Received: 16 March 2019 | Revised: 16 July 2019 | Accepted: 22 July 2019

DOI: 10.1002/humu.23880

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

DATABASES

Iranome: A catalog of genomic variations in the Iranian population

Zohreh Fattahi[1,2] | Maryam Beheshtian[1,2] | Marzieh Mohseni[1,2] | Hossein Poustchi[3] | Erin Sellars[4] | Sayyed Hossein Nezhadi[5] | Amir Amini[6] | Sanaz Arzhangi[1] | Khadijeh Jalalvand[1] | Peyman Jamali[7] | Zahra Mohammadi[3] | Behzad Davarnia[1] | Pooneh Nikuei[8] | Morteza Oladnabi[1] | Akbar Mohammadzadeh[1] | Elham Zohrehvand[1] | Azim Nejatizadeh[8] | Mohammad Shekari[8] | Maryam Bagherzadeh[4] | Ehsan Shamsi-Gooshki[9,10] | Stefan Börno[11] | Bernd Timmermann[11] | Aliakbar Haghdoost[12,13] | Reza Najafipour[14] | Hamid Reza Khorram Khorshid[1] | Kimia Kahrizi[1] | Reza Malekzadeh[3] | Mohammad R. Akbari[4,15,16] | Hossein Najmabadi[1,2]

[1]Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

[2]Kariminejad–Najmabadi Pathology & Genetics Center, Tehran, Iran

[3]Digestive Diseases Research Center, Digestive Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran

[4]Women's College Research Institute, University of Toronto, Toronto, Ontario, Canada

[5]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

[6]Information Technology Office, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

[7]Shahrood Genetic Counseling Center, Welfare Office, Semnan, Iran

[8]Molecular Medicine Research Center, Hormozgan University of Medical Sciences, Bandar Abbas, Iran

[9]Medical Ethics and History of Medicine Research Center, Tehran University of Medical Sciences, Tehran, Iran

[10]Department of Medical Ethics, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran

[11]Max Planck Institute for Molecular Genetics, Berlin, Germany

[12]Modeling in Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

[13]Regional Knowledge Hub, and WHO Collaborating Centre for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

[14]Cellular and Molecular Research Centre, Qazvin University of Medical Sciences, Qazvin, Iran

[15]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[16]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada

Correspondence
Mohammad R. Akbari, Women's College Hospital, University of Toronto, 76 Grenville Street, Room 6421, M5S 1B2 Toronto, Ontario, Canada.
Email: mohammad.akbari@utoronto.ca

Hossein Najmabadi, Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Daneshjoo Blvd, Koodakyar St., Evin, Tehran 1985713834, Iran.
Email: hnajm12@yahoo.com

Abstract

Considering the application of human genome variation databases in precision medicine, population-specific genome projects are continuously being developed. However, the Middle Eastern population is underrepresented in current databases. Accordingly, we established Iranome database (www.iranome.com) by performing whole exome sequencing on 800 individuals from eight major Iranian ethnic groups representing the second largest population of Middle East. We identified 1,575,702 variants of which 308,311 were novel (19.6%). Also, by presenting higher frequency for 37,384 novel or known rare variants, Iranome database can improve the power of molecular diagnosis. Moreover, attainable clinical information makes this database a good resource for

Funding information
Vice deputy for research and technology at Iran Ministry of Health and Medical Education, Grant/Award Number: 700/150; Iran Vice-President office for Science and Technology, Grant/Award Number: 11/66100

classifying pathogenicity of rare variants. Principal components analysis indicated that, apart from Iranian-Baluchs, Iranian-Turkmen, and Iranian-Persian Gulf Islanders, who form their own clusters, rest of the population were genetically linked, forming a super-population. Furthermore, only 0.6% of novel variants showed counterparts in "Greater Middle East Variome Project", emphasizing the value of Iranome at national level by releasing a comprehensive catalog of Iranian genomic variations and also filling another gap in the catalog of human genome variations at international level. We introduce Iranome as a resource which may also be applicable in other countries located in neighboring regions historically called Greater Iran (Persia).

KEYWORDS
Genome project, genomic variation database, Iran, Iranome, whole exome sequencing

1 | INTRODUCTION

Completion of the Human Genome project (HGP) was a turning point in the field of human genetics, providing the first human reference DNA sequence. With the advent of next generation sequencing (NGS) technologies, some personal genomes were sequenced and numerous single nucleotide polymorphisms (SNPs), mostly with lower allele frequencies, were identified which were not present in the dbSNP database at that time (Naidoo, Pawitan, Soong, Cooper, & Ku, 2011). To obtain a more complete picture of the rare variants in different human populations, the 1,000 genome (1KG) project expanded and provided the largest catalog of human genetic variations applying whole exome sequencing (WES) and whole genome sequencing (WGS) on 2,504 individuals from 26 different populations (Auton et al., 2015; Naidoo et al., 2011). This project revealed over 88 million variants in the human genome providing a valuable resource for research on the genetic basis of human disorders. The majority of these variants were rare (allele frequency <0.5%) and the total number of variants was different among the 26 populations with the rare ones limited to closely related populations, known as geographical clustering of the rare variants (Auton et al., 2015; Tennessen et al., 2012).

Furthermore, WES of 6,515 European American and African American individuals in the NHLBI GO Exome Sequencing Project (ESP) was used to infer mutation ages and determined that about 86% of single nucleotide variants (SNVs) had been created recently. The results showed that the majority of these variants were rare SNVs arising as a result of population growth, and the recently emerged rare deleterious SNVs were significantly increased in disease genes (Fu et al., 2013). Such variants are of great importance in the field of genetic diagnosis of Mendelian disorders. Basically, the pipelines used to identify the causal variant from the extensive list of genetic variations detected by NGS methods, include filtering the variants based on high allele frequency. Therefore, the presence of large-scale databases of genetic variations with representatives from many different ethnic groups is essential to provide a more complete picture of human genome variations and their allele frequencies. This provides a good resource for clinical and functional interpretation of the variants, distinguishing real disease-causing variants from polymorphisms. In line with this demand, a few collaborative projects were established recently and aggregated a large number of WES and WGS data, providing a more comprehensive summary of human genome variations. These included the Exome Aggregation Consortium (ExAC), aggregating 60,706 exome sequences and the later Genome Aggregation Database (gnomAD), aggregating 125,748 exome sequences in addition to 15,708 whole-genome sequences of unrelated individuals from various ancestries (Lek et al., 2016).

Data analysis of such large-scale projects led to the following conclusion that rare variants are more likely to be population-specific. This shows the necessity to conduct population-specific genome projects to identify their genetic backgrounds. Such attempts will help in building a more complete picture of genetic variations in the human genome by introducing novel and population-specific rare variants. In fact, this was the aim of the 1KG project which selected samples from 26 different populations. However, the need for population-specific genome projects still remains, because many ethnic groups are not represented in 1KG project or the number of individuals per population is insufficient to attain reliable allele frequencies (An, 2017; Dopazo et al., 2016; Yamaguchi-Kabata et al., 2015). Accordingly, lack of representatives from specific populations and ethnic groups in human genome databases may lead to marginalization of members of those populations in the era of genomic revolution, which might put them in danger of discrimination by depriving them of the benefits of new advances in genetic technologies and the associated medical advances. Creation of population-specific genomic variation databases will play an important role in genomic medicine and healthcare as the interpretation of a causal variant in the clinical setting requires knowledge of its frequency in the population the patient comes from (An, 2017; Auton et al., 2015; MacArthur et al., 2014). Therefore, from an ethical point of view, population-specific genome projects improve health equity at a population and global level (Boomsma et al., 2014; The Genome of the Netherlands Consortium, 2014).

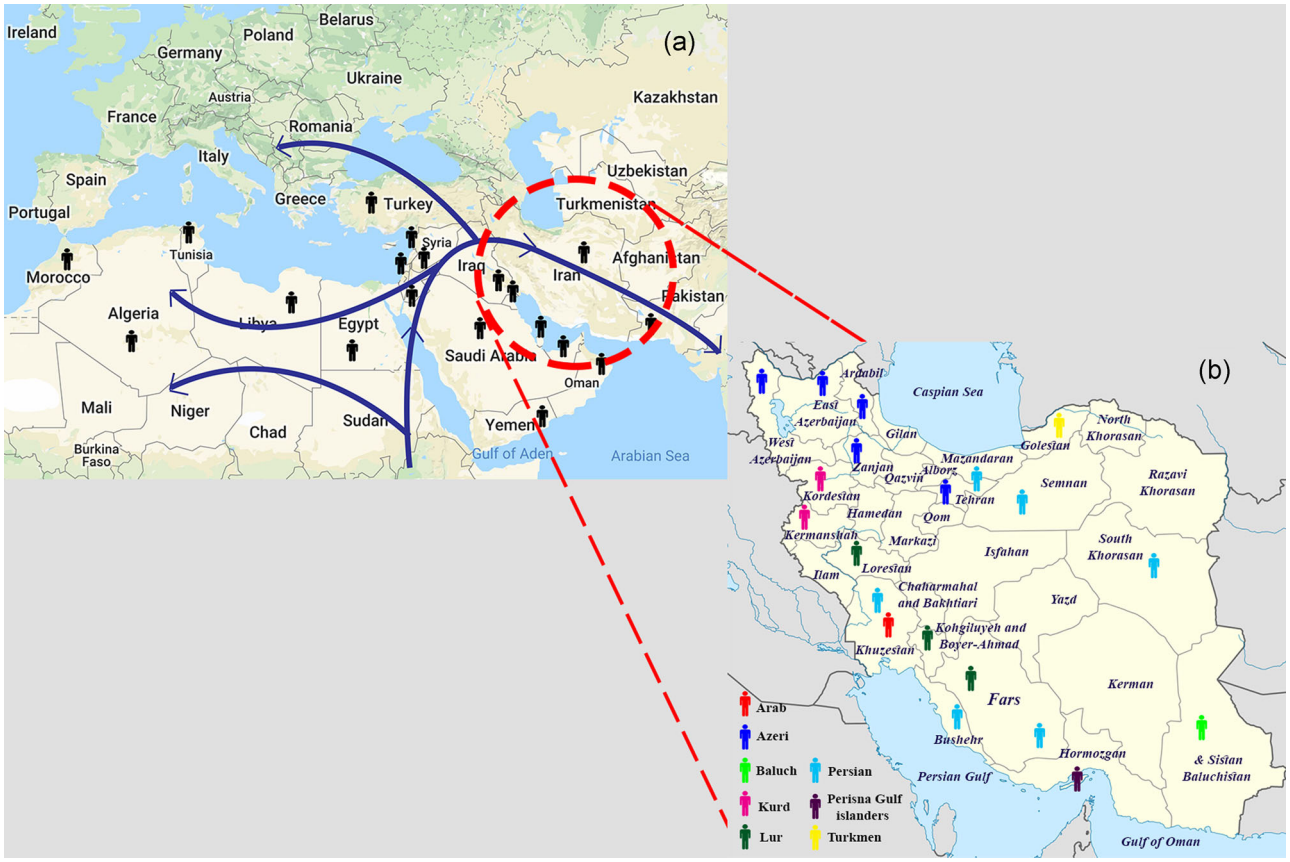

FIGURE 1 (a) Map of Iran, its neighboring countries and countries in the Middle East and North Africa. The blue arrows show the initial migrations out of Africa towards the Fertile Crescent (a region located in the Middle East which stretches from the Zagros Mountains in southwestern Iran to northern Mesopotamia and into southeast Anatolia) and then the migration of early humans from this region to Asia and Europe (Eurasia). In addition, the black man symbols show the countries and populations which were investigated in The Greater Middle East (GME) Variome Project. (b) Map of Iran with its provinces. The man symbols show where all 800 samples in the Iranome project were taken. The red, dark blue, light green, pink, dark green, light blue, purple and yellow man symbols are shown in provinces where samples were collected for Iranian–Arabs, Iranian–Azeris, Iranian–Baluchs, Iranian–Kurds, Iranian–Lurs, Iranian–Persians, Iranian–Persian Gulf Islanders, and Iranian–Turkmen, respectively

Middle Eastern genomes are completely absent from the most renowned human genome variation datasets (Auton et al., 2015; Lek et al., 2016). The Middle East; large region encompassing 17 countries from West Asia to North Africa, is renowned as "*Home to the cradle of civilization*" and also as an important gateway of modern human migratory routes out of Africa, thereafter populating the whole world (Henn, Cavalli-Sforza, & Feldman, 2012). Furthermore, its overall 411 million residents have come from diverse ethnic groups with rapid population growth (The Middle East Population (2018-05-20). Retrieved from http://worldpopulationreview.com/continents/the-middle-east-population/).

Iran, the second largest population in the Middle East, is geographically located in West Asia in a historical region known as the "Fertile Crescent" where the initial migrations out of Africa towards Asia and Europe (Eurasia) occurred (Map of Human Migration; Retrieved from https://genographic.nationalgeographic.com/human-journey/; Figure 1a; Alkan et al., 2014; Henn et al., 2012) (Figure 1a).

In addition, the important role and geographical location of Iran in the expansion and distribution of gene mutations during the Silk Road trade, an important period causing population admixture after the divergence of western and eastern Eurasia, cannot be overlooked (Comas et al., 1998; Derenko et al., 2013; Zarei & Alipanah, 2014). All these factors emphasize the valuable and significant additive information on human genome variation that can be gained by sequencing genomes from Middle Eastern population and especially from Iranian population, as is the main focus of this study. As a result of the tradition of consanguineous marriage, a high burden of recessive disorders is reported in the region, and attempts to clarify the genetic variants and their allele frequencies in the Middle Eastern population have recently been initiated aiming to improve precision medicine.

First, "The Greater Middle East (GME) Variome Project" included 2,497 WES data selected from the Greater Middle East (Figure 1a), the later "Al mena" project aggregated WES and WGS sequence data from 2,115 unrelated individuals from the Middle East and North Africa (MENA) region (Koshy, Ranawat, & Scaria, 2017; Scott et al., 2016). Although these large-scale aggregation projects have provided an excellent view of allelic frequencies in the Middle Eastern

population, the number of Iranian individuals included is insufficient to provide a detailed account of allelic frequencies in this specific population, particularly, because it is considered to be one of the ancient founder populations in the Middle East (Scott et al., 2016), and it is a multiethnic and multilinguistic community (Amanolahi, 2005). These Iranian ethnic groups are among the most under-represented populations in the human genomic variation databases currently available.

Here, we describe the design of "Iranome" as a population-specific project to address the aforementioned issues in medical genetics and precision medicine among Middle Easterns and in particular Iranian populations. With this in mind, we established the Iranome database (www.iranome.com and/or www.iranome.ir) by performing whole exome sequencing on 800 individuals from eight major ethnic groups in Iran, with 100 healthy individuals from each ethnic group. The eight ethnic groups were as follows: Iranian-Arabs, Iranian-Azeris, Iranian-Baluchs, Iranian-Kurds, Iranian-Lurs, Iranian-Persians, Iranian-Persian Gulf Islanders, and Iranian-Turkmen. They represent over 80 million Iranians and, to some degree, the half a billion individuals who live in the Middle East.

2 | MATERIALS AND METHODS

2.1 | Editorial policies and ethical considerations

The present study obtained approval from the Biomedical Research Ethics Committee, part of the National Institute for Medical Research Development (Certification number: IR.NIMAD.REC.1395.003). In addition, all of the volunteers were properly informed of the project's objective and signed consent forms approving the publication of their results anonymously in aggregation with others.

2.2 | Project design and selection of samples

The Iranome project was designed to reflect the demographic context of the country. Iran is geographically located in southwest Asia bordering the Caspian Sea, the Persian Gulf, and the Gulf of Oman (Figure 1b), with over 80 million residents, it is considered to be the 18th most populous country in the world comprising 1.07% of the world population. Throughout its history, Iran has been invaded by other countries on many occasions, each making their own contributions to the gene pool of the local population (Farhud et al., 1991). In addition, this country was a passageway between the Far East and the West as it lay on the Silk Road trade route and as a result, the Iranian people have interacted and intermarried with many different nations and races such as Greeks, Arabs, Mongols, Turks, and other tribes. So, the Iranian population is very heterogeneous and is composed of various ethnic groups (up to 26 in some reports) who live in geographically distinct regions of this vast country. Because the national census information is collected based on provinces and not on ethnicities, there are unofficial statistics about these ethnicities that are mostly categorized based on their language or religion and not race or biological factors (Amanolahi, 2005; Rashidvash, 2016). Although there are some discrepancies regarding the proportion of each ethnic group, it is well-known that Persians (Fars) are the dominant majority of the ethnic component in Iran, and Azeris or Azerbaijanis comprise the second largest ethnic group (largest ethnic minority) in the country (Material S1).

In some reports, the percentages of Iranian ethnicities are as follows: Persians (65%), Azeris (16%), Kurds (7%), Lurs (6%), Arabs (2%), Baluchs (2%), Turkmen (1%), Qashqai (1%), and non-Persian, non-Turkic groups such as Armenians, Assyrians, and Georgians (less than 1%). However, other reports consider Persians as the dominant population (51%), followed by the rest of the population consisting of Azeris (24%), Gilakis and Mazandaranis (8%), Kurds (7%), Arabs (3%), Lurs (2%), Baluchs (2%), Turkmen (2%), Zabolies (2%) and others (1%) (Banihashemi, 2009; Hassan, 2008; Majbouri & Fesharaki, 2017).

There are some inconsistencies in considering Gilaki and Mazandarani people who live on the Caspian Sea coast as a separate ethnicity. Some reports consider these people as originally Persian and that any differences present between Mazandarani and Gilaki people are not due to race, but to environmental differences. However, they speak Persian dialects which are distinctive compared to Persian speakers from the central plateau such as those in Tehran or Shiraz (Curtis, Hooglund, & Division, 2008; Rashidvash, 2012, 2013).

The geographical separation and different historical origins of these ethnic groups play a significant role not only in building specific languages, cultures and life styles but possibly also has a bearing on their varied genetic background. However, some reports claim that common genetic roots of the ethnicities present in Iran are not changed significantly by encounters with other races throughout history (Farjadian & Safi, 2013). As the objective of the Iranome project was to develop a population-specific framework of genomic variations, providing a good resource for classifying these differences and producing reasonable national health care plans, we decided to include 100 individuals from each of the following ethnic groups: Arabs, Azeris, Baluchs, Kurds, Lurs, Persians, Turkmen, and also Persian Gulf Islanders (Figure 1b).

Samples were obtained through a network of local physicians in different provinces who were trained to collect samples according to the project's criteria shown in Table 1 (See questionnaire as material S2). All participants, whose ancestors were born in Iran, were registered in the project anonymously upon having pure race up to at least two generations (four grandparents) and after clinical evaluation according to the completed questionnaire and the complete blood count (CBC) and urine test results. The familial relationship was also explored as far as possible to prevent including relatives who had similar genetic backgrounds in the study.

2.3 | Sequencing and data analysis

All 800 Iranian DNA samples underwent exome enrichment using Agilent SureSelectXT Human All Exon V6 (Agilent Technologies Inc, Santa Clara, CA) to capture 60 Mb of human genome and then paired-end sequencing was performed using different Illumina

TABLE 1 Detailed information examined for each participant in the study

Individual information	Criteria
Age	Inclusion criteria:>30 years
Ethnicity	Inclusion criteria: Pure race up to at least two generations for each ethnicity
Sex	~1:1 ratio in the final 100 samples from each ethnicity
Weight	Recorded (all included)
Height	Recorded (all included)
Blood pressure	Recorded as normal, hypertension, not-known (all included)
Smoking	Recorded as smoker, nonsmoker (all included)
Record of medication use	Recorded with drug information (all included)
Disease history (cardiovascular, nephrological, urological, nervous system, gastrointestinal, endocrine and metabolic, blood, connective tissue, cancer)	Recorded (If any); the known rare Mendelian disorders, cancer, seizure, and epilepsy were excluded, multifactorial phenotypes were not excluded
Record of physical disability	Excluded: individuals with apparent physical disability
Disease history in parents and relatives	Recorded (all included)
Relative relationship in parents	Recorded (all included)
Twin pregnancy	Recorded
CBC and urine test results	Excluded: Hemoglobin < 10, glucose in urine

Abbreviation: CBC, cell blood count.

sequencers (Illumina, San Diego, CA). The generated paired-end reads of 100–150 bp were then aligned to *Homo sapiens* (human) genome assembly GRCh37(hg19)-1KG-decoy using the Burrows-Wheeler Aligner (BWA; V0.7.5a) after proper quality control assessment using the FastQC toolkit (Li & Durbin, 2010). BAM processing was implemented by applying Picard tools (V2.2.1) and then GATK pipeline (V3.7) adhering to best practices (Van der Auwera et al., 2013), which included marking and filtering duplicate reads, filtering low quality reads, insertion/deletion realignment and base quality recalibration. The alignment metrics were assessed using Picard tools to perform quality control of the BAM files followed by coverage assessment using GATK pipeline. Variant calling of the WES samples was performed by joint genotyping followed by variant calling using the haplotypecaller module of GATK pipeline. The final variant recalibration and filtering were accomplished by GATK Variant Quality Score Recalibration (VQSR). The variants identified were then annotated using the last updated versions of various databases and tools using ANNOVAR package and SNP and Variation Suite (SVS; Wang, Li, & Hakonarson, 2010). Statistics for all of the samples were acquired by SVS, RTG tools, VCFtools, and awk programing (Danecek et al., 2011).

2.4 | Database design

All the variants identified were made publicly available to the scientific community through a web-based genomic variation browser at http://www.iranome.com/ and http://iranome.ir/. The Iranome Browser uses the open source code developed initially for the ExAC browser by the laboratory of Dr Daniel G. MacArthur at Broad Institute of MIT and Harvard Universities, Cambridge, MA, with some modifications made to it by Golden Helix Inc. (Bozeman, MT).

2.5 | Analysis of the population genetic structure

The Eigenstrat method was used for analysis of the population genetic structure among the samples. In this method, principal components analysis (PCA) is applied to SNVs to infer continuous axes of genetic variation. To avoid the clustering of individuals based on regions of linkage disequilibrium (LD), SNPs in two known high LD regions (the human leukocyte antigen (HLA) region in chromosome 6 and a polymorphic chromosome 8 inversion) and dependent SNPs with $r^2 \geq 0.2$ over a shifting window of 500 kb were excluded and the remaining genetic variants were used for PCA. The eigenvectors of the first two PCs with the largest eigenvalues of the individuals for each population were plotted to visualize the genetic structure of different ethnic groups in comparison with each other. Then to compare the genetic structure of the Iranian population with other populations, the common variants between the Iranome database and the 26 populations of the 1KG database were pooled together and PCA was applied to the pooled database.

2.6 | Runs of homozygosity identification

The Golden Helix SVS algorithm (version 8.8.3) was used for the Runs of homozygosity (ROH) estimations using the WES data for all 800 individuals. First, stringent filtering was applied. Variants which did not pass the following quality criteria were removed: VQSR filtering,

number of supporting reads (≥10 reads), genotype quality (≥40), and alternate allele read ratio (≤0.15 for Ref_Ref variants, >0.3 and <0.7 for Alt_Ref variants and >0.85 for Alt_Alt variants). Then variants located on chromosome X and Y were excluded. The filtered list of variants was then used to assess all of the possible runs per sample based on the following criteria specified to the algorithm: The minimum run length was taken as 500 kb with a minimum 25 variants per run. Also, one heterozygous and five missing calls per run were allowed. Next, the second algorithm provided a clustered list of ROHs for at least 20 samples in the data set. Finally, the highly overlapping ROH regions were calculated as optimal ROH clusters.

3 | RESULTS

3.1 | Demography of Iranome project samples

The total number of samples included in the project was 800 individuals who were not suffering from severe rare Mendelian disorders. Table 2 provides a summary of the demographic information for the Iranome samples. The approximate 1:1 ratio of female/male was reflected in the overall list of samples as well as in each ethnic group. To decrease the bias made by late-onset Mendelian disorders, samples were selected from individuals who were >30 years old. The mean age of individuals at blood draw was 50.61 (standard deviation [*SD*] 9.33 years). The mean age of female individuals was 50.65 (*SD* 9.08 years) and the mean age of male individuals was 50.59 (*SD* 9.56 years). The age range of individuals included in the project was 30–84 years. Also, as was mentioned before, representative sampling was according to the ethnicities and not the geographical regions. The provinces in which samples from each ethnicity were selected are shown in Table 2 and also in Figure 1b.

3.2 | Genomic structure of the Iranian population

The genetic structure and ancestry among the seven ethnicities and also the Persian Gulf Islanders across Iran were estimated using principal components analysis (PCA) and population clusters are shown in Figure 2. The population clusters of Arab, Azeri, Kurd, Lur, and Persian ethnicities look genetically very similar to each other (Figure 2a) which is more distinctive in Arabs and Azeris (Figure 2b). The other three populations including Baluchs, Turkmen, and Persian Gulf Islanders are genetically more distinct from the other five, which may be explained by the separation of these groups from the rest of the population through geographical and cultural isolation (Figure 2a). Comparison of the Iranian population to the five super populations in the 1KG project (African, American, East Asian, European and South Asian) showed that the population clusters of Arabs, Azeris, Kurds, Lurs, and Persians are genetically distinct and these should probably be considered to be the sixth super population (main Iranian cluster) with its own genetic background distinct from the other five already known super populations (Figure 2c). Interestingly, this main Iranian cluster is located between Europeans and South Asians, predictable from their geographical locations. In comparison to the five super populations of the 1KG project, the Baluchs and Persian Gulf Islanders are located genetically between the main Iranian cluster and South Asians. In addition, Turkmen are

TABLE 2 Demographic information of Iranome samples

	Number (n) and age (mean ± SD) (years)			
Ethnicity	**Female**	**Male**	**Total**	**Provinces in which samples were taken**
Arab	44	56	100	Khuzestan
	(50.05 ± 9.14)	(46.79 ± 8)	(48.22 ± 8.63)	
Azeri	44	56	100	Eastern Azerbaijan, Western Azerbaijan, Ardebil, Tehran, Zanjan
	(49.52 ± 9.27)	(50.48 ± 7.93)	(50.06 ± 8.51)	
Baluch	40	60	100	Sistan & Baluchistan
	(46.87 ± 11.73)	(47.55 ± 11.54)	(47.28 ± 11.56)	
Kurd	48	52	100	Kurdistan, Kermanshah
	(48.94 ± 7.13)	(49.77 ± 6.37)	(49.37 ± 6.72)	
Lur	59	41	100	Lorestan, Fars, Kohgiluyeh and Boyer-Ahmad
	(50.77 ± 8.69)	(53 ± 11.8)	(51.69 ± 10.08)	
Persian	50	50	100	Fars, Semnan, Tehran, Bushehr, Khuzestan, Razavi-Khorasan
	(52.24 ± 9.09)	(54.08 ± 9.67)	(53.16 ± 9.39)	
Persian Gulf Islanders	50	50	100	Hormozgan
	(52.38 ± 7.73)	(51.98 ± 8.17)	(52.18 ± 7.92)	
Turkmen	48	52	100	Golestan
	(53.44 ± 8.84)	(55.2 ± 10.01)	(52.96 ± 9.43)	
Total	383	417	800	Iran
	(50.65 ± 9.08)	(50.59 ± 9.56)	(50.61 ± 9.33)	

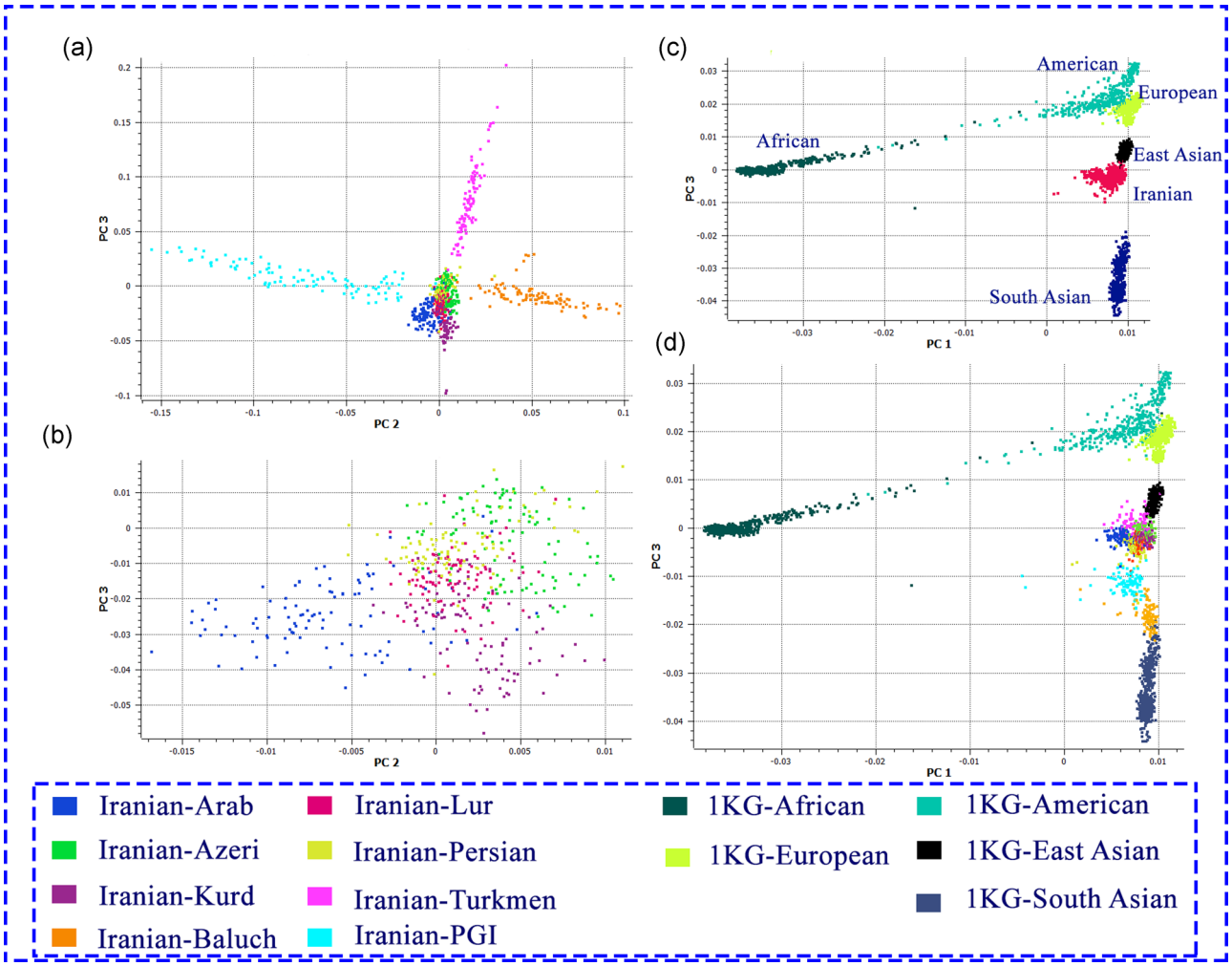

FIGURE 2 Results of principal components analysis (PCA) performed on the eight groups studied in Iranome project. Each color shows an ethnicity cluster as defined in the figure. (a) PCA of seven Iranian ethnicities and also Persian Gulf Islanders (PGI) show a common overlapping cluster for Arabs, Azeris, Kurds, Lurs and Persians, while Baluchs, Turkmen and Persian Gulf Islanders occur in separate clusters (PC2 and PC3 are shown). (b) PCA results for five Iranian ethnicities: Arabs, Azeris, Kurds, Lurs and Persians, showing similar genetic background, although Arabs and Azeris are more distinctive from the other three ethnic groups (PC2 and PC3 are shown). (c) Comparison of the Iranian population (shown as a super population including Arabs, Azeris, Kurds, Lurs and Persians) with the five super populations of the 1KG project (PC1 and PC3 are shown). (d) Comparison of the Iranian population (seven Iranian ethnicities and also Persian Gulf Islanders (PGI)) with the five super populations of 1KG project (PC1 and PC3 are shown)

located between the main Iranian cluster and East Asians (Figure 2d). Additional PCs for each of the abovementioned analyses are shown in the Figures S1 and S2.

3.3 | The iranome data set

For the 800 samples sequenced in this project, the mean depth of coverage for the exons of the human genome based on CCDS Release15 was 84X with 97% and 93% coverage at 10X and 20X or more, respectively.

In total, we identified 1,575,702 variants within protein coding regions captured by the SureSelect Human All Exon V6 kit, which passed a filter based on the following quality metrics: VQSR filtering (Using 99.0 tranche), depth of coverage (>4 reads for Ref_Ref and Alt_Alt variants, >8 reads for Ref_Alt variants), genotype quality (≥15), alternate allele read ratio (≤0.15 for Ref_Ref variants, >0.25 and <0.7 for Alt_Ref variants and > 0.8 for Alt_Alt variants) and Strand bias estimated using Fisher's Exact Test (FisherStrand <30). These high quality variants included 1,332,298 SNPs and 243,404 insertions/deletions (indels) and represent one variant in every approximately 38 bp of the captured 60 Mbp exome interval.

Among these 1,575,702 variants, 52.5% were singletons and 308,311 variants (including 240,256 SNPs and 68,055 indels) had no record in the following public databases: dbSNP catalog (version 149), dbSNP Common catalog (version 151), ExAC database, gnomAD database, NHLBI ESP6500 database, 1 kG Phase3, the Avon Longitudinal Study of Parents and Children (ALSPAC) data set, UK10K Twins data set and TOPMed data set; therefore, they were considered to be novel variants representing 19.6% of the entire detected variants (Table 3). As expected, the

majority of these novel variants were singletons (81%). On the other hand, an additional 50% (793,806) of the detected variants were observed in public databases but with an allele frequency less than 0.01 (rare variants). Therefore, about 70% of the variants identified in this data set belong to the category of rare/novel variants (Table 3). However, among these 1,102,117 rare/novel variants, we identified 37,384 variants (3.4%) with an alternate allele frequency of greater than 1% in the Iranome data set (Figure 3b). Therefore, in addition to introducing 308,311 novel variants to the catalog of human genome variation, the Iranome database can improve the power of molecular diagnosis by showing an alternative allele frequency of higher than 1% for 37,384 novel or previously known rare variants.

The average number of genomic variations per Iranian individual within regions covered by the SureSelect Human All Exon V6 kit was 92,162. This is also shown in Table 4 for all eight ethnicities investigated. In addition, the average number of singletons, transitions, and transversions per individual in the Iranome database was 796.67, 55481.25, and 23340, respectively. This represents the mean Ti/Tv ratio of 2.38 in the data set.

The total number of genetic variations did not differ significantly among the Iranian ethnic groups studied. Also, all ethnicities had similar proportions of the total number of novel genetic variations detected in the Iranian population with the highest detected in Azeri's and the lowest in Turkmen (Figure 3a,c). In addition, Baluchs and Turkmen had the lowest percentage of ethnic-specific novel variants among the total detected novel variants in the Iranian population whereas other ethnic groups showed nonsignificant differences. The detailed statistics for each ethnic group is shown in Table 4. This indicates that, although notable differences can be observed among these eight ethnicities in terms of their appearance, language and geographical location, they are genetically similar and contribute equally to the genetic pool of the Iranian population. So, we propose that the application of this current database can be

TABLE 3 Proportion of variants identified in Iranome project based on the alternate allele frequency and variant types

			Alternate allele frequency (AF)		
		Total	Novel	Rare (AF < 0.01)	Common (AF ≥ 0.01)
	All variants	1,575,702	308,311 (19.57%)	793,806 (50.38%)	473,585 (30.05%)
Clinically relevant effect					
Annotated by Refseq Genes 105 Interim v1, NCBI		1,440,997	280,691	733,334	426,972
LoF Effect	Exon_loss_variant	4	2	0	2
	Stop_lost	316	82	173	61
	Stop_gained	4805	1333	2949	523
	Initiator_codon_variant	703	178	388	137
	Frameshift_variant	11,065	6332	3586	1147
	Splice_acceptor_variant	1773	517	989	267
	Splice_donor_variant	1583	507	883	193
Total		20,249 (1.4%)	8,951	8968	2330
Missense Effect	5'-UTR_premature_start_codon_gain_ variant	1926	318	1205	403
	Disruptive_inframe_deletion	189	79	83	27
	Disruptive_inframe_insertion	60	18	34	8
	Missense	238,922	39,443	152,877	46,602
	Inframe_deletion	5871	1139	3137	1595
	Inframe_insertion	3103	581	1628	894
Total		250,071 (17.35%)	41,578	158,964	49,529
Other Effects	5′_UTR_variants	39,225	8452	19,904	10,869
	3′_UTR_variants	56,794	11,431	27,864	17,499
	Intron_variant	878,672	187,215	402,483	288,974
	Synonymous_variant	161,893	18,062	97,670	46,161
	Splice_region_variant	33,944	4,967	17,410	11,567
	Stop_retained_variant	149	35	71	43
Total		1,170,677 (81.24%)	230,162	565,402	375,113

Abbreviation: LoF, loss of function.

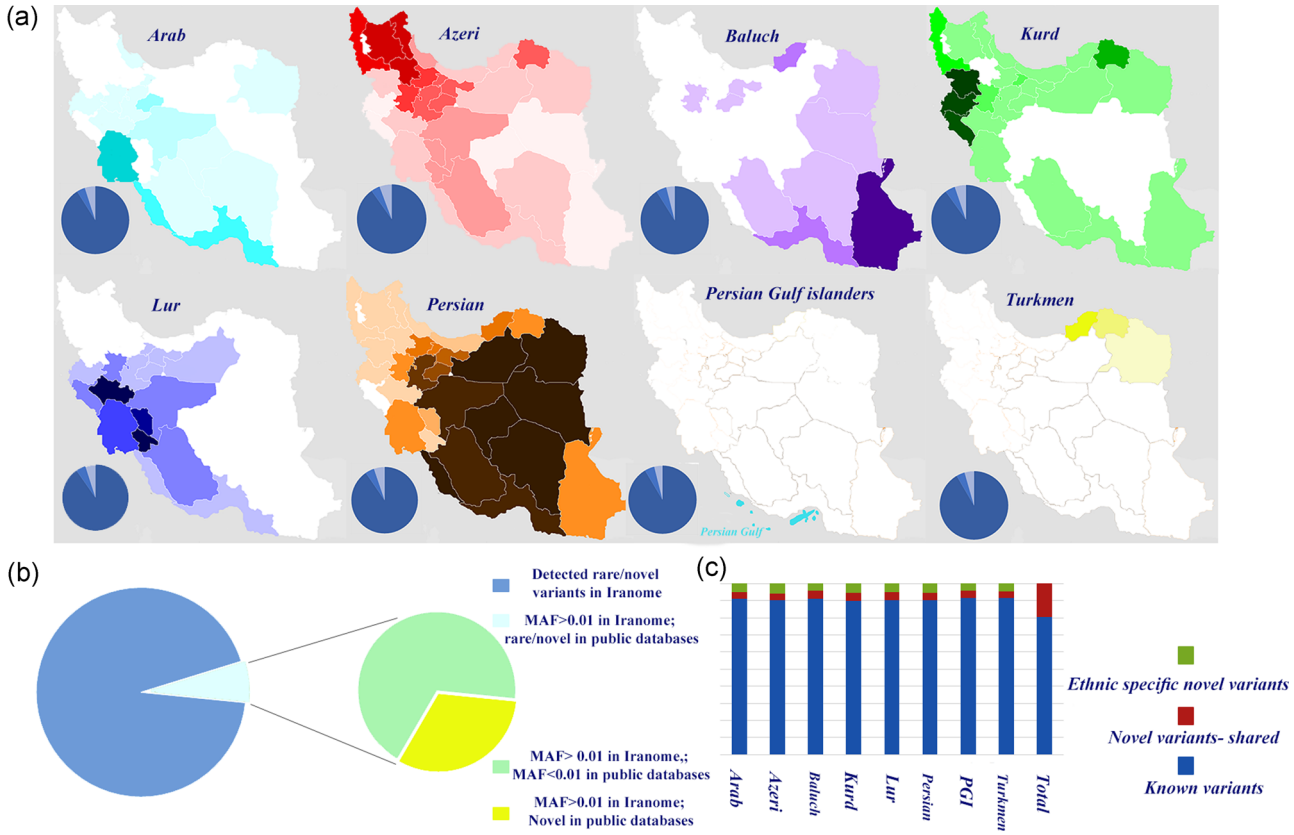

FIGURE 3 (a) Relative geographical distributions of individuals from each ethnicity are shown by colors (relative intensity) in each map (The Iran maps are created according to a poll in 2010 and are retrieved from Wikimedia Commons; the free media repository upon free permission to copy, distribute and/or modify under the terms of the GNU Free Documentation License). The area of each pi chart under each map represents the total number of variants identified in each ethnicity and Persian Gulf Islanders. Each pi chart is divided into three slices, showing known variants in each ethnicity (darkest blue), common novel variants in each ethnicity (medium blue) and ethnic-specific novel variants (light blue). (b) Portion of frequent variants (MAF > 0.01) in Iranome data set out of the rare/novel variants compared to public databases. (c) Proportion of known and novel variants per each ethnicity

TABLE 4 Proportion of variants identified in the Iranome project based on ethnicities, in addition to average number of genomic variations in each individual from eight ethnic groups

Ethnic group	No. of individuals	All variants detected in each ethnic group	Average no. of variants per individual	Novel variants detected	% of novel variants detected in each ethnic group / total number of variants detected in each ethnic group	Ethnic-specific novel variants	% of ethnic specific novel variants / total number of novel variants in Iranome
Arab	100	578,143	94,066	51,842	8.97%	28,808	9.34%
Azeri	100	526,479	91,486	51,794	9.84%	29,880	9.69%
Baluch	100	526,858	105,071	47,049	8.93%	20,961	6.80%
Kurd	100	502,970	94,135	50,057	9.95%	26,399	8.56%
Lur	100	482,663	87,215	47,056	9.75%	23,861	7.74%
Persian	100	490,559	85,060	47,017	9.58%	25,611	8.31%
Persian Gulf Islanders	100	564,481	103,916	48,281	8.55%	22,020	7.14%
Turkmen	100	462,059	76,351	38,891	8.42%	20,509	6.65%
All ethnicities	**800**	**1,575,702**	**92,162**	**308,311**	–	–	–

extended to the other ethnic minorities and the Iranian population in general.

3.4 | Functional annotation of variants in the iranome data set

The genomic variations detected in the Iranome database were then annotated by Refseq Genes 105 Interim v1, NCBI, which led to the identification of 1,440,997 variants annotated on verified mRNA transcripts of which 426,972 (29.63%) variants were frequently observed in public genomic databases with a frequency of greater than and equal to 0.01, and 733,334 (50.89%) variants were rarely observed with a frequency of less than 0.01, plus, an additional 280,691 (19.47%) novel variants (Table 3; Figure 4a).

Based on functional variant effects, all of these variants were categorized into three main groups of variants with Loss-of-function (LoF) effect constituting 1.4% of the total variants, variants with Missense effect and variants with Other effects constituting 17.35% and 81.24%, respectively (Figure 4a). These three groups were subcategorized into 19 different sequence ontology terms as described in Table 3.

In group 1 (LoF effects), 54.64% of the variants were Frameshifts whereas in group 2 (missense effects), 95.54% were missense variants. The most frequent variants in group 3 (other effects) were intronic variants (75.06%).

As it is shown in Table 3, variants with LoF effect were more prevalent among the rare/novel categories (Figure 4a,b). In total, LoF variants constituted 1.4% (20,249 variants) of the database of which 21% were located at Online Mendelian Inheritance in Man (OMIM) genes with a reported associated phenotype. These LoF variants were located in 8365 unique genes of which 75.5% were registered in the OMIM database. In addition, while the number of LoF variants was significantly decreased in common variants category, the percentage of LoF genes located at OMIM genes with a reported associated phenotype did not differ significantly among the three categories of novel, rare and common variants (Figure 4c).

3.5 | Genomic structure of the Iran data set compared to the Greater Middle East Variome dataset

Among the final list of 308,311 novel variants identified in the Iranian population through this project, we aimed to clarify the degree of population specificity by comparing these variants with the Greater Middle East (GME) Variome data set which includes samples

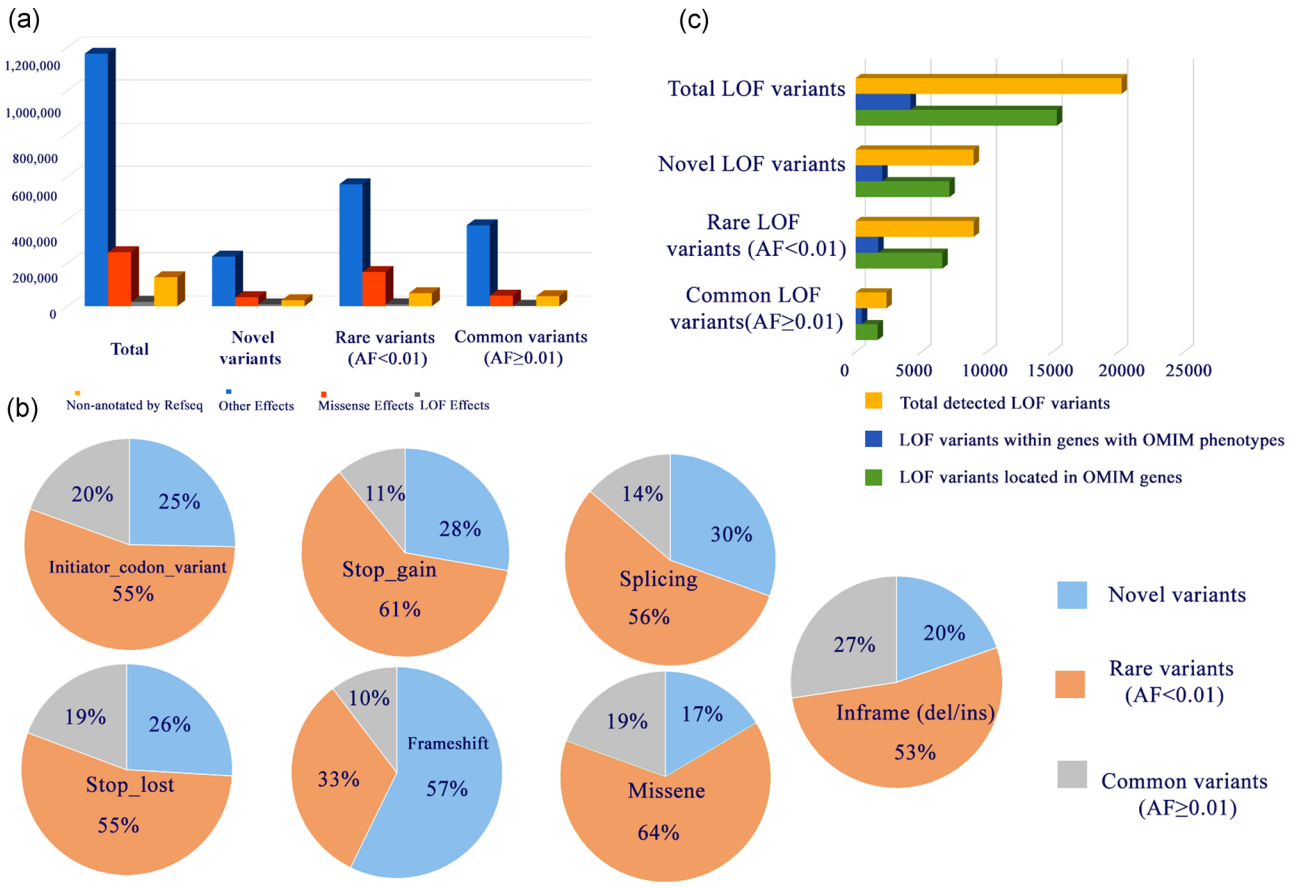

FIGURE 4 (a) Proportion of all three main groups of variants with Loss-of-function (LoF) effect, missense effect, and other effects is shown and categorized based on their allele frequency. (b) Pi charts show the proportion of some types of functional variants based on their allele frequency. (c) Distribution of identified LoF variants based on their frequency in public databases, their location on OMIM genes and the genes with OMIM phenotypes

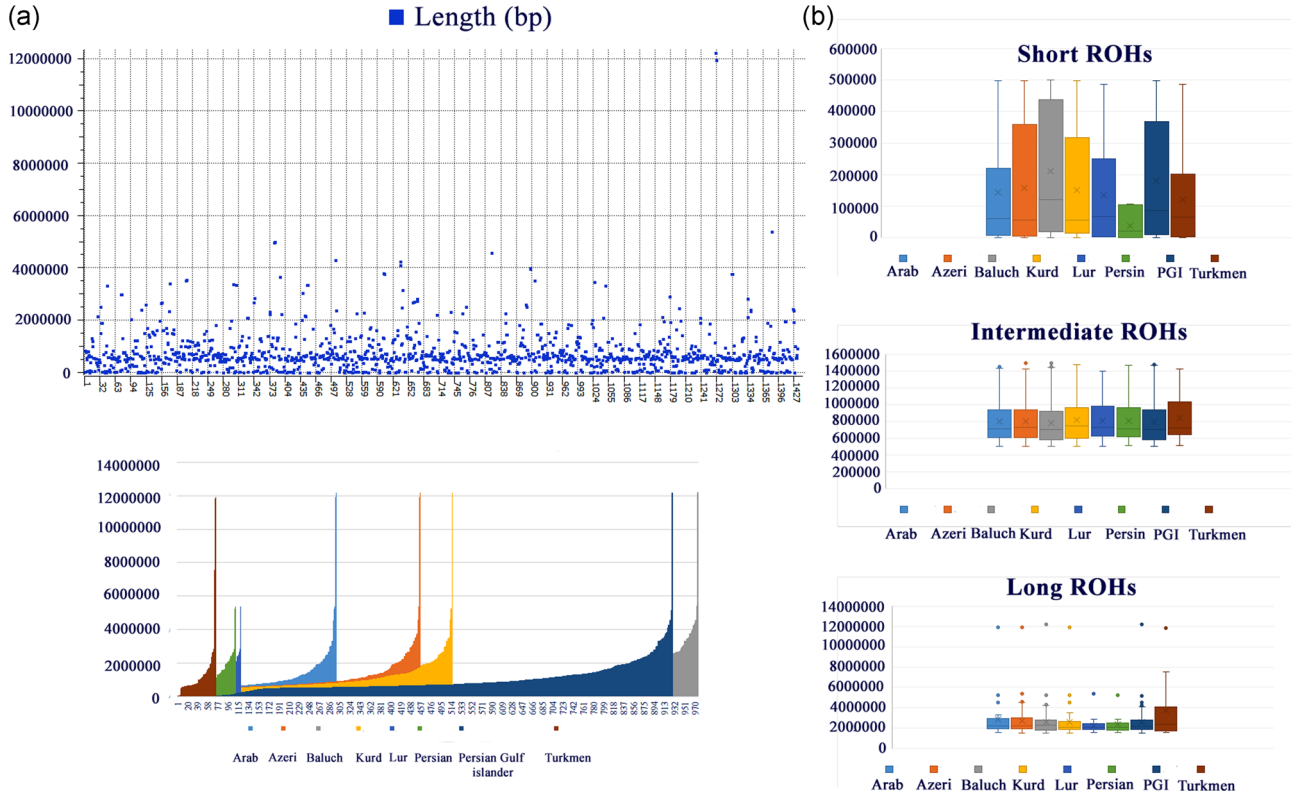

FIGURE 5 (a) Distribution of ROHs based on their length in all Iranians as well as each ethnicity. (b) Distribution of long, intermediate and short ROHs among the eight ethnicities investigated in this database. ROH, runs of homozygosity

from populations in closer geographical regions in the Middle East, compared to the other public databases.

Contrary to expectations, we only found about 1896 (0.6%) of the novel variants having overlap with the GME data set (including 601 variants with allele frequency greater than 0.01 and 1295 variants with allele frequencies < 0.01). So, apparently most of the novel variations identified in the Iranome project are not represented outside the corresponding population, even in closely located geographical regions. In fact, this is another indication of the importance of such ethnic-specific databases in the clinical setting and in molecular diagnosis.

3.6 | Known pathogenic/likely pathogenic variants observed in the Iranome data set

The contribution of known pathogenic variants in the Iranome database was assessed using the latest update of the ClinVar database (last updated on 2019-01-01). In total, 50,030 variants from the Iranome database were classified in ClinVar of which 721 were reported to be pathogenic and/or likely pathogenic. We assessed the allele frequency of these variants in the Iranian population and 668 of them (92.6%) were rare variants with alternative allele frequency of less than and equal to 1%, and only 53 variants had an alternate allele frequency of greater than 1%.

Further in-depth investigation of these frequent pathogenic variants as well as the rare ones appearing as homozygous (for rare recessive disorders) or heterozygous (for rare dominant disorders), along with exclusion of the variants known to be pathogenic for the phenotypes with a chance to be present in the Iranome database (refer to Table 1), led to the identification of 12 reported pathogenic/likely pathogenic variants in the ClinVar database that suggested cause in rare Mendelian disorders whereas they were seen in apparently normal individuals of the Iranome database. The additional information provided by the Iranome data set resulted in reclassification of these variants into variants of uncertain according to the *American College of Medical Genetics and Genomics (ACMG) Guideline* (Table S1; Richards et al., 2015).

3.6.1 | Iranome adds uncertainty to the pathogenicity of four variants observed with rare allele frequency in other similar public databases

The rare missense variant, p.Val1081Met, in the *KDM5C* gene is recognized as a pathogenic variant in the ClinVar database (with no supportive evidence) for mental retardation, X-linked, syndromic, Claes–Jensen type, XLR (MIM# 300534). This variant was detected as hemizygous in one of the Persian individuals in the Iranome database. Additional clinical follow-up revealed that this individual was apparently normal and unlikely to be related to this phenotype. Therefore, according to ACMG guidelines, this variant should be reclassified as a variant of uncertain significance based

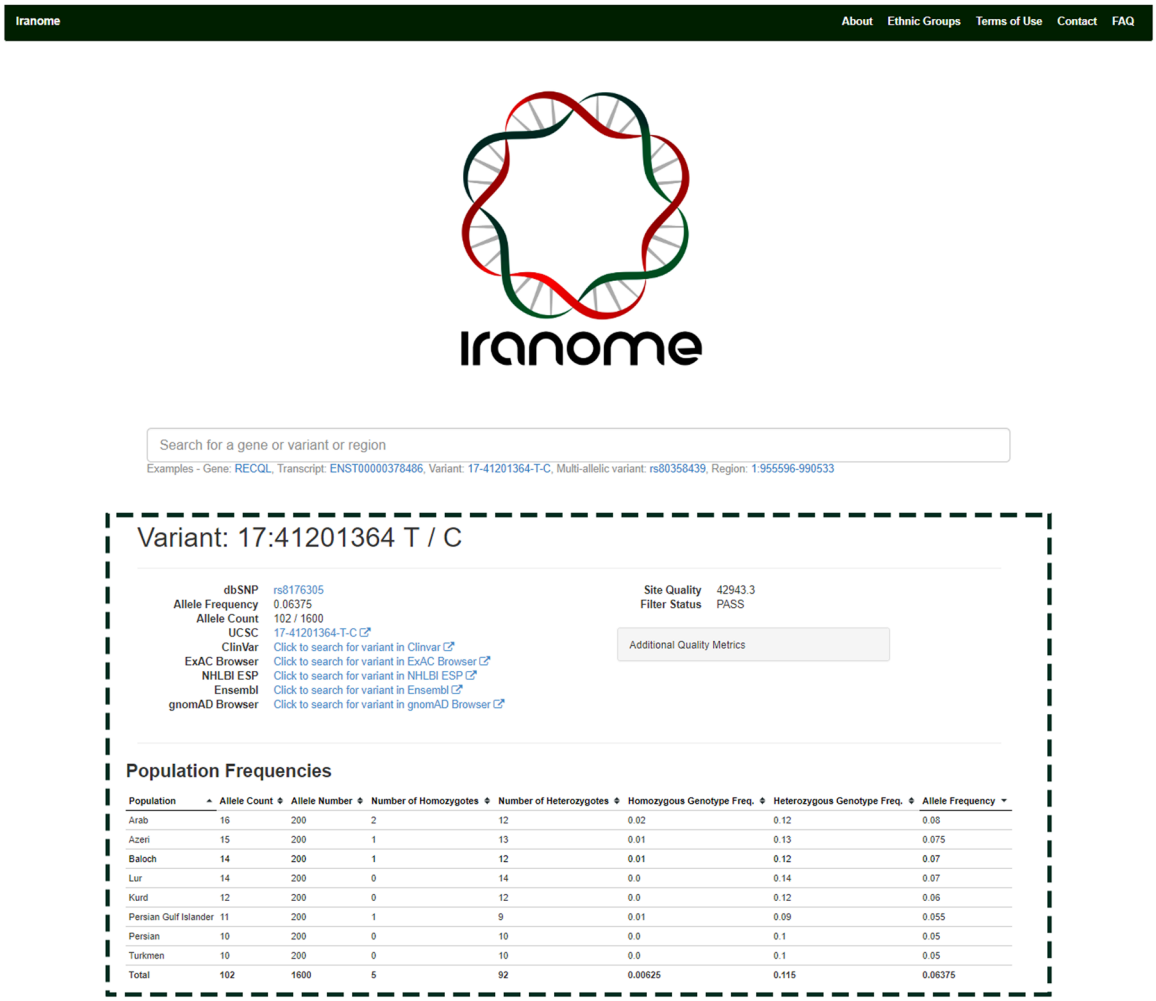

FIGURE 6 Web-based interface of the Iranome database

on the observation of a healthy adult individual in the Iranome database.

The rare missense variant, p.Ala338Val, in the *ALDOB* gene is recognized as a pathogenic/likely pathogenic variant in the ClinVar database for Fructose intolerance, hereditary, AR (MIM# 229600; Davit-Spraul et al., 2008; Esposito et al., 2010). This variant was identified as homozygous in one of the Baluch individuals in the Iranome database. Additional clinical follow-up revealed that this individual was apparently normal and unlikely to be related to this phenotype and therefore, this variant should be reclassified as a variant of uncertain significance based on the observation of a healthy adult individual in the Iranome database.

The rare heterozygous missense variant, p.Arg848Gln, in the *KIF1A* gene is recognized as likely pathogenic in the ClinVar database (with no supportive evidence) for Mental retardation, autosomal dominant 9 (MIM# 614255). Two heterozygous individuals (both Persian) were identified in the Iranome database. These individuals were apparently normal (unlikely to be related to mental retardation, autosomal dominant 9) and therefore, this variant should be reclassified as a variant of uncertain significance based on the observation of two healthy adult individuals in the Iranome database.

The rare heterozygous stop-gain variant, p.Tyr206Ter, in the *SCN8A* gene is recognized as likely pathogenic in the ClinVar database (with no supportive evidence) for Cognitive impairment with or without cerebellar ataxia, AD (MIM# 614306). Four heterozygous Lur individuals were identified in the Iranome database. These individuals were apparently normal (unlikely to be related to Cognitive impairment with or without cerebellar ataxia) and therefore, this variant should be reclassified as a variant of uncertain significance based on the observation of four healthy adult individuals in the Iranome database.

3.6.2 | Iranome adds uncertainty to the pathogenicity of two variants with homozygous genotype but not in trans with other causative variants

The rare homozygous variant, p. Gly674Arg, in the *WFS1* gene was observed in an individual with Arab ethnicity who was apparently

normal (unlikely to be related to Wolfram syndrome). The variant is considered to be pathogenic/likely pathogenic (Khanim, Kirk, Latif, & Barrett, 2001) but our data support the idea that this variant is polymorphism in the homozygous state and it can be considered to be causative only when occurring in trans with other variants in the *WFS1* gene (Häkli, Kytövuori, Luotonen, Sorri, & Majamaa, 2014). This variant should also be reclassified as of uncertain significance in its homozygous state.

The next rare homozygous variant, p.Phe55Leu, in the *PAH* gene, has been similarly reported as compound heterozygous along with another pathogenic variant in several patients presenting different types of PAH-related disorders, but especially in mild phenylketonuria (PKU) and hyperphenylalaninemia (HPA). The observation of one homozygous individual (Persian) in the Iranome database who was apparently normal for this phenotype supports the view that this variant should be reclassified as of uncertain significance in its homozygous state.

3.6.3 | Iranome confirms the uncertainty about *GPR161* being responsible for pituitary stalk interruption syndrome

Karaca et al. (2015) reported a homozygous missense variant, p.Leu19Gln, in the *GPR161* gene as the potential novel cause of pituitary stalk interruption syndrome. In addition, this variant is recognized as likely pathogenic in the ClinVar database. However, the observation of homozygous individuals in public databases brought into question its contribution to this phenotype. Two homozygous Baluch individuals were also observed in the Iranome database. Additional clinical follow-up revealed that these individuals were apparently normal and unlikely to be related to pituitary stalk interruption syndrome. Therefore, this variant should be reclassified as a variant of uncertain significance according to ACMG guideline (See Table S1 for supporting evidence).

3.6.4 | Iranome adds uncertainty to the pathogenicity of *ZC3H14* variant but does not exclude this gene as the cause of mental retardation, autosomal recessive 56

Pak et al. (2011) reported *ZC3H14* as a novel causative gene in two families presenting nonsyndromic intellectual disability. They confirmed the expression of this gene in adult and fetal human brain and showed a critical role of its ortholog for normal *Drosophila melanogaster* development and neuronal function. The homozygous 25 bp intronic deletion, c.2204+8_2204+32del25, observed in the second family is recognized as pathogenic in the ClinVar database. However, the observation of homozygous individuals in public databases brought into question its contribution to this phenotype. The two homozygous individuals from the Iranome database belong to the Arab and Baluch ethnicities. Additional clinical follow-up was assessed. These individuals were apparently normal (unlikely to be related to Mental retardation, autosomal recessive 56). Therefore, this variant should be reclassified as a variant of uncertain significance according to ACMG guideline (See Table S1 for supporting evidence).

3.6.5 | Iranome adds uncertainty to the pathogenicity of *CHD4* variant but does not exclude this gene as the cause of Sifrim-Hitz-Weiss syndrome

In 2016, Sifrim et al. (2016) identified de novo p.Val1608Ile in the *CHD4* gene in a patient with Sifrim–Hitz–Weiss syndrome, AD (MIM# 617159). However, observation of heterozygous individuals in public databases as well as three heterozygous individuals in the Iranome database (Lur ethnicity and Persian Gulf Islanders) does not support its pathogenicity. Therefore, this variant should be reclassified as a variant of uncertain significance based on the observation of three healthy adult individuals in the Iranome database.

3.7 | Assessment of runs of homozygosity in the Iranian population

Runs of homozygosity (ROH) regions are the nearby chromosomal segments of homozygous SNPs which are categorized into three classes based on their length as short (<500 kb), intermediate (500 kb–1.5 Mb), and long ROHs (>1.5 Mb). These three classes correspond to ancient haplotypes, population background relatedness and recent parental relatedness, respectively (Pemberton et al., 2012). ROH regions are considered to have functional significance in populations and are considered to be potential target regions having a tendency for rare and common disorders (Magi et al., 2014). Long ROHs are a consequence of recent parental relatedness and therefore can be observed more frequently in inbred populations, and are more likely to surround causative variants in individuals coming from such populations (Hu et al., 2018).

Due to the high rate of consanguinity in the Iranian population, we attempted to assess the ROHs of Iranian individuals using WES data, to determine the autozygome map of Iranian individuals. In total, 1,446 highly overlapping ROH clusters were calculated (optimal clusters) in which 35.3% were short, 55.3% had intermediate length, and 9.4% were long ROHs (Table S2). The distribution of these ROHs based on their length in the Iranome database and per each ethnicity is indicated in Figure 5a. The longest ROH was about 12 Mb and was located on chromosome 16 encompassing 28 genes in which, surprisingly, all were pseudogenes. We expected an increased burden and length of ROHs in the Iranian population similar to what was observed in the GME database. When the percentage of each ROH category in Iranome was compared with the agreed 55%, 35% and 10% of short, intermediate and long ROHs reported in different 1KG populations (Pippucci, Magi, Gialluisi, & Romeo, 2014), we could observe a 20.3% increase in intermediate ROHs and a 19.7% decrease in short ROHs, compatible with expectations. Interestingly, we observed no significant difference in the proportion of long ROHs.

The overall rate of consanguineous marriage in Iran is estimated as 38.6% with an approximately equal distribution in different ethnicities and with the highest rate observed in Baluchs (Saadat, Ansari-Lari, & Farhud, 2004). We assessed the number of ROHs per each ethnicity and compared the relative distributions of these ROHs to see if they reflect the distribution pattern of consanguinity in each ethnic group

(Figure 5a,b). Interestingly, the largest number of ROHs was observed in Baluchs and Persian Gulf Islanders, the two ethnicities that formed their own clusters separate from the rest of the population in PCA, thus showing an increased burden of ROHs due to inbreeding but not increased length of the ROH compared with the other ethnicities. The highest percentages of long ROHs were observed in Turkmen, Persian, Lur, and Arab ethnicities, who had 49%, 45%, 39.3%, and 49% consanguineous marriage rates, respectively (Saadat et al., 2004).

3.8 | Web-based interface of the Iranome database

The Iranome project has produced a reference panel of genomic variations in Iran. This database provides the allele frequency of all variants identified in the Iranian population and includes genomic coordinates of corresponding alternate alleles, quality scores of the variants (including Quality score, BaseQRankSum, Read Depth (DP), Strand Bias-Fisher's (FS), Mapping Qual (MQ), MQRankSum, Quality by Depth (QD), ReadPosRankSums), corresponding dbSNP ID, the alternate allele counts, the alternate allele frequencies, genotype (GT), and genotype quality (GQ), and the number of heterozygotes and homozygotes in all 800 samples as well as in each ethnic group. Each variant was annotated and it is clinically relevant information including the transcript name, Human Genome Variation Society (HGVS) nomenclature, sequence ontology, exon number, gene name, and bioinformatics prediction of most of the well-known programs and algorithms were made available.

To provide rapid and free access to such data, a web-based interface is also provided through the two following links: http://www.iranome.com/ and http://iranome.ir/. The web server is user-friendly and provides a search box on its home page to explore variants based on their genomic positions, or by searching their gene symbol, or by specifying a transcript, or dbSNP ID (#rs ID) and also by providing corresponding coordinates of the intended genomic region.

The gene pages provide the list of all variants identified among the entire data set and the variant classification is based on the most clinically relevant transcript of a gene or the longest one when the clinically relevant transcript was not known. Each variant is hyperlinked to its specific page. The variant page provides the following details: allele frequencies of each variant in all ethnic groups studied are shown separately in a table in addition to functional and clinical annotations along with bioinformatics prediction scores and quality metrics for the specified variant. The variant page also provides useful links to the corresponding public databases such as dbSNP, UCSC genome browser, ClinVar, ExAC Browser, NHLBI ESP database, Ensemble and gnomAD Browser, providing additional information for the variant specified (Figure 6).

4 | CONCLUSION

4.1 | Impact of the Iranome project at the national level

For the best characterization of rare variants in a population, sequencing as many individuals as possible is critical. To fulfill such an objective, in phase I of the Iranome project, we sequenced 800 individuals selected from the main Iranian ethnic groups. We released a comprehensive catalog of Iranian genomic variations and constructed the Iranome database, the first public database of allele frequencies of genomic variants in the Iranian population. This is the first genomic variant database specific to the Iranian population provided by whole exome sequencing on a reasonable number of Iranian individuals coming from the main ethnic groups in this population.

Approximately 70% of the variants identified in our database were novel or had frequencies less than 1% (rare variants) in public databases. With regard to the prominence of such variants in terms of diagnosis and management of patients suffering from rare Mendelian disorders, the Iranome database offers a comprehensive healthcare resource at the national level by providing population-specific allele frequencies of such variants. The data are also accessible from an ethnic-specific viewpoint, which can be useful while interpreting the variants identified in patients coming from specific Iranian ethnical groups.

The Iranian plateau has been exposed to invasions of different people throughout history including incursion of nomads from the central Asian steppes, Arab-Muslims, Seljuq Turks originating from Oghuz tribes, and then Mongols. The large number of invasions and migrations played a major role in generating the diverse demographic structure in the Iranian plateau, which is apparently influenced by generated gene flows. This genetic diversity is confirmed by mtDNA sequencing analysis (Derenko et al., 2013) which also proposed a "common maternal ancestral gene pool" among Iranian people speaking Indo-Iranian languages and Iranian people speaking Turkic Qashqais. This is in line with the results obtained from principal components analysis of Iranome samples where, apart from the Iranian Baluch, Turkmen and Persian Gulf Islander populations, which form their own clusters, the remaining populations are genetically very similar. We also observed that the proportion of variants and, in particular, ethnic-specific novel variants, did not differ in the different ethnic groups investigated. In addition, the inclusion in the Iranome project of 100 non-Arab people living in the Persian Gulf islands played a prominent role in clarifying the genetic background and diversity of this less-studied subpopulation in Iran.

Furthermore, our analysis showed that only 0.6% of novel variants in the Iranome data set have counterparts in databases of the Middle Eastern population, emphasizing the value of the Iranome project in clarifying the genetic background and allelic frequency of the Iranian population.

4.2 | Impact of the Iranome project at the international level

This project introduced 308,311 additional variants into the human genomic variation catalog and fills another little corner of the human genetic variation picture. Furthermore, this database is an excellent resource for other countries in the region, specifically, the neighboring countries located in a region which was

historically called Greater (Persia). This historical term refers to a region which included parts of the Caucasus, West Asia and the Middle East (Bahrain, Kurdistan, the modern state of Iran, and some parts of Iraq), Central Asia (Uzbekistan, Tajikistan, Turkmenistan, Xinjiang), and parts of South Asia (Afghanistan and Pakistan), which were under the control of the Persian Empire and therefore were historically influenced by Iranian culture (https://en.wikipedia.org/wiki/Greater_Iran). Iranic (Iranian) people, defined as people who speak Indo-Iranian languages and their dialects, are present in this region and are estimated to number about 150–200 million individuals (https://en.wikipedia.org/wiki/Iranian_peoples). Moreover, applying mtDNA sequencing analysis, Derenko et al. (2013) showed that there is a common set of maternal lineages between people living in Iran and people living in Anatolia, the Caucasus and the Arabian Peninsula.

The Iranome database can be considered to be a good resource for all of the Iranic people who not only live in the country of Iran but who also inhabit this historical region. The inclusion of 100 Iranian Persians in the database can be a useful resource for Persian-speaking people living in Afghanistan, Tajikistan, the Caucasus, Uzbekistan, Bahrain, Kuwait, and Iraq. The inclusion of 100 Iranian Kurds can be a useful resource for other Kurdish people living in Iraqi Kurdistan, Turkey, Syria, Armenia, Israel, Georgia, and Lebanon. The presence of 100 Iranian Baluchs can be considered to be a useful resource for Baluchi people living in Pakistan, Oman, Afghanistan, Turkmenistan, Saudi Arabia, and UAE. The genetic information on 100 Iranian Azeris can also be used for people living in Azerbaijan, Turkey, Russian, and Georgia. Although Iranian Arabs are admixed with other ethnicities in Iran such as Persians, Turks, and Lurs, the genetic variants identified in 100 Iranian Arabs investigated in this study can be a useful resource for the populations of Arab countries which are geographically close to Iran.

Furthermore, Iranians seem to be the most probable parent population of most of the Hungarian ethnic groups. After Slavs and Germans, Iranians and Turks are the most likely contributors to the Hungarian ethnic groups, because of the relatively short genetic distances and average admixture estimates (Guglielmino, Silvestri, & Beres, 2000).

In general, genomic databases are useful in clarifying the actual role of pathogenic/likely pathogenic variants in human diseases. This will be more valuable when dealing with rare homozygous variants in a small number of individuals in public databases. Knowing the clinical features of these individuals or at least having access to collect the required clinical data from such suspected individuals will be valuable for medical geneticists. Hopefully, the Iranome database will be very helpful in this matter as already discussed above where 12 pathogenic/likely pathogenic known variants could be reclassified with the help of this dataset.

In conclusion, we believe that the Iranome database, as the first national genomic effort to clarify the genetic background of the country, will be useful in medical genomics and in the healthcare system in Iran as well as in other Indo-Iranian speaking populations in the region. We plan to improve the database by adding more individuals from other Iranian ethnic groups who were not represented in the present phase of the project to comprehensively identify Iranian genomic variations.

ACKNOWLEDGMENTS

This national project could not be completed without the supports in University of Social Welfare & Rehabilitation Sciences in Tehran, Iran. The authors would like to acknowledge the support of Dr.Sorena Sattari, vice-presidency for Science and Technology, and Dr. Hossein Vatanpour the general manager of Technology at Ministry of Health and Medical Education. The authors would like to acknowledge all of the 800 individuals who participated in this project as volunteers and the network of medical experts who helped in sample collection throughout the country. This study was funded by Iran Vice-President Office for Science and Technology (grant number: 11/66100) and Vice deputy for Research and Technology at Iran Ministry of Health and Medical Education, grant number: 700/150.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ORCID

Hossein Najmabadi http://orcid.org/0000-0002-6084-7778

REFERENCES

Alkan, C., Kavak, P., Somel, M., Gokcumen, O., Ugurlu, S., Saygi, C., … Bekpen, C. (2014). Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics*, *15*, 963. https://doi.org/10.1186/1471-2164-15-963

Amanolahi, S. (2005). A note on ethnicity and ethnic groups in Iran. *Iran & the Caucasus*, *9*(1), 37–41.

An, J. Y. (2017). National human genome projects: An update and an agenda. *Epidemiology and Health*, *39*):e2017045. https://doi.org/10.4178/epih.e2017045

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). Current protocols in bioinformatics: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, https://doi.org/10.1002/0471250953.bi1110s43. 11.10.11-33

Banihashemi, K. (2009). Iranian human genome project: Overview of a research process among Iranian ethnicities. *Indian Journal of Human Genetics*, *15*(3), 88–92. https://doi.org/10.4103/0971-6866.60182

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., … van Duijn, C. M. (2014). The Genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics*, *22*(2), 221–227. https://doi.org/10.1038/ejhg.2013.118

Comas, D., Calafell, F., Mateu, E., Pérez-Lezaun, A., Bosch, E., Martínez-Arias, R., … Bertranpetit, J. (1998). Trading genes along the silk road: MtDNA

sequences and the origin of central Asian populations. *The American Journal of Human Genetics*, *63*(6), 1824–1838. https://doi.org/10.1086/302133

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825. https://doi.org/10.1038/ng.3021

Curtis, G. E., & Hooglund, E. J. Library of Congress. Federal Research Division. (2008). *Iran: A Country Study*. Washington, DC: Federal Research Division, Library of Congress.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davit-Spraul, A., Costa, C., Zater, M., Habes, D., Berthelot, J., Broué, P., … Baussan, C. (2008). Hereditary fructose intolerance: Frequency and spectrum mutations of the aldolase B gene in a large patients cohort from France--identification of eight new mutations. *Molecular Genetics and Metabolism*, *94*(4), 443–447. https://doi.org/10.1016/j.ymgme.2008.05.003

Derenko, M., Malyarchuk, B., Bahmanimehr, A., Denisova, G., Perkova, M., Farjadian, S., & Yepiskoposyan, L. (2013). Complete mitochondrial DNA diversity in Iranians. *PLOS One*, *8*(11):e80673. https://doi.org/10.1371/journal.pone.0080673

Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Alemán, A., García-García, F., … Antiñolo, G. (2016). 267 Spanish Exomes reveal population-specific differences in disease-related genetic variation. *Molecular Biology and Evolution*, *33*(5), 1205–1218. https://doi.org/10.1093/molbev/msw005

Esposito, G., Imperato, M. R., Ieno, L., Sorvillo, R., Benigno, V., Parenti, G., … Salvatore, F. (2010). Hereditary fructose intolerance: Functional study of two novel ALDOB natural variants and characterization of a partial gene deletion. *Human Mutation*, *31*(12), 1294–1303. https://doi.org/10.1002/humu.21359

Farhud, D. D., Mahmoudi, M., Kamali, M. S., Marzban, M., Andonian, L., & Saffari, R. (1991). Consanguinity in Iran. *Iranian Journal of Public Health*, *20*, 1–16.

Farjadian, S., & Safi, S. (2013). Genetic connections among Turkic-speaking Iranian ethnic groups based on HLA class II gene diversity. *International Journal of Immunogenetics*, *40*(6), 509–514. https://doi.org/10.1111/iji.12066

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., … Akey, J. M. (2013). Erratum: Corrigendum: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, *495*, 270. https://doi.org/10.1038/nature12022

Guglielmino, C. R., Silvestri, A., & Beres, J. (2000). Probable ancestors of Hungarian ethnic groups: An admixture analysis. *Annals of Human Genetics*, *64*(Pt 2), 145–159. https://doi.org/10.1017/S0003480000008010

Hassan, H. D. (2008). *Iran: Ethnic and Religious Minorities* (Report no. RL34021). Washington D.C. (https://digital.library.unt.edu/ark:/67531/metadc795725/): University of North Texas Libraries, Digital Library, https://digital.library.unt.edu; crediting UNT Libraries Government Documents Department.

Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. *Proceedings of the National Academy of Sciences*, *109*(44), 17758–17764. https://doi.org/10.1073/pnas.1212380109

Hu, H., Kahrizi, K., Musante, L., Fattahi, Z., Herwig, R., Hosseini, M., … Najmabadi, H. (2018). Genetics of intellectual disability in consanguineous families. *Molecular Psychiatry*, *24*, 1027–1039. https://doi.org/10.1038/s41380-017-0012-2

Häkli, S., Kytövuori, L., Luotonen, M., Sorri, M., & Majamaa, K. (2014). WFS1 mutations in hearing-impaired children. *International Journal of Audiology*, *53*(7), 446–451. https://doi.org/10.3109/14992027.2014.887230

Karaca, E., Buyukkaya, R., Pehlivan, D., Charng, W. L., Yaykasli, K. O., Bayram, Y., … Lupski, J. R. (2015). Whole-exome sequencing identifies homozygous GPR161 mutation in a family with pituitary stalk interruption syndrome. *The Journal of Clinical Endocrinology & Metabolism*, *100*(1), E140–E147. https://doi.org/10.1210/jc.2014-1984

Khanim, F., Kirk, J., Latif, F., & Barrett, T. G. (2001). WFS1/wolframin mutations, Wolfram syndrome, and associated diseases. *Human Mutation*, *17*(5), 357–367. https://doi.org/10.1002/humu.1110

Koshy, R., Ranawat, A., & Scaria, V. (2017). al mena: A comprehensive resource of human genetic variants integrating genomes and exomes from Arab, Middle Eastern, and North African populations. *Journal of Human Genetics*, *62*(10), 889–894. https://doi.org/10.1038/jhg.2017.67

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., … Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*(7497), 469–476. https://doi.org/10.1038/nature13127

Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., … Pippucci, T. (2014). H3M2: Detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics*, *30*(20), 2852–2859. https://doi.org/10.1093/bioinformatics/btu401

Majbouri, M., & Fesharaki, S. (2017). Iran's multi-ethnic mosaic: A 23-year perspective. *Social Indicators Research*, https://doi.org/10.1007/s11205-017-1800-4

Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N., & Ku, C. S. (2011). Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human genomics*, *5*(6), 577–622.

Pak, C., Garshasbi, M., Kahrizi, K., Gross, C., Apponi, L. H., Noto, J. J., … Kuss, A. W. (2011). Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in Drosophila and humans. *Proceedings of the National Academy of Sciences*, *108*(30), 12390–12395. https://doi.org/10.1073/pnas.1107103108

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics*, *91*(2), 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014

Pippucci, T., Magi, A., Gialluisi, A., & Romeo, G. (2014). Detection of runs of homozygosity from whole exome sequencing data: State of the art and perspectives for clinical, population and epidemiological studies. *Human Heredity*, *77*(1-4), 63–72. https://doi.org/10.1159/000362412

Rashidvash, V. (2012). The race of the Azerbaijani people in Iran (Atropatgan). *International Journal of Research In Social Sciences (IJRSS)*, *2*(3), 437–449.

Rashidvash, V. (2013). Iranian people: Iranian ethnic groups. *International Journal of Humanities and Social Science*, *3*(15), 216–226.

Rashidvash, V. (2016). Iranian people and the race of people settled in the Iranian plateau. *International Journal of Humanities & Social Science Studies*, *3*(1), 181–191.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). 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. *Genetics in Medicine*, *17*(5), 405–423. https://doi.org/10.1038/gim.2015.30

Saadat, M., Ansari-Lari, M., & Farhud, D. D. (2004). Short report consanguineous marriage in Iran. *Annals of Human Biology*, *31*(2), 263–269. https://doi.org/10.1080/03014460310001652211

Scott, E. M., Halees, A., Itan, Y., Spencer, E. G., He, Y., Azab, M. A., … Gleeson, J. G. (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nature Genetics*, *48*(9), 1071–1076. https://doi.org/10.1038/ng.3592

Sifrim, A., Hitz, M. P., Wilsdon, A., Breckpot, J., Turki, S. H. A., Thienpont, B., … Hurles, M. E. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature Genetics*, *48*(9), 1060–1065. https://doi.org/10.1038/ng.3627

Tennessen, J. A., Bigham, A. W., O'connor, T. D., Fu, W., Kenny, E. E., Gravel, S., … Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*(6090), 64–69. https://doi.org/10.1126/science.1219240

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603

Yamaguchi-Kabata, Y., Nariai, N., Kawai, Y., Sato, Y., Kojima, K., Tateno, M., … Nagasaki, M. (2015). iJGVD: An integrative Japanese genome variation database based on whole-genome sequencing. *Human Genome Variation*, *2*, 15050. https://doi.org/10.1038/hgv.2015.50

Zarei, F., & Alipanah, H. (2014). Mitochondrial DNA variation, genetic structure and demographic history of Iranian populations. *Molecular Biology Research Communications*, *3*(1), 45–65.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Fattahi Z, Beheshtian M, Mohseni M, et al. Iranome: A catalog of genomic variations in the Iranian population. *Human Mutation*. 2019;1–17. https://doi.org/10.1002/humu.23880