Received: 10 September 2021 | Revised: 16 December 2021 | Accepted: 29 January 2022

DOI: 10.1002/humu.24336

RESEARCH ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Qatar genome: Insights on genomics from the Middle East

Hamdi Mbarek[1] | Geethanjali Devadoss Gandhi[3,6] | Senthil Selvaraj[6] | Wadha Al-Muftah[1,4] | Radja Badji[1] | Yasser Al-Sarraj[1,5] | Chadi Saad[1] | Dima Darwish[1] | Muhammad Alvi[1] | Tasnim Fadl[1] | Heba Yasin[1] | Fatima Alkuwari[1] | Rozaimi Razali[2] | Waleed Aamer[6] | Fatemeh Abbaszadeh[7] | Ikhlak Ahmed[8] | Younes Mokrab[6] | Karsten Suhre[5] | Omar Albagha[3,9] | Khalid Fakhro[6] | Ramin Badii[7] | Said I. Ismail[1,3] | Asma Althani[1,10] | Qatar Genome Program Research Consortium

[1]Qatar Genome Program, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha, Qatar

[2]Department of Biomedical Sciences, College of Health Sciences, Qatar University, Doha, Qatar

[3]College of Health & Life Sciences, Hamad Bin Khalifa University, Education City, Doha, Qatar

[4]Department of Genetic Medicine, Weill Cornell Medicine-Qatar, Doha, Qatar

[5]Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar

[6]Human Genetics Department, Sidra Medicine, Doha, Qatar

[7]Molecular Genetics Laboratory, Hamad Medical Corporation, Doha, Qatar

[8]Sidra Medicine, Biomedical Informatics - Research Branch, Doha, Qatar

[9]Center of Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, UK

[10]Biomedical Research Center, Qatar University, Doha, Qatar

Correspondence

Hamdi Mbarek and Said I.Ismail, Qatar Genome Program, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha 5825, Qatar.
Email: hmbarek@qf.org.qa (H. M.); saismail@qf.org.qa (S. I. I.)

Abstract

Despite recent biomedical breakthroughs and large genomic studies growing momentum, the Middle Eastern population, home to over 400 million people, is underrepresented in the human genome variation databases. Here we describe insights from Phase 1 of the Qatar Genome Program with whole genome sequenced 6047 individuals from Qatar. We identified more than 88 million variants of which 24 million are novel and 23 million are singletons. Consistent with the high consanguinity and founder effects in the region, we found that several rare deleterious variants were more common in the Qatari population while others seem to provide protection against diseases and have shaped the genetic architecture of adaptive phenotypes. These results highlight the value of our data as a resource to advance genetic studies in the Arab and neighboring Middle Eastern populations and will significantly boost the current efforts to improve our understanding of global patterns of human variations, human history, and genetic contributions to health and diseases in diverse populations.

KEYWORDS
Arab ancestry, diversity, genetics, large-scale sequencing, Middle East

1 | INTRODUCTION

Several countries worldwide have initiated large-scale population genomics projects representing various regions from Africa, Europe, North and South America, South Asia, and Australia (Gudbjartsson, Helgason, et al., 2015; Gudbjartsson, Sulem, et al., 2015; Gurdasani et al., 2019; Manolio et al., 2019; Naslavsky et al., 2020; Stark et al., 2019; Turro et al., 2020; Wu et al., 2019). In addition t6o this groundbreaking work, there are also ongoing large collaborative efforts to increase diversity in human genetics, including the All of Us Research Program (Collins & Varmus, 2015), the Human Health and Heredity in Africa (H3Africa) Initiative (C. Rotimi et al., 2014), and the TOPMed Program (Taliun et al., 2021). Such studies provided valuable new insight into human disease, population structure, and history of migration (Boomsma et al., 2014; Chiang et al., 2018; Francioli et al., 2014; Gurdasani et al., 2019; Okada et al., 2018; Scott et al., 2016; Wu et al., 2019). Despite this notable focus on diversity, there is still considerable effort needed to cover the broad diversity of world ancestries to ensure that discoveries do not conserve historical disparities and to uncover the various diseases etiologies that remain uncharacterized to date (Bentley et al., 2017; Landry et al., 2018; Mills & Rahal, 2019). The Middle-East regions are still underrepresented in the public databases (Abou Tayoun & Rehm, 2020). For instance, the latest version of gnomAD database (3.1) contains data from only 158 Middle-eastern genomes (Karczewski et al., 2020). The Qatar Genome Program (QGP) is a population genome project based in Qatar aiming to sequence the genomes of local population for the purpose of supporting genomic medicine in the country and the region. As part of Phase 1, it has sequenced the whole genomes of 6045 subjects whose specimens were collected and biobanked by the Qatar Biobank (QBB) (Al Thani et al., 2019) (Figure 1a).

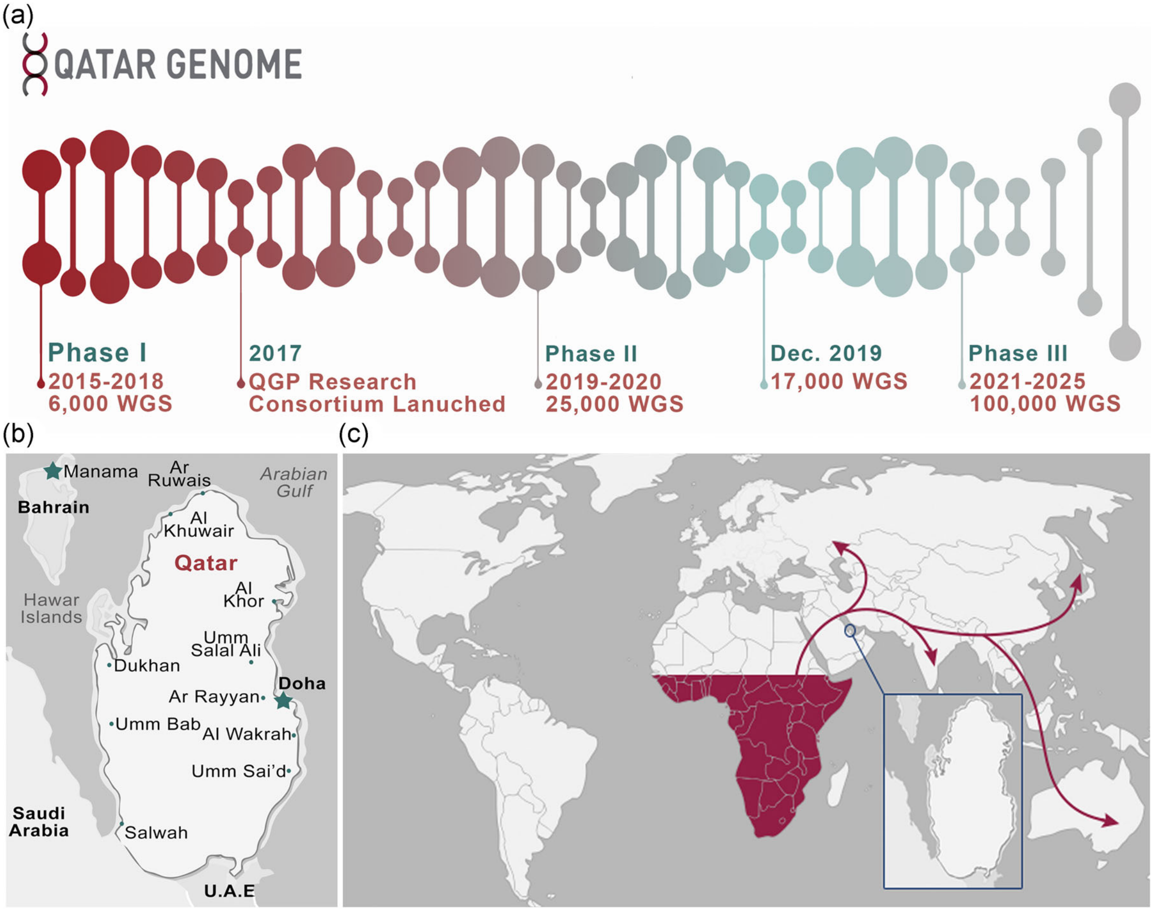

FIGURE 1 Qatar Genome Program, timelines, and regional context. (a) Three phases project timeline and current status. (b) Qatar Geographical map. Qatar is located in the north-eastern coast of the Arabian Peninsula with an area of 11,521 km^2 sharing borders with Saudi Arabia from the south and maritime borders with Bahrain, UAE, and Iran. (c) The Arabian Peninsula is believed to be the first stop in human migration out of Africa, and home for the first ancient Eurasian populations, whom later spread throughout Asia and Europe

Qatar occupies a relatively small surface area of 11,521 km^2 on the western coast of the Arabian Gulf. Qatar shares its southern border with Saudi Arabia and a maritime border with Bahrain, UAE, and Iran (Figure 1b) and has a population of approximately 2.8 million. The country is located at a historic intersection of ancient and recent migration and admixture (Arauna et al., 2017; Hellenthal et al., 2014). Similar to other countries in the region, it is known for its unique population structure that is characterized by a high consanguinity rate and increased prevalence of rare genetic diseases (Al-Gazali et al., 2006; Anwar et al., 2014; Hunter-Zinck et al., 2010; Rodriguez-Flores et al., 2014, 2016; Scott et al., 2016). Recent genetic studies identified indigenous Arabs as the direct descendants of the first Eurasian populations established by early migrations out of Africa (Bentley et al., 2017) (Figure 1c). Moreover, sizable proportions of the population have more recent Persian and African ancestry (Harkness & Khaled, 2014). QBB includes comprehensive phenotyping, providing excellent synergy for discovery when combined with the whole genome sequencing (WGS) data, that also enable accurate estimate of allele frequencies for rare and common variants, and well-defined polygenic risk scores for many disease traits. All such features of the local population potentiate discoveries, not only related to millions of people in the immediate neighboring region but also inform genetic studies in other parts of the world.

2 | MATERIALS AND METHODS

2.1 | QBB subject recruitment

The QBB is a longitudinal population-based cohort study examining a population sample of permanent Qatari residents (Qatari nationals, other Arabs and non-Arabs) with follow-up every 5 years (Al Thani et al., 2019). To achieve a representative sample of the permanent population that resides in Qatar, the inclusion criteria of the QBB are: (1.) To be Qatari nationals or resident in Qatar for at least 15 years and (2.) To be 18 years or older. QBB is inclusive and language specification and tribes name or origin are not part of the inclusion criteria. The participants are recruited from the general public via either social media and the QBB website or through personal recommendations of family and friends.

The study covers extensive baseline sociodemographic data, clinical, and behavioral phenotypic data, biological samples (i.e., blood, urine, saliva, DNA, RNA, viable cells, and others), as well as clinical biomarkers and Omics data (i.e., genomics, transcriptomics, proteomics, metabolomics, etc.) (Al Thani et al., 2019). Currently the QBB has reached 44.7% of the target population (60,000) and more than 2 million biological samples. For this study, data from 6218 Qatari nationals participants were available from QBB population cohort. The percentage female was 56.74% and the mean age was 40 years (SD: 12.7 years).

2.2 | Ethics statement

All QBB participants signed an Informed Consent Form before their participation; Ethical approval for QBB study protocol was obtained from the Hamad Medical Corporation (HMC) Ethics Committee in 2011 and continued with QBB Institutional Review Board (IRB) from 2017 onwards and it is renewed on an annual basis (IRB protocol number, QF-QGP-RES-PUB-002).

2.3 | QBB sample collection

Physical and clinical measurements were collected by the QBB, in addition to biological samples (approximately 60 ml of blood, 5 ml of saliva, and 10 ml of urine). Participants were instructed for 8 h fasting before the visit, but due to different visit shifts samples were mostly spot specimens. Blood samples were analyzed to assess 66 different biomarkers associated with disease risk factors. Hematology and blood chemistry biomarkers were analyzed at Hamad General Hospital laboratories. EDTA blood samples were separated by centrifugation into plasma, buffy coat (leukocytes), and erythrocytes. All collected samples were aliquoted and stored in three different locations (Al Thani et al., 2019).

2.4 | DNA isolation and quality control

Before DNA isolation, each buffy coat sample was registered into the Laboratory Information Management System (LIMS) and assigned with three identifiers: (i.) the aliquot code, (ii.) a subject-specific personal number, and (iii.) a sample-specific serial number. Samples were received in 2D-coded FluidX tubes (Brooks Life Sciences). Upon receiving, samples were scanned on a 2D FluidX Perception Barcode Reader to check for consistency against the sample submission form. The buffy coat samples were processed for DNA isolation using the automated QIASymphony SP instrument according to Qiagen MIDI kit protocol's recommendations. The assessment of DNA quantity and quality was carried out using NanoDrop 8000 (Thermofisher), FlexStation 3 (Molecular Devices), and LabChip GX (Perkin Elmer). The absorbance at 260 and 280 nm wavelength was measured on Nanodrop 8000 and used to check DNA purity. A fluorescence-based quantification was performed on FlexStation 3 using Quant-iT PicoGreen dsDNA Assay (Thermofisher). Briefly, an aqueous working solution of the Quant-iT PicoGreen reagent was prepared on the day of the quantification experiment by making a 200-fold dilution of the concentrated dimethyl sulfoxide solution in tris-EDTA (TE). TE buffer was also used for diluting DNA samples and in the assay itself. Sample measurement on FlexStation 3 was performed following the manufacturer's recommendations. DNA integrity was checked on LabChip GX. The Gel-Dye solution, DNA samples, and DNA ladder were prepared according to the manufacturer's instructions; the run data were compared with the electropherogram of a typical high-molecular-weight ladder and assessed for quality. A genomic DNA (gDNA) quality score (GQS) was calculated for each sample. The GQS is derived from the size distribution of the gDNA and it represents the degree of degradation of a given sample, with a score of 5 corresponding to intact gDNA and a score of 0 corresponding to a highly degraded gDNA. Figure S1 shows the GQS distribution across 50 samples assessed from Phase I. The distribution shows GQS>3.5.

2.5 | Whole genome sequencing

Library construction and sequencing were performed at the Sidra Clinical Genomics Laboratory Sequencing Facility. After extraction of gDNA, sample integrity was controlled using the gDNA assay on the Perkin Elmer Caliper Labchip GXII. Concentration was measured using Invitrogen Quant-iT dsDNA Assay on the FlexStation 3. Around 150 ng of DNA were used for library construction with the Illumina TruSeq DNA Nano kit. Each library was indexed using the Illumina TruSeq Single Indexes. Library quality and concentration were assessed using the DNA 1k assay on a Perkin Elmer GX2. Libraries were quantified using the KAPA HiFi Library quantification kit on a Roche LightCycler 480. Flow cells were loaded at 1 sample per lane and cluster generation was performed on a cBot 1.0 or 2.0 using the HiSeq X Ten Reagent Kit v2.5. Flow cells were loaded at a cluster density between 1255 and 1412 K/mm^2 and sequenced on an Illumina Hiseq X instrument to a minimum average coverage of 30x.

2.6 | Sequencing data processing methods

The Sidra Bioinformatics Core (SBC) developed a pipeline to perform the NGS analysis for QGP and other internal projects (Figure S2). The core also developed a framework to automate the processing of the samples. Data are received from the clinical genomic lab (CGL) in Fastq format. Quality control of Fastq files is performed using FastQC (v0.11.2), to calculate quality metrics and ensure that raw reads have good quality. Reads are then trimmed and aligned to hs37d5 reference genome using bwa.kit (v0.7.12) and a bam file is generated. Quality control on mapped reads (BAM files), to evaluate the coverage of each sample, is performed using Picard (v1.117) [CollectWgsMetrics]. The variant calling is performed following GATK 3.4 best practices: Indel realignment and base recalibration (BQSR) is performed on the initial bam then HaplotypeCaller run on each sample to generate an intermediate genomic gVCF (gVCF). Joint Genotyping is performed using all generated gVCF files at once. We first run GenomicsDB to combine the different samples by regions, then on each region, we run GenotypeGVCFs, apply SNP/Indel recalibration (VQSR), and then merge all regions. Annotation is performed using SnpEff/SnpSift (v4.3t). The following databases are used within SnpEff/SnpSift for the annotation of the multisamples VCF file:

- dbSNP build 151
- ClinVar 2019-02-11
- dbNSFP v2.9
- GWAS catalog
- msigDBdb v5.0

All variants are kept within the VCF file. Copy Number Variation analysis was performed using Canvas (v1.11.0) and structural variant analysis was performed using Manta (v0.29.6) and Delly (v0.7.8). Both analyses use bam file as input and were performed at the single sample level. Additionally, QGP VCF file was decomposed for multi allelic position and then normalize using vt (v0.5). QGP VCF file was split chromosome wise and this per chromosome VCF file was provided for further analysis as well. All pipeline references are in the Supporting Information Data.

To identify disease-causing variants in HGMD, ClinVar and OMIM, we used VCF file annotated with phenotype/disease information from these databases. To achieve that, we applied successive filtering on the variant list using different criteria (selecting only those located in known HGMD/OMIM gene, variants with minor allele frequency (MAF) <1% in all databases, except QGP, and the variant should be within or affecting the coding region; missense, nonsense, frameshift, and splice-site variants). Among the final list, we selected those that have been previously reported and flagged as disease-causing "disease-causing mutations (DM)/DM?" in HGMD or "Pathogenic/Likely_pathogenic" in ClinVar.

2.7 | Data quality control

QGP Phase I study included 6218 samples. We applicate downstream quality control on the multisample VCF using the PLINK v2.0 tool (Chang et al., 2015). After quality control, eight samples were removed for excess heterozygosity, one for low-call rates (less than 95%), 65 for gender mismatch, 87 for population outliers (individuals with more than four standard deviation (±4 SD) away from the mean of the first two multidimensional scaling component), and 10 for identical matching. After these exclusions (*N* = 171), a final set of 6047 subjects was obtained (Thareja et al., 2021).

2.8 | Statistical analyses

We compared the allele counts of QGP samples to allele counts present in gnomAD exome samples for HGMD DM variants. A Fisher's exact test was used to calculate variations that were significantly overrepresented in the QGP samples (due to founder effect) and corrected for multiple testing using the Bonferroni method.

2.9 | Hail genomic processing tool

Data preprocessing and analysis were performed using Hail 0.2. allele count, allele number, allele frequency, homozygous count calculation for each subpopulation was performed simultaneously using python scripts written using hail framework. Quality analysis for variant calls and individual sample were performed using variant_qc and sample_qc functions, respectively. Sample level statistics for each sample was generated using the Hail.

2.10 | QGP variant browser

QGP variant browser provides a mechanism for the researchers to be able to search, filter, and browse the QGP genomic variants data. This web-based browser supports fast database query response time for searching through more than 88 million records with search and filter functionality on the QGP gene variants and its attributes (e.g., allele frequency, homozygosity, etc.). The access procedure is described in Supporting Information Document 2.

3 | RESULTS

3.1 | Genetic variability of the Qatari population

We have identified a total of 88,191,239 variants, which includes 74,991,446 single-nucleotide variants (SNVs) (74,040,559 bi-allelic SNVs) with 939,405 multiallelic sites and 13,199,792 INDELS (8,389,562 bi-allelic INDELS) with 2,018,185 multiallelic sites/microsatellites (Figures 2a-c and S3). Importantly, twenty-eight percent (28%) of the total variants (24,620,313) were novel and not previously reported in single nucleotide polymorphism Ddatabase (dbSNP) build 151 or other population databases (gnomAD, 1000 Genomes, and Greater Middle East [GME]) (Figures 2b, S4a-b, and S5). Each individual genome presented a median of 3.4 million SNVs and 63,755 novel variants. We estimated the transition to transversion (ti/tv) ratio of 2.05 and heterozygotes to nonref homozygote (Het/Hom) ratio of 1.85, which is consistent with previous WGS studies (Auton et al., 2015). We found 23 million variants present as singletons which are less when compared with the number of variants falling under the MAF spectrum of <0.1% (2–12 alleles) which should be around 34 million variants (Figure 2c and Table S1). While considering the novel variants, singletons (45%) being slightly higher than the variants that

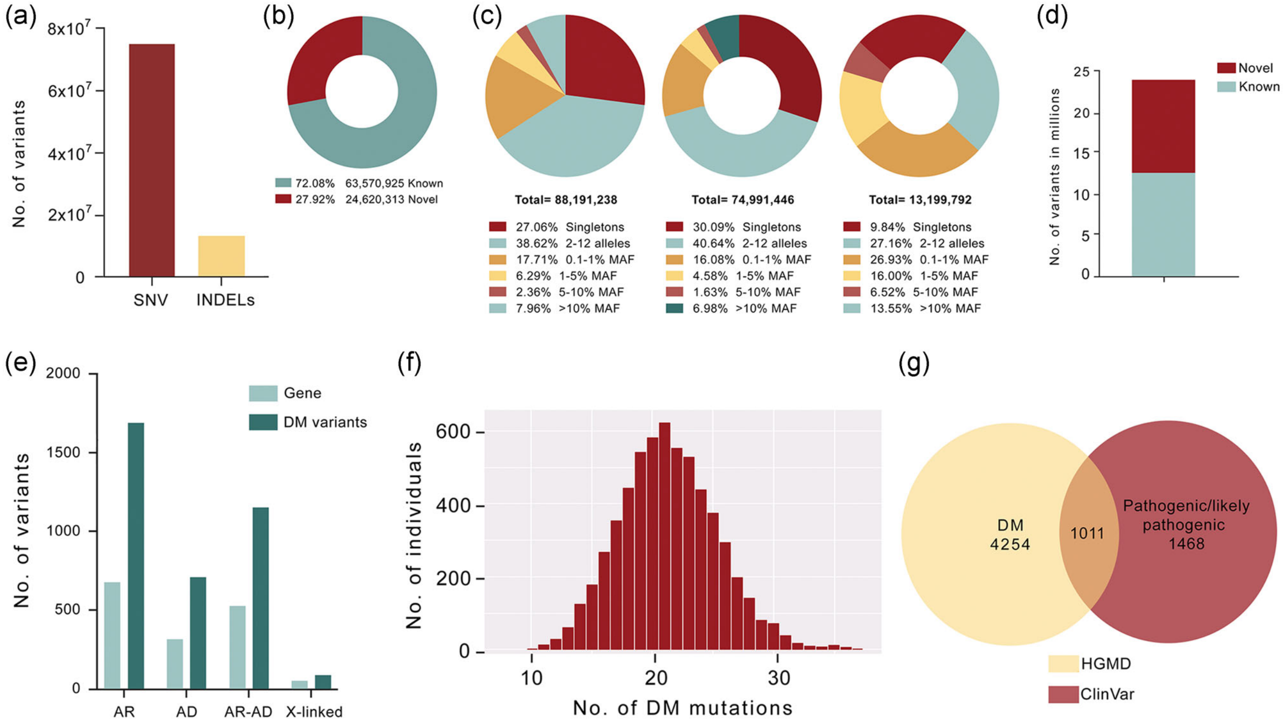

FIGURE 2 Variants distribution and allele frequency spectrum of QGP data. Number of SNVs and INDELS present within the QGP data. (b) Known and novel variants distribution of QGP data. (c) QGP variants classification based on minor allele frequency (MAF). (d) Proportion of known and novel singletons within the QGP data. (e) Classification of DM variants based on pattern of inheritance. Inheritance patterns of genes were derived from OMIM database. (f) Distribution of DM variants among individuals in QGP sub clusters. (g) QGP variants classified as both DM and pathogenic/likely pathogenic. QGP, Qatar Genome Program

fall in the category of 2–12 alleles (42%) and only 13% of the novel variants exceed the MAF > 0.1%. Half of the singletons present in QGP were already reported in dbSNP and, each individual carried a median of 1336 singletons (Figures 2d and S6).

To evaluate the impact and scale of disease-causing variants in our population, we annotated the variant list with disease/phenotype information from HGMD, ClinVar, and OMIM databases. In total, we found 4254 disease-causing mutations (DM), which includes 3970 SNVs and 284 INDELS (Figure S7a). These variants are located across 1672 genes that are linked to phenotypes with different modes of inheritance (678 follow autosomal recessive [AR]; 315 autosomal dominant [AD]; 526 both AR and AD; and 50 X-linked inheritance) (Figure 2e). The vast majority (97%) of these DM variants are rare with MAF <1%, and among these 30% observed as singletons (Figure S7b). Each individual in the QGP data set carries a median of 21 DM variants (range of 8–37) (Figures 2f and S7c), slightly less than what have been previously reported; 25 DMs/individual in the UK10k (Xue et al., 2012) and 29 DMs/individual in the Uganda genome studies (Gurdasani et al., 2019). Each individual also carries in the homozygous state a median of five DM variants (range of 1–11) compared with three homozygous DMs/individual in the Uganda genome and project (Auton et al., 2015; Gurdasani et al., 2019). Our data show that approximately 900 protein-coding genes have at least one DM mutation and 26 genes present 15 or more DM mutations (Figure S7d). When QGP data are classified according to ClinVar information (version February 11, 2019), we found that 1449 variants are classified as "pathogenic" or "likely pathogenic" (Figure S7e). Further classification considering both HGMD and ClinVar, revealed that 1011 variants were marked as DM and "pathogenic or likely pathogenic" (Figure 2g), with 160 variants unique to the Qatari population. Interestingly, only a subset of 14 variants, among the 1011 variants, are shared between the QGP samples and data from GME Variome Project (Scott et al., 2016) (Table 1). There are also 34 variants which confer protection against several diseases including malaria, obesity, and heart disease (Table S2).

We found some rare pathogenic variants present in Qatari population with high minor allele frequencies due to the founder effect. Some of the examples include variant in the *MPL* gene [MIM: 604498] (rs750046020, NM_005373.3:c.317C>T; NM_005373.3:p.Pro106Leu), previously associated with thrombocytosis, occurs at a MAF of 0.9%, and similarly, variants in the genes *CBS [MIM:*236200] (rs398123151, NM_001178008.3:c.1006C>T; NM_001178008.3:p.Arg336Cys) and *KRT5* [MIM: 148040] (rs267607448, NM_000424.4:c.1411C>T; NM_000424.4:p.Arg471Cys) associated with homocystinuria and Epidermolysis Bullosa, respectively, are observed at a MAF of 0.7%."

3.2 | Genetic ancestry and diversity of the Qatari population

To capture the genetic diversity of the Qatari population and understand its relationship with the world's populations in both modern and ancient times, we identified six major ancestries: General Arabs (QGP_GAR, 38%), Peninsular Arabs (QGP_PAR, 17%), Arabs of Western Eurasia and Persia (QGP_WEP, 22%), South Asian Arabs (QGP_SAS, 1%), African Arabs (QGP_AFR, 3%), and Admixed Arabs (QGP_ADM, 19%) (Razali et al., 2021) (Figure S8). These genetic clusters have distinct signatures in terms of Chr-Y haplogroups (Razali et al., 2021) (Figure 6). Notably, the J1a2b Chr Y haplogroup, seen previously in Southern Arabia, was observed in 1419 males making it the largest set of individuals ever sequenced for this haplogroup. We identified 29 novel subhaplogroups of J1a2b where the individuals were mainly of QGP_GAR and QGP_PAR ancestries (Razali et al., 2021) (Figure 6). In our study, the majority of QGP_PAR individuals are descendants of a tribe that originated from the historical homeland of ancient Arab tribes in Southern Arabia. These results suggest the richness in terms of the genomic diversity of the population, which can represent the whole Middle Eastern region. Thus, we use this to our advantage by creating a dedicated imputation reference panel for the Middle Eastern population, given their lack of representation in current publicly available imputation panels. We were able to show superior performance compared with existing imputation reference panels and an improved imputation rate for rare and common allele frequencies variants (Razali et al., 2021) (Figure 7).

We next characterized the spectrum of genetic variability based on the fine-scale population structure observed in the Qatari population. This analysis highlighted that 70% of the novel variants are cluster-specific, 5% are found in all subclusters, and the remaining 25% are shared between one or more subclusters (Figure S9a). Similarly, we found that about half (2139) of the DM variants are cluster-specific and only 68 out of 4254 DM variants were present in all subclusters (Figure S9b). Furthermore, individuals in the QGP_AFR subcluster have the highest heterozygotes to nonref homozygote (Het/Hom) ratio, whereas the ratio was found to be lowest for the QGP_PAR cluster. This reflects the high homozygosity and high consanguinity present within the individuals of this cluster (Figure S9c). Similarly, the median number of singletons is lower for PAR cluster compared with other subclusters reflects the closely related individual present in this cluster (Table 2).

Furthermore, runs of homozygosity (ROH) analysis of the QGP done by Razali et al. (2021), identified per population ROH boundary for short, medium, and long ROH. We observed that Peninsular Arabs (PAR) have the lowest median for short ROH after African-based populations. In addition, PAR has the highest median for long ROH, indicating recent consanguinity events. When we analyzed the relationship between genes and the ROH regions, we observed that there are more OMIM genes in ROH regions compared with non-OMIM genes regardless of the ROH classes. PAR was shown to have significantly more OMIM genes compared with the other QGP and 1KG populations.

3.3 | Burden of pathogenic variation

We then focused on the burden of pathogenic variants of recessively inherited disorders in the Qatari population. We found the most

TABLE 1 Pathogenic variants unique to the Middle East region

Chrom	Pos	ID	QGP_ MAF	GME_MAF	GENE	Ensembl Transcript ID	HGVS_C	HGVS_P	HGMD	CLINVAR	Disease phenotype
2	47604159	rs606231204	0.00223	0.00050	EPCAM	ENST00000405271	c.499dup	p.Gln167fs	DM	P	Congenital Tufting Enteropathy (CTE)
2	73676380	rs746640196	0.00017	0.00050	*ALMS1*	ENST00000264448	c.2723C>G	p.Ser908*	DM	LP	Alstrom syndrome
2	172305304	rs797045038	0.00728	0.00101	*DCAF17*	ENST00000375255	c.436delC	p.Ala147fs	DM	P	Woodhouse-Sakati syndrome
3	113119479	rs866096259	0.00331	0.00050	*WDR52*	ENST00000393845	c.1387G>T	p.Glu463*	DM	P	Spermatogenic failure
4	108866582	rs397514513	0.00223	0.00151	*CYP2U1*	ENST00000332884	c.947A>T	p.Asp316Val	DM	P	Spastic paraplegia
4	119736287	rs730882211	0.00149	0.00101	*SEC24D*	ENST00000379735	c.700G>C	p.Gly234Arg	DM	LP	Intellectual disability/Seizures
6	135776888	rs121434350	0.00017	0.00101	*AHI1*	ENST00000265602	c.1328T>A	p.Val443Asp	DM	P/LP	Joubert syndrome
8	145741257	.	8.27E–05	0.00050	*RECQL4*	ENST00000428558	c.1149G>A	p.Trp383*	DM	P	Rothmund-Thomson syndrome
9	111899809	rs878853280	0.00041	0.00101	*FRRS1L*	ENST00000561981	c.961C>T	p.Gln321*	DM	P	Epileptic encephalopathy
15	65295453	rs863224897	8.27E–05	0.00050	*MTFMT*	ENST00000220058	c.1116delT	p.Pro373fs	DM	LP	Moyamoya disease
16	77369781	rs148319220	8.27E–05	0.00101	*ADAMTS18*	ENST00000282849	c.1731C>G	p.Cys577Trp	DM	P	Microcornea,myopic chorioretinal atrophy,and telecanthus (MMCAT)
19	11304215	.	0.00256	0.00050	*KANK2*	ENST00000432929	c.541A>G	p.Ser181Gly	DM	P	Nephrotic syndrome,16
21	47805894	rs387906928	8.27E–05	0.00050	*PCNT*	ENST00000359568	c.3460G>T	p.Glu1154*	DM	P	Microcephalic osteodysplastic primordial dwarfism type 2 (MOPD2)
22	27012112	rs1064793935	0.00025	0.00050	*CRYBB1*	ENST00000215939	c.171delG	p.Asn58fs	DM	P	Cataract

Note: Pathogenic variants exclusively reported in QGP and GME (Greater Middle East) variome project. QGP_MAF - Minor allele frequency in QGP data.

Abbreviations: ClinVar, P-Pathogenic; GME_MAF, minor allele frequency in GME variome project; HGMD: DM, disease-causing mutation; LP, likely pathogenic.

TABLE 2 Median number of variant sites per genome

Annotation	QGP (*n* = 6,045, Depth = 32.4x)	ADM (*n* = 1,180, Depth = 32.2x)	AFR (*n* = 92, Depth = 31.9x)	GAR (*n* = 2,311, Depth = 32.2x)	PAR (*n* = 1,052, Depth = 32.2x)	SAS (*n* = 38, Depth = 31.9x)	WEP (*n* = 1, 372, Depth = 32.2x)
SNV	3,467,270	3,596,354	3,967,082	3,466,051	3,391,850	3,492,506	3,458,604
INDELS	1,107,288	1,128,043	1,207,016	1,105,836	1,094,173	1,113,900	1,101,075
Singletons	1336	9242	17,606	2484	408	20,056	3193
Novel SNV	18,453	21,311	25,993	16,419	12,022	23,814	20,788
Novel INDELS	45,756	46,263	48,406	45,752	46,061	46,195	45,107
Synonymous	10,657	11,094	12,285	10,643	10,372	10,768	10,635
Missense	10,681	11,241	12,464	10,921	10,684	10,997	10,895
Intron	1,617,713	1,659,957	1,833,502	1,618,061	1,586,985	1,632,466	1,613,603
Intergenic	1,760,161	1,806,271	1,989,241	1,759,321	1,726,938	1,774,147	1,756,828
Conserved: GERP>3	3751	3859	4233	3755	3667	3788	3738

Note: Novel SNV and INDELS: Variants, which are not reported in dbSNP or gnomAD or 1000G project. GERP (Genomic Evolutionary Rate Profiling) score: Scores >3 represent highly conserved positions.

Abbreviations: ADM, Admixed; AFR, Africans; GAR, general Arabs; PAR, Peninsular Arabs; SAS, South Asians; WEP, Arabs of Western Eurasia and Persia.

common recessive alleles are those linked to structural deformities and developmental disorders, consistent with the fact that such recessive traits prevail in societies where endogamy and consanguinity are practiced (Table S3). However, some of these identified alleles are too common to be classified as pathogenic variants (rs201818754, rs373804633, rs199768740, and rs80358230) as their frequencies in PAR subpopulation exceed 4%, far more than the associated disease prevalence.

A notable example of an AR disorder is Woodhouse–Sakati syndrome [WSS (MIM:241080)], a disease characterized by hypogonadism and hair thinning that often progresses to alopecia totalis. Of the less than 100 individuals reported globally with the disease, 30 are from Middle Eastern families (Bohlega & Alkuraya, 1993). WSS is caused by biallelic pathogenic variants in the [*DCAF17* (MIM: 612515)] (previously known as *C2ORF37*) gene. We identified NM_025000.4(*DCAF17*):c.436delC (p.Ala147fs) as the sole pathogenic variant of this gene in 88 individuals, in heterozygous state (MAF = 0.007) (Supporting Information Data). Although all heterozygous individuals were found to be clinically asymptomatic, the alternate allele in these individuals is associated with the decreased levels of Insulin (p value = 2.9E−02; β = −0.225; Figure S10) which could explain diabetes mellitus being one of the characteristic clinical phenotypes in WSS. We also found that c.436delC is enriched (fisher exact test p = 7.57E−34; OR = 18.45) in one of the founder populations, QGP_PAR subcluster, this is consistent with a previous report that identified NM_025000.4(*DCAF17*):c.436delC (rs797045038) as a founder variant in the Qatari population (Ben-Omran et al., 2011). This variant has also been reported in the Kingdom of Saudi Arabia (Alazami et al., 2008), which has a large number of tribes sharing common and similar carrier frequency with Qatar's native population. HMC is hosting the national molecular diagnostic laboratories of Qatar and has identified to date 34 WSS patients and 64 heterozygous carriers. Data from both QGP and HMC laboratories indicate that the carrier frequency for WSS in the Qatari population is approx. 1 in 42 individuals (2.5%) with MAF of 1.25%, which is the highest reported in the world. Remarkably, the carrier frequency of NM_025000.4(*DCAF17*):c.436delC (p.Ala147fs) is 7x higher in Qatar than in the same tribe living in neighboring Saudi Arabia and has not yet been reported in population frequency databases, such as gnomAD and 1000 genomes or the 100K Genomes Project that includes patients with rare genetic diseases (Turnbull et al., 2018).

4 | DISCUSSION

Here we characterized a broad spectrum of genetic variation in the Qatari population, in total over 88 million variants (1.86% of novel variants per individual genome and 24.6 M novel variants in the whole data set). This large-scale study allowed us to identify five nonadmixed subgroups in QGP (*n*=6045) compared with three in the previous study Fakhro et al., 2016 (*n* = 1005) (Fakhro et al., 2016). We found a larger number of DM variants carried per individual which could be explained by incomplete penetrance, or the individual might carry them in a heterozygous state (Francioli et al., 2014; Xue et al., 2012). We described the distribution of genetic variation across the subclusters and found the majority of the novel variants to be cluster-specific. These data support records of high consanguinity and founder effect but also identify a previously unstudied component of the Middle Eastern population. In an earlier work, we have performed the first genome wide association studies of a list of 45 quantitative traits in 6047 individuals from the Qatari population. We have replicated many previously known loci and we identified 17 novel and Qatari-specific signals across the studied traits. We have also showed that European-derived polygenic scores has reduced predictive performance when applied to the Middle Eastern population of Qatar (Thareja et al., 2021).

Recently, we have reported a total of 60 pathogenic and likely pathogenic variants in 25 ACMG genes in 141 unique individuals (Elfatih et al., 2021) and several other efforts are currently under way to build the catalogs of predicted loss-of-function variants and Mendelian disorders mutations and to characterize the pharmacogenomic (Jithesh et al., 2022) and the cancer landscapes of the Qatari population (Saad et al., in press). Furthermore, using a combination of whole genomes and exome sequence data and clinical reports, we developed a microarray with Qatari-specific pathogenic variants that could be used to rapidly, accurately and at low cost, screen the Qatari population for pathogenic variants of newborns, premarital couples, and patients presenting to the clinic (Rodriguez-Flores, 2022).

Previous genetic studies in the Middle East region have assessed the genomic variations linked to health and diseases mostly limited to whole exome sequencing on relatively small sample size (AlSafar et al., 2019; Fattahi et al., 2019; John et al., 2018; Monies et al., 2019; Scott et al., 2016). Our QGP data have a key advantage over these studies since we are performing large-scale population sequencing using a whole genome approach. Although our work provided various insights into the genomic of the Middle East, we should address one limitation of our approach is that we are including only Qatari nationals in the first phase. To overcome this limitation, we are including long-term residents in our next freezes.

In conclusion, this first phase of the QGP constitutes the largest comprehensive analysis of whole genomes representative of tens of millions of Arabian Peninsula and Middle East inhabitants. Such genetic information is largely lacking in global databases (Easteal et al., 2020). Our next phases will focus on specific diseases relevant to the Qatari population's health burden—for example, cancer, diabetes, and rare diseases—while accelerating the ability to use the genome sequencing data into clinical implementation. We anticipate our data will represent a valuable resource to advance genetic studies in the Arab and neighboring Middle Eastern populations and will significantly boost the current efforts to improve our understanding of global patterns of human variations, human history, and genetic contributions to health and diseases in diverse populations (C. N. Rotimi & Adeyemo, 2021).

ACKNOWLEDGMENTS

The Qatar Genome Program (QGP) and Qatar Biobank (QBB) are both Research and Development entities within Qatar Foundation for Education, Science and Community Development. The authors are thankful for everyone who contributed to this endeavor including the QGP and QBB team members, in addition to our partners at Hamad Medical Corporation (HMC), Sidra Medicine and other national stakeholders. The authors would like to especially thank all participants in this study for their continuous support.

THE QATAR GENOME PROGRAM RESEARCH CONSORTIUM

Qatar Genome Project Management: Said I. Ismail[1], Wadha Al-Muftah[1], Radja Badji[1], Hamdi Mbarek[1], Dima Darwish[1], Tasnim Fadl[1], Heba Yasin[1], Maryem Ennaifar[1], Rania Abdellatif[1], Fatima Alkuwari[1], Muhammad Alvi[1], Yasser Al-Sarraj[1], Chadi Saad[1], Asmaa Althani[1,16]

Biobank and Sample Preparation: Eleni Fethnou[2], Fatima Qafoud[2], Eiman Alkhayat[2], Nahla Afifi[2]

Sequencing and Genotyping group: Sara Tomei[3], Wei Liu[3], Stephan Lorenz[3]

Applied Bioinformatics Core: Najeeb Syed[4], Hakeem Almabrazi[4], Fazulur Rehaman Vempalli[4], Ramzi Temanni[4]

Data Management, Advanced Applications and Computing Infrastructure groups: Tariq Abu Saqri[5], Mohammedhusen Khatib[5], Mehshad Hamza[5], Tariq Abu Zaid[5], Ahmed El Khouly[5], Tushar Pathare[5], Shafeeq Poolat[5], Shafqat Baig[5], Anwar Haque[5], Mohamed Jama[5], Rashid Al-Ali[5]

Genetic Variability group: Geethanjali Devadoss Gandhi[6,8], Senthil Selvaraj[6], Najeeb Syed[4], Xavier Estivill[6], Hamdi Mbarek[1]

Population Structure and Genome Reference group: Rozaimi Mohamad Razali[6], Juan Rodriguez-Flores[17], Elbay Aliyev[6], Haroon Naeem[6], Waleed Aamer[6], Andrew Clark[18], Khalid Fakhro[6], Younes Mokrab[6]

GWAS group: Gaurav Thareja[7]*, Yasser Al-Sarraj*[1,8], Aziz Belkadi[7], Maryam Almotawa[9], Karsten Suhre[7+], Omar Albagha[+8,15] (*equally contributed, + jointly supervised)

Mendelian Disorders group: Waleed Aamer[6], Alya Al-Kurbi[6], Aljazi Al-Maraghi[6], Geethanjali Devadoss Gandhi[6,8], Najeeb Syed[4], Khalid Fakhro[6]

Loss of Function group: Fatemeh Abbaszadeh[10]*, Ikhlak Ahmed[5]*, Najeeb Syed[4], Mohammad Abuhaliqa[10], Rashid Al Ali[5], Khalid Fakhro[6], Zafar Nawaz[10], Ajayeb Al Nabet Al Marri[10], Xavier Estivill[6], Puthen V. Jithesh[8], Ramin Badii[10] (*equally contributed)

Consortium Lead Principal Investigators: Omar Albagha[8,15], Souhaila Al-Khodor[11], Mashael

Alshafai[12], Ramin Badii[10], Lotfi Chouchane[13], Xavier Estivill[6], Khalid Fakhro[6], Hamdi Mbarek[1], Younes Mokrab[6], Puthen V. Jithesh[8], Karsten Suhre[7], Zohreh Tatari[14]

Affiliations

1. Qatar Genome Program, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha, Qatar.
2. Qatar Biobank for Medical Research, Qatar Foundation, Building 317, Hamad Medical City, Doha, Qatar.
3. Sidra Medicine, Integrated Genomics Services, Out-Patient Clinic, Doha, Qatar.
4. Sidra Medicine, Applied Bioinformatics Core - Integrated Genomics Services - Research Branch, Doha, Qatar.
5. Sidra Medicine, Biomedical Informatics – Research Branch, Doha, Qatar.
6. Sidra Medicine, Human Genetics Department, Doha, Qatar.

7. Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar.
8. College of Health and Life Sciences, Hamad Bin Khalifa University, Education City, Doha, Qatar.
9. Qatar Biomedical Research Institute (QBRI), Hamad Bin Khalifa University, Doha, Qatar.
10. Molecular Genetics Laboratory, Hamad Medical Corporation, Doha, Qatar.
11. Sidra Medicine, Maternal and Child Health Program, Doha Qatar.
12. College of Health Sciences, Qatar University, Doha, Qatar.
13. Departments of Genetic Medicine, Microbiology and Immunology, Weill Cornell Medicine-Qatar, Doha, Qatar.
14. Sidra Medicine, Clinical Research Center, Doha, Qatar.
15. Center of Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, UK.
16. Biomedical Research Center, Qatar University, Doha, Qatar.
17. Department of Genetic Medicine, Weill Cornell Medicine, New York, U.S.A.
18. Department of Molecular Biology and Genetics, Cornell University, New York, U.S.A.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

The informed consent given by the study participants does not cover posting of participant level phenotype and genotype data of Qatar Biobank/Qatar Genome Project in public databases. However, access to QBB/QGP data can be obtained through an established ISO-certified process by submitting a project request at https://www.qatarbiobank.org.qa/research/how-to-apply, under licence QF-QGP-RES-PUB-002 for the current study. QGP variants browser access procedure is described in Supporting Information Document 2.

ORCID

Hamdi Mbarek http://orcid.org/0000-0002-1108-0371
Geethanjali Devadoss Gandhi https://orcid.org/0000-0002-6607-1595
Senthil Selvaraj https://orcid.org/0000-0002-0185-6033
Wadha Al-Muftah https://orcid.org/0000-0003-2549-082X
Radja Badji https://orcid.org/0000-0002-0512-3312
Yasser Al-Sarraj https://orcid.org/0000-0002-1505-6360
Chadi Saad https://orcid.org/0000-0001-6963-9126
Dima Darwish https://orcid.org/0000-0001-6201-8925
Muhammad Alvi https://orcid.org/0000-0002-3399-9303
Tasnim Fadl https://orcid.org/0000-0002-3863-0900
Heba Yasin https://orcid.org/0000-0002-8692-3687
Fatima Alkuwari https://orcid.org/0000-0002-0516-6750
Rozaimi Razali https://orcid.org/0000-0002-8996-3975
Waleed Aamer https://orcid.org/0000-0002-1324-3509
Fatemeh Abbaszadeh https://orcid.org/0000-0002-0964-9100
Ikhlak Ahmed https://orcid.org/0000-0002-5753-9627
Younes Mokrab https://orcid.org/0000-0003-1611-6692
Karsten Suhre https://orcid.org/0000-0001-9638-3912
Omar Albagha https://orcid.org/0000-0001-5916-5983
Khalid Fakhro https://orcid.org/0000-0002-3150-1276
Ramin Badii https://orcid.org/0000-0003-3992-0560
Said I. Ismail https://orcid.org/0000-0002-8425-4010
Asma Althani https://orcid.org/0000-0003-2715-3712

REFERENCES

Abou Tayoun, A. N., & Rehm, H. L. (2020). Genetic variation in the Middle East—an opportunity to advance the human genetics field. *Genome Medicine*, *12*(1), 116. https://doi.org/10.1186/s13073-020-00821-7

Al Thani, A., Fthenou, E., Paparrodopoulos, S., Al Marri, A., Shi, Z., Qafoud, F., & Afifi, N. (2019). Qatar biobank cohort study: Study design and first results. *American Journal of Epidemiology*, *188*(8), 1420–1433. https://doi.org/10.1093/aje/kwz084

Alazami, A. M., Al-Saif, A., Al-Semari, A., Bohlega, S., Zlitni, S., Alzahrani, F., Bavi, P., Kaya, N., Colak, D., Khalak, H., Baltus, A., Peterlin, B., Danda, S., Bhatia, K. P., Schneider, S. A., Sakati, N., Walsh, C. A., Al-Mohanna, F., Meyer, B., & Alkuraya, F. S. (2008). Mutations in C2orf37, encoding a nucleolar protein, cause hypogonadism, alopecia, diabetes mellitus, mental retardation, and extrapyramidal syndrome. *American Journal of Human Genetics*, *83*(6), 684–691. https://doi.org/10.1016/j.ajhg.2008.10.018

Al-Gazali, L., Hamamy, H., & Al-Arrayad, S. (2006). Genetic disorders in the Arab world. *British Medical Journal*, *333*(Issue 7573), 831–834. BMJ https://doi.org/10.1136/bmj.38982.704931.AE

AlSafar, H. S., Al-Ali, M., Elbait, G. D., Al-Maini, M. H., Ruta, D., Peramo, B., Henschel, A., & Tay, G. K. (2019). Introducing the first whole genomes of nationals from the United Arab Emirates. *Scientific Reports*, *9*(1), 14725. https://doi.org/10.1038/s41598-019-50876-9

Anwar, W. A., Khyatti, M., & Hemminki, K. (2014). Consanguinity and genetic diseases in North Africa and immigrants to Europe. *European Journal of Public Health*, *24*(SUPPL. 1), 57–63. https://doi.org/10.1093/eurpub/cku104

Arauna, L. R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G., & Comas, D. (2017). Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Molecular Biology and Evolution*, *34*(2), 318–329. https://doi.org/10.1093/molbev/msw218

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., & Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Ben-Omran, T., Ali, R., Almureikhi, M., Alameer, S., Al-Saffar, M., Walsh, C. A., Felie, J. M., & Teebi, A. (2011). Phenotypic heterogeneity in Woodhouse-Sakati syndrome: Two new families with a mutation in the C2orf37 gene. *American Journal of Medical Genetics, Part A*, *155*(11), 2647–2653. https://doi.org/10.1002/ajmg.a.34219

Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: Why the uneven progress? *Journal of Community Genetics*, *8*(4), 255–266. https://doi.org/10.1007/s12687-017-0316-6

Bohlega, S. A., & Alkuraya, F. S. (1993). Woodhouse-Sakati Syndrome. In GeneReviews®. http://www.ncbi.nlm.nih.gov/pubmed/27489925

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., Van Dijk, F., Francioli, L. C., Hottenga, J. J., Laros, J. F., Li, Q., Li, Y., Cao, H., Chen, R., Du, Y., Li, N., … 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

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chiang, C. W. K., Marcus, J. H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., Pitzalis, M., Busonero, F., Maschio, A., Pistis, G., Steri, M., Angius, A., Lohmueller, K. E., Abecasis, G. R., Schlessinger, D., Cucca, F., & Novembre, J. (2018). Genomic history of the Sardinian population. *Nature Genetics*, *50*(10), 1426–1434. https://doi.org/10.1038/s41588-018-0215-8

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, *372*(9), 793–795. https://doi.org/10.1056/nejmp1500523

Easteal, S., Arkell, R. M., Balboa, R. F., Bellingham, S. A., Brown, A. D., Calma, T., Cook, M. C., Davis, M., Dawkins, H., Dinger, M. E., Dobbie, M. S., Farlow, A., Gwynne, K. G., Hermes, A., Hoy, W. E., Jenkins, M. R., Jiang, S. H., Kaplan, W., Leslie, S., … Baynam, G. (2020). Equitable expanded carrier screening needs indigenous clinical and population genomic data. *American Journal of Human Genetics*, *107*(2), 175–182. https://doi.org/10.1016/j.ajhg.2020.06.005

Elfatih, A., Mifsud, B., Syed, N., Badii, R., Mbarek, H., Abbaszadeh, F., & Estivill, X. (2021). Actionable genomic variants in 6045 participants from the Qatar Genome Program. *Human Mutation*. https://doi.org/10.1002/humu.24278

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., Al-Marri, A. A. N., Khalil, C. A., Al-Shakaki, A., Chidiac, O., Stadler, D., Zirie, M., Jayyousi, A., Salit, J., Mezey, J. G., Crystal, R. G., & Rodriguez-Flores, J. L. (2016). The Qatar genome: A population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, *3*, 16016. https://doi.org/10.1038/hgv.2016.16

Fattahi, Z., Beheshtian, M., Mohseni, M., Poustchi, H., Sellars, E., Nezhadi, S. H., Amini, A., Arzhangi, S., Jalalvand, K., Jamali, P., Mohammadi, Z., Davarnia, B., Nikuei, P., Oladnabi, M., Mohammadzadeh, A., Zohrehvand, E., Nejatizadeh, A., Shekari, M., Bagherzadeh, M., … Najmabadi, H. (2019). Iranome: A catalog of genomic variations in the Iranian population. *Human Mutation*, *40*(11), 1968–1984. https://doi.org/10.1002/humu.23880

Francioli, L. C., Menelaou, A., Pulit, S. L., Van Dijk, F., Palamara, P. F., Elbers, C. C., Neerincx, P. B. T., Ye, K., Guryev, V., Kloosterman, W. P., Deelen, P., Abdellaoui, A., Van Leeuwen, E. M., Van Oven, M., Vermaat, M., Li, M., Laros, J. F. J., Karssen, L. C., Kanterakis, A., & Wijmenga, C. (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

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., & Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Gudbjartsson, D. F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S. A., Zink, F., Oddson, A., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Kong, A., Helgason, A., Masson, G., Magnusson, O. T., Thorsteinsdottir, U., & Stefansson, K. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. *Scientific Data*, *2*, 150011. https://doi.org/10.1038/sdata.2015.11

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., Mathieson, I., Ekoru, K., DeGorter, M. K., Nsubuga, R. N., Finan, C., Wheeler, E., Chen, L., Cooper, D. N., Schiffels, S., & Sandhu, M. S. (2019). Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell*, *179*(4), 984–1002. https://doi.org/10.1016/j.cell.2019.10.004

Harkness, G., & Khaled, R. (2014). Modern traditionalism: Consanguineous marriage in Qatar. *Journal of Marriage and Family*, *76*(3), 587–603. https://doi.org/10.1111/jomf.12106

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Hunter-Zinck, H., Musharoff, S., Salit, J., Al-Ali, K. A., Chouchane, L., Gohar, A., Matthews, R., Butler, M. W., Fuller, J., Hackett, N. R., Crystal, R. G., & Clark, A. G. (2010). Population genetic structure of the people of Qatar. *American Journal of Human Genetics*, *87*(1), 17–25. https://doi.org/10.1016/j.ajhg.2010.05.018

Jithesh, P. V., Abuhaliqa, M., Syed, S., Ahmed, I., El Anbari, M., Bastaki, K., Sherif, S., Umlai, U., Jan, Z., Ghandi, G., Manickam, C., Selvaraj, S., George, C., Bangarusamy, D., Abdel-latif, R., Al-Shafai, M., Tatari-Calderone, Z., Estivill, X., & Pirmohamed, M., The Qatar Genome Program Research Consortium. (2022). A population study of clinically actionable genetic variationaffecting drug response from the Middle East. *NPJ Genomic Medicine*. (In Press). https://doi.org/10.1038/s41525-022-00281-5

John, S. E., Antony, D., Eaaswarkhanth, M., Hebbar, P., Channanath, A. M., Thomas, D., Devarajan, S., Tuomilehto, J., Al-Mulla, F., Alsmadi, O., & Thanaraj, T. A. (2018). Assessment of coding region variants in Kuwaiti population: Implications for medical genetics and population genomics. *Scientific Reports*, *8*(1), 16583. https://doi.org/10.1038/s41598-018-34815-8

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, *37*(5), 780–785. https://doi.org/10.1377/hlthaff.2017.1595

Manolio, T. A., Bult, C. J., Chisholm, R. L., Deverka, P. A., Ginsburg, G. S., Jarvik, G. P., McLeod, H. L., Mensah, G. A., Relling, M. V., Roden, D. M., Rowley, R., Tamburro, C., Williams, M. S., & Green, E. D. (2019). Genomic medicine year in review: 2019. *American Journal of Human Genetics*, *105*(6), 1072–1075. https://doi.org/10.1016/j.ajhg.2019.11.006

Mills, M. C., & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, *2*(Issue 1), 9. https://doi.org/10.1038/s42003-018-0261-x

Monies, D., Abouelhoda, M., Assoum, M., Moghrabi, N., Rafiullah, R., Almontashiri, N., Alowain, M., Alzaidan, H., Alsayed, M., Subhani, S., Cupler, E., Faden, M., Alhashem, A., Qari, A., Chedrawi, A., Aldhalaan, H., Kurdi, W., Khan, S., Rahbeeni, Z., & Alkuraya, F. S. (2019). Erratum: Lessons learned from large-Scale, first-tier clinical exome sequencing in a highly consanguineous population. *American Journal of Human Genetics*, *104*, 1182–1201. https://doi.org/10.1016/j.ajhg.2019.09.019

Naslavsky, M. S., Scliar, M. O., Yamamoto, G. L., Wang, J. Y. T., Zverinova, S., Karp, T., Nunes, K., Ceroni, J. R. M., de Carvalho, D. L., da Silva Simões, C. E., Bozoklian, D., Nonaka, R., Silva, N., dos, S. B., Souza, A., da, S., Andrade, H., de, S., Passos, M. R. S., … Zatz, M. (2020). Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis

(São Paulo, Brazil). *bioRxiv*, *10*, 24. https://doi.org/10.1101/2020.09.15.298026

Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., Suematsu, M., Matsuda, K., Yamamoto, K., Kubo, M., Hirose, N., & Kamatani, Y. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature Communications*, *9*(1), 1631. https://doi.org/10.1038/s41467-018-03274-0

Razali, R. M., Rodriguez-Flores, J., Ghorbani, M., Naeem, H., Aamer, W., Aliyev, E., Jubran, A., Qatar Genome Program Research Consortium, Clark, A. G., Fakhro, K. A., & Mokrab, Y. (2021). Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nature Communications*, *12*(1), 5929. https://doi.org/10.1038/s41467-021-25287-y

Rodriguez-Flores, J. L., Fakhro, K., Agosto-Perez, F., Ramstetter, M. D., Arbiza, L., Vincent, T. L., Robay, A., Malek, J. A., Suhre, K., Chouchane, L., Badii, R., Al-Marri, A. A. N., Khalil, C. A., Zirie, M., Jayyousi, A., Salit, J., Keinan, A., Clark, A. G., Crystal, R. G., & Mezey, J. G. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Research*, *26*(2), 151–162. https://doi.org/10.1101/gr.191478.115

Rodriguez-Flores, J. L., Fakhro, K., Hackett, N. R., Salit, J., Fuller, J., Agosto-Perez, F., Gharbiah, M., Malek, J. A., Zirie, M., Jayyousi, A., Badii, R., Al-Nabet Al-Marri, A., Chouchane, L., Stadler, D. J., Mezey, J. G., & Crystal, R. G. (2014). Exome Sequencing identifies potential risk variants for mendelian disorders at high prevalence in Qatar. *Human Mutation*, *35*(1), 105–116. https://doi.org/10.1002/humu.22460

Rodriguez-Flores, J. L., Messai-Badji, R., Robay, A., Temanni, R., Syed, N., Markovic, M., Al-Khayat, E., Qafoud, F., Nawaz, Z., Badii, R., Al-Sarraj, Y., Mbarek, H., Al-Muftah, W., Alvi, M., Rostami, M. R., Cruzado, J., Mezey, J. G., Shakaki, A. A., Malek, J. A., … Crystal, R. G. (2022). The QChip1 knowledge base and micro array for precision medicine in Qatar. *NPJ Genomic Medicine*. *7*(1), 3. https://doi.org/10.1038/s41525-021-00270-0

Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., Adebiyi, E., Ademola, A. D., Adeyemo, A., Adu, D., Affolabi, D., Agongo, G., Ajayi, S., Akarolo-Anthony, S., Akinyemi, R., Akpalu, A., Alberts, M., Alonso Betancourt, O., Alzohairy, A. M., Ameni, G., & Zar, H. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science*, *344*(6190), 1346–1348. https://doi.org/10.1126/science.1251546

Rotimi, C. N., & Adeyemo, A. A. (2021). From one human genome to a complex tapestry of ancestry. *Nature*, *590*(7845), 220–221. https://doi.org/10.1038/d41586-021-00237-2

Saad, M., Mokrab, Y., Halabi, N., Shan, J., Razali, R., Kunji, K., Syed, N., Temanni, R., Subramanian, M., Ceccrelli, M., the Qatar Genome Program Research Consortium, Tabrizi, AR., Bedognetti, D., & Chouchane, L. (In Press). Genetic predisposition to cancer across people of different ancestries in Qatar: A population-based, cohort study. Lancet Oncology.

Scott, E. M., Halees, A., Itan, Y., Spencer, E. G., He, Y., Azab, M. A., Gabriel, S. B., Belkadi, A., Boisson, B., Abel, L., Clark, A. G., Rahim, S. A., Abdel-Hadi, S., Abdel-Salam, G., Abdel-Salam, E., Abdou, M., Abhytankar, A., Adimi, P., Ahmad, J., & Zhang, S. Y. (2016). Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. *Nature Genetics*, *48*(9), 1071–1079. https://doi.org/10.1038/ng.3592

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfied, M. J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., Boughtwood, T., Braithwaite, J., Goodhand, P., Birney, E., & North, K. N. (2019). Integrating genomics into healthcare: A global responsibility, *American Journal of Human Genetics* (104, pp. 13–20). Cell Press Issue 1 https://doi.org/10.1016/j.ajhg.2018.11.014

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., & Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Thareja, G., Al-Sarraj, Y., Belkadi, A., Almotawa, M., Ismail, S., Al-Muftah, W., Badji, R., Mbarek, H., Darwish, D., Fadl, T., Yasin, H., Ennaifar, M., Abdellatif, R., Alkuwari, F., Alvi, M., Al-Sarraj, Y., Saad, C., Althani, A., Fethnou, E., & Albagha, O. M. E. (2021). Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nature Communications*, *12*(1), 1250. https://doi.org/10.1038/s41467-021-21381-3

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereaux, A., Smith, K., Martin, A. R., Sosinsky, A., McDonagh, E. M., Sultana, R., & Caulfield, M. J. (2018). The 100 000 genomes project: Bringing whole genome sequencing to the NHS. *BMJ (Online)*, *361*. https://doi.org/10.1136/bmj.k1687

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H. L., Sanchis-Juan, A., Frontini, M., Thys, C., Stephens, J., Mapeta, R., Burren, O. S., Downes, K., Haimel, M., Tuna, S., Deevi, S., Aitman, T. J., Bennett, D. L., … Ouwehand, W. H. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, *583*(7814), 96–102. https://doi.org/10.1038/s41586-020-2434-2

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., Soon, W. W. J., Bertin, N., Lin, C. B., Khor, C. C., DeGiorgio, M., Cheng, S., Bao, L., Karnani, N., Hwang, W. Y. K., Davila, S., Tan, P., Shabbir, A., Moh, A., & Wang, C. (2019). Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*, *179*(3), 736–749. https://doi.org/10.1016/j.cell.2019.09.019

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., & Tyler-Smith, C. (2012). Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics*, *91*(6), 1022–1032. https://doi.org/10.1016/j.ajhg.2012.10.015

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Mbarek, H., Devadoss Gandhi, G., Selvaraj, S., Al-Muftah, W., Badji, R., Al-Sarraj, Y., Saad, C., Darwish, D., Alvi, M., Fadl, T., Yasin, H., Alkuwari, F., Razali, R., Aamer, W., Abbaszadeh, F., Ahmed, I., Mokrab, Y., Suhre, K., & Albagha, O. (2022). Qatar genome: Insights on genomics from the Middle East. *Human Mutation*, 43, 499–510. https://doi.org/10.1002/humu.24336