PLOS ONE

RESEARCH ARTICLE

# DALIA- a comprehensive resource of Disease Alleles in Arab population

**Aastha Vatsyayan[1,2☯‡], Parul Sharma[3☯‡], Shrey Gupta[3], Sumiti Sandhu[4], Seetha Lakshmi Venu[5], Vandana Sharma[6], Bouabid Badaoui[7], Kaidi Azedine[7], Serti Youssef[7], Anna Rajab[8], Alaaeldin Fayez[9], Seema Madinur[10], Anop Ranawat[1], Kavita Pandhare[1], Srinivasan Ramachandran[1,11], Sridhar Sivasubbu[1,11], Vinod Scaria[1,11]****

**1** CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India, **2** Amity University, Noida, India, **3** Indraprastha Institute of Information Technology Delhi (IIIT-D), Delhi, India, **4** Amity University, Rajasthan, Jaipur, India, **5** ACBR, Delhi, India, **6** Kalindi College, University of Delhi, New Delhi, India, **7** Mohammed V University, Rabat, Morocco, **8** National Genetic Center, Ministry of Health, Muscat, Oman, **9** Human Genetics and Genome Research Division, Center of Excellence for Human Genetics, National Research Center, Cairo, Egypt, **10** Independent Researcher, India, **11** Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

☯ These authors contributed equally to this work.
‡ These authors share first authorship on this work.
* vinods@igib.in

## Abstract

The Arab population encompasses over 420 million people characterized by genetic admixture and a consequent rich genetic diversity. A number of genetic diseases have been reported for the first time from the population. Additionally a high prevalence of some genetic diseases including autosomal recessive disorders such as hemoglobinopathies and familial mediterranean fever have been found in the population and across the region. There is a paucity of databases cataloguing genetic variants of clinical relevance from the population. The availability of such a catalog could have implications in precise diagnosis, genetic epidemiology and prevention of disease. To fill in the gap, we have compiled DALIA, a comprehensive compendium of genetic variants reported in literature and implicated in genetic diseases reported from the Arab population. The database aims to act as an effective resource for population-scale and sub-population specific variant analyses, enabling a ready reference aiding clinical interpretation of genetic variants, genetic epidemiology, as well as facilitating rapid screening and a quick reference for evaluating evidence on genetic diseases.

## Introduction

The Arab population encompasses twenty three Arab States with a total population of around 420 million individuals and roughly encompassing 5% of the world population [1, 2]. The rich genetic diversity in the Arab population has been contributed by a genetic admixture resulting from waves of migration, to form one of the largest geocultural units in the world [3]. A number of genetic diseases are prevalent in the Arab population. These include

OPEN ACCESS

**Citation:** Vatsyayan A, Sharma P, Gupta S, Sandhu S, Venu SL, Sharma V, et al. (2021) DALIA- a comprehensive resource of Disease Alleles in Arab population. PLoS ONE 16(1): e0244567. https://doi.org/10.1371/journal.pone.0244567

**Editor:** Obul Reddy Bandapalli, German Cancer Research Center (DKFZ), GERMANY

**Received:** August 27, 2020

**Accepted:** December 11, 2020

**Published:** January 13, 2021

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0244567

**Data Availability Statement:** The data used in the analysis is accessible at: https://clingen.igib.res.in/dalia/DALIA_Table.xlsx.

**Funding:** The work was funded by the Council of Scientific and Industrial Research, India [RareGen

Grant]. The funding body had no role in the preparation of the manuscript or decision to publish. AV acknowledges a Senior Research Fellowship from the Indian Council for Medical Research.

**Competing interests:** The authors have declared that no competing interests exist.

Hemoglobinopathies such as sickle cell disease (HbS) as well as thalassemias, Familial Mediterranean Fever, Osteopetrosis Syndromes, glucose-6-phosphate dehydrogenase (G6PD) deficiency and several metabolic disorders [4]. The prevalence of genetic disorders in the Arab population has been extensively reviewed by a number of authors [4–8]. It is widely believed that datasets with well-curated and annotated genetic datasets are urgently required to establish disease epidemiology, especially in genetically isolated subpopulations which bear a particularly heavy disease burden [4].

A few efforts in the past have attempted to systematically catalog the genetic diseases in the Arab population. The first and foremost major initiative was the Arab Genetic Disease Database, put together by Teebi and colleagues [9]. This was followed by a more systematic effort in later years, resulting in the creation of the Catalogue of Transmission Genetics in Arabs (CTGA), a comprehensive resource mapping genetic diseases in Arab population. The database created by the Centre for Arab Genomic Studies (CAGS) [10] provides information about phenotypes and related genes. In addition, a few resources have emerged that have collected the genetic variants in Arab populations. These include the Greater Middle East (GME) initiative [10, 11], which sequenced multiple individuals in the region to create a comprehensive variome resource. In addition, the Qatar genome programme (About Qatar Genome. Retrieved from https://qatargenome.org.qa/node/5) and a dataset of over 1005 Qatari exomes and genomes [12] aim to provide an overview of genetic diversity in the region. More recently, a genome-scale database, al mena [13] was created by our group, integrating datasets to provide a comprehensive allele frequency resource for the population, apart from providing allele frequencies of variants associated with diseases. A few more population-scale initiatives are underway which include the Saudi Human Genome Project [14] and the 1000 Arab Genome Project [14, 15].

With the increasing adoption of sequencing in clinical settings, a number of new genetic variants and diseases are being reported from the region. Nevertheless there has been no structured effort to curate these genetic variants in a systematic format to allow comparison and enable genetic epidemiology analyses. Our previous studies demonstrate how a systematic, manually curated resource of genetic variants could enable the establishment of genetic epidemiology of disease causing variants in the population, with implications in better diagnosis of the disease [13, 14, 16].

To fill in the gap, we have created a comprehensive manually curated resource of genetic variants in the Arab population. This database, DALIA—for Disease Alleles in Arabs provides a ready reference to genetic variants published from Arab populations for clinicians, patients, as well as researchers. In addition, the resource also serves as a resource for genetic epidemiology.

## Material and methods

### Literature coverage and curation

A list of relevant publications was retrieved using pubmed.mineR [17] tool using country names to query for publications describing "mutation", "variant" or "polymorphism". An exhaustive list of PubMed IDs was retrieved for each of the 23 countries which have a significant number of Arabs. These include 22 countries which were part of the Arab League and speak Arab- Algeria, Bahrain, the Comoros Islands, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Mauritania, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, the United Arab Emirates, Yemen apart from Israel which has a significant number of Arabs [18–23].

Each of the full-text articles were retrieved and manually curated, to include an extensive array of information, including country of origin and ethnicity apart from the variant type, the methods used for the annotation of variants etc. Special emphasis was made on including only papers that described variants in the countries, and not just reported them.

Each of the variants were further normalised to the GRCh37/hg19 build version of the Human genome for the genomic location and variant position, as well as according to the Human Genome Variation Society (HGVS) nomenclature for the representation of the variants. Gene names were similarly normalised to the Human Gene Nomenclature Committee (HGNC) nomenclature using Mutalyzer tool [24] which check for consistency for the normalised variants. Wherever applicable, dbSNP IDs and ClinVar IDs were added for variants mapping to respective databases. The disease names were also normalised according to the annotations in the Online Mendelian Inheritance in Man (OMIM) [25–27] database and linked wherever they could be consistently mapped.

For each of the variants, additional information on the zygosity of variant, and the method or assay used for identification of the variant were also collected. To ease the annotation, a prefilled list of techniques was used as a drop-down. The entire activity was performed on a template available to all annotators through Google docs online as well as offline through templates in Microsoft Excel. The annotators individually filled in variants for each country. Along with the template, the annotators were also given tutorial slides to train themselves on the curation guidelines so as to maintain uniformity across all curators.

## Quality control and compilation of annotations

Each of the annotation sheets corresponding to the variants from each annotating group were compiled. Each of these sets were then independently cross-checked by a different team member to eliminate manual errors. The corrected sets were compiled into the master sheet. Each of the entries were verified for (i) accuracy and consistency of annotations as per the HGVS nomenclature for variants. (ii) accuracy of the variant position and reference and alternate genotypes on the hg19 genome build and (iii) uniformity and consistency of annotations, using the Mutalyzer tool.

## Allele frequencies and genetic epidemiology

The allele and genotype frequencies of the variants were systematically compiled for the global as well as Middle Eastern populations. The choice of Middle Eastern datasets was based on the fact that no specific datasets are publicly available representative of the Arab populations and the dataset from the region could at least in part provide insights into the prevalence of variants in the region. The global allele frequencies were derived from the ExAC [28] as well as 1000 Genomes [29] databases, while the allele frequencies in Middle Eastern populations were compiled from the al mena database. The al mena database encompasses data from the Qatar subpopulations representing African, Arabian, Bedouin, Persian and South Asian ancestry subgroups as well as the data from the subpopulations which were represented in the Greater Middle East (GME) study [11] representing Asian Peninsula, Central Asian, Israel, Northeast Africa, Northwest Africa, Syrian Desert and Turkish Peninsula. The comparisons were tested for significance using Fisher's Exact Test. Comparisons were done for each subpopulation vis-a-vis the population average and with the global frequencies as derived from ExAC and 1000 Genomes databases.

## Analysis of genes under natural selection

We used two different metrics, the Integrated Haplotype Score (iHS) and Fixation index statistic (Fst) to identify genes under natural selection. The top 1% genes sorted by the iHS scores in

Qatar population were searched for the genes in the database. Pairwise Fst scores [30] were then computed for pairs of Qatari population (QALL) and the African (QAFR), Arab (QARB), Bedouin (QBED), Persian (QPER), and South Asian (QSAS) subpopulations.

### Database and web interface design

The database was developed using MongoDB, a popular open-source NOSQL database system in view of the flexibility offered and the scalability for population-scale variant curation. The web server was configured in Apache 2.4.12, and the interface was coded in AngularJS and PHP.

## Results & discussion

### Data compilation

We compiled a total of 3577 genetic variants from 368 genes and associated with 1984 diseases from over 1113 publications originating from 23 countries. Of these variants, a total of 2790 variants reported were unique, out of which a total of 2679 variants (96%) were exonic, while the remaining included 110 non-exonic (26 3'UTR, 24 intronic, 17 splicing, 12 UTR5, 11 upstream, 1 ncRNA intronic, and 1 a downstream variants).

Apart from the genetic context, associated information on the gene, disease, population from which the variant was reported, and a spectrum of computational scores predicting pathogenicity of the variants, including SIFT [31] Polyphen [27] and CADD [25] were also compiled for each of the variants. Allele frequencies of the variants across population-scale datasets like gnomAD [32] along with linkouts to other relevant databases including OMIM, ClinVar [33] and dbSNP [34] were also compiled systematically.

### Database features

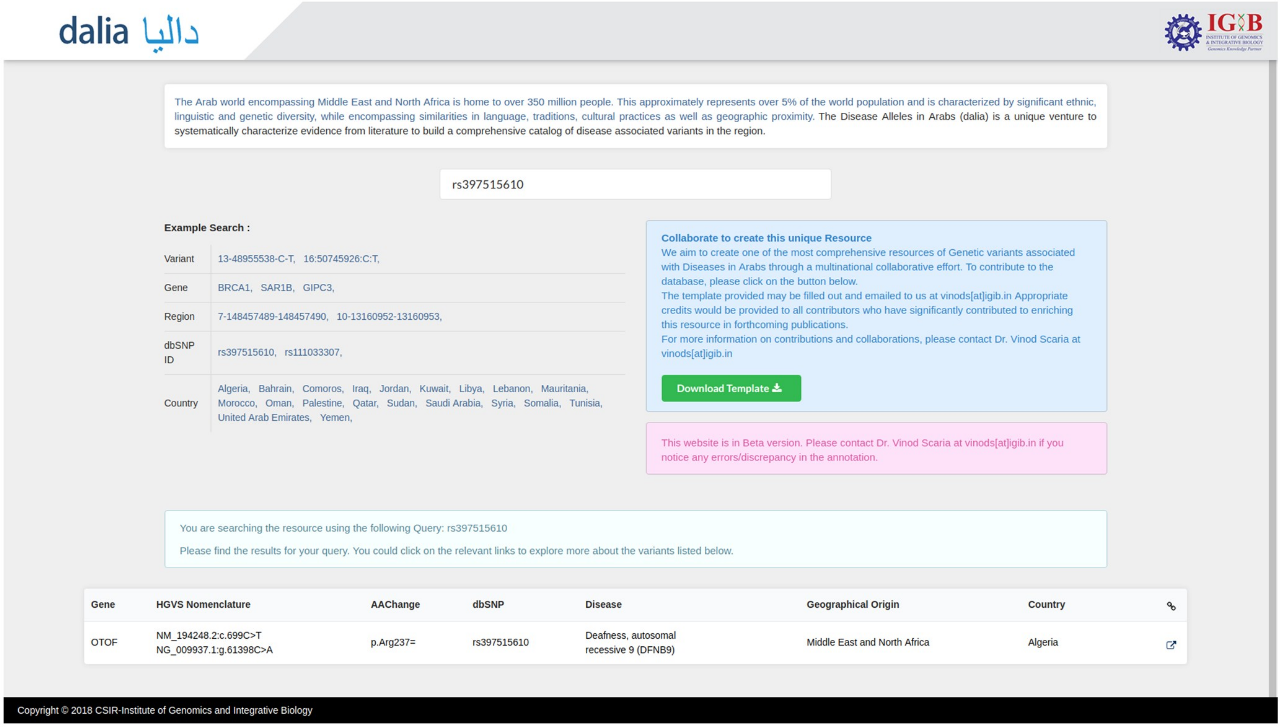
The database is designed to have a user-friendly interface to the data compilation. The prospective user can search the resource in multiple query formats, including genomic location, rsID or gene name. Population-specific variants can also be viewed by searching the name of the country. A complete list of example formats is available on the homepage (Fig 1). A comprehensive result containing details regarding the gene, disease caused, ethnic and geographic details of the patient, along with various allele frequencies as well as variant annotation scores have been linked to each search query (Fig 2).

### Comparison with other databases

The variants in DALIA were compared with variants in ClinVar and the HGMD-Public version to see whether the variants covered in the present compilation are also similarly covered in other databases. Comparison of the 2790 unique genetic variants in dalia with Clinvar revealed that 2074 variants were shared by ClinVar. A similar comparison of variants with HGMD revealed 2197 variants were shared with the Public version of HGMD.

### Allele frequency analysis

We performed a comparison of the allele frequencies of the variants across the population scale datasets of allele frequencies from 1000 Genomes, ExAC, Greater Middle Eastern, and Qatar. The former two represented global populations while the latter two represented population data from the middle east. The Qatar and GME allele frequencies were retrieved from the al mena database. Of the variants, a total of 564 variants had a reported allele frequency in at least one of the four population scale datasets considered, while 255 had frequencies across all the four datasets



The Arab world encompassing Middle East and North Africa is home to over 350 million people. This approximately represents over 5% of the world population and is characterized by significant ethnic, linguistic and genetic diversity, while encompassing similarities in language, traditions, cultural practices as well as geographic proximity. The Disease Alleles in Arabs (dalia) is a unique venture to systematically characterize evidence from literature to build a comprehensive catalog of disease associated variants in the region.

rs397515610

Example Search :

| | |
|---|---|
| Variant | 13-48955538-C-T, 16:50745926:C:T, |
| Gene | BRCA1, SAR1B, GIPC3, |
| Region | 7-148457489-148457490, 10-13160952-13160953, |
| dbSNP ID | rs397515610, rs111033307, |
| Country | Algeria, Bahrain, Comoros, Iraq, Jordan, Kuwait, Libya, Lebanon, Mauritania, Morocco, Oman, Palestine, Qatar, Sudan, Saudi Arabia, Syria, Somalia, Tunisia, United Arab Emirates, Yemen, |

Collaborate to create this unique Resource
We aim to create one of the most comprehensive resources of Genetic variants associated with Diseases in Arabs through a multinational collaborative effort. To contribute to the database, please click on the button below.
The template provided may be filled out and emailed to us at vinods[at]igib.in Appropriate credits would be provided to all contributors who have significantly contributed to enriching this resource in forthcoming publications.
For more information on contributions and collaborations, please contact Dr. Vinod Scaria at vinods[at]igib.in

This website is in Beta version. Please contact Dr. Vinod Scaria at vinods[at]igib.in if you notice any errors/discrepancy in the annotation.

You are searching the resource using the following Query: rs397515610

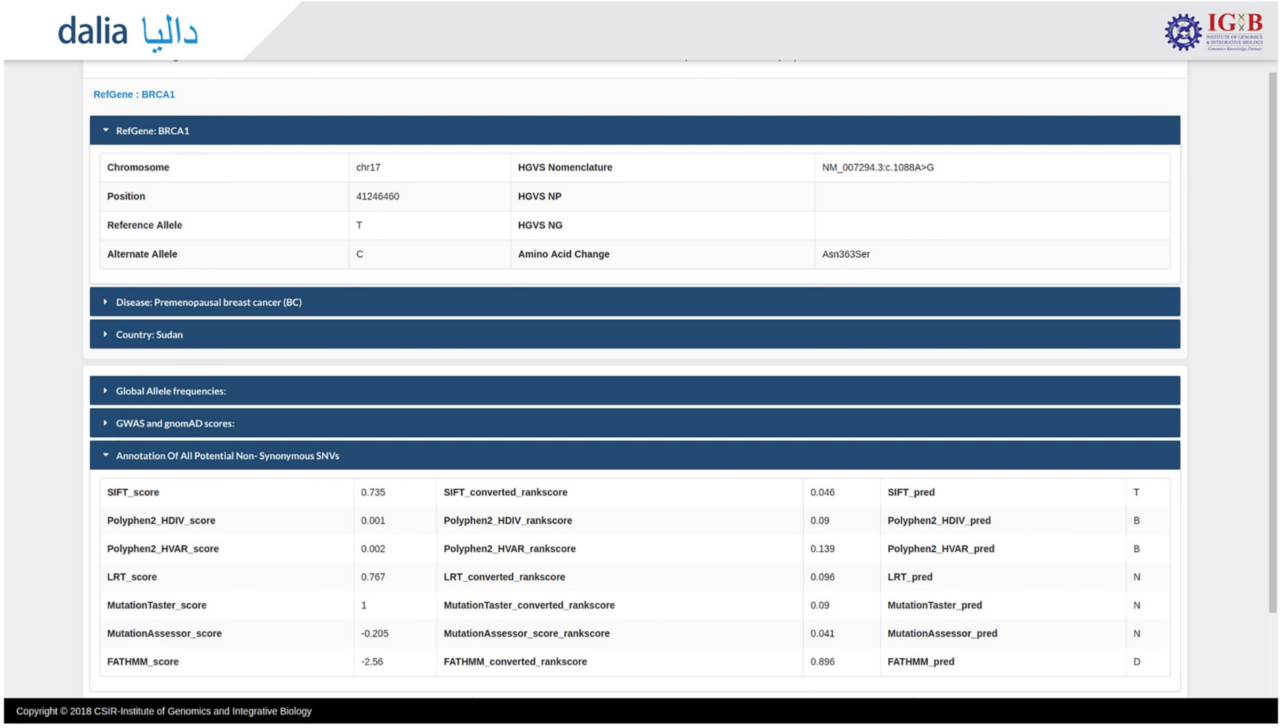
Please find the results for your query. You could click on the relevant links to explore more about the variants listed below.

| Gene | HGVS Nomenclature | AAChange | dbSNP | Disease | Geographical Origin | Country | |
|---|---|---|---|---|---|---|---|
| OTOF | NM_194248.2:c.699C>T<br>NG_009937.1:g.61398C>A | p.Arg237= | rs397515610 | Deafness, autosomal recessive 9 (DFNB9) | Middle East and North Africa | Algeria | |

Copyright © 2018 CSIR-Institute of Genomics and Integrative Biology

**Fig 1. The DALIA homepage.** It provides an easy to use interface for searching and retrieving the compiled information on genetic variants. The 'Example Search' section illustrates numerous query formats in which the database can be explored.

https://doi.org/10.1371/journal.pone.0244567.g001



RefGene : BRCA1

▾ RefGene: BRCA1

| | | | |
|---|---|---|---|
| Chromosome | chr17 | HGVS Nomenclature | NM_007294.3:c.1088A>G |
| Position | 41246460 | HGVS NP | |
| Reference Allele | T | HGVS NG | |
| Alternate Allele | C | Amino Acid Change | Asn363Ser |

▸ Disease: Premenopausal breast cancer (BC)

▸ Country: Sudan

▸ Global Allele frequencies:

▸ GWAS and gnomAD scores:

▾ Annotation Of All Potential Non- Synonymous SNVs

| | | | | | |
|---|---|---|---|---|---|
| SIFT_score | 0.735 | SIFT_converted_rankscore | 0.046 | SIFT_pred | T |
| Polyphen2_HDIV_score | 0.001 | Polyphen2_HDIV_rankscore | 0.09 | Polyphen2_HDIV_pred | B |
| Polyphen2_HVAR_score | 0.002 | Polyphen2_HVAR_rankscore | 0.139 | Polyphen2_HVAR_pred | B |
| LRT_score | 0.767 | LRT_converted_rankscore | 0.096 | LRT_pred | N |
| MutationTaster_score | 1 | MutationTaster_converted_rankscore | 0.09 | MutationTaster_pred | N |
| MutationAssessor_score | -0.205 | MutationAssessor_score_rankscore | 0.041 | MutationAssessor_pred | N |
| FATHMM_score | -2.56 | FATHMM_converted_rankscore | 0.896 | FATHMM_pred | D |

Copyright © 2018 CSIR-Institute of Genomics and Integrative Biology

**Fig 2. Expanded search results.** Each selected search result links to a page containing complete information about the variant, including information about the gene, disease, genome level information as well as allele frequencies and various predictive scores that could help establish pathogenicity. Further, links to various other resources including Pubmed, Clinvar, dbSNP and OMIM are also included for a more complete picture of the variant.

https://doi.org/10.1371/journal.pone.0244567.g002

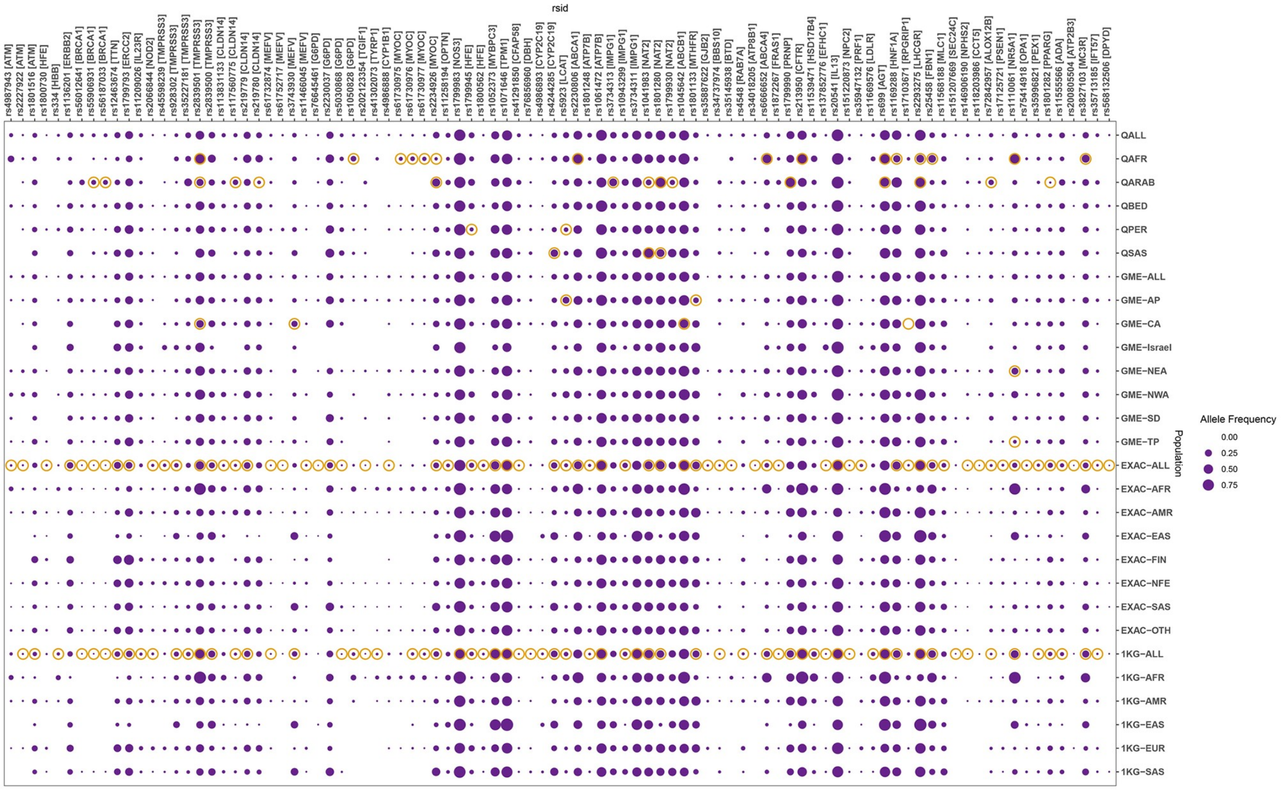


**Fig 3. Comparison of allele frequencies.** The figure represents a subset (p-value < 0.01) of all statistically significant variants obtained after applying the Fisher's Exact Test (S1 Fig). The orange circles highlight variants that were significantly different either within the Qatar or GME subpopulations, or with respect to global population averages ExAC and 1000 Genomes. The populations shown are, in order: Qatari population (QALL) and sub-populations African (AFR), Arab (QARB), Bedouin (QBED), Persian (QPER), and South Asian (QSAS), GME population (GME-ALL), and sub-populations northwest Africa (GME-NWA), northeast Africa (GME-NEA), Arabian peninsula (GME-AP), Israel (GME-Israel), Syrian desert (GME-SD), Turkish peninsula (GME-TP) and Central Asia (GME-CA), ExAC population (EXAC-ALL), sub-populations African/African American (EXAC-AFR), Latino (EXAC-AMR), East Asian (EXAC-EAS), Finnish (EXAC-FIN), Non-Finnish European (EXAC-NFE), South Asian (EXAC-SAS) and Other (EXAC-OTH), 1000 Genomes population(1KG-ALL) and sub-populations African(1KG-AFR), Ad Mixed American(1KG-AMR), East Asian(1KG-EAS), European(1KG-EUR) and South Asian(1KG-SAS).

https://doi.org/10.1371/journal.pone.0244567.g003

considered. Statistical analyses were performed for the allele frequencies (S1 Fig) and analysis revealed a total of 142 out of the 255 variants considered had a significant difference with global populations (p-value < 0.01). The variants and their frequencies are summarised in Fig 3.

## Genes and variants under natural selection

Since the individual-level genotypes were available only for the Qatar population, we analysed whether any of the genes harbouring the genetic variants showed signals of natural selection. Briefly, after retrieving the genes in the top 1% of |iHS| scores which encompass 1368 genes, we obtained a total of 10 genes which overlapped with the genes in our compendium. Pairwise Fst scores were further computed for each of the variants in the genes. The 10 genes in the DALIA database had a total of 12 genetic variants reported from the Arab population. The genes and annotations are summarised in S1 Table. Of specific mention would be the ATP7B gene which is associated with Wilson's disease. A total of 37 variants in the ATP7B gene associated with Wilson disease were reported in the DALIA database. The variants had an allele frequency ranging from 0 across QAFR, QARAB, QPER AND QSAS to 0.777 in the QBED Qatar sub-population, and from 0 across GME-CA, GME-NWA and GME-TP subpopulations, to

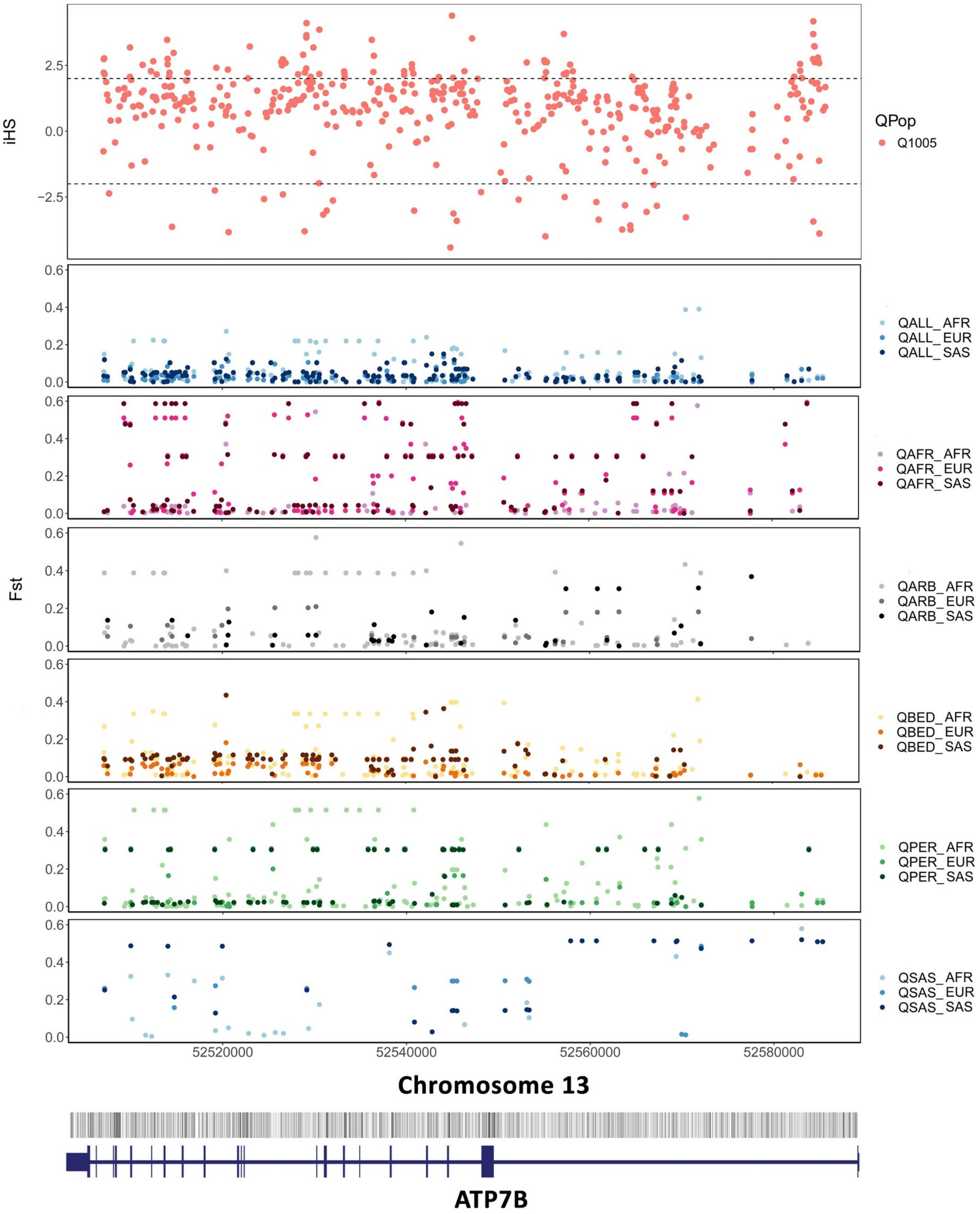


**Fig 4. Signatures of positive selection in the ATP7B gene in the Qatar population.** The figure depicts the iHS scores (top) for all known variants of the ATP7B gene plotted along the length of the gene loci shown below. The subsequent boxes depict all the pairwise Fst scores for the Qatar population as a whole (QALL), as well as all its subpopulations (QAFR, QARAB, QBED, QPER, and QSAS). The array of lines at the bottom represents all known variants along the length of the gene, the exon structure of which is shown beneath it.

https://doi.org/10.1371/journal.pone.0244567.g004

**Table 1. G6PD variants in the DALIA database.**

| Variant | HGVS nomenclature | ACMG Classification | Evidence | ExAC | Q1005 | GME | 1000 Genomes |
|---|---|---|---|---|---|---|---|
| rs5030868 | NM_001042351.1: c.563C>T | Pathogenic | PP3, BS2, PP5, PS3, PP1-M, PM1 | EXAC-ALL: 0.003<br>EXAC-AMR: 0.001<br>EXAC-EAS: 0<br>EXAC-FIN: 0<br>EXAC-NFE:0.001<br>EXAC-OTH: 0.011<br>EXAC-SAS: 0.016 | QALL: 0.042 QAFR: 0.061 QARAB: 0.004<br>QBED: 0.032 QPER: 0.079 QSAS: 0.122 | GME-ALL: 0.024<br>GME-AP: 0.077<br>GME-CA:0.042<br>GME-Israel:0<br>GME-NEA: 0.008<br>GME-NWA:0<br>GME-SD:0.018<br>GME-TP:0.008 | 1KG-ALL: 0.001<br>1KG-AFR: 0<br>1KG-AMR: 0<br>1KG-EAS: 0<br>1KG-EUR: 0<br>1KG-SAS: 0.004 |
| rs76645461 | NM_000402.3: c.233T>C | Likely pathogenic | PP3, PP5, PS3, PM1 | EXAC-ALL: 0<br>EXAC-AMR: 0<br>EXAC-EAS: 0<br>EXAC-FIN: 0<br>EXAC-NFE: 0<br>EXAC-OTH: 0<br>EXAC-SAS: 0 | QALL: 0.004 QAFR: 0.014 QARAB: 0.004<br>QBED: 0.004 QPER: 0 QSAS: 0 | GME-ALL: 0.003<br>GME-AP: 0.012<br>GME-CA: 0<br>GME-Israel: 0<br>GME-NEA: 0.003<br>GME-NWA: 0<br>GME-SD: 0<br>GME-TP: 0 | 1KG-ALL: 0<br>1KG-AFR: 0<br>1KG-AMR: 0<br>1KG-EAS: 0.001<br>1KG-EUR: 0<br>1KG-SAS: 0 |
| rs5030869 | NM_001360016.2: c.1003G>T | Likely pathogenic | PP3, PP5, PS3, PM1 | EXAC_ALL: 0.0002<br>EXAC_AFR: 0<br>EXAC_AMR: 0.0001<br>EXAC_EAS: 0.0001<br>EXAC_FIN: 0<br>EXAC_NFE: 0.0001<br>EXAC_OTH: 0.001<br>EXAC_SAS: 0.0006 | QALL: 0.00 QAFR: 0.00 QARAB: 0.00<br>QBED: 0.00 QPER: 0.01 QSAS: 0.00 | GME_ALL: 0.00<br>GME_NWA: 0.00<br>GME_NEA: 0.00<br>GME_AP: 0.01<br>GME_Israel: 0.00<br>GME_SD: 0.00<br>GME_TP: 0.00<br>GME_CA: 0.01 | |
| rs2515904 | NM_000402.3:c.576-60C>G | Benign | PP5, BS1, BS2 | | QALL: 0.03 QAFR: 0.19 QARAB: 0.00<br>QBED: 0.00 QPER: 0.00 QSAS: 0.00 | | 1KG-ALL: 0.05<br>1KG-AFR: 0.17<br>1KG-AMR: 0.01<br>1KG-EAS: 0.00<br>1KG-EUR: 0.00<br>1KG-SAS: 0.00 |
| rs5030872 | NM_001360016.2: c.542A>T | Likely pathogenic | BP4, PP5, PS3, PM1 | EXAC_ALL: 0.00<br>EXAC_AFR: 0.00<br>EXAC_AMR: 0.00<br>EXAC_EAS: 0.00<br>EXAC_FIN: 0.00<br>EXAC_NFE: 0.00<br>EXAC_OTH: 0.00<br>EXAC_SAS: 0.00 | | GME_ALL: 0.00<br>GME_NWA: 0.01<br>GME_NEA: 0.00<br>GME_AP: 0.00<br>GME_Israel: 0.00<br>GME_SD: 0.00<br>GME_TP: 0.00<br>GME_CA: 0.00 | |

(*Continued*)

Table 1. (Continued)

| Variant | HGVS nomenclature | ACMG Classification | Evidence | ExAC | Q1005 | GME | 1000 Genomes |
|---|---|---|---|---|---|---|---|
| rs137852339 | NM_001360016.2: c.949G>A | Likely pathogenic | BP4, PP5, PS3, PP1-M, PM1 | EXAC_ALL: 0.00<br>EXAC_AFR: 0.00<br>EXAC_AMR: 0.00<br>EXAC_EAS: 0.00<br>EXAC_FIN: 0.00<br>EXAC_NFE: 0.00<br>EXAC_OTH: 0.00<br>EXAC_SAS: 0.01 | QALL:0.01 QAFR:0.00<br>QARAB:0.00 QBED:0.01<br>QPER:0.00 QSAS:0.00 | GME_ALL: 0.00<br>GME_NWA: 0.00<br>GME_NEA: 0.00<br>GME_AP: 0.00<br>GME_Israel: 0.00<br>GME_SD: 0.00<br>GME_TP: 0.00<br>GME_CA: 0.00 | 1KG-ALL: 0.00<br>1KG-AFR: 0.00<br>1KG-AMR: 0.00<br>1KG-EAS: 0.00<br>1KG-EUR: 0.00<br>1KG-SAS: 0.01 |

The table contains all population and subpopulation allele frequencies associated with the four G6PD variants of statistical significance (p-value <0.01) present in the DALIA database as well as the pathogenic variants from other databases included in the heatmap. **Note**: For GME populations, the allele frequency has been calculated only for women.

https://doi.org/10.1371/journal.pone.0244567.t001

0.702 in GME-AP sub-population in the GME dataset. The iHS scores and the pairwise Fst scores for the variants are plotted in Fig 4. The iHS and Fst scores for the remaining 9 genes are included in the Supplementary file (S2–S10 Figs).

## Integration of datasets for understanding genetic epidemiology

We further explored whether the dataset could provide insights into the genetic epidemiology of mendelian diseases and traits. To this end, we analysed the allele frequencies of the variants in the Qatar exome/genome dataset as well as the GME datasets. Analysis of genetic variants with significant allele frequency differences revealed a total of 94 variants that were significantly different (Fisher's exact, p-value <0.01) in one or more populations or subpopulations. Upon analysis of these 94 variants, we discovered 4 variants mapping to the G6PD gene, associated with G6PD deficiency or favism, a prevalent disease in the Middle east. To provide a common standard for comparison of the pathogenicity annotations for the variants, we reclassified them according to the ACMG/AMP guidelines for interpretation of sequence variants [35]. Two of the four variants were found to have evidence to qualify being annotated as pathogenic or likely pathogenic. The variants, annotations and allele frequencies in each of the population and sub-populations are summarised in Table 1.

For further analysis, we additionally compiled all known pathogenic or likely pathogenic variants in the G6PD gene reported in Clinvar as well as other public databases. This resulted in a final list of 44 unique variants. We further analysed allele frequencies for these variants from across the four databases. Of the 44 variants, 21 had known allele frequencies in one or more of the four population datasets and 6 had allele frequencies in either Qatar, GME or both of the middle eastern population datasets.

Further analysis of these 6 variants revealed that the variants had allele frequencies ranging from 0.0 to 0.187 across different sub populations. Some of the variants showed significant population specificity, as in the case of rs5030872, which had allele frequencies in GME_ALL (0.001) and GME_NWA (0.01) but has not been reported in any of the Qatar or 1000 genome datasets. Similarly, rs2515904 has allele frequencies reported in Qatar (0–0.187) and 1000genomes (0–0.173) population datasets, but not in the GME and ExAC populations, indicating differences even among the Qatar and GME populations. Some variants showed intrapopulation variation, for example rs2515904, which had allele frequencies in QALL

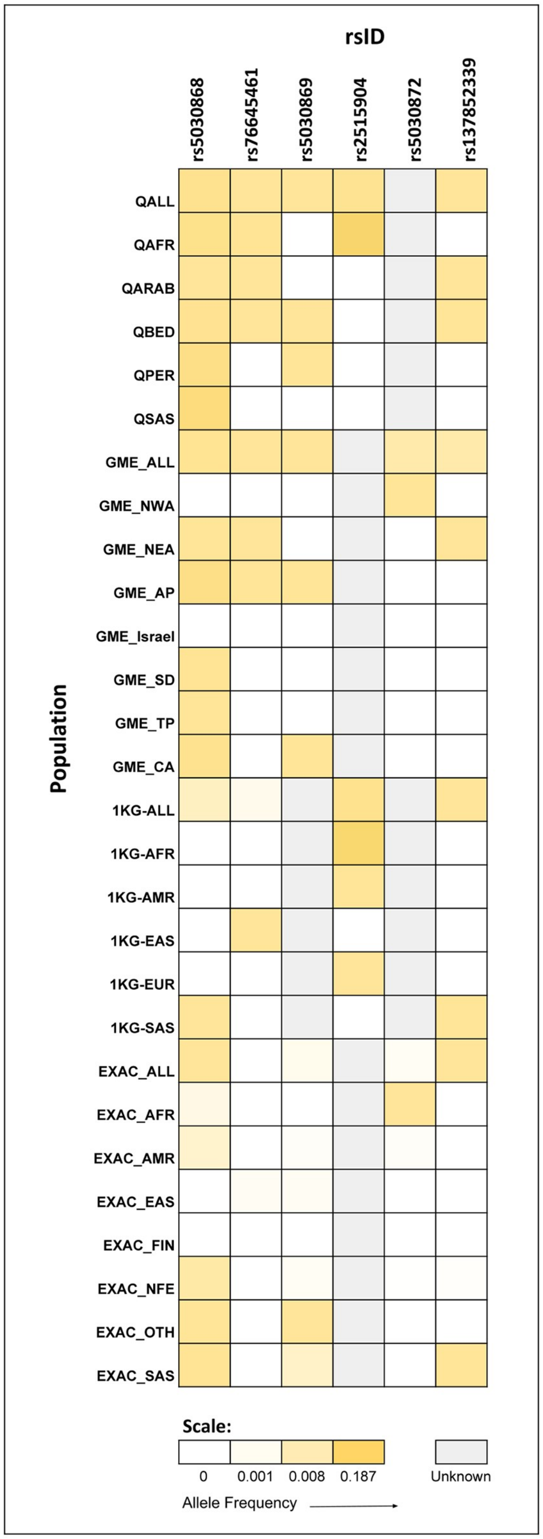


**Fig 5. Allele frequency heat map.** The figure depicts the variation in allele frequency distribution of all known G6PD pathogenic/likely pathogenic variants across four major datasets—Qatar, GME, 1000 Genomes and ExAC.

https://doi.org/10.1371/journal.pone.0244567.g005

(0.034) and QAFR (0.187), but had allele frequency 0 in QARAB, QBED, PER and QSAS. These variations are summarized in Fig 5 as an allele frequency heatmap.

## Conclusions

In summary, DALIA fills the need for a relevant, manually curated and annotated database of variants associated with genetic diseases in the Arab populations. The DALIA database provides searchable access to over 2700 genetic variants and associated information published from the region. Apart from being a ready reference to clinicians and clinical geneticists, we also suggest using examples how such a database could provide insights into genetic epidemiology of diseases as well as clues on prevalent genetic variants in subpopulations which is pertinent in developing approaches for cost effective diagnosis and screening of genetic diseases.

With population scale genome programmes now underway in the region [14, 15, 36] it is imperative that DALIA would become an important resource enabling appropriate analysis and interpretation of genomic data as well as a central resource for genetic epidemiology of diseases. Similarly with the advent of whole exome and whole genome sequencing becoming prevalent in clinical settings for the diagnosis of genetic diseases [36–39] it is expected that the accelerated discovery of genetic variants in diseases would also benefit the DALIA database in years to come.

In summary, to the best of our knowledge DALIA is one of the largest compilations of disease associated genetic variants from the Arab populations.

## Supporting information

**S1 Fig. Bubble plot displaying all the 143 significantly different variants found in the DALIA database upon comparison with both regional subpopulations as well as global population averages.**
(TIF)

**S2 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene ABCA4.** The 10 genes with top 1% |iHS| scores include ABCA4, THSD1, ATP7B, MEFV, ADA, CLDN14, MLC1, HSD17B4, NAT2, and ABCA1 (S1 Table). This plot depicts the iHS as well as pairwise Fst scores along the gene loci, for all the known variants for the gene ABCA4. The array of lines at the bottom represent all known variants, and the exon structure of the gene is shown beneath it.
(TIF)

**S3 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene THSD1.**
(TIF)

**S4 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene MEFV.**
(TIF)

**S5 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene ADA.**
(TIF)

**S6 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene CLDN14.**
(TIF)

**S7 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene MLC1.**
(TIF)

**S8 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene HSD17B4.**
(TIF)

**S9 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene NAT2.**
(TIF)

**S10 Fig. Plot depicting the iHS and pairwise Fst scores along the gene loci, for all the known variants for the gene ABCA1.**
(TIF)

**S1 Table. The table provides details of the ten genes in our database that had top 1% |iHS| scores, indicating positive selection.**
(XLSX)

## Acknowledgments

We acknowledge the research community working actively on genetic diseases for their wholehearted contribution, suggestions and contributions to the database without which it would have been impossible to be able to put together the resource.

## Author Contributions

**Conceptualization:** Sridhar Sivasubbu, Vinod Scaria.

**Data curation:** Aastha Vatsyayan, Shrey Gupta, Sumiti Sandhu, Seetha Lakshmi Venu, Vandana Sharma, Bouabid Badaoui, Kaidi Azedine, Serti Youssef, Anna Rajab, Alaaeldin Fayez, Seema Madinur.

**Formal analysis:** Aastha Vatsyayan.

**Funding acquisition:** Vinod Scaria.

**Investigation:** Aastha Vatsyayan, Vinod Scaria.

**Methodology:** Vinod Scaria.

**Project administration:** Aastha Vatsyayan, Vinod Scaria.

**Resources:** Anop Ranawat, Srinivasan Ramachandran, Vinod Scaria.

**Software:** Anop Ranawat, Kavita Pandhare, Srinivasan Ramachandran, Vinod Scaria.

**Supervision:** Parul Sharma, Vinod Scaria.

**Validation:** Aastha Vatsyayan.

**Visualization:** Aastha Vatsyayan.

**Writing – original draft:** Aastha Vatsyayan.

**Writing – review & editing:** Aastha Vatsyayan, Vinod Scaria.

## References

1. Arab Countries 2020 n.d. https://worldpopulationreview.com/country-rankings/arab-countries (accessed November 21, 2020).
2. Member states of the Arab League n.d. https://www.worlddata.info/alliances/arab-league.php (accessed November 21, 2020).
3. Tadmouri GO, Sastry KS, Chouchane L. Arab gene geography: From population diversities to personalized medical genomics. Glob Cardiol Sci Pract 2014; 2014:394–408. https://doi.org/10.5339/gcsp.2014.54 PMID: 25780794
4. Al-Gazali L, Hamamy H, Al-Arrayad S. Genetic disorders in the Arab world. BMJ 2006; 333:831–4. https://doi.org/10.1136/bmj.38982.704931.AE PMID: 17053236
5. Nakouzi G, Kreidieh K, Yazbek S. A review of the diverse genetic disorders in the Lebanese population: highlighting the urgency for community genetic services. J Community Genet 2015; 6:83–105. https://doi.org/10.1007/s12687-014-0203-3 PMID: 25261319
6. Tadmouri GO, Nair P, Obeid T, Al Ali MT, Al Khaja N, Hamamy HA. Consanguinity and reproductive health among Arabs. Reprod Health 2009; 6:17. https://doi.org/10.1186/1742-4755-6-17 PMID: 19811666
7. Al Rajeh S, Awada A, Bademosi O, Ogunniyi A. The prevalence of epilepsy and other seizure disorders in an Arab population: a community-based study. Seizure 2001; 10:410–4. https://doi.org/10.1053/seiz.2001.0602 PMID: 11700993
8. El-Hazmi MAF, Al-Hazmi AM, Warsy AS. Sickle cell disease in Middle East Arab countries. Indian J Med Res 2011; 134:597–610. https://doi.org/10.4103/0971-5916.90984 PMID: 22199098
9. Teebi AS, Teebi SA, Porter CJ, Cuticchia AJ. Arab genetic disease database (AGDDB): a population-specific clinical and mutation database. Hum Mutat 2002; 19:615–21. https://doi.org/10.1002/humu.10082 PMID: 12007218
10. Tadmouri GO, Al Ali MT, Al-Haj Ali S, Al Khaja N. CTGA: the database for genetic disorders in Arab populations. Nucleic Acids Res 2006; 34:D602–6. https://doi.org/10.1093/nar/gkj015 PMID: 16381941
11. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet 2016; 48:1071–6. https://doi.org/10.1038/ng.3592 PMID: 27428751
12. Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, et al. The Qatar genome: a population-specific tool for precision medicine in the Middle East. Hum Genome Var 2016; 3:16016. https://doi.org/10.1038/hgv.2016.16 PMID: 27408750
13. Koshy R, Ranawat A, Scaria V. al mena: a comprehensive resource of human genetic variants integrating genomes and exomes from Arab, Middle Eastern and North African populations. J Hum Genet 2017; 62:889–94. https://doi.org/10.1038/jhg.2017.67 PMID: 28638141
14. Project Team SG. The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse 2015; 6:22–6. https://doi.org/10.1109/MPUL.2015.2476541 PMID: 26583887
15. Al-Ali M, Osman W, Tay GK, AlSafar HS. A 1000 Arab genome project to study the Emirati population. J Hum Genet 2018; 63:533–6. https://doi.org/10.1038/s10038-017-0402-y PMID: 29410509
16. Hariprakash JM, Vellarikkal SK, Verma A, Ranawat AS, Jayarajan R, Ravi R, et al. SAGE: a comprehensive resource of genetic variants integrating South Asian whole genomes and exomes. Database 2018; 2018:1–10. https://doi.org/10.1093/database/bay080 PMID: 30184194
17. Rani J, Shah AR, Ramachandran S. pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. Journal of Biosciences 2015; 40:671–82. https://doi.org/10.1007/s12038-015-9552-2 PMID: 26564970
18. The Arab League n.d. https://www.cfr.org/backgrounder/arab-league (accessed November 21, 2020).
19. Geography of the Modern Middle East and North Africa n.d. http://www.middleeastpdx.org/resources/original/geography-of-the-modern-middle-east-and-north-africa/ (accessed November 21, 2020).
20. Khachfe HH, Refaat MM. Bibliometric analysis of Cardiovascular Disease Research Activity in the Arab World. ICFJ 2019; 15. https://doi.org/10.17987/icfj.v15i0.554
21. Danial-Farran N, Brownstein Z, Gulsuner S, Tammer L, Khayat M, Aleme O, et al. Genetics of hearing loss in the Arab population of Northern Israel. European Journal of Human Genetics 2018; 26:1840–7. https://doi.org/10.1038/s41431-018-0218-z PMID: 30139988
22. A BK, Nakash O, Goldberger N, Haklai Z, Geraisy N, Yatzkar U, et al. Completed suicide and suicide attempts in the Arab population in Israel. Soc Psychiatry Psychiatr Epidemiol 2016; 51. https://doi.org/10.1007/s00127-016-1219-2

23. Mokdad AH, Jaber S, Aziz MIA, AlBuhairan F, AlGhaithi A, AlHamad NM, et al. The state of health in the Arab world, 1990–2010: an analysis of the burden of diseases, injuries, and risk factors. Lancet 2014; 383:309–20. https://doi.org/10.1016/S0140-6736(13)62189-3 PMID: 24452042

24. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 2008; 29:6–13. https://doi.org/10.1002/humu.20654 PMID: 18000842

25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019; 47:D886–94. https://doi.org/10.1093/nar/gky1016 PMID: 30371827

26. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005; 33:D514–7. https://doi.org/10.1093/nar/gki033 PMID: 15608251

27. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;Chapter 7:Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928

28. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536:285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

29. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015; 526:68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

30. Sivadas A, Scaria V. Pharmacogenomic survey of Qatari populations using whole-genome and exome sequences. Pharmacogenomics J 2018; 18:590–600. https://doi.org/10.1038/s41397-018-0022-8 PMID: 29720721

31. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003; 31:3812–4. https://doi.org/10.1093/nar/gkg509 PMID: 12824425

32. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Genomics 2019: 806.

33. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018; 46:D1062–7. https://doi.org/10.1093/nar/gkx1153 PMID: 29165669

34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29:308–11. https://doi.org/10.1093/nar/29.1.308 PMID: 11125122

35. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015; 17:405–24. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

36. Zayed H. The Qatar genome project: translation of whole-genome sequencing into clinical practice. Int J Clin Pract 2016; 70:832–4. https://doi.org/10.1111/ijcp.12871 PMID: 27586018

37. Jithesh PV, Scaria V. From genomes to genomic medicine: enabling personalized and precision medicine in the Middle East. Per Med 2017; 14:377–82. https://doi.org/10.2217/pme-2017-0048 PMID: 29754564

38. Yousri NA, Fakhro KA, Robay A, Rodriguez-Flores JL, Mohney RP, Zeriri H, et al. Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. Nat Commun 2018; 9:333. https://doi.org/10.1038/s41467-017-01972-9 PMID: 29362361

39. Al-Dewik N, Mohd H, Al-Mureikhi M, Ali R, Al-Mesaifri F, Mahmoud L, et al. Clinical exome sequencing in 509 Middle Eastern families with suspected Mendelian diseases: The Qatari experience. Am J Med Genet A 2019; 179:927–35. https://doi.org/10.1002/ajmg.a.61126 PMID: 30919572