# CHARACTERIZATION OF THE FINE-SCALE GENETIC STRUCTURE OF THE TURKISH POPULATION

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MOLECULAR BIOLOGY AND GENETICS

By
Meltem Ece Kars
January 2022

Characterization of the fine-scale genetic structure of the Turkish Population
By Meltem Ece Kars
January 2022

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Tayfun Özçelik(Advisor)

Ayşe Nazlı Başak

Özlen Konu Karakayalı

Ali Osmay Güre

Serkan Belkaya

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

# ABSTRACT

## CHARACTERIZATION OF THE FINE-SCALE GENETIC STRUCTURE OF THE TURKISH POPULATION

Meltem Ece Kars
Ph.D. in Molecular Biology and Genetics
Advisor: Tayfun Özçelik
January 2022

The construction of population-based genetic resources plays a pivotal role in the study of human biology and disease. In this study, the fine-scale genetic structure of the Turkish (TR) population was characterized using the whole-exome (WES, $n$ = 2,589) and whole-genome (WGS, $n$ = 773) sequences of 3,362 unrelated individuals from Turkey. Significant levels of admixture from Balkan, Caucasus, Middle East, and Europe were detected in the TR subregions, consistent with the history of Anatolia. Results of the population structure analyses showed that the TR and European populations have a closer genetic relationship than previously appreciated. Inbreeding coefficient calculations and runs of homozygosity analysis reflected the unique effects of the high rate of consanguineous marriage on the TR genome. A TR Variome comprising over 40 million variants was constructed using the data generated in this study. Derived allele frequency (DAF) calculations revealed that 28% of TR-WES and 49% of TR-WGS variants in the very rare frequency bins (DAF $<$ 0.005) were not listed in the Genome Aggregation Database. The lists of clinically-relevant variants and human gene knockouts in the TR Variome were also listed in this study, presenting the potential of the TR Variome being an invaluable resource for future disease gene identification studies. Additionally, a reference panel for genotype imputation was generated using TR-WGS data. Since this panel significantly increased imputation accuracy in both TR and neighboring populations, it will probably facilitate genome-wide association studies in these populations. In the second part of the study, the sequencing data of a total of 3,599 unrelated TR individuals were assessed for previously reported pathogenic (RP) variants and predicted pathogenic (PP) variants in Online Inheritance in Men (OMIM) genes associated with a phenotype. Analyses revealed that no less than 70% of TR people have at least 1 RP variant, and all individuals possess at least one RP and/or PP variant in their genome. Moreover, 25% of individuals carried at least one RP variant in the newborn screening genes. Each individual in the study also had at least a 1 in 17 chance of carrying an RP variant in one of the 73 American College of Medical

Genetics recommended actionable genes. *MEFV*, *ABCA4*, *CYP21A2*, *PAH*, and *CFTR* displayed the highest cumulative carrier frequencies (CF), consistent with the high prevalence of the phenotypes they are responsible for. By estimating the CF and genetic prevalence in 3,251 OMIM genes using RP and PP variants, this study presents the most comprehensive data so far demonstrating the landscape of genetic disease in the TR population.



*Keywords:* Admixture, carrier frequency, population genetics, variation, whole exome sequencing, whole genome sequencing.

# ÖZET

# TÜRK TOPLUMUNUN İNCE ÖLCEKLİ GENETİK YAPISININ KARAKTERİZASYONU

Meltem Ece Kars
Moleküler Biyoloji ve Genetik, Doktora
Tez Danışmanı: Tayfun Özçelik
Ocak 2022

Toplum bazlı genetik kaynaklarının oluşturulması, insan biyolojisi ve hastalığının araştırılmasında çok önemli bir rol oynar. Bu çalışmada, Türkiye'den akraba olmayan 3.362 bireyin tam ekzom (TED, $n = 2.589$) ve tam genom (TGD, $n = 773$) dizileri kullanılarak Türk (TR) popülasyonunun ince ölçekli genetik yapısı karakterize edildi. Anadolu'nun tarihi ile uyumlu olarak, TR alt bölgelerinde Balkan, Kafkaslar, Orta Doğu ve Avrupa'dan önemli seviyelerde genetik karışım tespit edildi.Toplum yapısı analizlerinin sonuçları, TR ve Avrupa toplumlarının önceden tahmin edilenden daha yakın bir genetik ilişkiye sahip olduğunu gösterdi. Akraba çiftleşme katsayısı hesaplamaları ve homozigotluk serilerinin analizi, yüksek akraba evliliği oranının TR genomu üzerindeki benzersiz etkilerini yansıttı. Bu çalışmada üretilen veriler kullanılarak 40 milyondan fazla varyant içeren bir TR Variome oluşturuldu. Türev alel frekansı (TAF) hesaplamaları, çok nadir frekans aralığındaki (TAF <0.005) TR-TED varyantlarının %28'inin ve TR-TGD varyantlarının %49'unun Genome Aggregation Database'de listelenmediğini ortaya koydu. TR Variome'daki klinik etkisi olan varyantların ve insan gen nakavtlarının listeleri de bu çalışmada sunuldu ve bu listeler, TR Variome'un gelecekte gerçekleştirilecek hastalık geni tanımlama çalışmaları için paha biçilmez bir kaynak olma potansiyelini ortaya koydu.Ayrıca, TR-TGD verileri kullanılarak genotip atama için bir referans paneli oluşturuldu. Bu panel hem TR hem de komşu toplumlarda atama doğruluğunu önemli ölçüde arttırdığından, muhtemelen bu toplumlarda genom çapında ilişkilendirme çalışmalarını kolaylaştıracaktır. Çalışmanın ikinci bölümünde, bir fenotiple ilişkili Online Inheritance in Men (OMIM) genlerinde yer alan, daha önce bildirilen patojenik (BP) varyantlar ve öngörülen patojenik (ÖP) varyantlar için toplam 3.599 akraba olmayan TR bireyinin dizileme verileri değerlendirildi. Analizler, TR halkının en az %70'inin en az bir BP varyantına sahip olduğunu ve tüm bireylerin genomlarında en az bir BP ve/veya ÖP varyant taşıdığını ortaya koydu. Ayrıca bireylerin %25'i yenidoğan tarama genlerinde en az 1 BP varyantı taşıyordu. Ek olarak, çalışmadaki her bireyin, American College of Medical Genetics tarafından önerilen, eyleme geçilebilir 73 genden birinde bir BP varyant taşıma şansı en az

17'de 1'di. *ABCA4*, *CYP21A2*, *PAH* ve *CFTR*, sorumlu oldukları fenotiplerin yüksek prevalansı ile tutarlı olarak en yüksek kümülatif taşıyıcı frekanslarını (TF) sergiledi. Bu çalışmada, BP ve ÖP varyantlar kullanılarak hesaplanan 3.251 OMIM genindeki TF ve genetik prevalans ile TR toplumundaki genetik hastalıkları tasvir eden şimdiye kadarki en kapsamlı veriler sunuldu.

*Anahtar sözcükler*: Genetik karışım, taşıyıcı frekansı, toplum genetiği, tüm ekzom dizileme, tüm genom dizileme, varyasyon.

# Acknowledgement

This thesis is the pleasing outcome of several fortunate and compelling experiences. I would like to express my eternal gratitude to my Ph.D. supervisor, Prof. Tayfun Özçelik, who guided me through reshaping my career path, conceived the idea of this thesis, encouraged and supported me during my doctoral studies. This study would not come true without him.

I feel blessed and I am grateful for having Prof. Ayşe Nazlı Başak and Assoc. Prof. Özlen Konu Karakayalı in my thesis monitoring committee. Prof. Ayşe Nazlı Başak's invaluable guidance and data took this study to a whole other level. Assoc. Prof. Özlen Konu Karakayalı's precious suggestions and feedback helped me remarkably during analyses. I am also thankful for the Prof. Ali Osmay Güre and Asst. Prof. Serkan Belkaya for being on my dissertation committee and all professors in the Department of Molecular Biology and Genetics for sharing their expertise. I would like to also thank the collaborators of the study for sharing valuable knowledge and data.

I am thankful to the former senior researcher of the group, Dr. Onur Emre Onat, who introduced me to bioinformatics. I also thank him and Dr. Umut Inci Onat for their friendship.

I would also like to express my sincere gratitude to İclal Özçelik, whose kind support had a significant impact on me and my research.

I also extend my thanks to Nezahat Doğan and Demet Güzelsoy for their help in the study and emotional support.

I gratefully acknowledge that I am the recipient of the fellowship of 2211/A National Doctorate Scholarship Program of Turkey of Directorate of Science Fellowships and Grant Programmes (BİDEB) of Scientific and Technological Research Council (TÜBİTAK).

I thank my family for bringing up me and their invaluable support. Lastly, I want to express my heartfelt gratitude to my husband, my biggest supporter, Serhan Kars, for bearing with me during not just one, but two theses.

Meltem Ece Kars

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| 1000GP | 1000 Genomes Project |
| ACB | African Caribbeans in Barbados |
| ACHDNC | Advisory Committee on Heritable Disorders in Newborns and Children |
| ACMG | American College of Medical Genetics |
| AD | Autosomal dominant |
| ADHD | Attention deficit hyperactivity disorder |
| AF | Allele frequency |
| AFR | African |
| AMP | Association for Molecular Pathology |
| AMR | Admixed American |
| AP | Arabian Peninsula |
| AR | Autosomal recessive |
| ASW | Americans of African Ancestry in SW USA |
| BEB | Bengali from Bangladesh |
| BED | Browser Extensible Data |
| BLK | Balkan |
| BCL | Binary Base Call |
| bp | Base pairs |
| BWA | Burrows-Wheeler Aligner |
| CADD | Combined Annotation Dependent Depletion |
| CAH | Congenital adrenal hyperplasia |
| CAU | Caucasus |
| CCDS | Consensus coding sequence |
| CDX | Chinese Dai in Xishuangbanna |
| CEU | Utah Residents with Northern and Western European Ancestry |
| CF | Carrier frequency |
| CHB | Han Chinese in Beijing, China |
| CHS | Southern Han Chinese |
| CLM | Colombians from Medellin, Colombia |
| CNA | Central and North Asian |
| CNV | Copy number variation |
| DAF | Derived allele frequency |

| | |
|---|---|
| dbSNP | Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation |
| dbVar | Database of Genomic Structural Variation |
| DM | Disease-causing mutations |
| DM? | Probable/possible pathogenic mutations |
| DSPD | Delayed sleep phase disorder |
| EUR | European |
| EAS | East Asian |
| ESN | Esan in Nigeria |
| ESP | NHLBI GO Exome Sequencing Project |
| ExAC | Exome Aggregation Consortium |
| FIN | Finnish in Finland |
| FMF | Familial Mediterranean fever |
| $F_{\text{plink}}$ | Inbreeding coefficient |
| $F_{\text{ROH}}$ | The autosomal genome in runs of homozygosity |
| $F_{\text{ST}}$ | Wright's fixation index |
| GATK | Genome Analysis Tool Kit |
| GBR | British in England and Scotland |
| GIH | Gujarati Indian from Houston,Texas, USA |
| GME | Greater Middle East |
| gnomAD | The Genome Aggregation Database |
| GP | Genetic prevalence |
| GRCh | Genome Reference Consortium human build |
| GVCF | Genomic variant call format |
| GWAS | Genome-wide association studies |
| GWD | Gambian in Western Divisions in the Gambia |
| HC-pLoF | High-confidence pLoF |
| HGDP | Human Genome Diversity Project |
| HGMD | Human Gene Mutation Database |
| HGP | The Human Genome Project |
| H.O. array | Affymetrix Human Origins array |
| HWE | Hardy-Weinberg equilibrium |
| IBS | Iberian population in Spain |
| IQR | Interquartile range |
| Indel | Insertion/deletion variant |
| ISOGG | International Society of Genetic Genealogy |

| | |
|---|---|
| ITU | Indian Telugu from the UK |
| JPT | Japanese in Tokyo, Japan |
| Kb | Kilobase |
| KHV | Kinh in Ho Chi Minh City, Vietnam |
| LC-pLoF | Low-confidence pLoF |
| LD | Linkage disequilibrium |
| LoF | Loss of function |
| LWK | Luhya in Webuye, Kenya |
| MAF | Minor allele frequency |
| Mb | Megabase |
| MSL | Mende in Sierra Leone |
| mtDNA | Mitochondrial DNA |
| MXL | MXL Mexican Ancestry from Los Angeles, CA, USA |
| NBS | Newborn screening |
| NCBI | National Center for Biotechnology Information |
| NEA | North East African |
| NWA | North West African |
| NGS | Next-generation sequencing |
| NHLBI | The National Heart, Lung, and Blood Institute |
| NROH | Number of ROHs |
| OMIM | Online Inheritance in Man |
| P | Pathogenic |
| P/LP | Pathogenic/Likely pathogenic |
| PC | Principal component |
| PCA | Principal components analysis |
| PEL | Peruvians from Lima, Peru |
| Pext | proportion expression across transcripts |
| PJL | Punjabi from Lahore, Pakistan |
| PKU | Phenylketonuria |
| pLoF | Predicted loss-of-function |
| PP | Predicted pathogenic |
| PPV | Positive predictive value |
| PROMIS | The Pakistan Risk of Myocardial Infarction Study |
| PUR | Puerto Ricans from Puerto Rico |
| QC | Quality control |
| rCRS | Revised Cambridge Reference Sequence |

| | |
|---|---|
| ROH | Runs of homozygosity |
| RP | Reported pathogenic |
| SAS | South Asian |
| SD | Syrian Desert |
| SNP | Single nucleotode polymorphism |
| SNV | Single nucleotide variation |
| SO | Sequence ontology |
| SROH | Sum of ROHs |
| STU | Sri Lankan Tamil from the UK |
| TFBS | Transcription Factor Binding Site |
| TSI | Toscani in Italia |
| tSNE | t-distributed stochastic neighbor embedding |
| TR | Turkish |
| TR-B | Turkish with Balkan Ancestry |
| TR-C | Central Turkish |
| TR-E | Eastern Turkish |
| TR-N | Northern Turkish |
| TR-S | Southern Turkish |
| TR-U | Turkish with unknown origin |
| TR-W | Western Turkish |
| UMAP | Uniform manifold approximation and projection |
| UTR | Untranslated region |
| VCF | Variant call format |
| VQSR | Variant quality score recalibration |
| WES | Whole exome sequencing |
| WGS | Whole genome sequencing |
| YRI | Yoruba in Ibadan, Nigeria |

# Chapter 1

# Introduction

## 1.1 The sequence of the human genome

Investigating the relationship between genotype and phenotype has been a fundamental method for understanding human biology and disease. The establishment of the human reference genome was one of the milestones in the study of genetics, accelerating the pace of discoveries in different areas of biology and medicine. The Human Genome Project (HGP) launched in 1990 with the aims to map and sequence the human genome, to produce a genetic map and sequence DNA of model organisms, to develop software, algorithms, and database tools for the interpretation of genomic information, to identify ethical, legal, and social implications, and to support research training, technology development, and transfer [1]. The first phase of the HGP consisted of DNA cloning and shotgun sequencing and subsequently assembling sequence data from multiple clones by determining overlapping regions and establishing a contiguous sequence. However, the first drafts of the human genome (NCBI Build 34) that were released by both the International Human Genome Sequencing Consortium and Celera Genomics in 2001 contained a vast amount of gaps [2, 3]. The "finishing phase" of the HGP, by reducing the gaps to 341 and covering 99% of the euchromatic genome, generated the "near-complete sequence" of the human genome in 2004 (National Center for Biotechnology Information, NCBI Build 35) [4]. The human genome has become "more complete" in subsequent assemblies, namely NCBI Build 36.1

in 2006, Genome Reference Consortium human build 37, (GRCh37) in 2009, and GRCh38 in 2013 with the developments in high throughput sequencing technology and bioinformatics tools. Today, we know that the length of the human genome is over 3 billion base pairs and is estimated to contain 30,000 genes.

The accurate identification of the genes, and indexing their genomic coordinates in the human genome is essential for genetic analysis. The well-known resources for the reference assembly and annotation of genomic regions of the human genome (location of genes, exons, introns, regulatory and intergenic regions) are the Reference Sequence (RefSeq) [5] and Ensembl [6]. Refseq and Ensembl are the collections of sequences of genome, transcripts, and proteins for humans and several other organisms. They also provide annotations for coding regions, conserved domains, intronic and intergenic regions, and variation. Although they utilize the same reference assemblies, annotations provided by the two databases slightly differ because of the differences in their pipeline.

## 1.2 Human genetic variation

The human reference genome is an indispensable resource for the study of human genetics and undoubtedly expedited biomedical research. Yet, the extent of genetic variation in the human populations was understood with ensuing efforts [7]. Sequencing of genomes from diverse populations has shown that any individual genome, on average, differs from the reference human genome at 4 to 5 million sites, which correspond to about 20 million bases of sequence [8, 9].

### 1.2.1 Types of genetic variants

A genetic variant is the result of a mutational process, which leads to single or multiple nucleotide changes. Genetic variants can be divided into two main categories as a single nucleotide change (single nucleotide variation, SNV) or a structural variant, namely insertion/deletion variant (indel), block substitution, inversion, or copy number variation (CNV)(Figure 1.1). Indels are defined as the insertion or deletion of one or more base pairs (bp). Most indels are short

(composed of only a few bases) and usually less than 1 kilobase (Kb) in length. Block substitutions occur when a string of adjacent nucleotides changed with the same number of different nucleotides. An inversion variant describes the event in which the orientation of the bp is reversed. CNVs occur when the number of repeating sequences varies between individuals. Over 99.9% of human genetic variants are SNVs or short indels; however, the effect of structural variants becomes much bigger when the number of affected bases is considered [8].

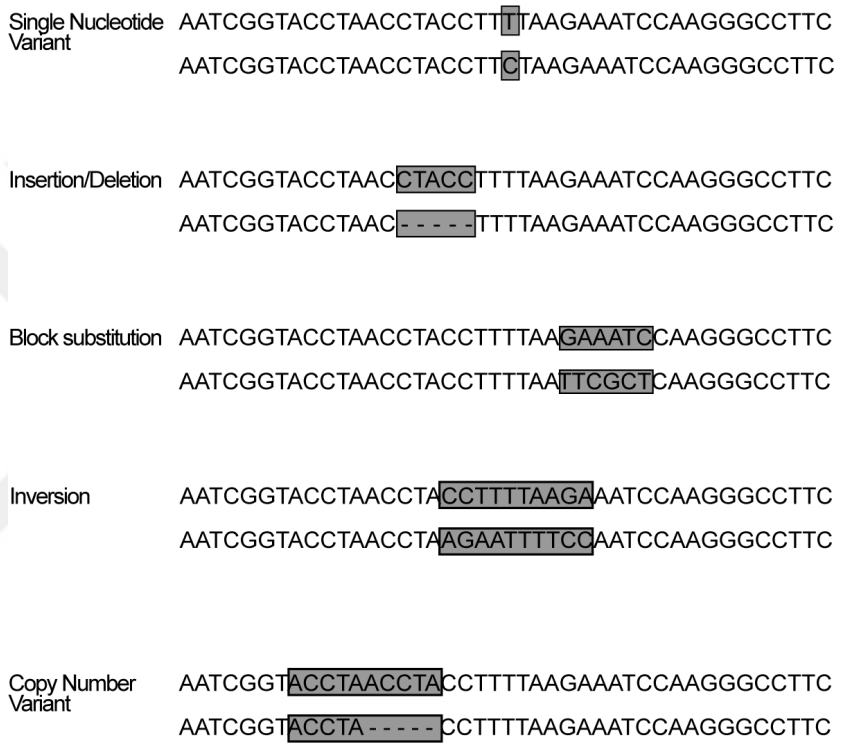



**Figure 1.1: Examples for genetic variant types.** The sequence above represents the wild type genotype, sequence below represents the mutational process.

### 1.2.2 Functional impact of genetic variants

SNVs in the protein-coding region of the genome are designated based on their effects on the amino acid sequence. If an SNV does not result in an alteration in the amino acid, it is called a synonymous mutation. When an SNV causes a change in the amino acid, it is named a missense mutation. Lastly, a nonsense (or stop gain) variant occurs when an SNV converts the codon of an amino acid into a stop codon. Indels can also be classified into two groups as frameshift

and non-frameshift. When the length of an indel is a multiple of three, it is named as a non-frameshift indel. Otherwise, it can cause a shift in the reading frame and is called a frameshift indel. A variant located in the splice region of the genome is named a splice-site variant. Splice-site variants can disrupt RNA splicing resulting in the exon skipping or the inclusion of introns and subsequently, an altered protein-coding sequence [10].

Most of the genetic variants are tolerated at the organism level, while some alter the function of a protein or RNA. The functional impacts can be the result of a change in [10]:

1. Enzymatic activity due to changes in catalytic sites or binding pockets of a protein
2. Enzymatic activity via affecting stability or mobility of a protein
3. Protein-protein interactions
4. The amount (expression level) of a gene product

The genetic variants can also be divided into two basic categories according to their effects on the organism: loss-of-function (LoF) and gain-of-function variants. LoF variants are genetic variants that are predicted to decrease or diminish the function of a gene product. Gain-of-function variants result in altered gene products that lead to a new molecular function or a new pattern of expression. Although there are exceptions, recessive mutations inactivate the affected gene and cause an LoF whereas dominant mutations lead to a gain of function [11].

### 1.2.3 Population frequency of genetic variants

Genetic variants are also categorized based on their frequency in the population as common and rare. Although the definition of “common” and “rare” differ among different populations, the broadly accepted cut-off is 1% [7]. Thus, common variants are observed in the general population with a frequency of 1% or higher, while the frequencies of rare variants are lower than 1%. Polymorphism is a term to describe common genetic variations, although it is not frequently used

anymore. The effect of most variants is predicted to be neutral, which means they do not have large functional effects on proteins and phenotype [7]. When they arise in the population, their frequency is altered by chance or demographic factors. Beneficial variants can be defined as the variants that have a fitness advantage. The frequency of such variants rapidly increases in the population [12]. On the contrary, variants that disrupt the function of a protein and affect the phenotype in an unfavorable way, deleterious variants, cannot reach a high frequency. Although deleteriousness has a prominent effect on the frequency of a variant in the population, it is also affected by demographic factors that are described in detail in the following sections.

## 1.3 Genetic structure of populations

The advancements in sequencing technology paved the way for the studies of genetic variation within populations, population history, ancestral components, and genealogy. Using the genetic information that is encoded in genomes as patterns of genetic variation and recombination events, the historical events and forms of natural selection that shaped the genetic substructure of human populations can be evaluated. The importance of genetic ancestry in the medical perspective is that genetic variants with different frequencies in diverged populations have various clinically-relevant functions. For example, global populations have different genetic determinants associated with diabetes mellitus, iron deficiency, and drug response [13–15]. Moreover, the prevalence of certain disorders differs in different ethnic or geographical groups. For instance, Tay-Sachs Disease among Ashkenazi Jews and sickle cell disease in Africans are much more prevalent when compared to other populations [9]. Studying the genetic structure of populations can be useful to identify the genes and variants that are responsible for such disorders. Another example of the importance of evaluating the genetic structure of a population is that research in consanguineous populations facilitates disease gene discovery due to increased levels of homozygosity [16, 17].

### 1.3.1 Demographic origins of genetic variation within human populations

Certain population events might severely alter the frequency of variants and overall genetic diversity in a population [9]. For example, changes in variant frequencies by chance, genetic drift, occur due to bottleneck or founder effect. Population bottleneck is a significant reduction in the population size due to natural disasters, epidemics, or famine. Under the bottleneck effect, only a few individuals can transmit their genetic information to their offspring. Thus, the new population formed after a bottleneck has much different variant frequencies in addition to a reduced diversity compared to the former population (Figure 1.2 A). The founder effect is low genetic diversity in a population due to descending from a small number of colonizing ancestors. A population with the founder effect has a reduced genetic variation; also, the representation of genetic variants in the population is significantly altered compared to the former population (Figure 1.2 B). Admixture is defined as intermixing between previously diverged populations or gene flow from a previously diverged donor population to a recipient population. Admixture elevates the genetic diversity and heterogeneity. As a result, some genetic variants can have increased or decreased frequency in the new admixed population (Figure 1.2 C). Lastly, large populations tend to have a higher genetic diversity, whereas small populations have a lower diversity.

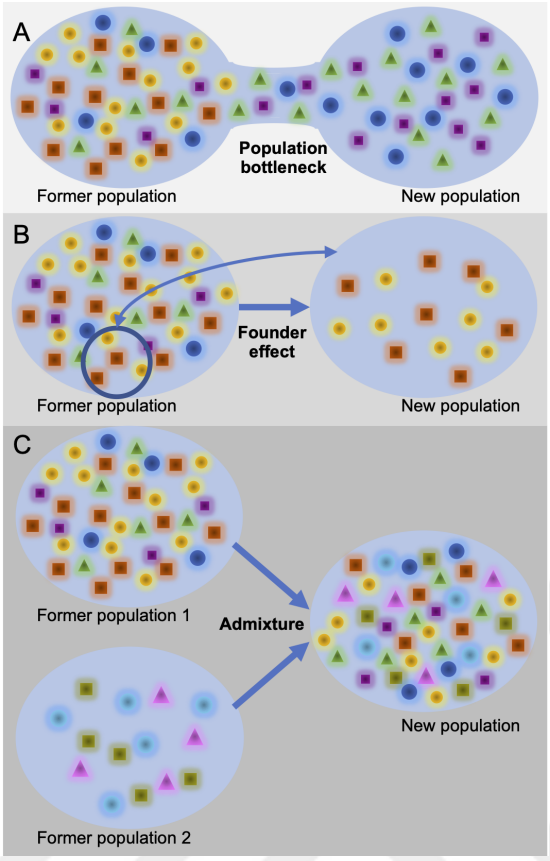



**Figure 1.2: The effect of population events on the genetic structure of the population.** Schematic drawing displaying the effects of **A** population bottleneck; **B** founder effect; **C** admixture on the genetic variation in the population. Shapes and colors represent genetic variations.

#### 1.3.1.1 The effect of consanguinity on genome

As indicated in the previous section, many ancient and historic population events, such as migration, isolation, bottleneck, having a large or small population size, have shaped the genetic landscape of modern-day human populations. Inbreeding is another factor that displays a prominent effect on the human genome by decreasing genetic diversity. Inbred populations might be formed due to isolation or a high rate of consanguineous matings. In general, inbreeding in human populations depends on geography and culture. In the geographical region that spans the Middle East and North Africa (or is defined as "Greater Middle East"), the practice of consanguineous marriage is highly frequent (Figure 1.3) [18]. Consanguinity results in having offspring with a higher risk for a recessive disorder because of the increased chance of having two copies of deleterious alleles. [16].

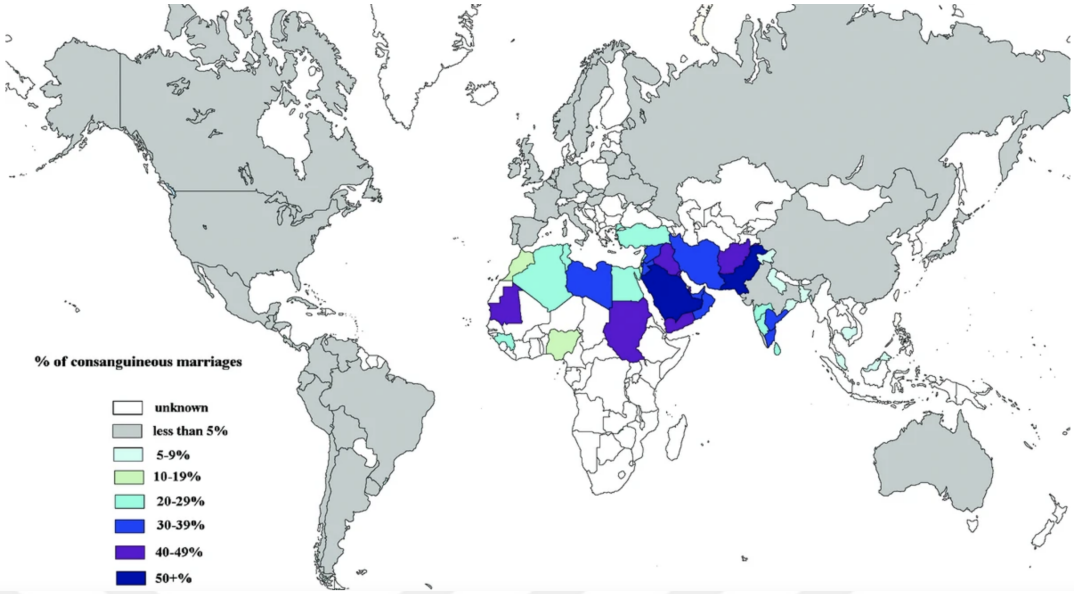



**Figure 1.3: Global rates of consanguinity.** Reprinted by permission from Springer Nature, [18], Copyright ©2011.

Inbreeding coefficient, simply the percentage of homozygous alleles in the genome, measures genetic relatedness and is expected to be high in a consanguineous population [19]. Also, the stretches of contiguous homozygous segments, runs of homozygosity (ROH), is another measure for inbreeding. The length and number of ROHs reflect the demographic events that previously took place in a population (Figure 1.4) [20]. Large populations tend to have a shorter and lower number of ROHs, whereas smaller populations have a larger and higher number of ROHs. Admixture results in shorter and very few ROHs. Consanguinity affects the length of ROHs by bringing longer ROHs. Bottlenecks, on the other hand, increase both the number and length of ROHs.

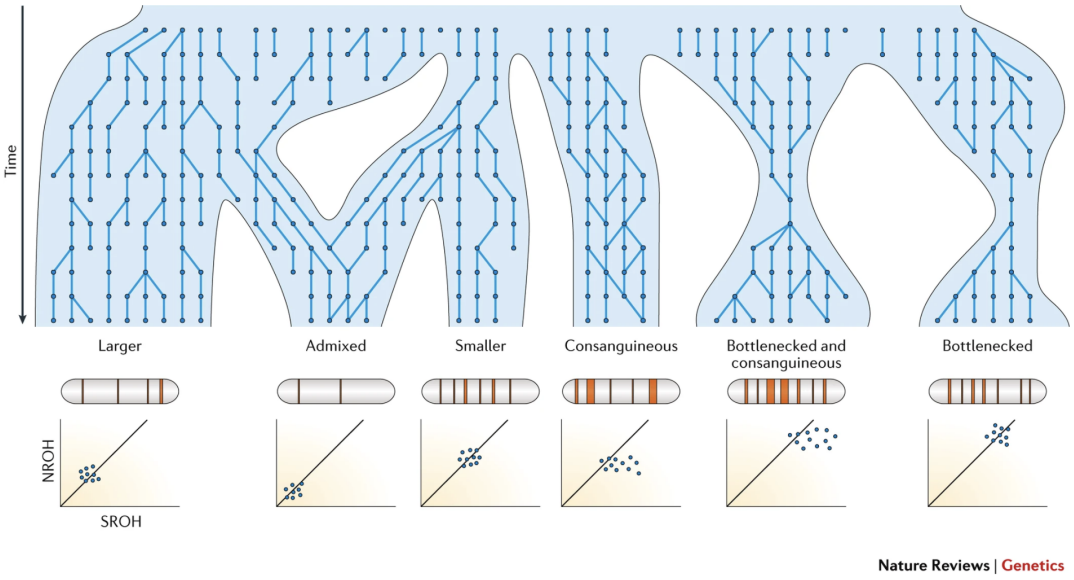



**Figure 1.4: Demographic origins of ROH.** The upper part (in blue) demonstrates the pedigree patterns demographic histories of populations. The lower part is the representation for the number of ROHs (NROH) versus sum total length of ROHs (SROH). Reprinted by permission from Springer Nature, [20],Copyright ©2018.

## 1.4 Resources of human genetic variants

The number of available variant resources is rapidly increasing. The most comprehensive NCBI databases that catalog human genetic variation are the Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation (dbSNP) and the Database of Genomic Structural Variation (dbVar) for short variants and structural variants, respectively [21, 22]. These are collections of variations and their annotations that are submitted from many resources.

One of the earliest studies aiming discovery of genetic variation in diverse populations is the International HapMap Project, which was launched in 2002 [23]. The goal of the International HapMap Project was to catalog common sequence variants in African (AFR), European (EUR), and East Asian (EAS) populations to be used in future genome-wide association studies (GWAS) of common disease. The project used tag SNPs that were genotyped using whole-genome, array-based genotyping. More than one million single nucleotide polymorphisms (SNP), haplotype diversities, and some insight into structural variation in 269 DNA samples

from four populations were presented in the results of the first phase in 2005 [24]. The second phase of the study provided 2.1 million additional SNPs and fine-scale structure of linkage disequilibrium of global populations [25]. The data of the International HapMap Consortium facilitated a remarkable amount of early GWAS.

Although HapMap provided invaluable insights for the common genetic variation, research in genetic disease required the investigation of variants in the lower frequency spectrum. Thus, the 1000 Genomes Project (1000GP) Consortium initiated the study of the common and rare genomic variation from diverse populations using high throughput sequencing technology. The aim of the pilot phase of 1000GP is to catalog over 95% of common variants, as well as variants with lower frequencies [26]. The results aid in many human genetic studies by enabling array design, genotype imputation for GWAS, and listing tolerated variants. The final phase of the 1000GP constituted 2,504 samples from 26 populations in Africa, East Asia, Europe, South Asia, and the Americas. Figure 1.5 demonstrates the population sampling of 1000GP. The high-quality variants of all study participants were obtained using whole-genome sequencing (WGS), targeted exome sequencing, and high-density SNP genotyping [27]. The data of the 1000GP has been a fundamental resource for many population genetics studies, GWAS, and studies of disease gene discovery.

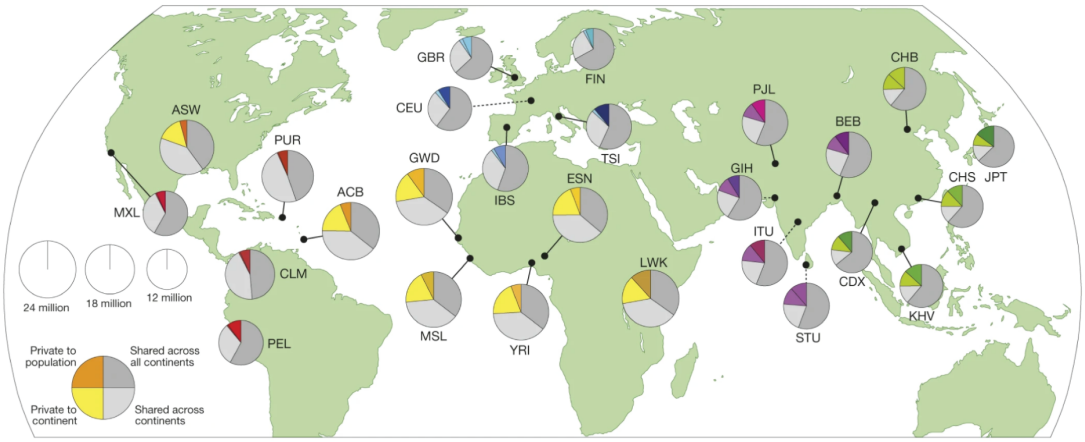



**Figure 1.5: The populations of the 1000GP.** Pie charts on the map demonstrate the number of polymorphic variants in populations (Publisher: Springer Nature, [27], licensed under CC BY-NC-SA 3.0, 2015).

There are many other resources that present genetic variations in diverse populations and can be utilized in studies of human genetics. NHLBI GO Exome Sequencing Project (ESP) was initially established for discovering novel genes and mechanisms contributing to heart, lung, and blood disorders and became an important resource for variant frequency data. ESP consists of whole-exome sequencing (WES) data of 6,503 individuals with EUR or AFR ancestry [28].

Exome Aggregation Consortium (ExAC) released the variant frequencies of 60,706 WES samples from EUR, AFR, South Asian (SAS), EAS, and admixed American (AMR) populations [29]. ExAC was the first release of the Genome Aggregation Database (gnomAD). The second release of gnomAD is composed of 125,748 WES and 15,708 WGS data from unrelated individuals who were initially sequenced as part of various disease-specific and population genetic studies [30]. The third release contains the data of 76,156 WGS. It has been the largest resource for the studies of human genetics.

Even though gnomAD provided a vast amount of genetic data, there have been many underrepresented populations in the available databases. Thus, population-specific studies and variant datasets from many diverse populations have begun to be released. One specific example is the Middle Eastern population. GME Variome was the first study that represents the genetic data from the geographical region [17]. GME contained samples from North East Africa (NEA), North West Africa (NWA), Arabian Peninsula (AP), Persia and Pakistan (PP), Syrian Desert (SD), and Turkish (TR) Peninsula (Figure 1.6. The characterization of the GME Variome has shown that studies conducted on these populations would enable disease gene identification, especially for recessive disorders. They also showed that the extensive admixture and high level of consanguinity affected the genetic diversity of Middle Eastern populations significantly.

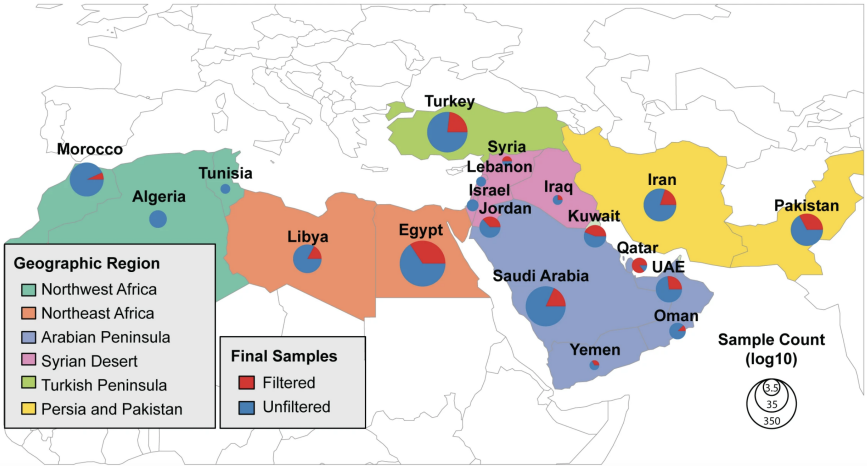



**Figure 1.6: The populations of the GME Variome.** Reprinted by permission from Springer Nature, [17], Copyright ©2016.

There have been other studies evaluating the genetic diversity in the Middle East. WES and WGS of the Qatari population introduced thousands of variants that were not represented in other variant databases [31]. Iranome employed 800 WES samples from 8 ethnic groups of Iran to demonstrate the genetic landscape of the country [32]. A recent study from the AP used137 WGS samples from 8 populations, namely Iraqi Arab, Iraqi Kurd, Jordanian, Emirati, Omani, Saudi, Syrian, Yemeni [33]. The study provided insights into population history and introduced millions of variants that were not cataloged previously.

GenomeAsia 100K study is composed of many underrepresented populations across the Asia continent. The pilot study consists of WGS data of 1,739 individuals from 219 Asian populations [34]. GenomeAsia 100K Consortium showed that future studies in these populations might facilitate rare and disease-associated variant and gene discovery.

Several other studies presented the genetic substructures of other underrepresented populations [35–40]. These efforts helped to characterize the population-specific variation and became a valuable resource for comprehensive genetic studies.

## 1.5 High throughput sequencing

The first drafts of the reference human genome were generated using shotgun sequencing, which is one of the precursors of WGS. In shotgun sequencing, DNA is fragmented in random short pieces and multiple overlapping reads are generated. DNA is assembled into contiguous sequences using multiple overlapping reads [2]. The method is based on the "chain-termination method" of Sanger sequencing, which only enables the sequencing of short DNA segments. Paired-end sequencing aided in determining overlapping regions. The need for faster and cheaper high-throughput sequencing of human DNA for routine applications led to the development of next-generation sequencing (NGS, or massively parallel sequencing) technology [41]. NGS produces millions of sequences simultaneously and facilitated the sequencing of coding regions of the human genome "whole exome" as well as entire genome "whole genome". Contrary to random shotgun sequencing, the DNA is amplified *in vitro* using clonal amplification by polymerase chain reaction (PCR), which prevents the possible loss of sequences in the bacterial cloning [42]. As the new technologies for NGS developed, multiple techniques emerged in the clonal amplification and sequencing steps.

Illumina platform has been the most widely used technology for short-read NGS [43]. In Illumina sequencing, after shearing DNA into small fragments and binding adapters to the ends, solid phase-bridge amplification is performed on a flow cell, which generates clonal clusters (Figure 1.7 A). The invention of reversible dye terminators enabled the sequence by synthesis method (Illumina/Solexa sequencing technology), in which reads are generated without termination until a specified length. As each fluorescent-labeled nucleotide is added to the oligonucleotide, a fluorescent signal is emitted, which is captured by total internal reflection fluorescence microscopy through laser channels (Figure 1.7 B).

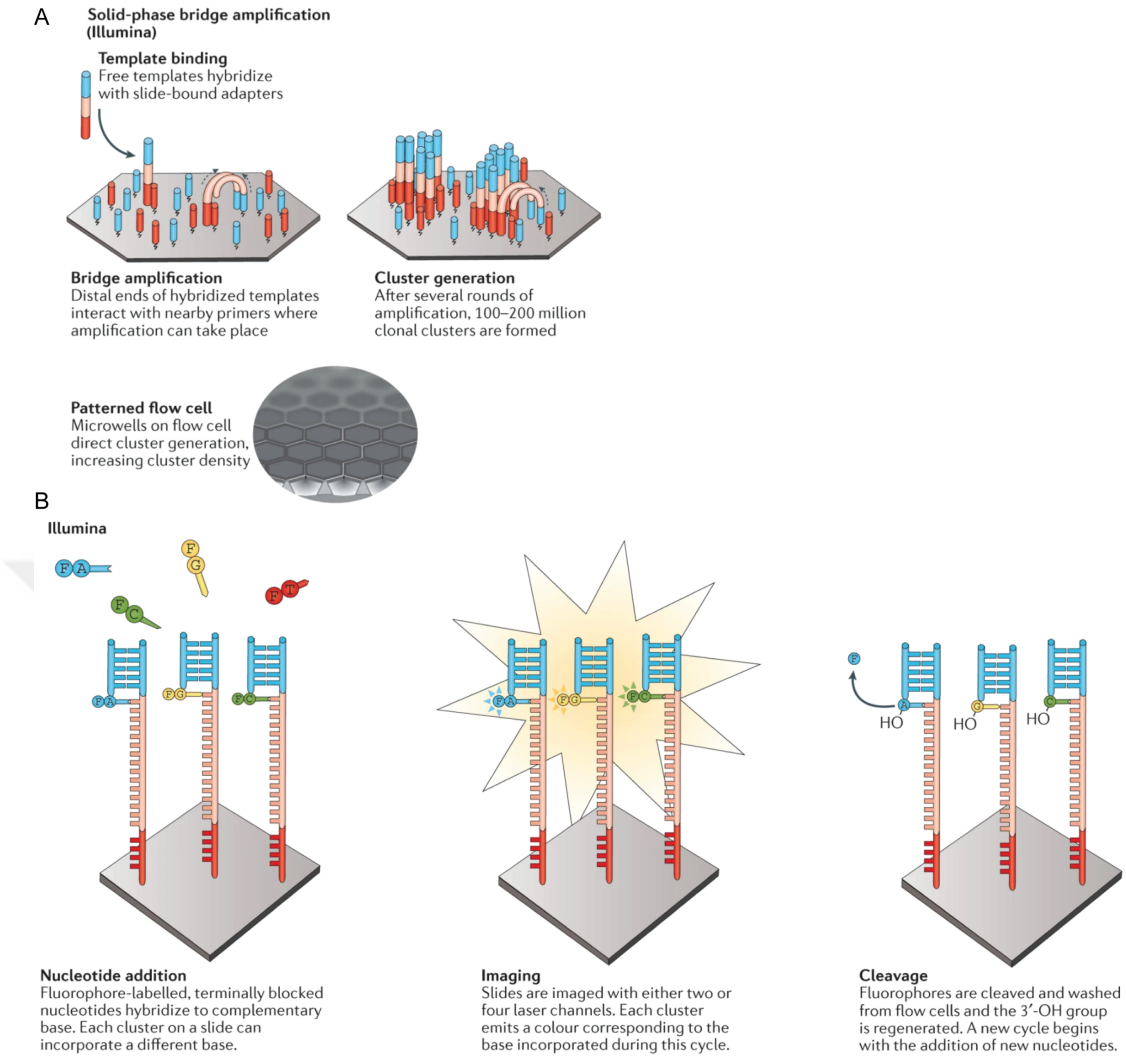



**Figure 1.7: Illumina NGS technology. A** Clonal amplification on patterned flow cell. **B** Sequencing by synthesis. Reprinted by permission from Springer Nature, [43], Copyright ©2016.

When a DNA sample is sequenced on an NGS machine, the output files of NGS are produced in the Binary Base Call (BCL) format. BCL files are converted to FASTQ format, which stores both raw sequence reads and quality scores. Then, the raw, unmapped reads were aligned to a reference genome using bioinformatics tools to generate the files that contain aligned reads. Variants of the sample are called after the quality control steps and variant call format (VCF) files are generated, which can be used in the subsequent analyses [44].

## 1.6 Implementation of sequencing data in medical genetics

### 1.6.1 Interpretation of sequencing variants for the assertion of clinical significance

The great leap in genomics research and increased diagnostic yield of genetic testing are undoubtedly linked to rapid advancement in NGS technology. WES and WGS are increasingly used in clinical settings, allowing physicians to determine the genetic cause for challenging cases with inherited disease [45, 46]. WES and WGS are also utilized in research to identify the novel variants associated with Mendelian and multifactorial disorders. Still, the assessment of the pathogenicity of a variant is often difficult [47]. American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) developed guidelines for the clinical interpretations of genetic variations [48]. The criteria and the evidence framework for variant classification proposed by ACMG-AMP were shown in Figure 1.8.

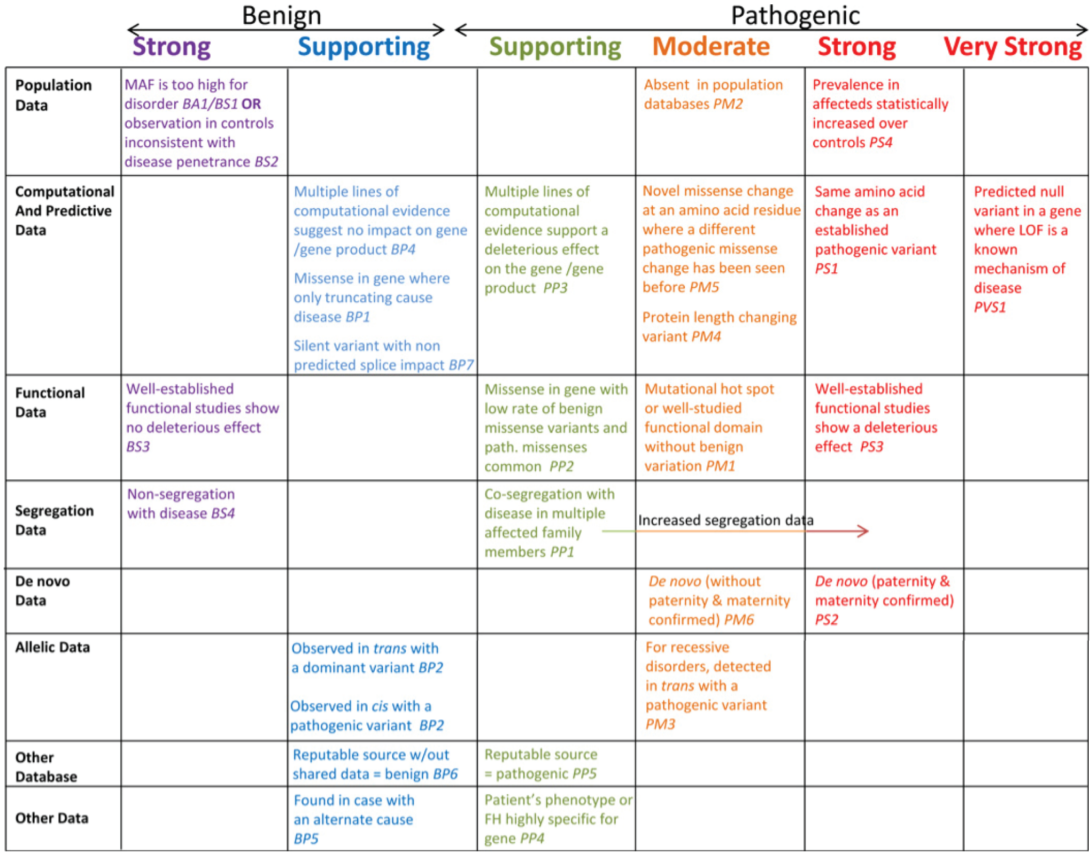



**Figure 1.8: The criteria and the evidence framework for variant classification.** Reprinted by permission from Springer Nature, Genetics in Medicine, [48], Copyright ©2015.

Using these lines of evidence and the following rules in Table 1.1, genetic variants are classified into five groups of clinical significance, namely pathogenic, likely pathogenic, benign, likely benign, and uncertain significance. Clinical laboratory geneticists perform variant classification in their practice by following these guidelines, as well as their professional judgment.

### 1.6.2 Disease databases for genetic variants

Disease databases contain data for genes, phenotype information for the associated disease and genetic variants as well as the pathogenicity assessment of the variants. The well-known disease databases are Online Mendelian Inheritance in Man (OMIM), ClinVar, and Human Gene Mutation Database (HGMD). There are also locus and disease-specific variant databases that contain information on sets of phenotypes or loci such as the Leiden Open (source) Variation Database.

**Table 1.1** Rules for combining criteria to classify sequence variants

| | |
|---|---|
| Pathogenic | (i) 1 Very Strong (PVS1) *AND*<br>(a) $\geq$ 1 Strong (PS1-PS4) *OR*<br>(b) $\geq$ 2 Moderate (PM1-PM6) *OR*<br>(c) 1 Moderate (PM1-PM6) and 1 Supporting (PP1-PP5) *OR*<br>(d) $\geq$ 2 Supporting (PP1-PP5)<br>(ii) $\geq$ 2 Strong (PS1-PS4) *OR*<br>(iii) 1 Strong (PS1-PS4) *AND*<br>(a) $\geq$ 3 Moderate (PM1-PM6) *OR*<br>(b) 2 Moderate (PM1-PM6) *AND* $\geq$ 2 Supporting (PP1-PP5) *OR*<br>(c) 1 Moderate (PM1-PM6) *AND* $\geq$ 4 Supporting (PP1-PP5) |
| Likely pathogenic | (i) 1 Very Strong (PVS1) *AND* 1 Moderate (PM1-PM6) *OR*<br>(ii) 1 Strong (PS1-PS4) *AND* 12 Moderate (PM1-PM6) *OR*<br>(iii) 1 Strong (PS1-PS4) *AND* $\geq$ 2 Supporting (PP1-PP5) *OR*<br>(iv) $\geq$ 3 Moderate (PM1-PM6) *OR*<br>(v) 2 Moderate (PM1-PM6) *AND* $\geq$ 2 Supporting (PP1-PP5) *OR*<br>(vi) 1 Moderate (PM1-PM6) *AND* $\geq$ 4 Supporting (PP1-PP5) |
| Benign | (i) 1 Stand-Alone (BA1) *OR*<br>(ii) $\geq$ 2 Strong (BS1-BS4) |
| Likely Benign | (i) 1 Strong (BS1-BS4) and 1 Supporting (BP1-BP7) *OR*<br>(ii) $\geq$ 2 Supporting (BP1-BP7) |
| Uncertain significance | (i) Other criteria shown above are not met *OR*<br>(ii) the criteria for benign and pathogenic are contradictory |



OMIM is a comprehensive public resource for human genes and their associated phenotypes. It provides information on gene structure and function, genotype/phenotype correlations, allelic variants, and associated publications, where available [49]. ClinVar and HGMD provide interpretations for probabilities of pathogenicity of variants that underlie or are closely associated with human disorders. ClinVar is a public database that accepts submissions from researchers, using both standard terms of ACMG/AMP for clinical significance and non-standard terms that cannot be met by ACMG/AMP [50].HGMD contains manually curated data of germline variants from peer-reviewed publications of human inherited disease [51]. These types of automated/curated databases are frequently consulted in both clinical practice and research. The variant categories in ClinVar and HGMD are listed in Table 1.2.

**Table 1.2** Classes of variants in ClinVar and HGMD

| Clinvar | HGMD |
|---|---|
| • Benign | • Frameshift or truncating variants (FTV) |
| • Likely Benign | • Functional polymorphisms (FP) |
| • Variant of unknown significance | • Disease-associated polymorphisms with |
| • Likely pathogenic | supporting functional evidence (DFP) |
| • Pathogenic | • Disease-associated polymorphisms (DP) |
| • Drug response | • Probable/possible pathogenic mutations |
| • Association | (DM?) |
| • Risk factor | • Disease-causing mutations (DM) |
| • Protective | |
| • Affects | |
| • Conflicting data from submitters | |

### 1.6.3 *In silico* prediction algorithms

There are a number of *in silico* prediction tools that aid the interpretation of the different types of genetic variants. These algorithms determine the effect of variants at the nucleotide and gene product level by predicting their potential impact on canonical or alternative transcripts [48]. Some also provide data on the expression of the affected transcripts at the tissue level, which can give a clue about the possible effects on the phenotype [52]. Such tools are used to predict the impact of missense variants, the variants that might affect splicing or have an LoF effect, and noncoding variants.

The main criteria for prediction tools while assessing missense variants are the physical properties of the altered amino acid, the location of the amino acid on the protein, and the evolutionary conservation. The measurement of one or a combination of these criteria is used in various *in silico* algorithms that assess the predicted impact of a missense change. The predictive accuracy for the previously-known disease variants was shown to be around 65-80% for most tools [48]. Because the different tools use different criteria, the predictive accuracy can be raised if *in silico* prediction tools are used in combinations with one another [53,54]. The predictive tools for splicing depend on the prediction of creation of a new splice site or loss of an existing one based on probabilistic methods of splicing motifs and the potential effect on primary or alternative transcripts [55]. LoF prediction tools assess the effect of protein-truncating variants, namely nonsense (stop gain) variants, frameshift indels, and canonical splice site variants using the

combination of certain criteria such as evolutionary conservation, possible rescue sites, location on functional domains, and biological networks [56, 57].

ACMG/AMP recommends that *in silico* prediction tools should be used in combinations, yet they should be counted as a single piece of evidence for variant pathogenicity. Variant interpretation using software programs should be implemented carefully because they provide only predictions [48]. Despite the debates on the adequacy of *in silico* prediction tools for variant classification, they are practical for predicting the mutational effects in large datasets.

### 1.6.4 Causative variant and gene identification for Mendelian disease

High throughput sequencing reveals a huge amount of genetic variants, which complicates the prioritization process of causal variants for Mendelian disease. Variant prioritization is a method for decreasing the millions of variants of an individual with a genetic disease to a small subset of deleterious variants that might be causative of the phenotype [58]. Variant prioritization tools are *in silico* prediction algorithms that utilize phylogenetic conservation scores, possible effects on protein structure, functional genomics data, and variant frequencies [59]. Variant prioritization is typically a previous step to gene prioritization for the Mendelian disease gene discovery. Gene prioritization tools use the combination of information on genotype frequencies in affected and unaffected individuals, variant frequencies in databases, inheritance models, pedigree information, patient phenotypes, and network analysis to prioritize likely damaged genes associated with a phenotype [60].

Although variant frequencies are efficient measures for variant and gene prioritization, population stratification should also be considered for the candidate gene approach. Although there are exceptions, pathogenic variants that underlie Mendelian diseases should be rare in all human subpopulations. It is a common observation that the average frequency of a variant can be lower than its frequency in certain subpopulations [60]. For instance, a very rare variant in a EUR population can be observed with a very high frequency in an EAS population. Thus, variant frequencies should be examined using variant resources that

contain samples with diverse ancestral origins when performing variant and gene prioritization. Therefore, the generation of variant databases including samples from underrepresented populations in studies of Mendelian disorders will be crucial. For example, GME Variome was shown to reduce the number of prioritized variants by four to sevenfold when it is applied to studies of unsolved recessive disorders in the Middle East population [17].

### 1.6.5 GWAS of complex traits

GWAS are the primary approach to identify common genetic variants that are associated with predisposition to complex traits. Thousands of GWAS with up to millions of study samples have been published and uncovered the genetic predisposition to many types of cancer, metabolic, cardiovascular, and rheumatic diseases and genetic influence on continuous traits such as height and blood lipid level [61]. In 2020, National Human Genome Research InstituteEuropean Bioinformatics Institute (NHGRI-EBI) GWAS Catalog listed more than 4,346 published studies across more than 4,933 diseases and traits, and the numbers are rapidly increasing [62]. The main steps of GWAS are shown in Figure 1.9. Briefly, DNA samples of affected and unaffected individuals from the target population are collected and their genotypes were obtained using sequencing (mostly using SNP arrays). After quality control steps, the untyped variants were imputed using a reference panel for imputation and statistical tests were performed. The most common plots for the evaluation of the results were the quantile-quantile plots and Manhattan plots to demonstrate SNPs that reach genome-wide significance, which have typically a $P$ value lower than 5x10-8 [63]. According to the density of sequenced or imputed genotypes, this method generally reveals the loci that can be associated with the disorder, rather than direct causal gene or variant.

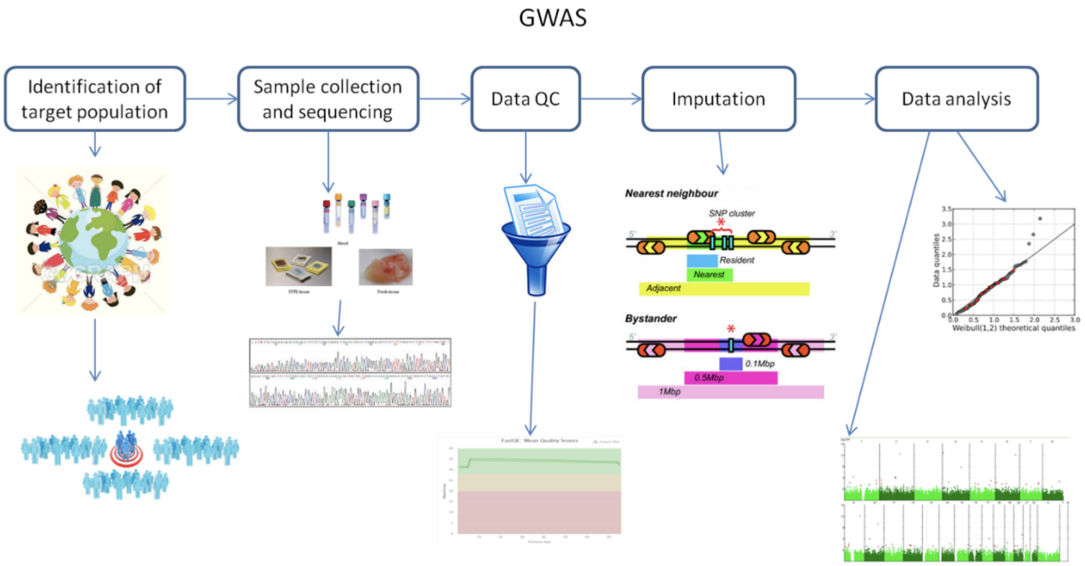



**Figure 1.9: Flowchart for GWAS.** Publisher: MDPI, Basel, Switzerland, [63],licensed under (CC BY 4.0), 2020

GWAS are usually performed using standard arrays that detect SNPs with a minor allele frequency (MAF) >5%. According to the "common variant-common disease" hypothesis, studying common variants was thought to identify most of the genetic risk for common, complex diseases. Yet, the heritability of complex diseases could not be fully explained by this approach. Two main reasons behind the missing heritability might be [64]:

1. The International HapMap Project facilitated the GWAS by enabling array design using the information on common SNPs and patterns of LD. However, the vast majority of GWAS did not evaluate the effect of variants with lower frequencies, which led researchers to consider the "common disease-rare variants hypothesis. The hypothesis is based on the possibility that multiple rare variants (MAF <1%) cause predisposition to common diseases. The implementation of next-generation sequencing data to GWAS allowed the discovery of more disease-associated loci, though results show that neither hypothesis can be accepted standalone.

2. Most of the GWAS was conducted on EUR-descent populations (88.98% https://gwasdiversitymonitor.com/ by October 6th, 2021). The lack of diversity prevents the identification of genetic variants that contribute to disease in populations worldwide, and subsequently, the predictive performance decreases. The inclusion of samples from the different ancestral

backgrounds helps to detect informative markers and true associations more easily.

#### 1.6.5.1 Population-based reference panels for genotype imputation

High throughput sequencing provided two approaches to be utilized in GWAS: design of next-generation genotyping arrays and improved imputation reference panels [65]. Genotype imputation is a statistical method to 'predict' the genotypes that were not sequenced in the first place. The algorithm evaluates the observed genotypes and uses similar stretches of reference panels of densely genotyped or completely sequenced samples to infer untyped variants (Figure 1.10. The most beneficial 'completely sequenced' dataset has been that of 1000GP, which allows the imputation and subsequent testing of association for millions of variants, including SNVs, short indels, and structural variations.

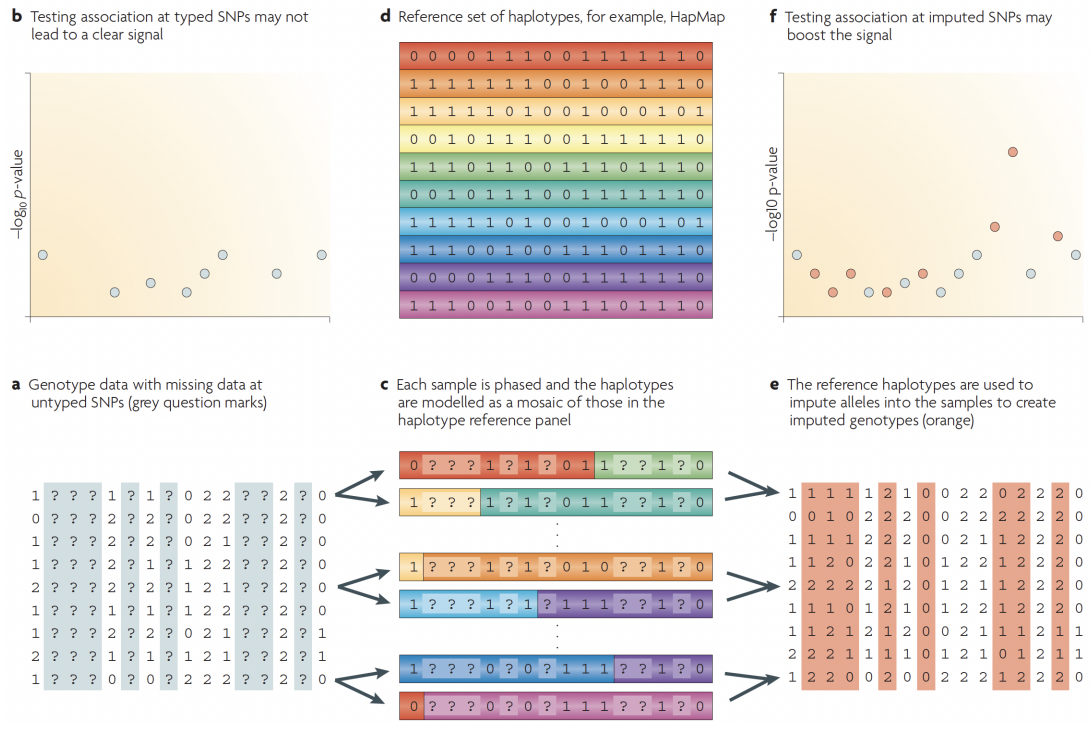



**Figure 1.10: The principles of imputation.** The figure shows that association testing using only typed SNPs might not result in a significant association. Imputation using a reference set of haplotypes helps to predict untyped genotypes. Using both typed and imputed genotypes can lead to detecting significant associations. Reprinted by permission from Springer Nature, [66], Copyright ©2015.

Generation of population-specific reference panels for genotype imputation aids to increase the number of accurately imputed variants. Specifically, the accuracy of imputation increases especially for rare variants when the population-specific reference panels are combined with a pre-existing panel such as 1000GP and HapMap [27, 35, 67].

### 1.6.6 Carrier frequencies for recessive disorders and the secondary findings in large sequencing datasets

Large-scale genomic databases and population-specific variomes catalog high-resolution variant data and variant burden in distinct ethnic groups, which aid in assessing the carrier frequencies (CFs) for recessive disorders [53, 68, 69]. One specific example is the CF calculations for the genes associated with the diseases in the newborn screening (NBS) program. NBS programs aim to detect disorders that benefit from early diagnosis, which are mostly diagnosed using metabolic and/or genetic tests [70]. Detecting carriers for a recessive disease is helpful for both the national health system to develop strategies for possible citizens with the disease and individuals who might want to consider their reproductive choices.

Moreover, secondary findings, variants in genes that are not sequenced for the primary clinical diagnosis, yet can have an impact on the individual or their family members, can be identified in sequence databases. Recently, ACMG has expanded the list of 59 actionable genes to 73 genes with monogenic disorders for which certain preventive measures can be undertaken [71]. Several studies in diverse population cohorts have assessed the numbers of previously reported pathogenic (RP), predicted pathogenic (PP) variants in genes responsible for rare Mendelian disorders, in genes recommended by ACMG, or in sets of genes specific to a particular phenotype(s) [68,72,73]. These efforts depicted the genetic landscape of the respective populations in terms of medical genetics by calculating CFs and incidence rates of some genetic disorders, also the per-individual burden of RP and/or PP variants.

## 1.7 Anatolia and the TR Population

From prehistoric times until today, Anatolia (or Asia Minor)has been a center for many civilizations and has served as a hub of gene flow and admixture between populations due to its geographical location between Asia, Africa, and Europe. Even before the establishment of ancient empires, genetic findings of admixture between Anatolian, Caucasus, and Northern Levantine populations, and a long-distance migration from Central Asia to Anatolia were documented [74]. Because of its strategic position between Asia and Europe, Anatolia was exposed to many expansions and conquests. Hattians, Hurrians, Assyrians, Hittites, Greeks, Thracians, Phrygians, Urartians, Persians, Armenians, Turks, and many other civilizations lived in Anatolia and transferred their traditions, culture, art, and genetic heritage to the present-day inhabitants of the region [75].

The Turkic peoples are a collection of ethnolinguistic groups, originating from Central Asia. The migration of the Turkic tribes in Western Asia and Eastern Europe occurred between the 6th and 11th centuries. With the victory of the Battle of Manzikert in 1071, Anatolia, whose doors were opened to the Seljuks, became the home of TR people for the first time. Following the fall of the Seljuk Empire, several Anatolian principalities (beyliks) were founded. The Ottoman beylik dominated the rest of the beyliks and transformed into the Ottoman Empire by rapidly expanding into Europe and Anatolia and had a multinational character after the conquest of Constantinople in 1453. The peak of the Ottoman power was achieved between the 16th and 17th centuries when 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. After the decline period in the late 18th century and the dissolution period in the 19th century, World War 1 led to falling of the Ottoman Empire. The Republic of Turkey was founded in 1923 following the Turkish War of Independence (1919-1922) and is currently home to more than 80 million people. Turkish-speaking people constitute the major ethnolinguistic group in Turkey. Additionally, there are five independent Turkic countries in Central Asia, where more than 70 million people live. Turkic people also live as a minority in many other countries in Asia, Europe, and North Africa. Studies investigating the genetic contribution of ancient or modern-day Central Asian populations

to modern-day Anatolia detected that the proportion of gene flow was between 3% to 30% [76–78]. These findings demonstrate the possibility that the modern-day TR population in Turkey is derived from intermixing between pre-existing Anatolian and Turkic peoples.

## 1.8 Aim of the study

The GME Variome has been the largest resource representing the genetic variation in Turkey by containing 140 WES samples from the TR Peninsula [17]. The population structure of Turkey and its relationship with Europeans, Middle Easterns, and Central Asians were also investigated using SNP array data of 64 samples from Turkey [79]. Another study investigated the genetic variation within Turkey using WGS data of 16 TR individuals [80]. The array sequencing data of 56 TR individuals were also presented in a study investigating the origin of farming in the Ancient Near East using ancient and modern Near East samples [81]. These efforts revealed signatures of a high level of admixture, previously uncaptured variation, and a high rate of consanguinity in the TR population, as well as demonstrated the importance of studying the genetic structure of the TR population for the understanding of population and human genetics. Turkey has a higher population size compared to its immediate neighbors; hence the previous small-scale genetic studies do not provide a comprehensive representation of the genetic structure and variation of the TR population. The prevalence of consanguineous marriages is very high in the TR population [18]. The rate is up to 34% in Eastern Turkey, and the majority (75%) of consanguineous marriages occur between first-degree cousins [82]. This situation leads to elevated homozygosity in the population, and consequently brings a higher risk for recessive disease. Thus, studies conducted in inbred populations contribute to identifying disease gene discovery for recessive Mendelian disorders [16]. The declining cost of NGS technology helped researchers generate a huge amount of sequencing data from diverse populations. The contribution of large-scale variant datasets and reference panels containing population-based haplotypes are indisputable for the study of human genetics [27,35,67,83]. Therefore, the generation of such resources for the TR population is indispensable.

Generation of a large-scale variant resource subsequently enables researchers to assess CFs for recessive disorders and secondary findings in the dataset. There are many studies for the discovery of such findings in large variant databases such as ESP, gnomAD, or population-based resources [53, 72, 84, 85]. Although most prevalent disorders such as phenylketonuria (PKU) and biotinidase deficiency are detected through the national NBS program in Turkey, up-to-date information on the CF of the mutations in the causative genes or disease incidence are currently unavailable [86]. Moreover, the rate of secondary findings or a comprehensive examination of carrier rates for Mendelian disorders in the TR population has not been investigated yet. Accordingly, the aims of this study are:

1. To investigate the fine-scale genetic structure of the TR population using WES and WGS data.
2. To produce a resource that catalog genetic variation in the TR population for future disease gene identification studies.
3. To generate a reference panel for genotype imputation that contains phased haplotypes of TR individuals for future GWAS.
4. To present CFs for recessive disorders and the rate of secondary findings in the TR population.

# Chapter 2

# Materials and Methods

## 2.1 Study Samples

The study consisted of 3,864 TR individuals who were either whole-exome ($n$ = 3,072) or whole-genome ($n$ = 792) sequenced. The blood samples of these individuals were collected to investigate the genetic basis of obesity, polycystic ovarian syndrome (PCOS), delayed sleep phase disorder (DSPD), amyotrophic lateral sclerosis (ALS), essential tremors, ataxia, Parkinson's disease, various immunological disorders, and various neurological disorders. (Table 2.1) Written informed consents, which provided the permission to use the DNA samples and demographic information for research purposes and to share the data, were obtained from all study participants during the sampling.

## 2.2 DNA sequencing and variant calling

WES was performed at the Yale Center for Genome Analysis, TUBITAK, or Macrogen on the HiSeq2000, HiSeq2500, or HiSeq4000 platforms with 100-bp paired-end reads. IDT xGen Exome Research 392 Panel v1.0 capture, Roche SeCap EZ Whole Exome V1, xGen Exome Research 392 Panel v1.0 capture, Roche

**Table 2.1** 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 | 765 | WES |
| Parkinson's Disease | 53 | WES |
| Polycystic ovarian syndrome | 15 | WES |
| Various neurological and immunological disorders | 1,559 | WES |
| Amyotrophic lateral sclerosis | 792 | WGS |
| Total | 3,072 + 792 = 3,864 | |

Reprinted from [87].

SeCap EZ Whole Exome V2, Roche SeCap EZ Whole Exome V3, or Agilent SureSelect Human All Exon V6 kits were used for WES according to the manufacturers protocol. Base-calling, read filtering, and demultiplexing were performed by following the Illumina run processing pipeline. The read pairs were mapped to the human genome build GRCh37 using Burrows-Wheeler Aligner (BWA) v.0.7.17 [88]. Duplicate reads were marked and removed via Picard tools [89]. The variant discovery was performed following Best Practices workflows of Genome Analysis Toolkit v.3.7 [90]. Base quality score recalibration (BQSR) and local realignment around indels were carried out with GATK. HaplotypeCaller was employed to generate GVCF files, followed by joint genotyping using GenotypeGVCFs. Multiallelic variants were split into biallelic with LeftAlignAndTrimVariants. Since multiple platforms were used to sequence different cohorts, genotype calling was limited to the intersection of target regions of exome sequencing kits that overlap with consensus coding sequence (CCDS) build 15 coding exons to overcome potential batch effects [91]. In total, 2,694,125 variants were identified. 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. Variant calling was carried out using the Isaac variant caller [92]. The GVCF files for all WGS samples were jointly genotyped using Illumina gvcfgenotyper [93]. Normalization, realignment around indels, and splitting multiallelic variants were performed using BCFtools [94]. The final joint VCF file was lifted over to human genome build GRCh37 with Picard tools using hg19 to b37 chain file. 72,982,375 variants were identified using WGS. Coverage calculations of the WES and WGS samples were performed using mosdepth,

SAMtools and BCFtools [94].

## 2.3 Sample-, genotype-, and variant-based QC filtering

Statistical outliers of WES and WGS samples were evaluated separately using BCFtools stats. By following a similar approach to sample-based QC filtering of gnomAD, four measures were evaluated, namely number of singletons, transition/transversion ratio, average depth and the total number of variants [30]. The medians and median absolute deviations of these four measures were calculated and 89 WES samples were removed from the dataset because they fell outside five absolute deviations from the median. WGS batch did not contain any outlier samples.

For the selection of high-quality genotypes and variants from the WES and WGS data, genotype- and variant-based QC filtering were applied. First, genotypes with a depth $< 8$ and a genotype quality (GQ) $< 20$ were identified and converted to missing genotypes. Second, variants with a Phred-scaled quality score $< 30$ were removed from the dataset. Then, variants with a missingness rate higher than 20% across all individuals were filtered out. 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. The analyses were restricted to genomic regions called by gnomAD and all sites failing QC according to the filtering method of gnomAD were excluded ($n$ = 1,244,833) [30]. For the WGS samples, Sample-based quality control measures and the mean number of novel variants (not present in dbSNP build 151) for the exome regions of the WES and WGS datasets were computed using VariantEval walker of GATK [21, 31].

Relatedness status of TR individuals was determined using KING [95]. Due to high rates of consanguineous marriage in the TR population, a high level of inbreeding was expected. Also, only coding sequence variants were used to

infer genetic relations in the TR population. These two factors were previously assumed to cause an overestimation of relatedness. [31, 82, 95] Thus, following a liberal approach, second degree and closer relatives were removed from the dataset by using a kinship coefficient threshold of 0.0884 [95]. 413 samples were removed after this step. Finally, 773 WGS and 2,589 WES samples corresponding to a total of 3,362 individuals remained. The cohort demographics were shown in Table 2.2.

The study cohort is composed of the sequencing data of individuals whose samples were originally recruited to identify the genetic basis of various disorders, therefore, the AF of certain disease-associated variants might be overestimated. In order to prevent possible inflation, we determined the variants in the genes that are causally associated with the phenotypes that were originally intended to be investigated. 206 disease-associated variants were excluded from the dataset (Table A.1). In total, 1,123,248 WES and 45,981,720 WGS "high quality" variants were used for the downstream analyses. The LiftoverVcf utility of Picard tools was employed to obtain the GRCh38 positions of the variants. The hg19 to GRCh38 chain file was downloaded from the UCSC website. 48,141,045 (99.57%) variants were successfully lifted over to GRCh38 [96].

**Table 2.2** Demographics after exclusion of low quality and related samples

| Cohort | Mean age | Gender (Female/Male) | Affected | Unaffected | Unknown |
|---|---|---|---|---|---|
| Amyotrophic lateral sclerosis | 54.71 ± 14.77 | 395/514 | 698 | 211 | 0 |
| Ataxia | 44.3 ± 14.82 | 75/73 | 101 | 47 | 0 |
| Delayed sleep phase disorder | 28.38 ± 6.99 | 8/10 | 18 | 0 | 0 |
| Essential tremors | 52.37 ± 19.87 | 44/35 | 57 | 22 | 0 |
| Obesity | 38.87 ± 12.39 | 491/181 | 560 | 112 | 0 |
| Parkinsons disease | 47.18 ± 19.43 | 13/16 | 19 | 10 | 0 |
| Polycystic ovarian syndrome | 29 ± 10.11 | 9/0 | 9 | 0 | 0 |
| Various neurological and immunological disorders | 42.12 ± 17.88 | 673/825 | 32 | 26 | 1,440 |
| Total | 47.41± 16.41 | 1,708/1,654 | 1,494 | 428 | 1,440 |

Reprinted from [87].

## 2.4 Evaluation of variant calling and filtering

The accuracy of the variant calling pipeline and the QC filtering method was evaluated by producing technical replicates of WES and WGS samples. GenomeAsia 100K study used a similar method to evaluate their variant calling and QC filtering method. [34] In total, 38 technical replicates were produced by sequencing TR samples in different batches. Among those, 11 and 6 samples were sequenced in two different batches of WES and WGS, respectively. Additionally, 21 samples were sequenced using both WES and WGS platforms. Variant sensitivity and specificity, variant positive predictive value (PPV), genotype concordance, and non-ref genotype concordance were calculated for SNVs and indels using the VariantEval tool of GATK after calling variants and applying the QC filtering method. On average, highly accurate genotypes were detected based on sensitivity, specificity, positive predictive value, and concordance rates. These results were similar to the previous studies investigating technical accuracy of sequencing data [97, 98].

**Table 2.3** Concordance results of the technical replicates

| Technical Replicates | Variant Type | Variant Sensitivity | Variant PPV | Variant Specificity | Genotype Concordance | Non-REF Genotype Concordance |
|---|---|---|---|---|---|---|
| Exome/Exome | SNV | 0.98± 0.02 | 0.97± 0.02 | 0.95± 0.03 | 0.97± 0.02 | 0.97± 0.02 |
| (n = 11) | INDEL | 0.79± 0.18 | 0.85± 0.09 | 0.84± 0.09 | 0.87± 0.16 | 0.87± 0.16 |
| Exome/Genome | SNV | 0.93± 0.09 | 0.99± 0.002 | 0.98± 0.003 | 0.97± 0.007 | 0.97± 0.007 |
| (n = 21) | INDEL | 0.61± 0.16 | 0.86± 0.04 | 0.87± 0.06 | 0.85± 0.02 | 0.85± 0.02 |
| Genome/Genome | SNV | 0.99± 0.003 | 0.99± 0.0001 | 0.99± 0.0001 | 0.96± 0.005 | 0.97± 0.004 |
| (n = 6) | INDEL | 0.89± 0.08 | 0.97± 0.003 | 0.96± 0.005 | 0.76± 0.07 | 0.80± 0.07 |

Reprinted from [87].

## 2.5 Population structure analyses

Four different datasets were generated to evaluate the genetic structure of the TR population in a regional and global context. Populations used in the analyses and their sample sizes were listed in Table S4. The TR dataset (n = 3,362) that was used for the analysis of genetic variation within the TR population was produced by the following steps. Exome data was previously shown to provide accurate results when analyzing the genetic structure of populations [99]. Therefore, a BED file that contains the overlapping regions of the target regions of WES kits and CCDS regions was produced. Using this BED file, the exome region of WGS samples was extracted and combined with that of WES samples. For the global dataset, 13 populations from 1000GP were selected [27]: AFR populations YRI and LWK; EUR populations GBR, TSI, IBS, and FIN; SAS populations GIH, BEB, PJL, and ITU; EAS populations CHB, CHS, and JPT ($n$ = 1,299). Afterward, WGS data of the selected 1000GP populations and TR-WGS samples ($n$ = 773) and genomic variations of Near-East populations from Lazaridis *et al.* [81] ($n$ = 1,430) were combined. For the samples sequenced in both 1000GP and Lazaridis et al. , sequence data of Lazaridis et al. ($n$ = 156) were removed from the dataset. Third, the regional dataset was produced (n = 1,805) by listing the populations with the closest relationship with the TR population according to Wrights fixation index ($F_{ST}$, pairwise $F_{ST}$ of <0.01 were included.). Lastly, the phylogeny dataset ($n$ = 5,357) was generated by combining the AF data of all TR-WES and TR-WGS samples, Middle Eastern populations from Scott *et al.* [17], and 1000GP to conduct phylogenetic tree analysis.

**Table 2.4** Populations included in the study

| Population | | Super population | | Sequencing method | *n* | Study |
|---|---|---|---|---|---|---|
| YRI | Yoruba in Ibadan, Nigeria | AFR | African | WGS | 108 | 1000GP |
| LWK | Luhya in Webuye, Kenya | AFR | African | WGS | 96 | 1000GP |
| GWD | Gambian in Western Divisions in the Gambia | AFR | African | WGS | 113 | 1000GP |
| MSL | Mende in Sierra Leone | AFR | African | WGS | 85 | 1000GP |
| ESN | Esan in Nigeria | AFR | African | WGS | 99 | 1000GP |
| ASW | Americans of African Ancestry in SW USA | AFR | African | WGS | 61 | 1000GP |
| ACB | African Caribbeans in Barbados | AFR | African | WGS | 96 | 1000GP |
| MXL | Mexican Ancestry from Los Angeles, CA, USA | AMR | American | WGS | 64 | 1000GP |
| PUR | Puerto Ricans from Puerto Rico | AMR | American | WGS | 104 | 1000GP |
| CLM | Colombians from Medellin, Colombia | AMR | American | WGS | 94 | 1000GP |
| PEL | Peruvians from Lima, Peru | AMR | American | WGS | 85 | 1000GP |
| - | Albanian | BLK | Balkan | H.O. array*, WGS | 6, 1 | Lazaridis *et al.* [81], SGDP |
| - | Bulgarian | BLK | Balkan | H.O. array*, WGS | 10, 1 | Lazaridis *et al.* [81], SGDP |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | Crete | BLK | Balkan | WGS | 2 | SGDP |
| - | Croatian | BLK | Balkan | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Greek | BLK | Balkan | H.O. array*, WGS | 20, 2 | Lazaridis *et al.* [81], SGDP |
| - | Abkhazian | CAU | Caucasus | H.O. array*, WGS | 9, 2 | Lazaridis *et al.* [81], SGDP |
| - | Adygei | CAU | Caucasus | H.O. array*, WGS, Illumina HuHap 650k | 17, 2, 17 | Lazaridis *et al.* [81], SGDP, HGDP |
| - | Armenian | CAU | Caucasus | H.O. array*, WGS | 10, 2 | Lazaridis *et al.* [81], SGDP |
| - | Balkar | CAU | Caucasus | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Chechen | CAU | Caucasus | H.O. array*, WGS | 9, 1 | Lazaridis *et al.* [81], SGDP |
| - | Georgian | CAU | Caucasus | H.O. array*, WGS | 10, 2 | Lazaridis *et al.* [81], SGDP |
| - | Kumyk | CAU | Caucasus | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Lezgin | CAU | Caucasus | H.O. array*, WGS | 9, 2 | Lazaridis *et al.* [81], SGDP |
| - | Nogai | CAU | Caucasus | H.O. array* | 9 | Lazaridis *et al.* [81] |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | North Ossetian | CAU | Caucasus | H.O. array*, WGS | 10, 2 | Lazaridis *et al.* [81], SGDP |
| - | Altaian | CNA | Central and North Asian | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Dolgan | CNA | Central and North Asian | H.O. array* | 3 | Lazaridis *et al.* [81] |
| - | Even | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Kalmyk | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Kyrgyz | CNA | Central and North Asian | H.O. array* | 9 | Lazaridis *et al.* [81] |
| - | Mansi | CNA | Central and North Asian | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Mongola | CNA | Central and North Asian | H.O. array* | 6 | Lazaridis *et al.* [81] |
| - | Selkup | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Tajik | CNA | Central and North Asian | H.O. array* | 8 | Lazaridis *et al.* [81] |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | Tubalar | CNA | Central and North Asian | H.O. array* | 22 | Lazaridis *et al.* [81] |
| - | Turkmen | CNA | Central and North Asian | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Tuvinian | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Ulchi | CNA | Central and North Asian | H.O. array* | 25 | Lazaridis *et al.* [81] |
| - | Uygur | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Uzbek | CNA | Central and North Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Yakut | CNA | Central and North Asian | H.O. array*, Illumina HuHap 650k | 20, 25 | Lazaridis *et al.* [81], HGDP |
| - | Yukagir | CNA | Central and North Asian | H.O. array* | 19 | Lazaridis *et al.* [81] |
| CHB | Han Chinese in Beijing, China | EAS | East Asian | WGS | 103 | 1000GP |
| JPT | Japanese in Tokyo, Japan | EAS | East Asian | WGS | 104 | 1000GP |
| CHS | Southern Han Chinese | EAS | East Asian | WGS | 104 | 1000GP |
| CDX | Chinese Dai in Xishuangbanna, China | EAS | East Asian | WGS | 93 | 1000GP |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| KHV | Kinh in Ho Chi Minh City, Vietnam | EAS | East Asian | WGS | 99 | 1000GP |
| - | Daur | EAS | East Asian | H.O. array* | 9 | Lazaridis *et al.* [81] |
| - | Oroqen | EAS | East Asian | H.O. array* | 9 | Lazaridis *et al.* [81] |
| - | Tu | EAS | East Asian | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Xibo | EAS | East Asian | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Chuvash | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| CEU | Utah Residents (CEPH) with Northern and Western European Ancestry | EUR | European | WGS | 99 | 1000GP |
| TSI | Toscani in Italia | EUR | European | WGS | 107 | 1000GP |
| FIN | Finnish in Finland | EUR | European | WGS | 99 | 1000GP |
| GBR | British in England and Scotland | EUR | European | WGS | 89 | 1000GP |
| IBS | Iberian population in Spain | EUR | European | WGS | 107 | 1000GP |
| - | Basque | EUR | European | H.O. array* | 29 | Lazaridis *et al.* [81] |
| - | Belarusian | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Czech | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Estonian | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | French | EUR | European | H.O. array* | 61 | Lazaridis *et al.* [81] |
| - | Hungarian | EUR | European | H.O. array* | 20 | Lazaridis *et al.* [81] |
| - | Icelandic | EUR | European | H.O. array* | 12 | Lazaridis *et al.* [81] |
| - | italian North | EUR | European | H.O. array* | 20 | Lazaridis *et al.* [81] |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | Italian South | EUR | European | H.O. array* | 6 | Lazaridis *et al.* [81] |
| - | Lithuanian | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Maltese | EUR | European | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Mordovian | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Norwegian | EUR | European | H.O. array* | 11 | Lazaridis *et al.* [81] |
| - | Orcadian | EUR | European | H.O. array* | 13 | Lazaridis *et al.* [81] |
| - | Romanian | EUR | European | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Russian | EUR | European | H.O. array* | 22 | Lazaridis *et al.* [81] |
| - | Saami | EUR | European | H.O. array* | 1 | Lazaridis *et al.* [81] |
| - | Sardinian | EUR | European | H.O. array* | 27 | Lazaridis *et al.* [81] |
| - | Sicillian | EUR | European | H.O. array* | 11 | Lazaridis *et al.* [81] |
| - | Ukranian | EUR | European | H.O. array* | 9 | Lazaridis *et al.* [81] |
| NEA | North East African | GME | Middle Eastern | WES** | 368 | Scott *et al.* [17] |
| NWA | North West African | GME | Middle Eastern | WES** | 99 | Scott *et al.* [17] |
| AP | Arabian Peninsula | GME | Middle Eastern | WES** | 171 | Scott *et al.* [17] |
| SD | Syrian Desert | GME | Middle Eastern | WES** | 58 | Scott *et al.* [17] |
| - | Algerian | GME | Middle Eastern | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Assyrian | GME | Middle Eastern | H.O. array* | 11 | Lazaridis *et al.* [81] |
| - | Bedouin | GME | Middle Eastern | H.O. array*, WGS, Illumina HuHap 650k | 44, 2, 40 | Lazaridis *et al.* [81], SGDP, HGDP |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | Cypriot | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Druze | GME | Middle Eastern | H.O. array*, WGS, Illumina HuHap 650k | 39, 2, 47 | Lazaridis *et al.* [81], SGDP, HGDP |
| - | Egyptian | GME | Middle Eastern | H.O. array* | 18 | Lazaridis *et al.* [81] |
| - | Hazara | GME | Middle Eastern | H.O. array* | 14 | Lazaridis *et al.* [81] |
| - | Iranian | GME | Middle Eastern | H.O. array*, WGS | 38, 2 | Lazaridis *et al.* [81], SGDP |
| - | Iranian Bandari | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Jew Ashkenazi | GME | Middle Eastern | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Jew Cochin | GME | Middle Eastern | H.O. array* | 5 | Lazaridis *et al.* [81] |
| - | Jew Ethiopian | GME | Middle Eastern | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Jew Georgian | GME | Middle Eastern | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Jew Iranian | GME | Middle Eastern | H.O. array* | 9 | Lazaridis *et al.* [81] |
| - | Jew Iraqi | GME | Middle Eastern | H.O. array*, WGS | 6, 2 | Lazaridis *et al.* [81], SGDP |
| - | Jew Libyan | GME | Middle Eastern | H.O. array* | 9 | Lazaridis *et al.* [81] |
| - | Jew Moroccan | GME | Middle Eastern | H.O. array* | 6 | Lazaridis *et al.* [81] |
| - | Jew Tunisian | GME | Middle Eastern | H.O. array* | 7 | Lazaridis *et al.* [81] |
| - | Jew Turkish | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| - | Jew Yemenite | GME | Middle Eastern | H.O. array*, WGS | 8, 2 | Lazaridis *et al.* [81], SGDP |
| - | Jordanian | GME | Middle Eastern | H.O. array*, WGS | 9, 3 | Lazaridis *et al.* [81], SGDP |
| - | Lebanese | GME | Middle Eastern | H.O. array* | 28 | Lazaridis *et al.* [81] |
| - | Libyan | GME | Middle Eastern | H.O. array* | 6 | Lazaridis *et al.* [81] |
| - | Moroccan | GME | Middle Eastern | H.O. array* | 10 | Lazaridis *et al.* [81] |
| - | Mozabite | GME | Middle Eastern | H.O. array*, WGS, Illumina HuHap 650k | 21, 2, 30 | Lazaridis *et al.* [81], SGDP, HGDP |
| - | Palestinian | GME | Middle Eastern | H.O. array*, WGS, Illumina HuHap 650k | 38, 3, 51 | Lazaridis *et al.* [81], SGDP, HGDP |
| - | Samartian | GME | Middle Eastern | WGS | 1 | SGDP |
| - | Saharawi | GME | Middle Eastern | H.O. array* | 6 | Lazaridis *et al.* [81] |
| - | Saudi | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Syrian | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Tunisian | GME | Middle Eastern | H.O. array* | 8 | Lazaridis *et al.* [81] |
| - | Yemeni | GME | Middle Eastern | H.O. array* | 6 | Lazaridis *et al.* [81] |
| GIH | Gujarati Indian from Houston, Texas, USA | SAS | South Asian | WGS | 100 | 1000GP |
| PJL | Punjabi from Lahore, Pakistan | SAS | South Asian | WGS | 95 | 1000GP |
| BEB | Bengali from Bangladesh | SAS | South Asian | WGS | 86 | 1000GP |

*(continued on next page)*

*Table 2.4 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| STU | Sri Lankan Tamil from the UK | SAS | South Asian | WGS | 102 | 1000GP |
| ITU | Indian Telugu from the UK | SAS | South Asian | WGS | 101 | 1000GP |
| - | Balochi | SAS | South Asian | H.O. array* | 20 | Lazaridis *et al.* [81] |
| - | Brahui | SAS | South Asian | H.O. array* | 21 | Lazaridis *et al.* [81] |
| - | Burusho | SAS | South Asian | H.O. array* | 23 | Lazaridis *et al.* [81] |
| - | Kalash | SAS | South Asian | H.O. array* | 18 | Lazaridis *et al.* [81] |
| - | Makrani | SAS | South Asian | H.O. array* | 20 | Lazaridis *et al.* [81] |
| - | Pathan | SAS | South Asian | H.O. array* | 19 | Lazaridis *et al.* [81] |
| - | Sindhi | SAS | South Asian | H.O. array* | 18 | Lazaridis *et al.* [81] |
| TR-B | Turkish with Balkan Ancestry | TR | Turkish | WGS, WES | 68, 22 | Current study |
| TR-W | Western Turkish | TR | Turkish | WGS, WES | 97, 60 | Current study |
| TR-C | Central Turkish | TR | Turkish | WGS, WES | 75, 366 | Current study |
| TR-N | Northern Turkish | TR | Turkish | WGS, WES | 166, 206 | Current study |
| TR-S | Southern Turkish | TR | Turkish | WGS, WES | 79, 37 | Current study |
| TR-E | Eastern Turkish | TR | Turkish | WGS, WES | 162, 122 | Current study |
| TR-U | Turkish with unknown origin | TR | Turkish | WGS, WES | 126, 1,776 | Current study |
| - | Turkish | TR | Turkish | H.O. array* | 56 | Lazaridis *et al.* [81] |

*: Affymetrix Human Origins array. **: Included only in Treemix analysis using AF data of Scott *et al.* [17].

SGDP: Simons Genome Diversity Project. HGDP: Human Genome Diversity Project. Reprinted from [87].

The same steps for all four datasets were followed to generate variant sets for the analyses. First, SNVs were extracted from the VCF files of TR and 1000GP samples using BCFtools and converted to PLINK binary file format [100]. The sequencing data of Lazaridis *et al.* was also converted to PLINK binary file format using the convertf utility of EIGENSOFT [101]. The binary files were merged and variants were filtered using PLINK v.1.9 and following thresholds. Variants with a missingness rate higher than 20%), variants that deviated from Hardy-Weinberg equilibrium (HWE) with a $P$ of <0.00005, variants with a MAF lower than 0.05. Variants displaying linkage disequilibrium (r2 = 0.5) were also pruned. All plots were generated with the aid of ggplot2, and reshape, dplyr and stringr packages of R software were used for generating necessary data formats for plotting [102–105].

### 2.5.1 Origin of alleles

Grand-maternal and grand-paternal birthplace of the TR individuals were obtained from patient records where available and the numbers of chromosomes from Balkan (TR-B), Central (TR-C), East (TR-E), North (TR-N), South (TR-S), and West (TR-W) subregions were depicted on a map of Turkey. To test any possible bias due to different maternal and paternal lineage, principal component analysis (PCA) was repeated by using both grand-maternal and grand-paternal origins.

### 2.5.2 PCA

The extent of genetic variation among populations was investigated PCA. PCA is a dimensionality reduction method displaying multi-dimensional data in two dimensions and used to visualize the variance between samples.EIGENSOFT SmartPCA tool was used for PCA in the three datasets [101]. The first dataset, the origin-known TR samples from the TR dataset, was used for evaluating the variation in the TR subregions. The second PCA was performed using the global dataset for displaying the genetic variation in a global context. The last PCA was performed with the regional dataset to evaluate the genetic variance in the TR

population as well as genetically closer populations to Turkey. The first four principal components (PCs) explaining the highest degree of variation were plotted for demonstration purposes.

### 2.5.3 Procrustes analysis

Procrustes analysis is a statistical method that transforms data by translating, scaling, and dilating to reduce the sum of squared Euclidean distances between landmark points while preserving the pairwise distances. This technique is used to detect the similarity between maps of genetic variation and geographical maps, and the significance is tested using a permutation test [106]. The first symmetric Procrustes analysis with 100,000 permutations was carried out using the PC1 and PC2 of the PCA with the origin-known TR samples from the TR dataset and the unprojected geographic coordinates (landmark points) of the TR subregions. The midpoints of latitude and longitude of the TR subregions were determined on the map of Turkey. The second Procrustes analysis was performed using the PC1 and PC2 of the PCA with the regional dataset. The geographical coordinates of the capital cities were determined for the populations included in the regional dataset. The R package vegan was employed for the Procrustes analyses [107].

### 2.5.4 tSNE and UMAP analyses

To evaluate the genetic variation within the TR population and in the global context, two additional dimensionality reduction techniques were used. t-distributed stochastic neighbor embedding (tSNE) and Uniform Manifold Approximation and Projection(UMAP) are generally used for single-cell RNA-seq data; however, can also be used to display population stratification as in the case of PCA [108, 109]. Rtsne and umap packages of R software were utilized for the tSNE and UMAP analyses, respectively [110, 111]. Both analyses were performed for the TR and global datasets using PC1-10.

### 2.5.5 Admixture

Population stratification was evaluated with ADMIXTURE, which calculates population AFs and ancestry proportions using a maximum likelihood approach [112]. Ancestral contributions of the TR subregions were determined by running the ADMIXTURE with $k$ from 2 to 8 for the TR dataset where $k = 4$ revealed the lowest cross-validation error. Additionally, analyses with $k$ from 2 to 14 was performed for the global dataset where $k = 12$ revealed the lowest cross-validation error.

### 2.5.6 Phylogenetic tree

Population splits of the diverged populations in the phylogeny dataset were assessed by Treemix software [113]. Treemix constructs a maximum likelihood phylogenetic tree based on the distribution of AFs in the populations compared to the ancestral AF to evaluate the topology of relationships between populations. YRI was selected as the 'root' according to the out-of-Africa hypothesis and previous publications [17]. Additionally, a second Treemix analysis was performed for the TR subregions using the origin-known TR individuals from the TR dataset to demonstrate the effects of local migration events and gene flow on the genetic structure of the TR population. Three migration events were allowed in the analysis. Both analyses repeated several times, which resulted in similar tree structures with the same branch lengths being obtained.

### 2.5.7 Wright's $F_{ST}$

Weir and Cockerham estimation, which is defined as the ratio of the genetic variance between populations to the total genetic variance in the most common ancestral population, was used to calculate the pairwise Wright's $F_{ST}$ values to assess the level of genetic similarity between populations included in the regional dataset [114]. Wright's $F_{ST}$ values were calculated using EIGENSOFT SmartPCA and the results were displayed on a heatmap. Also, the genetic similarity

between the TR subregions and the populations in the regional dataset were calculated using Wright's $F_{ST}$ for evaluating the effect of geographical locations on genetic drift.

### 2.5.8 Linkage disequilibrium decay

Linkage disequilibrium (LD) is the cosegregation of different loci in a nonrandom way. LD has been used to study genotype-phenotype correlations and selection pressures in various organisms. The level of LD ($r^2$) is affected by the effective population size and the number of recombination events [115]. Thus, the rate of genome-wide LD decay reflects the demographic history of a population. The rates of the LD decay in the global populations were analyzed using the PLINK –$r^2$ option. Pairwise correlations without a $r^2$ limit were estimated using a sliding window with 70 Kb in length. The bins of $r^2$ were generated according to the genomic distance between SNPs (up to 70 Kb). The average values of $r^2$ were determined for each bin and plotted against the genomic distance. In the analysis, EUR and BLK samples were combined as EUR because of the relatively low number of samples in the BLK population, which resulted in a very low rate of LD decay in the BLK population and prevented the interpretation of the results.

## 2.6 Inbreeding status and estimation of ROH

### 2.6.1 Inbreeding coefficient

The level of inbreeding of the global populations was determined using the –het algorithm of PLINK. The inbreeding coefficient ($F_{plink}$) is calculated using the ratio of the observed and expected number of homozygous LD pruned genotypes. There were several individuals who have large negative $F_{plink}$ levels, which might be due to a recent intermixing between previously diverged populations or sampling bias [31]. The results were also stratified according to the reported status of parental relationships for the TR samples. The reported parental relationship of the TR-WGS samples was as follows: 538 unrelated, 56 endogamous, 95 consanguineous, and 84 unknown. Individuals were categorized as endogamous

if they are the offspring of two individuals who are not known to be related through their ancestors but are from the same small geographical region where marriages occur recurrently within that small population. Inbreeding coefficients of TR subregions were determined using origin-known TR samples. The statistical significance between inbreeding coefficients of TR and global populations and between TR subregions were determined using the Kruskal-Wallis test or one-way ANOVA after testing normal distribution and homogeneity of variances using Kolmogorov-Smirnov test and Levene's test, respectively. Pairwise comparisons were done using the Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing adjustment.

### 2.6.2 Analysis of ROH

The blocks of homozygous regions (ROHs) in autosomes of TR-WGS and 1000GP samples were determined using PLINK –homozyg algorithm. First, SNPs with MAF <0.05 and those that diverted from HWE with a $P$ <0.00005 were excluded from the dataset. The selected options for the –homozyg algorithm were as follows [116]:

1. *homozyg-snp 50*: Minimum number of SNPs in a ROH
2. *homozyg-window-snp 50*: Minimum density (1 SNP in 50Kb) to consider a segment as a ROH
3. *homozyg-kb 300*: Minimum length of a ROH
4. *homozyg-window-missing 5*: Maximum number of missing genotypes in a ROH
5. *homozyg-window-het 3*: Maximum number of heterozygous genotypes in a ROH (to tolerate possible errors during variant calling or genotyping)
6. *homozyg-window-threshold 0.05*: Proportion of overlapping windows during scanning of homozygous SNPs (to overcome the probability of a window being homozygous by chance)

ROH length categories were determined based on previously published ranges: short ROH (<0.515 Mb), medium-length ROH (0.516-1.606 Mb), and long ROH (>1.607 Mb) [17, 117]. For each individual, the sum of ROHs (SROH) in all length categories was calculated. The statistical significance between sROHs of TR and global populations and between TR subregions were determined using the Kruskal-Wallis test after testing normal distribution and homogeneity of variances using Kolmogorov-Smirnov test and Levene's test, respectively. Pairwise comparisons were done using the Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing adjustment. For the frequency calculations, ROHs up to 4 Mb in length were binned together with a bin size of 100Kb while ROHs $\geq$ 4 Mb were binned together and plotted using a histogram.

The autosomal genome in ROH was determined using the formula [118]:

$$F_{roh} = \frac{\sum L_{roh}}{L_{auto}}$$

where $\sum$ $L_{roh}$ is the sum of an individuals ROHs above a threshold while $L_{auto}$ is the total length of the autosomal genome except centromeres. According to GRCh37 assembly and the length of the genome of TR-WGS samples covered at 8X, Lauto was determined as 2,643,316 Kb. $L_{roh}$ was calculated using a) all ROH classes, b) long ROHs.

## 2.7 Y-chromosome and mitochondrial DNA haplogroups

Y-chromosome haplogroups have been investigated in many genetic studies of population history because variations in the non-recombining portion of Y-chromosome can be tracked in many generations and directly indicates patrilineal heritage [119]. The haplogroups are assigned based on SNP and short tandem repeat markers. Similarly, variations in the specific regions of mitochondrial DNA (mtDNA) have been used to determine the matrilineal origins of individuals due to lack of recombination [120]. The distributions of Y-chromosome and mtDNA haplogroups in human populations are geographically stratified. The current study

contains WGS data of 773 individuals of which 439 were males, providing an opportunity to infer the distributions of Y-chromosome and mtDNA haplogroups in the TR population.

To analyze Y-chromosome haplotypes of individuals from the global dataset, high-quality variants in the 10.3 Mb of the Y-chromosome were extracted [27]. Y-chromosome haplotypes were determined by Y-Lineage Tracker, which uses the International Society of Genetic Genealogy (ISOGG) Y-DNA tree (2019 version) as a reference [119]. The proportions of C-RPS4Y and O3-M122 sublineages, which were previously reported as Central Asian specific, were calculated to investigate the level of Central Asian contribution to the modern-day TR population. To identify mtDNA haplogroups of TR individuals, mtDNA variants were aligned to the Revised Cambridge Reference Sequence (rCRS) of the human mtDNA (NC_ 012920). rCRS-aligned mtDNA variants of individuals from 1000GP and the Human Genome Diversity Project (HGDP) were also downloaded [121]. Haplogrep2 (v2.2, which employs Phylotree v17 as a reference, mtDNA haplogroups were identified [120, 122]. D4c and G2a mtDNA haplotypes were previously proposed as Central Asian specific [123]. The frequencies of these mtDNA haplotypes were calculated in the TR samples to evaluate the proportion of matrilineal contribution from Central Asia.

## 2.8 Variome characterization

### 2.8.1 Derived allele frequencies

An ancestral allele is defined as the allele that persists in its ancestral state. Ancestral alleles are carried by the last common ancestor of the taxon. A derived allele, on the other hand, is the 'younger' allele that is diverged from its initial state through the evolutionary process [124]. Derived alleles have been used to understand LD, selective pressures, and differences in the frequency of disease-causing variants among populations. Furthermore, disease-associated alleles were shown to be more likely to be derived alleles with low-frequency [125]. Therefore, derived allele frequencies (DAFs) were investigated in the TR dataset and compared to that of gnomAD and GME Variome. First, ancestral sequences of

Ensembl compara for Homo sapiens, which involve multiple sequence alignment of six primates and mapped on the GRCh37, were obtained from the 1000GP FTP site. Then, Jvarkit, vcfancestralalleles tool was employed to annotate the variants from TR-WES, TR-WGS, gnomAD, and GME datasets with the ancestral alleles [126]. DAFs were not estimated for the sites where an ancestral allele is not present. The relationships between TR, gnomAD, and GME DAFs were assessed using Pearson Product Moment Correlation.

### 2.8.2 Functional annotation

The functional impact of genetic variants was obtained using Ensembl v.87 annotations by employing SnpEff v.4.4 [127, 128]. Standardized terminology of effects of sequence changes provided by Sequence Ontology(SO) and impacts of variants provided by SnpEff is listed in Table 2.5 [129].

The effect of predicted LoF variants (pLoFs: frameshift, essential splice site, stop gain, stop loss, and start loss) was further investigated using LOFTEE. LOFTEE is a Variant Effect Predictor plugin to identify high-confidence pLoFs (HC-pLoFs) using the information of ancestral state, splice prediction, and transcript information. LOFTEE flags the low-confidence pLoF (LC-pLoF) variants: variants in the ancestral state across primates; frameshift and stop gain variants located in the last 5% of the transcript or in an exon surrounded by noncanonical splice sites, and splice site variants in introns shorter than 15 bp or in an intron with a noncanonical splice site or rescued by nearby in-frame splice sites [130]. Moreover, a transcript expression-aware annotation for single nucleotide pLoF variants was performed using proportion expression across transcripts (pext) values. Pext is the proportion of the total transcriptional output from a gene that would be affected by the variant and can be utilized as an indicator of the functional impact of pLoF variants [52]. Additionally, pLOF variants were annotated with gnomAD_ pLI scores, which indicates the probability of a gene to be tolerant to LoF variants using the expectation-maximization algorithm. [29].

Missense variants were also further categorized based on their predicted deleteriousness using *in silico* prediction tools, namely PolyPhen-2, SIFT, and Combined Annotation Dependent Depletion (CADD). PolyPhen-2 evaluates the effect

**Table 2.5** SnpEff annotation

| Effect (SO) | Impact | Effect (SO) | Impact |
|---|---|---|---|
| bidirectional gene fusion | HIGH | 3 prime UTR truncation + exon loss variant | MODERATE |
| chromosome | HIGH | 5 prime UTR truncation + exon loss variant | MODERATE |
| duplication | HIGH | disruptive inframe deletion | MODERATE |
| exon loss variant | HIGH | disruptive inframe insertion | MODERATE |
| feature ablation | HIGH | inframe deletion | MODERATE |
| frameshift variant | HIGH | inframe insertion | MODERATE |
| gene fusion | HIGH | missense variant | MODERATE |
| inversion | HIGH | sequence feature + exon loss variant | MODERATE |
| protein protein contact | HIGH | 3 prime UTR variant | MODIFIER |
| rare amino acid variant | HIGH | 5 prime UTR variant | MODIFIER |
| rearranged at DNA level | HIGH | coding sequence variant | MODIFIER |
| splice acceptor variant | HIGH | conserved intergenic variant | MODIFIER |
| splice donor variant | HIGH | conserved intron variant | MODIFIER |
| start lost | HIGH | downstream gene variant | MODIFIER |
| stop gained | HIGH | exon variant | MODIFIER |
| stop lost | HIGH | gene variant | MODIFIER |
| structural interaction variant | HIGH | intergenic region | MODIFIER |
| 5 prime UTR premature start codon gain variant | LOW | intragenic variant | MODIFIER |
| initiator codon variant | LOW | intron variant | MODIFIER |
| splice region variant | LOW | miRNA | MODIFIER |
| start retained | LOW | regulatory region variant | MODIFIER |
| stop retained variant | LOW | transcript variant | MODIFIER |
| synonymous variant | LOW | upstream gene variant | MODIFIER |

of amino acid substitutions that result in amino acid alterations based on multiple sequence alignments, and their effects on the 3D-protein structure and classifies them as B (benign), P (possibly damaging), or D (probably damaging) [131]. SIFT analyzes missense variants according to sequence homology and physical properties of amino acids and classifies them as T (Tolerated) or D (deleterious) [132]. CADD combines mutation rates of local genomic regions, functional genomic data, transcript-level effects, multiple conservation scores, and protein-level deleteriousness and scores variants by using a machine-learning algorithm. These scores are interpreted as ranks of variants according to their damaging effect. For example, variants at the top 10% according to deleteriousness have a CADD score higher than 10, while variants at the top 1% have a CADD score

higher than 20 [133]. Using these three prediction tools in combination, missense variants were categorized as deleterious if they were annotated as D in both PolyPhen-2 and SIFT and had a CADD score >20. Under other instances, they were classified as "other missense".

Variants were annotated by the ANNOVAR v.2019Oct24 using the data from dbnsfp35a, which includes annotations of PolyPhen-2, SIFT, CADD, and GERP++ scores [134, 135]. AFs from diverse populations were retrieved from gnomAD, 1000GP, ESP, and GME databases and also annotated using ANNOVAR [17, 27, 28, 30]. "High-quality WES ($n$ = 1,123,248) and WGS (n = 45,981,720) variants were annotated separately and then combined. AFs of 365,489 shared variants of WES and WGS datasets were re-calculated. Variants were also classified according to their presence in other variant databases as 'novel' if they had no record in dbSNP build 151, gnomAD, 1000GP, and ESP; as 'rare' if they had an AF lower than 1% in all these databases, and 'common' in rest of the outcomes.

### 2.8.3 Homozygous pLoF mutations

The homozygous LoF mutations (or gene knockouts) contribute to the understanding of gene function and genotype-phenotype relationship [136]. Sequencing of individuals from inbred populations enables researchers to identify naturally occurring human gene knockouts [17, 57, 136]. There were 1,294 HC-pLoF variants that are found in homozygous state in TR individuals and AF of 723 HC-pLoFs were less than 1% in the TR population. Homozygous HC-pLoFs variants that were found in phase with another variant and complicated the functional interpretation were removed from the list. To compare TR homozygous HC-pLoFs with the previously published lists of human knockouts, the lists of rare human-knockouts provided by Iceland, GME, PROMIS, British Pakistani, and GenomeAsia studies were downloaded and homozygous pLoFs of gnomAD and 1000GP were extracted [17,27,30,34,57,136,137]. Furthermore, common homozygous pLoFs (frequency in the TR population $\geq$ 0.01) were identified and compared with the previously published list of common knockouts in the ExAC and gnomAD [138]. Three homozygous HC-pLoFs (p.Ser177fs in *PSG4*, p.Leu251fs in *FAM166A*, c.1259-1G>C in *ACOT9*) were validated using the 38 technical

replicates in the TR Variome.

### 2.8.4 Clinically relevant variants

To evaluate the clinical significance of the TR variants, variants were first investigated whether they were located in the OMIM genes and the patterns of inheritance of their associated phenotypes were assessed, where applicable (Accessed December 10th, 2019) [49]. Then, all TR variants were annotated using Human Gene Mutation Database (HGMD) Professional v.2020.2 and ClinVar (Accessed September 9th, 2020) [50,51]. Variants in the classes of disease-causing pathological mutations (DMs) in HGMD or pathogenic (P) and pathogenic/likely pathogenic (P/LP) in ClinVar were extracted.

## 2.9 Per-genome variant summary and Imputation panel

### 2.9.1 Per-genome variant summary

The level of genome-wide variation of the TR individuals was compared with that of 1000GP populations. First, high-quality variants of the 773 WGS samples were cataloged. Then, the number of variant sites and singletons in the TR WGS samples and 2,504 samples from 1000GP were calculated using BCFtools. To validate singletons detected in the TR population, the concordance rate between the 6 technical replicates of the WGS data was calculated and determined as $0.985 \pm 0.003$.

### 2.9.2 Imputation panel

To generate a TR reference panel for genotype imputation for future GWAS, WGS data of 773 TR samples were utilized. BEAGLE v5.1 was employed to generate autosomal haplotypes [139]. To obtain more accurate haplotypes for the TR

reference panel, the BEAGLE-phased genotypes were re-phased using SHAPEIT v2 [140]. SHAPEIT was run with the default parameters except for a window size of 0.5, which allows better phasing accuracy for the sequencing data. The predictive accuracy of the TR reference panel was assessed by the following steps: First, 73 TR samples were randomly selected and their haplotypes were removed from the TR reference panel. Then, genotypes of these 73 samples were obtained from the unphased WGS data. Chromosome 20 variants of 73 individuals were used to generate a pseudo-GWAS panel, which was composed of 44,367 SNPs from the Infinium Omni2.5-8 kit for array sequencing. The rest of the genotypes were 'masked'. Afterward, the phased haplotypes of 1000GP were downloaded from IMPUTE2 website [141]. Three different imputations were conducted on chromosome 20 of 73 TR samples using IMPUTE 2. Chromosome 20 haplotypes that was split into 5 Mb chunks and buffer regions with 250 Kb in length were used during imputation. The first imputation was performed using the 1000GP reference panel, the second was performed using the TR reference panel, and the last one was performed by employing both TR and 1000GP reference panels to fill the masked genotypes. The default parameters of IMPUTE2 were used except for setting k_ hap, (the number of reference haplotypes used as templates) to 10,000. Increasing the number of reference haplotypes was recommended by the developers of IMPUTE2 because diverse reference haplotypes could increase imputation accuracy.

To evaluate the accuracy of three different imputation methods, the correlation between the masked sequence genotypes (0,1,2) and the imputed genotype dosages (0-2)were assessed using squared Pearsons correlation coefficients (R2). The R2 results were stratified into non-overlapping AF bins. Wilcoxon rank-sum test was employed to test statistical significance between the results. Additionally, IMPUTE2 produces a summary output, which contains the number of imputed variants with corresponding expected R2 results and expected AF bins. The summary file was used to compare the expected number of accurately imputed variants in each imputation method.

The implementation of the TR reference panel for imputation of samples from neighboring populations was tested using the phased WGS data of the Simons Genome Diversity Project (SGDP) [142]. The phased WGS data of SGDP were downloaded and chromosome 20 variants of the Balkan (BLK), Caucasus (CAU),

and GME populations were selected to generate a pseudo-GWAS panel. Three imputations were performed using 1000GP, TR, and TR+1000GP reference panels, as was done for the TR samples, and the aforementioned steps were followed for the evaluation of imputation accuracy.

## 2.10 Molecular findings for Mendelian and complex traits

Generation of the TR Variome provided an opportunity to study the current status of Mendelian and complex traits in the TR population. By including the 330 new WES samples, a total of 4,194 TR individuals who either yielded WES ($n$ = 3,402) or WGS ($n$ = 792) data were involved in this section of the study. The demographics of the new cohort were shown in Table 2.6. After the same sample-, genotype-, and variant-based QC steps defined in the previous sections and subsequent removal of second-degree relatives, 1,369,087 WES and 47,420,907 WGS variants were obtained using 2,589 WES and 773 WGS samples.

### 2.10.1 Variant classification and pathogenicity assesment

First, TR variants located in the OMIM genes and associated with a phenotype were selected (Accessed December 10th, 2019) [49]. Three different variant datasets were generated using previously reported pathogenic variants (RP) in HGMD or ClinVar and the variants that were predicted to be pathogenic (PP) using *in silico* prediction tools. Dataset 1 consisted of RP variants that were classified as disease-causing mutation (DM) or probable/possible pathogenic mutation (DM?) in HGMD and P or P/LP in ClinVar. Dataset 2 covers the RP variants that were classified as DM or DM? in HGMD or P or P/LP in ClinVar. Variants that received contradicting interpretations (e.g., pathogenic in one database and benign in the other database) were eliminated from the datasets. Also, the ClinVar variants with conflicting interpretations of pathogenicity that had at least one submission as likely benign or benign were removed. To capture additional potential pathogenic variants, HC-pLoFs predicted to be deleterious

by LOFTEE, and deleterious missense variants predicted to be deleterious according to PolyPhen-2, SIFT, and CADD using the aforementioned cut-offs were selected as PP variants and included in Dataset 3 together with Dataset 2 variants [130–133]. For the RP variants, if HGMD or ClinVar phenotype of a variant differed from the one listed in OMIM, the variant was based on HGMD or ClinVar phenotype. The categorization of the PP variants was solely on OMIM phenotypes.

**Table 2.6** Demographics for assessing the molecular findings for Mendelian and complex traits

| Cohort | $n$ | Method | Age | Gender* | Affected | Unaffected | Unknown | Exclusion** |
|---|---|---|---|---|---|---|---|---|
| Amyotrophic lateral sclerosis | 136 | WES | 49.79 ± 17.45 | 75/61 | 95 | 41 | 0 | Neuromuscular, Neurological, Congenital |
| Ataxia | 148 | WES | 44.3 ± 14.82 | 75/73 | 101 | 47 | 0 | Neuromuscular, Neurological, Congenital |
| Delayed sleep phase disorder | 18 | WES | 28.38 ± 6.99 | 8/10 | 18 | 0 | 0 | Psychiatric |
| Essential tremors | 79 | WES | 52.38 ± 19.88 | 35/44 | 57 | 22 | 0 | Neurological |
| Obesity | 806 | WES | 38.48 ± 12.57 | 221/585 | 693 | 113 | 0 | Obesity |
| Polycystic ovarian syndrome | 111 | WES | 29 ± 7.83 | 2/109 | 109 | 2 | 0 | - |
| Various congenital and neurological disorders | 1,245 | WES | 42.12 ± 17.88 | 691/554 | 32 | 26 | 1,187 | Neurological, Neuromuscular, Congenital |
| Various immune system disorders | 253 | WES | - | 134/119 | - | - | 253 | Immune system/Infection |
| Amyotrophic lateral sclerosis | 773 | WGS | 55.44 ± 14.21 | 439/334 | 603 | 170 | 0 | Neurological, Neuromuscular, Congenital |

*: Male/Female

**: Samples from the respective cohort were removed from the respective dataset(s).

The study cohort consisted of individuals with congenital/multi-system disorders, neurological, neuromuscular, psychiatric, immunological/infectious diseases, and obesity. First, the variants were divided into 23 disease groups based on their associated phenotypes in OMIM, HGMD, and ClinVar (Table 2.7). To prevent over-representation of disease-associated variants, the data of the study cohorts were removed from the datasets of the respective disease group; for example, samples of the neurological disease cohort were excluded from the neurological disease dataset. AFs were recalculated after the removal.

**Table 2.7** Disease groups

| Disease groups | Number of samples |
|---|---|
| Cardiovascular | 3,599 |
| Congenital/Multi-system disorders | 1,268 |
| Connective tissue | 3,599 |
| Dental | 3,599 |
| Dermatologic | 3,599 |
| Drug metabolism | 3,599 |
| Ear | 3,599 |
| Endocrine | 3,599 |
| Eye | 3,599 |
| Gastrointestinal | 3,599 |
| Hematologic | 3,599 |
| Immune system/Infection | 3,346 |
| Kidney/Urinary system | 3,599 |
| Metabolic | 3,599 |
| Neoplasm | 3,599 |
| Neurological | 1,189 |
| Neuromuscular | 1,268 |
| Obesity | 2,628 |
| Psychiatric | 3,580 |
| Pulmonary | 3,599 |
| Reproductive system | 3,599 |
| Rheumatic | 3,599 |
| Skeletal | 3,599 |

For the PP variants, an additional filtering step was performed by excluding variants with a MAF >1% in the TR cohort. Additionally, variants with a MAF >1% in gnomAD were removed, if associated with a recessive phenotype, and the variants with a MAF >0.1% in gnomAD were removed, if associated with a dominant phenotype. *ABCA4* variants previously reported as hypomorphic (variants that show their effect only when observed together with a severe variant): rs76157638, rs56357060, rs1800548, rs61753019, rs61748549, rs61748532,

rs146786552, rs61750563, rs61754056, and rs369973540 were removed from the datasets [72]. NBS genes were derived from Advisory Committee on Heritable Disorders in Newborns and Children (ACHDNC) website [70]. 13 additional genes associated with diseases that can also be screened in newborns were also included in the NBS gene list (Table 2.8). The ACMG SF v3.0 list of 73 actionable genes was used to report the secondary findings of TR individuals (Table 2.9) [71].

**Table 2.8** NBS genes included in the study

| NBS Gene | ACHDNC | RP and/or PP variant in the TR Variome |
|---|---|---|
| *ABCD1* | Yes | Yes |
| *ACADM* | Yes | Yes |
| *ACADS* | No | Yes |
| *ACADSB* | No | Yes |
| *ACADVL* | Yes | Yes |
| *ACAT1* | Yes | Yes |
| *ADA* | Yes | Yes |
| *ARG1* | No | Yes |
| *ASL* | Yes | Yes |
| *ASS1* | Yes | Yes |
| *BCKDHB* | Yes | Yes |
| *BTD* | Yes | Yes |
| *CBS* | Yes | Yes |
| *CFTR* | Yes | Yes |
| *CPT2* | No | Yes |
| *CYP21A2* | Yes | Yes |
| *DBT* | Yes | Yes |
| *DUOX2* | Yes | Yes |
| *ETFA* | No | Yes |
| *ETFDH* | No | Yes |
| *FAH* | Yes | Yes |
| *GAA* | Yes | Yes |
| *GALC* | No | Yes |
| *GALE* | Yes | Yes |

*(continued on next page)*

*Table 2.8 continued*

| | | |
|---|---|---|
| *GALK1* | Yes | Yes |
| *GALT* | Yes | Yes |
| *GBA* | No | Yes |
| *GCDH* | Yes | Yes |
| *GCH1* | No | Yes |
| *GJB2* | Yes | Yes |
| *GJB3* | Yes | Yes |
| *GJB6* | Yes | Yes |
| *HADHA* | Yes | Yes |
| *HBB* | Yes | Yes |
| *HLCS* | Yes | Yes |
| *HMGCL* | Yes | Yes |
| *HPD* | Yes | Yes |
| *IDUA* | Yes | Yes |
| *IVD* | Yes | Yes |
| *MAT1A* | No | Yes |
| *MCCC1* | Yes | Yes |
| *MCCC2* | Yes | Yes |
| *MCEE* | Yes | Yes |
| *MMAA* | Yes | Yes |
| *MMAB* | Yes | Yes |
| *MMACHC* | No | Yes |
| *MMADHC* | Yes | Yes |
| *MTHFR* | Yes | Yes |
| *MTR* | Yes | Yes |
| *MTRR* | Yes | Yes |
| *MMUT* | Yes | Yes |
| *NPC1* | No | Yes |
| *NPC2* | No | Yes |
| *PAH* | Yes | Yes |
| *PAX8* | Yes | Yes |
| *PCCA* | Yes | Yes |
| *PCCB* | Yes | Yes |
| *SLC22A5* | Yes | Yes |

*(continued on next page)*

*Table 2.8 continued*

| | | |
|---|---|---|
| *SLC25A13* | Yes | Yes |
| *SLC5A5* | Yes | Yes |
| *TAT* | Yes | Yes |
| *TG* | Yes | Yes |
| *TPO* | Yes | Yes |
| *TSHB* | Yes | Yes |
| *TSHR* | Yes | Yes |
| *BCKDHA* | Yes | No |
| *HADHB* | Yes | No |
| *SMN1* | Yes | No |
| *SMN2* | Yes | No |

ACHDNC: Advisory Committee on Heritable Disorders in Newborns and Children

**Table 2.9** ACMG recommended actionable genes included in the study

| ACMG recommended actionable gene | RP and/or PP variant in the TR Variome |
|---|---|
| *ACTC1* | Yes |
| *ACVRL1* | Yes |
| *APC* | Yes |
| *APOB* | Yes |
| *ATP7B* | Yes |
| *BMPR1A* | Yes |
| *BRCA1* | Yes |
| *BRCA2* | Yes |
| *BTD* | Yes |
| *CACNA1S* | Yes |
| *CASQ2* | Yes |
| *COL3A1* | Yes |
| *DSC2* | Yes |
| *DSG2* | Yes |
| *DSP* | Yes |

*(continued on next page)*

*Table 2.9 continued*

| | |
|---|---|
| *ENG* | Yes |
| *FBN1* | Yes |
| *FLNC* | Yes |
| *GAA* | Yes |
| *HFE* | Yes |
| *HNF1A* | Yes |
| *KCNH2* | Yes |
| *KCNQ1* | Yes |
| *LDLR* | Yes |
| *LMNA* | Yes |
| *MAX* | Yes |
| *MEN1* | Yes |
| *MLH1* | Yes |
| *MSH2* | Yes |
| *MSH6* | Yes |
| *MUTYH* | Yes |
| *MYBPC3* | Yes |
| *MYH11* | Yes |
| *MYH7* | Yes |
| *MYL2* | Yes |
| *MYL3* | Yes |
| *NF2* | Yes |
| *OTC* | Yes |
| *PALB2* | Yes |
| *PCSK9* | Yes |
| *PKP2* | Yes |
| *PMS2* | Yes |
| *PRKAG2* | Yes |
| *PTEN* | Yes |
| *RB1* | Yes |
| *RET* | Yes |
| *RPE65* | Yes |
| *RYR1* | Yes |
| *RYR2* | Yes |

*(continued on next page)*

*Table 2.9 continued*

| | |
|---|---|
| *SCN5A* | Yes |
| *SDHB* | Yes |
| *SDHC* | Yes |
| *SDHD* | Yes |
| *SMAD3* | Yes |
| *STK11* | Yes |
| *TGFBR1* | Yes |
| *TGFBR2* | Yes |
| *TMEM127* | Yes |
| *TMEM43* | Yes |
| *TNNI3* | Yes |
| *TNNT2* | Yes |
| *TP53* | Yes |
| *TPM1* | Yes |
| *TRDN* | Yes |
| *TSC1* | Yes |
| *TSC2* | Yes |
| *TTN* | Yes |
| *VHL* | Yes |
| *WT1* | Yes |
| *ACTA2* | No |
| *GLA* | No |
| *SDHAF2* | No |
| *SMAD4* | No |

### 2.10.2 Calculations of CF and genetic prevalence

Cumulative CF for each gene was calculated by dividing the number of heterozygotes to the total number of individuals in each disease group. Individuals with more than one heterozygous variant in a single gene were counted only once during calculation. The genetic prevalence (GP; the proportion of individuals in the population who are expected to be affected based on their genotype) was calculated for each gene associated with an autosomal recessive disorder by estimating

the probability of two individuals who have a heterozygous mutation in the same gene to have a homozygous offspring [72]:

$$GP = \frac{\sum_i i, i + \sum_{i,j} i, j}{4}$$

where $i$ and $j$ are CFs of different variants in a gene. The formula provided the sum of the probabilities of homozygous and compound heterozygous mutations. The expected frequency of affected individuals under HWE was also calculated using the aggregate CF for each gene:

$$p^2 + 2pq + q^2 = 1$$

where $p$ and $q$ are the frequencies of wild type and alternate alleles, respectively, and $q^2$ represents the frequency of affected individuals for an autosomal recessive disorder. The proportion of affected individuals for X-linked disorders was calculated by adding $q$ to $q^2$. Pearsons product-moment correlation and linear regression were used to assess the degree of relationship between reported CF and disease prevalence with estimated CF and GP.

# Chapter 3

# Results

## 3.1 Population structure of Turkey

1,123,248 variants from 2,589 WES and 45,981,721 variants from 773 WGS samples were obtained following QC steps and filtering according to familial relatedness. The plots demonstrating the QC measures after excluding low-quality samples were shown in Figure 3.1 and Figure 3.2. The mean target base coverage for the exons of CCDS build 15 of the WES samples was 70X with 95.32%, 93.81%, and 88.45% coverage at 8X, 10X and 20X or more, respectively. The mean depth of coverage for the WGS samples was 34X with 93.9%, 93.7% and 93.4% coverage at 8X, 10X and 20X or more, respectively. Exome regions of WGS samples were combined with WES samples for the TR dataset, which was used in the investigation of the population structure of Turkey. QC measures for the integration of the coding regions of WES and WGS data were calculated to eliminate the potential batch effects between the two datasets [31]. Results did not manifest a prominent batch effect (Table 3.1).

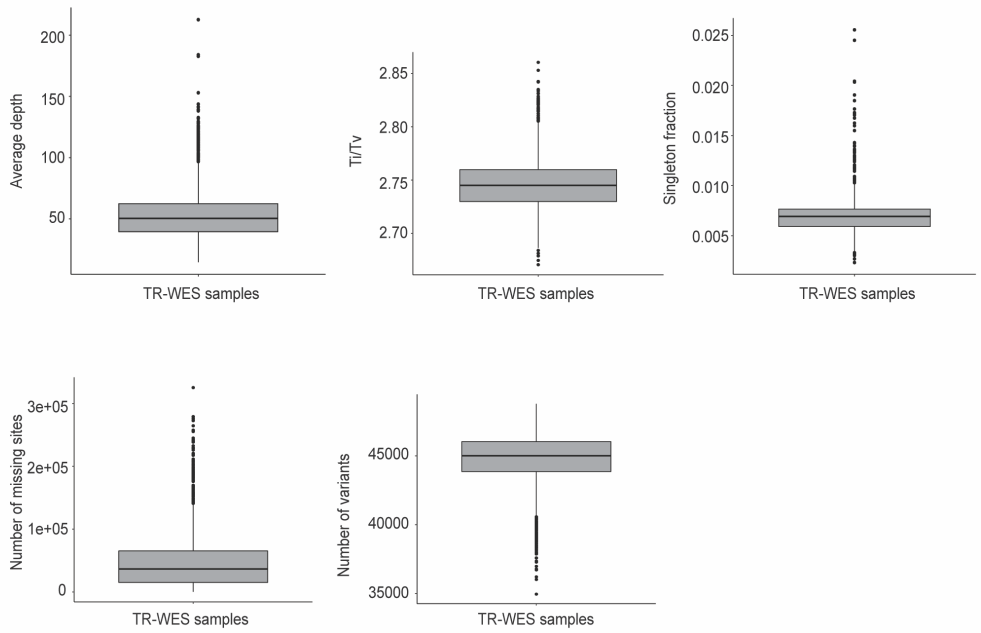



**Figure 3.1: QC metrics for the TR-WES samples.** Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).Adapted from [87].

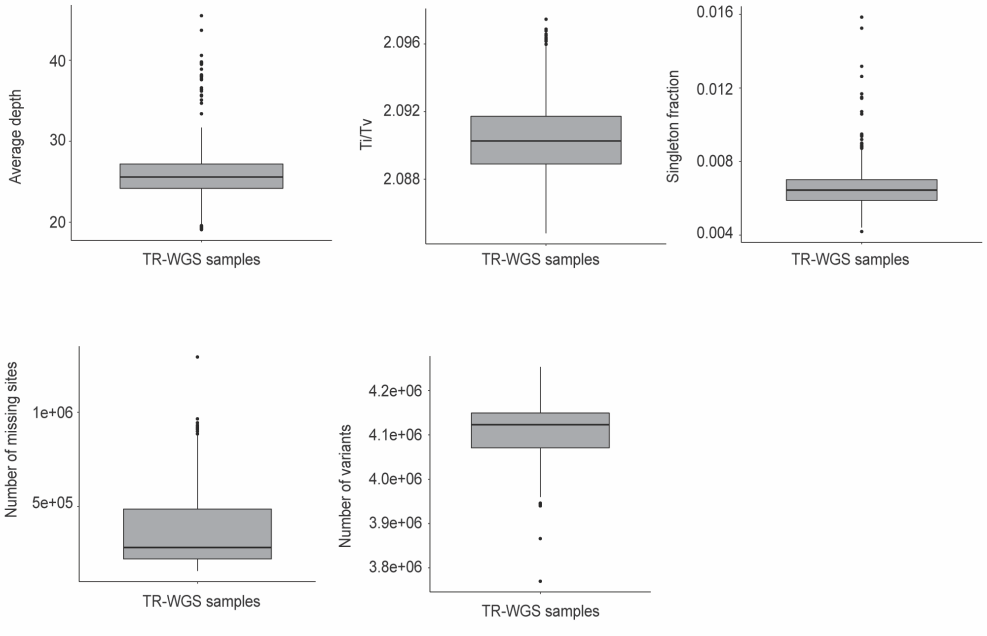



**Figure 3.2: QC metrics for the TR-WGS samples.** Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).Adapted from [87].

**Table 3.1** QC measures for the the integration of coding regions of WES and WGS data

| Sequencing type | Whole genome | | Whole exome | |
|---|---|---|---|---|
| Sample size | 773 | | 2,589 | |
| Minimum variant depth | 8 | | 8 | |
| Minimum allele count | 1 | | 1 | |
| Mean variant depth | 25 | | 54 | |
| Novel SNPs (% not in dbSNP 151) | 0.49% | | 0.37% | |
| Transition/transversion | 2.43 | | 2.55 | |
| | All | Novel | All | Novel |
| Number of variants | 67,644 | 334 | 42,483 | 161 |
| Heterozygotes | 42,077 | 330 | 25,738 | 164 |
| Variant homozygotes | 25,566 | 4 | 16,745 | 5 |

Reprinted from [87].

The birthplaces of maternal and paternal grandparents of 1,460 TR samples were obtained from records and grouped into six different subregions, namely TR-B, TR-C, TR-E,TR-N, TR-S, and TR-W. 123 TR samples (8.42%) have differed in terms of subregions that maternal and paternal origins belong. Two PCAs were performed using TR individuals of known grandparental origin according to geographical origins of maternal or paternal grandparent to test the effect of potential alterations in the genetic variability of the TR subregions (Table 3.2). No remarkable difference was observed between the groups formed using maternal or paternal grandparents (Figure 3.3). Thus, TR subregions based on maternal grandparental origins were used for the subsequent population structure, inbreeding, and ROH analyses.

**Table 3.2** Geographical origins of TR individuals

| Subregion | Group | Origin based on maternal grandparent | Origin based on paternal grandparent |
|---|---|---|---|
| Balkan | TR-B | 90 | 86 |
| West | TR-W | 157 | 160 |
| Central | TR-C | 441 | 433 |
| North | TR-N | 372 | 376 |
| South | TR-S | 116 | 121 |
| East | TR-E | 284 | 284 |
| Total | | 1,460 | 1,460 |

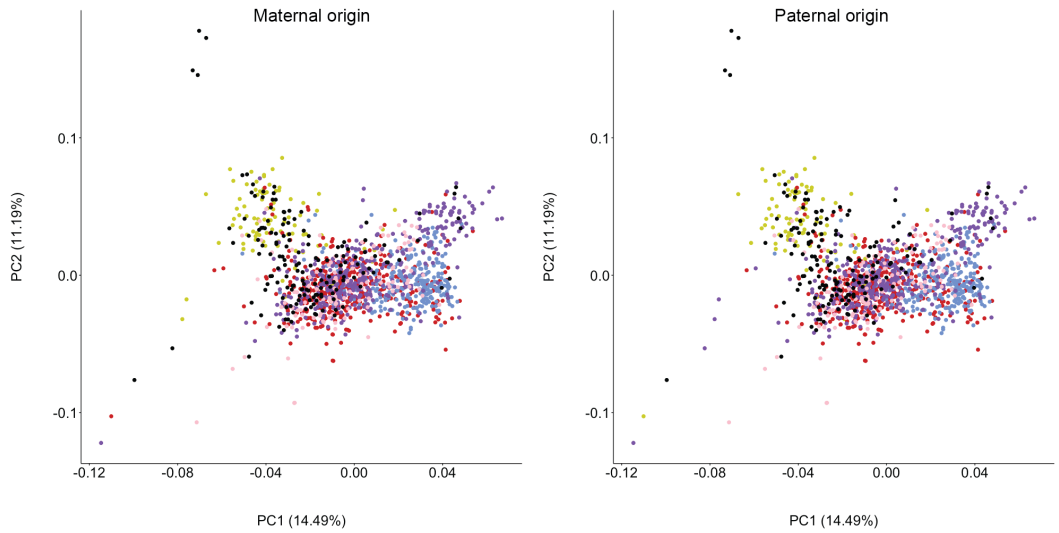



**Figure 3.3: PCA on TR individuals with known origin.** Plots for the PC1 and PC2, which explain 14.49% and 11.19% of the total variation observed in the TR population ($n$ =1,460). Each dot represents one individual. PCA plot on the left side was reprinted from [87].

PCA using only TR individuals of known origin revealed the 14.49% and 11.19% of the genetic variability in the first two PCs. The TR subregions did not form distinct clusters along PC axes; however, their distribution was similar to their geographical location in the east-west direction (Figure 3.3. Three Procrustes analyses were performed to further understand the effect of geography on the genetic variation of the TR population using: PC1 and PC2, PC2 and PC3, and PC3 and PC4. The highest correlation for Procrustes rotation was observed when the first two PCs were used. The results were consistent with the PCA; there was no clear-cut separation between the TR subregions,; however, a significant mild positive correlation was detected (Correlation in Procrustes rotation = 0.49, $P$ <1  $10^{-5}$, Figure 3.4).

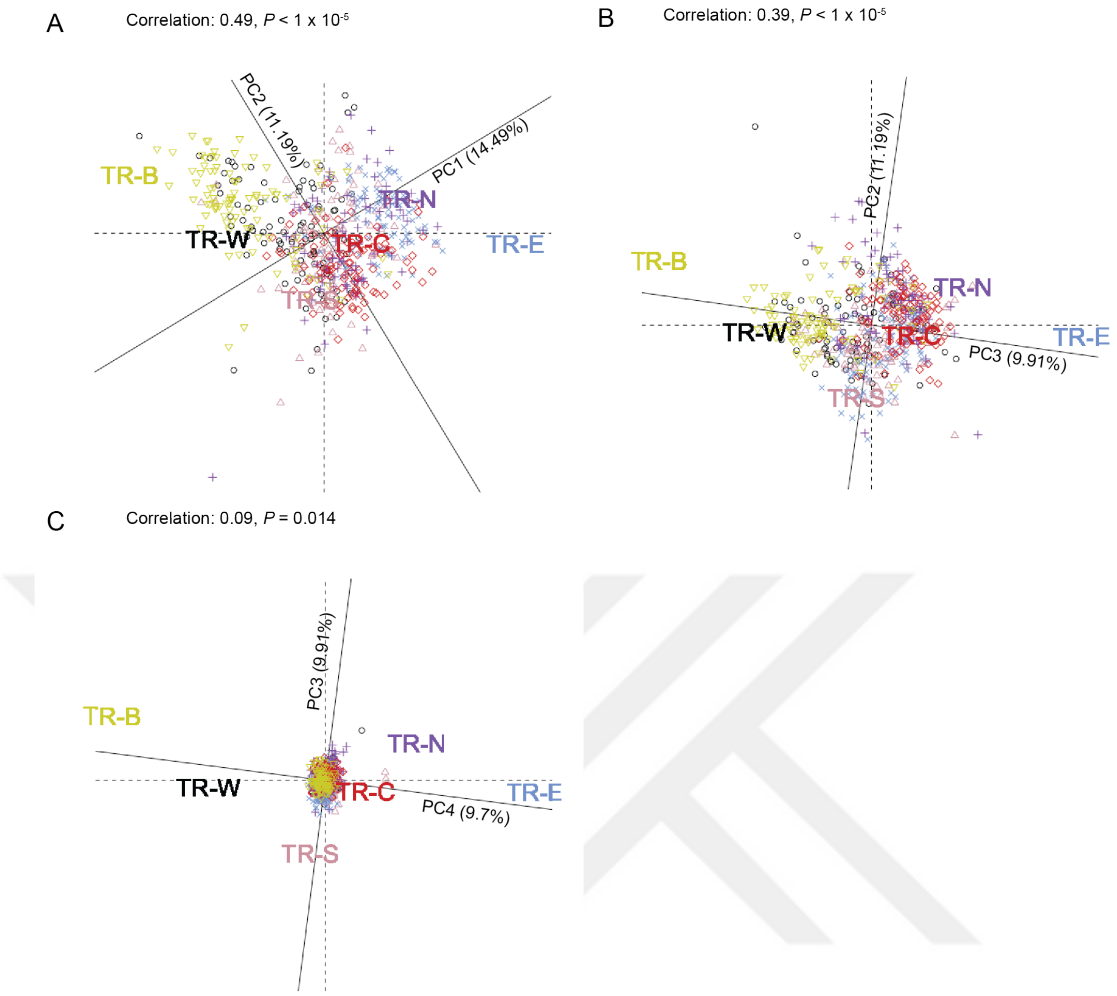



**Figure 3.4: Procrustes analysis on TR individuals with known origin.** Procrustes analysis using **A** PC1 and PC2; **B** PC2 and PC3; **C** PC3 and PC4, ($n$ = 1,460). Each dot represents one individual.

Population stratification in Turkey was also tested using two additional dimensionality reduction techniques, tSNE and UMAP, by employing PC1-PC10. Small clusters of samples from TR-B, TR-E, and TR-N were distinguished using both techniques; however, signs of a remarkable intermixing were observed between these subregions (Figure 3.5).

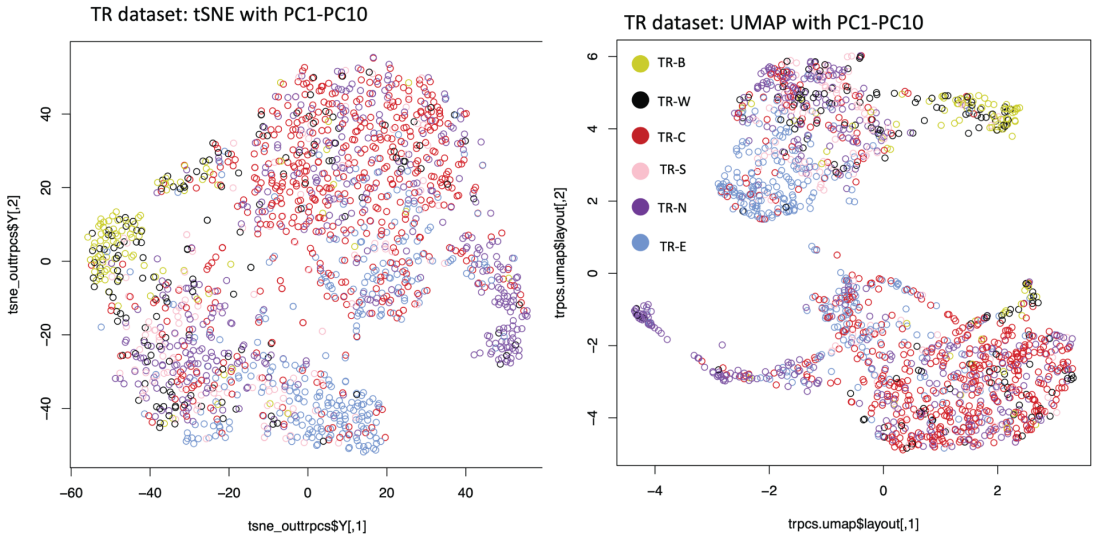



**Figure 3.5: Results of the tSNE and UMAP analyses using the TR dataset.** Evaluation of the population stratification using PC1-PC10 ($n$ = 1,460). Each dot represents one individual.

To assess the genetic association/differentiation of the TR population on a global scale, variants of individuals from diverse populations were merged with those of TR individuals (Table 2.4) [27,81]. Initially, a PCA was performed using the TR and eight super-populations: AFR, EUR, BLK, CAU, GME, SAS, EAS, and Central and North Asia (CNA). PC1 separated EAS and CNA populations whereas PC2 and PC3 distinguished AFR and SAS populations, respectively. The rest of the populations (GME, CAU, TR, BLK, and EUR) displayed an east to west cline in PC4 (Figure 3.6).

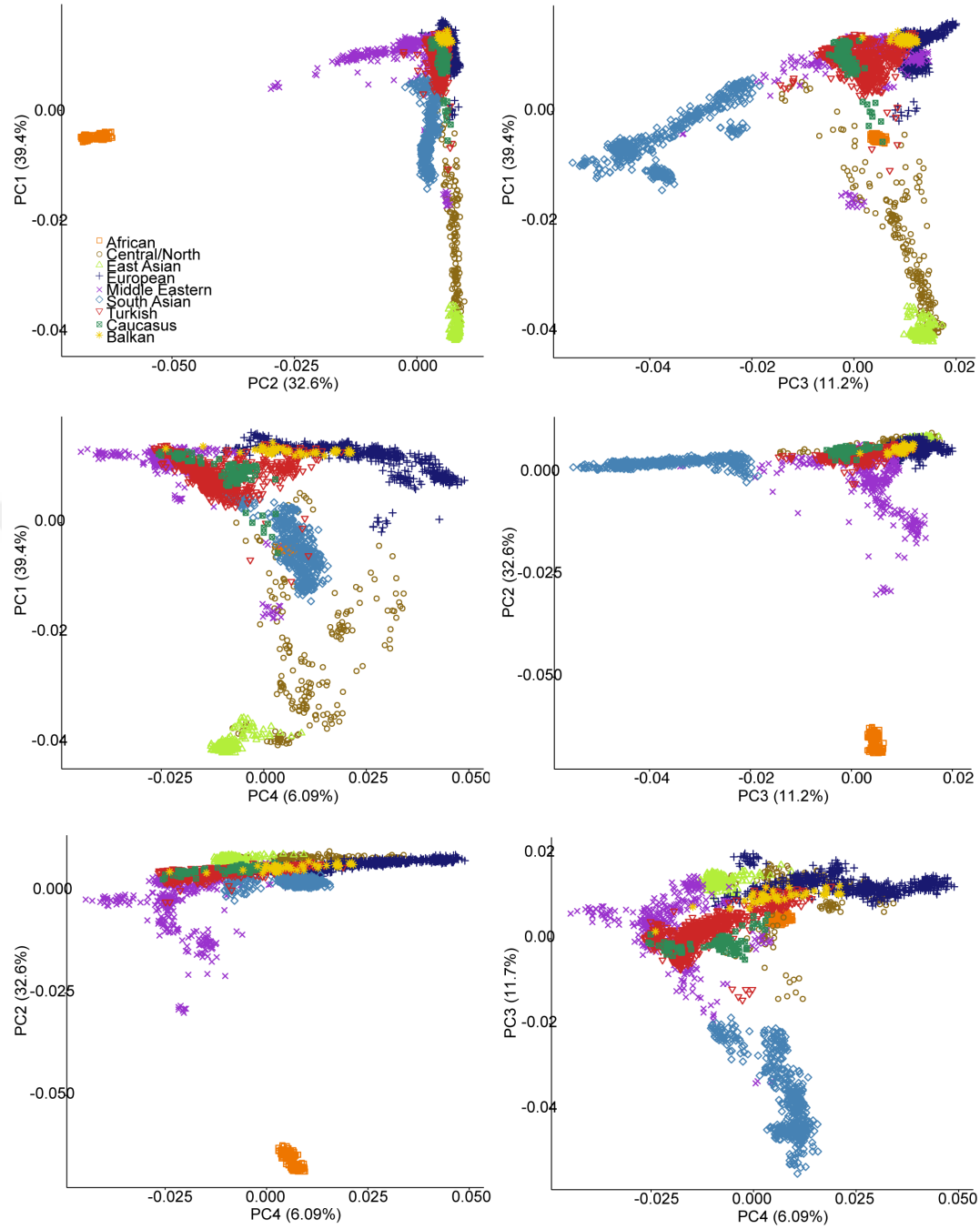



**Figure 3.6: PCA on the global dataset.** Plots show the first four PCs and the percentages of the variance explained. Samples in the figure are from Turkey ($n$ = 773), Lazaridis *et al.* ($n$ = 1,304) [81] and 1000GP ($n$ = 1,299) [27]. Each dot represents one individual. Reprinted from [87].

Using the first 10 PCs of the PCA on the global dataset, tSNE and UMAP were carried out. These analyses revealed more distinct clusters of populations and indicated a close genetic relationship between TR and BLK, CAU, GME,

and EUR populations (Figure 3.7).

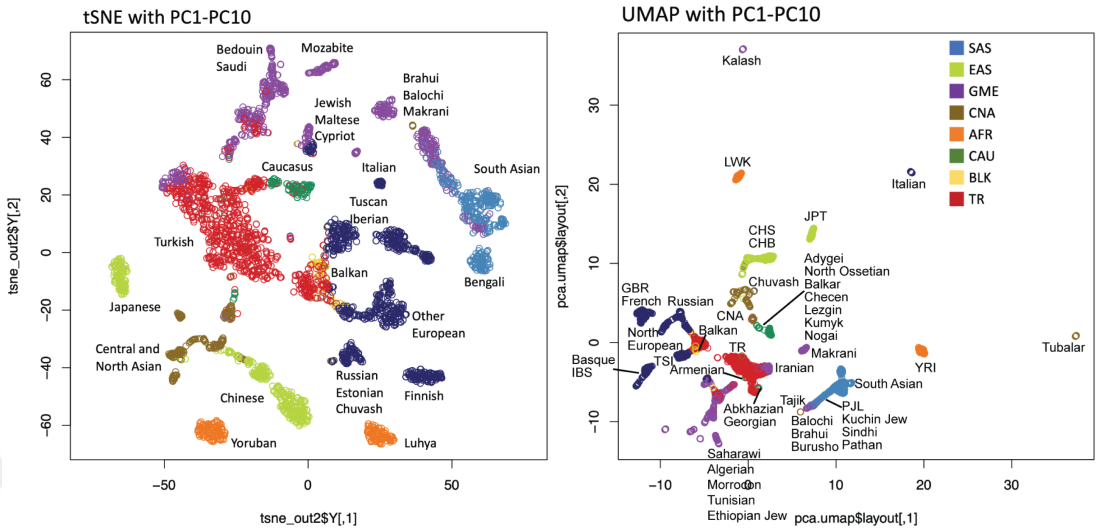



**Figure 3.7: Results of the tSNE and UMAP analyses using the global dataset.]** tSNE and UMAP were employed using first 10 PCs. Samples in the figure are from Turkey ($n$ = 773), Lazaridis *et al.* ($n$ = 1,304) [81] and 1000GP ($n$ = 1,299) [27]. Each dot represents one individual.

Population stratification in the TR subregions was then evaluated using ADMIXTURE [112]. ADMIXTURE was run using the number of ancestral contributions ($k$) from 2 to 8, where $k = 4$ resulted in the lowest cross-validation error (Figure 3.8). Each TR subregion exhibited all four ancestral components, albeit with different proportions. Similar to TR individuals assigned to a TR subregion, all four ancestral components were also present in individuals with unknown ancestral birthplaces (TR-U). (Figure 3.9).

A second ADMIXTURE was run using the global dataset to uncover the genetic substructure of the TR population by evaluating global ancestral contributions. Although the cross-validation errors were slightly lower in $k = 12$, $k = 8$ was a better representation for global ancestries (Figures 3.8 and 3.10). According to ADMIXTURE with $k = 8$, four major ancestral components were predominantly observed in EUR (navy and yellow), BLK(yellow), CAU (dark green), and GME (purple) populations; and these components formed the genetic substructure of TR individuals (Figure 3.10). Prominent effects of geography on the global ancestral contributions to the TR subregions were observed. Specifically, TR-B and TR-W subregions had higher proportions of the primary ancestral

components from the EUR and BLK populations. On the other hand, the shared ancestral component of the TR, CAU, and non-Arab populations of GME displayed a remarkable increase in the east direction for the TR subregions (Figure 3.11). Furthermore, the average contribution of the CNA population to the TR population was estimated as 9.59%. The TR subregions had different amounts of CNA contribution: TR-W, 12%; TR-S, 11.2%; TR-N, 10.6%; TR-C, 10.1%; TR-B, 7.69%; and TR-E, 6.48%.

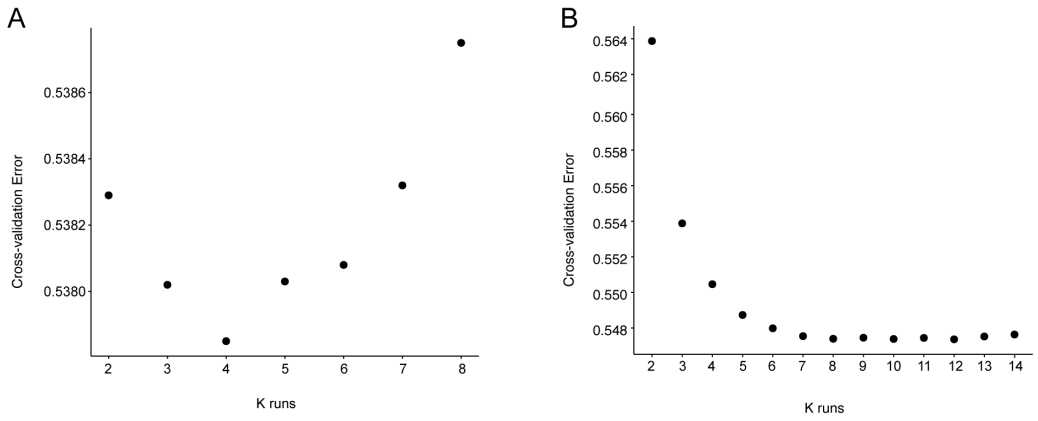



**Figure 3.8: ADMIXTURE cross-validation errors.** **A** Cross-validation errors for the TR subregions. ADMIXTURE analysis with $k = 4$ resulted with the lowest cross-validation error. $k = 4$ gave the lowest cross-validation error. **B** Cross-validation errors for the populations from the global dataset. The lowest cross-validation error was obtained when ($k = 12$) was used. Reprinted from [87].

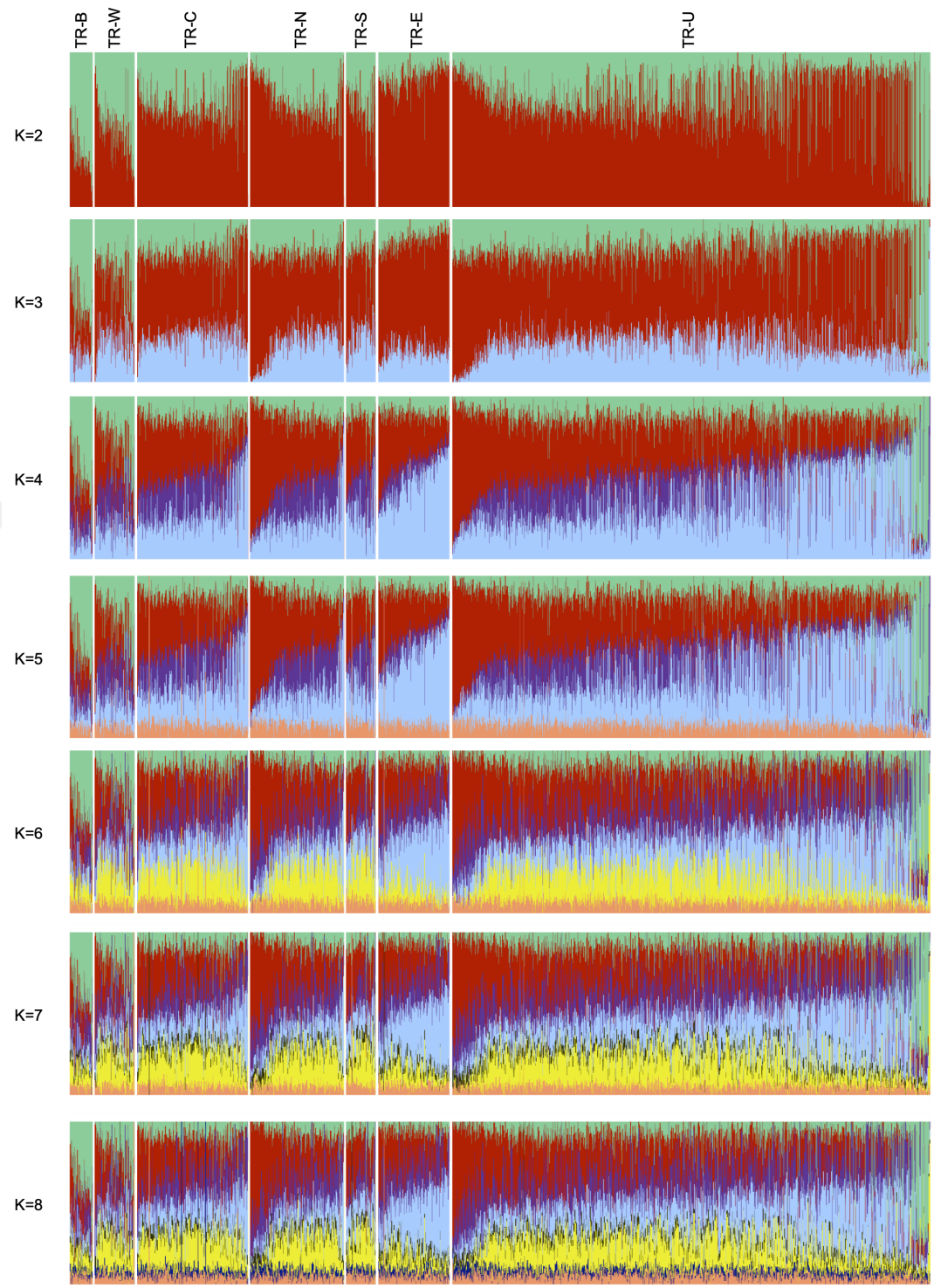



**Figure 3.9: ADMIXTURE analysis of the TR subregions for clusters $k = 2$ to $k = 8$.** The TR samples ($n$ = 3,362) were grouped according to TR subregions and organized from west (left) to east (right) based on the coordinates of the geographical regions. Each vertical line represents the proportions of ancestral components of a single individual. More color suggests more ancestral components.Reprinted from [87].

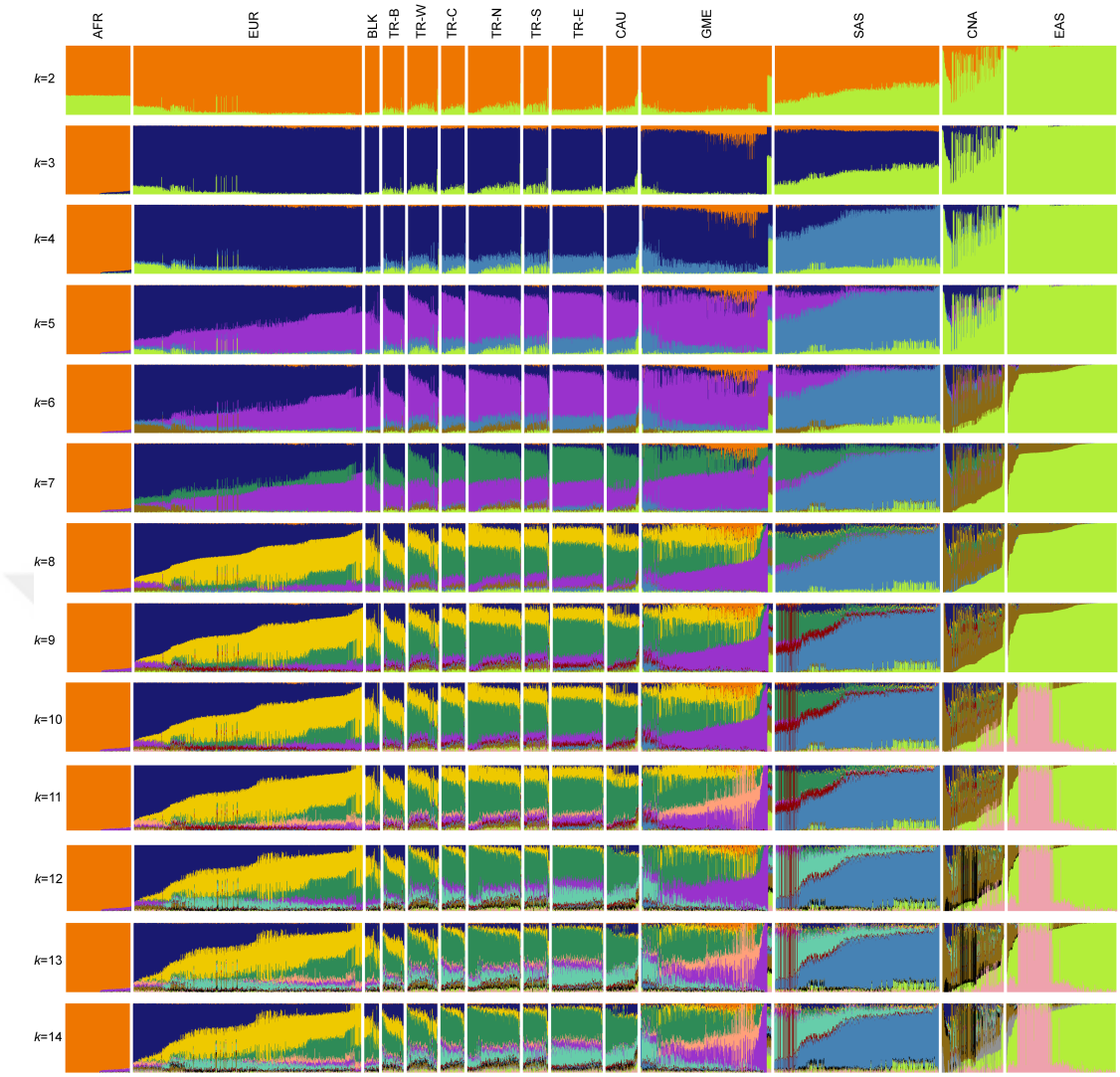



**Figure 3.10: ADMIXTURE analysis of the global dataset for clusters $k = 2$ to $k = 14$** Each vertical line represents the proportions of ancestral components of a single individual. Samples from Turkey ($n = 647$), Lazaridis *et al.* ($n = 1{,}304$) and 1000GP populations ($n = 1{,}299$) grouped by geographical region and organized from west (left) to east (right), reflecting the trends of overlap.Reprinted from [87].

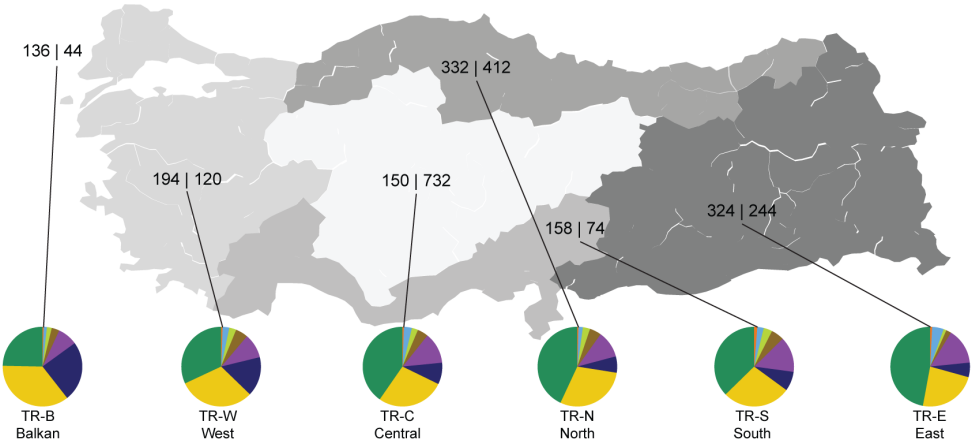



**Figure 3.11: The number of chromosomes originating from each TR subregion and global ancestral contributions to the TR subregions.** Map of TR showing the number of chromosomes (WGS/WES) and the mean admixture proportions of individuals with known birthplaces who originated from present-day TR and former Ottoman Empire territories. Reprinted from [87].

To further investigate the populations that have a close genetic connection with the TR population, a third PCA was performed using the regional dataset. Again, the location of the populations along PC axes reflected the importance of geography in shaping the genetic structure of populations. Also, the results indicated that the TR population has a very high level of genetic variability compared to other populations in the regional dataset. (Figure 3.12). As expected, EUR and TR populations were linked through TR-B and TR-W, while the genetic connections of TR population between CAU and GME populations were established by the other TR subregions (Figure 3.13).

The effect of geographical distances on the regional dataset was further tested using Procrustes analysis. The highest correlation in the Procrustes rotation was observed when PC1 and PC2 are used in the analysis, which revealed a strong positive correlation (Correlation in Procrustes rotation = 0.75, $P$ <1 x $10^{-5}$, Figure 3.14).

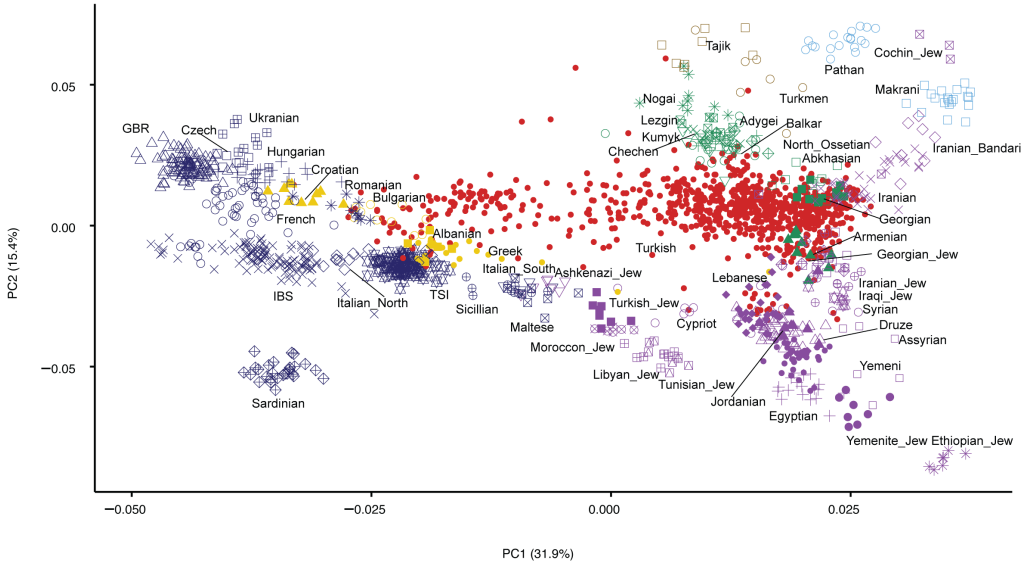



**Figure 3.12: PCA on the regional dataset.** The populations with the lowest pairwise Wrights FST (<0.01) values were included. Each dot represents a single individual, while colors indicate the superpopulations ($n$ = 1,805). Reprinted from [87].

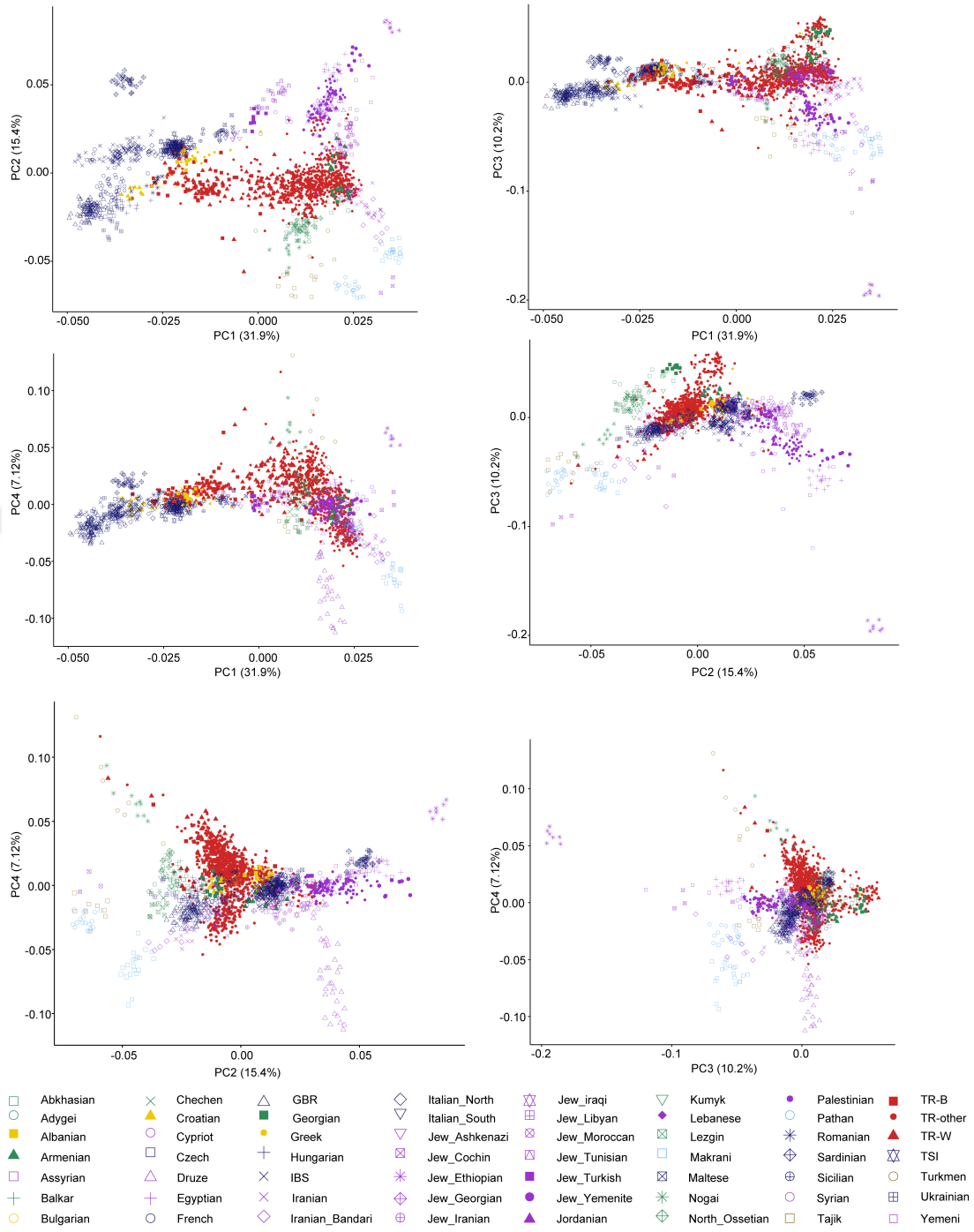



**Figure 3.13: PCA on TR subregions and control populations in a regional context.** Plots for the first four principal components and percentages of variance explained. Shapes represent individuals from different subpopulations, while colors represent the super populations ($n$ = 1,805). Reprinted from [87].

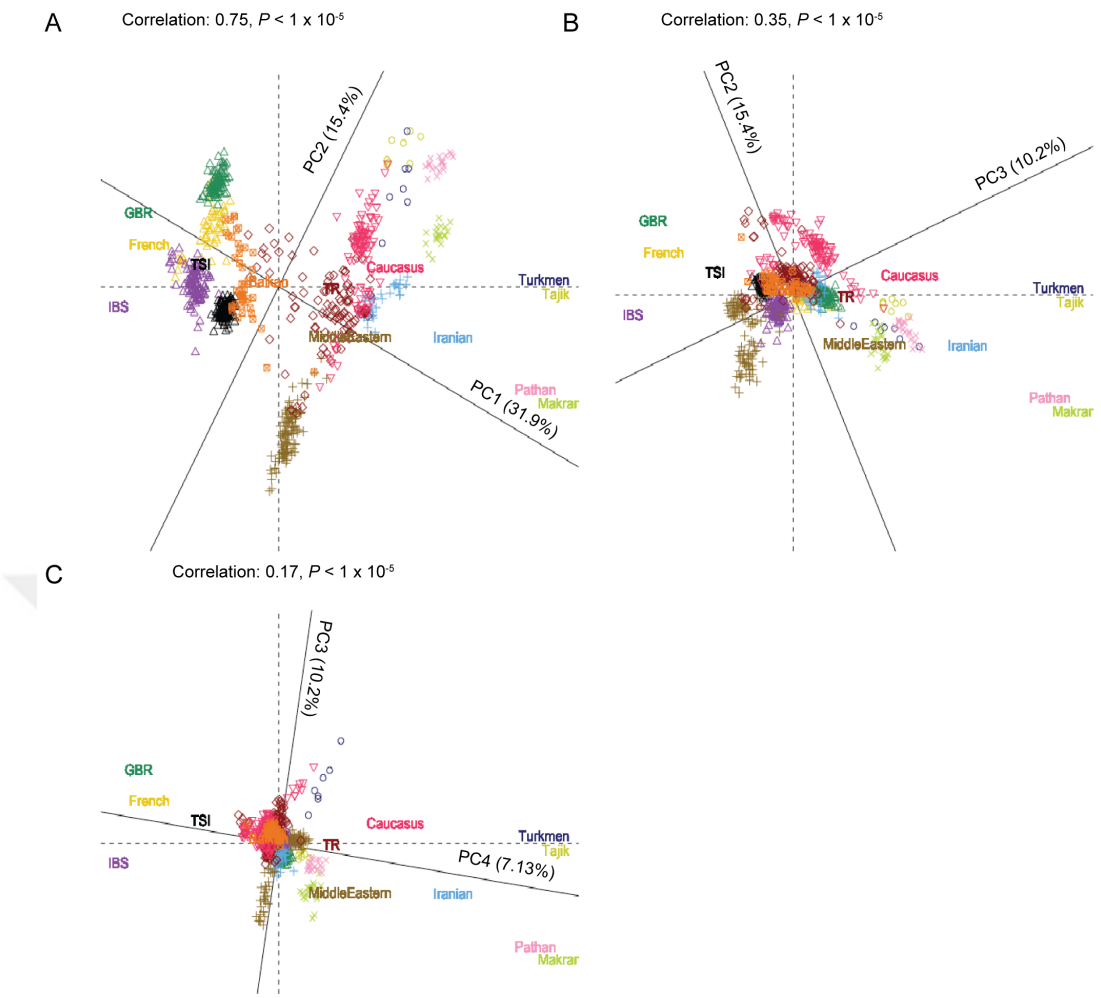



**Figure 3.14: Procrustes analysis on TR individuals and control populations.**Procrustes analysis using **A** PC1 and PC2; **B** PC2 and PC3; **C** PC3 and PC4. Individuals from TR were randomly selected and a Procrustes analysis was performed based on unprotected coordinates of geographical locations and PC1 and PC2 coordinates of TR,1000GP EUR, and populations from Lazaridis *et al.*

A maximum-likelihood phylogenetic tree was generated to investigate the effect of the internal migrations between the TR subregions on their genetic composition using Treemix [113]. The highest genetic drift was observed in the TR-B subregion. Moreover, gene flows from TR-E to TR-C and TR-S, and from TR-B to TR-W were detected (Figure 3.15). The position of Turkey along historical routes of human migration and the divergence patterns were also evaluated using Treemix, by including samples from the 1000GP and the GME populations (Figure 3.16). The TR population was located between EUR and GME branches. Also, the ordering of populations on the inferred tree was coherent with the "out-of-Africa" hypothesis [143].

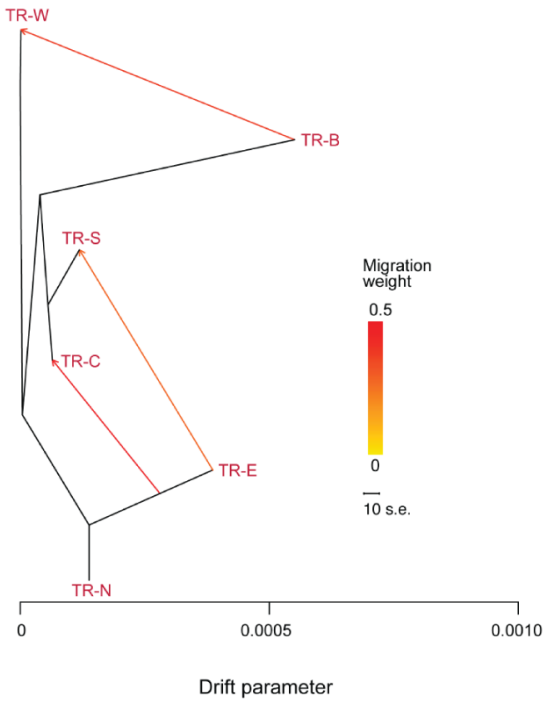



**Figure 3.15: Maximum-likelihood phylogenetic tree - Treemix in TR subregions.** Treemix phylogeny of the TR subregions. Three migration events were allowed during the analysis. The red color indicates higher degrees of gene flow. The lengths of branches are proportional to the extent of population drift.

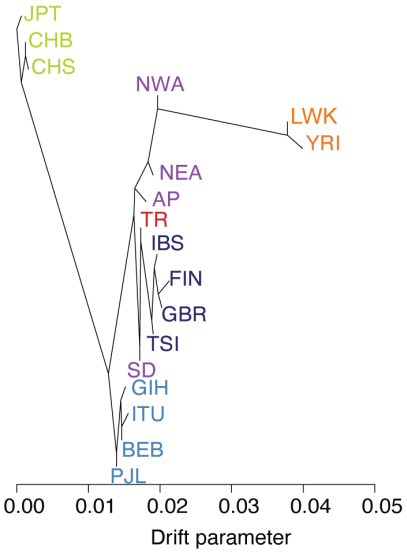



**Figure 3.16: Maximum-likelihood phylogenetic tree - Treemix.** Treemix phylogeny of the TR ($n$ = 3,362), the 1000GP ($n$ = 1,299), and the GME populations ($n$ = 696). The lengths of branches are proportional to the extent of population drift. Colors indicate superpopulations. Reprinted from [87].

The genetic associations of the populations in the regional dataset were further analyzed with $F_{ST}$. The TR population displayed the closest relationship with the neighboring populations in the West and East, followed by TSI. The results probably reflect the high levels of BLK, CAU, EUR, and GME admixture. When the pairwise $F_{ST}$ values of the TR subregions were evaluated, TR-B and TR-E had the most distant relationship ($F_{ST}$= 0.003) among the TR subregions, while they were closely related to their neighboring BLK and CAU-GME populations, respectively. Results emphasized the prominent effect of geographical distance on the genetic structure (Table 3.3).

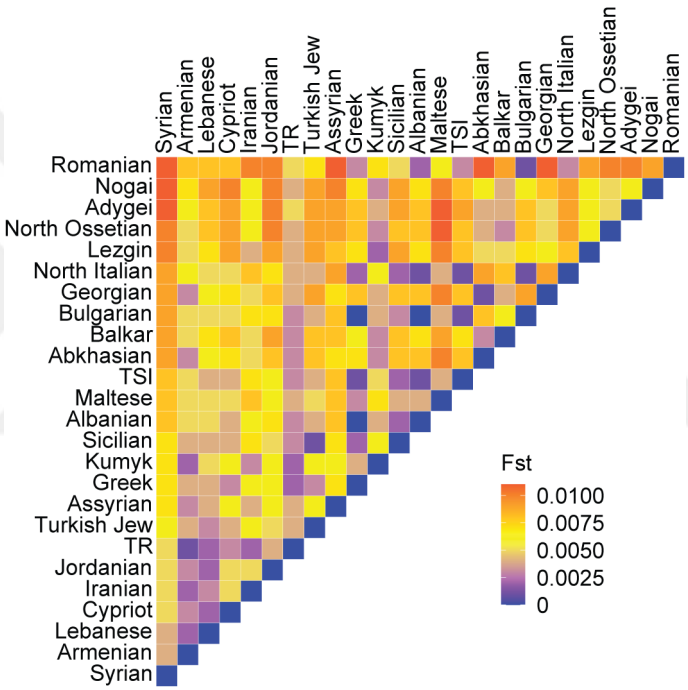



**Figure 3.17: Heatmap of pairwise $F_{ST}$ values in the regional dataset.**The blue color indicates a closer, while red shows a more distant genetic relationship. Reprinted from [87].

**Table 3.3** $F_{ST}$ for the TR subregions

| | TR-B | TR-W | TR-C | TR-N | TR-S | TR-E |
|---|---|---|---|---|---|---|
| TR-B | - | 0.001 | 0.002 | 0.002 | 0.002 | 0.003 |
| TR-W | 0.001 | - | 0.001 | 0.001 | 0 | 0.002 |
| TR-C | 0.002 | 0.001 | - | 0 | 0 | 0 |
| TR-N | 0.002 | 0.001 | 0 | - | 0 | 0.001 |

*(continued on next page)*

*Table 3.3 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| TR-S | 0.002 | 0 | 0 | 0 | - | 0.001 |
| TR-E | 0.003 | 0.002 | 0 | 0.001 | 0.001 | - |
| TR-U | 0.001 | 0 | 0 | 0 | 0 | 0.001 |
| Abkhasian | 0.006 | 0.004 | 0.003 | 0.003 | 0.004 | 0.004 |
| Adygei | 0.006 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 |
| Albanian | 0 | 0.002 | 0.003 | 0.003 | 0.003 | 0.005 |
| Armenian | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 |
| Assyrian | 0.007 | 0.005 | 0.004 | 0.004 | 0.004 | 0.003 |
| Balkar | 0.005 | 0.004 | 0.003 | 0.003 | 0.004 | 0.004 |
| Bulgarian | 0 | 0.002 | 0.003 | 0.004 | 0.003 | 0.005 |
| Chechen | 0.007 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 |
| Croatian | 0.003 | 0.005 | 0.007 | 0.008 | 0.007 | 0.009 |
| Cypriot | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.004 |
| Czech | 0.003 | 0.005 | 0.007 | 0.008 | 0.008 | 0.009 |
| Druze | 0.009 | 0.008 | 0.007 | 0.007 | 0.007 | 0.007 |
| Egyptian | 0.009 | 0.007 | 0.006 | 0.007 | 0.006 | 0.007 |
| French | 0.003 | 0.005 | 0.007 | 0.007 | 0.007 | 0.009 |
| GBR | 0.004 | 0.006 | 0.008 | 0.009 | 0.008 | 0.01 |
| Georgian | 0.007 | 0.005 | 0.004 | 0.003 | 0.004 | 0.004 |
| Greek | 0.001 | 0.002 | 0.003 | 0.003 | 0.003 | 0.004 |
| Hungarian | 0.002 | 0.004 | 0.006 | 0.007 | 0.006 | 0.008 |
| IBS | 0.003 | 0.004 | 0.006 | 0.007 | 0.006 | 0.008 |
| Iranian | 0.005 | 0.003 | 0.002 | 0.002 | 0.002 | 0.001 |
| Iranian Bandari | 0.01 | 0.007 | 0.006 | 0.007 | 0.006 | 0.006 |
| Italian North | 0.002 | 0.003 | 0.004 | 0.005 | 0.004 | 0.006 |
| Italian South | 0.014 | 0.014 | 0.015 | 0.015 | 0.015 | 0.016 |
| Jew Ashkenazi | 0.005 | 0.005 | 0.006 | 0.006 | 0.006 | 0.007 |
| Jew Cochin | 0.019 | 0.016 | 0.016 | 0.017 | 0.016 | 0.016 |
| Jew Ethiopian | 0.036 | 0.033 | 0.033 | 0.034 | 0.031 | 0.033 |
| Jew Georgian | 0.011 | 0.009 | 0.008 | 0.008 | 0.008 | 0.008 |
| Jew Iranian | 0.01 | 0.008 | 0.007 | 0.008 | 0.007 | 0.007 |
| Jew iraqi | 0.009 | 0.007 | 0.006 | 0.006 | 0.006 | 0.006 |
| Jew Libyan | 0.012 | 0.011 | 0.011 | 0.011 | 0.011 | 0.012 |
| Jew Moroccan | 0.007 | 0.006 | 0.007 | 0.007 | 0.006 | 0.007 |

*(continued on next page)*

*Table 3.3 continued*

| | | | | | | |
|---|---|---|---|---|---|---|
| Jew Tunisian | 0.012 | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| Jew Turkish | 0.004 | 0.004 | 0.004 | 0.004 | 0.003 | 0.005 |
| Jew Yemenite | 0.017 | 0.015 | 0.014 | 0.014 | 0.013 | 0.014 |
| Jordanian | 0.006 | 0.004 | 0.004 | 0.004 | 0.003 | 0.004 |
| Kumyk | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
| Lebanese | 0.004 | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 |
| Lezgin | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
| Makrani | 0.013 | 0.01 | 0.009 | 0.01 | 0.01 | 0.008 |
| Maltese | 0.004 | 0.004 | 0.005 | 0.005 | 0.004 | 0.006 |
| Nogai | 0.005 | 0.003 | 0.004 | 0.004 | 0.004 | 0.005 |
| North Ossetian | 0.006 | 0.005 | 0.004 | 0.004 | 0.005 | 0.005 |
| Palestinian | 0.008 | 0.007 | 0.006 | 0.006 | 0.005 | 0.006 |
| Pathan | 0.011 | 0.009 | 0.008 | 0.009 | 0.008 | 0.008 |
| Romanian | 0.002 | 0.004 | 0.006 | 0.006 | 0.006 | 0.008 |
| Sardinian | 0.01 | 0.011 | 0.013 | 0.013 | 0.012 | 0.015 |
| Sicilian | 0.003 | 0.003 | 0.003 | 0.004 | 0.003 | 0.005 |
| Syrian | 0.007 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 |
| Tajik | 0.009 | 0.008 | 0.008 | 0.009 | 0.008 | 0.008 |
| TSI | 0.001 | 0.002 | 0.003 | 0.004 | 0.003 | 0.005 |
| Turkmen | 0.009 | 0.007 | 0.008 | 0.009 | 0.008 | 0.009 |
| Ukrainian | 0.003 | 0.006 | 0.008 | 0.009 | 0.009 | 0.01 |
| Yemeni | 0.01 | 0.008 | 0.007 | 0.008 | 0.006 | 0.007 |

Reprinted from [87].

The effect of demographic events that shaped the genetic structure of the TR population was assessed using the rate of LD decay. LD decays faster in large populations with a high reproduction rate, whereas it decays slower if a population bottleneck occurs. AFR showed the highest rate for LD decay, while the other populations in the global dataset had closer rates. The rate of LD decay was slower in the TR population when compared to that of AFR, GME and SAS populations, while it was faster than the rest of the populations (Figure 3.18). Thus, results might be reflecting the occurrence of a shared ancient population bottleneck in the global dataset, except for the AFR population.

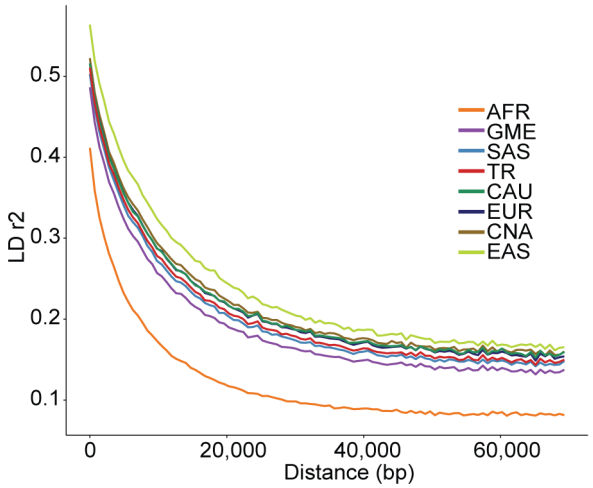



**Figure 3.18: The rate of LD decay in the global dataset.**Mean variant correlations (r2) are shown for each 700bp bin over 70,000 bp. Reprinted from [87].

## 3.2 Inbreeding status and estimation of ROH

Inbreeding coefficients of the populations in the global dataset were calculated and compared(Figure 3.19). The median and interquartile range (IQR) values of the global populations were shown in Table 3.4). The difference in the $F_{\text{plink}}$ of the global populations was statistically significant (Kruskal-Wallis test, H(8) = 576.76, $P$ <2.2 x 10-16). The TR population had higher $F_{\text{plink}}$ compared to that of other populations, except for CNA, GME, and SAS. $P$ values of pair-wise comparisons of $F_{\text{plink}}$ in global populations were calculated using Wilcoxon rank-sum test with Benjamini-Hochberg adjustment method (Table 3.5). Also, individuals with very high $F_{\text{plink}}$ values (up to 0.21) were detected in the TR population. These high values probably reflect recurrent consanguineous marriages in the family since the inbreeding coefficient is approximately half of the familial relationship between the parents. On the contrary, the large negative $F_{\text{plink}}$ values possibly demonstrate the offspring of pairs of unrelated but inbred individuals. Inbreeding coefficient $\geq$ 0.0156 is a cut-off for a kinship greater than that of a second cousin marriage [19]. 29.6% of the TR individuals had an $F_{\text{plink}}$ above this threshold, while the percentages for the other populations for the same threshold were for AFR, 0%; BLK, 0%; CAU, 1.98%; CNA, 51.5%; and EAS 4.91% EUR, 14.9%; GME, 53%; SAS, 41.1%; populations. The differences between TR subregions in terms of inbreeding coefficients were also evaluated, and found to

be statistically significant (Kruskal-Wallis test, H(5) = 11.33, $P = 0.045$) The medians for inbreeding coefficients of TR-N and TR-S were significantly higher compared to that of TR-B (Table 3.6 and Figure 3.20).

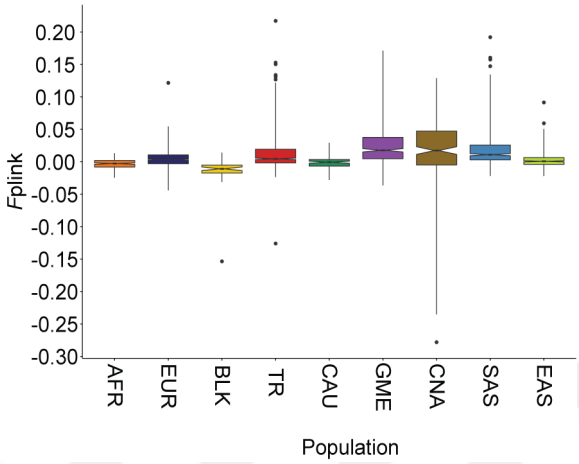



**Figure 3.19: Distributions of $F_{plink}$ in the global dataset.** Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).Reprinted from [87].

**Table 3.4** Medians and IQRs for $F_{\text{plink}}$ of the populations from the global dataset

| Population | Median | IQR |
|---|---|---|
| AFR | -0.00274 | 0.0101 |
| BLK | -0.0108 | 0.0121 |
| CAU | -0.000559 | 0.0103 |
| CNA | 0.0174 | 0.0522 |
| EAS | 0.000726 | 0.0103 |
| EUR | 0.00327 | 0.0135 |
| GME | 0.0177 | 0.0329 |
| SAS | 0.011 | 0.0229 |
| TR | 0.00459 | 0.0212 |

**Table 3.5** $P$ values of pairwise comparisons of $F_{\text{plink}}$

| | AFR | BLK | CAU | CNA | EAS | EUR | GME | SAS |
|---|---|---|---|---|---|---|---|---|
| BLK | 1.70E-06 | - | - | - | - | - | - | - |
| CAU | 0.24357 | 4.20E-06 | - | - | - | - | - | - |
| CNA | 2.10E-14 | 1.20E-08 | 2.60E-09 | - | - | - | - | - |
| EAS | 6.60E-08 | 1.50E-11 | 0.00462 | 4.20E-12 | - | - | - | - |
| EUR | <2.00E-16 | 2.40E-14 | 1.40E-07 | 2.20E-10 | 1.60E-05 | - | - | - |
| GME | <2.00E-16 | <2.00E-16 | <2.00E-16 | 0.18921 | <2.00E-16 | <2.00E-16 | - | - |
| SAS | <2.00E-16 | <2.00E-16 | <2.00E-16 | 0.50204 | <2.00E-16 | <2.00E-16 | 0.0008 | - |
| TR | <2.00E-16 | <2.00E-16 | 2.80E-11 | 0.00055 | 7.40E-13 | 3.30E-06 | <2.00E-16 | 6.50E-11 |

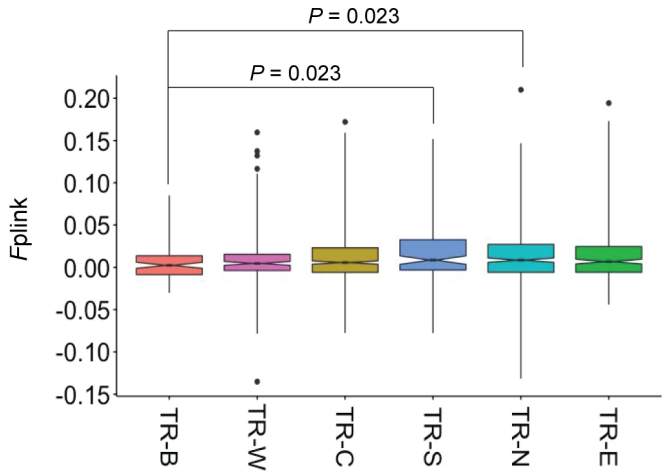



**Figure 3.20: Distributions of $F_{\text{plink}}$ in the TR subregions.** Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). $P$ values were obtained from pairwise comparisons using Wilcoxon rank sum test with Benjamini-Hochberg adjustment.

**Table 3.6** Medians and IQRs for $F_{\text{plink}}$ of the TR subregions

| Population | Median | IQR |
|---|---|---|
| TR-B | 0.0024 | 0.0222 |
| TR-C | 0.00568 | 0.0289 |
| TR-E | 0.00681 | 0.0303 |
| TR-N | 0.00846 | 0.0329 |
| TR-S | 0.00854 | 0.0358 |
| TR-W | 0.00473 | 0.0191 |

The impact of the reported parental relatedness on $F_{\text{plink}}$ was also investigated. Kruskal-Wallis test indicated that different degrees of the parental relationship significantly affect the level of $F_{\text{plink}}$ (H (3) = 557.46, P <2.2e-16). As expected, the medians for $F_{\text{plink}}$ of consanguineous or endogamous groups were significantly higher compared to that of unrelated marriages (Table 3.7 and Figure 3.21).

**Table 3.7** Medians and IQRs for $F_{\text{plink}}$ according to reported parental relatedness

| Relatedness | Median | IQR |
|---|---|---|
| Consanguineous | 0.0392 | 0.0556 |
| Endogamous | 0.00523 | 0.0236 |
| Unknown | 0.0192 | 0.0628 |
| Unrelated | 0.00208 | 0.0222 |

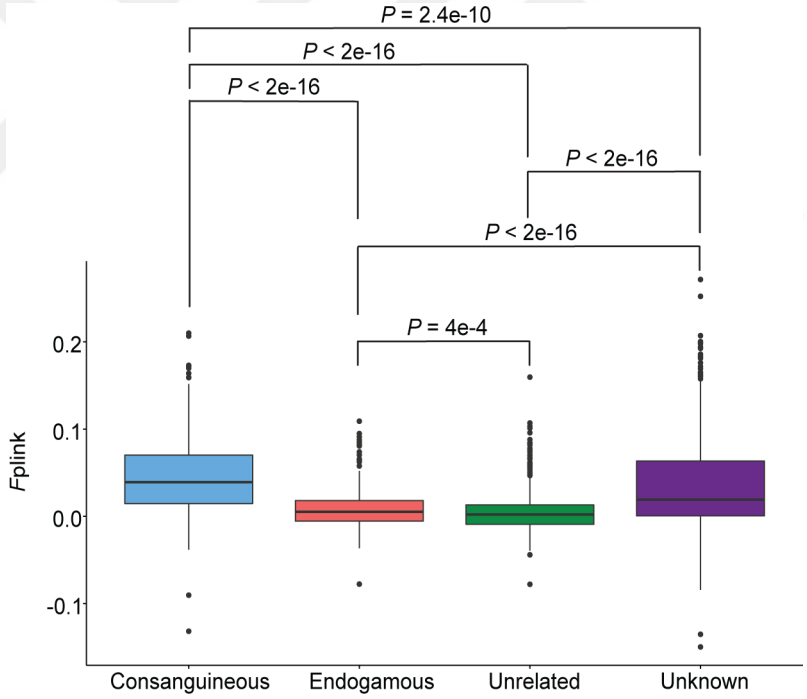



**Figure 3.21: Effects of consanguinity and endogamy on $F_{\text{plink}}$.** Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). $P$ values were obtained from pairwise comparisons using Wilcoxon rank-sum test with Benjamini-Hochberg adjustment.Reprinted from [87].

Increased inbreeding coefficients are associated with long ROHs [17]. To evaluate the number and length of ROHs in the TR population and compare them

with that of the 1000GP populations, ROHs were detected using PLINK [100]. Initially, the number of ROHs (NROH) were plotted against sum total length of ROHs (SROH) and the results were grouped based on reported parental relationship. Both NROH and SROH were elevated in the offspring of consanguineous marriages; however, overlaps between groups of parental relatedness were observed (Figure 3.22).

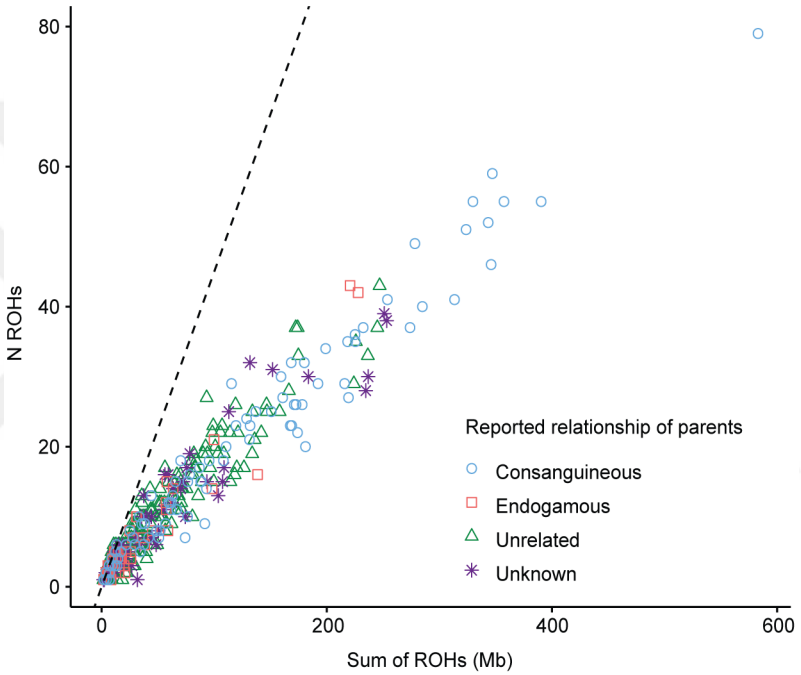



**Figure 3.22: Effect of consanguinity and endogamy on NROH and SROH.** Number of ROHs were plotted against the total length of ROHs. Shapes and colors indicate individuals with different levels of parental relationships. ($n$ = 773. Reprinted from [87].

When compared with the 1000GP populations, the TR population had the highest median while the YRI and LWK populations displayed the smallest medians for the total length of ROHs as previously reported (Figure 3.23 and Table 3.8) [117]. Short ROHs were enriched in the EAS and SAS populations (Figure 3.24), whereas medians for medium-length and long ROHs were highest in the TR population (Figures 3.25 and 3.26 and Table 3.8). There were statistically significant differences between populations in all classes of ROHs according to Kruskal-Wallis test (H(13) = 1431.1, $P$ <2x10-16; H(13) = 1408.6, $P$ <2x10-16;

H(13) = 1190.6, $P$ <2x10-16; H(13) = 332.67, $P$ <2x10-16 for total length, short, medium-length, and long ROHs, respectively). $P$ values of pairwise comparisons of all classes of ROHs in the 1000GP and TR populations using the WIlcoxon rank-sum test with the Benjamini-Hochberg adjustment were shown in Tables 3.9, 3.10, 3.11, and 3.12. Importantly, a TR individual carried the longest ROH in the dataset, which was 41 Mb in length.

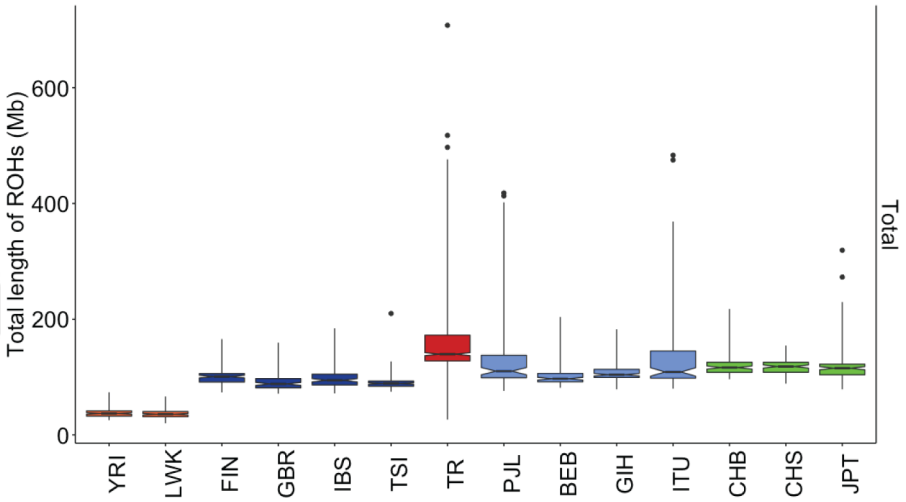



**Figure 3.23: Distributions of total ROH in TR and 1000GP populations.** Burden in samples of total length of ROH. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). Reprinted from [87].

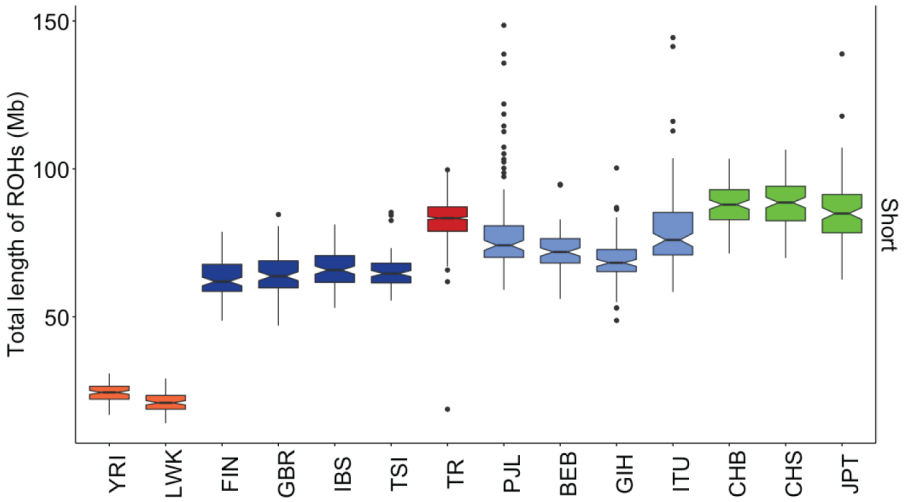



**Figure 3.24: Distributions of short ROH in TR and 1000GP populations.** Burden in samples of short ROHs (<516 Kb). Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). Reprinted from [87].

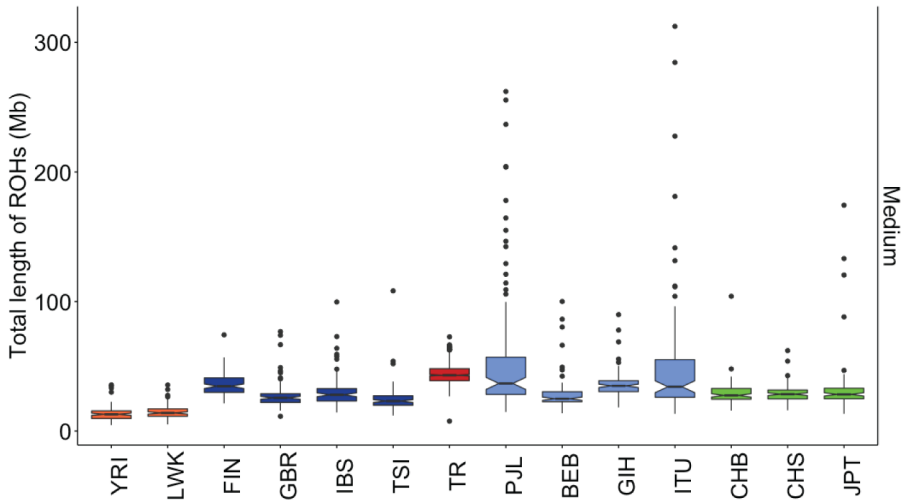



**Figure 3.25: Distributions of medium-length ROH in TR and 1000GP populations.** Burden in samples of medium-length ROHs (516-1,606Kb).Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). Reprinted from [87].

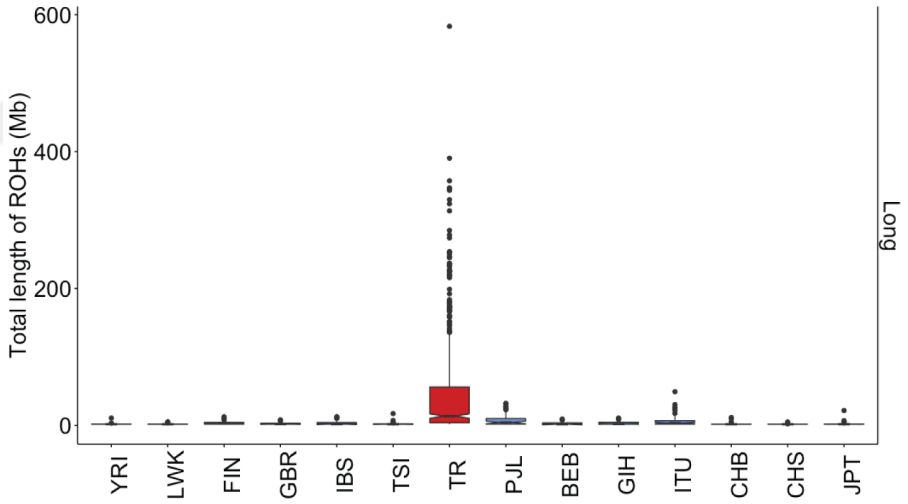



**Figure 3.26: Distributions of long ROH in TR and 1000GP populations.** Burden in samples of long ROHs (>1,606 Kb).Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers). Reprinted from [87].

**Table 3.8** Medians and IQRs for different classes of ROHs

| | Total ROH (Mb) | | Short ROH (Mb) | | Medium ROH (Mb) | | Long ROH (Mb) | |
|---|---|---|---|---|---|---|---|---|
| Population | Median | IQR | Median | IQR | Median | IQR | Median | IQR |
| BEB | 97.2 | 14.4 | 71.9 | 8.25 | 25 | 7.69 | 1.77 | 2.11 |
| CHB | 116 | 17.4 | 87.9 | 10.1 | 27.5 | 8.33 | 1.77 | 0.28 |
| CHS | 118 | 17.2 | 88.6 | 11.7 | 28.5 | 6.87 | 1.77 | 0.244 |
| FIN | 101 | 14.8 | 61.9 | 9.14 | 34.7 | 11.2 | 3.34 | 2.55 |
| GBR | 88.2 | 16 | 63.7 | 9.15 | 25.6 | 6.76 | 2.02 | 1.67 |
| GIH | 104 | 13.9 | 68.2 | 7.44 | 34.9 | 8.42 | 2.16 | 2.74 |
| IBS | 94.4 | 18.6 | 65.8 | 9.03 | 28 | 9.64 | 1.94 | 2.44 |
| ITU | 109 | 46.8 | 76 | 14.4 | 34.2 | 29 | 3.47 | 5.18 |
| JPT | 115 | 18.2 | 84.9 | 12.9 | 28.3 | 8.19 | 1.85 | 0.433 |
| LWK | 35.8 | 8.95 | 20.9 | 4.62 | 14 | 5.73 | 1.81 | 0.321 |
| PJL | 110 | 39 | 74.2 | 10.7 | 36.8 | 28.7 | 4.33 | 8.27 |
| TR | 140 | 44.6 | 83.3 | 8.25 | 43.1 | 9.3 | 13.7 | 52.3 |
| TSI | 88.9 | 9.15 | 64.6 | 6.67 | 23.2 | 7.33 | 1.79 | 0.719 |
| YRI | 36.9 | 8.96 | 24.5 | 4.38 | 13 | 5.91 | 1.88 | 0.329 |

**Table 3.9** *P* values of pairwise comparisons of sum of total length of ROHs

| | BEB | CHB | CHS | FIN | GBR | GIH | IBS | ITU | JPT | LWK | PJL | TR | TSI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHB | <2E-16 | - | - | - | - | - | - | - | - | - | - | - | - |
| CHS | <2E-16 | 0.62959 | - | - | - | - | - | - | - | - | - | - | - |
| FIN | 0.62959 | <2E-16 | <2E-16 | - | - | - | - | - | - | - | - | - | - |
| GBR | 1.50E-06 | <2E-16 | <2E-16 | 3.70E-07 | - | - | - | - | - | - | - | - | - |
| GIH | 4.90E-05 | 1.90E-09 | 6.30E-10 | 0.00074 | 6.80E-14 | - | - | - | - | - | - | - | - |
| IBS | 0.0454 | <2E-16 | <2E-16 | 0.01888 | 0.00492 | 8.10E-08 | - | - | - | - | - | - | - |
| ITU | 3.70E-07 | 0.17069 | 0.17785 | 1.00E-06 | 9.30E-16 | 0.03198 | 5.80E-11 | - | - | - | - | - | - |
| JPT | 6.50E-11 | 0.14917 | 0.05859 | 4.70E-11 | <2E-16 | 8.20E-05 | 2.90E-13 | 0.92672 | - | - | - | - | - |
| LWK | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | - | - | - | - |
| PJL | 4.60E-07 | 0.23561 | 0.21437 | 6.00E-07 | 1.30E-14 | 0.01145 | 3.90E-10 | 0.77451 | 0.98722 | <2E-16 | - | - | - |
| TR | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 2.10E-14 | - | - |
| TSI | 1.10E-10 | <2E-16 | <2E-16 | 3.80E-11 | 0.87676 | <2E-16 | 0.00018 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | - |
| YRI | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 0.17665 | <2E-16 | <2E-16 | <2E-16 |

**Table 3.10** *P* values of pairwise comparisons of sum of short ROHs

| | BEB | CHB | CHS | FIN | GBR | GIH | IBS | ITU | JPT | LWK | PJL | TR | TSI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHB | <2E-16 | - | - | - | - | - | - | - | - | - | - | - | - |
| CHS | <2E-16 | 0.47546 | - | - | - | - | - | - | - | - | - | - | - |
| FIN | <2E-16 | <2E-16 | <2E-16 | - | - | - | - | - | - | - | - | - | - |
| GBR | 5.60E-12 | <2E-16 | <2E-16 | 0.27408 | - | - | - | - | - | - | - | - | - |
| GIH | 0.00059 | <2E-16 | <2E-16 | 7.00E-10 | 1.00E-06 | - | - | - | - | - | - | - | - |
| IBS | 1.10E-08 | <2E-16 | <2E-16 | 0.00107 | 0.02642 | 0.00223 | - | - | - | - | - | - | - |
| ITU | 0.00017 | 1.10E-11 | 4.60E-12 | <2E-16 | <2E-16 | 6.90E-11 | <2E-16 | - | - | - | - | - | - |
| JPT | <2E-16 | 0.01314 | 0.003 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 3.40E-06 | - | - | - | - | - |
| LWK | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | - | - | - | - |
| PJL | 0.02954 | 4.30E-12 | 6.40E-12 | <2E-16 | 9.00E-16 | 2.80E-07 | 1.50E-12 | 0.24645 | 1.40E-07 | <2E-16 | - | - | - |
| TR | <2E-16 | 2.70E-09 | 4.60E-11 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 4.40E-09 | 0.06414 | <2E-16 | 7.40E-13 | - | - |
| TSI | 5.70E-13 | <2E-16 | <2E-16 | 0.01377 | 0.25333 | 2.00E-06 | 0.18996 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | - |
| YRI | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 3.90E-09 | <2E-16 | <2E-16 | <2E-16 |

**Table 3.11** *P* values of pairwise comparisons of sum of medium-length ROHs

| | BEB | CHB | CHS | FIN | GBR | GIH | IBS | ITU | JPT | LWK | PJL | TR | TSI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHB | 0.01117 | - | - | - | - | - | - | - | - | - | - | - | - |
| CHS | 0.0138 | 0.89752 | - | - | - | - | - | - | - | - | - | - | - |
| FIN | 1.10E-10 | 1.50E-08 | 8.20E-09 | - | - | - | - | - | - | - | - | - | - |
| GBR | 7.66E-01 | 0.00937 | 0.00576 | 2.20E-11 | - | - | - | - | - | - | - | - | - |
| GIH | 1.50E-12 | 3.00E-10 | 1.90E-10 | 8.98E-01 | 1.90E-13 | - | - | - | - | - | - | - | - |
| IBS | 1.17E-01 | 0.71974 | 0.69793 | 1.60E-07 | 0.03803 | 1.00E-08 | - | - | - | - | - | - | - |
| ITU | 1.70E-07 | 5.90E-05 | 1.20E-04 | 0.88769 | 4.70E-08 | 7.64E-01 | 4.50E-05 | - | - | - | - | - | - |
| JPT | 0.01875 | 0.80074 | 0.88868 | 1.30E-07 | 0.00641 | 4.40E-09 | 0.66067 | 3.20E-04 | - | - | - | - | - |
| LWK | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | - | - | - | - |
| PJL | 2.20E-09 | 1.50E-07 | 1.30E-07 | 0.12462 | 3.00E-10 | 2.06E-01 | 2.50E-07 | 0.26438 | 8.00E-07 | <2E-16 | - | - | - |
| TR | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 7.20E-06 | <2E-16 | <2E-16 | 3.24E-03 | - | - |
| TSI | 2.04E-03 | 1.80E-08 | 1.50E-08 | <2E-16 | 0.02789 | <2E-16 | 3.70E-06 | 3.30E-14 | 1.70E-07 | <2E-16 | 1.40E-15 | <2E-16 | - |
| YRI | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | <2E-16 | 2.64E-02 | <2E-16 | <2E-16 | <2E-16 |

**Table 3.12** $P$ values of pairwise comparisons of sum of long ROHs

| | BEB | CHB | CHS | FIN | GBR | GIH | IBS | ITU | JPT | LWK | PJL | TR | TSI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHB | 0.71262 | - | - | - | - | - | - | - | - | - | - | - | - |
| CHS | 0.45423 | 0.9271 | - | - | - | - | - | - | - | - | - | - | - |
| FIN | 3.50E-01 | 6.00E-02 | 8.94E-03 | - | - | - | - | - | - | - | - | - | - |
| GBR | 1.00E+00 | 0.46605 | 0.2288 | 1.64E-01 | - | - | - | - | - | - | - | - | - |
| GIH | 4.74E-01 | 9.03E-02 | 2.26E-02 | 7.42E-01 | 3.57E-01 | - | - | - | - | - | - | - | - |
| IBS | 5.44E-01 | 0.1643 | 0.05809 | 6.65E-01 | 0.47403 | 8.64E-01 | - | - | - | - | - | - | - |
| ITU | 1.24E-01 | 1.97E-02 | 4.35E-03 | 0.41397 | 5.48E-02 | 2.27E-01 | 2.20E-01 | - | - | - | - | - | - |
| JPT | 0.88352 | 0.47403 | 0.41397 | 9.02E-02 | 0.77653 | 1.61E-01 | 0.34666 | 2.86E-02 | - | - | - | - | - |
| LWK | 0.62908 | 0.82453 | 0.74204 | 0.03183 | 0.41397 | 0.05663 | 0.16107 | 0.01254 | 0.69743 | - | - | - | - |
| PJL | 3.52E-02 | 4.35E-03 | 6.90E-04 | 0.05481 | 7.71E-03 | 2.26E-02 | 3.28E-02 | 0.43321 | 5.16E-03 | 0.00229 | - | - | - |
| TR | 1.70E-05 | 1.00E-06 | 8.10E-08 | 2.00E-12 | 2.10E-09 | 2.20E-15 | 3.50E-10 | 4.30E-09 | 3.40E-09 | 2.40E-07 | 5.70E-06 | - | - |
| TSI | 7.42E-01 | 7.77E-01 | 7.15E-01 | 0.1281 | 0.74204 | 2.43E-01 | 4.31E-01 | 6.02E-02 | 9.20E-01 | 0.7965 | 1.25E-02 | 2.30E-06 | - |
| YRI | 0.9196 | 0.66535 | 0.46605 | 0.20753 | 0.80705 | 0.31811 | 0.47403 | 0.12427 | 0.93663 | 6.97E-01 | 0.05481 | 0.0001 | 0.93663 |

The results of the ROH analyses were also stratified according to the TR subregions (Figure 3.27 and Table 3.13). For the medium-length of ROHs, one-way ANOVA did not reveal a statistically significant difference between the TR subregions ($F(5,641) = 1.291$, $P = 0.266$). However, there were statistically significant differences between subregions in other classes of ROHs according to Kruskal-Wallis test ($H(5) = 12.508$, $P = 0.028$; $H(5) = 38.298$, $P = 3.28\text{x}10\text{-}7$; $H(5) = 37.189$, $P$ 5.49x10-7 for total length, short, and long ROHs,respectively). $P$ values of pairwise comparisons of the TR subregions for total, short, and long ROHs were shown in Tables 3.14, 3.15, and 3.16. The TR-B subregion had the statistically lowest long and total length of ROHs while the TR-W subregion had significantly lower values for the same categories compared to the TR-C, TR-E, and TR-N subregions. On the contrary, the TR-B subregion had significantly the highest value for the short ROHs, whereas the TR-W subregion had a higher value compared to the TR-N and TR-E subregions for the same category.

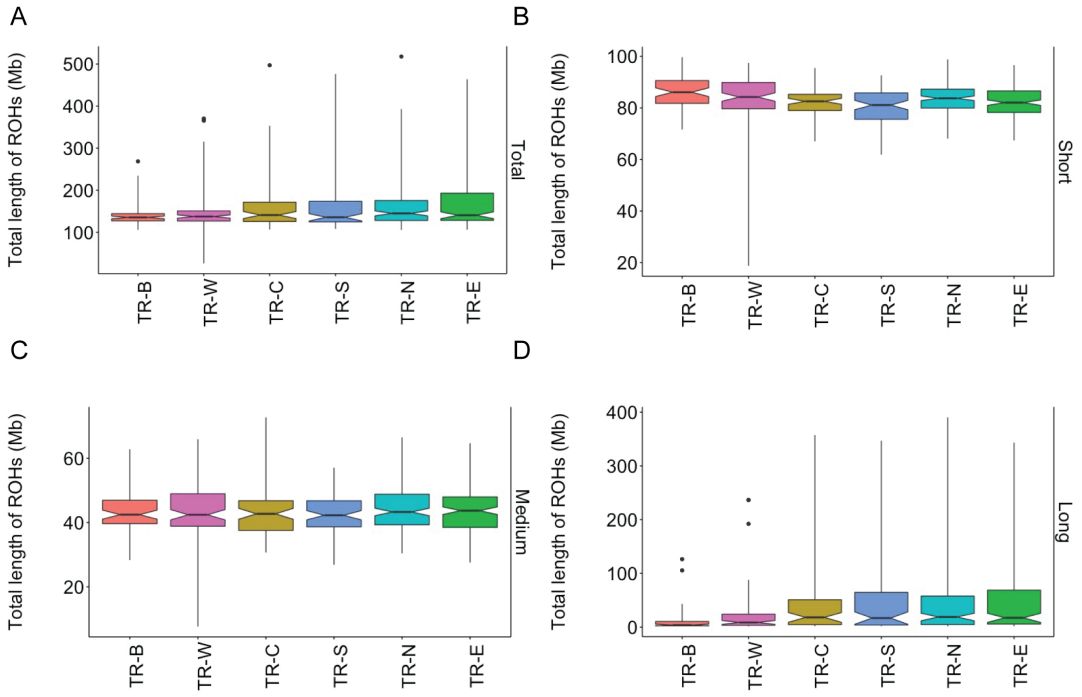



**Figure 3.27: Distributions of ROH in the TR subregions.** Burden in samples of **A** total length of ROHs; **B** short ROHs; **C** medium-length ROHs; **D**long ROHs.Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

The frequencies of ROHs were calculated and binned according to their lengths. ROHs longer than 4Mb in length were binned together. Results showed that 385 (49.81%) TR individuals had ROHs longer than 4Mb that were exclusively found in the TR population (Figure 3.28).

**Table 3.13** Medians and IQRs for different classes of ROHs

| | Total ROH (Mb) | | Short ROH (Mb) | | Medium ROH (Mb) | | Long ROH (Mb) | |
|---|---|---|---|---|---|---|---|---|
| Subregion | Median | IQR | Median | IQR | Median | IQR | Median | IQR |
| TR-B | 135 | 17.1 | 86.1 | 8.84 | 42.5 | 7.27 | 3.82 | 8.01 |
| TR-C | 141 | 45.8 | 82.6 | 6.28 | 42.7 | 9.25 | 18.1 | 46.4 |
| TR-E | 141 | 64.5 | 82 | 8.37 | 43.7 | 9.49 | 17.3 | 63.1 |
| TR-N | 145 | 47.4 | 83.7 | 7.34 | 43.3 | 9.51 | 18.8 | 53 |
| TR-S | 136 | 48.7 | 81.1 | 10.2 | 42.3 | 8.11 | 16.8 | 60.7 |
| TR-W | 138 | 24 | 84.2 | 10.2 | 42.4 | 10.1 | 8.69 | 20.8 |

**Table 3.14** $P$ values of pairwise comparisons of sum of total length of ROHs in the TR subregions

| | TR-B | TR-C | TR-E | TR-N | TR-S |
|---|---|---|---|---|---|
| TR-C | 4.1E-05 | - | - | - | - |
| TR-E | 3.6E-06 | 0.8879 | - | - | - |
| TR-N | 1.9E-05 | 8.88E-01 | 6.67E-01 | - | - |
| TR-S | 3E-04 | 0.8879 | 0.6665 | 9.07E-01 | - |
| TR-W | 4.24E-02 | 2.59E-02 | 3.40E-03 | 1.93E-02 | 5.57E-02 |

**Table 3.15** $P$ values of pairwise comparisons of sum of short ROHs in the TR subregions

| | TR-B | TR-C | TR-E | TR-N | TR-S |
|---|---|---|---|---|---|
| TR-C | 1E-03 | - | - | - | - |
| TR-E | <1E-3 | 1 | - | - | - |
| TR-N | 7.90E-02 | 7.63E-01 | 3.77E-01 | - | - |
| TR-S | <1E-3 | 1 | 1 | | - |
| TR-W | 4.77E-01 | 4.47E-01 | 2.38E-01 | 1.93E-02 | 5E-03 |

**Table 3.16** $P$ values of pairwise comparisons of sum of long ROHs in the TR subregions

| | TR-B | TR-C | TR-E | TR-N | TR-S |
|---|---|---|---|---|---|
| TR-C | 4.10E-05 | - | - | - | - |
| TR-E | 3.60E-06 | 0.8879 | - | - | - |
| TR-N | 1.90E-05 | 8.88E-01 | 6.67E-01 | - | - |
| TR-S | 3E-04 | 0.8879 | 0.6665 | 9.07E-01 | - |
| TR-W | 4.24E-02 | 2.59E-02 | 3.40E-03 | 1.93E-02 | 5.57E-02 |

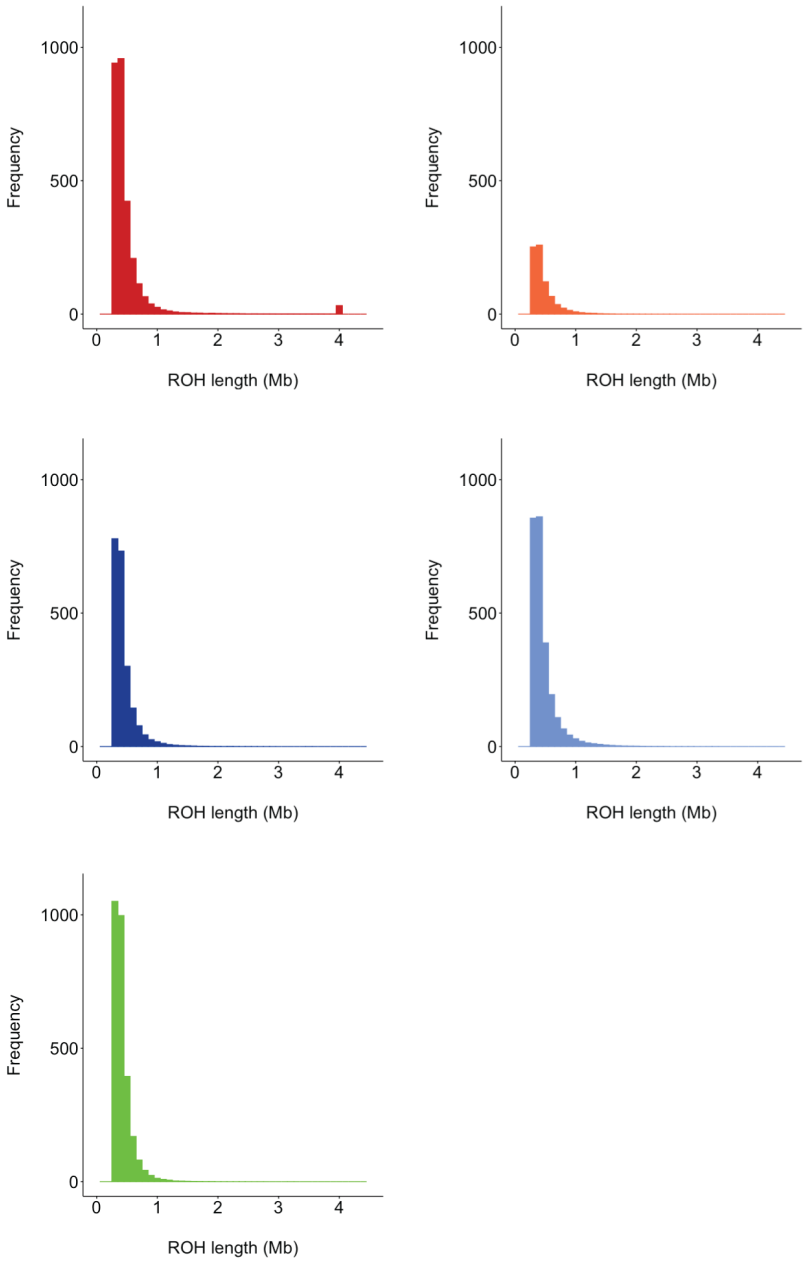



**Figure 3.28: Histograms of the frequencies of ROHs in the TR and 1000GP populations.** Frequencies were calculated by dividing the number of ROH by the population size. Reprinted from [87].

$F_{\mathrm{ROH}}$, the length of the autosomal genome in ROH, was calculated as a measure of autozygosity using the long and total length of ROHs. Both $F_{\mathrm{ROH\ (total)}}$ and $F_{\mathrm{ROH\ (long)}}$ had a strong positive correlation with $F_{\mathrm{plink}}$ ($P$ <2.2x $10^{-16}$, Figure 3.29 A and B). The medians for both $F_{\mathrm{ROH\ (total)}}$ and $F_{\mathrm{ROH\ (long)}}$ were higher in the offspring of consanguineous or endogamous marriages, similar to the results

of the $F_{\text{plink}}$ analysis. (Figure 3.29 C and D).

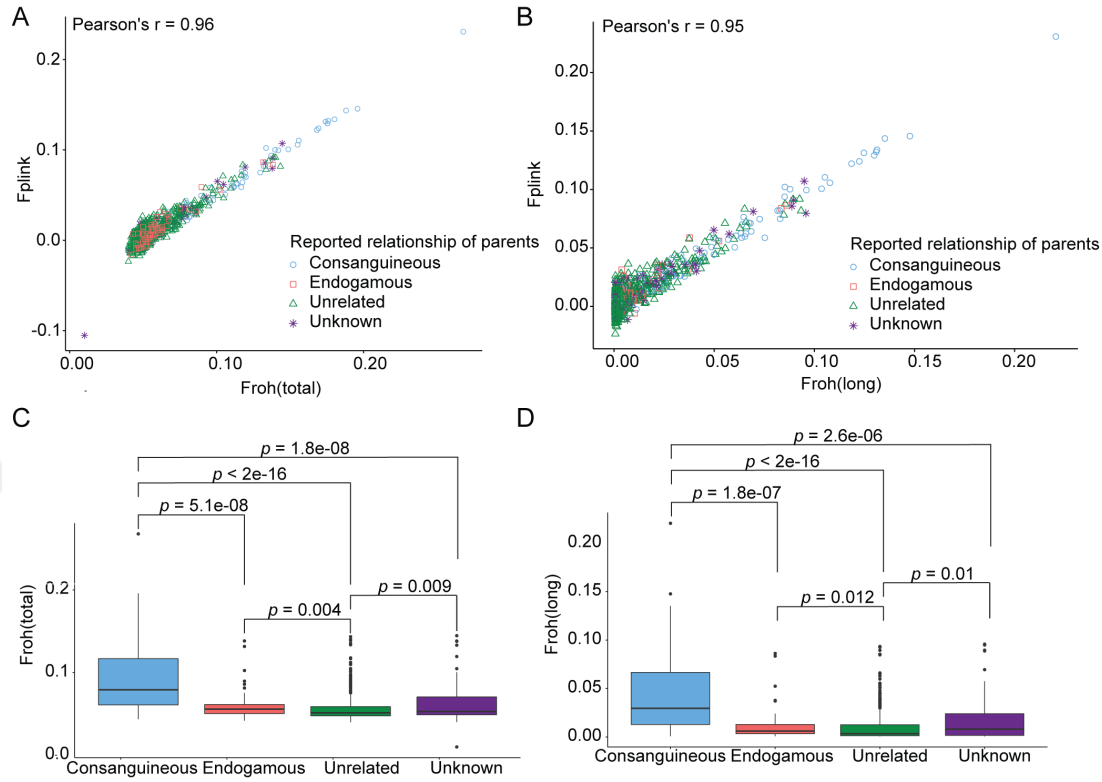



**Figure 3.29: $F_{\text{ROH}}$, $F_{\text{plink}}$ and the effect of parental relationship. A** The correlation of $F_{\text{plink}}$ and $F_{\text{ROH}}$(total). **B** The correlation of $F_{\text{plink}}$ and $F_{\text{ROH}}$(long). **C** Kruskal-Wallis test indicated that different degrees of the parental relationship significantly affect the level of $F_{\text{ROH}}$ (total) (H (3) = 112.2, P <2.2e-16). **D** Kruskal-Wallis test indicated that different degrees of the parental relationship significantly affect the level of $F_{\text{ROH}}$ (long) (H (3) = 97.135, P <2.2e-16). Reprinted from [87].

## 3.3 The distribution of Y-chromosome and mtDNA haplotypes

Y-chromosome and mtDNA haplotypes were inferred using TR WGS data and plotted along with the haplotypes of individuals from 1000GP, Lazaridis *et al.*, and HGDP for comparison [27, 81, 121]. J2a (18.4%), R1b (14.9%), and R1a (12.1%) constituted the most common Y-chromosome haplogroups in TR males,

consistent with the previous reports (Figure 3.30) [76]. The distribution of Y-chromosome haplogroups was similar in the TR subregions, except for TR-B where I2a (20%) was the most frequent followed by R2a (17.1%) and E1b (14.3%) (Figure 3.31). When the mtDNA haplotypes were evaluated, the most common mtDNA haplogroups in the TR population were from the H sublineage (27.55%), followed by U (19.53%) and T (10.99%), in line with the previous findings (Figure 3.32) [144]. The distribution of mtDNA haplotypes was also similar in the TR subregions except for TR-B, in which the proportion of the T lineage was very low (Figure 3.33). Additionally, the paternal and maternal gene flow from Central Asia were assessed using the proportion of Y-chromosome and mtDNA haplogroups that are suggested to be restricted to Central Asia [76]. The frequency of Y-chromosome sublineages C-RPS4Y and O3-M122 in the Central Asian populations were reported as 33% and 18% [76]. 13 (2.81%) TR individuals had these haplotypes, therefore, their contributions were calculated as from 0.0281/0.329100=8.5% to 0.0281/0.180100=15.6%. mt-DNA haplogroups D4c and G2a were also previously suggested as Central Asian-specific and observed at 8% in the Central Asian population in a previous study [123]. There were 5 TR individuals (0.65%) individuals with these haplogroups, therefore, their contribution was calculated as 0.0065/0.08100= 8.13%. High-resolution assignments of Y-chromosome and mtDNA haplogroups were listed in Table A.2.

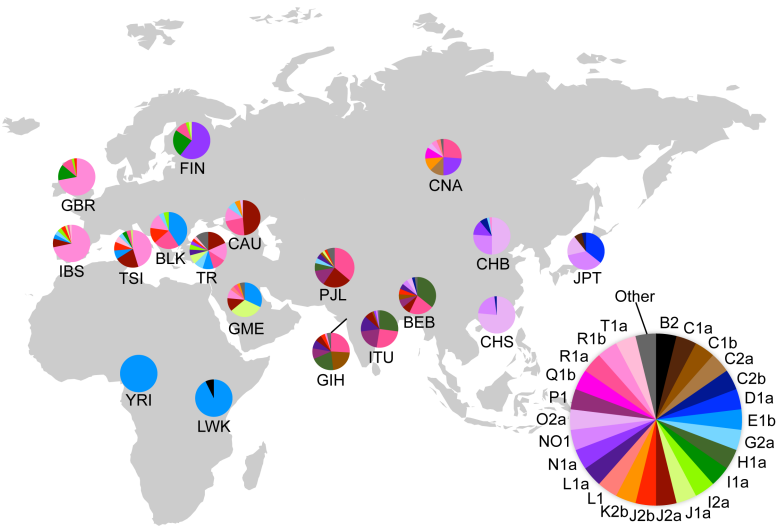



**Figure 3.30: Y-haplogroup distribution in the TR and control populations** Only main haplogroups are shown. Reprinted from [87].

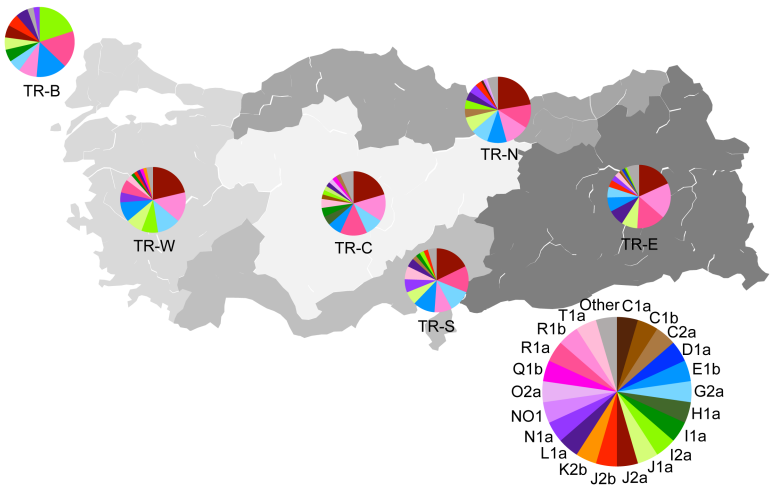



**Figure 3.31: Y-haplogroup distribution in the TR subregions.** The Y-chromosome haplogroups of TR males with known ancestral origin ($n = 370$). Only main haplogroups are shown. Reprinted from [87].

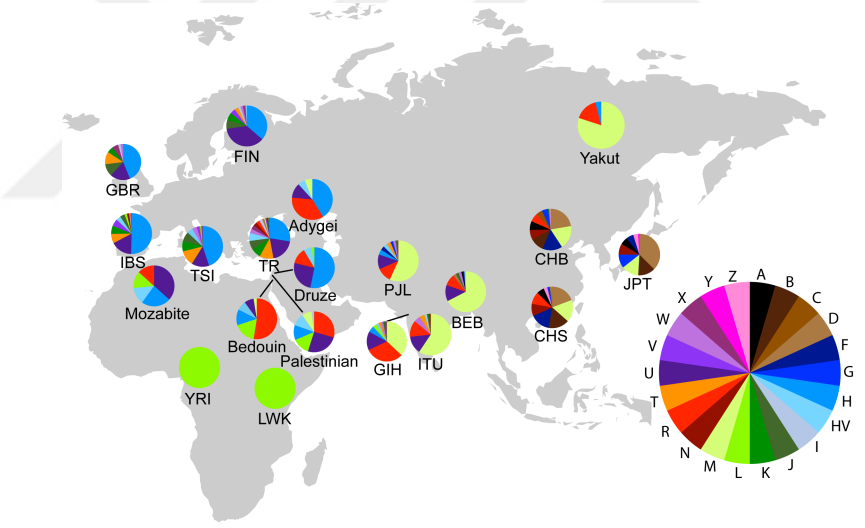



**Figure 3.32: mtDNA-haplogroup distribution in the TR and control populations.** mtDNA sequences of Adygei, Bedouin, Druze, Mozabite, Palestinian, and Yakut populations of the HGDP [121] were used in addition to the TR and 1000GP populations. Reprinted from [87].

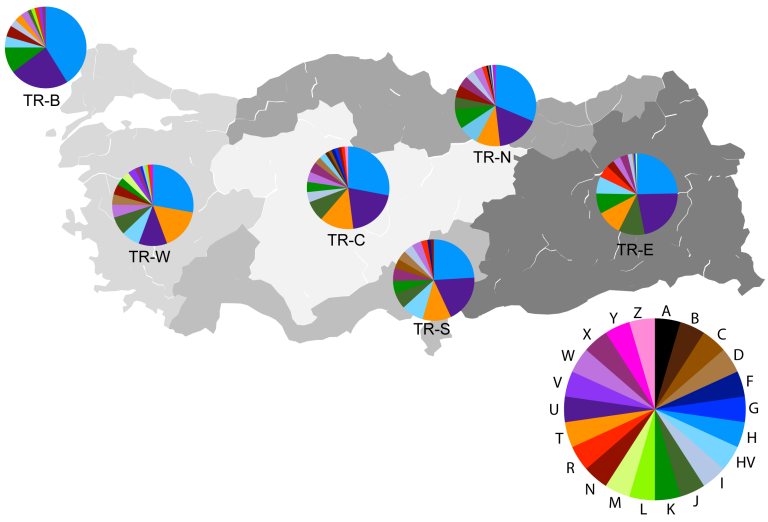



**Figure 3.33: mtDNA-haplogroup distribution in the TR subregions.** The mt-DNA haplogroups of TR individuals with known ancestral origin ($n$ = 647). Only main haplogroups are shown. Reprinted from [87].

## 3.4 The TR Variome

Since the identification of population-specific variants significantly contributes to the discovery of disease genes, one of the goals of the study was to generate the TR Variome that lists high-quality variants of the 3,346 TR individuals [17]. First, DAFs were calculated using TR WES and WGS data then compared with those of gnomAD WES, gnomAD WGS, and GME Variome. Roughly 28% of the WES and 49% of the WGS DAFs in the very rare DAF bins (AF <0.005) were specific to the TR Variome (Figure 3.34 A and B). Also, GME Variome did not contain about 79% of the very rare DAFs of the TR Variome (Figure 3.34 C). The frequencies of derived alleles in the rare DAF bins (<0.01) of the TR Variome searched in the gnomAD and GME. Although Pearson's test resulted in high correlations, heatmaps revealed a remarkable amount of TR DAFs (especially DAFs in TR-WGS) could be accurately calculated using neither datasets (Figure 3.35). Therefore, results suggest that the GME Variome is an inadequate representation of the TR population, albeit containing 170 TR samples.

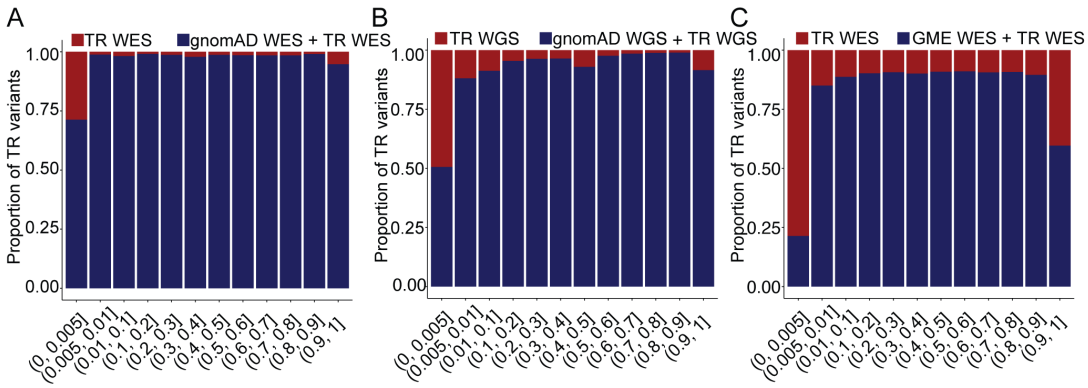



**Figure 3.34: DAFs in the TR population.** The proportion of TR derived alleles represented in the TR Variome versus gnomAD and GME. **A** TR WES versus gnomAD WES, **B** TR WGS versus gnomAD WGS, **C** TR WES versus GME WES. Reprinted from [87].

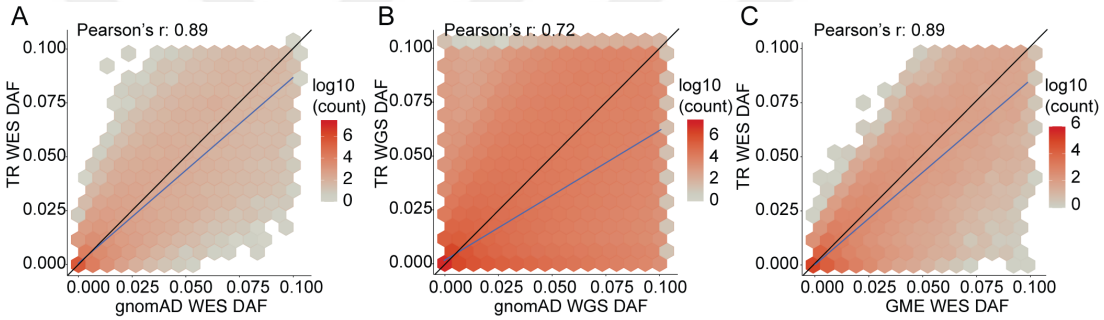



**Figure 3.35: The correlation of rare TR DAFs with those of gnomAD and GME. A** TR WES versus gnomAD WES, **B** TR-WGS versus gnomAD WGS, and **C** TR-WES versus GME WES. The log-transformed number of variants in hexagonal bins was shown in shades of red. Reprinted from [87].

Then, the variants in the TR Variome were categorized based on their functional impact on gene product into seven main groups: HC-pLoFs, LC-pLoFs, missense variants, non-frameshift indels, synonymous variants, non-coding variants, and other effects (Table 3.17). The missense variants were further classified into two according to their predicted deleteriousness as deleterious missense and other missense using a combination of CADD, SIFT, and Polyphen-2. Additionally, variants were categorized based on their AFs in other publicly-available variant datasets. In total, 9,999,451 novel variants were detected of which 37,123 were predicted to be HC-pLoF or deleterious missense. Surprisingly, 839,775 (2.55%) rare or novel variants were observed with an AF >1% in the TR population.

**Table 3.17** Functional annotation and AF distribution of TR variants

| | | AF in other public databases | | |
|---|---|---|---|---|
| | | Novel | Rare (AF<0.01) | Common (AF≥0.01) |
| All variants | | 9,999,451 | 22,932,246 | 13,807,782 |
| Functional consequence | | | | |
| High-confidence pLoFs | Frameshift variant | 4,271 | 3,453 | 490 |
| | Splice site variant | 2,932 | 3,084 | 225 |
| | Start loss variant | 1 | - | - |
| | Stop gain variant | 2,829 | 5,053 | 223 |
| | Stop loss variant | 4 | 1 | 2 |
| Low-confidence pLoFs | Frameshift variant | 2,110 | 2,518 | 667 |
| | Splice site variant | 1,795 | 3,035 | 1,784 |
| | Start loss variant | 445 | 949 | 158 |
| | Stop gain variant | 1,221 | 2,860 | 345 |
| | Stop loss variant | 322 | 503 | 149 |
| Missense variants | Deleterious missense | 27,086 | 64,177 | 1,728 |
| | Other missense | 53,768 | 192,554 | 41,172 |
| Non-frameshift indels | | 2,621 | 6,712 | 1,728 |
| Synonymous variants | | 53,768 | 192,554 | 41,172 |
| Other effects | Protein-protein contact | 149 | 348 | 40 |
| | Exon loss variant | - | 2 | - |
| | Gene fusion | 12 | 26 | 7 |
| | Structural interaction variant | 3,585 | 11,044 | 1,289 |
| | Bidirectional gene fusion | 15 | 36 | 15 |
| | Transcription Factor Binding Site (TFBS) ablation | 116 | 243 | 97 |
| | Non-essential splice site variant | 22,578 | 54,573 | 20,594 |
| | Initiator codon variant | 34 | 60 | 9 |
| | Stop retained variant | 81 | 180 | 53 |
| Non-coding variants | Intergenic region | 3,617,719 | 8,210,394 | 5,318,191 |
| | Intragenic variant | 846 | 1,767 | 994 |
| | Intron variant | 3,538,574 | 8,102,039 | 4,871,476 |
| | Upstream gene variant | 1,377,873 | 3,124,448 | 1,837,118 |
| | TFBS variant | 10,101 | 22,928 | 10,170 |
| | Sequence feature | 70,500 | 163,949 | 93,076 |
| | Downstream gene variant | 980,383 | 2,261,777 | 1,360,894 |
| | Non-coding transcript exon variant | 26,319 | 63,387 | 38,964 |
| | Untranslated region (UTR) variant | 148,801 | 345,901 | 163,203 |

Reprinted from [87].

When the variants were stratified using both functional impact and frequency, rare and novel categories contained higher proportions of HC-pLoF and deleterious missense variants than the common category. Similarly, gnomAD WES and WGS data contained higher proportions of HC-pLoF and deleterious missense variants in the rare category (Figure 3.36). Besides, the list of private variants (variants that are unique to a single individual either in the heterozygous or the homozygous state) of the TR Variome was generated. The list comprised 23,403,893 private variants of which 8,898,088 (38 %) were also novel. The number of private variants that were predicted to be HC-pLoFs or deleterious missense was calculated as 79,947 (0.34%). Among those, 32,687 (0.14%) were specific to the TR Variome.

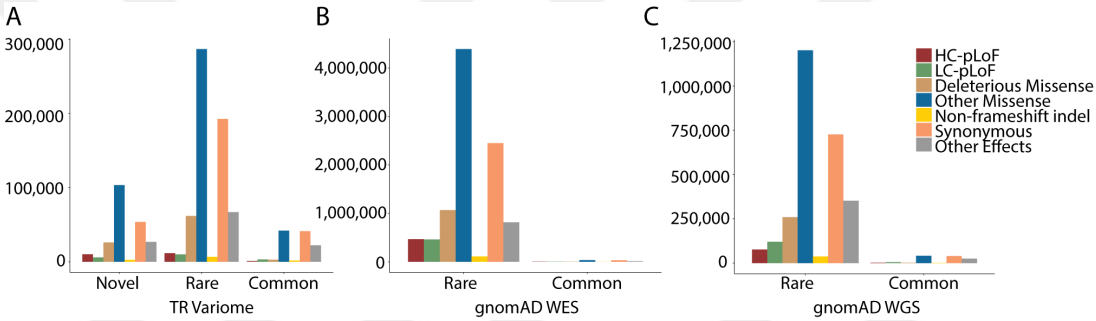



**Figure 3.36: Variant categories based on functional impact and frequency. A** in the TR Variome, **B** in the gnomAD WES data, **C** in the gnomAD WGS data. Non-coding variants were not shown. Panel A was reprinted from [87].

## 3.5 Homozygous predicted loss of function mutations

Populations with a high level of inbreeding enable the investigation of homozygous pLoFs [17, 57, 136]. Initially, rare homozygous HC-pLoFs (AF <1%) were searched in the TR Variome because rare homozygous HC-pLoFs are more likely to affect the function of a gene product [130]. 704 rare homozygous HC-pLoFs were identified in the TR Variome, which were located in 626 different genes (Table A.3). 680 (20.22%) individuals carried homozygous HC-pLoF variants while each had up to 4 genes with these variants. The list of homozygous HC-pLoF variants in the TR Variome was then compared with the previously published lists of rare human knockouts and homozygous pLoFs in the gnomAD and 1000GP

data [34, 57, 136, 137]. 173 genes with homozygous pLoFs were not previously reported in publications or present in the gnomAD and 1000GP and thus were specific to TR Variome.

Usually, common pLoFs are not subject to purifying selection and do not have a disruptive impact. Indeed, common homozygous pLoF variants (population frequency >1%) might be the conclusions of selective advantage or gene redundancy. A previous report generated the list of common homozygous pLoFs in ExAC and gnomAD [138].The list of common HC-pLoF variants in the TR Variome, which comprised 307 variants in 268 genes, was generated therefore as well (Table A.4). 48 genes (15.64%) in the list of common homozygous HC-pLoFs were also reported in the published list of ExAC/gnomAD. Notably, 259 variants in 227 genes, which were previously designated as rare knockouts, were observed in the TR population with a frequency higher than 1%. The expected number of homozygous pLoFs was calculated based on HWE. Results showed that the numbers of rare homozygous HC-pLoFs were exceeded than estimated and thus reflected the effect of the high level of consanguinity in the TR population.

Pext values of both rare and common single-nucleotide homozygous pLoFs were estimated to evaluate the effect of on transcriptional output [52]. High pext scores indicate a higher impact on transcriptional output. There were 463 rare single-nucleotide homozygous pLoFs of which 71 (15.3%) had a low (<0.1), 228 (49.2%) had a medium (0.1 -0.9), and 164 (35.4%), had a high (>0.9) pext value. On the other hand, there were 105 common single-nucleotide homozygous pLoFs of which 51 (48.6%) had a low, 42 (40%) had a medium, and 12 (11.4%) had a high pext value. Therefore, rare single-nucleotide homozygous pLoFs demonstrated a higher impact on transcriptional output.

## 3.6 Clinically relevant variants

OMIM, HGMD, and ClinVar were used to unravel clinically relevant variants in the TR Variome. LoF variants more likely cause a deleterious impact on proteins and have a top priority to be investigated as disease-causing variants; therefore the list of HC-pLoF variants in the TR Variome was generated. 22,570 HC-pLoF variants in 9,081 genes were identified in the TR Variome. 76.1% of these variants

were located under 6,831 OMIM-listed genes and 25.37% of them were located under 2,197 OMIM-listed genes with an associated phenotype. Then, these HC-pLoF variants were searched in other public variant resources and categorized as novel, rare or common. Novel and rare categories contained a significantly higher number of HC-pLoF variants compared to the common category. But the proportion of these HC-pLoFs in OMIM-listed genes and OMIM-listed genes with an associated clinical phenotype was similar (Figure 3.37).

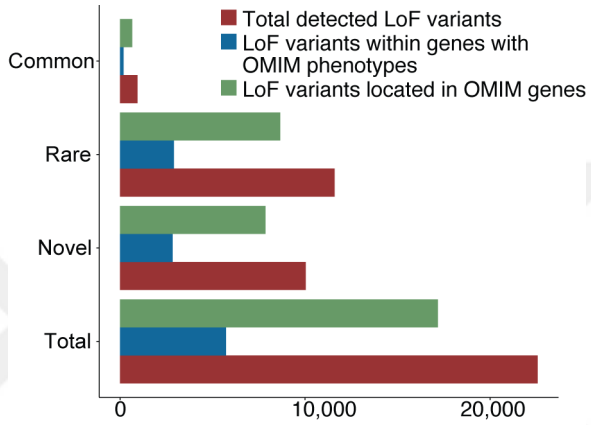



**Figure 3.37: Distribution of variants according to OMIM annotation.** Proportions of HC-pLoF variants, which were grouped based on their frequency, their location on OMIM genes and the OMIM genes associated with a phenotype. Reprinted from [87].

Afterwards, variants of the TR population were searched in HGMD [51] and ClinVar [50]. There were 6,537 variants in 2,188 genes classified as DM in HGMD. All TR individuals carried at least 1 DMs in their genome (range = 1-30, average = 12, 0-5 in the homozygous state)(Figure 3.38 A and Table A.5). Furthermore, 1,636 variants in 929 genes were reported as P or P/LP in ClinVar. 3,355 (99.79%) individuals possessed 1 to 19 such variants with an average of 6 (0-10 in the homozygous state) (Figure 3.38 B and Table A.6). 1,376 variants in 782 genes were classified as DM in HGMD and P or P/LP in ClinVar (Figure 3.39).

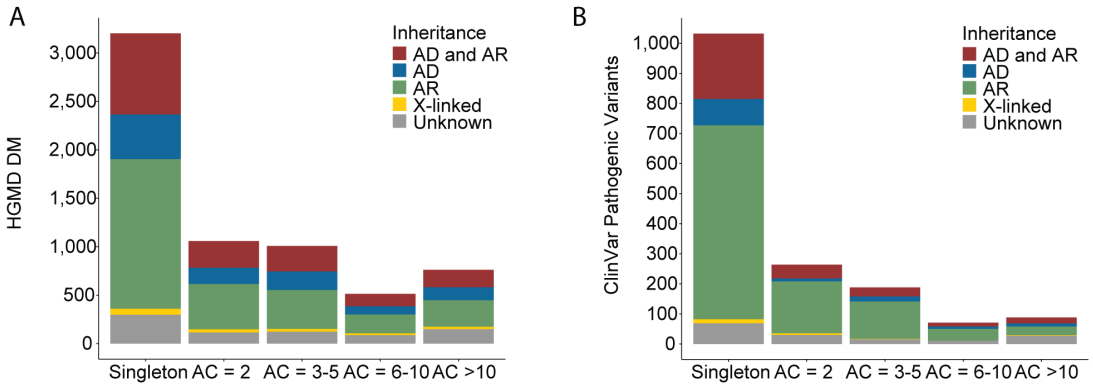



**Figure 3.38: Distribution of variants in HGMD and ClinVar variant classes. A** Number of DMs from HGMD and **B** P and P/LP variants from the ClinVar database, categorized based on their frequency in the TR population and inheritance type as AD, AR, X-linked or unknown. Reprinted from [87].

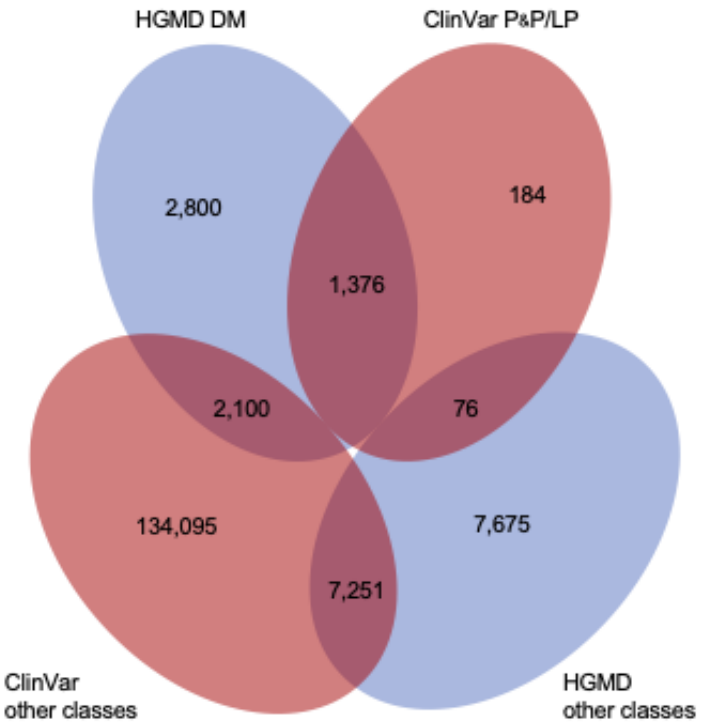



**Figure 3.39: The Venn Diagram demonstrating the number of the TR variants previously reported in HGMD and/or ClinVar.**HGMD had a higher number of records in the DM class compared to the number of variants in the P or P/LP classes in ClinVar. Reprinted from [87].

## 3.7 Per-genome variant summary and imputation panel

The genome-wide variation of TR individuals on a global scale was evaluated using per-genome variant sites and singletons (Figure 3.40). Consistent with the previous studies, different levels of genetic variation were observed in global populations. [27]. The number of variant sites per genome varied in TR individuals similar to what was observed in the recently admixed American populations. Importantly, the TR population had a remarkably higher mean for per-genome variant sites compared to the EUR populations.

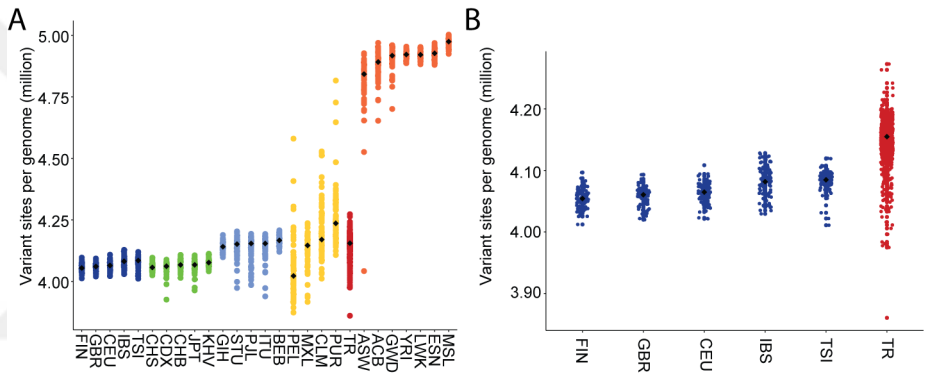



**Figure 3.40: Genome-wide variation in the TR and 1000GP populations. A** The number of variant sites per genome for autosomes. **B** The average number of variant sites per genome is higher in the TR population than in the EUR populations. Reprinted from [87].

Interestingly, the average number of variants observed in a single individual 'singletons' was higher in the TR and LWK populations than that of the other 1000GP populations (Figure 3.41).

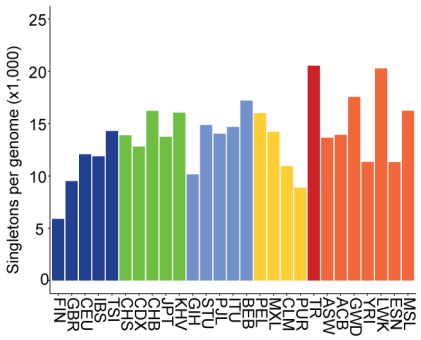



**Figure 3.41: The number of singletons of the 1000GP and TR populations.** The average number of singletons per genome for autosomes. Reprinted from [87].

Using the WGS data of 773 TR individuals, a haplotype reference panel was generated to be used in future GWAS. The predictive accuracy of the newly-generated TR reference panel was evaluated with a pseudo-GWAS panel and compared with that of 1000GP. The aggregate squared Pearson correlation coefficient ($R^2$) revealed that employing the TR reference panel alone resulted in significantly higher imputation accuracy, especially for the variants with AF <5% compared to the imputation with 1000GP reference panel. (Two-tailed Wilcoxon rank-sum test, $P = 0.002$ for the $R^2$ of the TR reference panel (mean ± SD: 0.88 ± 0.12) versus the $R^2$ of 1000GP reference panel (mean ± SD: 0.86 ± 0.14). When the both panels were applied simultaneously, the level of imputation accuracy was further improved (Two-tailed Wilcoxon rank-sum test, $P = 0.002$ for the $R^2$ of the TR+1000GP reference panel (mean ± SD: 0.89 ± 0.09) versus the $R^2$ of the 1000GP reference panel (mean ± SD: 0.86 ± 0.14). (Figure 3.42 A). Besides, the expected number of high-confidence imputed variants ($R^2$ >0.8) with expected AF <1% was higher when the TR reference panel was employed. On the other hand, the combined panel yielded a higher number of expected high-confidence variants with expected AF ≥ 1% (Figure 3.42 B). Overall, the TR reference panel produced 3,911 expected high-confidence rare variants (expected AF <1%) that were not imputed by the 1000GP panel while the combined panel revealed 20,951 and 3,902 expected high-confidence variants (expected AF ≥ %1) that were not captured by the TR and the 1000GP, respectively.

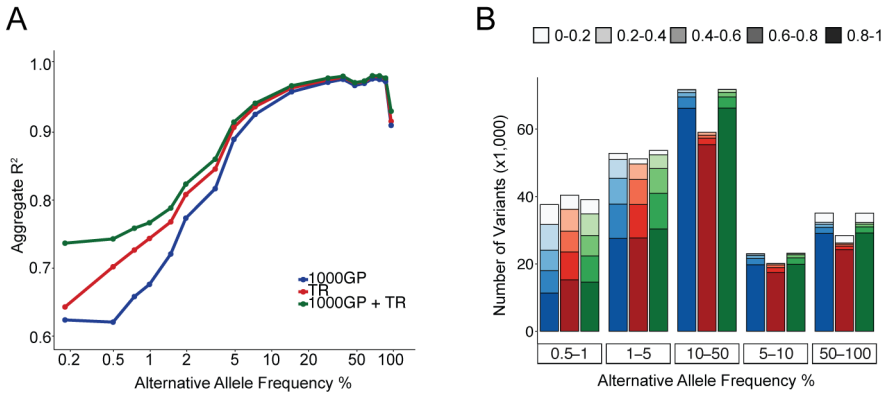



**Figure 3.42: Evaluation of imputation performance of TR, 1000GP, and TR + 1000GP reference panels on the TR samples. A** $R^2$ levels of variants imputed with TR, 1000GP, and TR + 1000GP reference panels in different frequency bins **B** The number of imputed variants as a function of expected alternative allele frequency. The density of shading demonstrates the level of expected imputation accuracy. Colors represent the reference panels used for the imputation: 1000GP (blue), TR (red), and TR + 1000GP (green). Reprinted from [87].

The imputation performance of the TR reference panel in the neighboring populations (BLK, CAU, and GME) was also assessed [142]. The TR reference panel alone resulted in the highest $R^2$ for the imputation of the genotypes of the CAU population (two-tailed Wilcoxon rank-sum test, $P = 0.041$ for the $R^2$ of the TR+1000GP reference panels (mean ± SD: 0.96 ± 0.05) versus the $R^2$ of the 1000GP reference panel (mean ± SD: 0.95 ± 0.06, $P$ >0.05 for the TR + 1000GP (mean ± SD: 0.95 ± 0.06) versus 1000GP. ). TR + 1000GP panel revealed statistically higher accuracy for the BLK population ($P$ >0.05 for the TR (mean ± SD: 0.96 ± 0.04) versus 1000GP (mean ± SD: 0.96 ± 0.04); $P = 0.009$ for the TR + 1000GP (mean ± SD: 0.96 ± 0.04) versus 1000GP) and the GME population ($P = 0.009$ for the TR (mean ± SD: 0.93 ± 0.07) versus 1000GP (mean ± SD: 0.95 ± 0.06); $P = 0.009$ for the TR + 1000GP (mean ± SD: 0.95 ± 0.05) versus 1000GP.) (Figure 3.43).

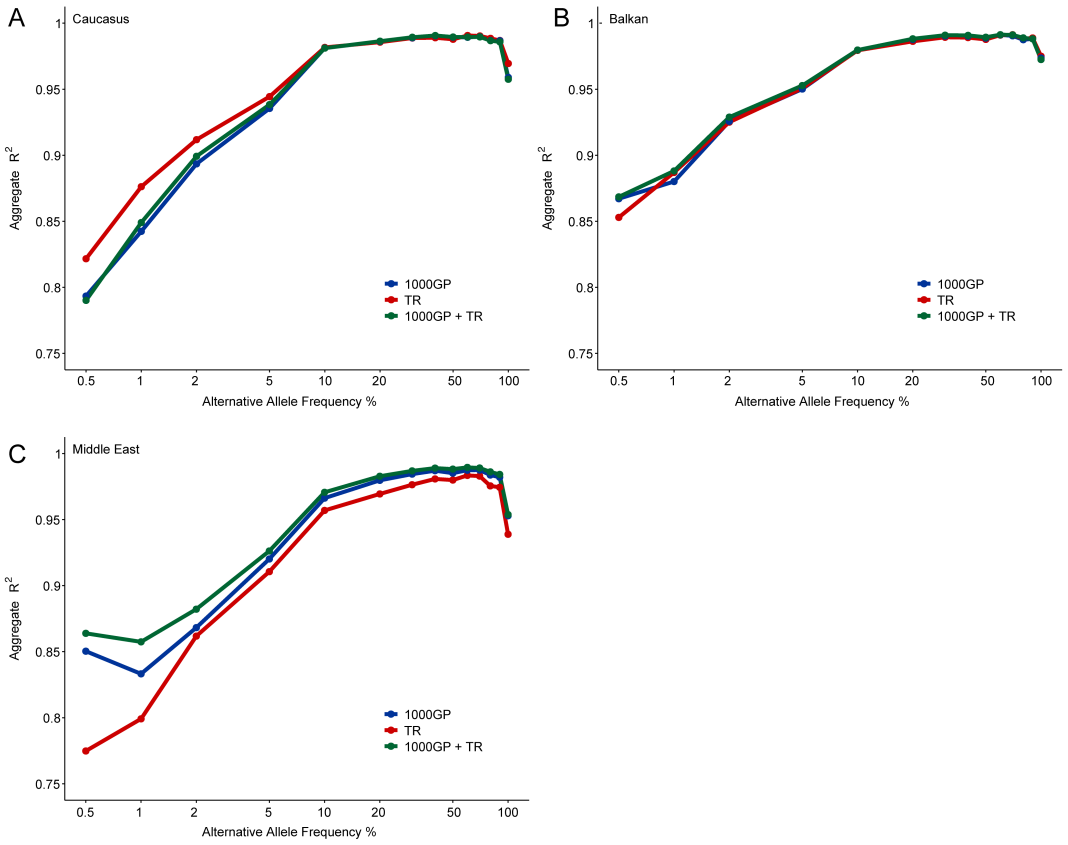



**Figure 3.43: Imputation accuracy for the neighboring populations.** Evaluation of imputation performance on chromosome 20. $R^2$ levels were calculated for genotypes called from WGS and imputed genotypes and plotted against alternative allele frequency for the three reference haplotype panels. **A** Imputation accuracy for the CAU population ($n = 13$). **B** Imputation accuracy for the BLK population ($n = 6$). **C** Imputation accuracy for the GME population ($n = 19$). Reprinted from [87].

## 3.8 Molecular findings for Mendelian and complex traits

In this section of the study, 1,981,939 WES and 72,982,375 WGS variants were identified using the data of 3,599 unrelated TR individuals. The mean target base coverage for the exons of CCDS build 15 of the WES samples was 68X with 95.33%, 93.83%, and 88.82% coverage at 8X, 10X, and 20X or more, respectively [91]. The mean depth of coverage for the WGS samples was 34X with 93.9%, 93.7%, and 93.4% coverage at 8X, 10X, and 20X or more, respectively.

A total of 48,308,918 variants were obtained following the procedure to identify high-quality samples, genotypes, and variants. 5,440 genes associated with a phenotype in OMIM were identified in the OMIM database. 3,599 TR individuals carried a total of 5,988,713 variants in 3,484 OMIM genes. Among those, 123,193 variants were previously reported in HGMD (n = 15,068) and/or ClinVar (n =119,403). 1,306 variants were reported as P or P/LP in ClinVar, while 5,844 variants were classified as DM or DM? in HGMD. These variants constituted the reported pathogenic (RP) variants in the current study. Also, following a variant annotation process to evaluate other possible pathogenic variants that were not previously reported in ClinVar or HGMD, 17,127 variants in 3,028 genes were identified as predicted pathogenic (PP) variants. Three datasets of RP and PP variants were generated, using approaches from most strict to most liberal (Figure 3.44 and Table A.7). Then, RP and PP variants were categorized based on the inheritance pattern of their associated phenotype(s). The highest percentage in all three datasets belonged to autosomal recessive (AR) phenotypes, followed by autosomal recessive/autosomal dominant (AR/AD), autosomal dominant (AD), X-linked, and Y-linked, except in Dataset 3, where the percentage of AD phenotypes was higher than that of AR/AD (Figure 3.45).

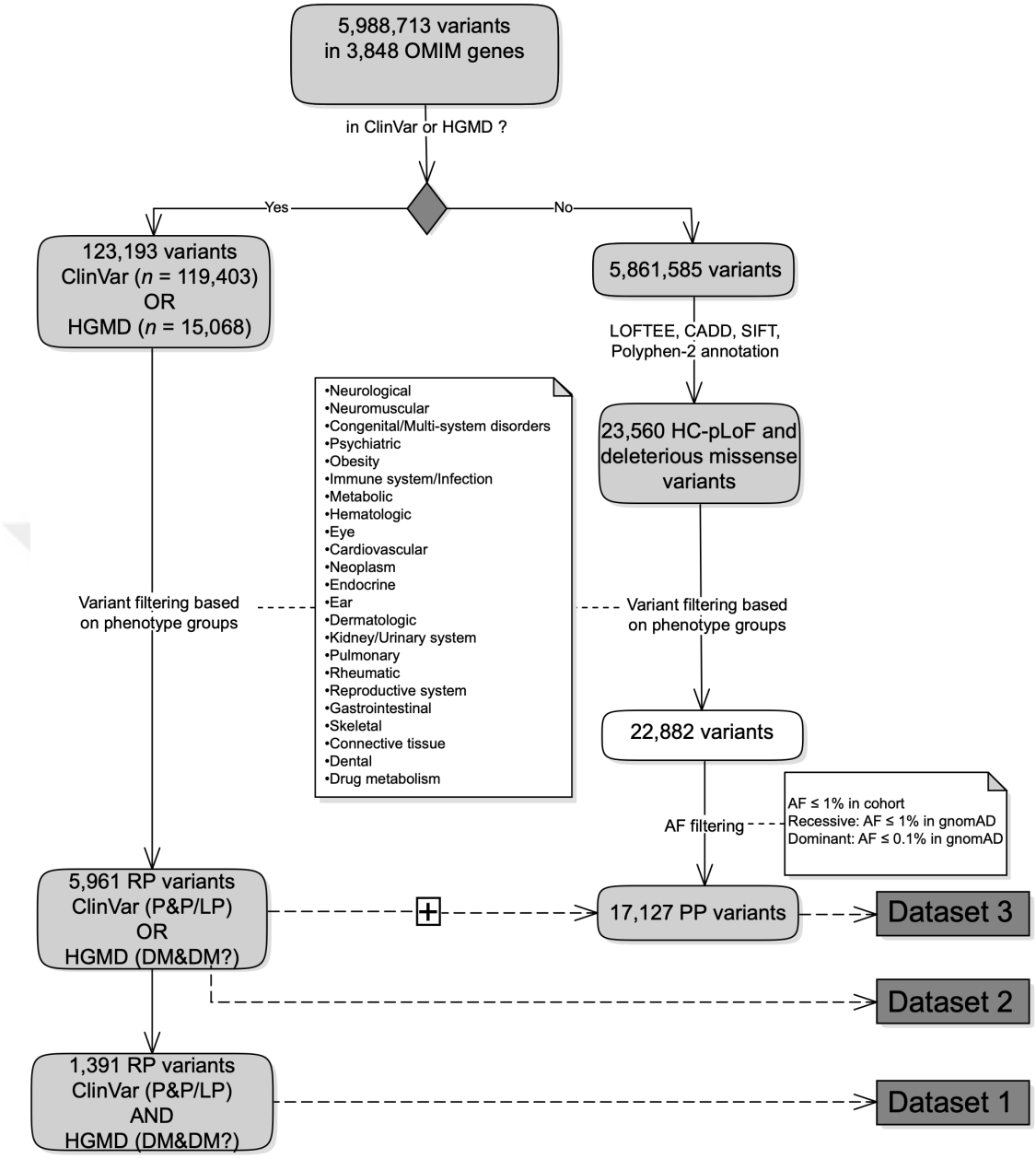



**Figure 3.44: Variant classification and selection of RP and PP variants for Datasets 1-3.** First, high quality variants that are located in OMIM genes associated with a phenotype were divided into two groups based on their presence in ClinVar and/or HGMD. Variants that are previously reported as pathogenic in both ClinVar and HGMD constituted Dataset 1. Variants that are previously reported in both ClinVar and/or HGMD constituted Dataset 2. To obtain putative pathogenic variants that are not previously reported, LOFTEE was used for HC-pLoFs; CADD, SIFT, and Polyphen-2 were used for deleterious missense variants. After filtration according to MAF, the PP variants were obtained. Dataset 3 consisted of Dataset 2 variants in addition to PP variants.

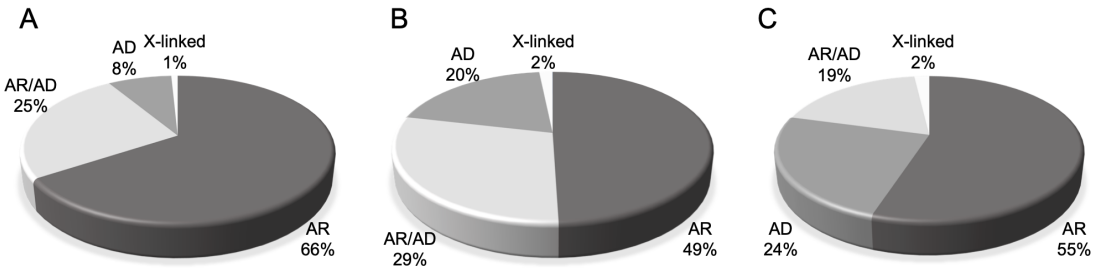



**Figure 3.45: Pie charts showing the proportion of variants associated with disorders with different inheritance patterns. A** in Dataset 1; **B** in Dataset 2; **C** in Dataset 3.

### 3.8.1 Number of variants per individual

Each individual carried 0 to 7 variants with a mean of 1.23 ± 1.11 in Dataset 1 (Table 3.18). The number of variants per individual ranged from 0 to 14 and from 1 to 31 in Dataset 2 and Dataset 3, respectively, while the means were 4.82 ± 2.27 and 12.31 ± 4.34 (Figure 3.46). Then, OMIM genes were classified into 23 disease groups, which enabled the determination of the percentage of TR individuals who possess a variant that belongs to a specific disease group. When the disease groups were sorted according to the percentage of carrying at least one Dataset 1 variant, the top three were metabolic (28.65%), eye (12.95%), and neurological (11.52%) diseases. The order was metabolic (54.1%), neurological (48.8%), and eye (36.2%) when the Dataset 2 variants were used, while it was neurological (91.68%), congenital (87.07%), and metabolic (76.47%) when Dataset 3 was employed. (Figure 3.47, Table A.8). Overall, the percentage of individuals carrying variants in AR and AR/AD inheritance categories were higher in almost all disease groups (Figure 3.48).

**Table 3.18** Number of variants per individual in Dataset 1

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total | Mean | SD | CI 95% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All OMIM genes | 1,060 | 1,299 | 773 | 330 | 109 | 23 | 4 | 1 | 3,599 | 1.23 | 1.11 | 1.19-1.26 |
| Male | 523 | 602 | 352 | 158 | 50 | 7 | 3 | 1 | 1,696 | 1.2 | 1.12 | 1.15-1.26 |
| Female | 537 | 697 | 421 | 172 | 59 | 16 | 1 | - | 1,903 | 1.25 | 1.11 | 1.19-1.29 |
| Newborn screening genes | 2,714 | 766 | 112 | 5 | 2 | - | - | - | 3,599 | 0.28 | 0.53 | 0.26-0.29 |
| ACMG recommended genes | 3,390 | 204 | 5 | - | - | - | - | - | 3,599 | 0.06 | 0.24 | 0.05-0.07 |

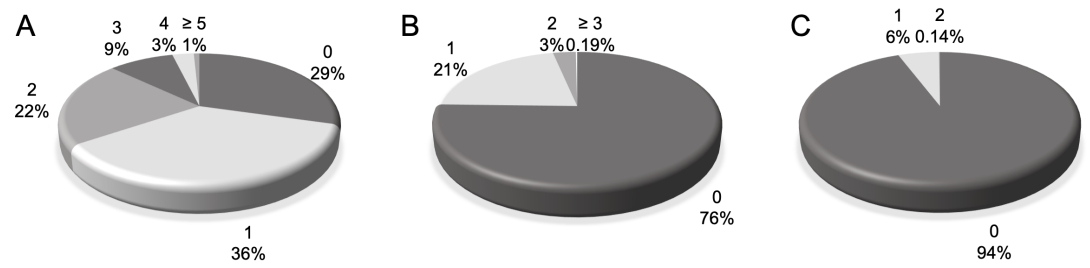



**Figure 3.46: Pie charts showing the distribution of the number of variants per individual in Dataset 1.** **A** in all OMIM genes; **B** b in NBS genes; **C** in ACMG actionable genes.

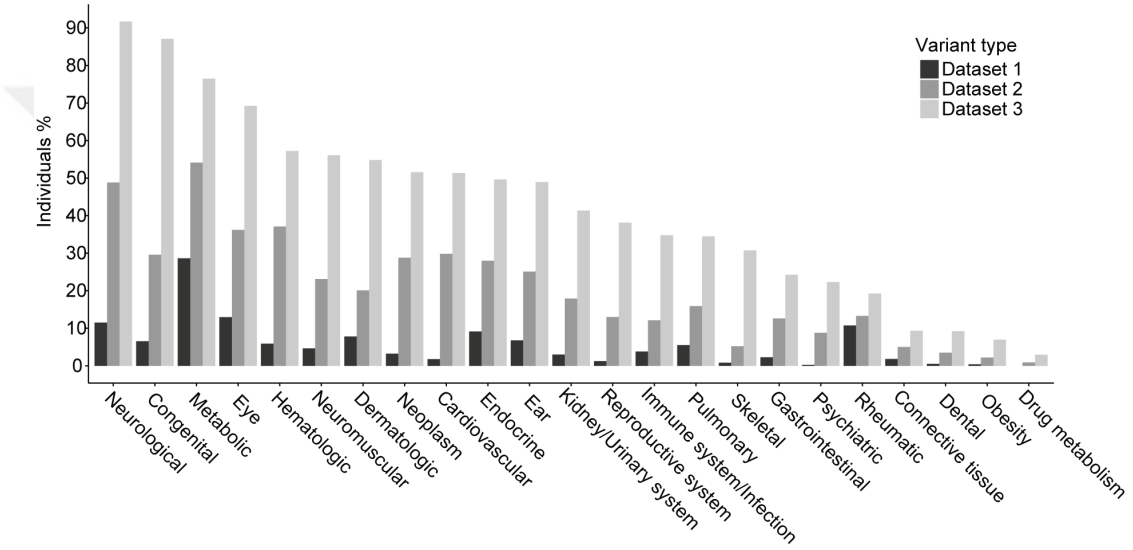



**Figure 3.47: Distribution of the percentages of individuals carrying pathogenic variants across disease groups.**The percentage of individuals harboring at least one RP or PP variant for each disease group. Colors indicate the three datasets.

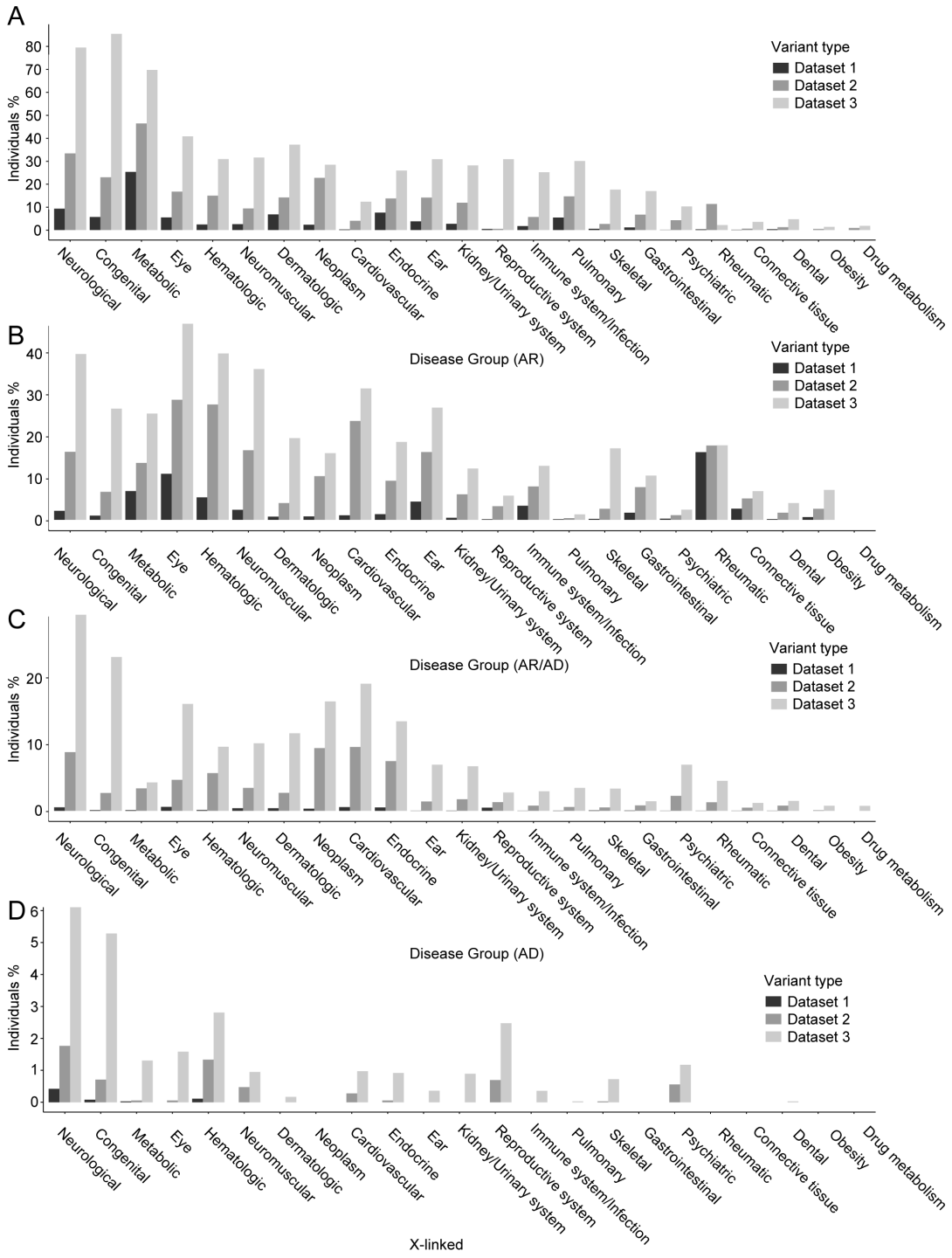



**Figure 3.48: Distribution of the percentages of individuals carrying pathogenic variants across disease groups in per inheritance category.** Bars show the percentage of individuals harboring at least one RP or PP variant for each disease group in Dataset 1-3. **A** for AR disorders **B** for AR/AD disorders **C** AD disorders **D** for X-linked disorders.

### 3.8.2 CF and GP of recessive disorders

Aggregate CFs of genes associated with recessive disorders were calculated using each dataset (Table A.9). The top 20 genes with the highest CF in Dataset 1 for each inheritance category were listed in Tables 3.19, 3.20, 3.21, 3.22. The highest CFs were observed for *PAH*, *CYP21A2*, and *CFTR* in the AR category, whereas *MEFV*, *ABCA4*, and *GJB2* took the lead in the AR/AD category. The top three genes according to CF in 14 disease groups and their associated phenotypes are listed in Table 3.23. GPs were calculated for AR and AR/AD disorders as previously described [72]. The expected frequency of affected individuals under HWE for AR, AR/AD, and X-linked disorders were also estimated (Table A.9).

Then, the estimated versus reported CF and GP were evaluated for the genes associated with AR disorders, where the CF and prevalence data were available. The highest correlation coefficient between the reported and estimated CFs and between reported prevalence and GP were yielded using Dataset 1 variants (Figure 3.49). Pairwise correlation coefficients were shown with a heatmap in Figure 3.50.

Additionally, three simple linear regressions were calculated to predict reported CF based on CF calculations using Dataset 1, Dataset 2, and Dataset 3. Significant regression equations were found: $F(1, 38) = 95.9$, $P = 6.09$ x 10-12, with an $R^2$ of 0.72) for Dataset 1, $F(1, 38) = 63.6$, $P = 1.23$ x 10-9, with an $R^2$ of 0.63 for Dataset 2, and $F(1, 38) = 59.84$, $P = 2.55$ x 10-9, with an $R^2$ of 0.61 for Dataset 3. Similarly, three simple linear regressions were calculated to predict reported prevalence based on GP calculations using Dataset 1, Dataset 2, and Dataset 3. Significant regression equations were found: $F(1, 60) = 86.65$, $P = 2.99$ x 10-13, with an $R^2$ of 0.59) for Dataset 1, $F(1, 60) = 66.56$, $P = 2.63$ x 10-11, with an $R^2$ of 0.53 for Dataset 2, and $F(1, 60) = 66.21$, $P = 2.86$ x 10-11, with an $R^2$ of 0.53 for Dataset 3. Therefore, CF and GP estimations using Dataset 1 provided the best results.

**Table 3.19** Top 20 genes according to cumulative CF in Dataset 1 for AR category

| Gene | MIM | Phenotype (Recessive) | Individual | Variant | Het | Hom | CF (1/x) | GP (1/x) |
|---|---|---|---|---|---|---|---|---|
| *PAH* | 612349 | Phenylketonuria (including hyperphenylalaninemia)* | 3,599 | 22 (29) | 68 (146) | 0 (3) | 53 (25) | 20,302 (4,368) |
| *CYP21A2* | 613815 | Congenital adrenal hyperplasia due to 21-hydroxylase deficiency | 3,599 | 5 | 146 | 1 | 25 | 3,133 |
| *CFTR* | 602421 | Cystic fibrosis | 3,599 | 41 | 120 | 1 | 30 | 6,502 |
| *PADI3* | 606755 | Uncombable hair syndrome | 3,599 | 3 | 94 | 3 | 38 | 7,165 |
| *SERPINA1* | 107400 | Alpha-1 antitrypsin deficiency | 3,599 | 8 | 68 | 0 | 53 | 17,915 |
| *MCCC2* | 609014 | 3-Methylcrotonyl-CoA carboxylase 2 deficiency | 3,599 | 2 | 57 | 5 | 63 | 17,060 |
| *MUTYH* | 604933 | Colorectal adenomatous polyposis | 3,599 | 8 | 56 | 1 | 64 | 26,734 |
| *CTH* | 607657 | Cystathioninuria | 3,599 | 1 | 50 | 1 | 72 | 20,724 |
| *CEP290* | 610142 | Bardet-Biedl syndrome 14, Leber congenital amaurosis 10, Senior-Loken syndrome 6, Meckel syndrome 4, Joubert syndrome 5 | 3,599 | 5 | 43 | 0 | 84 | 29,251 |
| *PRODH* | 606810 | Hyperprolinemia type I | 3,599 | 1 | 39 | 0 | 92 | 34,064 |

*(continued on next page)*

*Table 3.19 continued*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *ALDOB* | 612724 | Hereditary fructose intolerance | 3,599 | 6 | 38 | 0 | 95 | 55,651 |
| *MYO9A* | 604875 | Congenital presynaptic myasthenic syndrome 24 | 1,268 | 1 | 13 | 0 | 98 | 33,461 |
| *PARK2* | 602544 | Juvenile Parkinson disease type 2 | 1,189 | 7 | 12 | 0 | 99 | 45,975 |
| *ATP7B* | 606882 | Wilson disease | 3,599 | 16 | 36 | 1 | 100 | 73,387 |
| *EYS* | 612424 | Retinitis pigmentosa 25 | 3,599 | 10 | 36 | 0 | 100 | 58,346 |
| *CYP24A1* | 126065 | Infantile hypercalcemia | 3,599 | 6 | 36 | 0 | 100 | 58,876 |
| *PKHD1* | 606702 | Polycystic kidney disease 4 | 3,599 | 7 | 34 | 1 | 106 | 67,639 |
| *ACY1* | 104620 | Aminoacylase 1 deficiency | 3,599 | 4 | 33 | 0 | 109 | 62,650 |
| *GNRHR* | 138850 | Hypogonadotropic hypogonadism 7 without anosmia | 3,599 | 3 | 33 | 0 | 109 | 51,967 |
| *RNASEH2B* | 610326 | Aicardi-Goutieres syndrome 2 | 1,189 | 1 | 10 | 0 | 119 | 56,549 |

*: Numbers in parentheses were calculated including the variants associated with Hyperphenylalaninemia phenotype in HGMD

**Table 3.20** Top 20 genes according to cumulative CF in Dataset 1 for AR/AD category

| Gene | MIM | Phenotype (Recessive) | Phenotype (Dominant) | Individual | Variant | Het | Hom | CF (1/x) | GP (1/x) |
|---|---|---|---|---|---|---|---|---|---|
| *MEFV* | 608107 | Familial mediterrenean fever | Familial mediterrenean fever | 3,599 | 12 | 359 | 5 | 10 | 608 |
| *ABCA4* | 601691 | Severe early onset retinal dystrophy, Stargardt disease 1, Fundus flavimaculatus, Cone-rod dystrophy 3, Retinitis pigmentosa 19 | Age related macular degeneration 2, susceptibility to | 3,599 | 27 | 173 | 7 | 21 | 2,060 |
| *GJB2* | 121011 | Deafness, autosomal recessive 1A, | Deafness 3A, Bart-Pumphrey syndrome, Vohwinkel syndrome, Palmoplantar keratoderma with deafness, Keratitis-ichthyosis-deafness syndrome, Hystrix-like ichthyosis with deafness | 3,599 | 10 | 88 | 1 | 41 | 8,849 |
| *ABCC6* | 603234 | Pseudoxanthoma elasticum, Generalized arterial calcification of infancy 2 | Pseudoxanthoma elasticum, forme fruste | 3,599 | 13 | 60 | 0 | 60 | 21,588 |

*(continued on next page)*

*Table 3.20 continued*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *VWF* | 613160 | von Willebrand disease, types 2A, 2B, 2M, 2N, and 3 | von Willebrand disease, types 1, 2A, 2B, 2M, and 2N | 3,599 | 5 | 51 | 2 | 71 | 28,405 |
| *MPO* | 606989 | Myeloperoxidase deficiency | Alzheimer disease, susceptibility to | 3,346 | 3 | 33 | 1 | 101 | 55,287 |
| *DHTKD1* | 614984 | 2-aminoadipic 2-oxoadipic aciduria | Axonal Charcot-Marie-Tooth disease type 2Q | 3,599 | 2 | 31 | 1 | 116 | 51,813 |
| *MVK* | 251170 | Hyper-IgD syndrome, Mevalonic aciduria | Porokeratosis 3 | 3,346 | 2 | 26 | 0 | 129 | 68,791 |
| *ATM* | 607585 | Ataxia-telangiectasia | Breast cancer, susceptibility to | 1,189 | 7 | 8 | 0 | 149 | 343,001 |
| *GBA* | 606463 | Gaucher disease types I,II, III, and IIIC | Late onset Parkinson disease, susceptibility to, Lewy body dementia, susceptibility to | 3,599 | 6 | 24 | 0 | 150 | 121,909 |
| *POLG* | 174763 | Mitochondrial DNA depletion syndrome 4A, 4B (MNGIE, Alpers types, SANDO and SCAE), Progressive external ophthalmoplegia 1 | Progressive external ophthalmoplegia 1 | 3,599 | 6 | 21 | 0 | 171 | 62,519 |
| *SPINK1* | 167790 | Tropical calcific pancreatitis | Tropical calcific pancreatitis, Hereditary pancreatitis | 3,599 | 2 | 21 | 0 | 171 | 128,564 |
| *SLC7A9* | 604144 | Cystinuria | Cystinuria | 3,599 | 2 | 20 | 1 | 180 | 159,419 |

*(continued on next page)*

*Table 3.20 continued*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *ABCB4* | 171060 | Gallbladder disease 1, Intrahepatic cholestasis of pregnancy 3, Cholestasis, progressive familial intrahepatic 3 | Gallbladder disease 1, Intrahepatic cholestasis of pregnancy 3 | 3,599 | 2 | 17 | 0 | 212 | 226,250 |
| *LDLR* | 606945 | Familial hypercholesterolemia 1 | Familial hypercholesterolemia 1 | 3,599 | 11 | 16 | 0 | 225 | 362,316 |
| *NR2E3* | 604485 | Enhanced S-cone syndrome, Retinitis pigmentosa 37 | Retinitis pigmentosa 37 | 3,599 | 3 | 16 | 0 | 225 | 252,738 |
| *SLC3A1* | 104614 | Cystinuria | Cystinuria | 3,599 | 5 | 16 | 1 | 225 | 283,121 |
| *HBB* | 141900 | Beta Thalassemia, Sickle cell anemia | Erythrocytosis 6, Heinz body anemia, Delta-beta thalassemia, Hereditary persistence of fetal hemoglobin, Beta thalassemia inclusion-body, Methmoglobinemia, beta type | 3,599 | 11 | 15 | 0 | 240 | 414,490 |
| *MC4R* | 155541 | Obesity | Obesity | 2,682 | 2 | 11 | 0 | 244 | 279,345 |

*(continued on next page)*

*Table 3.20 continued*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *F5* | 612309 | Factor V deficiency, Budd-Chiari syndrome | Recurrent pregnancy loss 1, susceptibility to, Thrombophilia due to activated protein C resistance, Thrombophilia due to factor V Leiden, susceptibility to | 3,599 | 1 | 13 | 0 | 277 | 306,575 |

**Table 3.21** Top 20 genes according to cumulative CF in Dataset 1 for AD category

| Gene | MIM | Phenotype (Dominant) | Individual | Variant | Het | Hom | CF (1/x) |
|---|---|---|---|---|---|---|---|
| *FLG* | 135940 | Ichthyosis vulgaris | 3,599 | 5 | 19 | 0 | 189 |
| *TGIF1* | 602630 | Holoprosencephaly 4 | 1,189 | 1 | 6 | 0 | 198 |
| *INSL3* | 146738 | Cryptorchidism | 3,599 | 1 | 17 | 0 | 212 |
| *AGBL1* | 615496 | Fuchs endothelial dystrophy | 3,599 | 2 | 13 | 0 | 277 |
| *PROK2* | 607002 | Hypogonadotropic hypogonadism 4 | 3,599 | 1 | 12 | 0 | 300 |
| *TARDBP* | 605078 | Frontotemporal lobar degeneration, Amyotrophic lateral sclerosis 10 | 1,268 | 1 | 4 | 0 | 317 |
| *PROKR2* | 607123 | Hypogonadotropic hypogonadism 3 | 3,599 | 3 | 9 | 0 | 400 |
| *OPTN* | 602432 | Open angle glaucoma 1 E, Amyotrophic lateral sclerosis 12 | 1,268 | 1 | 3 | 1 | 423 |
| *BRIP1* | 605882 | Fanconi anemia, complementation group J, Breast cancer, early-onset, susceptibility to | 3,599 | 4 | 8 | 0 | 450 |
| *MEF2A* | 600660 | Coronary artery disease 1 | 3,599 | 1 | 8 | 0 | 450 |
| *GATA4* | 600576 | Testicular anomalies, Tetralogy of Fallot, Atrioventricular septal defect 4, Atrial septal defect 2, Ventricular septal defect 1 | 3,599 | 2 | 6 | 0 | 600 |

*(continued on next page)*

*Table 3.21 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *GDF3* | 606522 | Isolated microphthalmia 7, Microphthalmia with coloboma 6, Klippel-Feil syndrome 3 | 3,599 | 2 | 5 | 0 | 720 |
| *NANOS1* | 608226 | Spermatogenic failure 12 | 3,599 | 1 | 4 | 0 | 900 |
| *KHL10* | 608778 | Spermatogenic failure 11 | 3,599 | 1 | 4 | 0 | 900 |
| *ZNF687* | 610568 | Paget disease of bone 6 | 3,599 | 1 | 4 | 0 | 900 |
| *MIB1* | 608677 | Left ventricular noncompaction 7 | 3,599 | 1 | 4 | 0 | 900 |
| *MYOC* | 601652 | Primary open angle glaucoma 1A | 3,599 | 2 | 4 | 0 | 900 |
| *TNNT2* | 191045 | Familial restrictive cardiomyopathy 3, Hypertrophic cardiomyopathy 2, Left ventricular noncompaction 6, Dilated cardiomyopathy 1D | 3,599 | 2 | 4 | 0 | 900 |
| *RERE* | 605226 | Neurodevelopmental disorder with or without anomalies of the brain, eye, or heart | 1,189 | 1 | 1 | 0 | 1,189 |
| *TENM4* | 610084 | Hereditary essential tremor | 1,189 | 1 | 1 | 0 | 1,189 |

**Table 3.22** Top 20 genes according to cumulative CF in Dataset 1 for X-linked category

| Gene | MIM | Phenotype (X-linked)* | Individual (M/F) | Variant | Het | Hemi | CF (1/x) | HWE (1/x)** |
|---|---|---|---|---|---|---|---|---|
| *ZC4H2* | 300897 | Wieacker-Wolff syndrome | 367/822 | 1 | 1 | 0 | 1,189 | 2,376 |
| *PQBP1* | 300463 | Renpenning syndrome | 367/822 | 1 | 1 | 0 | 1,189 | 2,376 |
| *PGK1* | 311800 | Phosphoglycerate kinase 1 deficiency | 367/822 | 1 | 1 | 0 | 1,189 | 2,376 |
| *CUL4B* | 300304 | Mental retardation, syndromic 15 | 367/822 | 1 | 1 | 0 | 1,189 | 2,376 |
| *F8* | 300841 | Hemophilia A | 1,711/1,888 | 2 | 2 | 1 | 1,800 | 3,597 |
| *OTC* | 300461 | Ornithine transcarbamylase deficiency | 1,711/1,888 | 1 | 1 | 0 | 3,599 | 7,196 |
| *WAS* | 300392 | Wiskott-Aldrich syndrome | 1,711/1,888 | 1 | 1 | 0 | 3,599 | 7,196 |

*: All genes associated with X-linked phenotypes are included.

**: Expected number of hemi/homozygotes were calculated using Hardy-Weinberg equilibrium for X-linked disorders.

**Table 3.23** Top three genes according to cumulative CF in Dataset 1 for each disease group

| Disease group | Gene | Phenotype (AR) | Phenotype (AD) | Het | Hom | CF (1/x) | GP (1/x) |
|---|---|---|---|---|---|---|---|
| Cardiovascular | *MEF2A* | | Coronary artery disease | 8 | - | 450 | - |
| | *MYH7* | Myosin storage myopathy | Dilated cardiomyopathy 1S, Left ventricular noncompaction 5, Hypertrophic cardiomyopathy 1, Laing distal myopathy, Myosin storage myopathy, Myopathic type scapuloperoneal syndrome | 8 | - | 450 | 1,233,600 |
| | *GATA4* | | Atrioventricular septal defect 4, Atrial septal defect 2, Ventricular septal defect 1, Tetralogy of Fallot | 6 | - | 600 | - |
| Congenital | *DHCR7* | Smith-Lemli-Opitz syndrome | | 7 | - | 181 | 189,156 |

*(continued on next page)*

*Table 3.23 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *POLR1C* | Treacher Collins syndrome 3, Hypomyelinating leukodystrophy 11 | | 6 | - | 198 | 267,072 |
| | *ESCO2* | Roberts syndrome, SC phocomelia syndrome | | 6 | - | 211 | 207,461 |
| Endocrine | *CYP21A2* | Congenital adrenal hyperplasia due to 21-hydroxylase deficiency | | 146 | 1 | 25 | 3,133 |
| | *CYP24A1* | Infantile hypercalcemia 1 | | 36 | - | 100 | 58,876 |
| | *GNRHR* | Hypogonadotropic hypogonadism 7 without anosmia | | 33 | - | 109 | 51,967 |
| Eye | *ABCA4* | Early-onset severe retinal dystrophy, Stargardt disease 1, Fundus flavimaculatus, Cone-rod dystrophy 3, Retinitis pigmentosa 19 | Age-related macular degeneration 2 | 173 | 7 | 21 | 2,060 |
| | *CEP290* | Leber congenital amaurosis 10 | | 43 | - | 84 | 29,251 |
| | *EYS* | Retinitis pigmentosa 25 | | 36 | - | 100 | 58,346 |
| Gastrointestinal | *KRT8* | Cryptogenic cirrhosis | | 27 | 1 | 133 | 71,072 |

*(continued on next page)*

*Table 3.23 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *SPINK1* | Tropical calcific pancreatitis | Hereditary pancreatitis, Tropical calcific pancreatitis | 21 | - | 171 | 128,564 |
| | *ABCB4* | Progressive familial intrahepatic cholestasis, Gallbladder disease 1, Intrahepatic cholestasis of pregnancy 3 | Gallbladder disease 1, Intrahepatic cholestasis of pregnancy 3 | 17 | - | 212 | 226,250 |
| Hematologic | *VWF* | von Willebrand disease types 2A, 2B, 2M, 2N, and 3 | von Willebrand disease types 1, 2A, 2B, 2M, and 2N | 51 | 2 | 71 | 28,405 |
| | *HBB* | Beta-thalassemia, Sickle cell anemia | Hereditary persistence of fetal hemoglobin, Heinz body anemia, Delta-beta thalassemia, Beta type methmoglobinemia, Inclusion body beta-thalassemia, Erythrocytosis 6 | 15 | - | 240 | 414,490 |

*(continued on next page)*

*Table 3.23 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *F5* | Factor V deficiency | Thrombophilia due to activated protein C resistance, Susceptibility to Factor V Leiden thrombophilia, Susceptibility to recurrent pregnancy loss 1 | 13 | - | 277 | 306,575 |
| Immune system/Infectious | *MPO* | Myeloperoxidase deficiency | | 33 | 1 | 101 | 55,287 |
| | *MVK* | Hyper-IgD syndrome, Mevalonic aciduria | Porokeratosis 3 | 26 | - | 129 | 68,791 |
| | *C8B* | Type II C8 deficiency | | 24 | 1 | 139 | 80,982 |
| Kidney/Urinary system | *PKHD1* | Polycystic kidney disease 4, with or without hepatic disease | | 34 | 1 | 106 | 67,639 |
| | *SLC12A3* | Gitelman syndrome | | 15 | - | 240 | 398,548 |
| | *NPHS2* | Type 2 nephrotic syndrome | | 9 | - | 400 | 893,297 |
| Metabolic | *PAH* | Phenylketonuria, Hyperphenylalaninemia, non-PKU mild | | 146 | 3 | 25 | 4,368 |

*(continued on next page)*

*Table 3.23 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *MCCC2* | 3-Methylcrotonyl-CoA carboxylase 2 deficiency | | 57 | 5 | 63 | 17,060 |
| | *CTH* | Cystathioninuria | | 50 | 1 | 72 | 20,724 |
| Neoplasm | *MUTYH* | Multiple colorectal adenomas | | 56 | 1 | 64 | 26,734 |
| | *NTHL1* | Familial adenomatous polyposis 3 | | 11 | - | 327 | 428,192 |
| | *BRCA2* | Fanconi anemia, complementation group D1, Glioblastoma 3, Medulloblastoma | Familial breast-ovarian cancer 2, Prostate cancer, Medullablastoma, Wilms tumor, | 7 | - | 514 | 1,786,593 |
| Neurological | *PARK2* | Juvenile Parkinson disease type 2 | | 12 | - | 99 | 45,975 |
| | *RNASEH2B* | Aicardi-Goutieres syndrome 2 | | 10 | - | 119 | 56,549 |
| | *ATM* | Ataxia-telangiectasia | | 8 | - | 149 | 343,001 |
| Neuromuscular | *MYO9A* | Congenital myasthenic syndrome 24 | | 13 | - | 98 | 33,461 |
| | *CAPN3* | Limb-girdle muscular dystrophy 1 | Limb-girdle muscular dystrophy 4 | 4 | - | 317 | 565,488 |

*(continued on next page)*

*Table 3.23 continued*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *RYR1* | Central core disease, Minicore myopathy with external ophthalmoplegia, Congenital neuromuscular disease with uniform type 1 fiber | King-Denborough syndrome, Malignant hyperthermia susceptibility 1, Central core disease, Minicore myopathy with external ophthalmoplegia, Congenital neuromuscular disease with uniform type 1 fiber | 4 | - | 317 | 514,080 |
| Pulmonary | *CFTR* | Cystic fibrosis | | 120 | 1 | 30 | 6,502 |
| | *SERPINA1* | Alpha-1 antitrypsin deficiency | | 68 | - | 53 | 17,915 |
| | *CCDC103* | Primary ciliary dyskinesia 17 | | 7 | - | 514 | 1,057,372 |
| Rheumatic* | *MEFV* | Familial Mediterranean fever | Familial Mediterranean fever | 359 | 5 | 10 | 608 |
| | *CECR1* | Vasculitis, autoinflammation, immunodeficiency, and hematologic defects syndrome | | 10 | - | 360 | 672,873 |

*: Rest of the genes in the rheumatic disease group had only one heterozygote.

Although Dataset 1 was a significant indicator for reported CF and prevalence, and revealed highest correlations, results varied among diseases. With the variants in Dataset 1, the difference between estimated and reported CF was not larger than other sets for *CFTR*, *PAH*, *GJB2, CYP21A2*, *SERPINA1, PKHD1*, *MUTYH, ATP7B*, *ALDOB*, *CTH*, *MCCC1,* and *MCCC2*. Dataset 2 resulted in better CF estimations for *ACADM*, *DHCR7*, *SLC12A3*, *BTD*, *GAA, SLC22A5*, *PYGM*, and *GALC* while Dataset 3 was the best for *MEFV*, *GBA*, *GLDC*, *G6PC*, and *HEXA*. Similarly, employing the Dataset 1 resulted in the highest correlation coefficient between the estimated prevalence and GP. GP for *MEFV*, *CYP21A2*, *GAA*, *MCCC1*, *MCCC2*, *PMM2*, *GNPTAB*, *EYS*, and *MMACHC* were closer to the reported prevalence when Dataset 1 variants were used; however, reported prevalence of *CFTR*, *PKHD1*, *ATP7B*, *SLC12A3*, *AGXT*, *PYGM*, *SLC26A4*, and *SMPD1* were closer to the Dataset2 estimations while reported prevalence of *GJB2*, *SERPINA1*,*ACADM*, *DHCR7*, *CTH*, *BTD*, *GLDC*, *SLCA22A5*, and *ACADS* were closer to the Dataset 3 estimations. (Table 3.24).

**Table 3.24** Reported and estimated CFs and prevalence of recessive phenotypes

| | | Reported | | | | Dataset 1 | | Dataset 2 | | Dataset 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Phenotype | Gene | CF | [Ref.] | Prevalence | [Ref.] | CF | GP | CF | GP | CF | GP |
| Familial Mediterranean fever | *MEFV* | 5 | [145] | 1,075 | [146] | 10 | 608 | 9 | 518 | 9 | 515 |
| Cystic Fibrosis | *CFTR* | 25 | [147] | 3,333 | [148] | 30 | 6,502 | 13 | 1,219 | 13 | 1,194 |
| Phenylketonuria | *PAH* | 26 | [149] | 5,988 | [150] | 25 | 4,368 | 19 | 1,078 | 19 | 1,078 |
| GJB2-related congenital deafness | *GJB2* | 43 | [85] | 7,143 | [149] | 41 | 8,849 | 30 | 5,475 | 30 | 5,475 |
| Congenital adrenal hyperplasia due to 21-hydroxylase deficiency | *CYP21A2* | 50 | [151] | 10,000 | [149] | 25 | 3,133 | 22 | 2,579 | 20 | 2,353 |
| Alpha-1-antitrypsin deficiency | *SERPINA1* | 70 | [152] | 2,632 | [152] | 53 | 17,915 | 42 | 12,032 | 39 | 10,692 |
| Polycystic kidney disease | *PKHD1* | 70 | [149] | 20,000 | [153] | 106 | 67,639 | 34 | 7,668 | 27 | 5,377 |
| Gaucher Disease | *GBA* | 77 | [85] | 88,495 | [149] | 150 | 106,729 | 97 | 17,721 | 97 | 17,721 |
| Colorectal adenomatous polyposis, | *MUTYH* | 78 | [154] | 20,000 | [149] | 64 | 26,734 | 41 | 11,816 | 41 | 11,816 |
| Wilson disease | *ATP7B* | 90 | [155] | 30,000 | [149] | 100 | 73,387 | 42 | 12,560 | 37 | 10,239 |
| Hereditary fructose intolerance | *ALDOB* | 97 | [149] | 38,000 | [149] | 95 | 55,651 | 65 | 28,235 | 56 | 21,561 |
| Medium chain acyl-CoA dehydrogenase deficiency | *ACADM* | 100 | [149] | 18,868 | [149] | 129 | 96,303 | 92 | 54,653 | 82 | 44,703 |
| Smith-Lemli-Opitz syndrome | *DHCR7* | 100 | [149] | 40,000 | [149] | 181 | 189,156 | 106 | 72,772 | 91 | 54,929 |
| Gitelman syndrome | *SLC12A3* | 100 | [156] | 40,000 | [156] | 240 | 398,548 | 61 | 27,692 | 49 | 18,386 |

*(continued on next page)*

*Table 3.24 continued*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cystathioninuria | *CTH* | 120 | [157] | 14,000 | [49] | 72 | 20,724 | 72 | 20,724 | 68 | 19,507 |
| Biotinidase deficiency | *BTD* | 120 | [149] | 61,067 | [149] | 156 | 161,406 | 109 | 84,659 | 100 | 72,261 |
| Pompe disease | *GAA* | 132 | [85] | 283,000 | [158] | 189 | 222,366 | 112 | 82,766 | 92 | 58,478 |
| Glycine encephalopathy | *GLDC* | 156 | [149] | 95,000 | [149] | 514 | 1,523,859 | 225 | 357,319 | 95 | 68,806 |
| Systemic primary carnitine deficiency | *SLC22A5* | 182 | [85] | 70,000 | [149] | 277 | 503,021 | 200 | 281,583 | 138 | 141,949 |
| 3-methylcrotonyl-CoA carboxylase deficiency | *MCCC1, MCCC2* | 186 | [68] | 30,000 | [149] | 57 | 15,225 | 46 | 11,480 | 40 | 10,470 |
| Primary hyperoxaluria type I | *AGXT* | 195 | [159] | 151,887 | [159] | 257 | 392,509 | 133 | 119,656 | 116 | 93,861 |
| Short chain acyl-CoA dehydrogenase deficiency | *ACADS* | 229 | [85] | 35,000 | [149] | 327 | 588,764 | 88 | 42,890 | 78 | 36,206 |
| McArdle disease | *PYGM* | 243 | [85] | 170,000 | [149] | 720 | 3,238,200 | 75 | 33,170 | 51 | 17,015 |
| Glycogen storage disease Ia | *G6PC* | 261 | [85] | 100,000 | [149] | 360 | 569,354 | 360 | 569,354 | 257 | 359,800 |
| Congenital disorder of glycosylation, type Ia | *PMM2* | 263 | [160] | 286,726 | [160] | 171 | 140,791 | 144 | 112,145 | 133 | 100,604 |
| Galactosemia | *GALT* | 288 | [85] | 48,000 | [149] | 189 | 191,185 | 144 | 127,300 | 138 | 119,656 |
| Tay-Sachs disease | *HEXA* | 300 | [161] | 320,000 | [153] | 400 | 1,057,372 | 360 | 878,156 | 277 | 545,381 |
| Mucolipidosis types II and III | *GNPTAB* | 316 | [149] | 454,545 | [149] | 360 | 518,112 | 225 | 248,978 | 129 | 104,215 |
| Tyrosinemia type I | *FAH* | 465 | [85] | 120,000 | [149] | 400 | 996,369 | 300 | 609,544 | 257 | 458,506 |
| Pendred syndrome | *SLC26A4* | 465 | [85] | 100,000 | [153] | 277 | 539,700 | 62 | 28,816 | 55 | 23,213 |
| Fanconi Anemia type C | *FANCC* | 488 | [85] | 1,400,000 | [149, 162] | 1,200 | 7,401,601 | 400 | 977,570 | 400 | 977,570 |
| Metachromatic Leurkodystrophy | *ARSA* | 527 | [85] | 160,000 | [156] | 1,189 | 5,654,884 | 396 | 942,481 | 396 | 942,481 |

*(continued on next page)*

*Table 3.24 continued*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Niemann-Pick disease types A and B | *SMPD1* | 555 | [85] | 250,000 | [153] | 360 | 785,018 | 63 | 24,485 | 57 | 20,892 |
| VLCAD deficiency | *ACADVL* | 576 | [85] | 100,000 | [149] | 327 | 719,600 | 144 | 151,495 | 129 | 122,485 |
| Zellweger syndrome, PEX1 related | *PEX1* | 581 | [85] | - | - | 720 | 3,454,080 | 180 | 200,819 | 180 | 200,819 |
| Krabbe disease | *GALC* | 586 | [85] | 100,000 | [149] | 1,200 | 7,401,601 | 156 | 181,158 | 129 | 124,247 |
| Cystinosis | *CTNS* | 608 | [85] | 200,000 | [149] | 720 | 3,238,200 | 514 | 1,786,593 | 400 | 1,126,331 |
| Alkaptonuria | *HGD* | 632 | [85] | 1,000,000 | [149] | 720 | 3,047,718 | 400 | 1,079,400 | 240 | 417,832 |
| Isovaleric acidemia | *IVD* | 832 | [85] | 526,000 | [156] | 514 | 1,363,453 | 360 | 797,095 | 360 | 797,095 |
| Niemann-Pick disease type C | *NPC1, NPC2* | 3,151 | [163] | 150,000 | [149] | 327 | 729,735 | 144 | 141,175 | 97 | 89,716 |
| Congenital sucrase-isomaltase deficiency | *SI* | - | - | 5,000 | [153] | 327 | 609,544 | 129 | 118,833 | 52 | 21,079 |
| Retinitis pigmentosa 25 | *EYS* | - | - | 40,000 | [149] | 100 | 58,346 | 46 | 13,728 | 36 | 8,755 |
| Achromatopsia 3 | *CNGB3* | - | - | 50,000 | [153] | 450 | 1,328,492 | 360 | 878,156 | 138 | 137,430 |
| Citrullinemia | *ASS1* | - | - | 57,000 | [149] | 120 | 94,892 | 62 | 24,613 | 61 | 23,931 |
| Usher syndrome, type 2A | *USH2A* | - | - | 48,387 | [149] | 120 | 113,621 | 28 | 5,730 | 20 | 5,566 |
| Congenital presynaptic myasthenic syndrome 24 | *MYO9A* | - | - | 80,000 | [149] | 98 | 33,461 | 98 | 34,980 | 55 | 16,817 |
| Spastic paraplegia 11 | *SPG11* | - | - | 80,000 | [149] | 595 | 1,884,961 | 198 | 198,379 | 198 | 49,241 |
| Type VII oculocutenous albinism | *C10orf11* | - | - | 100,000 | [149] | 450 | 809,550 | 327 | 545,381 | 277 | 431,760 |
| Methylmalonic aciduria, mut(0) type | *MMUT* | - | - | 102,564 | [149] | 450 | 1,400,303 | 164 | 178,045 | 88 | 52,869 |
| Familial adenomatous polyposis 3 | *NTHL1* | - | - | 114,770 | [149] | 327 | 428,192 | 164 | 160,904 | 150 | 140,410 |

*(continued on next page)*

*Table 3.24 continued*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Methylmalonic aciduria and homocystinuria, cblC type | *MMACHC* | - | - | 200,000 | [149] | 189 | 199,274 | 124 | 95,769 | 120 | 90,738 |
| Congenital ichthyosis | *TGM1* | - | - | 200,000 | [149] | 277 | 479,733 | 189 | 246,720 | 124 | 113,125 |
| Chronic granulomatous disease due to deficiency of NCF-1 | *NCF1* | - | - | 227,273 | [149] | 418 | 878,095 | 48 | 9,570 | 48 | 8,417 |
| Glutaric acidemia IIC | *ETFDH* | - | - | 250,000 | [149] | 238 | 838,229 | 149 | 424,406 | 91 | 248,676 |
| Progressive myoclonic epilepsy 2B (Lafora) | *NHLRC1* | - | - | 250,000 | [149] | 396 | 628,320 | 396 | 628,320 | 396 | 628,320 |
| Factor VII deficiency | *F7* | - | - | 500,000 | [156] | 400 | 785,018 | 120 | 97,757 | 95 | 62,878 |
| Achromatopsia 2 | *CNGA3* | - | - | 600,000 | [149] | 450 | 1,400,303 | 92 | 48,971 | 84 | 42,089 |
| Plasminogen deficiency, type I | *PLG* | - | - | 625,000 | [153] | 300 | 450,532 | 106 | 76,193 | 62 | 28,158 |
| Pyridoxine-dependent epilepsy | *ALDH7A1* | - | - | 687,000 | [156] | 396 | 942,481 | 297 | 565,488 | 170 | 201,960 |
| Cerebrotendinous xanthomatosis | *CYP27A1* | - | - | 769,231 | [149] | 595 | 1,884,961 | 396 | 796,434 | 149 | 145,925 |
| Beta-ureidopropionase deficiency | *UPB1* | - | - | 1,000,000 | [153] | 450 | 908,968 | 225 | 301,228 | 164 | 178,045 |
| Peeling skin syndrome 2 | *TGM5* | - | - | 1,000,000 | [153] | 450 | 996,369 | 360 | 719,600 | 225 | 334,266 |
| Hermansky-Pudlak syndrome 1 | *HPS1* | - | - | 1,333,333 | [153] | 400 | 893,297 | 189 | 216,783 | 112 | 89,794 |

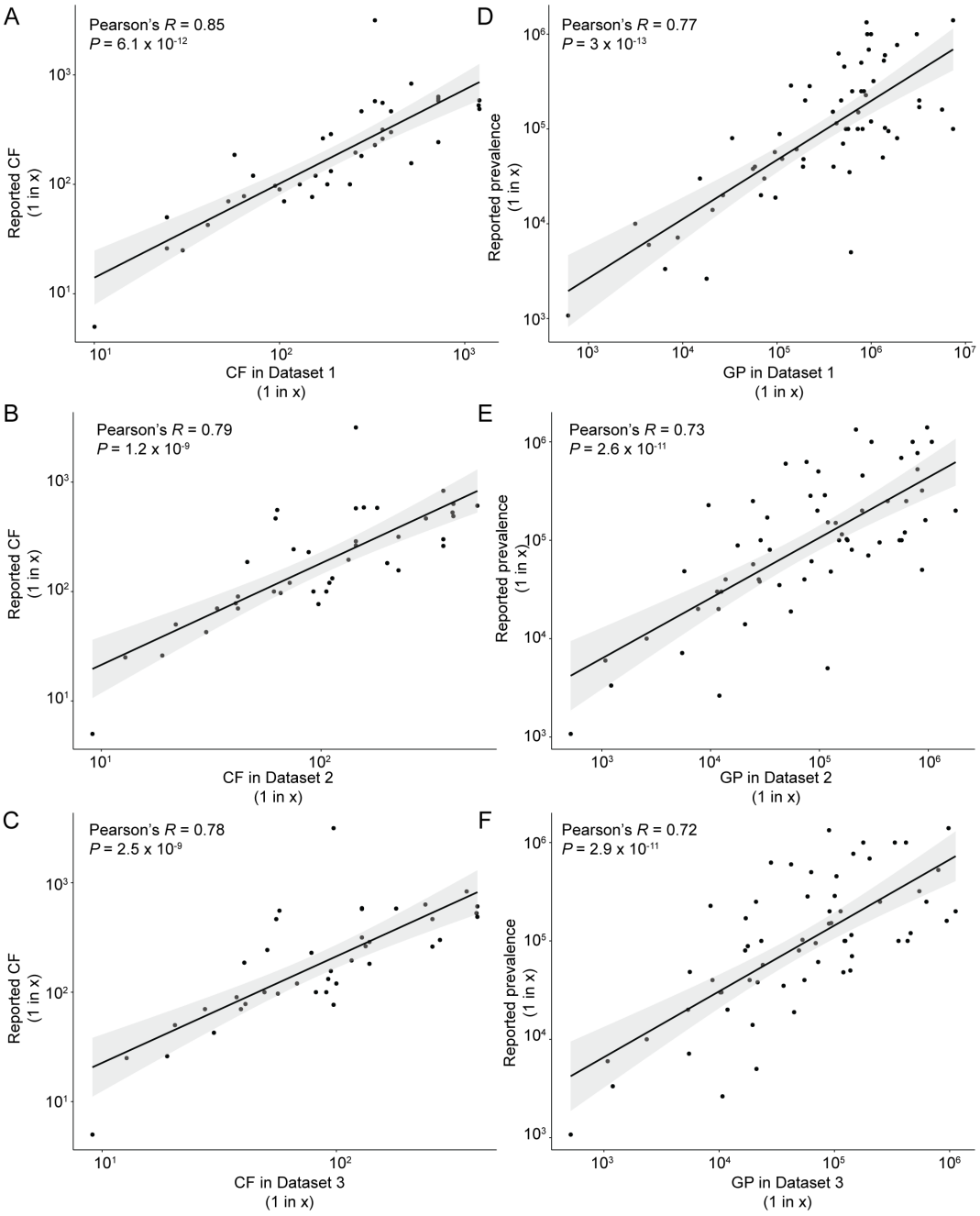



**Figure 3.49: Correlation of estimated and reported disease frequencies.** Correlation of reported and estimated CFs in **A** Dataset 1; **B** Dataset 2; **C** Dataset3. Correlation of reported and estimated disease prevalence in **D** Dataset 1; **E** Dataset 2; **F** Dataset 3.

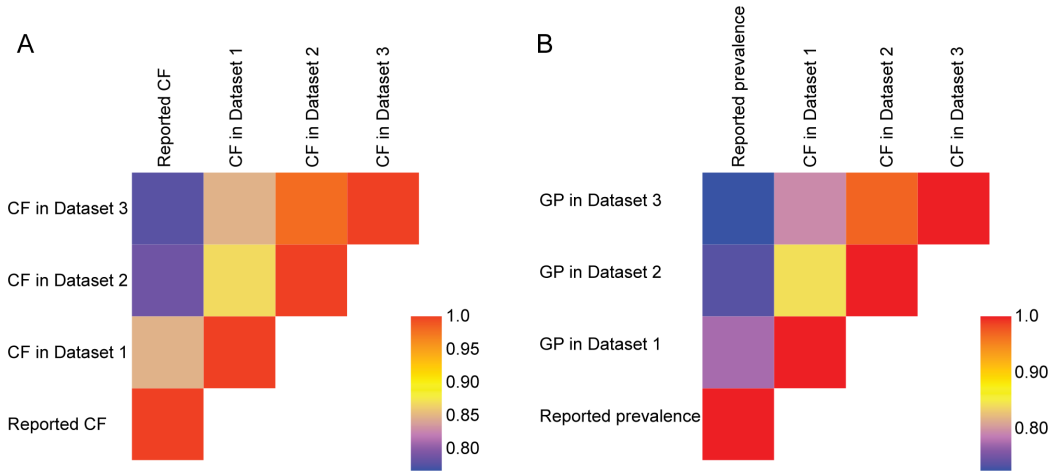



**Figure 3.50: Heatmap for pairwise correlations of Reported disease frequencies and Datasets 1-3. A** Pairwise correlation coefficients of reported and estimated CFs in Datasets 1-3; **B** Pairwise correlation coefficients of reported prevalence and estimated GPs in Datasets 1-3.

### 3.8.3 NBS and ACMG recommended actionable genes

The analyses included all variants located in genes associated with an OMIM phenotype; therefore, there was an opportunity to assess the status of NBS genes in the TR population. The gene list of Advisory Committee on Heritable Disorders in Newborns and Children, which consists of 56 genes, were used for this purpose as well as 13 genes that are associated with a childhood-onset metabolic disease such as arginase deficiency (Table 2.8) [70]. The proportion of individuals who had at least one variant was 25% when the Dataset 1 variants were used. The proportions went up to 43% and 55% when Dataset 2 and Dataset 3 variants were used. The number of variants in NBS genes per individual ranged from 0 to 4, 0 to 5, and 0 to 6 for Datasets 1-3, respectively, while the means were 0.28 $\pm$ 0.53, 0.57 $\pm$ 0.76, and 0.8 $\pm$ 0.89 (Table A.8). The highest CFs were observed for *PAH*, *CYP21A2*, and *CFTR* (Table A.10). ACMG has recently expanded the actionable gene list from 59 to 73 genes [71]. Accordingly, the carrier status of these 73 ACMG recommended actionable genes were evaluated in the cohort (Table 2.9). The proportion of individuals having at least one pathogenic variant in ACMG recommended actionable genes was estimated to be 5.81, 26.27%, and 33.59% for Datasets 1-3, respectively. The number of variants in ACMG genes

per individual varied from 0 to 3, 0 to 4, and 0 to 5 for Datasets 1-3, respectively, while the means were $0.06 \pm 0.24$, $0.31 \pm 0.56$, and $0.41 \pm 0.65$. (Table A.8). The highest CFs were observed for *MUTYH*, *ATP7B*, and *BTD* in Dataset 1, whilst *TTN* and *RYR1* had higher CF than *BTD* in Datasets 2 and 3, respectively (Table A.11).

# Chapter 4

# Discussion

## 4.1 The genetic structure of the TR population

This study is the most comprehensive report that describes the fine-scale genetic structure of the TR population thus far. Consistent with the previous findings in small-scale representations of the TR population, the results of the current study indicated a remarkable level of admixture, which reflects the position of Turkey at the crossroads of many ancient and historical population migrations. Ancient DNA studies on the genetic origins of Anatolian farmers pointed to contributions from Iran/Caucasus and the ancient Levant to local Anatolian hunter-gatherers during the late Pleistocene period. [164]. Some other studies contrarily proposed that the early Neolithic Central Anatolian population was most likely descended from local hunter-gatherers instead of immigrants from the Levant or Iran. [165]. Based on the result of previous research, western and eastern Anatolia became genetically homogenized approximately 8,500 years ago as a result of the gene flows between Anatolian and the neighboring South Caucasus and North Levantine populations. [74]. In agreement with those findings, the current study revealed a prominent shared ancestry in modern-day TR, CAU, non-Arab GME populations.

Numerous expansions and migration events took place in Anatolia during classical antiquity and the Middle Ages. Previous studies showed that individuals living in modern-day Anatolia have signs of intermixing between Middle Eastern,

Central Asian, and Siberian populations in their genomes [166]. One significant event that contributed to the formation of the modern-day TR population was the migration of Turkic tribes to Anatolia in the 11th century. Previous reports that used Alu insertion polymorphisms, mitochondrial or Y-chromosome loci of ancient and modern-day individuals suggested that the gene flow from Central Asia to the modern-day TR population was between 3% to 30% [76–78]. Similarly, this study revealed 10% autosomal, 8-15% paternal, and 8% maternal contribution from Central Asia to the TR population using modern-day samples.

Besides, more recent external and internal migration events occurred in Anatolia. During the late 19th and early 20th centuries, economic conditions and urbanization caused a vast amount of permanent internal migrations from the Eastern and Northern Anatolia to the Central, Southern, and Western regions [167]. Moreover, a population exchange with Balkan countries led to approximately 400,000 Balkan refugees settling in Western Anatolia in 1914 [168]. Consistent with these recent demographic events, Treemix analysis indicated gene flows from TR-E to TR-C and TR-S and from TR-B to TR-W. Dimensionality reduction techniques, Procrustes, ADMIXTURE, and $F_{\mathrm{ST}}$ analyses also showed that the geographical regions had a mild but significant impact on the genetic substructure of the TR population. However, no clear-cut separation between TR subregions was observed, which might have reflected the effect of more recent migration events.

The late paleolithic admixture events in Anatolia also spread to Europe. Early Neolithic farmers migrated to Europe and subsequently formed one of the major ancestral components of present-day EUR populations, especially those of Southern Europe [81, 166, 169, 170]. Traces of migration of later Neolithic Anatolian farmers to Europe were also detected by ancient DNA studies [171]. Furthermore, mtDNA studies indicated the recurrent gene flows between Europe and Near East during the past 10,000 years [172]. Hence, population structure analyses in the current study unraveling the close genetic relationship between modern-day TR and EUR populations have also supported ancient DNA studies. Also, the distributions of Y-chromosome and mtDNA haplogroups corroborated both the migration events in Turkey and the genetic connections between TR, BLK, CAU, GME, and EUR populations.

According to the out-of-Africa hypothesis, Homo sapiens originated from East Africa and dispersed across the world. In line with the out-of-Africa hypothesis, a study suggested that the genetic variation in human populations was shaped by two ancient population bottlenecks after they migrated from Africa [173]. Based on the findings of that study, the first bottleneck took place in the Middle East about 50,000 to 60,000 years ago while the second occurred when people migrated through the ancient land bridge on the Bering Strait from Asia to North America. Similarly, the positions of populations on the inferred phylogenetic tree by Treemix analysis are consistent with the out-of-Africa hypothesis. Also, the rates of LD decay in the global populations suggest a shared bottleneck in non-AFR populations. Both analyses demonstrated once more that the TR population genetically links GME and EUR populations.

The results of the previous smaller-scale studies presented the genetic variation in the TR population displayed a similar pattern of population substructure in Turkey with the current study. Hodoğlugil and Mahley, using over 500,000 SNP genotypes of 64 TR samples from Aydın, Kayseri, and İstanbul provinces, showed that TR samples were genetically homogenous and clustered closely with EUR and Middle Eastern samples of HGDP in PCA [79]. They also detected in supervised STRUCTURE analyses that at $k = 3$, ancestral contributions were 40%, 45%, and 15% from EUR, Middle Eastern, and Central Asian populations to TR individuals, respectively; and at $k = 4$ ancestral contributions were 38%, 35%, 18%, and 9% from EUR, Middle Eastern, SAS, and Central Asian populations, respectively [17, 80]. $F_{ST}$, on the other hand, revealed that the genetically closest populations to the TR population were, in order of magnitude, Adygei, Tuscan, Italian, French, Palestinian, and Druze populations. The patterns of LD decay in global populations of HGDP used in that study were almost identical to that of the current study. In 2014, Alkan *et al.* demonstrated the TR population structure and variation using WGS data of 16 TR individuals from 16 different cities [80]. Their results showed that the TR population was clustered with southern EUR populations, TSI and IBS, in Treemix analysis, and a migration event from the EAS branch to the TR population occurred. TThis migration event was interpreted as a gene flow from an ancient North Eurasian ancestry. Additionally, they detected an enrichment of non-synonymous private alleles in the TR population. The analysis of those 16 TR WGS samples identified 651,936 novel SNVs and 542,508 novel indels.

In conclusion, the data in the current study has a substantially higher sample size and provides both WES and WGS samples from all geographical regions in Turkey. Furthermore, the current study analyzed several other neighboring populations, therefore demonstrating the genetic structure of the TR population and genetically related populations in a higher resolution.

## 4.2 Effect of consanguinity on the TR genome

The rate of consanguineous marriage in Turkey has been previously reported as 22-36% [82, 174]. The comparable percentage in Western Europe and the Americas was <2%. Consistent with the demographic information, inbreeding coefficient calculations demonstrated a significantly higher level of consanguinity in the TR population compared to those in AFR, EUR, BLK, CAU, and EAS populations. SAS, GME, and CNA populations had higher medians of inbreeding coefficient compared to that of the TR population. High levels of inbreeding in these populations were also previously reported [175, 176]. As expected, consanguinity and endogamy significantly increased the inbreeding coefficients of TR individuals. The practice of consanguineous marriage has been reported to be higher in the eastern region of Turkey. [174]. However, the current study revealed that the highest medians of inbreeding coefficient belonged to TR-N and TR-S, while the only statistically significant differences were between TR-B vs. TR-N and TR-B vs. TR-S. These findings might reflect the remarkable effect of inbreeding on all TR subregions except TR-B, but also be due to internal migration events or a sampling bias.

The number and length of ROHs were also found to be high in offspring of consanguineous TR couples. In agreement with the high inbreeding coefficients of TR individuals, the TR population demonstrated high levels of total length, medium length, and long ROHs. Especially, long ROHs with >4 Mb in length were specifically observed in the TR population. These results were compatible with the findings of the GME study [17]. Since admixture derives the formation of shorter and fewer ROHs while the consanguinity causes longer ROHs, the variation observed within classes of ROHs in the TR population is plausible [20]. TR-C, TR-N, and TR-E had significantly higher medians of long and

total length ROHs compared to those of TR-B and TR-W, which probably reflected the higher levels of consanguinity in TR-C, TR-N, and TR-E subregions. Genome-wide autozygosity calculated with the long and total length of ROHs was also significantly correlated with the inbreeding coefficient and reported parental status. Therefore, autozygous regions of TR individuals are caused by relatedness between parents instead of lack of genetic diversity in the population.

## 4.3 The TR Variome and reference panel for genotype imputation

The generation of population-based variant resources has contributed to causative gene identification studies and GWAS, as well as has a strong potential for facilitating precision medicine. These resources presented thousands to millions of variants to the human genetics and genomics community. Iranome, for example, provided 308,311 novel variants and showed that 37,384 variants that were previously designated as rare or novel have a population frequency >1% in the Iranian population [32]. Another resource is the GenomeAsia 100K study, which also provided 194,585 novel variants with a frequency higher than 0.1% and 144,329 novel variants with a frequency higher than 1% in 1,739 individuals of 219 population groups across Asia [34]. Furthermore, Uganda genome resource, which comprised genome-wide data from 6,400 individuals and WGS data from 1,978 individuals, contributed the investigation of the heritability of cardiometabolic traits and novel loci associated with hematological, anthropometric, and certain metabolic traits [35]. Similar to the above-mentioned and several other population-specific variant resources, this study enabled the generation of a TR variant dataset using WES and WGS from 3,362 unrelated individuals from Turkey.

Investigation of high throughput sequencing data of 3,362 TR individuals led to the identification of 9,999,451 novel variants which correspond to 21% of all variants in the dataset. Importantly, 37,123 novel variants had a disruptive effect on proteins. When the variants were stratified using both functional impact and their frequency in public databases, rare and novel categories contained higher proportions of HC-pLoF and deleterious missense variants than the common category did. IIndeed, these were expected results since deleterious variants typically

tend to have a lower population frequency, and the Iranome study produced similar results [32]. Furthermore, the TR Variome presented 839,775 novel or previously known rare variants with an AF higher than 1% in the TR population. Evaluation of private variants identified that 38% of them were both novel and had a deleterious effect. DAF calculations showed that gnomAD, now the largest variant resource for humans, was not sufficient to represent the genetic variation seen in Turkey. This conclusion was also valid for GME; GME is not a sufficient estimator for TR DAFs, though it contains 140 samples from the TR population. Variant frequency information is essential for variant and gene prioritization during exclusion of low-probability candidates; therefore, the TR Variome will be beneficial for studies of disease genes [60].

The function of genes has been investigated by introducing knockouts and evaluating their phenotypic consequences [136]. Naturally-occurring homozygous LoF variants in humans, 'human knockouts', enable researchers to study the phenotypic effect of such variants and gene function in humans using sequencing. Still, their implication for clinical interpretation is often difficult because the assignment of LoF effects is based on just predictions. The challenges also arise due to imperfections in sequencing data analysis and annotation, incomplete penetrance, and variance in the impact of knocking out different genes [177]. Exploring large variant databases such as ExAC and gnomAD or studying consanguineous populations are the two main approaches for expanding the list of human knockouts [57, 177]. Hence, the TR Variome was assessed for homozygous HC-pLoFs as well. 20% of TR individuals carried at least one of the 704 rare homozygous HC-pLoFs in 626 genes, of which 173 were not reported in previous publications nor present in 1000GP and gnomAD. Additionally, there were 259 HC-pLoFs, which were previously listed as rare knockouts but have a frequency >1% in the TR population. There were a total of 307 variants in 268 genes categorized as common knockouts in the TR Variome. These common knockouts unravel gene redundancy or advantageous effects; therefore, the TR Variome contributed to the list of 166 genes with 179 common knockouts of ExAC and gnomAD with 220 new genes [138]. Deep phenotyping is required to fully understand the phenotypic consequences of homozygous HC-pLoFs [177]. In the current study, reported phenotypes comprised only the primary phenotypes that brought the family to medical attention. Therefore, experimental verification of predicted knockouts and deep phenotyping of individuals carrying such variants are still necessary to

completely understand their effect on phenotype.

The clinical relevance of the variants in the TR Variome was assessed using OMIM, ClinVar, and HGMD. Identification of HC-pLoF variants located in OMIM genes provided insight to extent of possible disease-causing mutations of TR individuals. Overall, 2,197 OMIM-listed genes with an associated phenotype contained 25.37% of all HC-pLoFs detected in the TR Variome. HC-pLoFs located in OMIM genes associated with a disorder, especially in novel and rare frequency bins, could be prioritized as disease-causing variants and require further investigation. Moreover, variants reported as pathogenic or disease-causing by HGMD and ClinVar were searched in the TR Variome to obtain variants with a certain level of clinical significance. Each TR individual possessed 2-30 variants classified as DMs in HGMD and 0-19 variants classified as P or P/LP variants in ClinVar. These findings might have yielded the secondary findings in the cohort or the carriers for variants associated with recessive disorders. However, these individuals might not have clinical manifestations even though they possess variants that are deemed to be pathogenic. Being heterozygous for a pathogenic variant of a recessive disease will not cause phenotypic alteration. The variable expression could affect the severity of signs and symptoms of a disease, whilst incomplete penetrance in dominant disorders might not affect phenotype despite carrying a pathogenic allele [178, 179]. Also, disease databases can contain false-positive variants. Contrarily, these individuals still can have additional pathogenic variants that were not previously defined in these databases [180]. Hence, variant resources from underrepresented populations such as the TR Variome should be employed in further investigations for the inclusion of both novel disease-causing variants and the exclusion of low probability candidates [181].

Genome-wide variation of TR individuals displayed a high number of variant sites and a remarkable amount of singletons. Although consanguineous populations are expected to have a lower level of genetic diversity, admixture exacerbates genetic variation in the population. Therefore, these high numbers have possibly arisen due to the high level of admixture in Turkey and further emphasized the potential of the TR Variome in rare variant discovery.

Since the overwhelming majority of GWAS was performed in the EUR descent populations, it is crucial to expand GWAS to underrepresented populations to

both replicate findings in these populations and uncover the missing heritability of complex traits. Population-specific reference panels for genotype imputation play a pivotal role to facilitate GWAS. For example, the reference panel generated using WGS data of 175 Mongolian individuals significantly increased the power of imputation for the Mongolian population [67]. Also, reference panels produced by Uganda genome resource and GenomeAsia100K study improved imputation accuracy in these populations [34, 35]. These studies showed that combining pre-existing panels such as 1000GP and HapMap with population-specific panels further improves imputation accuracy. Similarly, the results of the current study revealed that the TR reference panel, when used in combination with the 1000GP reference panel, significantly improves imputation power for the TR and neighboring populations. Therefore, the TR reference panel can be applicable to BLK, CAU, and GME populations and facilitate GWAS in these populations as well.

## 4.4 Utilization of TR Variome in reverse phenotyping

The TR Variome has a huge potential to be also employed in reverse phenotyping. Reverse phenotyping, or genotype first approach, is an alternative method to assign phenotypes in the research setting by surpassing the uncertainties that arise during diagnosis in the clinical setting. The traditional method in disease gene identification is the phenotype-first approach in which patients and families are identified, their clinical data are obtained, a research diagnosis is made, and lastly, their genomic data are collected for analysis. On the contrary, the genotype-first approach begins with the analysis of genomic data, continues with the determination of candidate loci, and subsequently, deep phenotyping of additional patients and their families to evaluate genotype-phenotype correlations in larger cohorts (Figure 4.1).

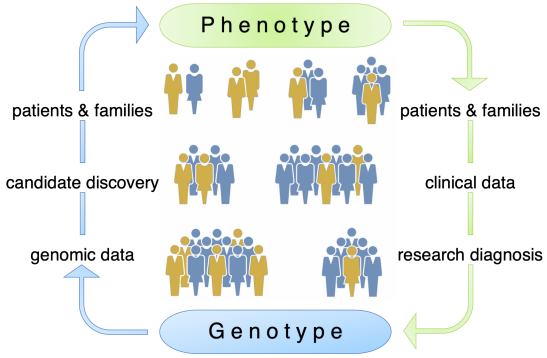



**Figure 4.1: The phenotype-first versus the genotype-first approach.** Reprinted by permission from American Society for Clinical Investigation, [182], Copyright ©2020.

Our group previously uncovered the association of two human *CRY1* variants, c.1657+3A>C (*CRY1*Δ11) and c.825+1G>A (*CRY1*Δ6), with attention deficit hyperactivity disorder (ADHD) by following the genotype-first approach in the TR Variome [182] (Figure 4.2). ADHD is a common heritable disorder of impulse control deficit and hyperkinesis. *CRY1*Δ11 is the causative variant for DSPD. This causation had been previously confirmed using reverse phenotyping in 6 TR families. During the comprehensive phenotyping of these patients, a high incidence of behavioral endophenotypes had been observed, which encouraged us to investigate the relationship of the *CRY* gene with psychiatric disorders. Initially, 96 individuals from 12 TR families were evaluated for psychiatric disorders,. The evaluation revealed a high incidence of signs and symptoms related to ADHD in addition to DSPD in *CRY1*Δ11 carriers. Then, to validate our findings in a larger cohort, we used the WES data of 447 TR individuals with obesity from TR Variome. A variant-based gene burden test revealed that only CRY1 had genome-wide statistical significance. We detected that AF of *CRY1*Δ11 was 0.0124 in 5,465 TR individuals from the sequence data of the TR Variome, Scientific and Technological Research Council of Turkey, and Ankara University Brain Research Center. The overall frequency of this variant was 1 in 44 in TR individuals and 1 in 103 in the EUR populations. We then performed a phenome-wide association study using BioMe BioBank, which contained phenotypic data of 324 *CRY1*Δ11 carriers and 9,114 non-carriers. The analysis showed that the strongest associations of *CRY1*Δ11 were with major depressive disorder, insomnia, anxiety, glaucoma, and nicotine dependence. *CRY1*Δ6, a private variant in a single TR family, was also segregated with ADHD and DSPD. Functional studies showed

that *CRY1*Δ6 results in an arrhythmic phenotype that disrupted the circadian rhythm. In conclusion, the results of this study demonstrate the high potential of the TR Variome for utilization in future disease gene identification studies.

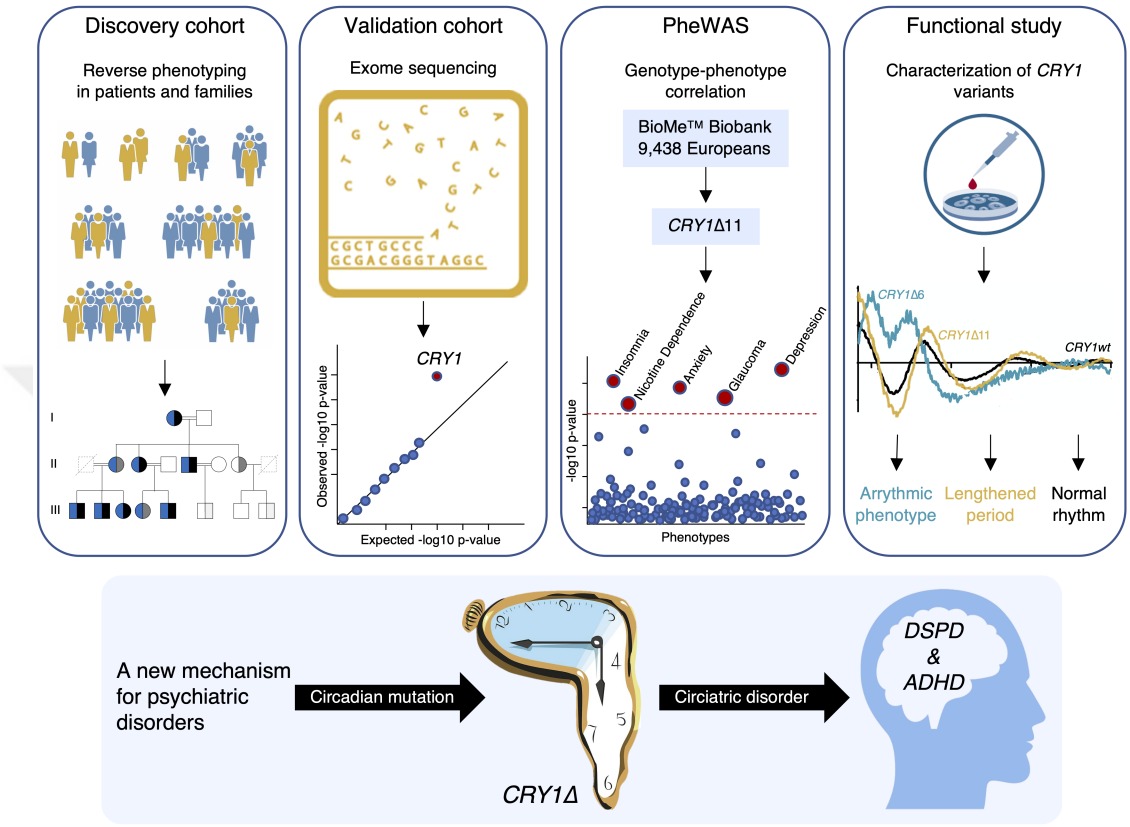



**Figure 4.2: Reverse phenotyping using the TR Variome to identify association of *CRY1* with ADHD.** Reprinted by permission from American Society for Clinical Investigation, [182], Copyright ©2020.

## 4.5 CF for recessive disorders and secondary findings in the TR population

The TR Variome provides a unique opportunity to confer a comprehensive report on the frequency of various genetic disorders in the TR population. In this section of the study, RP and PP variants, CFs, and estimated prevalence of genetic disorders in Turkey were presented using 3,599 unrelated TR individuals. Despite having individuals with various disorders in the TR Variome, a proper representation of the general TR population was obtained by excluding probands

and their family members from the relevant disease groups. Utilizing curated databases containing RP variants and in silico prediction algorithms are more efficient than manual curation for massive data from WES and WGS [53]. Although the curated databases can contain false positives or false negatives, they follow certain standards for variant classification, which should eliminate subjective decisions [180,183]. 70.54% of individuals carried at least 1 RP variant in the most conservative dataset (Dataset 1). The percentage of individuals carrying an RP variant increases to 99.03% when the Dataset 2 variants were used. It should also be noticed that the aim to identify PP variants was intentionally limited to rare HC-pLoF and deleterious missense variants since a low MAF is a predictor of pathogenicity. Although the best efforts were put forth to predict pathogenicity, the deleterious effect of these variants is no more than prediction; therefore, their pathogenic role is needed to be experimentally verified. When Dataset 3, which includes both RP and PP variants, was considered, all TR individuals possessed at least one RP or PP variant. In a previous study, EUR individuals have been shown to carry a higher number of RP variants compared to AFR individuals, which likely reflected the reporting bias in non-EUR descent populations [53]. Thus, the list of PP variants in the current study is important to decrease the potential bias in the TR population.

Another opportunity was to assess the presence of pathogenic variants in NBS and ACMG-recommended actionable genes. Surprisingly, 1 in 4 individuals possessed at least one RP variant in one of the 69 genes related to NBS disorders when the most strict dataset was considered. Furthermore, each individual in this study sample had at least a 1 in 17 chance of carrying an RP variant in one of the 73 ACMG actionable genes. Previous estimates of the percentage of carrying at least one RP variant in 56 ACMG actionable genes ranged from 1.2% to 5.6% in multiethnic cohorts [53, 84]. The conservative estimate resulted in a comparable percentage (5.8%) for the new list of ACMG, which consists of 76 genes.

*MEFV*, *ABCA4*, *CYP21A2*, *PAH*, and *CFTR* displayed the highest cumulative CFs, consistent with the high prevalence of the phenotypes that they are responsible for in the TR population. These highly prevalent diseases and their associated variants are discussed below in detail. *MEFV* encodes a protein named pyrin, which induces the release of pro-inflammatory cytokines and inflammatory cell death. *MEFV* is the causative gene for familial Mediterranean fever (FMF),

which is an AR inflammatory disease. In FMF, recurrent sterile inflammation of serosal membranes occurs together with febrile episodes and eventually leads to amyloidosis [146]. It has been recently reported that *Yersinia pestis* infection might have a role in a recent positive selection of FMF mutations in the Mediterranean basin [184]. FMF is highly prevalent in the TR population, as high as 1 in 1075, and was reported to have a CF of 1 in 5 [145, 146]. In the current study, 1 in 10 to 1 in 9 individuals was found to be a carrier of RP or PP variants in *MEFV*, while GP was calculated as 1 in 608 to 1 in 518. Dataset 1 revealed 359 heterozygous and 5 homozygous individuals for 12 RP variants. Dataset 2 revealed 36 additional heterozygotes for 18 additional RP variants, whereas Dataset 3 provided 1 heterozygote individual carrying 1 PP variant (p.Gln555*). Five most common *MEFV* mutations and their AFs among 100 healthy TR individuals were previously reported as M694V, 3%; M680I, 5%; V726A, 2%; M694I, 0%; and E148Q 12% and among 450 TR FMF patients as M694V 51.55%, M680I 9.22%, V726A 2.88%, M694I 0.44%, and E148Q 3.55% [146]. However, the most common *MEFV* variants and their AFs In the current study were M694V, 1.79%; V726A, 1.46%; M680I, 0.71%; E148Q, 0.56%; and R761H, 0.38%. The AF of M694I in the TR Variome was detected as 0.11%. The different frequencies of *MEFV* variants in the two studies might be attributable to the lower sample size of the former.

*ABCA4* variants account for 30% of all cases with inherited retinal disorders and have a EUR cumulative CF of 2.5%. *ABCA4* mutations manifest with different retinal phenotypes. Recessive phenotypes of *ABCA4* variants in OMIM are severe early-onset retinal dystrophy, Stargardt disease 1, fundus flavimaculatus, cone-rod dystrophy 3, and retinitis pigmentosa 19 whereas the heterozygous mutations lead to susceptibility to age-related macular degeneration 2. However, no clear genotype-phenotype correlations have been established for this gene. One in 21 TR individuals are carriers for *ABCA4* variants according to Dataset 1, while cumulative CF increases up to 1 in 12 in Datasets 2 and 3. G1961E was reported to be the most common *ABCA4* variant in the EUR (AF = 0.39%) and SAS (AF = 1.39%) populations of gnomAD when the hypomorphic variants are not considered. [72]. This variant was also the most prevalent variant of *ABCA4* in the TR population and AF in the TR Variome was much higher (2.05%) than that of the EUR population. Although hypomorphic variants were removed from the calculations, it is also possible that additional hypomorphic variants are still present

in the TR datasets. These possible hypomorphic variants might have elevated the cumulative CF in the TR population significantly because such variants were reported to display the most prominent effect on the EUR-descent populations and were understudied in other populations [72].

*CYP21A2* is the causative gene for congenital adrenal hyperplasia (CAH) due to 21-hydroxylase deficiency. This AR disease presents with adrenocortical insufficiency and hyperandrogenism. CAH due to 21-hydroxylase deficiency is the most common cause of CAH and is screened in all newborns in Turkey. NBS for CAH due to 21-hydroxylase deficiency and CAH due to 11$\beta$-hydroxylase deficiency are performed by measuring 17-hydroxyprogesterone, 21-deoxycortisol, cortisol, 11-deoxycortisol, and androstenedione in heel-prick blood samples [151]. The classical forms of CAH due to 21-hydroxylase deficiency are salt-wasting type, which is life-threatening, and simple virilizing type. The mildest form of CAH is the non-classic type, which is harder to diagnose and might cause hirsutism and irregular menstruation in females as they get older. The prevalence of the classical form of CAH due to 21-hydroxylase deficiency is 1 in 16,000 to 1 in 10,000 in Europe and North America, while the CF and incidence in Turkey were previously estimated as 1 in 50 and 1 in 15,067, respectively. The non-classical form is far more common than the classical form, with a 1 in 1000 incidence in Caucasians [185]. Most of the variants that cause the classical form of CAH due to 21-hydroxylase deficiency are deletions [49]. In the current study, the cumulative CF of *CYP21A2* mutations was calculated as 1 in 25 to 1 in 20 in Datasets 1-3. GP for *CYP21A2* mutations was estimated as 1 in 3,133 as a conservative estimate and up to 1 in 2,353 when Dataset 3 was used. These calculations were performed only using short variants of *CYP21A2*, which might be mostly associated with the non-classical form of CAH.

*PAH* encodes the enzyme, phenylalanine hydroxylase, which converts phenylalanine to tyrosine. PKU is a recessive inborn error of metabolism, caused by the mutations in *PAH*. In PKU, the diminished function of phenylalanine hydroxylase prevents the catalyzation of phenylalanine to tyrosine and subsequently causes excess amounts of phenylalanine in the blood. High amounts of phenylalanine have neurotoxic effects that decrease cognitive functions [150]. Although the classical form of PKU is the most severe type, mutations in *PAH* could also result in mild PKU and mild hyperphenylalaninemia. The prevalence of PKU varies

between populations e.g. 1 in 10,000 in Northern EUR and EAS population, 1 in 14,000 in Finland, and 1 in 143,000 in the Japanese population [149]. NBS for PKU is practiced across the globe. NBS program in Turkey includes PKU due to its high prevalence in the population. The prevalence of PKU in the TR population was reported as 1 in 2,600 in 2001 [186]. However, a more recent study reported the PKU prevalence as 1 in 5,988 and the prevalence of PKU + hyperphenylalaninemia as 1 in 4,057 in the TR population [150]. Similarly, CF for PKU was very high in the previous studies (1 in 26). Likewise, the pathogenic variants in *PAH* displayed a very high frequency in the current study. These variants might be associated with either PKU or hyperphenylalaninemia. The cumulative CF for *PAH* variants were calculated as 1 in 25 in Dataset 1 and 1 in 19 in Datasets 2 and 3. GP for *PAH* mutations in Dataset 1 was estimated as 1 in 4,368 and as 1,078 in Dataset 2 and 3. No PP variant was detected for the *PAH* gene. Therefore, additional RP variants in Dataset 2 and Dataset 3, which are reported as pathogenic in either HGMD or ClinVar, are possibly false positives. A previous study investigated the mutation profile of 66 TR newborns with amino acid disorders, of which 62 were PKU. Five most common *PAH* variants in these 66 subjects were A403V (AF = 13.64%), A300S (AF= 12.88%), V230I (AF = 11.36%), c.1066-11G>A (AF = 9.09%), and c.441+5G>T (AF = 7.58%) [187]. In studies investigating the mutation spectrum of PKU patients from Iran, the c.1066-11G>A variant had the highest frequency: AF = 27.5% and 19.3% in North Iran and Azerbaijani population, respectively [188, 189]. As far as we know, there are no previous studies investigating the mutation spectrum of *PAH* gene in the general TR population. In the current study, five most common *PAH* mutations in Dataset 1 were A300S (AF = 0.45%), T380M(AF = 0.29%), A403V (AF= 0.25%), c.1066-11G>A (AF = 0.21%), c.1169A>G (AF = 0.14%), and c.782G>A (AF = 0.11%). Dataset 2 and 3 also contained frequent *PAH* variants that were not present in Dataset 1: c.*19G>T (AF = 0.89%) and V230I (AF = 0.43%).

Cystic fibrosis is the most common AR disease in the EUR-descent populations and caused by mutations in *CFTR*. The CF of *CFTR* mutations is 1 in 25 in the EUR populations, while the prevalence of cystic fibrosis was reported as 1 in 3,333 in Turkey in 1973 using sweat chloride test [147, 148]. In the current study, the cumulative CF of *CFTR* variants were 1 in 30 to 1 in 13 in Datasets 1-3 while GPs ranged from 1 in 6,502 to 1 in 1,194. The most common mutation in the EUR

cystic fibrosis cases is ΔF508 ( 70%); however, cystic fibrosis displays a high level of genetic heterogeneity [147]. A previous study investigating the heterogeneity of *CFTR* mutations in TR cystic fibrosis patients revealed 36 mutations in 125 (75%) of the total 166 chromosomes [147]. Five most common mutations and their frequencies in that study were ΔF508, 23.5%; c.1677delTA, 7.2%; c.2183AA>G, 4.2%; G542X, 3.6%; F1052V, 3%. Current study detected a total of 105 RP and PP variants in *CFTR*: 41 RP variants in Dataset 1, 61 additional RP variants in Dataset 2, and 3 additional PP variants in Dataset 3. Five most common variants of *CFTR* were c.1210-11T>G (AF = 0.91%), c.-1044_ -1043insT (AF = 0.71%), ΔF508 (AF = 0.28%), P1013L (AF = 0.24%), K68E (AF = 0.23%).

Overall, Dataset 1 provided closer estimates of CF and GP to observed numbers compared to those of Dataset 2 and Dataset 3. However, none of the datasets were identified as a good estimator for a number of diseases. For instance, the estimated cumulative CF for *MCCC1* and *MCCC2* by all three datasets were higher than the reported prevalence and the reported CF of their consequent phenotype, 3-methylcrotonyl-CoA carboxylase deficiency. On the contrary, for *SERPINA1*, the cause of Alpha-1 antitrypsin deficiency, the RP and/or PP variants were inadequate to elucidate the reported prevalence rate. Various factors contribute to the dissimilarity between the estimated and reported frequency of disease. The GP calculation used in this study does not take inbreeding, isolated groups or assortative mating into consideration [72]. Since the TR population has high inbreeding coefficients, it could lead to underestimation of the GP in the current study. Determining the true prevalence of a disease, especially that of a rare disorder, is another significant challenge if there is a lack of a national system for epidemiological statistics. Lastly, misdiagnosis would be an obstacle to accurate calculations of disease frequency.

Inconsistencies between estimated and reported CF or between estimated GP and reported prevalence might be attributable to incomplete penetrance, locus heterogeneity, environmental factors for complex diseases, and under- or overestimation of frequencies due to false negatives or false positives [69]. Particularly, incomplete penetrance probably underlies the increased proportion of variants associated with AD disease such as *TGIF1* associated holoprosencephaly in the least conservative dataset, Dataset 3 [190]. It should also be noted that large indels, copy number variations, and trinucleotide repeat disorders, which are the

prominent causes of certain diseases such as Duchenne muscular dystrophy, alpha thalassemia, spinal muscular atrophy, and Huntington disease, cannot be detected by the short variant discovery in the current study.

## 4.6 Conclusion and Future Perspectives

In conclusion, this study demonstrates the effect of geographical location, admixture, and inbreeding on the TR genome by investigating the fine-scale genetic structure of the TR population. Using the data produced in this study along with ancestral sequences from Anatolia will further contribute to the research of population genetics. The TR Variome, which contains allele counts and annotations of over 40 million variants, is now a publicly available variant resource for future studies of human genetics. The TR reference panel for genotype imputation will also facilitate GWAS in Turkey, as well as in neighboring populations.

This data will also probably serve as the principal resource for the genetic counseling of TR people until a larger resource is generated. The findings can be useful in epidemiological studies that will shape public health decisions related to screening, diagnosis, and treatment. The data also provides an invaluable opportunity to design a custom SNP array for carrier screening, which will facilitate precision medicine in Turkey. If such kind of testing is performed regularly in Turkey, it will be beneficial for the TR population from both administrative and individual perspectives. From an administrative perspective, governors will be able to anticipate the health-related economic burden of their country through the frequency estimates of potential patients with a genetic disease. From an individuals perspective, one can learn about their genetic risk to have a chance to take preventive measures or start lifestyle modifications. Such kind of testing would also be useful for prenatal screening or NBS.

The differences between estimated and reported frequencies will be lowered as the number of RP variants increases with the newer versions of ClinVar and HGMD. It is conceivable that these departures will be fixed and pathogenic variants will be more easily identified in the future with the improvements in the variant annotation and interpretation strategies in addition to clinical and experimental verification. As the cost of sequencing decreases, it can be foreseen that

sequencing every newborn at birth will be possible soon. Therefore, more precise information on the frequencies of genetic disease and genetic susceptibility will be obtained in the future.

# Bibliography


[1] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters, "New goals for the U.S. Human Genome Project: 1998-2003," *Science*, vol. 282, no. 5389, pp. 682–9, 1998.

[2] J. D. McPherson, M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis *et al.*, "A physical map of the human genome," *Nature*, vol. 409, no. 6822, pp. 934–41, 2001.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–51, 2001.

[4] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–45, 2004.

[5] W. Li, K. R. O'Neill, D. H. Haft, M. DiCuccio, V. Chetvernin, A. Badretdin *et al.*, "RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation," *Nucleic Acids Res*, vol. 49, no. D1, pp. D1020–D1028, 2021.

[6] K. L. Howe, P. Achuthan, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean *et al.*, "Ensembl 2021," *Nucleic Acids Res*, vol. 49, no. D1, pp. D884–D891, 2021.

[7] F. S. Collins, M. S. Guyer, and A. Charkravarti, "Variations on a theme: Cataloging human DNA sequence variation," *Science*, vol. 278, no. 5343, pp. 1580–1, 1997.

[8] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nat Rev Genet*, vol. 10, no. 4, pp. 241–51, 2009.

[9] R. L. Nussbaum, R. R. McInnes, and H. F. Willard, *Thompson & Thompson genetics in medicine*. Philadelphia: Elsevier, 2016.

[10] R. A. Studer, B. H. Dessailly, and C. A. Orengo, "Residue mutations and their impact on protein structure and function: Detecting beneficial and pathogenic changes," *Biochem J*, vol. 449, no. 3, pp. 581–94, 2013.

[11] H. F. Lodish, A. Berk, C. Kaiser, M. Krieger, A. Bretscher, H. L. Ploegh *et al.*, *Molecular Cell Biology*. New York: W.H. Freeman, 2016.

[12] H. A. Orr, "Fitness and its role in evolutionary genetics," *Nat Rev Genet*, vol. 10, no. 8, pp. 531–9, 2009.

[13] R. L. Hanson, R. Rong, S. Kobes, Y. L. Muller, E. J. Weil, J. M. Curtis *et al.*, "Role of established type 2 diabetes-susceptibility genetic variants in a high prevalence American Indian population," *Diabetes*, vol. 64, no. 7, pp. 2646–57, 2015.

[14] M. W. Jallow, C. Cerami, T. G. Clark, A. M. Prentice, and S. Campino, "Differences in the frequency of genetic variants associated with iron imbalance among global populations," *PLoS One*, vol. 15, no. 7, p. e0235141, 2020.

[15] J. L. Yen-Revollo, D. J. Van Booven, E. J. Peters, J. M. Hoskins, R. M. Engen, H. D. Kannall *et al.*, "Influence of ethnicity on pharmacogenetic variation in the Ghanaian population," *Pharmacogenomics J*, vol. 9, no. 6, pp. 373–9, 2009.

[16] X. Yang, S. Al-Bustan, Q. Feng, W. Guo, Z. Ma, M. Marafie *et al.*, "The influence of admixture and consanguinity on population genetic diversity in Middle East," *J Hum Genet*, vol. 59, no. 11, pp. 615–622, 2014.

[17] E. M. Scott, A. Halees, Y. Itan, E. G. Spencer, Y. He, M. A. Azab *et al.*, "Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery," *Nat Genet*, vol. 48, no. 9, pp. 1071–6, 2016.

[18] H. Hamamy, S. E. Antonarakis, L. L. Cavalli-Sforza, S. Temtamy, G. Romeo, L. P. Kate *et al.*, "Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report," *Genet Med*, vol. 13, no. 9, pp. 841–7, 2011.

[19] F. C. Ceballos and G. Alvarez, "Royal dynasties as human inbreeding laboratories: the Habsburgs," *Heredity*, vol. 111, no. 2, pp. 114–21, 2013.

[20] F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, "Runs of homozygosity: Windows into population history and trait architecture," *Nat Rev Genet*, vol. 19, no. 4, pp. 220–34, 2018.

[21] S. T. Sherry, M. Ward, and K. Sirotkin, "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation," *Genome Res*, vol. 9, no. 8, pp. 677–9, 1999.

[22] I. Lappalainen, J. Lopez, L. Skipper, T. Hefferon, J. D. Spalding, J. Garner *et al.*, "DbVar and DGVa: Public archives for genomic structural variation," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D936–41, 2013.

[23] International HapMap Consortium, "The International HapMap Project," *Nature*, vol. 426, no. 6968, pp. 789–96, 2003.

[24] International Hapmap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–320, 2005.

[25] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs *et al.*, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–61, 2007.

[26] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–73, 2010.

[27] A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[28] NHLBI GO Exome Sequencing Project (ESP), https://evs.gs.washington.edu/EVS/, Accessed: 2019-11-10.

[29] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–91, 2016.

[30] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang *et al.*, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[31] K. A. Fakhro, M. R. Staudt, M. D. Ramstetter, A. Robay, J. A. Malek, R. Badii *et al.*, "The Qatar genome: A population-specific tool for precision medicine in the Middle East," *Hum Genome Var*, vol. 3, no. 16016, p. 16016, 2016.

[32] Z. Fattahi, M. Beheshtian, M. Mohseni, H. Poustchi, E. Sellars, S. H. Nezhadi *et al.*, "Iranome: A catalog of genomic variations in the Iranian population," *Hum Mutat*, vol. 40, no. 11, pp. 1968–1984, 2019.

[33] M. A. Almarri, M. Haber, R. A. Lootah, P. Hallast, S. Al Turki, H. C. Martin *et al.*, "The genomic history of the Middle East," *Cell*, vol. 184, no. 18, pp. 4612–4625.e14, 2021.

[34] GenomeAsia100K Consortium, "The GenomeAsia 100K Project enables genetic discoveries across Asia," *Nature*, vol. 576, no. 7785, pp. 106–111, 2019.

[35] D. Gurdasani, T. Carstensen, S. Fatumo, G. Chen, C. S. Franklin, J. Prado-Martinez *et al.*, "Uganda Genome Resource enables insights into population history and genomic discovery in Africa," *Cell*, vol. 179, no. 4, pp. 984–1002.e36, 2019.

[36] A. Jain, R. C. Bhoyar, K. Pandhare, A. Mishra, D. Sharma, M. Imran *et al.*, "IndiGenomes: A comprehensive resource of genetic variants from over 1000 Indian genomes," *Nucleic Acids Res*, vol. 49, no. D1, pp. D1225–D1232, 2021.

[37] C. C. Wang, H. Y. Yeh, A. N. Popov, H. Q. Zhang, H. Matsumura, K. Sirak *et al.*, "Genomic insights into the formation of human populations in East Asia," *Nature*, vol. 591, no. 7850, pp. 413–419, 2021.

[38] I. K. Jordan, L. Rishishwar, and A. B. Conley, "Native American admixture recapitulates population-specific migration and settlement of the continental United States," *PLoS Genet*, vol. 15, no. 9, p. e1008225, 2019.

[39] E. Gilbert, S. O'Reilly, M. Merrigan, D. McGettigan, V. Vitart, P. K. Joshi *et al.*, "The genetic landscape of Scotland and the Isles," *Proc Natl Acad Sci U S A*, vol. 116, no. 38, pp. 19 064–70, 2019.

[40] A. Bergstrom, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek *et al.*, "Insights into human genetic variation and population history from 929 diverse genomes," *Science*, vol. 367, no. 6484, 2020.

[41] T. Tucker, M. Marra, and J. M. Friedman, "Massively parallel sequencing: The next big thing in genetic medicine," *Am J Hum Genet*, vol. 85, no. 2, pp. 142–54, 2009.

[42] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire *et al.*, "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, no. 7189, pp. 872–6, 2008.

[43] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333–51, 2016.

[44] M. Pirooznia, M. Kramer, J. Parla, F. S. Goes, J. B. Potash, W. R. McCombie *et al.*, "Validation and assessment of variant calling pipelines for next-generation sequencing," *Hum Genomics*, vol. 8, p. 14, 2014.

[45] Z. Stark, T. Y. Tan, B. Chong, G. R. Brett, P. Yap, M. Walsh *et al.*, "A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders," *Genet Med*, vol. 18, no. 11, pp. 1090–6, 2016.

[46] Y. Yang, D. M. Muzny, F. Xia, Z. Niu, R. Person, Y. Ding *et al.*, "Molecular findings among patients referred for clinical whole-exome sequencing," *JAMA*, vol. 312, no. 18, pp. 1870–9, 2014.

[47] J. C. Taylor, H. C. Martin, S. Lise, J. Broxholme, J. B. Cazier, A. Rimmer *et al.*, "Factors influencing success of clinical genome sequencing across a broad spectrum of disorders," *Nat Genet*, vol. 47, no. 7, pp. 717–26, 2015.

[48] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster *et al.*, "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genet Med*, vol. 17, no. 5, pp. 405–24, 2015.

[49] Online Mendelian Inheritance in Man (OMIM), https://www.omim.org/, Accessed: 2019-12-10.

[50] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla *et al.*, "ClinVar: Improving access to variant interpretations and supporting evidence," *Nucleic Acids Res*, vol. 46, no. D1, pp. D1062–D1067, 2018.

[51] P. D. Stenson, M. Mort, E. V. Ball, M. Chapman, K. Evans, L. Azevedo *et al.*, "The Human Gene Mutation Database (HGMD): Optimizing its use in a clinical diagnostic or research setting," *Hum Genet*, vol. 139, no. 10, pp. 1197–207, 2020.

[52] B. B. Cummings, K. J. Karczewski, J. A. Kosmicki, E. G. Seaby, N. A. Watts, M. Singer-Berk *et al.*, "Transcript expression-aware annotation improves rare variant interpretation," *Nature*, vol. 581, no. 7809, pp. 452–8, 2020.

[53] T. Gambin, S. N. Jhangiani, J. E. Below, I. M. Campbell, W. Wiszniewski, D. M. Muzny *et al.*, "Secondary findings and carrier test frequencies in a large multiethnic sample," *Genome Med*, vol. 7, no. 1, p. 54, 2015.

[54] Q. Xiao and V. M. Lauschke, "The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders," *NPJ Genom Med*, vol. 6, no. 1, p. 41, 2021.

[55] J. P. Desvignes, M. Bartoli, V. Delague, M. Krahn, M. Miltgen, C. Broud *et al.*, "VarAFT: A variant annotation and filtration system for human next generation sequencing data," *Nucleic Acids Res*, vol. 46, no. W1, pp. W545–W553, 2018.

[56] S. Balasubramanian, Y. Fu, M. Pawashe, P. McGillivray, M. Jin, J. Liu *et al.*, "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes," *Nat Commun*, vol. 8, no. 1, p. 382, 2017.

[57] V. M. Narasimhan, K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. R. Barnes *et al.*, "Health and population effects of rare gene knockouts in adult humans with related parents," *Science*, vol. 352, no. 6284, pp. 474–7, 2016.

[58] D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis *et al.*, "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–76, 2014.

[59] S. Chakravorty and M. Hegde, "Gene and variant annotation for Mendelian disorders in the era of advanced sequencing technologies," *Annu Rev Genomics Hum Genet*, vol. 18, pp. 229–56, 2017.

[60] K. Eilbeck, A. Quinlan, and M. Yandell, "Settling the score: Variant prioritization and Mendelian disease," *Nat Rev Genet*, vol. 18, no. 10, pp. 599–612, 2017.

[61] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis *et al.*, "Genome-wide association studies for complex traits: Consensus, uncertainty and challenges," *Nat Rev Genet*, vol. 9, no. 5, pp. 356–69, 2008.

[62] M. C. Mills and C. Rahal, "The GWAS Diversity Monitor tracks diversity by disease in real time," *Nat Genet*, vol. 52, no. 3, pp. 242–3, 2020.

[63] M. A. Jurj, M. Buse, A. A. Zimta, A. Paradiso, S. S. Korban, L. A. Pop *et al.*, "Critical analysis of Genome-Wide Association Studies: Triple Negative Breast Cancer," *Int J Mol Sci*, vol. 21, no. 16, 2020.

[64] M. Sharma, R. Kruger, and T. Gasser, "From genome-wide association studies to next-generation sequencing: Lessons from the past and planning for the future," *JAMA Neurol*, vol. 71, no. 1, pp. 5–6, 2014.

[65] H. Kilpinen and J. C. Barrett, "How next-generation sequencing is transforming complex disease genetics," *Trends Genet*, vol. 29, no. 1, pp. 23–30, 2013.

[66] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nat Rev Genet*, vol. 11, no. 7, pp. 499–511, 2010.

[67] H. Bai, X. Guo, N. Narisu, T. Lan, Q. Wu, Y. Xing *et al.*, "Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia," *Nat Genet*, vol. 50, no. 12, pp. 1696–1704, 2018.

[68] H. K. Tabor, P. L. Auer, S. M. Jamal, J. X. Chong, J. H. Yu, A. S. Gordon *et al.*, "Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: Implications for the return of incidental results," *Am J Hum Genet*, vol. 95, no. 2, pp. 183–93, 2014.

[69] Y. Yamaguchi-Kabata, J. Yasuda, A. Uruno, K. Shimokawa, S. Koshiba, Y. Suzuki *et al.*, "Estimating carrier frequencies of newborn screening disorders using a whole-genome reference panel of 3552 Japanese individuals," *Hum Genet*, vol. 138, no. 4, pp. 389–409, 2019.

[70] American College of Medical Genetics Newborn Screening Expert Group, "Newborn screening: Toward a uniform screening panel and system–executive summary," *Pediatrics*, vol. 117, no. 5 Pt 2, pp. S296–307, 2006.

[71] D. T. Miller, K. Lee, W. K. Chung, A. S. Gordon, G. E. Herman, T. E. Klein *et al.*, "ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG)," *Genet Med*, vol. 23, no. 8, pp. 1381–90, 2021.

[72] M. Hanany, C. Rivolta, and D. Sharon, "Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases," *Proc Natl Acad Sci U S A*, vol. 117, no. 5, pp. 2710–6, 2020.

[73] G. A. Yanus, T. A. Akhapkina, A. J. Whitehead, I. V. Bizin, A. G. Iyevleva, E. S. Kuligina *et al.*, "Exome-based search for recurrent disease-causing alleles in Russian population," *Eur J Med Genet*, vol. 62, no. 7, p. 103656, 2019.

[74] E. Skourtanioti, Y. S. Erdal, M. Frangipane, F. Balossi Restelli, K. A. Yener, F. Pinnock *et al.*, "Genomic history of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus," *Cell*, vol. 181, no. 5, pp. 1158–75, 2020.

[75] T. Ozcelik, "Medical genetics and genomic medicine in Turkey: A bright future at a new era in life sciences," *Mol Genet Genomic Med*, vol. 5, no. 5, pp. 466–72, 2017.

[76] C. Cinnioglu, R. King, T. Kivisild, E. Kalfoglu, S. Atasoy, G. L. Cavalleri *et al.*, "Excavating Y-chromosome haplotype strata in Anatolia," *Hum Genet*, vol. 114, no. 2, pp. 127–48, 2004.

[77] C. C. Berkman, H. Dinc, C. Sekeryapan, and I. Togan, "Alu insertion polymorphisms and an assessment of the genetic contribution of Central Asia to Anatolia with respect to the Balkans," *Am J Phys Anthropol*, vol. 136, no. 1, pp. 11–8, 2008.

[78] G. Di Benedetto, A. Erguven, M. Stenico, L. Castr, G. Bertorelle, I. Togan *et al.*, "DNA diversity and population admixture in Anatolia," *Am J Phys Anthropol*, vol. 115, no. 2, pp. 144–56, 2001.

[79] U. Hodoglugil and R. W. Mahley, "Turkish population structure and genetic ancestry reveal relatedness among Eurasian populations," *Ann Hum Genet*, vol. 76, no. 2, pp. 128–41, 2012.

[80] C. Alkan, P. Kavak, M. Somel, O. Gokcumen, S. Ugurlu, C. Saygi *et al.*, "Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa," *BMC Genomics*, vol. 15, p. 963, 2014.

[81] I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick *et al.*, "Genomic insights into the origin of farming in the ancient Near East," *Nature*, vol. 536, no. 7617, pp. 419–24, 2016.

[82] S. Akbayram, N. Sari, C. Akgun, M. Dogan, O. Tuncer, H. Caksen *et al.*, "The frequency of consanguineous marriage in eastern Turkey," *Genet Couns*, vol. 20, no. 3, pp. 207–14, 2009.

[83] T. Ozcelik, M. Kanaan, K. B. Avraham, D. Yannoukakos, A. Mgarban, G. O. Tadmouri *et al.*, "Collaborative genomics for human health and cooperation in the Mediterranean region," *Nat Genet*, vol. 42, no. 8, pp. 641–5, 2010.

[84] M. O. Dorschner, L. M. Amendola, E. H. Turner, P. D. Robertson, B. H. Shirts, C. J. Gallego *et al.*, "Actionable, pathogenic incidental findings in 1,000 participants' exomes," *Am J Hum Genet*, vol. 93, no. 4, pp. 631–40, 2013.

[85] G. A. Lazarin, I. S. Haque, S. Nazareth, K. Iori, A. S. Patterson, J. L. Jacobson *et al.*, "An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: Results from an ethnically diverse clinical sample of 23,453 individuals," *Genet Med*, vol. 15, no. 3, pp. 178–86, 2013.

[86] B. Tezel, D. Dilli, H. Bolat, H. Sahman, S. Ozbas, D. Acican *et al.*, "The development and organization of newborn screening programs in Turkey," *J Clin Lab Anal*, vol. 28, no. 1, pp. 63–9, 2014.

[87] M. E. Kars, A. N. Basak, O. E. Onat, K. Bilguvar, J. Choi, Y. Itan *et al.*, "The genetic structure of the Turkish population reveals high levels of variation and admixture," *Proc Natl Acad Sci U S A*, vol. 118, no. 36, p. e2026076118, 2021.

[88] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–60, 2009.

[89] Broad Institute, "Picard tools," http://broadinstitute.github.io/picard/, 2019, Accessed: 2019-11-12; version 2.21.2.

[90] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky *et al.*, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res*, vol. 20, no. 9, pp. 1297–303, 2010.

[91] C. M. Farrell, N. A. O'Leary, R. A. Harte, J. E. Loveland, L. G. Wilming, C. Wallin *et al.*, "Current status and new features of the Consensus Coding Sequence database," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D865–72, 2014.

[92] C. Raczy, R. Petrovski, C. T. Saunders, I. Chorny, S. Kruglyak, E. H. Margulies *et al.*, "Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms," *Bioinformatics*, vol. 29, no. 16, 2013.

[93] J. O'Connell, "Illumina gvcfgenotyper," https://github.com/Illumina/gvcfgenotyper, 2019, Accessed: 2019-07-23; version 2019.02.2.

[94] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard *et al.*, "Twelve years of SAMtools and BCFtools," *Gigascience*, vol. 10, no. 2, 2021.

[95] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, vol. 26, no. 22, pp. 2867–73, 2010.

[96] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson *et al.*, "The UCSC Genome Browser Database: Update 2006," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D590–8, 2006.

[97] H. Fang, Y. Wu, G. Narzisi, J. A. O'Rawe, L. T. Barrn, J. Rosenbaum *et al.*, "Reducing INDEL calling errors in whole genome and exome sequencing data," *Genome Med*, vol. 6, no. 10, p. 89, 2014.

[98] R. L. Goldfeder, J. R. Priest, J. M. Zook, M. E. Grove, D. Waggott, M. T. Wheeler *et al.*, "Medical implications of technical accuracy in genome sequencing," *Genome Med*, vol. 8, no. 1, p. 24, 2016.

[99] A. Belkadi, V. Pedergnana, A. Cobat, Y. Itan, Q. B. Vincent, A. Abhyankar *et al.*, "Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage," *Proc Natl Acad Sci U S A*, vol. 113, no. 24, pp. 6713–8, 2016.

[100] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: Rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, p. 7, 2015.

[101] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet*, vol. 2, no. 12, p. e190, 2006.

[102] H. Wickham, *ggplot2: Elegant graphics for data analysis*. Springer Science & Business Media, 2009.

[103] Z. Zhang, "Reshaping and aggregating data: An introduction to reshape package," *Ann Transl Med*, vol. 4, no. 4, p. 78, 2016.

[104] H. Wickham, R. Francois, L. Henry, and K. Muller, "dplyr: A grammar of data manipulation." 2018, R package version 1.0.7, https://cran.r-project.org/web/packages/dplyr/.

[105] H. Wickham, "stringr: Simple, consistent wrappers for common string operations," 2019, R package version 1.4.0, https://cran.r-project.org/web/packages/stringr/.

[106] C. Wang, S. Zollner, and N. A. Rosenberg, "A quantitative comparison of the similarity between genes and geography in worldwide human populations," *PLoS Genet*, vol. 8, no. 8, p. e1002886, 2012.

[107] J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn *et al.*, "vegan: Community ecology package," 2020, R package version 2.5-7, https://CRAN.R-project.org/package=vegan.

[108] W. Li, J. E. Cerise, Y. Yang, and H. Han, "Application of t-SNE to human genetic data," *J Bioinform Comput Biol*, vol. 15, no. 4, p. 1750017, 2017.

[109] A. Diaz-Papkovich, L. Anderson-Trocm, C. Ben-Eghan, and S. Gravel, "UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts," *PLoS Genet*, vol. 15, no. 11, p. e1008432, 2019.

[110] J. H. Krijthe, "Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation," 2015, R package version 0.15, https://github.com/jkrijthe/Rtsne.

[111] T. Konopka, "umap," https://github.com/tkonopka/umap, 2019, Accessed: 2020-11-13; version 0.2.7.0.

[112] D. H. Alexander, J. Novembre, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," *Genome Res*, vol. 19, no. 9, pp. 1655–64, 2009.

[113] J. K. Pickrell and J. K. Pritchard, "Inference of population splits and mixtures from genome-wide allele frequency data," *PLoS Genet*, vol. 8, no. 11, p. e1002967, 2012.

[114] G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price, "Estimating and interpreting FST: The impact of rare variants," *Genome Res*, vol. 23, no. 9, pp. 1514–21, 2013.

[115] L. Park, "Linkage disequilibrium decay and past population history in the human genome," *PLoS One*, vol. 7, no. 10, p. e46603, 2012.

[116] F. C. Ceballos, S. Hazelhurst, and M. Ramsay, "Assessing runs of Homozygosity: A comparison of SNP Array and whole genome sequence low coverage data," *BMC Genomics*, vol. 19, no. 1, p. 106, 2018.

[117] T. J. Pemberton, D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li, "Genomic patterns of homozygosity in worldwide human populations," *Am J Hum Genet*, vol. 91, no. 2, pp. 275–92, 2012.

[118] R. McQuillan, A. L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc *et al.*, "Runs of homozygosity in European populations," *Am J Hum Genet*, vol. 83, no. 3, pp. 359–72, 2008.

[119] H. Chen, Y. Lu, D. Lu, and S. Xu, "Y-LineageTracker: A high-throughput analysis framework for Y-chromosomal next-generation sequencing data," *BMC Bioinformatics*, vol. 22, no. 1, p. 114, 2021.

[120] M. van Oven and M. Kayser, "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation," *Hum Mutat*, vol. 30, no. 2, pp. E386–94, 2009.

[121] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran *et al.*, "Worldwide human relationships inferred from genome-wide patterns of variation," *Science*, vol. 319, no. 5866, pp. 1100–4, 2008.

[122] H. Weissensteiner, D. Pacher, A. Kloss-Brandsttter, L. Forer, G. Specht, H. J. Bandelt *et al.*, "HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing," *Nucleic Acids Res*, vol. 44, no. W1, pp. W58–63, 2016.

[123] D. Comas, S. Plaza, R. S. Wells, N. Yuldaseva, O. Lao, F. Calafell *et al.*, "Admixture, migrations, and dispersals in Central Asia: Evidence from maternal DNA lineages," *Eur J Hum Genet*, vol. 12, no. 6, pp. 495–504, 2004.

[124] J. C. Fay and C. I. Wu, "Hitchhiking under positive Darwinian selection," *Genetics*, vol. 155, no. 3, pp. 1405–13, 2000.

[125] J. Lachance, "Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations," *BMC Med Genomics*, vol. 3, p. 57, 2010.

[126] P. Lindenbaum, "JVarkit: java-based utilities for bioinformatics," 2015, https://github.com/lindenb/jvarkit.

[127] S. E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard *et al.*, "Ensembl variation resources," *Database (Oxford)*, vol. 2018, 2018.

[128] P. Cingolani, A. Platts, l. L. Wang, M. Coon, T. Nguyen, L. Wang *et al.*, "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)*, vol. 6, no. 2, pp. 80–92, 2012.

[129] M. G. Reese, B. Moore, C. Batchelor, F. Salas, F. Cunningham, G. T. Marth *et al.*, "A standard variation file format for human genome sequences," *Genome Biol*, vol. 11, no. 8, p. R88, 2010.

[130] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter *et al.*, "A systematic survey of loss-of-function variants in human protein-coding genes," *Science*, vol. 335, no. 6070, pp. 823–8, 2012.

[131] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork *et al.*, "A method and server for predicting damaging missense mutations," *Nat Methods*, vol. 7, no. 4, pp. 248–9, 2010.

[132] R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng, "SIFT missense predictions for genomes," *Nat Protoc*, vol. 11, no. 1, pp. 1–9, 2016.

[133] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: Predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res*, vol. 47, no. D1, pp. D886–D894, 2019.

[134] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res*, vol. 38, no. 16, p. e164, 2010.

[135] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs," *Hum Mutat*, vol. 37, no. 3, pp. 235–41, 2016.

[136] D. Saleheen, P. Natarajan, I. M. Armean, W. Zhao, A. Rasheed, S. A. Khetarpal *et al.*, "Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity," *Nature*, vol. 544, no. 7649, pp. 235–9, 2017.

[137] P. Sulem, H. Helgason, A. Oddson, H. Stefansson, S. A. Gudjonsson, F. Zink *et al.*, "Identification of a large set of rare complete human knockouts," *Nat Genet*, vol. 47, no. 5, pp. 448–52, 2015.

[138] A. Rausell, Y. Luo, M. Lopez, Y. Seeleuthner, F. Rapaport, A. Favier *et al.*, "Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes," *Proc Natl Acad Sci U S A*, vol. 117, no. 24, pp. 13 626–36, 2020.

[139] B. L. Browning, Y. Zhou, and S. R. Browning, "A one-penny imputed genome from next-generation reference panels," *Am J Hum Genet*, vol. 103, no. 3, pp. 338–48, 2018.

[140] O. Delaneau, J. F. Zagury, and J. Marchini, "Improved whole-chromosome phasing for disease and population genetic studies," *Nat Methods*, vol. 10, no. 1, pp. 5–6, 2013.

[141] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet*, vol. 5, no. 6, p. bay119, 2009.

[142] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo *et al.*, "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations," *Nature*, vol. 538, no. 7624, pp. 201–6, 2016.

[143] B. M. Henn, L. R. Botigu, S. Gravel, W. Wang, A. Brisbin, J. K. Byrnes *et al.*, "Genomic ancestry of North Africans supports back-to-Africa migrations," *PLoS Genet*, vol. 8, no. 1, p. e1002397, 2012.

[144] Z. Bnfai, B. I. Melegh, K. Sumegi, K. Hadzsiev, A. Miseta, M. Ksler *et al.*, "Revealing the Genetic Impact of the Ottoman Occupation on Ethnic

Groups of East-Central Europe and on the Roma Population of the Area," *Front Genet*, vol. 10, p. 558, 2019.

[145] E. Yilmaz, S. Ozen, B. Balci, A. Duzova, R. Topaloglu, N. Besbas *et al.*, "Mutation frequency of Familial Mediterranean Fever and evidence for a high carrier rate in the Turkish population," *Eur J Hum Genet*, vol. 9, no. 7, pp. 553–5, 2001.

[146] V. Cobankara, G. Fidan, T. Turk, M. Zencir, M. Colakoglu, and S. Ozen, "The prevalence of familial Mediterranean fever in the Turkish province of Denizli: A field study with a zero patient design," *Clin Exp Rheumatol*, vol. 22, no. 4 Suppl 34, pp. S27–30, 2004.

[147] M. O. Kilinc, V. N. Ninis, E. Dagli, M. Demirkol, F. Ozkinay, Z. Arikan *et al.*, "Highest heterogeneity for cystic fibrosis: 36 mutations account for 75% of all CF chromosomes in Turkish patients," *Am J Med Genet*, vol. 113, no. 3, pp. 250–7, 2002.

[148] C. T. Gurson, H. Sertel, M. Gurkan, and S. Pala, "Newborn screening for cystic fibrosis with the chloride electrode and neutron activation analysis," *Helv Paediatr Acta*, vol. 28, no. 2, pp. 165–74, 1973.

[149] M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, G. Mirzaa *et al.*, "GeneReviews [Internet]," 1993-2021, https://www.ncbi.nlm.nih.gov/books/NBK1116/, Accessed: 2021-10-22.

[150] A. El-Metwally, L. Yousef Al-Ahaidib, A. Ayman Sunqurah, K. Al-Surimi, M. Househ, A. Alshehri *et al.*, "The prevalence of phenylketonuria in Arab countries, Turkey, and Iran: A Systematic Review," *Biomed Res Int*, vol. 2018, p. 7697210, 2018.

[151] T. Guran, B. Tezel, M. Cakir, A. Akinci, Z. Orbak, M. Keskin *et al.*, "Neonatal screening for congenital adrenal hyperplasia in Turkey: Outcomes of Extended Pilot Study in 241,083 Infants," *J Clin Res Pediatr Endocrinol*, vol. 12, no. 3, pp. 287–94, 2020.

[152] H. Simsek, A. Pinar, A. Altinbas, A. Alp, Y. H. Balaban, Y. Buyukasik *et al.*, "Cutoff level to detect heterozygous alpha 1 antitrypsin deficiency in Turkish population," *J Clin Lab Anal*, vol. 25, no. 4, pp. 296–9, 2011.

[153] "Orphanet: An online database of rare diseases and orphan drugs. Copyright, INSERM," 1997, http://www.orpha.net, Accessed: 2021-11-3.

[154] E. Arslan Ates, A. Turkyilmaz, O. Yildirim, C. Alavanda, H. Polat, S. Demir *et al.*, "Secondary findings in 622 Turkish clinical exome sequencing data," *J Hum Genet*, vol. 66, no. 11, pp. 1113–9, 2021.

[155] O. Simsek Papur, S. A. Akman, R. Cakmur, and O. Terzioglu, "Mutation analysis of ATP7B gene in Turkish Wilson disease patients: Identification of five novel mutations," *Eur J Med Genet*, vol. 56, no. 4, pp. 175–9, 2013.

[156] National Organization for Rare Disorders, "Rare Disease Database," https://rarediseases.org/, Accessed: 2021-11-3.

[157] C. Espinos, A. Garcia-Cazorla, D. Martinez-Rubio, E. Martinez-Martinez, M. A. Vilaseca, B. Perez-Duenas *et al.*, "Ancient origin of the CTH alelle carrying the c.200C >T (p.T67I) variant in patients with cystathioninuria," *Clin Genet*, vol. 78, no. 6, pp. 554–9, 2010.

[158] B. Schoser, P. Lafort, M. E. Kruijshaar, A. Toscano, P. A. van Doorn, A. T. van der Ploeg *et al.*, "Minutes of the European POmpe Consortium (EPOC) Meeting March 27 to 28, 2015, Munich, Germany," *Acta Myol*, vol. 34, no. 2-3, pp. 141–3, 2015.

[159] K. Hopp, A. G. Cogal, E. J. Bergstralh, B. M. Seide, J. B. Olson, A. M. Meek *et al.*, "Phenotype-genotype correlations and estimated carrier frequencies of primary Hyperoxaluria," *J Am Soc Nephrol*, vol. 26, no. 10, pp. 2559–70, 2015.

[160] Y. Yildiz, M. Arslan, G. Celik, C. Kasapkara, S. Ceylaner, A. Dursun *et al.*, "Genotypes and estimated prevalence of phosphomannomutase 2 deficiency in Turkey differ significantly from those in Europe," *Am J Med Genet A*, vol. 182, no. 4, pp. 705–712, 2020.

[161] M. M. Kaback, *Tay Sachs Disease*, in Encyclopedia of Genetics, 1st ed., S Brenner and J.H. Miller, Eds., New York: Elsevier, 2001, pp. 1941-43.

[162] S. Kesici, S. Ünal, B. Kuskonmaz, S. Aytac, M. Cetin, and F. Gumruk, "Fanconi anemia: A single center experience of a large cohort," *Turk J Pediatr*, vol. 61, no. 4, 2019.

[163] C. A. Wassif, J. L. Cross, J. Iben, L. Sanchez-Pulido, A. Cougnoux, F. M. Platt *et al.*, "High incidence of unrecognized visceral/neurological late-onset Niemann-Pick disease, type C1, predicted by analysis of massively parallel sequencing data sets," *Genet Med*, vol. 18, no. 1, pp. 41–8, 2016.

[164] M. Feldman, E. Fernndez-Domnguez, L. Reynolds, D. Baird, J. Pearson, I. Hershkovitz *et al.*, "Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia," *Nat Commun*, vol. 10, no. 1, p. 1218, 2019.

[165] G. M. Kilinc, D. Koptekin, C. Atakuman, A. P. Sumer, H. M. Donertas, R. Yaka *et al.*, "Archaeogenomic analysis of the first steps of Neolithization in Anatolia and the Aegean," *Proc Biol Sci*, vol. 284, no. 1867, pp. 2017–64, 2017.

[166] A. Omrak, T. Gunther, C. Valdiosera, E. M. Svensson, H. Malmstrom, H. Kiesewetter *et al.*, "Genomic evidence establishes Anatolia as the source of the European neolithic gene pool," *Curr Biol.*, vol. 26, no. 2, pp. 270–5, 2016.

[167] C. C. Clay, "Labour migration and economic conditions in nineteenth-century Anatolia," *Middle East Stud*, vol. 34, no. 4, pp. 1–32, 1998.

[168] A. Icduygu, S. Toktas, and B. A. Soner, "The politics of population in a nation-building process: Emigration of non-Muslims from Turkey," *Ethn Racial Stud*, vol. 31, no. 2, pp. 358–89, 2008.

[169] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg *et al.*, "Genome-wide patterns of selection in 230 ancient Eurasians," *Nature*, vol. 528, no. 7583, pp. 499–503, 2015.

[170] A. Raveane, S. Aneli, F. Montinaro, G. Athanasiadis, S. Barlera, G. Birolo *et al.*, "Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe," *Sci Adv*, vol. 5, no. 9, p. eaaw3492, 2019.

[171] G. M. Kilinc, A. Omrak, F. Ozer, T. Gunther, A. M. Buyukkarakaya, E. Bicakci *et al.*, "The demographic development of the first farmers in Anatolia," *Curr Biol.*, vol. 26, no. 19, pp. 2659–66, 2016.

[172] M. Richards, V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida *et al.*, "Tracing European founder lineages in the Near Eastern mtDNA pool," *Am J Hum Genet*, vol. 67, no. 5, pp. 1251–76, 2000.

[173] W. Amos and J. I. Hoffman, "Evidence that two main bottleneck events shaped modern human genetic diversity," *Proc of the Royal Soc B: Biological Sciences*, vol. 277, no. 1678, pp. 131–7, 2010.

[174] E. Tuncbilek, "Genetic services in Turkey," *Eur J Hum Genet*, vol. 5 Suppl 2, pp. 178–82, 1997.

[175] N. Marchi, P. Mennecier, M. Georges, S. Lafosse, T. Hegay, C. Dorzhu *et al.*, "Close inbreeding and low genetic diversity in inner asian human populations despite geographical exogamy," *Sci Rep*, vol. 8, no. 1, p. 9397, 2018.

[176] T. J. Pemberton and N. A. Rosenberg, "Population-genetic influences on genomic estimates of the inbreeding coefficient: A global perspective," *Hum Hered*, vol. 77, no. 1-4, pp. 37–48, 2014.

[177] V. M. Narasimhan, Y. Xue, and C. Tyler-Smith, "Human knockout carriers: Dead, diseased, healthy, or improved?" *Trends Mol Med*, vol. 22, no. 4, pp. 341–51, 2016.

[178] Y. Xue, Y. Chen, Q. Ayub, N. Huang, E. V. Ball, M. Mort *et al.*, "Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing," *Am J Hum Genet*, vol. 91, no. 6, pp. 1022–32, 2012.

[179] D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, and H. Kehrer-Sawatzki, "Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease," *Hum Genet*, vol. 132, no. 10, pp. 1077–130, 2013.

[180] C. F. Wright, R. Y. Eberhardt, P. Constantinou, M. E. Hurles, D. R. FitzPatrick, H. V. Firth *et al.*, "Evaluating variants classified as pathogenic in ClinVar in the DDD Study," *Genet Med*, vol. 23, no. 3, pp. 571–5, 2021.

[181] M. Abouelhoda, T. Faquih, M. El-Kalioby, and F. S. Alkuraya, "Revisiting the morbid genome of Mendelian disorders," *Genome Biol*, vol. 17, no. 1, pp. 235–5, 2016.

[182] O. E. Onat, M. E. Kars, S. Gul, K. Bilguvar, Y. Wu, A. Ozhan *et al.*, "Human CRY1 variants associate with attention deficit/hyperactivity disorder," *J Clin Invest*, vol. 130, no. 7, pp. 3885–900, 2020.

[183] N. Shah, Y. C. Hou, H. C. Yu, R. Sainger, C. T. Caskey, J. C. Venter *et al.*, "Identification of misclassified ClinVar variants via disease population prevalence," *Am J Hum Genet*, vol. 102, no. 4, pp. 609–19, 2018.

[184] Y. H. Park, E. F. Remmers, W. Lee, A. K. Ombrello, L. K. Chung, Z. Shilei *et al.*, "Ancient familial Mediterranean fever mutations in human pyrin and resistance to Yersinia pestis," *Nat Immunol*, vol. 21, no. 8, pp. 857–67, 2020.

[185] S. Livadas and C. Bothou, "Management of the female with non-classical congenital adrenal hyperplasia (NCCAH): A patient-oriented approach," *Front Endocrinol (Lausanne)*, vol. 10, p. 366, 2019.

[186] I. Ozalp, T. Coskun, A. Tokatli, H. S. Kalkanoglu, A. Dursun, S. Tokol *et al.*, "Newborn PKU screening in Turkey: At present and organization for future," *Turk J Pediatr*, vol. 43, no. 2, pp. 97–101, 2001.

[187] O. Oz, E. D. Akbulut, M. E. Karadag, A. Gonel, and . Koyuncu, "Amino acid metabolism disorders and PAH gene mutations in Southeastern Anatolia Region," *Turkish J Biochem*, vol. 46, no. 4, pp. 387–92, 2021.

[188] M. Bonyadi, O. Omrani, S. M. Moghanjoghi, and S. Shiva, "Mutations of the phenylalanine hydroxylase gene in Iranian Azeri Turkish patients with phenylketonuria," *Genet Test Mol Biomarkers*, vol. 14, no. 2, pp. 233–5, 2010.

[189] D. Zamanfar, H. Jalali, M. R. Mahdavi, M. Maadanisani, H. Zaeri, and E. Asadpoor, "Investigation of five common mutations on phenylalanine hydroxylase gene of phenylketonuria patients from two provinces in north of Iran," *Int J Prev Med*, vol. 8, p. 89, 2017.

[190] K. B. El-Jaick, S. E. Powers, L. Bartholin, K. R. Myers, J. Hahn, I. M. Orioli *et al.*, "Functional analysis of mutations in TGIF associated with holoprosencephaly," *Mol Genet Metab*, vol. 90, no. 1, pp. 97–111, 2007.

# Appendix A

# Data

**Table A.1** Variants in the genes that are causally associated with the phenotypes in the study (Separate file)

**Table A.2** Chromosome Y and mtDNA haplogroups of the TR samples

| SampleID | Gender | Mt DNA Haplogroup | Chr Y Haplogroup |
|---|---|---|---|
| TR-B1 | Male | H5a3a2 | R1b1a1b1b3a1a1 |
| TR-B2 | Male | L5a1 | R1b1a1b1b3a1a1 |
| TR-B3 | Male | T1a1l | R1b1a1b1a1a1b1a1a |
| TR-B4 | Male | H5e1 | R1a1a1b2 |
| TR-B5 | Male | H44b | R1a1a1b1a2b3a3a2 |
| TR-B6 | Male | H1 | R1a1a1b1a2b3a |
| TR-B7 | Male | H55+153 | R1a1a1b1a2b3a |
| TR-B8 | Male | U5b1b1a | R1a1a1b1a2a |
| TR-B9 | Male | V3 | R1a1a1b1a1a1c |
| TR-B10 | Male | U3b | N1a1a1a1a1a1a |
| TR-B11 | Male | K1a7 | L1a2 |
| TR-B12 | Male | U5b2a5 | L1a2 |
| TR-B13 | Male | H11a2 | J2b2a1a1a1a1b |
| TR-B14 | Male | U5b1b1a | J2b2a1a1a1a1a |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-B15 | Male | I1b | J2a1a1a2b2a3b1b2a1a |
| TR-B16 | Male | H7a1a | J2a1a1a2b2a2 |
| TR-B17 | Male | H13a2b | J1a2a1a2d2b2b2c4b |
| TR-B18 | Male | T2a1b1a | J1a2a1a2d2b2b2c4a |
| TR-B19 | Male | N1a1a | I2a1a2b1a1a1d |
| TR-B20 | Male | HV+16311 | I2a1a2b1a1a |
| TR-B21 | Male | R2 | I2a1a2b1a1a |
| TR-B22 | Male | U4b | I2a1a2b1a1a |
| TR-B23 | Male | U5a2b | I2a1a2b1a1a |
| TR-B24 | Male | U5b1c2 | I2a1a2b1a1a |
| TR-B25 | Male | X2+225+@16223 | I2a1a2b1a1a |
| TR-B26 | Male | H1b2 | I1a2a1a1d2a1a |
| TR-B27 | Male | H1h1 | I1a2a1a1d2a1a |
| TR-B28 | Male | U5a1a2a | G2b2b |
| TR-B29 | Male | H3h5 | G2a2b2a1a1a1a1 |
| TR-B30 | Male | U2e1h | G2a2b2a1a1a1 |
| TR-B31 | Male | K1c1 | E1b1b1b2a1a6 |
| TR-B32 | Male | H | E1b1b1b2a1a1a1a1f1b1a3 |
| TR-B33 | Male | H1+152 | E1b1b1a1b1a |
| TR-B34 | Male | H44b | E1b1b1a1b1a |
| TR-B35 | Male | N1b1a | E1b1b1a1b1a |
| TR-C1 | Male | B4c1b2 | T1a1a1b2b2b1a1a |
| TR-C2 | Male | J1c | T1a1a1 |
| TR-C3 | Male | T1 | R1b1a1b1b3 |
| TR-C4 | Male | F1b1f | R1b1a1b1b |
| TR-C5 | Male | T2 | R1b1a1b1b |
| TR-C6 | Male | X2+225+@16223 | R1b1a1b1b |
| TR-C7 | Male | H13c1a | R1b1a1b |
| TR-C8 | Male | J2b1 | R1a1a1b2a2a1 |
| TR-C9 | Male | HV4 | R1a1a1b2 |
| TR-C10 | Male | H1c+152 | R1a1a1b1a1a1c |
| TR-C11 | Male | U2d2 | R1a1a1b1a |
| TR-C12 | Male | K1a17 | Q2b3a |
| TR-C13 | Male | I5c | Q2 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-C14 | Male | U1a1a | Q2 |
| TR-C15 | Male | K1a+195 | Q1b2a1a1a |
| TR-C16 | Male | T | O2a1b1a1 |
| TR-C17 | Male | U7a3a | L1a2 |
| TR-C18 | Male | U5b1b1 | J2a1a1b1a |
| TR-C19 | Male | U2e1h | J2a1a1a2b2a3b1b2 |
| TR-C20 | Male | U5b1e1 | J2a1a1a2b2a2b3a |
| TR-C21 | Male | I5a2 | J2a1a1a2b2a1a1c2 |
| TR-C22 | Male | H7 | J2a1a1a2b2a1a1 |
| TR-C23 | Male | R0a1a | J2a1a1a2a2b2d1 |
| TR-C24 | Male | X2+225+@16223 | J2a1a1a2a2b2d1 |
| TR-C25 | Male | U1b2 | J2a1a1a2a1a |
| TR-C26 | Male | J1b1b1 | J1b1a1a |
| TR-C27 | Male | H7 | J1a2a1a2d2b2b2c4b |
| TR-C28 | Male | H13a2c | I2a1a2b1a1a |
| TR-C29 | Male | H1ap1 | I1a2a1a1d2a1a |
| TR-C30 | Male | W6 | I1a1b1a4a1 |
| TR-C31 | Male | H7a1a | H1a1a4b2 |
| TR-C32 | Male | H107 | H1a1a4b |
| TR-C33 | Male | Z3a1 | G2a2b2a4a1b1 |
| TR-C34 | Male | J1b1b1 | G2a2b1a1b |
| TR-C35 | Male | H1c+152 | G2a1a |
| TR-C36 | Male | J1b8 | G2a1a |
| TR-C37 | Male | D4g2b | G1b |
| TR-C38 | Male | K1a12a | E1b1b1b2a1a1a1a1f1b1a1 |
| TR-C39 | Male | W3b | E1b1b1b2a1a1a1a1f1b1a1 |
| TR-C40 | Male | HV1a'b'c | E1b1b1a1a1c1a |
| TR-C41 | Male | H101 | C1b1a1a1 |
| TR-E1 | Male | HV12a1 | T1a2b1 |
| TR-E2 | Male | T2g1 | T1a1a1b2b2b1a1a |
| TR-E3 | Male | X2p | T1a1a1b2b2b1a1a |
| TR-E4 | Male | U4a1 | R1b1a2 |
| TR-E5 | Male | U7 | R1b1a1b1b3a |
| TR-E6 | Male | T1b3 | R1b1a1b1b3 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-E7 | Male | U1a2 | R1b1a1b1b3 |
| TR-E8 | Male | H107 | R1b1a1b1b |
| TR-E9 | Male | H11 | R1b1a1b1b |
| TR-E10 | Male | H13a2b4 | R1b1a1b1b |
| TR-E11 | Male | I1 | R1b1a1b1b |
| TR-E12 | Male | J2b1 | R1b1a1b1b |
| TR-E13 | Male | K1a4 | R1b1a1b1b |
| TR-E14 | Male | T2c1a | R1b1a1b1b |
| TR-E15 | Male | U1a1a | R1b1a1b1b |
| TR-E16 | Male | U2d1 | R1b1a1b1b |
| TR-E17 | Male | U3b1 | R1b1a1b1b |
| TR-E18 | Male | H4 | R1b1a1b1a1a2b |
| TR-E19 | Male | U4b1b1 | R1b1a1b1a1a1c2b2b1a1a |
| TR-E20 | Male | H5 | R1b1a1b |
| TR-E21 | Male | HV12b1 | R1b1a1b |
| TR-E22 | Male | HV1b3b | R1b1a1b |
| TR-E23 | Male | N1b1a5 | R1b1a1b |
| TR-E24 | Male | U4'9 | R1a2a |
| TR-E25 | Male | R2 | R1a1b |
| TR-E26 | Male | H1j1a | R1a1a1b2a3b |
| TR-E27 | Male | T | R1a1a1b2a3 |
| TR-E28 | Male | T2e | R1a1a1b2a2a2 |
| TR-E29 | Male | K1a12a1a | R1a1a1b2a2a1d9 |
| TR-E30 | Male | U1a1a3 | R1a1a1b2a2a1c3 |
| TR-E31 | Male | K1b1c | R1a1a1b2a1 |
| TR-E32 | Male | U2e1a1 | R1a1a1b2a1 |
| TR-E33 | Male | H4a1 | R1a1a1b2 |
| TR-E34 | Male | HV14a | R1a1a1b2 |
| TR-E35 | Male | J1c | R1a1a1b2 |
| TR-E36 | Male | H5m | R1a1a1b1a |
| TR-E37 | Male | R0a1a | R1a1a1b1a |
| TR-E38 | Male | U8b1a1 | R1a1a1b1a |
| TR-E39 | Male | W4a | R1a1a1b1a |
| TR-E40 | Male | HV1b3b | Q2b |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-E41 | Male | J1b1a | Q2 |
| TR-E42 | Male | J1b1a2a | Q2 |
| TR-E43 | Male | T1b3 | Q2 |
| TR-E44 | Male | T2a | Q2 |
| TR-E45 | Male | H+152 | N1a2a1a |
| TR-E46 | Male | H15a1 | N1a2a1a |
| TR-E47 | Male | R0a | N1a2a1a |
| TR-E48 | Male | H14a | L1a2 |
| TR-E49 | Male | HV4c | L1a2 |
| TR-E50 | Male | HV6 | L1a2 |
| TR-E51 | Male | R2 | L1a2 |
| TR-E52 | Male | T1 | L1a2 |
| TR-E53 | Male | T2d2 | L1a2 |
| TR-E54 | Male | U2d2 | L1a2 |
| TR-E55 | Male | U5a1a1 | L1a2 |
| TR-E56 | Male | X2 | L1a2 |
| TR-E57 | Male | I5a2 | J2b2a1a1a1 |
| TR-E58 | Male | H2a3 | J2b1b |
| TR-E59 | Male | U3b | J2b1b |
| TR-E60 | Male | W9 | J2b1b |
| TR-E61 | Male | H107 | J2a2 |
| TR-E62 | Male | J1c17 | J2a2 |
| TR-E63 | Male | U3b | J2a2 |
| TR-E64 | Male | K1a12 | J2a1a1b2a1a2a1 |
| TR-E65 | Male | M4 | J2a1a1b2a1a2a1 |
| TR-E66 | Male | U2e1b2 | J2a1a1b2a1a2a1 |
| TR-E67 | Male | K1a19 | J2a1a1b1a1a |
| TR-E68 | Male | N1a1a1a1 | J2a1a1b1a1a |
| TR-E69 | Male | T2e | J2a1a1b1a1a |
| TR-E70 | Male | H92 | J2a1a1a2b2a3b1b |
| TR-E71 | Male | K1a | J2a1a1a2b2a1a1c2 |
| TR-E72 | Male | T2b | J2a1a1a2b2a1a1c2 |
| TR-E73 | Male | U1a1b | J2a1a1a2b2a1a1a |
| TR-E74 | Male | H | J2a1a1a2b2a1 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-E75 | Male | R1a | J2a1a1a2b1b3a2 |
| TR-E76 | Male | H1c3 | J2a1a1a2b1b3 |
| TR-E77 | Male | HV1b | J2a1a1a2a2b2d1 |
| TR-E78 | Male | U2e1a1 | J2a1a1a2a2b2a |
| TR-E79 | Male | H | J2a1a1a2a2b |
| TR-E80 | Male | X2+225+@16223 | J2a1a1a2a1a |
| TR-E81 | Male | U5a1a1 | J1a2b1 |
| TR-E82 | Male | U3b | J1a2a1a2d2b2b2c4b |
| TR-E83 | Male | HV7 | J1a2a1a2d2b2b2c4a |
| TR-E84 | Male | J1b1a3 | J1a2a1a2d2b2b2c4a |
| TR-E85 | Male | X2n | J1a2a1a2d2b2b2c4a |
| TR-E86 | Male | H7 | J1a2a1a2d2b2b2c4 |
| TR-E87 | Male | J1b1b1a | J1a2a1a2d2b2b2c4 |
| TR-E88 | Male | J2a1 | J1a2a1a2d2b2b2c4 |
| TR-E89 | Male | U7b | J1a2a1a2d2b2b2c4 |
| TR-E90 | Male | H | I2a1a2b1a1a |
| TR-E91 | Male | H5m | H1a1a4b |
| TR-E92 | Male | H6a1a | G2b2 |
| TR-E93 | Male | D4g1b | G2a2b2a4a1b1 |
| TR-E94 | Male | J2b1f | G2a2b2a1a1a1 |
| TR-E95 | Male | J1d6 | G2a2a |
| TR-E96 | Male | K1a31 | G2a2a |
| TR-E97 | Male | U1a1a | G2a2a |
| TR-E98 | Male | J2a1a1 | G2a1a |
| TR-E99 | Male | H | E1b1b1b2a1a1a |
| TR-E100 | Male | H+195+146 | E1b1b1b2a1a1a |
| TR-E101 | Male | H13a2c | E1b1b1b2a1a1a |
| TR-E102 | Male | HV+16311 | E1b1b1b2a1a1a |
| TR-E103 | Male | U1a1a | E1b1b1b2a1a1a |
| TR-E104 | Male | U7 | E1b1b1a1b2a3 |
| TR-E105 | Male | J1d6 | E1b1b1a1b1a9 |
| TR-E106 | Male | G2a2a | E1b1b1a1b1a |
| TR-E107 | Male | R2 | D1a1b1a1 |
| TR-E108 | Male | H | C2a1a1b1 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-E109 | Male | U3b2a1 | C1b1a1a1a |
| TR-E110 | Male | R0a1a | B2b1a1 |
| TR-N1 | Male | K1b1c | T1a2 |
| TR-N2 | Male | W3a1d | R1b1a2a1 |
| TR-N3 | Male | J1b1a | R1b1a2 |
| TR-N4 | Male | T2h | R1b1a1b1b |
| TR-N5 | Male | U1a1a+16129 | R1b1a1b1b |
| TR-N6 | Male | U3b2a1 | R1b1a1b1b |
| TR-N7 | Male | U7b | R1b1a1b1b |
| TR-N8 | Male | X2+225+@16223 | R1b1a1b1b |
| TR-N9 | Male | H13a1a1 | R1b1a1b1a1a2b |
| TR-N10 | Male | N1b1a7 | R1b1a1b1a1a2a |
| TR-N11 | Male | HV1a1 | R1b1a1b1a1a1b1a1a |
| TR-N12 | Male | T | R1b1a1b1a1a1b1a1a |
| TR-N13 | Male | X2d1 | R1a1a1b2a4 |
| TR-N14 | Male | H5m | R1a1a1b2a2a1c3 |
| TR-N15 | Male | I5c | R1a1a1b2a2a1 |
| TR-N16 | Male | U7 | R1a1a1b2a2a1 |
| TR-N17 | Male | A12 | R1a1a1b2 |
| TR-N18 | Male | H+195+146 | R1a1a1b2 |
| TR-N19 | Male | M7b1a1a1 | R1a1a1b1a2b3a3a |
| TR-N20 | Male | U4c1 | R1a1a1b1a1a1c |
| TR-N21 | Male | N1a2 | R1a1a1b1a |
| TR-N22 | Male | N1a3a | R1a1a1b1a |
| TR-N23 | Male | U1a1a+16129 | R1a1a1b1a |
| TR-N24 | Male | T1a1 | O2a1a1b1 |
| TR-N25 | Male | T2g | NO1 |
| TR-N26 | Male | H13a2b3 | N1a2a1a |
| TR-N27 | Male | H4c | N1a2a1a |
| TR-N28 | Male | J1d1b1 | N1a2a1a |
| TR-N29 | Male | U3c | N1a2a1a |
| TR-N30 | Male | H47 | L1a2 |
| TR-N31 | Male | HV4a2a | L1a2 |
| TR-N32 | Male | R0a2k | L1a2 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-N33 | Male | U1a1a | L1a2 |
| TR-N34 | Male | W3a1 | J2b2a2b |
| TR-N35 | Male | U3b2c | J2b2a2a1 |
| TR-N36 | Male | J1c2i | J2b1 |
| TR-N37 | Male | H46 | J2a2 |
| TR-N38 | Male | X2i | J2a1a1b2a1b1b3a2 |
| TR-N39 | Male | H | J2a1a1b2a1b1b3a |
| TR-N40 | Male | W3b | J2a1a1b2a1b1b2 |
| TR-N41 | Male | K1a19 | J2a1a1b2a1b1b |
| TR-N42 | Male | W3 | J2a1a1b2a1b |
| TR-N43 | Male | H2a1 | J2a1a1b2a1a2 |
| TR-N44 | Male | H79a | J2a1a1b2a1a2 |
| TR-N45 | Male | H14a2 | J2a1a1b1a |
| TR-N46 | Male | K1a12a | J2a1a1a2b2a3b1b2 |
| TR-N47 | Male | W6c1a | J2a1a1a2b2a3b1a1 |
| TR-N48 | Male | N1a2 | J2a1a1a2b2a2 |
| TR-N49 | Male | N9a1 | J2a1a1a2b2a1a1c2b1 |
| TR-N50 | Male | X2n | J2a1a1a2b2a1a1c2 |
| TR-N51 | Male | H1au | J2a1a1a2b2a1a1 |
| TR-N52 | Male | K1a+150 | J2a1a1a2b2a1a1 |
| TR-N53 | Male | HV1a1a | J2a1a1a2b2a1a |
| TR-N54 | Male | K1b1c | J2a1a1a2b2a1 |
| TR-N55 | Male | H92 | J2a1a1a2a2b |
| TR-N56 | Male | W3b | J2a1a1a2a2 |
| TR-N57 | Male | H20a | J1a2a1a2d2b2b2c4b1c3a |
| TR-N58 | Male | K1a8a | J1a2a1a2d2b2b2c4b |
| TR-N59 | Male | K1a8b | J1a2a1a2d2b2b2c4a |
| TR-N60 | Male | T1 | J1a2a1a2d2b2b2c4 |
| TR-N61 | Male | H5 | J1a2a1a2d2b2 |
| TR-N62 | Male | HV4a2b | J1a2a1a2d |
| TR-N63 | Male | T2h | J1a2a1 |
| TR-N64 | Male | K1a | I2a1a2b1a1a2a |
| TR-N65 | Male | H10 | I2a1a2b1a1a |
| TR-N66 | Male | H1c | I2a1a2b1a1a |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-N67 | Male | U1a1c1 | I2a1a2b1a1a |
| TR-N68 | Male | H29a | G2b2 |
| TR-N69 | Male | I1c1a | G2a2b2a1a1a1 |
| TR-N70 | Male | U5a1f1 | G2a2b2a1a1a1 |
| TR-N71 | Male | H2a3 | G2a2b2a1 |
| TR-N72 | Male | H33a | G2a2b2a1 |
| TR-N73 | Male | U3b2a1 | G2a2b1a1b |
| TR-N74 | Male | K1a4f | G2a2b1a |
| TR-N75 | Male | H13a1a2b | G2a2a |
| TR-N76 | Male | H1at | G2a2a |
| TR-N77 | Male | H5 | G1a1a2a1c |
| TR-N78 | Male | H5m | G1a1a2a1c |
| TR-N79 | Male | U3a3 | G1a1a2a1c |
| TR-N80 | Male | K1a28 | E1b1b1b2a1a6 |
| TR-N81 | Male | H1q | E1b1b1b2a1a1a1a1f1b1a3 |
| TR-N82 | Male | H14a2 | E1b1b1b2a1a1a1a1f1b1a1 |
| TR-N83 | Male | R1a | E1b1b1b2a1a1a1a1f1b1a1 |
| TR-N84 | Male | H | E1b1b1b2a1a1a |
| TR-N85 | Male | H1c13 | E1b1b1a1b1a6 |
| TR-N86 | Male | HV1a1 | E1b1b1a1b1a |
| TR-N87 | Male | HV1b3b | E1b1b1a1b1a |
| TR-N88 | Male | H5 | C2a1a1b1 |
| TR-N89 | Male | H70 | C2a1a1b1 |
| TR-N90 | Male | HV1a'b'c | C2a1a1b1 |
| TR-N91 | Male | T2b | C2a |
| TR-N92 | Male | H20 | C1a2 |
| TR-S1 | Male | H13a2c1 | T1a1a1b2b2b1a1a |
| TR-S2 | Male | T1a1b | T1a1a1b2b2b |
| TR-S3 | Male | U8b1a1 | T1a1a |
| TR-S4 | Male | J1b2 | R1b1a1b1b3 |
| TR-S5 | Male | H7 | R1b1a1b1b |
| TR-S6 | Male | H13a1c | R1b1a1b1a1a1c1b |
| TR-S7 | Male | H2a1 | R1b1a1b |
| TR-S8 | Male | W3a1 | R1a1b |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-S9 | Male | N2a1 | R1a1a1b2a2b1a2 |
| TR-S10 | Male | H+152 | R1a1a1b2a2a2 |
| TR-S11 | Male | W | R1a1a1b2a2a1d7 |
| TR-S12 | Male | I4a | R1a1a1b2a2a1c3 |
| TR-S13 | Male | H | R1a1a1b2a2a1b1 |
| TR-S14 | Male | T2d2 | Q2 |
| TR-S15 | Male | HV1a2 | N1a2a1a |
| TR-S16 | Male | I2 | N1a2a1a |
| TR-S17 | Male | X2n | N1a2a1a |
| TR-S18 | Male | T2c1f | L1a1b3a1b |
| TR-S19 | Male | U1a1a1 | L1a1b |
| TR-S20 | Male | H13a1c | J2b2a1a1a1a1a1b1 |
| TR-S21 | Male | H | J2a1a1b2a1b1b2 |
| TR-S22 | Male | H15a1 | J2a1a1b2a1a2 |
| TR-S23 | Male | D4j6 | J2a1a1a2b2a2 |
| TR-S24 | Male | H1c | J2a1a1a2b2a1a1c2 |
| TR-S25 | Male | H7 | J2a1a1a2b2a1a1c2 |
| TR-S26 | Male | H51 | J2a1a1a2b2a1a1a |
| TR-S27 | Male | HV1b3b | J1a3b1 |
| TR-S28 | Male | U3b | J1a2a1a2d2b2b2c4a |
| TR-S29 | Male | C4d | J1a2a1a2d2b2b2c4 |
| TR-S30 | Male | J1b2 | I2a1b1 |
| TR-S31 | Male | J1c3k | I1a2a1a1d2a1a |
| TR-S32 | Male | J1c15a | G2a2b2a1a1a1b1b1b1 |
| TR-S33 | Male | K1a4j | G2a2b1 |
| TR-S34 | Male | T2b | G2a2a |
| TR-S35 | Male | C4b | G2a1a |
| TR-S36 | Male | H | G2a1a |
| TR-S37 | Male | U1b1 | E1b1b1b2a1a1a |
| TR-S38 | Male | H | E1b1b1a1b2a3 |
| TR-S39 | Male | J1c2e1 | E1b1b1a1b1a |
| TR-S40 | Male | T2g1b | C2a1a1b1 |
| TR-U1 | Male | T2a1a | T1a1a1b2b |
| TR-U2 | Male | J1b8 | R1b1a2 |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-U3 | Male | U2e1a1 | R1b1a1b1b3a1a1 |
| TR-U4 | Male | U5a2a1 | R1b1a1b1b3a1a1 |
| TR-U5 | Male | F1b1e | R1b1a1b1b3a |
| TR-U6 | Male | H5 | R1b1a1b1b3 |
| TR-U7 | Male | H | R1b1a1b1b |
| TR-U8 | Male | T2e+152 | R1b1a1b1b |
| TR-U9 | Male | U2e1e | R1b1a1b1b |
| TR-U10 | Male | U3b | R1b1a1b1b |
| TR-U11 | Male | U5a1+@16192 | R1b1a1b1b |
| TR-U12 | Male | K1a4j1 | R1b1a1b1a1a1b1a1a |
| TR-U13 | Male | K1a+150 | R1b1a1b |
| TR-U14 | Male | U2d2 | R1b1a1b |
| TR-U15 | Male | U3b | R1b1a1b |
| TR-U16 | Male | U4b3 | R1b1a1b |
| TR-U17 | Male | J1c | R1a1a1b2a2a2 |
| TR-U18 | Male | U3b3 | R1a1a1b2a2a2 |
| TR-U19 | Male | U7 | R1a1a1b2a2a2 |
| TR-U20 | Male | H71 | R1a1a1b2 |
| TR-U21 | Male | U2+152 | R1a1a1b2 |
| TR-U22 | Male | T1b | R1a1a1b1a2b |
| TR-U23 | Male | U4b2a1 | R1a1a1b1a2b |
| TR-U24 | Male | J2a2a | N1a2a1a |
| TR-U25 | Male | B4a1+16311 | L1a2 |
| TR-U26 | Male | HV0 | L1a2 |
| TR-U27 | Male | U3b1a1 | L1a2 |
| TR-U28 | Male | U5a2b | L1a2 |
| TR-U29 | Male | J1b3a | J2b2 |
| TR-U30 | Male | K1a19 | J2a2 |
| TR-U31 | Male | K2a5 | J2a1a1b2a1b1b3a |
| TR-U32 | Male | U7b | J2a1a1b2a1b |
| TR-U33 | Male | H14b4 | J2a1a1b2a1a2 |
| TR-U34 | Male | H1e5b | J2a1a1b2a1a2 |
| TR-U35 | Male | K1a17a | J2a1a1b2a1a1 |
| TR-U36 | Male | D4j12 | J2a1a1b1a |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-U37 | Male | J1b5 | J2a1a1a2b2a3 |
| TR-U38 | Male | H1cf | J2a1a1a2b2a1a1 |
| TR-U39 | Male | U1a1a | J2a1a1a2b2a1a1 |
| TR-U40 | Male | K1a | J2a1a1a2a2b |
| TR-U41 | Male | U7 | J1b1a1a |
| TR-U42 | Male | H13a2b4 | J1a2b1b2 |
| TR-U43 | Male | H14a2c | J1a2b1b2 |
| TR-U44 | Male | M30b | J1a2b1b2 |
| TR-U45 | Male | H47 | J1a2b1 |
| TR-U46 | Male | H26b | J1a2a1a2d2b2b2c4b |
| TR-U47 | Male | T2g1b | J1a2a1a2d2b2b2c4b |
| TR-U48 | Male | H1t | J1a2a1a2d2b2b2c4 |
| TR-U49 | Male | K1a+150 | J1a2a1a2d2b2b2c2a |
| TR-U50 | Male | L0a2c | J1a2a1a2d2b2 |
| TR-U51 | Male | U2e1b2 | J1a2a1a2c1 |
| TR-U52 | Male | H6a1a | I2a1b1a2b2 |
| TR-U53 | Male | X1 | I2a1b1a2a |
| TR-U54 | Male | U7a4a1 | H1a1b1a |
| TR-U55 | Male | X2 | G2b2 |
| TR-U56 | Male | K1a4 | G2a2b2a1a1a1 |
| TR-U57 | Male | H5 | G2a2a |
| TR-U58 | Male | T2b4 | G2a2a |
| TR-U59 | Male | D4e4a | G2a1a |
| TR-U60 | Male | U1a1a+16129 | G2a |
| TR-U61 | Male | H13a1a1 | E1b1b1b2a1a6 |
| TR-U62 | Male | U1a1a | E1b1b1b2a1a1a1a1f1b1a1 |
| TR-U63 | Male | K1a | E1b1b1b2a1a1a |
| TR-U64 | Male | T2h | E1b1b1b2a1a1a |
| TR-U65 | Male | U4b3 | E1b1b1b2a1a1a |
| TR-U66 | Male | T2 | E1b1b1a1b2a3 |
| TR-U67 | Male | HV4a2a | E1b1b1a1b1a6a1 |
| TR-U68 | Male | H6a1a | E1b1b1a1b1a10a2g |
| TR-U69 | Male | K1a | C2a1a2a |
| TR-W1 | Male | T2 | T1a1a1b2b2b1a1a |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-W2 | Male | H1 | T1a1a1b2 |
| TR-W3 | Male | T2b | R1b1a1b1b3a1a1a |
| TR-W4 | Male | H104 | R1b1a1b1b3a1a1 |
| TR-W5 | Male | T2d2 | R1b1a1b1b3a1a1 |
| TR-W6 | Male | HV+16311 | R1b1a1b1b |
| TR-W7 | Male | HV0e | R1b1a1b1b |
| TR-W8 | Male | T2b25 | R1b1a1b1b |
| TR-W9 | Male | H14a | R1b1a1b1a1a2c1a1e |
| TR-W10 | Male | K1b1b1 | R1b1a1b1a1a2b |
| TR-W11 | Male | H1 | R1a1a1b2a2a1 |
| TR-W12 | Male | V | R1a1a1b1a |
| TR-W13 | Male | U3b | Q1b1a |
| TR-W14 | Male | H14b | Q1a2a2 |
| TR-W15 | Male | U3b2a1 | N1a2a1a |
| TR-W16 | Male | Y1a1 | N1a2a1a |
| TR-W17 | Male | U5a2b | L1a2 |
| TR-W18 | Male | X2e2a2 | J2b2a1a1a1a1a1a1 |
| TR-W19 | Male | HV0 | J2a2 |
| TR-W20 | Male | L2'3'4'6+ | J2a2 |
| TR-W21 | Male | H6a2 | J2a1a1b2a1b1b3a |
| TR-W22 | Male | U2e1b1 | J2a1a1b2a1b1b3 |
| TR-W23 | Male | D4e4 | J2a1a1a2b2a3b1a1 |
| TR-W24 | Male | N1b1a8a | J2a1a1a2b2a2 |
| TR-W25 | Male | T2a1b1 | J2a1a1a2b2a2 |
| TR-W26 | Male | H1ap1 | J2a1a1a2b2a1a1 |
| TR-W27 | Male | H5 | J2a1a1a2b1 |
| TR-W28 | Male | T1 | J2a1a1a2a2 |
| TR-W29 | Male | I5 | J2a1 |
| TR-W30 | Male | T2a1b1 | J1a2b1b2 |
| TR-W31 | Male | W3a1 | J1a2a1a2d2b2b2c4b |
| TR-W32 | Male | T2e2 | J1a2a1a2d2b2 |
| TR-W33 | Male | X2+225+@16223 | J1a2a1a2d2b2 |
| TR-W34 | Male | W3b | J1a2a1 |
| TR-W35 | Male | H13a1c | I2a1a2b1a1a |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-W36 | Male | M5a1b | I2a1a2b1a1a |
| TR-W37 | Male | T1a7 | I2a1a2b1a1a |
| TR-W38 | Male | U5a1d | I2a1a2b1a1a |
| TR-W39 | Male | HV7 | I2a1a1a1a1a1a1c5 |
| TR-W40 | Male | HV18 | I1a2a1a1d2a1a |
| TR-W41 | Male | N1a1a1a1 | G2a2b2a1a1b1a1a1a1 |
| TR-W42 | Male | T2h | G2a2b2a1a1a1b2a |
| TR-W43 | Male | H1c4 | G2a2b2a1a1a1b1a1 |
| TR-W44 | Male | U5b1e1 | G2a2b2a1a1a1 |
| TR-W45 | Male | HV4 | G2a2b1a1b |
| TR-W46 | Male | T2d2 | G2a1a1a1a1a1a1 |
| TR-W47 | Male | H47 | G2a1a |
| TR-W48 | Male | H1j | G1a1a2a1c |
| TR-W49 | Male | H33 | E1b1b1b2a1a6c |
| TR-W50 | Male | V7a | E1b1b1a1b1a6a1 |
| TR-W51 | Male | H2a5b | E1b1b1a1b1a |
| TR-W52 | Male | H6a1a2a | E1b1b1a1b1a |
| TR-B36 | Female | H | - |
| TR-B37 | Female | H | - |
| TR-B38 | Female | H101 | - |
| TR-B39 | Female | H11a+152 | - |
| TR-B40 | Female | H13a1a1c | - |
| TR-B41 | Female | H1b | - |
| TR-B42 | Female | H20a2 | - |
| TR-B43 | Female | H28 | - |
| TR-B44 | Female | H44b | - |
| TR-B45 | Female | H55+153 | - |
| TR-B46 | Female | H5n | - |
| TR-B47 | Female | H6a2 | - |
| TR-B48 | Female | H70 | - |
| TR-B49 | Female | H87 | - |
| TR-B50 | Female | HV+16311 | - |
| TR-B51 | Female | HV9 | - |
| TR-B52 | Female | I1c1a | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-B53 | Female | J1c3 | - |
| TR-B54 | Female | K1a12a1a | - |
| TR-B55 | Female | K1a1a | - |
| TR-B56 | Female | K1a3a | - |
| TR-B57 | Female | K1a7 | - |
| TR-B58 | Female | K1b | - |
| TR-B59 | Female | N1a3a | - |
| TR-B60 | Female | U1a1a | - |
| TR-B61 | Female | U4c1 | - |
| TR-B62 | Female | U5a1a1a | - |
| TR-B63 | Female | U5a1d1 | - |
| TR-B64 | Female | U5a1d2a1 | - |
| TR-B65 | Female | U5a1g | - |
| TR-B66 | Female | U5a2a | - |
| TR-B67 | Female | W1c | - |
| TR-B68 | Female | W1h | - |
| TR-C42 | Female | C4a1a+195 | - |
| TR-C43 | Female | D4g2 | - |
| TR-C44 | Female | G2a2 | - |
| TR-C45 | Female | H+152 | - |
| TR-C46 | Female | H+195+146 | - |
| TR-C47 | Female | H1 | - |
| TR-C48 | Female | H10a | - |
| TR-C49 | Female | H11 | - |
| TR-C50 | Female | H13a2c | - |
| TR-C51 | Female | H14b | - |
| TR-C52 | Female | H1ak | - |
| TR-C53 | Female | H55+153 | - |
| TR-C54 | Female | H7b | - |
| TR-C55 | Female | H8b | - |
| TR-C56 | Female | I4 | - |
| TR-C57 | Female | J1c2e | - |
| TR-C58 | Female | N1b1a3 | - |
| TR-C59 | Female | T | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-C60 | Female | T1 | - |
| TR-C61 | Female | T1 | - |
| TR-C62 | Female | T2 | - |
| TR-C63 | Female | T2 | - |
| TR-C64 | Female | T2a1b | - |
| TR-C65 | Female | T2a3 | - |
| TR-C66 | Female | U1b2 | - |
| TR-C67 | Female | U3a3 | - |
| TR-C68 | Female | U3b1a1 | - |
| TR-C69 | Female | U4a2a | - |
| TR-C70 | Female | U5a1+@16192 | - |
| TR-C71 | Female | U5a1f1a | - |
| TR-C72 | Female | U5a1g | - |
| TR-C73 | Female | U7 | - |
| TR-C74 | Female | W3a1 | - |
| TR-C75 | Female | X2e2a2 | - |
| TR-E111 | Female | H | - |
| TR-E112 | Female | H | - |
| TR-E113 | Female | H | - |
| TR-E114 | Female | H | - |
| TR-E115 | Female | H+195+146 | - |
| TR-E116 | Female | H1 | - |
| TR-E117 | Female | H1+16355 | - |
| TR-E118 | Female | H101 | - |
| TR-E119 | Female | H13a2b4 | - |
| TR-E120 | Female | H29 | - |
| TR-E121 | Female | H44b | - |
| TR-E122 | Female | H46 | - |
| TR-E123 | Female | H5 | - |
| TR-E124 | Female | H6b2 | - |
| TR-E125 | Female | H8b1 | - |
| TR-E126 | Female | HV1b2 | - |
| TR-E127 | Female | I5a3 | - |
| TR-E128 | Female | J1b1a | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-E129 | Female | J2a2a | - |
| TR-E130 | Female | J2a2c | - |
| TR-E131 | Female | J2b1a2a | - |
| TR-E132 | Female | J2b1e | - |
| TR-E133 | Female | K1a | - |
| TR-E134 | Female | K1a19 | - |
| TR-E135 | Female | K1a31 | - |
| TR-E136 | Female | K1a4c1 | - |
| TR-E137 | Female | K1a4i | - |
| TR-E138 | Female | K1a4j1 | - |
| TR-E139 | Female | N1a3 | - |
| TR-E140 | Female | N1b1a8b | - |
| TR-E141 | Female | N3 | - |
| TR-E142 | Female | R30a1c | - |
| TR-E143 | Female | T1 | - |
| TR-E144 | Female | T1a1b1 | - |
| TR-E145 | Female | T1b3 | - |
| TR-E146 | Female | T2d2 | - |
| TR-E147 | Female | T2g | - |
| TR-E148 | Female | U1a1a | - |
| TR-E149 | Female | U1a1a1a | - |
| TR-E150 | Female | U2e1a1 | - |
| TR-E151 | Female | U3b | - |
| TR-E152 | Female | U3b | - |
| TR-E153 | Female | U3b1a1 | - |
| TR-E154 | Female | U3b2a1a | - |
| TR-E155 | Female | U4d2 | - |
| TR-E156 | Female | U5a1a1 | - |
| TR-E157 | Female | U7 | - |
| TR-E158 | Female | U8b1a1 | - |
| TR-E159 | Female | W3b | - |
| TR-E160 | Female | W6 | - |
| TR-E161 | Female | W6 | - |
| TR-E162 | Female | X2d1 | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-N93 | Female | D4j+16311 | - |
| TR-N94 | Female | G2a2 | - |
| TR-N95 | Female | H | - |
| TR-N96 | Female | H | - |
| TR-N97 | Female | H | - |
| TR-N98 | Female | H13a2b1 | - |
| TR-N99 | Female | H13a2b4 | - |
| TR-N100 | Female | H13a2c | - |
| TR-N101 | Female | H13c1a | - |
| TR-N102 | Female | H14a | - |
| TR-N103 | Female | H15b | - |
| TR-N104 | Female | H20 | - |
| TR-N105 | Female | H24 | - |
| TR-N106 | Female | H41a | - |
| TR-N107 | Female | H47 | - |
| TR-N108 | Female | H47 | - |
| TR-N109 | Female | H5 | - |
| TR-N110 | Female | H5 | - |
| TR-N111 | Female | H5’36 | - |
| TR-N112 | Female | H5f | - |
| TR-N113 | Female | H5m | - |
| TR-N114 | Female | H7 | - |
| TR-N115 | Female | H85 | - |
| TR-N116 | Female | HV | - |
| TR-N117 | Female | HV14a | - |
| TR-N118 | Female | HV1a1 | - |
| TR-N119 | Female | HV1a1a | - |
| TR-N120 | Female | HV1a1a | - |
| TR-N121 | Female | HV4a2a | - |
| TR-N122 | Female | I1b | - |
| TR-N123 | Female | I1b | - |
| TR-N124 | Female | I1c | - |
| TR-N125 | Female | I1c1a | - |
| TR-N126 | Female | J1b1b1 | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-N127 | Female | J1b1b1 | - |
| TR-N128 | Female | J1b3b | - |
| TR-N129 | Female | J1d1b1 | - |
| TR-N130 | Female | J1d3a | - |
| TR-N131 | Female | K1a+195 | - |
| TR-N132 | Female | K1a4c1 | - |
| TR-N133 | Female | K1a4j | - |
| TR-N134 | Female | N1a1a1 | - |
| TR-N135 | Female | N1b1a+195 | - |
| TR-N136 | Female | N1b1a3 | - |
| TR-N137 | Female | R0a2 | - |
| TR-N138 | Female | T1 | - |
| TR-N139 | Female | T1a10a | - |
| TR-N140 | Female | T2 | - |
| TR-N141 | Female | T2a1a1 | - |
| TR-N142 | Female | T2a1b1a1b | - |
| TR-N143 | Female | T2b | - |
| TR-N144 | Female | T2b4+152 | - |
| TR-N145 | Female | T2d2 | - |
| TR-N146 | Female | T2e | - |
| TR-N147 | Female | U1b1 | - |
| TR-N148 | Female | U1b2 | - |
| TR-N149 | Female | U2e1e | - |
| TR-N150 | Female | U3a'c | - |
| TR-N151 | Female | U3a'c | - |
| TR-N152 | Female | U3b2a1 | - |
| TR-N153 | Female | U3b2a1 | - |
| TR-N154 | Female | U4a2a | - |
| TR-N155 | Female | U4b | - |
| TR-N156 | Female | U4b1a1a1 | - |
| TR-N157 | Female | U5a1a1 | - |
| TR-N158 | Female | U5a1d2b | - |
| TR-N159 | Female | U5a2+16362 | - |
| TR-N160 | Female | U5b1b1+@16192 | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-N161 | Female | U7 | - |
| TR-N162 | Female | V+@16298 | - |
| TR-N163 | Female | X2n | - |
| TR-N164 | Female | X2o | - |
| TR-N165 | Female | X4 | - |
| TR-N166 | Female | Y1a1 | - |
| TR-S41 | Female | C4a1 | - |
| TR-S42 | Female | D4c2b | - |
| TR-S43 | Female | D4j | - |
| TR-S44 | Female | F1b1+@152 | - |
| TR-S45 | Female | H | - |
| TR-S46 | Female | H | - |
| TR-S47 | Female | H20a | - |
| TR-S48 | Female | H4b | - |
| TR-S49 | Female | H4b | - |
| TR-S50 | Female | HV0a | - |
| TR-S51 | Female | HV0f | - |
| TR-S52 | Female | HV1a1 | - |
| TR-S53 | Female | HV1b | - |
| TR-S54 | Female | HV1b3b | - |
| TR-S55 | Female | I5c1 | - |
| TR-S56 | Female | K1a+150 | - |
| TR-S57 | Female | K1a4 | - |
| TR-S58 | Female | K1a4a | - |
| TR-S59 | Female | R0a1a | - |
| TR-S60 | Female | R0a1a | - |
| TR-S61 | Female | T1a10 | - |
| TR-S62 | Female | T2c1 | - |
| TR-S63 | Female | T2c1a2 | - |
| TR-S64 | Female | T2h | - |
| TR-S65 | Female | U1a1a1a | - |
| TR-S66 | Female | U2e1a1 | - |
| TR-S67 | Female | U2e1e | - |
| TR-S68 | Female | U2e1h | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-S69 | Female | U3a3 | - |
| TR-S70 | Female | U3a3 | - |
| TR-S71 | Female | U3b | - |
| TR-S72 | Female | U5a1a1 | - |
| TR-S73 | Female | U6a1a1 | - |
| TR-S74 | Female | U7a4a1 | - |
| TR-S75 | Female | U7b | - |
| TR-S76 | Female | W1h | - |
| TR-S77 | Female | X2e2b | - |
| TR-S78 | Female | X2f | - |
| TR-S79 | Female | X4 | - |
| TR-U70 | Female | C5a1 | - |
| TR-U71 | Female | H11a1 | - |
| TR-U72 | Female | H11a2 | - |
| TR-U73 | Female | H13c1a | - |
| TR-U74 | Female | H13c2 | - |
| TR-U75 | Female | H14b | - |
| TR-U76 | Female | H1ag | - |
| TR-U77 | Female | H1bs | - |
| TR-U78 | Female | H2a1i | - |
| TR-U79 | Female | H5 | - |
| TR-U80 | Female | H55 | - |
| TR-U81 | Female | H92 | - |
| TR-U82 | Female | HV12 | - |
| TR-U83 | Female | HV1b | - |
| TR-U84 | Female | HV2a1 | - |
| TR-U85 | Female | HV4a2b | - |
| TR-U86 | Female | I1 | - |
| TR-U87 | Female | I1a | - |
| TR-U88 | Female | J1b1a2a | - |
| TR-U89 | Female | J1b1b1 | - |
| TR-U90 | Female | J1b3b | - |
| TR-U91 | Female | J1c+16261 | - |
| TR-U92 | Female | J1c1b | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-U93 | Female | J2a1a1 | - |
| TR-U94 | Female | K1a | - |
| TR-U95 | Female | K1a19 | - |
| TR-U96 | Female | K1a19 | - |
| TR-U97 | Female | K1a4 | - |
| TR-U98 | Female | K1a7 | - |
| TR-U99 | Female | K2a9 | - |
| TR-U100 | Female | R0a | - |
| TR-U101 | Female | R0a1a | - |
| TR-U102 | Female | R1a | - |
| TR-U103 | Female | R1a1a1 | - |
| TR-U104 | Female | R2 | - |
| TR-U105 | Female | T1 | - |
| TR-U106 | Female | T1 | - |
| TR-U107 | Female | T1b | - |
| TR-U108 | Female | T2 | - |
| TR-U109 | Female | T2b | - |
| TR-U110 | Female | T2c1a | - |
| TR-U111 | Female | T2c1c2 | - |
| TR-U112 | Female | T2d2 | - |
| TR-U113 | Female | T2h | - |
| TR-U114 | Female | U1a1 | - |
| TR-U115 | Female | U3a3 | - |
| TR-U116 | Female | U3b | - |
| TR-U117 | Female | U3b2a1 | - |
| TR-U118 | Female | U4b1a1a | - |
| TR-U119 | Female | U5a2b | - |
| TR-U120 | Female | U5b2a5 | - |
| TR-U121 | Female | U7 | - |
| TR-U122 | Female | V | - |
| TR-U123 | Female | W1 | - |
| TR-U124 | Female | W1c | - |
| TR-U125 | Female | X2e | - |
| TR-U126 | Female | X2m'n | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-W53 | Female | D4 | - |
| TR-W54 | Female | D4j7 | - |
| TR-W55 | Female | D4m2a | - |
| TR-W56 | Female | G2a2a | - |
| TR-W57 | Female | H11a1 | - |
| TR-W58 | Female | H1b1a | - |
| TR-W59 | Female | H1b1b | - |
| TR-W60 | Female | H1h1 | - |
| TR-W61 | Female | H20a | - |
| TR-W62 | Female | H46 | - |
| TR-W63 | Female | H47 | - |
| TR-W64 | Female | H4a1a1a | - |
| TR-W65 | Female | H5a1q | - |
| TR-W66 | Female | H7b1 | - |
| TR-W67 | Female | H7g | - |
| TR-W68 | Female | H92 | - |
| TR-W69 | Female | HV15 | - |
| TR-W70 | Female | J1b1a | - |
| TR-W71 | Female | J1b1b1 | - |
| TR-W72 | Female | J1b9 | - |
| TR-W73 | Female | J1c2 | - |
| TR-W74 | Female | J1c2f | - |
| TR-W75 | Female | J1c2f | - |
| TR-W76 | Female | J1d1b1 | - |
| TR-W77 | Female | K1a23 | - |
| TR-W78 | Female | K1a2b | - |
| TR-W79 | Female | M5a1b | - |
| TR-W80 | Female | M5a1b | - |
| TR-W81 | Female | N1b1a | - |
| TR-W82 | Female | N1b1a+16129 | - |
| TR-W83 | Female | R1a1 | - |
| TR-W84 | Female | T2 | - |
| TR-W85 | Female | T2a1b1a | - |
| TR-W86 | Female | T2b | - |

*(continued on next page)*

*Table A.2 continued*

| | | | |
|---|---|---|---|
| TR-W87 | Female | T2b1 | - |
| TR-W88 | Female | T2c1a | - |
| TR-W89 | Female | U3b1 | - |
| TR-W90 | Female | U3b3 | - |
| TR-W91 | Female | U4a1 | - |
| TR-W92 | Female | U4b1a4 | - |
| TR-W93 | Female | U4d2 | - |
| TR-W94 | Female | V | - |
| TR-W95 | Female | W1 | - |
| TR-W96 | Female | W3a1 | - |
| TR-W97 | Female | W3a1d | - |

**Table A.3** List of rare homozygous HC-pLoFs (Separate file)

**Table A.4** List of rare homozygous HC-pLoFs (Separate file)

**Table A.5** TR variants that are listed as DM in HGMD (Separate file)

**Table A.6** TR variants that are listed as Pathogenic or Pathogenic/Likely pathogenic in ClinVar (Separate file)

**Table A.7** RP and PP variants in the study (Separate file)

**Table A.8** Number of variants per individual in each disease group (Separate file)

**Table A.9** Cumulative number of variants, CF and prevalence for each gene (Separate file)

**Table A.10** Cumulative number of variants, CF and prevalence for NBS genes (Separate file)

**Table A.11** Cumulative number of variants, CF and prevalence for ACMG recommended actionable genes (Separate file)

# Appendix B

# Copyright permissions

1. **Figures 3.1-3.3, 3.6, 3.8-3.13, 3.16-3.21-3.26, 3.28-3.43, Tables 2.1-2.4, 3.1, 3.3, 3.17, A.1-A.6 and Publications:** M. E. Kars, A. N. Basak, O. E. Onat, K. Bilguvar, J. Choi, Y. Itan *et al.*, "The genetic structure of the Turkish population reveals high levels of variation and admixture," *Proc Natl Acad Sci U S A*, vol. 118, no. 36, p. e2026076118, 2021.

2. **Figures 4.1, 4.2, and Publications:** O. E. Onat, M. E. Kars, S. Gul, K. Bilguvar, Y. Wu, A. Ozhan *et al.*, "Human CRY1 variants associate with attention deficit/hyperactivity disorder," *J Clin Invest*, vol. 130, no. 7, pp. 3885-3900, 2020.

3. **Figure 1.3:** H. Hamamy, S. E. Antonarakis, L. L. Cavalli-Sforza, S. Temtamy, G. Romeo, L. P. Kate *et al.*, "Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report," *Genet Med*, vol. 13, no. 9, pp. 841-847, 2011.

4. **Figure 1.4:** F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, "Runs of homozygosity: Windows into population history and trait architecture," *Nat Rev Genet*, vol. 19, no. 4, pp. 220-234, 2018.

5. **Figure 1.6**: E. M. Scott, A. Halees, Y. Itan, E. G. Spencer, Y. He, M. A. Azab *et al.*, "Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery," *Nat Genet*, vol. 48, no. 9, pp. 107-106, 2016.

6. **Figure 1.7:** S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333-51, 2016.

7. **Figure 1.8 and Table 1.1:** S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster *et al.*, "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genet Med*, vol. 17, no. 5, pp. 405-424, 2015.

8. **Figure 1.10:** J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nat Rev Genet*, vol. 11, no. 7, pp. 499-511, 2010.

CCC Marketplace™



| | | | |
|---|---|---|---|
| Order Date | 12-Oct-2021 | Type of Use | Republish in a thesis/dissertation |
| Order License ID | 1153567-2 | | |
| ISSN | 0021-9738 | Publisher | AMERICAN SOCIETY FOR CLINICAL INVESTIGATION |
| | | Portion | Chapter/article |

## LICENSED CONTENT

| | | | |
|---|---|---|---|
| Publication Title | The journal of clinical investigation | Rightsholder | American Society for Clinical Investigation |
| Article Title | Human CRY1 variants associate with attention deficit/hyperactivity disorder. | Publication Type | Journal |
| | | Start Page | 3885 |
| | | End Page | 3900 |
| Author/Editor | AMERICAN SOCIETY FOR CLINICAL INVESTIGATION., Robinson, George Canby | Issue | 7 |
| | | Volume | 130 |
| Date | 01/01/1924 | | |
| Language | English | | |
| Country | United States of America | | |

## REQUEST DETAILS

| | | | |
|---|---|---|---|
| Portion Type | Chapter/article | Rights Requested | Main product |
| Page range(s) | 3885-3900 | Distribution | Worldwide |
| Total number of pages | 16 | Translation | Original language of publication |
| Format (select all that apply) | Print, Electronic | Copies for the disabled? | No |
| Who will republish the content? | Author of requested content | Minor editing privileges? | No |
| | | Incidental promotional use? | No |
| Duration of Use | Life of current edition | | |
| Lifetime Unit Quantity | Up to 499 | Currency | USD |

## NEW WORK DETAILS

| | | | |
|---|---|---|---|
| Title | Characterization of the fine-scale genetic structure of the Turkish population | Institution name | Ihsan Dogramaci Bilkent University |
| | | Expected presentation date | 2021-12-31 |
| Instructor name | Tayfun Özçelik | | |

## ADDITIONAL DETAILS

| | | | |
|---|---|---|---|
| Order reference number | N/A | The requesting person / organization to appear on the license | Meltem Ece Kars |

## REUSE CONTENT DETAILS

| | | | |
|---|---|---|---|
| Title, description or numeric reference of the portion(s) | Full article | Title of the article/chapter the portion is from | Human CRY1 variants associate with attention deficit/hyperactivity disorder. |
| Editor of portion(s) | Onat, O. Emre; Kars, M. Ece; Gül, Şeref; Bilguvar, Kaya; Wu, Yiming et. al. | Author of portion(s) | Onat, O. Emre; Kars, M. Ece; Gül, Şeref; Bilguvar, Kaya; Wu, Yiming et. al. |
| Volume of serial or monograph | 130 | Issue, if republishing an article from a serial | 7 |
| Page or page range of portion | 3885-3900 | Publication date of portion | 2020-07-01 |

## CCC Terms and Conditions

1. Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the "Work(s)"). Copyright Clearance Center, Inc. ("CCC") grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the "Rightsholder"). "Republication", as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the Order Confirmation. "User", as used herein, means the person or entity making such republication.

2. The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a "freelancer" or other third party independent of User and CCC, such party shall be deemed jointly a "User" for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.

3. Scope of License; Limitations and Obligations.

   3.1. All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.

   3.2. General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2% per month or, if less, the maximum rate allowed by applicable law. Unless otherwise specifically set forth in the Order Confirmation or in a separate written agreement signed by CCC, invoices are due and payable on "net 30" terms. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void, as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company.

   3.3. Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is "one-time" (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period).

   3.4. In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such

material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of such third party materials; without a separate license, such third party materials may not be used.

3.5. Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: "Republished with permission of [Rightsholder's name], from [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc. " Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

3.6. User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

4. Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

5. Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

6. Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7. Effect of Breach. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8. Miscellaneous.

8.1. User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these

terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2. Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here:https://marketplace.copyright.com/rs-ui-web/mp/privacy-policy

8.3. The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed under this Service.

8.4. No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5. The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court.If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.

v 1.1

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 12, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5166470827859 |
| License date | Oct 12, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Genetics in Medicine |
| Licensed Content Title | Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report |
| Licensed Content Author | Hanan Hamamy et al |
| Licensed Content Date | May 6, 2011 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |

| | |
|---|---|
| Will you be translating? | no |
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Figure 1 |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |

Terms and Conditions

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of

another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where '**reuse in a dissertation/thesis**' has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 12, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5166470400042 |
| License date | Oct 12, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Reviews Genetics |
| Licensed Content Title | Runs of homozygosity: windows into population history and trait architecture |
| Licensed Content Author | Francisco C. Ceballos et al |
| Licensed Content Date | Jan 15, 2018 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |
| Will you be translating? | no |

| | |
|---|---|
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Figure 1 |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |

Terms and Conditions

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also

seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that

affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 12, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5166470150265 |
| License date | Oct 12, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Genetics |
| Licensed Content Title | Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery |
| Licensed Content Author | Eric M Scott Sohair Abdel Rahim et al |
| Licensed Content Date | Jul 18, 2016 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |

| | |
|---|---|
| Will you be translating? | no |
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Supplementary Figure 1 |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |

Terms and Conditions

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of

another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where '**reuse in a dissertation/thesis**' has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 14, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5167551486748 |
| License date | Oct 14, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Reviews Genetics |
| Licensed Content Title | Coming of age: ten years of next-generation sequencing technologies |
| Licensed Content Author | Sara Goodwin et al |
| Licensed Content Date | May 17, 2016 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 2 |
| High-res required | no |
| Will you be translating? | no |

| | |
|---|---|
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Figure 1b and Figure 3a |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |

Terms and Conditions

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also

seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that

affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 20, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5172921287060 |
| License date | Oct 20, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Genetics in Medicine |
| Licensed Content Title | Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology |
| Licensed Content Author | Sue Richards PhD et al |
| Licensed Content Date | Mar 5, 2015 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 2 |

| | |
|---|---|
| High-res required | no |
| Will you be translating? | no |
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Figure 1 and Table 5 |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |
| Terms and Conditions | |

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 22, 2021

---

This Agreement between Dr. Meltem Kars ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5174131158851 |
| License date | Oct 22, 2021 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Reviews Genetics |
| Licensed Content Title | Genotype imputation for genome-wide association studies |
| Licensed Content Author | Jonathan Marchini et al |
| Licensed Content Date | Jun 2, 2010 |
| Type of Use | Thesis/Dissertation |
| Requestor type | non-commercial (non-profit) |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |
| Will you be translating? | no |

| | |
|---|---|
| Circulation/distribution | 100 - 199 |
| Author of this Springer Nature content | no |
| Title | Characterization of the fine-scale genetic structure of the Turkish population |
| Institution name | Bilkent University |
| Expected presentation date | Dec 2021 |
| Portions | Figure in Box 1 |
| Requestor Location | Dr. Meltem Kars<br>Bilkent University<br><br>Ankara, other 06800<br>Turkey<br>Attn: Dr. Meltem Kars |
| Total | 0.00 USD |

Terms and Conditions

**Springer Nature Customer Service Centre GmbH**
**Terms and Conditions**

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

**1. Grant of License**

**1. 1.** The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

**1. 2.** The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

**1. 3.** If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also

seek permission from that source to reuse the material.

**2. Scope of Licence**

**2. 1.** You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

**2. 2.** A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

**2. 3.** Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

**2. 4.** Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

**2. 5.** An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

**3. Duration of Licence**

**3. 1.** A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

| Scope of Licence | Duration of Licence |
|---|---|
| Post on a website | 12 months |
| Presentations | 12 months |
| Books and journals | Lifetime of the edition in the language purchased |

**4. Acknowledgement**

**4. 1.** The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

**5. Restrictions on use**

**5. 1.** Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that

affect the meaning, intention or moral rights of the author are strictly prohibited.

**5. 2.** You must not use any Licensed Material as part of any design or trademark.

**5. 3.** Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

**6. Ownership of Rights**

**6. 1.** Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

**7. Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

**8. Limitations**

**8. 1.** ***BOOKS ONLY:*** Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

**8. 2.** For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

**9. Termination and Cancellation**

**9. 1.** Licences will expire after the period shown in Clause 3 (above).

**9. 2.** Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

**Appendix 1 — Acknowledgements:**

**For Journal Content:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

**For Adaptations/Translations:**
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

**Note: For any republication from the British Journal of Cancer, the following credit line style applies:**

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

**For Book content:**
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

**Other Conditions**:

Version 1.3

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**

# Appendix C

# Publications

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

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

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

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

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

Turkish Variome | admixture | sequencing | population genetics | variation

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

## Significance

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

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

The authors declare no competing interest.

This article is a PNAS Direct Submission.



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

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

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

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

Published August 23, 2021.

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

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

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

## Results

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

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

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

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

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

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

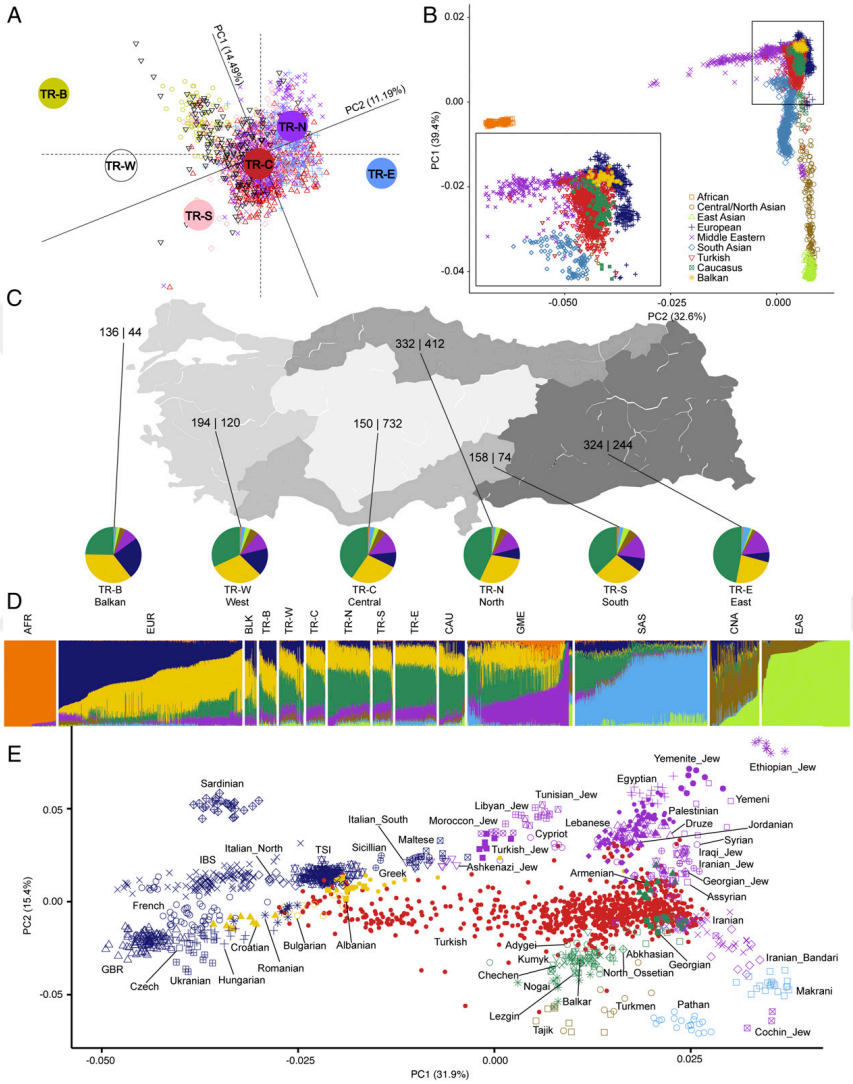



**Fig. 1.** TR as a hub of extensive human admixture. (*A*) Procrustes analysis based on unprojected coordinates of geographical locations and PC1 and PC2 coordinates of 1,460 TR individuals with known origin. (*B*) PCA for individuals from the TR WGS ($n = 773$), the populations from Lazaridis et al.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

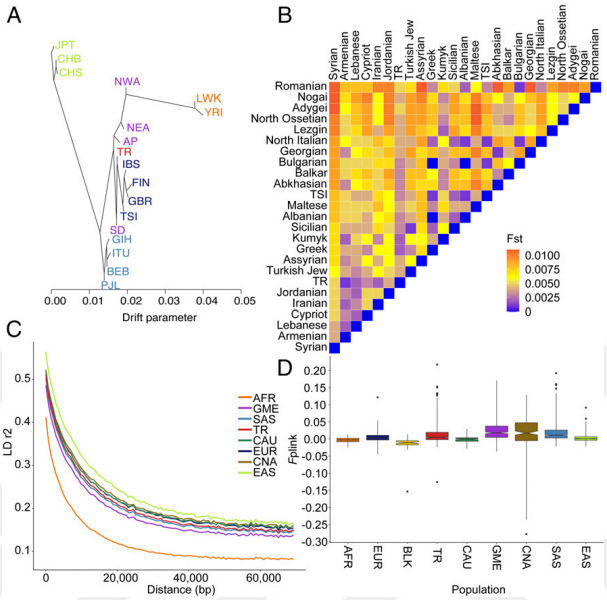



**Fig. 2.** The Turkish Peninsula as a bridge in the migration trajectories and high inbreeding levels in the TR population. (*A*) TreeMix phylogeny of the TR population ($n$ = 3,362) along with the 1000GP controls ($n$ = 1,299) and the GME populations ($n$ = 696) representing divergence patterns. The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

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

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

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

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

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

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

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

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

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

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

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

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

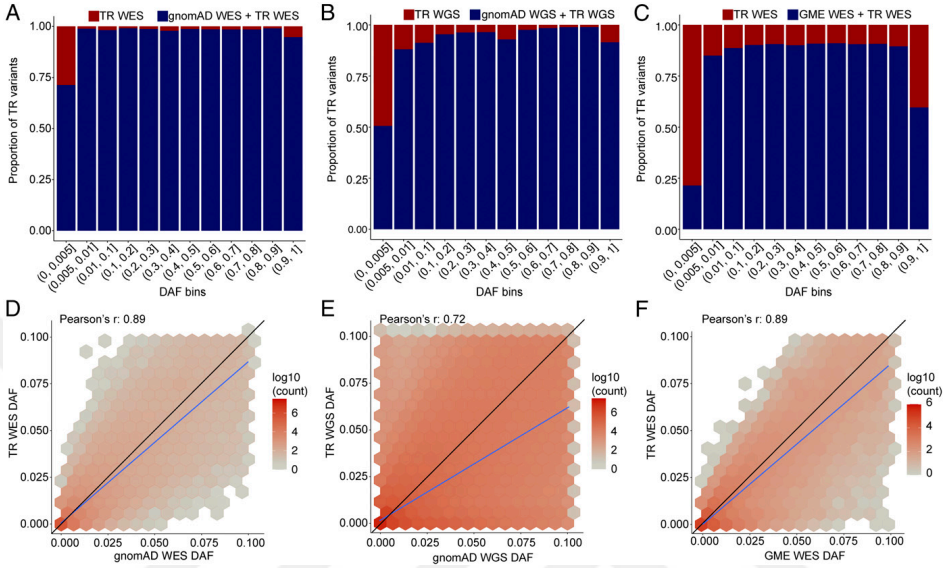



**Fig. 3.** The TR Variome possesses a significant number of very rare and unique variants that are poorly represented in gnomAD and GME. The proportion of TR variants represented in the TR Variome and other databases. (*A*) TR WES versus gnomAD WES, (*B*) TR WGS versus gnomAD WGS, (*C*) TR WES versus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

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

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

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

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

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

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

## Discussion

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



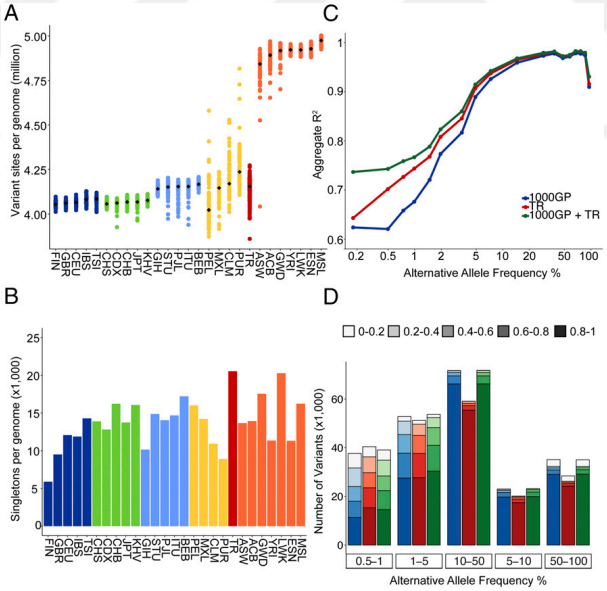



**Fig. 4.** Per-genome variant summary and imputation. (*A*) The number of variant sites per genome for autosomes. (*B*) The average number of singletons per genome for autosomes. (*C*) Evaluation of imputation performance on chromosome 20. The aggregate squared Pearson correlation coefficient ($R^2$)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

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

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

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

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

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

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

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

## Materials and Methods

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

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

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

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

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

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

JCI The Journal of Clinical Investigation

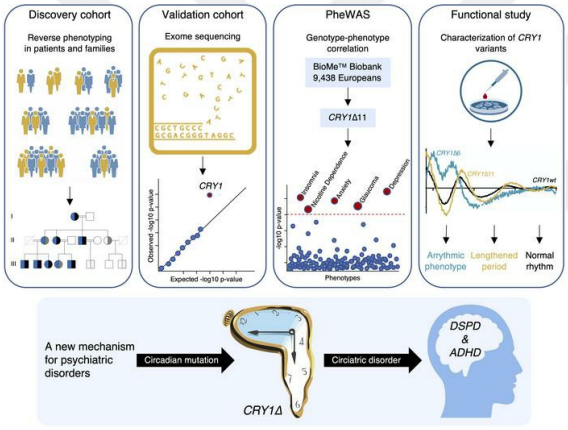
# Human *CRY1* variants associate with attention deficit/hyperactivity disorder

O. Emre Onat, … , İ. Halil Kavaklı, Tayfun Özçelik

*J Clin Invest.* 2020;130(7):3885-3900. https://doi.org/10.1172/JCI135500.

Research Article Genetics

## Graphical abstract



# Human *CRY1* variants associate with attention deficit/hyperactivity disorder

O. Emre Onat,[1] M. Ece Kars,[1] Şeref Gül,[2,3] Kaya Bilguvar,[4] Yiming Wu,[5] Ayşe Özhan,[1] Cihan Aydın,[2,3] A. Nazlı Başak,[6] M. Allegra Trusso,[7] Arianna Goracci,[7] Chiara Fallerini,[8] Alessandra Renieri,[8,9] Jean-Laurent Casanova,[10,11,12,13,14] Yuval Itan,[5,15] Cem E. Atbaşoğlu,[16] Meram C. Saka,[16] İ. Halil Kavaklı,[2] and Tayfun Özçelik[1,17,18]

[1]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey. [2]Department of Chemical and Biological Engineering and [3]Department of Molecular Biology and Genetics, Koç University, Istanbul, Turkey. [4]Department of Genetics, Yale Center for Genome Analysis, Yale University School of Medicine, New Haven, Connecticut, USA. [5]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [6]Neurodegeneration Research Laboratory, Suna and Inan Kıraç Foundation, KUTTAM, Koç University, Istanbul, Turkey. [7]Division of Psychiatry, Department of Molecular Medicine and Development, Azienda Ospedaliera Universitaria Senese, Siena, Italy. [8]Medical Genetics, University of Siena, Siena, Italy. [9]Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy. [10]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, New York, USA. [11]Laboratory of Human Genetics of Infectious Diseases, Necker Branch INSERM U1163, Necker Hospital for Sick Children, Paris, France. [12]Imagine Institute, University of Paris, Paris, France. [13]Pediatric Immunology-Hematology Unit, Necker Hospital for Sick Children, Paris, France. [14]Howard Hughes Medical Institute (HHMI), Rockefeller University, New York, New York, USA. [15]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [16]Department of Psychiatry, Ankara University Medical School, Ankara, Turkey. [17]Neuroscience Program, Graduate School of Engineering and Science, and [18]Institute of Materials Science and Nanotechnology, National Nanotechnology Research Center (UNAM), Bilkent University, Ankara, Turkey.

**Attention deficit/hyperactivity disorder (ADHD) is a common and heritable phenotype frequently accompanied by insomnia, anxiety, and depression. Here, using a reverse phenotyping approach, we report heterozygous coding variations in the core circadian clock gene cryptochrome 1 in 15 unrelated multigenerational families with combined ADHD and insomnia. The variants led to functional alterations in the circadian molecular rhythms, providing a mechanistic link to the behavioral symptoms. One variant, *CRY1*Δ11 c.1657+3A>C, is present in approximately 1% of Europeans, therefore standing out as a diagnostic and therapeutic marker. We showed by exome sequencing in an independent cohort of patients with combined ADHD and insomnia that 8 of 62 patients and 0 of 369 controls carried *CRY1*Δ11. Also, we identified a variant, *CRY1*Δ6 c.825+1G>A, that shows reduced affinity for BMAL1/CLOCK and causes an arrhythmic phenotype. Genotype-phenotype correlation analysis revealed that this variant segregated with ADHD and delayed sleep phase disorder (DSPD) in the affected family. Finally, we found in a phenome-wide association study involving 9438 unrelated adult Europeans that *CRY1*Δ11 was associated with major depressive disorder, insomnia, and anxiety. These results defined a distinctive group of circadian psychiatric phenotypes that we propose to designate as "circiatric" disorders.**

## Introduction

Sleep is genetically regulated by the circadian rhythm. Disruption of this rhythm leads to aberrant sleep patterns (1–4). Sleep is also frequently disturbed in individuals with psychiatric disorders. For example, dim light melatonin onset, a reliable marker of circadian function, is delayed in children with attention deficit/hyperactivity disorder (ADHD) (5, 6). However, a causal relationship between inherited circadian mutations and psychiatric traits has not been established (7–12). More generally, psychiatric and sleep disorders proved to be extremely difficult to solve by classic phenotype-first studies because of complex factors that are at both the phenotype and genotype definition and characterization levels. For example, phenotype misclassifications arise from problems of distinguishing health from disease, the episodic nature of symptoms, and establishing accurate diagnostic criteria. Likewise, interpretation of genotype data becomes a challenge because of early postzygotic mutations, incomplete penetrance, variable expressivity, or high levels of genetic heterogeneity (11, 13). Several levels of causation have been implicated in the emergence of heterogeneity: (a) the existence of many different rare and severe mutations of the same gene in unrelated individuals, (b) the same mutation leading to different phenotypic outcomes in different individuals, (c) mutations in different genes leading to the same disorder, and (d) a collective effect of many individual gene events (14). As a result, the genomic landscape of ADHD and more generally of psychiatric disorders remains largely unknown, and much of the genetic risk is unexplained.

Reverse phenotyping is an alternative approach to overcome the uncertainties inherent to clinical diagnoses in the patient care setting and promises to achieve accurate phenotype assignments in the research setting (15–18). Also termed the genotype-first approach, reverse phenotyping consists of 3 consecutive steps: (a) collection of genomic data and candidate discovery, (b) determination of causality by phenotype and segregation analyses in

**Conflict of interest:** The authors have declared that no conflict of interest exists.

**Submitted:** December 5, 2019; **Accepted:** April 16, 2020; **Published:** June 15, 2020.
**Reference information:** *J Clin Invest*. 2020;130(7):3885–3900.
https://doi.org/10.1172/JCI135500.

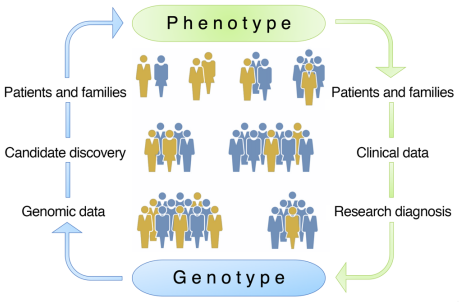



**Figure 1. Reverse phenotyping.** Schematic of the phenotype-first (green) versus the genotype-first (also referred to as reverse phenotyping, blue) approaches for identification of causal gene mutations. The phenotype-first approach relies on identification of patients and families, collection of clinical data, accurate research diagnosis, and, finally, collection of genotype data steps. In the genotype-first approach, the process is reversed and starts with the analysis of genomic data and selection of candidate variants followed by comprehensive clinical phenotyping of patients and families to make accurate genotype-phenotype correlations.

families, and (c) population screening for the mutant locus in phenotypically well-characterized cohorts (Figure 1).

In this context, using reverse phenotyping, we recently identified a gain-of-function *CRY1* variant (*CRY1*Δ11, *CRY1* c.1657+3A>C, rs184039278) that provides a mechanistic link to delayed sleep phase disorder (DSPD), a common form of insomnia, in 6 large multigenerational Turkish families (1). CRY1 is an essential component of the core molecular clock and represses the activity of the transcription factors CLOCK and BMAL1 transactivation (19). As we observed a high incidence of behavioral endophenotypes including a history of depression mainly in mutation carriers, we decided to further characterize the clinical features of individuals from these 6 families, as well as individuals from 6 additional Turkish families with DSPD.

## Results

*Reverse phenotyping in the discovery cohort*. We first performed clinical evaluations of a cohort of 96 individuals from 12 families from Turkey. A systematic psychiatric assessment was conducted using the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5, 5th edition) (20) and questionnaires (21, 22). For this evaluation, 2 board-certified psychiatrists with extensive experience in psychiatric analysis, interviewed the study participants without knowing their genotype status. The interviews with each study participant were conducted over a period of 60–90 minutes and were designed to ascertain the psychosocial, functional, and mental status of each of the subjects. The ADHD Child Evaluation (ACE) interview questionnaires "ACE – A diagnostic interview of ADHD in children" and "ACE+ – A diagnostic interview of ADHD in adults" were used to support the interviews (http://www.psychology-services.uk.com). A polysomnographic sleep recording of a *CRY1*Δ11 carrier using continuous electroencephalography, electromyography, and electrooculography was also done and is reported elsewhere (1).

This comprehensive phenotyping revealed that the symptoms and signs that define ADHD were present in addition to DSPD in 46 of 48 mutation-positive individuals and absent in 44 of 48 mutation-negative relatives or spouses. Two carriers and 3 WT individuals were classified as ADHD spectrum and 1 WT individual as affected. This corresponded to a total of 47 affected individuals, 37 of whom displayed patterns of behavior consistent with a combined presentation, 2 who were found to be predominantly hyperactive and impulsive, and 8 who were found to be predominantly inattentive. We independently confirmed our observations by characterizing 5 mutation carriers from 2 Italian families using the criteria described above for the Turkish families (Figure 2). This evaluation was carried out by 2 board-certified psychiatrists. All mutation carriers were diagnosed with ADHD, and 3 carriers also experienced sleep disturbances (Figure 3, A–N, and Supplemental Tables 1–3; supplemental material available online with this article; https://doi.org/10.1172/JCI135500DS1). Visual inspection of segregation patterns in the pedigrees suggested an autosomal dominant inheritance of the phenotype, and no significant difference in ADHD and DSPD symptoms was observed between homozygous and heterozygous carriers. These results strongly suggested that circadian dysfunction, as exemplified by the *CRY1*Δ11 mutation, has a very strong association with combined ADHD and DSPD (OR 281, $P = 1.99 \times 10^{-21}$, Fisher's exact test).

*ADHD comorbidities*. ADHD is possibly an extreme expression of continuous heritable traits significantly correlated with educational outcomes, psychiatric or personality disorders, obesity-related phenotypes, smoking or smoking-related cancer, reproductive success, longevity, and insomnia (12). Therefore, we searched for ADHD comorbidities in the families using the phenotype information documented during the examination, which included demographics, history of depression, and smoking status. We documented an overrepresentation of a recurrent history of depression in *CRY1*Δ11$^+$ adults (34 of 53, 64.2%; $n = 4$ homozygous, $n = 30$ heterozygous) compared with *CRY1*Δ11$^-$ adults (5 of 48, 10.4%; OR 15.4, $P = 1.65 \times 10^{-8}$) (Supplemental Table 4). These findings are consistent with epidemiological studies, which report that ADHD may sometimes be the underlying cause for features of clinical depression, especially in adults (23). Also, we documented an increase in smoking (ever vs. never) in *CRY1*Δ11 carriers compared with their WT family members (83% vs. 43% in males and 46% vs. 33% in females). According to the 2016 Global Adult Tobacco Survey in Turkey (24), this increase holds when compared with the general Turkish population (44% in males and 19% in females), and warrants further investigation in a larger cohort (Supplemental Table 4) and warrants further investigation in a larger cohort (Supplemental Table 4).

*Sunlight exposure*. Epidemiological studies suggest a link between sleep, mood, and sunlight. Clinicians used bright light

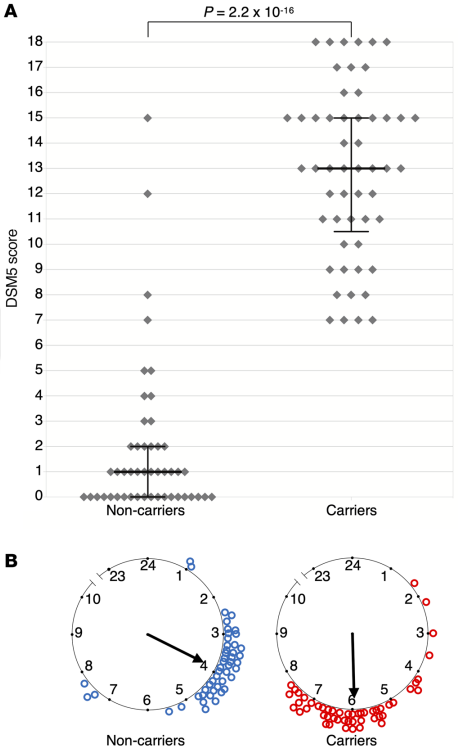



**Figure 2. DSM-5 ADHD scores, mid-sleep point on free days, and *CRY1*Δ11 mutation status of the 14-family discovery cohort.** (**A**) Each dot represents the DSM-5 score of 101 individuals. Medians and interquartile ranges are marked for 53 *CRY1*Δ11 carriers and 48 intrafamilial noncarriers. A Mann-Whitney *U* test indicated that the DSM-5 score for the carriers (median = 13) was greater than that for the noncarriers (median = 1). $U = 64.5$, $P = 2.2 \times 10^{-16}$. (**B**) The mid-sleep point on free days (MSF) for the same subjects are plotted on a discontinuous clock face from 2300 to 1000 hours for noncarriers (left, blue) and carriers (right, red). No subject data fell within the gap time (1000 to 2300 hours) not represented in the plot.

(25, 26) or high solar intensity (27) as a controlled intervention for circadian rhythm sleep problems as well as ADHD. We performed a systematic analysis of direct sunlight exposure durations in the 52 *CRY1*Δ11 carriers of these 14 families, and investigated the correlation with mid-sleep points and ADHD severity as judged by clinical observations and questionnaire scores (Supplemental Table 5). Although we did not observe a correlation for mid-sleep points, we noted a milder presentation of ADHD in 33 individuals, 19 of whom had longer periods of daily sunlight exposure compared with the 19 individuals with severe ADHD. A scatter plot of Adult ADHD Self-Report Scale (ASRS) scores and the mean duration of sun exposure revealed a moderate negative correlation (Figure 4, Spearman's rho = –0.46, $P = 0.0005$). This observation highlights the potential value of light in addition to medications and talk therapies in the management of ADHD (25–27). Accordingly, we propose including the Munich ChronoType Questionnaire (MCTQ) (21) as part of any ADHD diagnostic evaluation to document sunlight exposure durations and sleep.

*CRY1*Δ11 *penetrance and ADHD phenotype heterogeneity*. High degrees of allelic or locus heterogeneity, a presence of phenocopies, and, most important, difficulties inherent to psychiatric nosological definitions — especially due to the fluctuating nature of particular manifestations in an individual evaluated at specific time points — pose major challenges in identifying the genomic determinants of complex disorders (16–18, 28, 29), and ADHD is no exception. Two individuals from the Turkish families, who were heterozygous for *CRY1*Δ11, displayed an ADHD spectrum of symptoms but failed to fulfill the DSM-5 criteria. We also noted that 3 Turkish individuals from the families had signs and symptoms of ADHD, yet were WT for *CRY1*Δ11 (Figures 2 and 3, and Supplemental Table 1). For example, in the Turkish cohort, patient 17-010, who is a highly successful student preparing for medical school exams, had a relatively high score on the ASRS self-reporting questionnaire, but was unlikely to have ADHD based on her clinical evaluation. Her condition was more consistent with an anxiety disorder or performance anxiety. We made a similar observation for subject 17-091, who had a borderline DSM-5 score but, based on the clinical evaluation, was unlikely to have ADHD. This individual's symptoms were secondary to a recent coronary artery bypass surgery. Both of these individuals had normal sleep patterns. Therefore, these 2 subjects and an additional subject, 16-082, were classified as probably not affected. However, 17-368 and 16-027, who are carriers, were classified as probably affected on the basis of clinical observations indicating that they could be in partial remission.

*Phenotype-first approach*. To complement the reverse phenotyping, we selected 447 unrelated adults from our in-house database of Turkish families with obesity/metabolic phenotypes and designated them as the validation cohort (Supplemental Figure 1, A–C). We next reviewed their medical charts and contacted all of them for MCTQ and ASRS questionnaire evaluations (21, 22). Supplemental Table 6 presents the demographic data for this cohort (i.e., age, sex, education, occupation, marital status, and number of children). This database is not publicly accessible, but ethics and consent procedures for the subjects allowed for recontact. During this first step, 108 of 447 (24.2%) individuals self-identified as having excessive inattention and/or hyperactivity and impulsivity, and 185 of 447 (41.3%) as having delayed sleep patterns on free days. These high numbers were expected, since a majority of the families were recruited to study the genetic basis for obesity. BMI data revealed a marked positive correlation with ADHD, and obesity is associated with sleep problems (12). The next step of a physician-led interview using the DSM-5 criteria with the 108 ASRS-positive individuals confirmed that 78 of them met the ADHD diagnostic criteria. Exclusion criteria were autism, epilepsy, intellectual disability, psychosis/schizophrenia, ADHD symptoms due to personality disorders, adoption, sexual or phys-

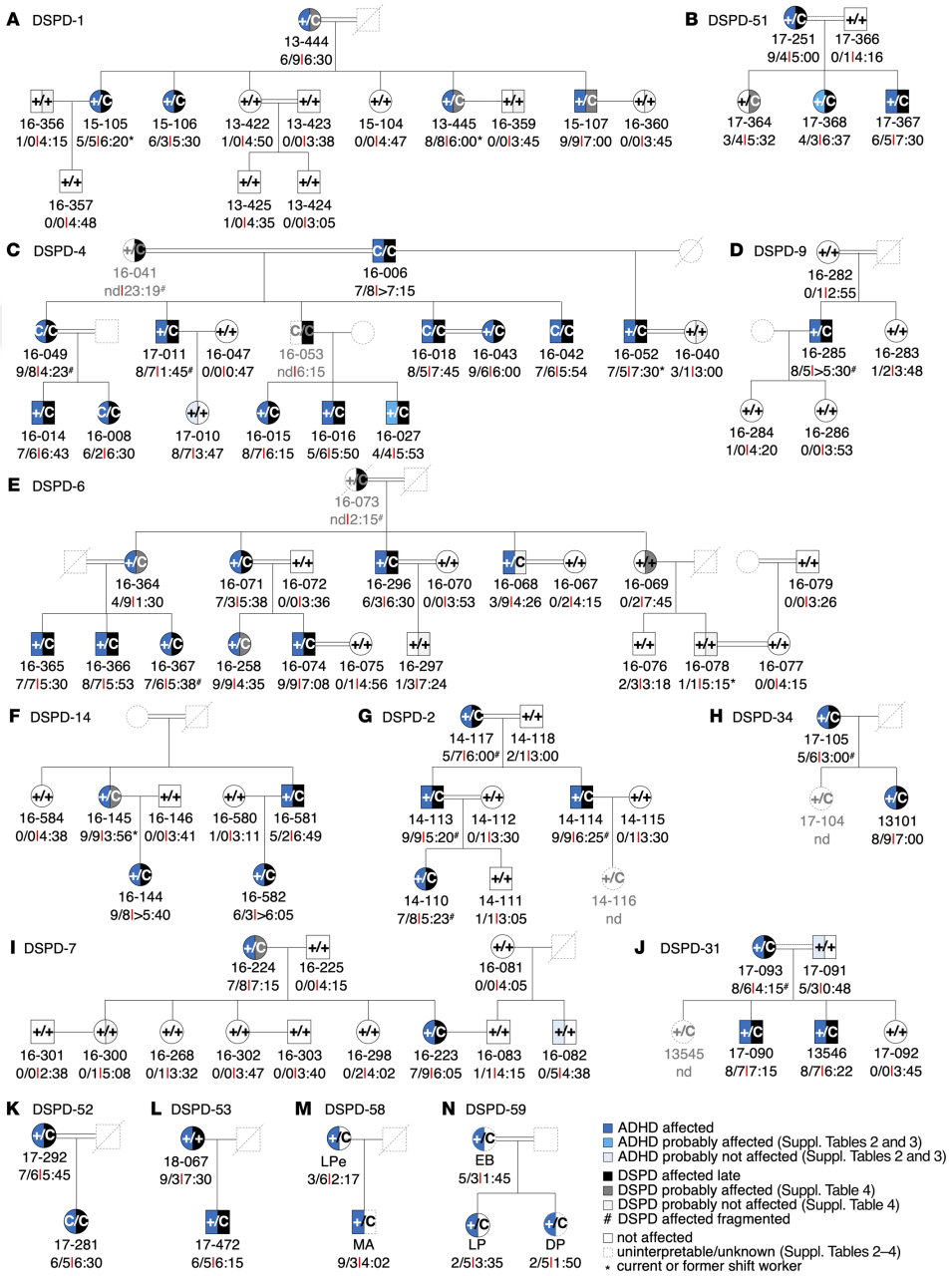

A DSPD-1
13-444 6/9|6:30
16-356 1/0|4:15
15-105 5/5|6:20*
15-106 6/3|5:30
13-422 1/0|4:50
13-423 0/0|3:38
15-104 0/0|4:47
13-445 8/8|6:00*
16-359 0/0|3:45
15-107 9/9|7:00
16-360 0/0|3:45
16-357 0/0|4:48
13-425 1/0|4:35
13-424 0/0|3:05
B DSPD-51
17-251 9/4|5:00
17-366 0/1|4:16
17-364 3/4|5:32
17-368 4/3|6:37
17-367 6/5|7:30
C DSPD-4
16-041 nd|23:19#
16-006 7/8|>7:15
16-049 9/8|4:23#
17-011 8/7|1:45#
16-047 0/0|0:47
16-053 nd|6:15
16-018 8/5|7:45
16-043 9/6|6:00
16-042 7/6|5:54
16-052 7/5|7:30*
16-040 3/1|3:00
16-014 7/6|6:43
16-008 6/2|6:30
17-010 8/7|3:47
16-015 8/7|6:15
16-016 5/6|5:50
16-027 4/4|5:53
D DSPD-9
16-282 0/1|2:55
16-285 8/5|>5:30#
16-283 1/2|3:48
16-284 1/0|4:20
16-286 0/0|3:53
E DSPD-6
16-073 nd|2:15#
16-364 4/9|1:30
16-071 7/3|5:38
16-072 0/0|3:36
16-296 6/3|6:30
16-070 0/0|3:53
16-068 3/9|4:26
16-067 0/2|4:15
16-069 0/2|7:45
16-079 0/0|3:26
16-365 7/7|5:30
16-366 8/7|5:53
16-367 7/6|5:38#
16-258 9/9|4:35
16-074 9/9|7:08
16-075 0/1|4:56
16-297 1/3|7:24
16-076 2/3|3:18
16-078 1/1|5:15*
16-077 0/0|4:15
F DSPD-14
16-584 0/0|4:38
16-145 9/9|3:56*
16-146 0/0|3:41
16-580 1/0|3:11
16-581 5/2|6:49
16-144 9/8|>5:40
16-582 6/3|>6:05
G DSPD-2
14-117 5/7|6:00#
14-118 2/1|3:00
14-113 9/9|5:20#
14-112 0/1|3:30
14-114 9/9|6:25#
14-115 0/1|3:30
14-110 7/8|5:23#
14-111 1/1|3:05
14-116 nd
H DSPD-34
17-105 5/6|3:00#
17-104 nd
13101 8/9|7:00
I DSPD-7
16-224 7/8|7:15
16-225 0/0|4:15
16-081 0/0|4:05
16-301 0/0|2:38
16-300 0/1|5:08
16-268 0/1|3:32
16-302 0/0|3:47
16-303 0/0|3:40
16-298 0/2|4:02
16-223 7/9|6:05
16-083 1/1|4:15
16-082 0/5|4:38
J DSPD-31
17-093 8/6|4:15#
17-091 5/3|0:48
13545 nd
17-090 8/7|7:15
13546 8/7|6:22
17-092 0/0|3:45
K DSPD-52
17-292 7/6|5:45
17-281 6/5|6:30
L DSPD-53
18-067 9/3|7:30
17-472 6/5|6:15
M DSPD-58
LPe 3/6|2:17
MA 9/3|4:02
N DSPD-59
EB 5/3|1:45
LP 2/5|3:35
DP 2/5|1:50
ADHD affected
ADHD probably affected (Suppl. Tables 2 and 3)
ADHD probably not affected (Suppl. Tables 2 and 3)
DSPD affected late
DSPD probably affected (Suppl. Table 4)
DSPD probably not affected (Suppl. Table 4)
# DSPD affected fragmented
not affected
uninterpretable/unknown (Suppl. Tables 2–4)
* current or former shift worker

**Figure 3. DSPD and ADHD phenotypes in *CRY1Δ*11 carrier families of Turkish and Italian descent. (A–N)** Families DSPD-1, -2, -4, -6, -7, -9, -14, -31, -34, -51, -52, -53, -58, and -59 underwent psychiatric evaluation during personal interviews and using sleep and ChronoType questionnaires (1, 21, 22). Individual ID numbers followed by DSM-5 scores and the MSF are shown below and the genotype status (+ denotes the WT allele; C denotes the mutant allele) in the pedigree symbols. See Supplemental Tables 1–3 for further details.

ical abuse, birth weight under 1.5 kg, and other neurological or systemic disorders that might explain ADHD symptoms (11, 13). Sixty-two of 78 (79.5%) of these individuals also had DSPD (Supplemental Table 6).

*Whole-exome sequencing in the validation cohort*. Exome sequencing was performed for all of the 447 individuals and revealed that only *CRY1* reached genome-wide statistical significance after variant prioritization using a variant-based, gene-based burden analysis and an optimal unified sequence kernel association test (SKAT-O) (ref. 30, Figure 5, A–C, and Supplemental Table 7). We identified 8 *CRY1*Δ11 carriers who were classified in the combined ADHD and DSPD category and 1 carrier who had isolated DSPD. These results validate the observations in the discovery cohort, suggesting that as many as 1 in 8 (8 of 62, 13%) patients with combined ADHD and DSPD and 1 in 21 (9 of 185, 5%) with DSPD may carry *CRY1*Δ11.

*CRY1 c.1657+3A>C allele frequency in the Turkish and Italian populations*. We combined the data from 6 different databases, which corresponds to a total of 5465 individuals, and observed a Turkish minor allele frequency (MAF) of 0.0124 for *CRY1* c.1657+3A>C (Scientific and Technological Research Council of Turkey [TÜBİTAK]: $n$ = 1082; Yale Center for Genome Analysis [YCGA]: $n$ = 1193; Rockefeller University, Casanova Lab, Turkish individuals: $n$ = 253; Koç University, Başak lab, Turkish individuals: $n$ = 1191; Bilkent, Bilkent University database: $n$ = 1013; Ankara University Brain Research Center [AUBAUM]: $n$ = 733). In the Italian population, the allele frequency is 0.01678 (Network for Italian Genomes database [NIG]: $n$ = 447) (Figure 6A and Supplemental Table 8).

*Age estimation of CRY1*Δ11 *in the Turkish and European populations*. In the Genome Aggregation Database (gnomAD), the *CRY1*Δ11 allele frequency is 0.0044 (0.03%–3.3%), which is, for example, 1 in 103 individuals in the European-derived general population (31). In the Turkish population, the frequency is higher (0.0124; Figure 6A and Supplemental Table 8). To determine whether all occurrences of *CRY1*Δ11 descended from a single ancestral mutational event or arose independently, we combined the haplotypes of 297 Europeans from the 1000 Genomes Project data (32) with those of the 447-individual validation cohort. We observed a shared haplotype block of 571.6 kb, suggesting a common founder effect (Figure 7). Assuming a 25-year intergeneration interval, the age of *CRY1*Δ11 is estimated to be approximately 11,175 years (95% CI: 6550–13,700), corresponding to 447 generations for the Turkish population, and 6425 years (95% CI: 5500–9500) and 257 generations for the European population (Figure 6, B and C). Based on these data, *CRY1*Δ11 probably spread to Europe during the arrival of Neolithic Anatolian farmers approximately 8500 years ago (33).

*Phenome-wide association study of CRY1*Δ11 *in the BioMe BioBank*. To test whether distinct circadian psychiatric outcomes define *CRY1*Δ11 mutations, we consulted the BioMe BioBank of the Institute for Personalized Medicine at the Icahn School of Medicine at Mount Sinai for a phenome-wide association study (PheWAS). We investigated the electronic medical record–linked phenotypes (ICD-10-CM codes; Centers for Disease Control and Prevention's International Classification of Diseases, 10th Revision, Clinical Modification) across 9438 unrelated adult European-only samples from the BioMe BioBank and observed 324 *CRY1*Δ11 carriers and 9114 noncarriers (MAF = 0.017). The initial PheWAS did not reveal a positive association of *CRY1*Δ11 with a distinct phenotype. However, when we filtered phenotype codes that interfere with an accurate ADHD diagnosis (11, 13), 37 of 80 individuals with ADHD were excluded, and we detected a 2.2-fold increase in the OR (95% CI: 0.42–6.87) for ADHD in *CRY1*Δ11 carriers with respect to the OR for controls, though the remaining sample size ($n$ = 43) was not sufficient to reach statistical significance. Filtered ICD-10-CM codes correspond to major mental and neurological disorders, congenital malformations of the nervous system, chromosomal abnormalities, and endocrine and metabolic diseases that lead to intellectual impairment (Supplemental Table 9). We did not filter out ADHD comorbidities (20, 34). After filtering, 238 *CRY1*Δ11 carriers and 6825 noncarriers with 6820 ICD-10-CM codes remained, and a repeat PheWAS revealed the strongest associations with major depressive disorder (MDD) (single episode: OR 1.91, $P = 7.87 \times 10^{-4}$; recurrent: OR 2.55, $P = 1.28 \times 10^{-2}$); insomnia (OR 1.84, $P = 3.87 \times 10^{-3}$); anxiety (OR 1.68, $P = 4.56 \times 10^{-3}$); glaucoma (OR 3.65, $P = 7.11 \times 10^{-3}$); and nicotine dependence (OR 2.01, $P = 2.52 \times 10^{-2}$) (Table 1 and Supplemental Tables 10 and 11). Glaucoma has been reported with increased frequency in individuals with sleep problems, and an association with *CRY1*Δ11 requires further investigation (35).

*Identification of CRY1*Δ6. Exome sequencing in the validation cohort identified 1 more individual heterozygote for a rare *CRY1* variant (c.825+1G>A, rs780614131, Supplemental Table 12) in the

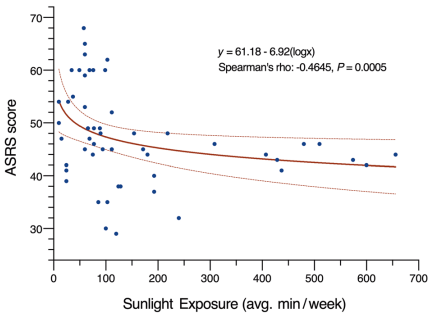



**Figure 4. Plot showing the correlation between sunlight exposure and ADHD severity.** The ADHD severity of 53 *CRY1*Δ11 carriers was assessed using ASRS scores, which are plotted against the mean duration of sunlight exposure in minutes per week ($R$ = –0.44). The line of best fit demonstrates the negative correlation, and the dashed lines represent the 95% CIs.

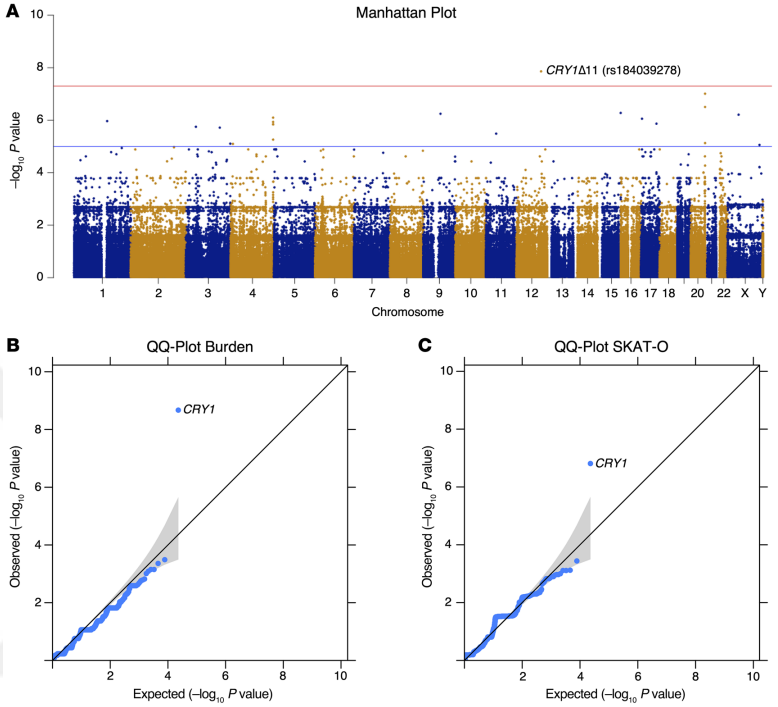



**Figure 5. GWAS analyses of ADHD$^+$ versus ADHD$^-$ groups.** ADHD$^+$, affected, $n$ = 78; ADHD$^-$, unaffected, $n$ = 369. (**A**) Manhattan plot for genome-wide association of single nucleotide variants (MAF <0.05). Plots show the $-\log_{10}$ ($P$ value) on the $y$ axis and the chromosomal position of each variant on the $x$ axis. Genes are ranked by uncorrected $P$ values. Red line shows the genome-wide significance cutoff determined by Bonferroni's correction. (**B**) Q-Q plot showing the observed and expected $P$ values for gene-based burden analysis. (**C**) Q-Q plot showing the observed and expected $P$ values for gene-based SKAT-O analysis. In the Q-Q plots, the expected null distribution (no association) is plotted along the black diagonal with the corresponding 95% CIs, and the entire distribution of the observed minimum achievable $P$ value–adjusted (MAP-adjusted) $-\log_{10}$ ($P$ value) is plotted in blue.

photolyase homology region (PHR), which leads to the skipping of exon 6 (hence *CRY1*Δ6). A genotype-phenotype correlation revealed that this variant segregates with ADHD and DSPD in the family (Figure 8, A–C, and Supplemental Table 13). When the phenotype of *CRY1*Δ6 carriers was compared with that of *CRY1*Δ11 carriers, we noted more severe psychiatric symptoms such as anxiety and oppositional defiant characteristics.

*Functional characterization of CRY1*Δ*6.* We characterized the functional consequences of the *CRY1*Δ6 variant at the molecular level. The secondary pocket of CRY1 is partially encoded by exon 6 (Figure 8D) and interacts with the Per-Arnt-Sim (PAS-B) domain of CLOCK (36). Docking simulations of modeled CRY1Δ6 and WT CRY1 with CLOCK, using HADDOCK (high-ambiguity–driven protein-protein docking) (37, 38), suggested that R256 and F257, encoded by exon 6 of CRY1, are essential for the interaction with CLOCK (Figure 8E). Further analysis indicated that CLOCK does not fit into CRY1Δ6 as strongly as it does into WT CRY1 (Supplemental Table 14). We hypothesized that exon 6 of CRY1 could be critical for the interaction with BMAL1 and CLOCK (BMAL1/CLOCK) proteins; therefore, it may be unable to repress BMAL1/CLOCK transactivation. We tested this hypothesis by transfecting human embryonic kidney 293T cells (HEK293T cells) with a *Per1::Luc* reporter and other appropriate plasmids. Repressor activity of CRY1Δ6, which lacks amino acid residues from 229 to 275, was substantially less than that of full-length (FL) WT *CRY1* and *CRY1*Δ11 (Figure 8F). Then, we performed coimmunoprecipitation (co-IP) assays to assess the interaction between CRY1Δ6 and BMAL1/CLOCK. The CRY1Δ6 variant had a severe deficit in its ability to coimmunoprecipitate with the BMAL1/CLOCK heterodimer (Figure 8G), independent of PER2 (Figure 8H). We next used a rescue assay to determine the effect of CRY1Δ6 function on the circadian rhythm (39). When expressed under the control of its endogenous promoter and an intronic element, *mCry1* can rescue rhythms in the bioluminescence reporter *Per2:Luc* (*Per2* promoter fused with the luciferase gene) in *Cry1*$^{-/-}$ *Cry2*$^{-/-}$ double-KO (DKO) mouse embryonic fibroblasts (MEFs) (37). Human WT *CRY1* and the *CRY1*Δ11 variant res-

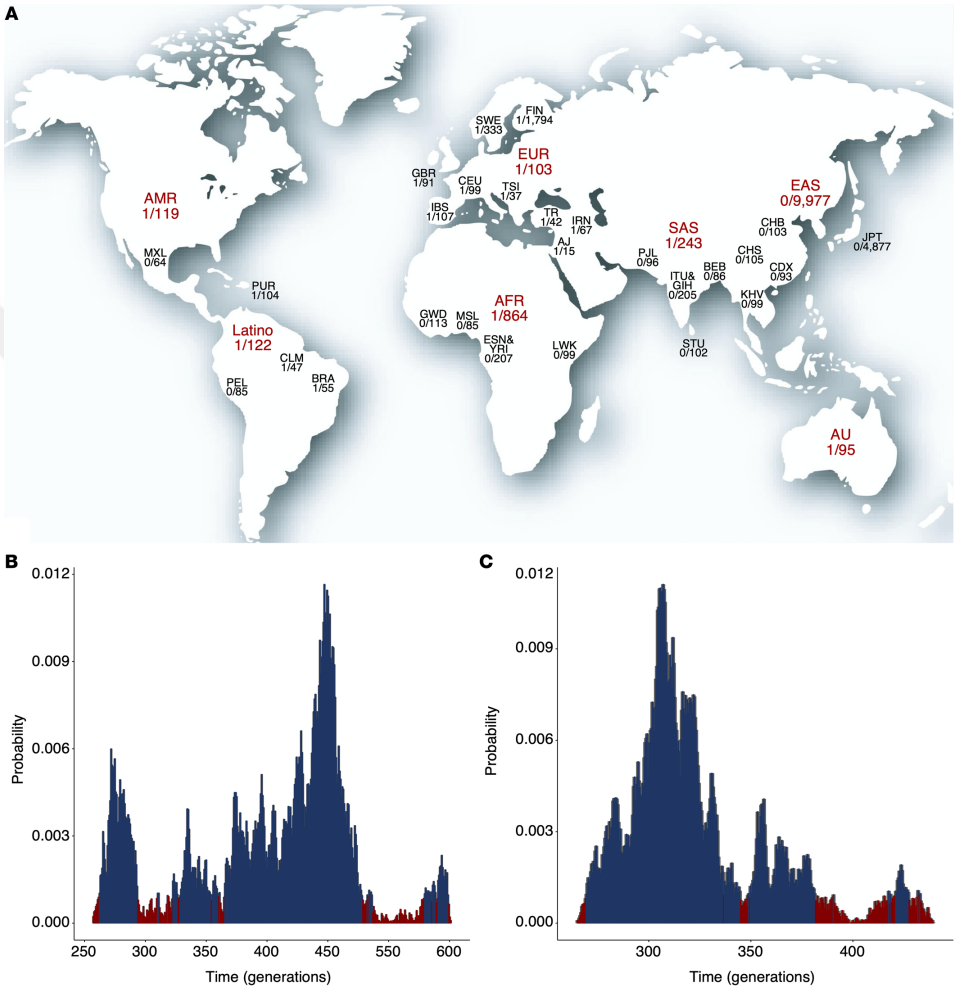



**Figure 6. Allele frequency and age distributions of *CRY1*Δ11 in different populations.** (**A**) The highest allele frequencies were observed in Ashkenazi Jewish, Italian, Turkish, Brazilian, Iranian, and non-Finnish European populations. The world map is reproduced from 3D Geography. Posterior probability distribution plots depict peaks at (**B**) 447 generations (TR; 9 carriers and 438 noncarriers; 95% CI, 262–548 generations) and (**C**) 257 generations (GBR, IBS, and CEU; 3 carriers and 294 noncarriers; 95% CI, 219–382 generations).

cued the circadian rhythm, but not the *CRY1*Δ6 variant (Figure 8I). We also confirmed that the circadian period increased for *CRY1*Δ11 by approximately 26 minutes compared with that for WT *CRY1*, consistent with the previous study (1).

CRY1 stability affects the periodicity of the circadian rhythm (40, 41). In order to test the degradation rate of CRY1Δ6, we expressed a CRY1Δ6::Luc fusion protein in HEK293T cells and monitored the decay in luminescence as a reporter for protein degradation. Our results indicated that the half-life of CRY1Δ6 (~3 hours) was significantly higher than that of WT CRY1 (~1.9 hours) and CRY1Δ11 (~2 hours) (Figure 8J). Collectively, these data show that although CRY1Δ6 is more stable than WT CRY1, its reduced affinity for BMAL1/CLOCK caused an arrhythmic phenotype when it was expressed in the DKO MEFs.

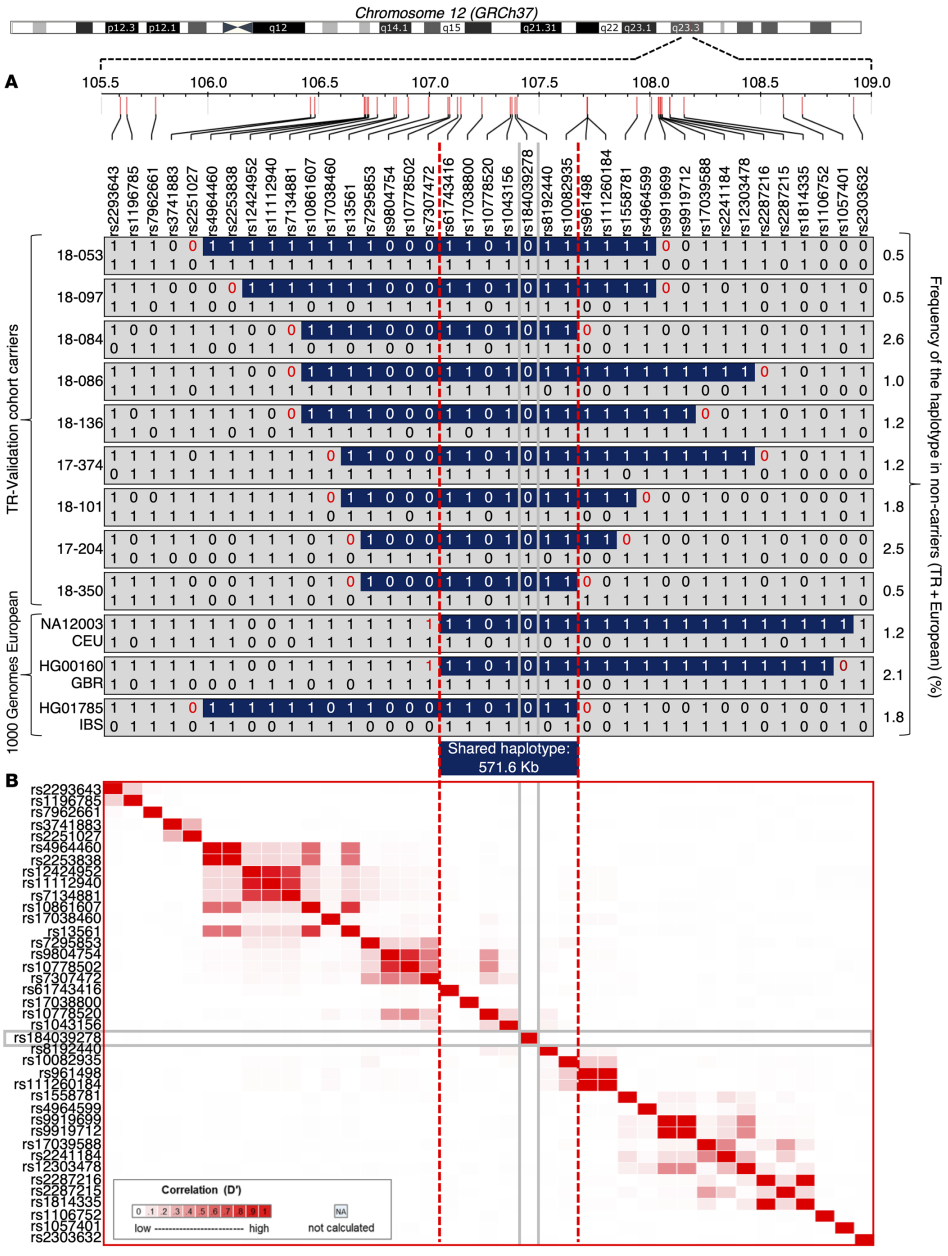

Chromosome 12 (GRCh37)
A
105.5
106.0
106.5
107.0
107.5
108.0
108.5
109.0
rs2293643
rs1196785
rs7962661
rs3741883
rs2251027
rs4964460
rs2253838
rs12424952
rs11112940
rs7134881
rs10861607
rs17038460
rs13561
rs7295853
rs9804754
rs10778502
rs7307472
rs61743416
rs17038800
rs10778520
rs1043156
rs184039278
rs8192440
rs10082935
rs961498
rs111260184
rs1558781
rs4964599
rs9919699
rs9919712
rs17039588
rs2241184
rs12303478
rs2287216
rs2287215
rs1814335
rs1106752
rs1057401
rs2303632
TR-Validation cohort carriers
18-053
18-097
18-084
18-086
18-136
17-374
18-101
17-204
18-350
1000 Genomes European
NA12003
CEU
HG00160
GBR
HG01785
IBS
Frequency of the haplotype in non-carriers (TR + European) (%)
Shared haplotype:
571.6 Kb
B
Correlation (D')
low
high
not calculated

**Figure 7. Shared haplotypes for 39 SNPs at 12q23.3 and LDmatrix. (A)** Shared haplotypes in the 3-Mb region spanning the *CRY1* gene in carriers of *CRY1*Δ11 ($n$ = 12) and noncarriers ($n$ = 732) from the 447-individual validation and the 1000 Genomes Project cohorts. Only the haplotypes of the carriers and the frequency of the haplotypes in noncarriers are shown in the figure. All the markers on chromosome 12 were phased using SHAPEIT, version 2.17, but only 39 are represented. **(B)** Heatmap of pairwise LD statistics for 39 SNP targets that were determined with the LDmatrix module of LDlink software using the 1000 Genomes Project European subpopulation. Pairwise LD values between the SNPs are described by white-red shading: $R^2$ = 0, white; $R^2$ = 1, red.

## Discussion

Our genotype-first analysis identified the first coding variants to our knowledge to be significantly associated with ADHD and indicates an important role for a common variant that alters the circadian rhythm. ADHD is a highly heritable psychiatric disorder that affects nearly 5% of children and teenagers and 2.5% of adults globally. ADHD is considered primarily a disorder of impulse control deficit, difficulties in delaying gratification, altered patterns of motivation, and hyperkinesis (42). Interestingly, functional connectivity of brain networks involved in behavior and cognition are implicated in the etiology of ADHD. More specifically, neuroimaging research has revealed functional and maturational abnormalities such as a delay in reaching peak thickness of much of the cerebrum including the prefrontal cortex. Also, delayed maturation and atypical interactions of brain networks, which are involved in the regulation of attentional resources, were shown using functional neuroimaging. In addition, reduced activation of the dopaminergic mesolimbic system was observed in individuals with ADHD (see ref. 42 for a comprehensive review).

Core clock proteins act on nearly 30% of all genes by binding to the E-box sequence (CAGGTG) in their promoters and hence exhibit daily rhythmicity (43). Clock, Bmal, Per, and Cry genes are expressed broadly throughout the brain, including in several limbic regions responsible for mood regulation and brain reward. Experiments using *Drosophila* and mice demonstrated that mutations in circadian genes result in depressive or manialike symptoms and affect sensitization to addiction (7). Polymorphisms in *CLOCK* and *PER3* were also suggested to be associated with an increased rate of depressive relapses in patients with bipolar disorder. Therefore, it is conceivable that mutations in circadian genes can interfere with the function of a substantial number of genes that underlie various psychiatric phenotypes. It is tempting to speculate that disturbances in the daily rhythmicity of clock-controlled genes are critical for adaptive plasticity of the brain and optimal functioning of neuronal structures that are important for an ongoing process of assessing the environment, coping with it, and enabling the individual to anticipate and deal with future challenges.

*CRY* genes are essential cogs in the core clock machinery. Our characterization of *CRY1*Δ11 and *CRY1*Δ6 mutations in individuals with combined ADHD and DSPD has uncovered what we believe to be a novel mechanism for a distinct combination of behavioral phenotypes in humans. In addition to ADHD and DSPD, we detected high rates of ADHD comorbidities including depression and smoking behavior in *CRY1*Δ11 carriers. Furthermore, we identified significant associations of *CRY1*Δ11 with MDD, anxiety, nicotine dependence, and glaucoma. Our results contribute to an unveiling of mechanisms behind the biological overlap of sleep traits with psychiatric traits, as well as support the previously reported findings on ADHD comorbidities (10, 12).

Both the *CRY1*Δ11 and *CRY1*Δ6 mutations affect the periodicity of the circadian rhythm. Whereas *CRY1*Δ6 causes an arrhythmic phenotype, *CRY1*Δ11 lengthens the circadian period by approximately half an hour. We did not observe a significant difference in the severity of ADHD in individuals carrying *CRY1*Δ11 or *CRY1*Δ6; however, we noted more severe psychiatric symptoms related to anxiety and oppositional defiant disorder in *CRY1*Δ6 carriers. In terms of frequency, *CRY1*Δ6 is a private variant, whereas *CRY1*Δ11 appears with high frequency in the Eastern Mediterranean, European, and European-derived populations. Allele frequency data and haplotypes of *CRY1*Δ11 carriers from different populations indicate that the mutation might have originated in individuals from the Eastern Mediterranean region and expanded to the West into Europe and to the East into Persia, consistent with the Neolithic migration of Anatolian farmers (33).

Multiple studies report lower rates of ADHD in regions with high solar intensity (27). In order to determine whether exposure to sunlight ameliorates ADHD symptoms, we recorded the mean durations of sun exposure in *CRY1* mutation carriers. Interestingly, we found that carriers with longer sun exposure durations, especially work-imposed exposure, either had milder phenotypes or were in partial remission. These results suggest that sunlight exerts a protective effect for ADHD symptoms.

**Table 1. Association of CRY1Δ11 with the BioMe BioBank phenotypes**

| Phenotype name | No. of cases | Carrier case/control | WT | OR (95% CI) | $P$ | MAF | Incidence (1 in ...) | ICD-10-CM |
|---|---|---|---|---|---|---|---|---|
| MDD, single episode | 636 | 37/201 | 599/6226 | 1.91 (1.29–2.75) | $7.87 \times 10^{-4}$ | 0.03 | 17 | F32.0, F32.1, F32.2, F32.4, F32.5, F32.9, F32.89 |
| Insomnia | 488 | 28/210 | 460/6365 | 1.84 (1.18–2.78) | $3.87 \times 10^{-3}$ | 0.03 | 17 | G47.00, F51.01, F51.09 |
| Anxiety disorder | 690 | 36/202 | 654/6171 | 1.68 (1.13–2.43) | $4.56 \times 10^{-3}$ | 0.03 | 19 | F41.1, F41.8, F41.9 |
| Glaucoma | 54 | 6/232 | 48/6777 | 3.65 (1.26–8.66) | $7.11 \times 10^{-3}$ | 0.06 | 9 | H40.9 |
| MDD, recurrent | 100 | 8/230 | 92/6733 | 2.55 (1.05–5.31) | $1.28 \times 10^{-2}$ | 0.04 | 13 | F33.0, F33.1, F33.2, F33.40, F33.41, F33.42, F33.8, F33.9 |
| Nicotine dependence | 204 | 12/226 | 176/6649 | 2.01 (1.00–3.66) | $2.52 \times 10^{-2}$ | 0.03 | 16 | F17.200 |
| ADHD | 43 | 3/235 | 40/6785 | 2.17 (0.42–6.87) | $1.20 \times 10^{-1}$ | 0.03 | 14 | F90.0, F90.1, F90.2, F90.8, F90.9 |

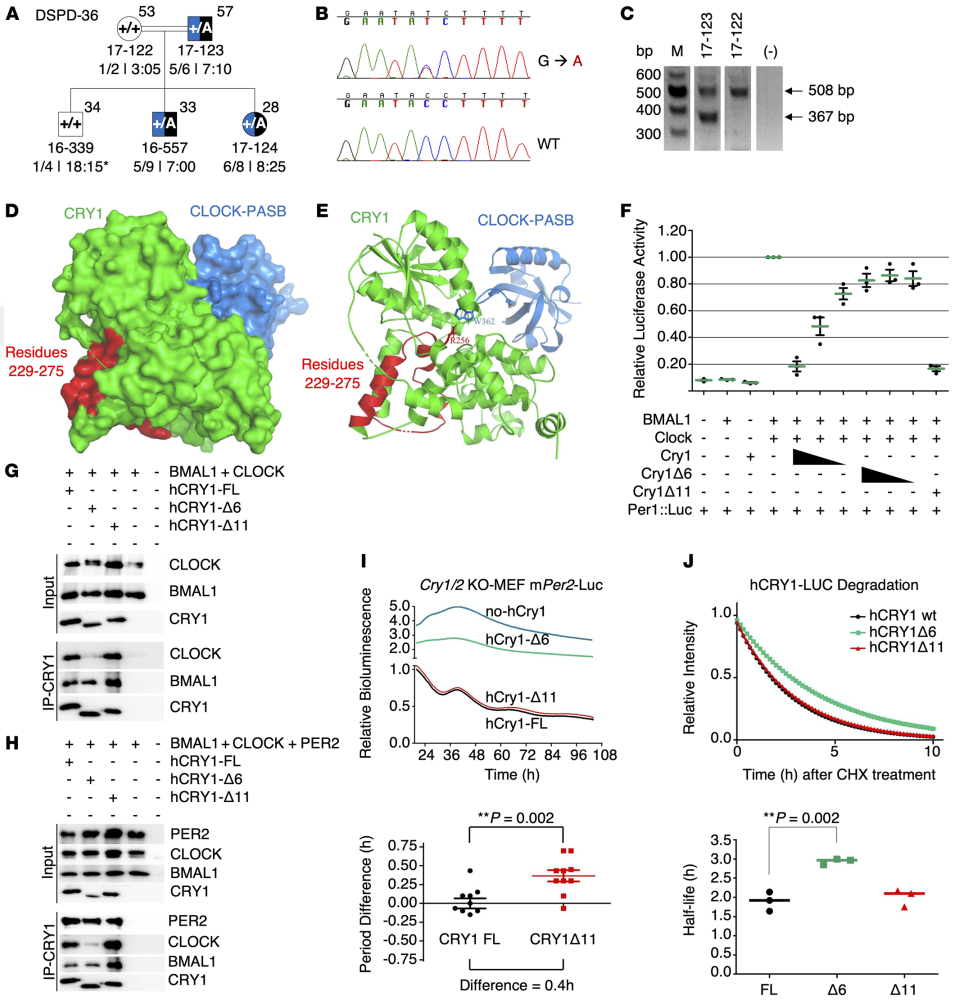



**Figure 8. Phenotype-genotype characterization of family DSPD-36 and functional characterization of *CRY1*Δ6.** (**A**–**C**) Phenotype-genotype characterization of family DSPD-36. (**A**) The family DSPD-36 was assessed as described in the legend to Figure 3. (**B** and **C**) c.825+1G>A causes skipping of 141-bp exon 6, leading to an in-frame deletion of 47 residues in the middle of the PHR and clock-binding domains of the CRY1 protein. M, DNA marker. (**D**–**J**) Functional characterization of *CRY1*Δ6. (**D**) Docking analysis of CRY1 and the CLOCK PAS-B domain. (**E**) The critical amino acid residue of CRY1 (R256) interacts with CLOCK PAS-B (W362). (**F**) Analysis of the effect of FL WT *CRY1*, *CRY1*Δ6, and *CRY1*Δ11 on CLOCK/BMAL1-driven transcription with the *Per1:Luc* assay. An E-box–driven luciferase reporter plasmid *Per1:Luc* was coexpressed in HEK293T cells along with plasmids consisting of *CLOCK* and *BMAL1* cDNAs and decreasing amounts of WT *CRY1* or *CRY1*Δ6. Data represent the mean induction of bioluminescence over basal levels from duplicate transfections 48 hours after transfection. Error bars represent the SD from at least 3 biological experiments. Co-IP assay with (**G**) CLOCK, BMAL1, (**H**) PER2, CLOCK, BMAL1, and human CRY1s (hCRY1). Blots in **G** and **H** are representative of at least 3 independent experiments. (**I**) Rescue assays were performed with human *CRY1*s with at least 4 biological replicates. Samples were normalized to the initial luminescence signal. The graph below indicates the period differences, with the whiskers representing the mean ± SEM values. *n* = 5 per group. Data were pooled from 2 independent experiments. $^{**}P = 0.002$, by unpaired *t* test. m*Per2-Luc*, mouse *Per2-Luc*. (**J**) Degradation assay with human CRY1::Luc and its variants. The graph below was generated by fitting a 1-phase decay curve to the data, which indicate the half-life, with the horizontal line representing the mean. *n* = 3 per condition, pooled from 3 independent experiments. $^{**}P = 0.002$, by unpaired *t* test.

In conclusion, we describe a monogenic form of combined ADHD and DSPD frequently accompanied by a history of depression due to pathogenic *CRY1* mutations. Furthermore, our findings provide a mechanistic explanation for the development of these behavioral phenotypes by linking a common and causal genetic variant with a compromised circadian period due to disturbances in the negative feedback loop of the core molecular clock. Although these observations are consistent with a substantial epidemiological comorbidity of psychiatric disorders, we believe they provide a novel perspective on their genomic architecture. Psychiatric disorders are characterized by a polygenic nature with many genetic loci contributing to risk (13). However, in the case of *CRY1*, mutations at a single locus could lead to what may be one of the most common autosomal dominant disorders. Therefore, *CRY1*Δ11 has significant potential in diagnostic testing (44) and presents a target for therapeutic intervention. Reverse phenotyping of individuals and families with damaging mutations in core clock genes and genotyping for circadian rhythmicity in well-characterized cohorts of psychiatric disorders could pave the way to dissect the constitutional determinants of a distinct group of circadian psychiatric phenotypes that we propose to designate as circiatric disorders.

## Methods

*Patient evaluations and genetic material.* Clinical information included medical history, a physical examination, a psychiatric evaluation, pedigree drawings, a complete blood count, blood and urine biochemistry analysis, and height and weight measurements for BMI determination. Genomic DNA was extracted from blood cells using standard procedures and the NucleoSpin Blood L Kit (Macherey-Nagel). For those families living outside of Ankara, several of the investigators traveled to the families' hometowns in Konya (DSPD-1), Urfa (DSPD-4), and Kayseri (DSPD-6) to perform the clinical evaluations. Families in Italy were evaluated at Siena University. All participants completed the adult ASRS questionnaire (22), developed by the WHO to measure symptoms of ADHD, and the MCTQ (21). The psychiatric analysis began from the childhood period, and paid special attention to establishment of trust to minimize the drive to give appropriate rather than candid answers. Clarity of communication was equally important to make sure that correct words were chosen in the expression of emotions by each subject. Special attention was paid to dissect whether symptoms were secondary to another psychiatric disorder. Several of the investigators reviewed the ASRS questionnaires and the MCTQs completed by each participant. Clinical data on the family members (DSM-5 ADHD and ASRS scores, ADHD severity, demographics, ADHD symptoms, and sleep behavior) are presented in Supplemental Tables 1–3. ADHD diagnosis according to the DSM-5 requires 6 symptoms for children younger than 17 years of age and 5 or more symptoms for older adolescents and adults. Phenotype components in the families included excessive inattention and/or hyperactivity and impulsivity as well as executive dysfunction, lack of emotional self-control, and motivation frequently present with characteristics of oppositional defiance. The current severity of ADHD was specified as mild (few if any symptoms, which result in only minor functional impairments); moderate (functional impairments or symptoms between mild and severe); and severe (presence of symptoms that result in marked impairments in social or occupational functioning). In Figure 3, ADHD is represented with blue and DSPD with black colors. Individuals for whom a definitive ADHD diagnosis was made were further classified as combined (when all 3 core features of inattention, hyperactivity, and impulsivity were present); predominantly inattentive (diagnosed if ≥5 symptoms of inattention but <5/6 symptoms of hyperactivity/impulsivity had persisted for ≥6 months); and predominantly hyperactive/impulsive (diagnosed if ≥5/6 symptoms of hyperactivity/impulsivity but <5/6 symptoms of inattention had persisted for ≥6 months). Supplemental Table 2 provides data on the DSM-5 symptoms of inattention (questions 1–9), DSM-5 symptoms of hyperactivity and impulsivity (questions 10–18), and ASRS scores for the 14 families and individual family members. Questions 1–4 of Part A and questions 7–11 of Part B are for inattention, and questions 5 and 6 of Part A and questions 12–18 of Part B are for hyperactivity/impulsivity evaluation. Sleep behavior was independently assessed by several of the investigators through a sleep interview, which included sleep and ChronoType questionnaires. As previously reported (1), for the families DSPD-1, -4, -6, -7, -9, and -14, DSPD is part of the behavioral phenotype and is also present in families DSPD-2, -31, -34, -51, -52, -53, -58, and -59 (Supplemental Table 3) reported in the current study. Note that data on additional individuals from families DSPD-1, -4, -6, and -9 are now presented in Supplemental Table 3. Nine families (DSPD-1, -2, -4, -14, -6, -9, -51, -52, and -31) are consanguineous or endogamous (Figure 3) and are from different cities or towns located in Anatolia. An important observation emerging from DSPD-4 and DSPD-52 is that there is no marked phenotypic difference for DSPD or ADHD between the homozygotic (16-006, 16-008, 16-018, 16-042, 16-049, 17-281) and heterozygotic (16-014, 16-015, 16-016, 16-027, 16-043, 16-052, 17-011, 17-292) individuals.

*CRY1 c.1657+3A>C amplification and genotyping.* *CRY1* c.1657+3A>C genotype status was determined by amplifying genomic DNA using hCry1i10F (5′-GTCAACACTTCTGTGAGCCT-3′) and hCry1i12R (5′-CAGATGCATGTCTCTTGACC-3′) and restriction digestion analysis (1). The PCR yielded a 623-bp product of the genomic locus containing exon 11 and was digested with Hpy188I (+ allele: no cut, variant c.1657+3A>C: 276 bp + 347 bp; Supplemental Figure 2).

*Whole-exome sequencing.* Whole-exome sequencing (WES) was performed on genomic DNA from the 447-individual validation cohort at the YCGA (Supplemental Table 15). Exome capture was done using the xGen Exome Research Panel (version 1.0) Capture Kit (Integrated DNA Technologies [IDT]) according to the manufacturer's protocol. Samples were sequenced on the HiSeq 4000 platform (Illumina) with 100-bp paired-end reads. WES data and associated sample information described in Supplemental Table 6 have been deposited in the NCBI's dbGAP repository (accession ID: BioProject PRJNA624188; https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA624188). Base calling, read filtering, and demultiplexing were performed with the standard Illumina processing pipeline. The read pairs were mapped to the human genome build GRCh37 with the Burrows-Wheeler Aligner (BWA), version 0.7.17, with default settings (45). Aligned duplicate reads were marked using Mark Duplicates in Picard tools. GATK, version 3.7 (Genome Analysis Toolkit [GATK]), was used for base quality score recalibration (BQSR) and local realignment around indels to refine alignment artifacts around putative insertions or deletions (46). Variant discovery was performed in 2 steps, beginning with variant calling with GATK HaplotypeCaller (https://gatk.broadinstitute.org/

hc/en-us/articles/360037225632-HaplotypeCaller) followed by joint genotyping using GATK GenotypeGVCFs (https://gatk.broadinstitute.org/hc/en-us/articles/360037057852-GenotypeGVCFs). Variants with a Phred quality score below 30 were removed. The resulting variant call set was refined using variant quality score recalibration (VQSR) as implemented in GATK VariantRecalibrator (https://gatk.broadinstitute.org/hc/en-us/articles/360036510892-VariantRecalibrator). Variant recalibration was applied by the GATK ApplyRecalibration walker using a tranche sensitivity of 99.5% for SNPs and 99.0% for indels. VQSR was used to define low-quality variants for downstream processing (Supplemental Table 16).

Variants were trimmed and left-aligned around indels, and multiallelics were split using GATK, version 3.7. A total of 886,935 variants were obtained. Sample-based quality control was carried out using PLINK, version 1.9, software (47). No low-quality samples with more than 10% missing genotypes were identified. Sex verification and kinship analysis were performed using KING software (48). No related individuals were detected (degree = 2, kinship coefficient ≥0.0625) in the validation cohort (Supplemental Figure 1, A and B). To determine outliers of the population, a principal component (PC) analysis was conducted on a subset of the common biallelic variants ($n$ = 43,557) pruned for linkage disequilibrium (LD) using the PLINK, version 1.9, "indep-pairwise" command (window = 50 SNPs, step size = 5 SNPs, maximum r2 = 0.5). The first 10 PCs were calculated using the "smartpca" module of the EIGENSTRAT method (EIGENSOFT package) (49), and no outlier samples were observed (Supplemental Figure 1C).

*Variant annotation and prioritization*. Variants in protein coding genes were identified by SnpEff, version 4.4 (50), which uses the ENSEMBL, version 87, gene models to determine variant functional region and impact on the assigned gene. Variants were annotated using ANNOVAR (version 2019Oct24) (51). Variants were subsequently filtered out on the basis of quality control scores, MAFs, deleteriousness/functional impact, and variants at low-complexity regions.

Briefly, common variants defined by a MAF of more than 0.1% in GnomAD, version 2.1.1 (31), Kaviar (52), 1000 Genomes Project (32), or ESP6500 (53), and an in-house Turkish unrelated control database of 2628 whole-exome and 773 whole-genome data sets were excluded from the analysis. The potential impact of missense variants was predicted using MetaSVM (54) and Combined Annotation–Dependent Depletion (CADD) (55) tools, and that of splice site variants located ±3 bp of exon-intron junctions was predicted using dbscSNV-ADA/-RF (56) and Spidex scores (ref. 57 and Supplemental Table 16).

*Candidate gene prioritization approach*. We performed analyses of associations between ADHD and rare deleterious mutations using 31,432 variants on 11,528 genes. Single variant–based analysis was carried out using PLINK, version 1.9. Small-sample adjustments and rare variant weights were used for gene-based analysis with SKAT-O and the burden test with Bonferroni's multiple testing correction (30). The results of the single-variant association test and gene-based statistical analyses were used to create Manhattan and quantile-quantile (Q-Q) plots, respectively (Figure 5, A and B). Although small-sample adjustments were applied, the Q-Q plots still had a slightly anticonservative pattern.

*Other core clock genes*. The presence of additional polymorphisms was not unexpected, given the large degree of variation commonly found in human clock genes (58, 59). In addition to the *CRY1* c.1657+3A>C variant, other coding, rare, and deleterious variations of core clock genes (*CRY1*, *CRY2*, *PER1*, *PER2*, *PER3*, *ARNTL*, *CLOCK*, and *CSNK1D*) and additional candidate clock genes (*CSNK1E*, *ARNTL2*, *FBXL3*, *FBXL21*, *BHLHE40*, *BHLHE41*, *NR1D1*, and *RORA*) were identified, but none of the genes other than *CRY1* were statistically significant for the association with the phenotypically distinguished groups tested (Supplemental Tables 12 and 17).

*Haplotype analysis*. We inferred haplotypes from chromosome 12 of the 447-individual validation cohort and European populations from 1000 Genomes Project, Phase 3 using phased SNPs by SHAPEIT, version 2.17 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html). The same SNPs were used to calculate the linkage disequilibrium across ancestral populations using LDlink (https://ldlink.nci.nih.gov/?tab=home) and to estimate age using DMLE+. SHAPEIT uses a set of genotypes and a genetic map and produces a set of estimated haplotypes 60). We investigated the 1000 Genomes Project, Phase 3 haplotype data and noted 3 *CRY1*Δ11 carriers in the European populations of Utah residents with Northern and Western European ancestry (CEU), the British in England and Scotland (GBR), and the Iberian population in Spain (IBS). All 594 European and 894 Turkish (TR) haplotypes were combined. Thirty-nine phased, biallelic, informative SNPs (MAF >0.05) in the 3-Mb region spanning *CRY1*Δ11 were evaluated, and 12 different haplotypes in carriers were identified. These haplotypes share a 571.6 kb common segment, which involves *CASC18*, *NUAK1*, *TCP11L2*, *POLR3B*, *RFX4*, *MTERF2*, *TMEM263*, *CRY1*, *BTBD11*, *PWP1*, *PRDM4*, *WSCD2*, and *CMKLR1*. The frequencies of 12 haplotypes were in the range of 0.5% to 2.6% in 732 noncarriers (Figure 7A).

To assess LD across ancestral populations, the LDmatrix module of LDlink (61) was used, and an interactive heatmap matrix of pairwise linkage disequilibrium using 39 SNPs was created (Figure 7B). The 1000 Genomes Project, Phase 3 haplotype data on populations from CEU, GBR, IBS, and Tuscany in Italy (TSI) were extracted.

*Age estimation of CRY1*Δ11 *in Turkish and European populations*. DMLE+, version 2.3 (62), was used to estimate the age of the *CRY1*Δ11 mutation, with the recommended burn-in and sampling intervals and a variety of parameter ranges. We used the haplotypes generated from 38 phased SNPs spanning 3 Mb around the *CRY1*Δ11 mutation in unrelated individuals. The population growth rate (r) was estimated (e) using the equation: $T_1 = T_0 e^{(gr)}$, in which $T_1$ is the estimated size of the current population, $T_0$ is the estimated size of the ancestral population, and g is the number of generations between these 2 time points (63). The growth rate of the Turkish population was estimated as 0.009 ($T_1$ = 81.81 million, $T_0$ = 12 million [200 BCE] and g = 88.7) (64), and the growth rate of the European (GBR, IBS, and CEU) populations in the 1000 Genomes Project was estimated as 0.016 ($T_1$ = 116.26 million, $T_0$ = 20.75 million on 1 CE, and g = 80.7) (65). A 25-year intergeneration interval was used for calculations. The "proportion of disease-bearing chromosomes sampled" was estimated as $3.11 \times 10^{-6}$ for the Turkish cohort and $5.47 \times 10^{-6}$ for the European cohort, using the population sizes ($T_1$) and carrier frequencies (1 of 42 and 1 of 100). The mutation density depicted a peak at 447 generations (95% credible set = 262–548) and 257 generations (95% credible set = 219–382) for the Turkish and the European populations, respectively.

*PheWAS of CRY1*Δ11 *in the BioMe BioBank*. PheWAS analyses for *CRY1*Δ11 were performed on the basis of electronic medical record–linked phenotypes (ICD-10-CM codes) in the BioMe BioBank of the Institute for Personalized Medicine at the Icahn School of Medicine at

Mount Sinai. The phenotype information and *CRY1*Δ11 status of 9438 unrelated adult Europeans were analyzed. The *CRY1*Δ11 phenotype association was tested independently using Fisher's exact test.

*Reverse transcription PCR analysis of CRY1 mRNA*. Fresh venous blood samples were collected into PAXgene Blood RNA tubes (PreAnalytiX), and total RNA was isolated from subjects 17-122 and 17-123 using the QIAamp RNA Blood Mini Kit with on-column DNase digestion (QIAGEN). Equal amounts of total RNA were used for first-strand cDNA synthesis using the RevertAid First-Strand cDNA Synthesis Kit with Oligo(dT)$_{18}$ priming followed by RNase H digestion (Thermo Fisher Scientific). The resulting coding change of CRY1 c.825+1G>A was tested by amplifying the part of the cDNAs between exons 5 and 8 using CR508F (5′-GGAGAAACTGAAGCACTTACTC-3′) and CR508R (5′-CAAATACCTTCATTCCTTCTTCCC-3′). The PCR yielded a FL 508-bp product in the WT individual and an additional 367-bp product in the heterozygous proband (Figure 8C).

*Human CRY1 cloning and mutagenesis*. The WT coding sequence for *CRY1* was obtained from a human cDNA sample designated as 17-125. Two microliters of this cDNA sample was amplified via touchdown PCR using Phusion polymerase (Thermo Fisher Scientific) and the primers provided in the Supplemental Table 18. Amplified fragments were initially cloned to pJET1.2/blunt vector (Thermo Fisher Scientific) using the CloneJET PCR Cloning Kit (Thermo Fisher Scientific) for individual clones destined for pcDNA4A and pMU2. Sequence verification of amplified inserts was performed via automated Sanger sequencing (Macrogen) to confirm the absence of mutation(s) possibly introduced during the PCR amplification. Sequence-verified inserts were subcloned to their respective vectors (pMU2 and pcDNA4/Myc-His A) via restriction/ligation. Mutagenesis was performed for both *CRY1* integrated constructs via Phusion-based, site-directed mutagenesis using the oligonucleotide primers listed in Supplemental Table 19. Sequence verification was repeated as described above. PCR reactions for each construct were prepared in 50 μL of total volume using the reaction conditions in Supplemental Table 20.

To clone *CRY1* into pMU2 and pcDNA4/Myc-His A, flanking XbaI and NotI sites were added to the primers. For the pcDNA4/Myc-His A construct, a stop codon was removed, and 2 extra nucleotides were added to include the His tag present in the plasmid.

Touchdown PCR was performed on a T100 Thermal Cycler (Bio-Rad). PCR reactions were set up on ice and transferred to a preheated thermocycler. The cycling conditions for the touchdown PCR reaction are described in Supplemental Table 21.

The sizes of the PCR products were verified by agarose gel electrophoresis, and the band corresponding to *CRY1* was excised and purified using a NucleoSpin PCR and Gel Purification Kit (Macherey Nagel). The purified *CRY1* fragment was then ligated to an empty pJET1.2/blunt vector using a CloneJET PCR Cloning Kit (Thermo Fisher Scientific) following the manufacturer's protocol. Ligation reactions (5 μL) were transformed to DH5α cells, and transformed cells were spread on Luria-Bertani (LB) broth agar plates supplemented with 100 μg/mL ampicillin and then incubated overnight at 37°C. Colonies from plates were selected the next day and grown in 2 mL liquid LB, and plasmids were purified using a plasmid purification kit (Macherey Nagel). Plasmids were then digested with the appropriate restriction endonucleases to confirm the presence of the insert with gel electrophoresis and plasmids. The plasmids were then sequenced to confirm the absences of mutation(s) of the *CRY1* gene via automated Sanger sequencing (Macrogen). Next, plasmids were double digested with XbaI/NotI for pMU2 inserts and EcoRV/NotI for pcDNA4/Myc-His A inserts. pMU2 and pcDNA4/Myc-His A were also double-digested with XbaI/NotI and EcoRV/NotI, respectively. Digested destination vectors were treated with FastAP (Thermo Fisher Scientific) to limit self-annealing. Both inserts and vectors were gel purified and ligated using T4 DNA ligase (Thermo Fisher Scientific) to generate pMU2-h*CRY1* and pcDNA4A-h*CRY1* plasmids and then transformed into DH5α cells. pMU2-*CRY1*–transformed cells were plated onto LB agar supplemented with 34 μg/mL chloramphenicol, and pcDNA4A-*CRY1*–transformed cells were plated onto LB agar supplemented with 100 μg/mL ampicillin and incubated overnight at 37°C. The presence of the human *CRY1* cDNAs in these vectors was verified with gel electrophoresis using the appropriate restriction endonucleases.

The deletion of exons 6 and 11 from human *CRY1* cDNA was performed with a PCR-based strategy using Phusion Polymerase and asymmetric oligonucleotides incorporating 20-nt homology designed to incorporate missense mutations and deletions on pMU2-*CRY1* and pcDNA4A-*CRY1* constructs (the conditions and forward/reverse primers for PCR mutagenesis are listed in the Supplemental Tables 22 and 23).

Phusion-based mutagenesis PCR reactions were performed on a T100 Thermocycler (Bio-Rad). PCR reactions were set up on ice and transferred to a preheated thermocycler. The cycling conditions for the mutagenesis PCR reaction are provided in Supplemental Table 23.

After mutagenesis with PCR, the reactions were treated with DpnI at 37°C for 3 hours to eliminate parental template plasmid DNA and then transformed into DH5α cells. pMU2-*CRY1*–transformed cells were plated onto LB agar supplemented with 34 μg/mL chloramphenicol and pcDNA4A-*CRY1*–transformed cells were plated onto LB-agar supplemented with 100 μg/mL ampicillin and incubated overnight at 37°C. The presence of mutations was verified with automated Sanger sequencing (Macrogen).

*Real-time bioluminescence rescue assay*. *Cry1*$^{-/-}$ *Cry2*$^{-/-}$ MEFs (CRY-DKO MEFs) ($3 \times 10^5$) were seeded in 35-mm clear tissue culture plates. Cells were transfected with 4000 ng pGL3-*Per2-Luc* (luciferase reporter) and 150 ng *CRY1* plasmid [pMU2-P(*CRY1*)-(intron 336)]designed to rescue the circadian rhythm using FuGENE 6 Transfection Reagent (Promega) according to the manufacturer's protocols. The total DNA amount was equalized to 4150 ng with pSport6 plasmid if only the reporter plasmid was transfected. Seventy-two hours after transfection, cells were synchronized with 0.1 μM dexamethasone for 2 hours. Medium was replaced with bioluminescence recording medium (1% DMEM powder [w/v], 0.035% sodium bicarbonate, 0.35% D[+] glucose powder, 1% mL 1M HEPES buffer, 0.25% penicillin/streptomycin, 5% FBS), in which luciferin was freshly supplemented (0.1 mM final concentration). Plates were sealed with vacuum grease and placed into the LumiCycle (Actimetrics). Bioluminescence monitoring was performed for 70 seconds every 10 minutes for 7 days via photomultiplier tubes. Luminescence values were recorded and processed using LumiCycle Analysis software. The first 20 hours of data were discarded from the analysis due to transient luminescence upon medium change. Period and amplitude values were obtained using damped sine wave based on the running average option for each sample.

*CRY1-mediated repression of BMAL1-CLOCK transcription activity*. Low-passage-number HEK293T cells ($5 \times 10^6$) were seeded onto a 10-cm plate containing 10 mL DMEM (Gibco, Thermo Fisher Scien-

tific) supplemented with 10% heat-inactivated FBS (Gibco, Thermo Fisher Scientific), 1× penicillin/streptomycin (Gibco, Thermo Fisher Scientific), and 4 mM L-glutamine (Gibco, Thermo Fisher Scientific) (denoted as 1× DMEM). Cells were incubated overnight at 37°C in 5% $CO_2$ until they reached 70%–80% confluence. Cells were washed with 5 mL 1× PBS, trypsinized, and resuspended in DMEM supplemented with 20% heat-inactivated FBS, 2× penicillin/streptomycin, and 8 mM L-glutamine (denoted as 2× DMEM), such that the total concentration was $8 \times 10^5$ cells/mL. Diluted HEK293T cells were distributed to white Costar 96-well culture plates (50 μL/well), rendering the cell concentration $4 \times 10^4$ cells/well.

For each well, a mixture of 50 ng pSport6-*BMAL1*, 125 ng pSport6-*CLOCK*, 50 ng pGL3-Per1::Luc, 1 ng pRL-TK, 4 ng pcDNA4A-*CRY1*, and 120 ng empty pSport6 was prepared in DMEM (without FBS or antibiotics). For a positive control, the mixture was supplemented with 4 ng empty pcDNA4/Myc-His A. This mixture was supplemented with 0.9 μL, 22-kDa linear polyethylenimine (PEI) (Polysciences), vortexed briefly and incubated at room temperature for 20 minutes. Fifty microliters of the mixture was added on top of each well in triplicate. The plates were incubated for 24 hours at 37°C in 5% $CO_2$. Firefly luciferase and *Renilla* luciferase expression was determined using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's protocol.

*CRY1-Luc degradation assay*. Low-passage-number HEK293T cells ($5 \times 10^6$) were seeded onto a 10-cm plate containing 10 mL 1× DMEM. Cells were incubated overnight at 37°C in 5% $CO_2$ until 70%–80% confluence. Forty nanograms of the expression vector (*CRY1*-Luc plasmid) was reverse-transfected into $4 \times 10^4$ HEK293T cells on a white 96-well plate with a flat-bottomed via PEI transfection reagent. Cells were treated with luciferin (0.4 mM final) and HEPES (10 mM final and pH = 7.2) after 48 hours of transfection for 2 hours. Cycloheximide (CHX) (20 μg/mL final) was added to wells to halt the protein synthesis. The plate was sealed with optically clear film. Luminescence readings were collected via Synergy H1 reader every 10 minutes at 32°C for 24 hours. The protein half-life was calculated using 1-phase exponential decay fitting (GraphPad Prism 5, GraphPad Software).

*Immunoprecipitation*. HEK293T cells ($4 \times 10^5$ per well) were seeded onto 6-well tissue plates 24 hours before the transfection. Cells were transfected via PEI with *CRY1*-His-Myc, *CRY1*-Δ11-His-Myc, or *CRY1*-Δ6-His-Myc in pcDNA4-A or empty sport6 with Flag-CMV-*BMAL1*, or Flag-CMV-*CLOCK* for IP with BMAL1, CLOCK, and CRY. The Flag-*PER2*-CMV plasmid was also transfected along with *BMAL1*, *CLOCK*, and *CRY* plasmids to immunoprecipitate 4 clock proteins. For negative control *BMAL1*, *CLOCK* or *BMAL1*, *CLOCK*, and *PER2* were transfected with empty sport6 plasmid instead of *CRY* via PEI transfection reagent. Twenty-four hours after transfection, the cells were harvested via ice-cold PBS. After centrifugation, the pellets were lysed in 300 μL passive lysis buffer (PLB) (15 mM HEPES, 300 mM NaCl, 5 mM NaF, 1% NP40 supplemented with fresh protease inhibitor) for 20 minutes on ice. To get rid of cell debris, the samples were centrifuged for 15 minutes at 13,000 × *g* at 4°C. Ten percent of the supernatant was saved as input. Ni-NTA agarose resin (15 μL) (QIAGEN) per sample was equilibrated by washing 2 times with TBS-300 (15 mM Tris, 300 mM NaCl) supplemented with 25 mM imidazole and 2 times with PLB. The remaining supernatant was added onto the equilibrated resins with 25 mM imidazole. The cell lysates and resins were incubated for 1.5 hours to pull down CRYs. The resins were washed 4 times with TBS-300 (300 μL) with 25 mM imidazole. Proteins were isolated from resins by boiling in Laemmli buffer (31.5 mM Tris-HCl, pH 6.8 buffer 10% glycerol, 1% SDS, 0.005% bromophenol blue, and freshly added 5% β-mercaptoethanol).

Anti-Flag antibody (MilliporeSigma, A9469) was used to detect BMAL1, CLOCK, and PER2. Blots were stripped (Advansta Strip-It Buffer [R03722-D50]) and incubated with anti-Myc antibody (Abcam, ab18185) to detect CRYs. In all Western blot analyses, the blots were incubated overnight with a primary antibody. The murine IgGκ–binding protein (m–IgGκ BP) conjugated to HRP (Santa Cruz Biotechnology, sc-516102) was used as the secondary antibody. To capture the chemiluminescent signals, WesternBright ECL HRP substrate (Advansta, K-12045-D20) and Bio-Rad ChemiDoc Imaging System were used.

*Docking with HADDOCK*. The CLOCK-PASB domain (residues 261–384 from 4f3l.pdb) was docked into the secondary pocket of mCRY1 (PDB ID 5T5X) as described previously (36) via HADDOCK 2.2 server (37, 38). Active residues (directly involved in the interaction) for CRY were: G106, R109, E383, and E382; and for CLOCK PAS-B were: G332, H360, Q361, W362, and E367. Passive residues were defined automatically around the active residues. Docking was performed with default parameters.

*Study approval*. For the Turkish and Italian cohorts, we obtained written informed consent from all of the study participants at the time of blood sampling. The institutional ethics committee of Bilkent University and Siena University approved the study. A code and family ID number deidentified each individual. The consent procedure allows recontact for the collection of individual-level phenotypic data, which are different from the primary reason for referral.

*Statistics*. The subjects were classified into the following behavioral categories: (a) affected hyperactive/impulsive; (b) affected inattentive; (c) affected combined; (d) ADHD spectrum (probably not affected); (e) ADHD spectrum (probably affected); (f) not affected; and (g) unknown/uninterpretable. They were also classified according to the severity of ADHD as severe, moderate, mild, or in partial remission (Supplemental Table 1). For the statistical analysis of the association between ADHD and *CRY1* allele status, the first, second, third, and fourth categories were combined as "affected," and the fifth and sixth categories were combined as "unaffected." Subjects deemed "unknown/uninterpretable" were excluded from the analysis. The statistical analyses were performed using SPSS, version 23 (GraphPad Prism, version 6.0e, GraphPad Software) and in-house Perl scripts. Normality of the data was assessed with a Shapiro-Wilk test. For categorical variables, a Fisher's exact test, OR, and 95% CI were calculated. Since the data were not normally distributed, a Mann-Whitney *U* test was used for comparison of the groups. A 2-tailed, unpaired *t* test was used for statistical evaluation of the CRY1 rescue and degradation assays. For further information, refer to Supplemental Table 24.

## Author contributions

TO conceived of the project. IHK designed the functional characterization studies of the mutants and WT human CRY1. OEO, MEK, SG, KB, YW, AO, CA, ANB, MAT, AG, CF, AR, JLC, YI, CEA, MCS, IHK, and TO collected and analyzed the data. MCS, CA, MAT, and AG, board-certified psychiatrists, conducted psychiatric evaluations of the study participants. For the families living outside of Ankara, CA, MCS, OEO, and TO traveled to the families' hometowns in Konya (DSPD-1), Urfa (DSPD-4), and Kayseri (DSPD-6)

to perform clinical evaluations. CA, MCS, MAT, and AG reviewed the ASRS questionnaires, and TO, OEO, MAT, and AG reviewed the MCTQs completed by each participant. TO, OEO, MAT, and AG conducted independent sleep behavior assessments. TO and IHK wrote the manuscript with input from all authors.

## Acknowledgments

We are indebted to İclal Büyükdevrim Özçelik, Nezahat Doğan, and Eyyüp-Emel Göllü for their dedication at every step of the project and for their insightful communications with the families. We gratefully acknowledge Jeffrey M. Friedman, Aziz Sancar, Michael W. Young, Alina Patke, and Turgay Dalkara for their invaluable comments on the manuscript. We thank Andrew C. Liu and Hiroki Ueda for the gift of the *Cry1*$^{-/-}$ and *Cry2*$^{-/-}$ MEFs and the *Cry1* rescue vector. The Turkish Academy of Sciences-TÜBA supported this work. WES of the replication cohort was performed at the Yale Center for Mendelian Genomics, funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute (NHLBI), NIH (UM1HG006504). The GSP Coordinating Center (U24 HG008956) contributed to the cross-program scientific initiatives and provided logistical and general study coordination. This work was funded in part by the National Center for Advancing Translational Sciences (NCAST); the NIH Clinical and Translational Science Award (CTSA) program (UL1TR001866); the NIH (R01AI088364, R01AI127564, R37AI095983, P01AI61093); the French National Research Agency (ANR) under the "Investments for the future" program (ANR-10-IAHU-01); the Integrative Biology of Emerging Infectious Diseases Laboratoire d'Excellence (ANR-10-LABX-62-IBEID); an IEIHSEER grant (ANR-14-CE14-0008-01); a SEAe-Host Factors grant (ANR-18-CE15-0020 02); a PNEUMOID project grant (ANR 14-CE15-0009-01); and by grants from the INCA/Cancéropole Ile-de-France (2013-1-PL BIO-11-INSERM 5-1); the Rockefeller University, INSERM; the HHMI, University of Paris; the St. Giles Foundation; and the Charles Bronfman Institute for Personalized Medicine at the Icahn School of Medicine at Mount Sinai. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Address correspondence to: Tayfun Özçelik, Bilkent University, Department of Molecular Biology and Genetics, Bilkent, Ankara 06800, Turkey. Phone: 90.312.266.4380; Email: tozcelik@bilkent.edu.tr.

1. Patke A, et al. Mutation of the human circadian clock gene CRY1 in familial delayed sleep phase disorder. *Cell*. 2017;169(2):203–215.e13.
2. Toh KL, et al. An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science*. 2001;291(5506):1040–1043.
3. Xu Y, et al. Functional consequences of a CKIdelta mutation causing familial advanced sleep phase syndrome. *Nature*. 2005;434(7033):640–644.
4. Hirano A, et al. A cryptochrome 2 mutation yields advanced sleep phase in humans. *Elife*. 2016;5:e16695.
5. Rahman SA, Kayumov L, Tchmoutina EA, Shapiro CM. Clinical efficacy of dim light melatonin onset testing in diagnosing delayed sleep phase syndrome. *Sleep Med*. 2009;10(5):549–555.
6. Thapar A, Cooper M. Attention deficit hyperactivity disorder. *Lancet*. 2016;387(10024):1240–1250.
7. Wulff K, Gatti S, Wettstein JG, Foster RG. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nat Rev Neurosci*. 2010;11(8):589–599.
8. Sehgal A, Mignot E. Genetics of sleep and sleep disorders. *Cell*. 2011;146(2):194–207.
9. Pagani L, et al. Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. *Proc Natl Acad Sci USA*. 2016;113(6):E754–E761.
10. Lane JM, et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat Genet*. 2017;49(2):274–281.
11. Brainstorm Consortium, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395):eaap8757.
12. Demontis D, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet*. 2019;51(1):63–75.
13. Sullivan PF, Geschwind DH. Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell*. 2019;177(1):162–183.
14. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010;141(2):210–217.
15. Schulze TG, McMahon FJ. Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping. *Hum Hered*. 2004;58(3-4):131–138.
16. Stessman HA, Bernier R, Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. *Cell*. 2014;156(5):872–877.
17. Dal GM, et al. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet*. 2014;51(7):455–459.
18. Özçelik T, Onat OE. Genomic landscape of the Greater Middle East. *Nat Genet*. 2016;48(9):978–979.
19. Ye R, Selby CP, Chiou YY, Ozkan-Dagliyan I, Gaddameedhi S, Sancar A. Dual modes of CLOCK:BMAL1 inhibition mediated by cryptochrome and period proteins in the mammalian circadian clock. *Genes Dev*. 2014;28(18):1989–1998.
20. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC, USA: American Psychiatric Association; 2012.
21. Roenneberg T, Wirz-Justice A, Merrow M. Life between clocks: daily temporal patterns of human chronotypes. *J Biol Rhythms*. 2003;18(1):80–90.
22. Adler LA, et al. Validity of pilot adult ADHD Self-Report Scale (ASRS) to rate adult ADHD symptoms. *Ann Clin Psychiatry*. 2006;18(3):145–148.
23. Bron TI, Bijlenga D, Verduijn J, Penninx BW, Beekman AT, Kooij JJ. Prevalence of ADHD symptoms across clinical stages of major depressive disorder. *J Affect Disord*. 2016;197:29–35.
24. Global Tobacco Surveillance System Data (GTSSData). Centers for Disease Control and Prevention Web Site. https://nccd.cdc.gov/GTSSDataSurveyResources/Ancillary/DataReports.aspx?CAID=1. Updated August 6, 2019. Accessed September 17, 2019.
25. Rybak YE, McNeely HE, Mackenzie BE, Jain UR, Levitan RD. An open trial of light therapy in adult attention-deficit/hyperactivity disorder. *J Clin Psychiatry*. 2006;67(10):1527–1535.
26. Fargason RE, et al. Correcting delayed circadian phase with bright light therapy predicts improvement in ADHD symptoms: A pilot study. *J Psychiatr Res*. 2017;91:105–110.
27. Arns M, van der Heijden KB, Arnold LE, Kenemans JL. Geographic variation in the prevalence of attention-deficit/hyperactivity disorder: the sunny perspective. *Biol Psychiatry*. 2013;74(8):585–590.
28. Insel TR. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *Am J Psychiatry*. 2014;171(4):395–397.
29. Hennekam RC, Biesecker LG. Next-generation sequencing demands next-generation phenotyping. *Hum Mutat*. 2012;33(5):884–886.
30. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5–23.
31. Karczewski KJ, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Posted on bioRxiv April 8, 2020. https://doi.org/10.1101/531210.
32. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–1073.
33. Lazaridis I, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*. 2016;536(7617):419–424.
34. Green M, et al. *Diagnosis of Attention-Deficit/Hyperactivity Disorder*. Rockville Maryland, USA: Agency for Health Care Policy and Research; 1999.
35. Qiu M, Ramulu PY, Boland MV. Association between sleep parameters and glaucoma in the United States population: National Health and

Nutrition Examination Survey. *J Glaucoma*. 2019;28(2):97–104.
36. Michael AK, et al. Formation of a repressive complex in the mammalian circadian clock is mediated by the secondary pocket of CRY1. *Proc Natl Acad Sci USA*. 2017;114(7):1560–1565.
37. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*. 2003;125(7):1731–1737.
38. van Zundert GCP, et al. The HADDOCK2.2 Web server: user-friendly integrative modeling of biomolecular complexes. *J Mol Biol*. 2016;428(4):720–725.
39. Ukai-Tadenuma M, Yamada RG, Xu H, Ripperger JA, Liu AC, Ueda HR. Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell*. 2011;144(2):268–281.
40. Siepka SM, et al. Circadian mutant overtime reveals F-box protein FBXL3 regulation of cryptochrome and period gene expression. *Cell*. 2007;129(5):1011–1023.
41. Gao P, et al. Phosphorylation of the cryptochrome 1 C-terminal tail regulates circadian period length. *J Biol Chem*. 2013;288(49):35277–35286.
42. Posner J, Polanczyk GV, Sonuga-Barke E. Attention-deficit hyperactivity disorder. *Lancet*. 2020;395(10222):450–462.
43. Sancar A. Mechanisms of DNA Repair by photolyase and excision nuclease (Nobel Lecture). *Angew Chem Int Ed Engl*. 2016;55(30):8502–8527.
44. Manolio TA, et al. Bedside back to bench: building bridges between basic and clinical genomic research. *Cell*. 2017;169(1):6–12.
45. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
46. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303.
47. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575.
48. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867–2873.
49. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):e190.
50. Cingolani P, et al. 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)*. 2012;6(2):80–92.
51. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
52. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*. 2011;27(22):3216–3217.
53. NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server. http://evs.gs.washington.edu/EVS. Updated April 23, 2019. Accessed May 4, 2020.
54. Dong C, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125–2137.
55. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310–315.
56. Jian X, Liu X. In silico prediction of deleteriousness for nonsynonymous and splice-altering single nucleotide variants in the human genome. *Methods Mol Biol*. 2017;1498:191–197.
57. Xiong HY, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. 2015;347(6218):1254806.
58. Ciarleglio CM, et al. Genetic differences in human circadian clock genes among worldwide populations. *J Biol Rhythms*. 2008;23(4):330–340.
59. Hawkins GA, Meyers DA, Bleecker ER, Pack AI. Identification of coding polymorphisms in human circadian rhythm genes PER1, PER2, PER3, CLOCK, ARNTL, CRY1, CRY2 and TIMELESS in a multi-ethnic screening panel. *DNA Seq*. 2008;19(1):44–49.
60. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011;9(2):179–181.
61. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31(21):3555–3557.
62. Reeve JP, Rannala B. DMLE+: Bayesian linkage disequilibrium gene mapping. *Bioinformatics*. 2002;18(6):894–895.
63. Borroni B, et al. Founder effect and estimation of the age of the Progranulin Thr272fs mutation in 14 Italian pedigrees with frontotemporal lobar degeneration. *Neurobiol Aging*. 2011;32(3):555.e1–555.e8.
64. Ottoni C, Ricaut FX, Vanderheyden N, Brucato N, Waelkens M, Decorte R. Mitochondrial analysis of a Byzantine population reveals the differential impact of multiple historical events in South Anatolia. *Eur J Hum Genet*. 2011;19(5):571–576.
65. Maddison A. *The World Economy: A Millennial Perspective*. Paris, France: Development Centre of the Organisation for Economic Co-operation and Development; 2001.