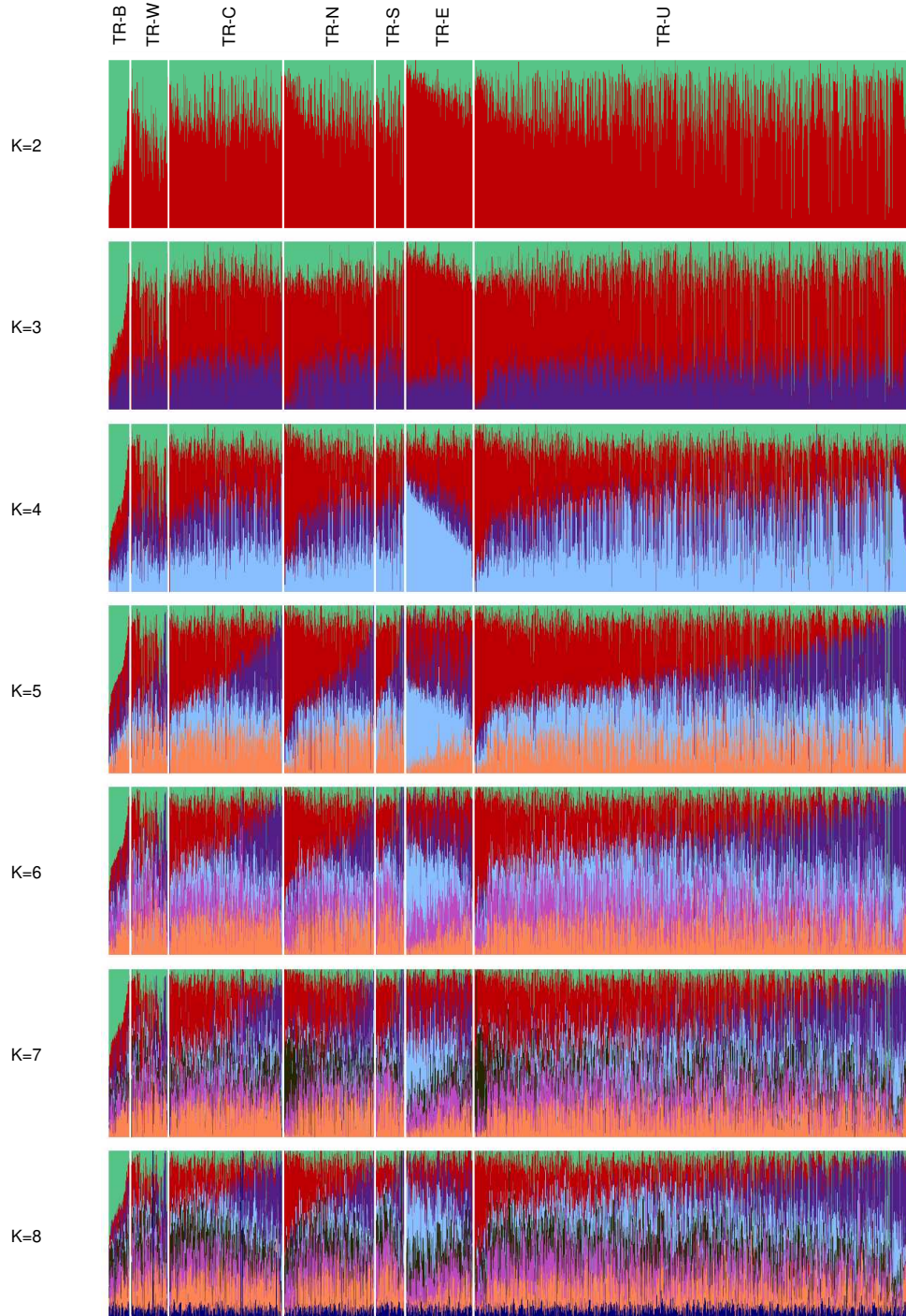
918 used for the imputation: 1000GP (blue), TR (red), or combined 1000GP and TR (green).



919
920 **Figure S1. Principal component analysis on TR individuals with known origin.** Plots for the PC1 and
921 PC2, which explain 14.42% and 11.11% of the total variation seen in Turkey. PC1 distinguishes TR-W,
922 TR-B and TR-E subregions.
923



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.



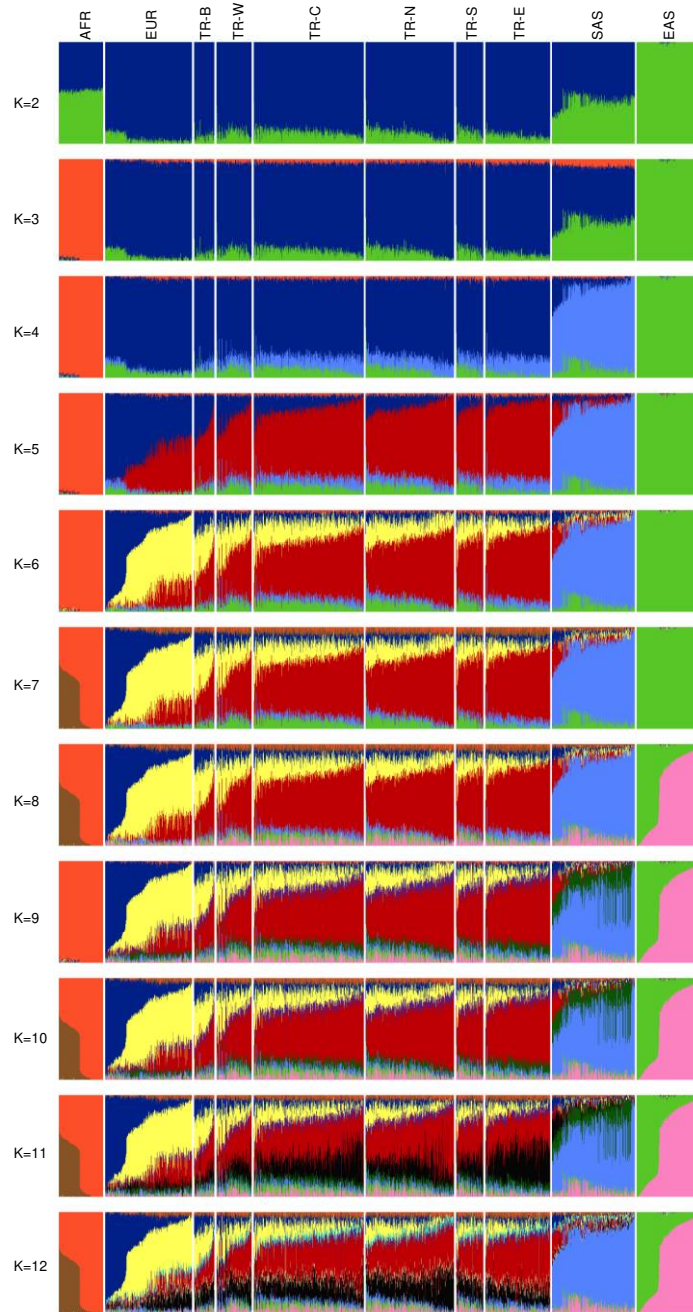
929

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

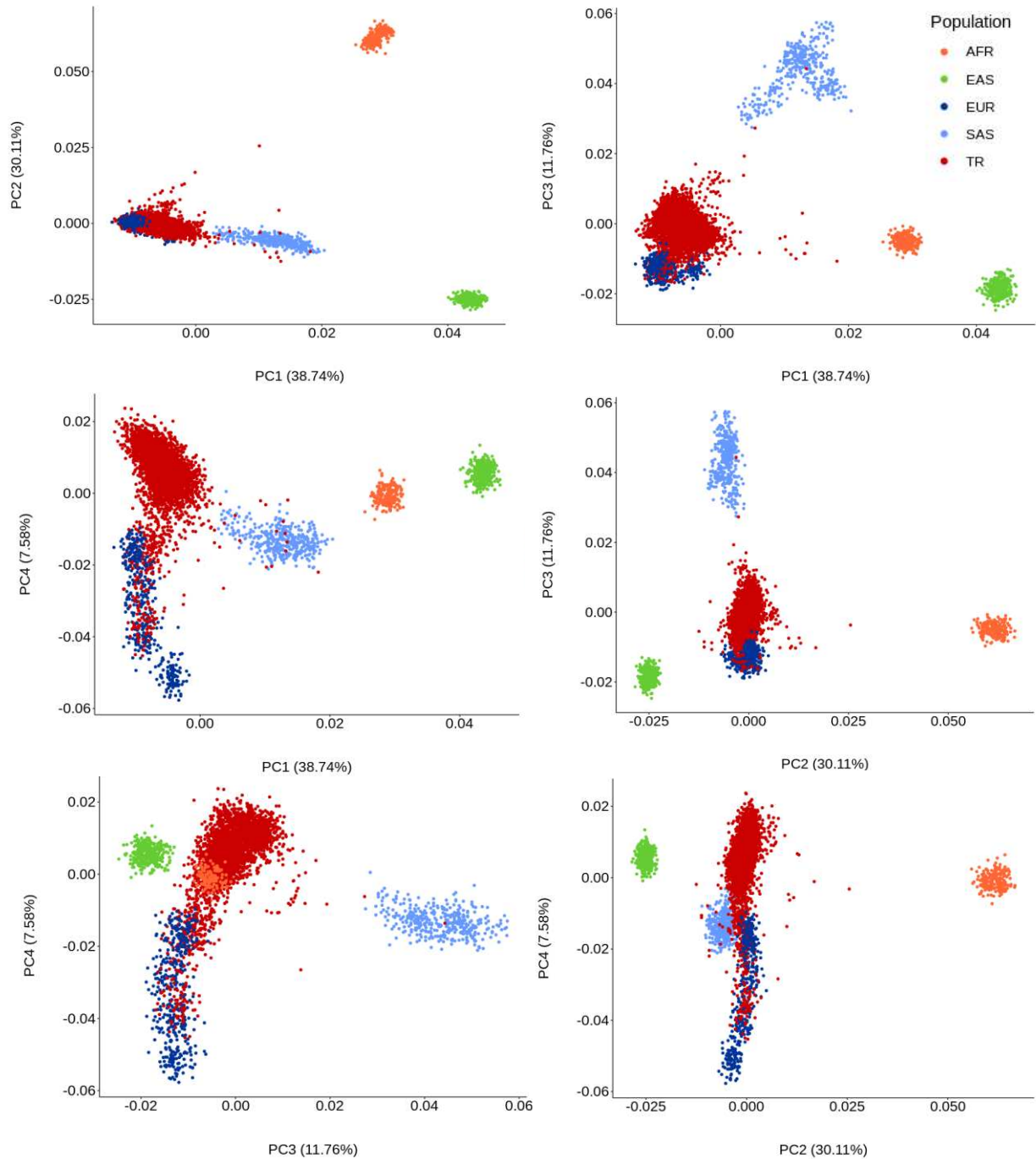
931 Samples grouped by geographical region and organized from west (left) to east (right). Each column

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

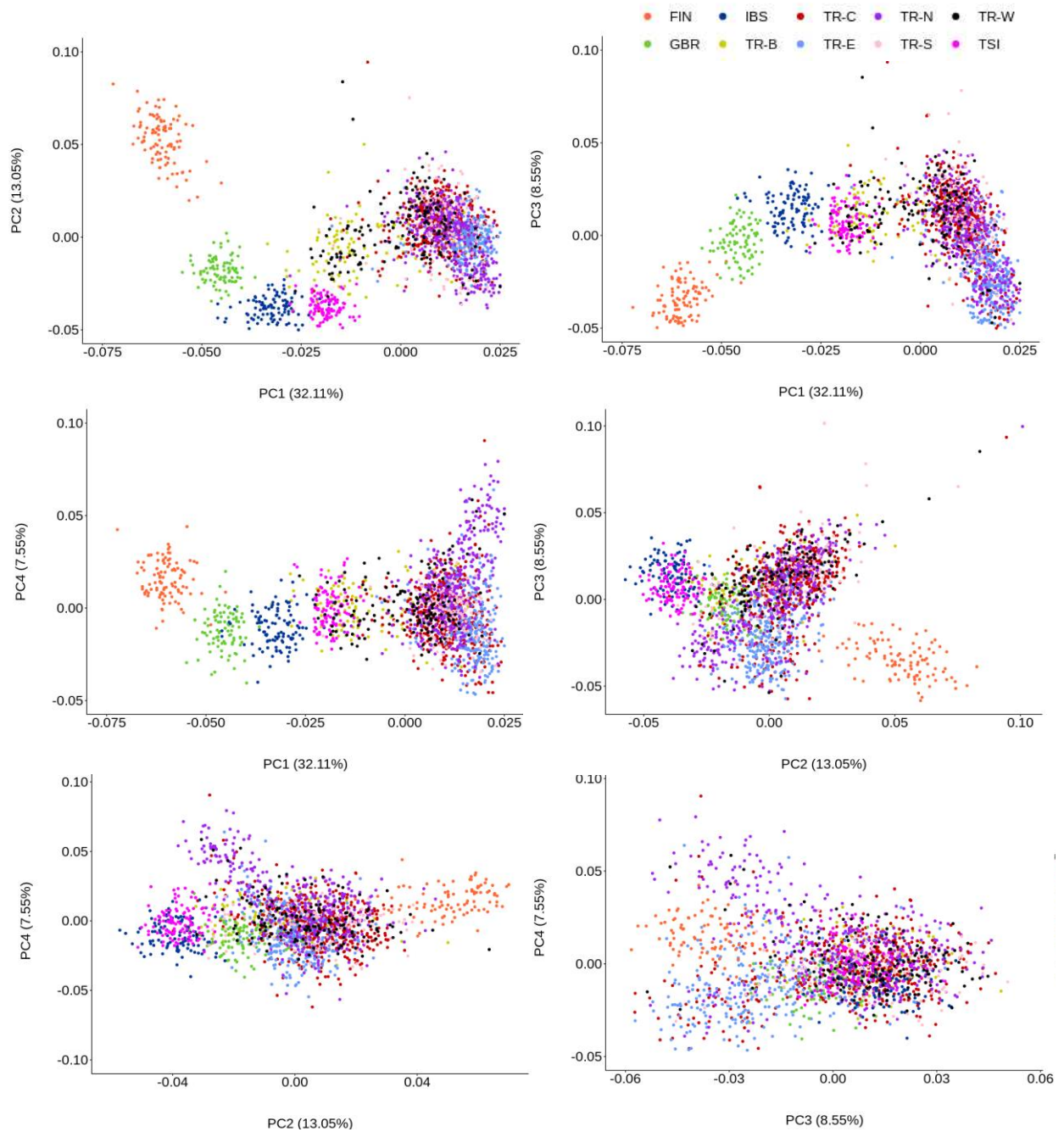
933 multiple ancestral components.



934
 935 **Figure S4. Unsupervised ADMIXTURE analysis of the TR population in a global context for clusters**
 936 **$k = 2$ to $k = 12$.** Samples from Turkey and 1000GP populations grouped by geographical region and
 937 organized from west (left) to east (right), showing trends of overlap. All subregions of the TR Peninsula
 938 appear to be represented by multiple ancestral components, including a single ancestry unique to Turkey,
 939 two ancestries predominant in European populations, as well as small proportions of Asian and African
 940 components.



941
 942 **Figure S5. Principal-component analysis on the TR and 1000GP populations.** Plots for the first four
 943 principal components and percentages of variance explained. PC1 (38.74%) and PC2 (30.11%) separate
 944 Africans and East Asians respectively from the other populations, while PC3 separates SAS from the TR
 945 population. PC4 demonstrated the degree of variance between the TR and European Populations,
 946 although it failed to generate distinct clusters.



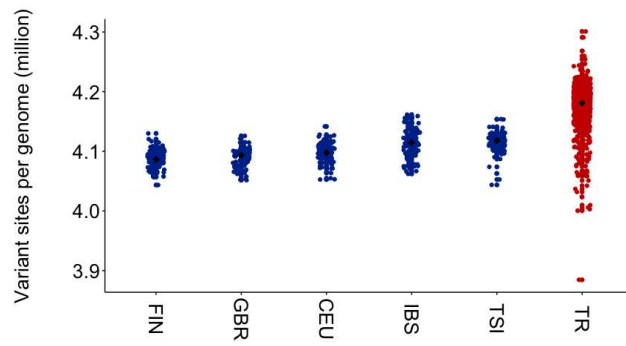
947

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

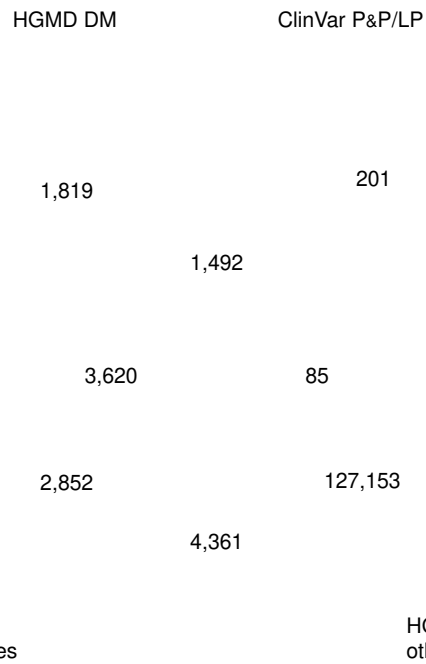
950 explained. PC1 (32.11%) separated the TR and European populations, although there was a significant

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



952

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



956

957

958

959

960

961

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.

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

Cohort	<i>n</i>	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	

963

964 **Table S2. Sample based-quality control measures for the integration of the coding**
 965 **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

966

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

