

COMMENT

Open Access



# Massive underrepresentation of Arabs in genomic studies of common disease

Romit Bhattacharya<sup>1,2,3†</sup>, NingNing Chen<sup>4,5†</sup>, Injeong Shim<sup>2,3,6</sup>, Hiroyuki Kuwahara<sup>4,5</sup>, Xin Gao<sup>4,5†</sup>, Fowzan S. Alkuraya<sup>7\*†</sup> and Akl C. Fahed<sup>1,2,3\*†</sup> 

## Abstract

Arabs represent 5% of the world population and have a high prevalence of common disease, yet remain greatly underrepresented in genome-wide association studies, where only 1 in 600 individuals are Arab. We highlight the persistent and unaddressed underrepresentation of Arabs in genomic databases and discuss its impact on public health genomics and missed opportunities for biological discovery.

**Keywords** Arab, Middle East, Underrepresentation, Diversity, Genomic, Polygenic score, GWAS

<sup>†</sup>Romit Bhattacharya and NingNing Chen are co-first authors.

<sup>†</sup>Xin Gao, Fowzan S. Alkuraya, and Akl C. Fahed jointly supervised this work.

\*Correspondence:

Fowzan S. Alkuraya

FAlkuraya@kfshrc.edu.sa

Akl C. Fahed

afahed@mgh.harvard.edu

<sup>1</sup> Department of Medicine, Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, 185 Cambridge Street | CPZN 3.128, Boston, MA 02114, USA

<sup>2</sup> Department of Medicine, Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

<sup>3</sup> Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>4</sup> Computational Biosciences Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

<sup>5</sup> Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

<sup>6</sup> Department of Digital Health, Samsung Medical Center, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of South Korea

<sup>7</sup> Department of Translational Genomics, Center for Genomic Medicine, King Faisal Specialist Hospital and Research Center, Riyadh, Kingdom of Saudi Arabia

## Arabs, common complex disease, and healthcare disparities

Arab people represent about 5% of the world population and are an ancestrally diverse group with unique health considerations. Arabs can be defined by their geographic histories, religious practices, ethnicities, and countries of origin. The countries of the Arab world span the Arabian Peninsula, the Levant, and North Africa and are very diverse economically — including some of the wealthiest and poorest countries in the world — and genetically through extensive evidence of admixture found among present-day Arabs. However, while Arabs are ancestrally diverse, they have (1) genetically distinct signatures [1] and (2) unique risk factors for common diseases. These include waterpipe smoking — a key driver for risk of heart disease and cancer, some of the highest rates of vitamin D deficiency worldwide — attributed to dark skin and conservative clothing [2] and consanguinity leading to autozygosity.

Arabs are also present in diaspora, and as such, the risk of propagating healthcare disparities is a problem for not only the Arab world but also the West. Several Arab countries have witnessed large waves of immigration to the West due to conflict and political or economic breakdown. In the US alone, it is estimated that there are 3.7 million Arabs although it is likely that actual estimates



are larger due to underreporting. Limited studies on Arab Americans observed a higher age-adjusted mortality rate than non-Arab Americans raising concerns for possible disparities in those populations [3]. With extensive European, Asian, and African admixture, Arabs can present as White, or non-White which allows misidentification in Western datasets and confounds attempts at disaggregation of data. In fact, even the US Census does not acknowledge Arabs as a separate category and instead sees them as “White.”

There is a rising epidemic of complex diseases in countries of the Arab world. The increasing rates of cardiovascular disease, diabetes, dyslipidemia, and cancer are often ascribed to environmental, behavioral, or economic factors. For decades, communicable diseases in low- and middle-income countries of the Arab world were the public health priority, but today nearly 80% of deaths are attributable to non-communicable diseases [4]. Common diseases are highly heritable and it remains unclear how genomic background may drive risk among Arabs for common diseases where there is a large public health burden. The interplay of genomic and nongenomic risk for many common diseases is increasingly appreciated in European-ancestry datasets and, more recently, other world populations with growing genomic and clinical datasets. For example, recent insights from individuals of South Asian ancestry identified increased risk for cardiovascular disease not captured by known conventional risk factors [5]. Whether this is true for Arabs is unknown.

### **Massive and persistent underrepresentation of Arabs in genome-wide association studies**

Genomic databases globally have suffered from limited ancestral diversity, and Martin et al. highlighted how this could propagate health disparities through predictable overrepresentation of genetic minorities — demonstrating that in 2019 despite only 16% of the global population being of European descent, ~79% of all genome-wide association studies (GWAS) participants were of European descent [6]. Arab populations are among the populations least represented in GWAS and large population-based biobanks, along with peoples of Oceania, Southeast Asia, Native American nations, and parts of Africa. Given high levels of endogamy and consanguinity in the Arab population, the inclusion of Arabs in such datasets has been an area of interest, but has not been a focus globally. Currently, in the GWAS Catalog, 88% of individuals are of European ancestry and 5.8% are East Asian (Biobank Japan and China Kadoorie Biobank being among the largest contributors from East Asia). By contrast, the relative inclusion of other ancestry groups has consistently languished at <1.6%, with Hispanic, South

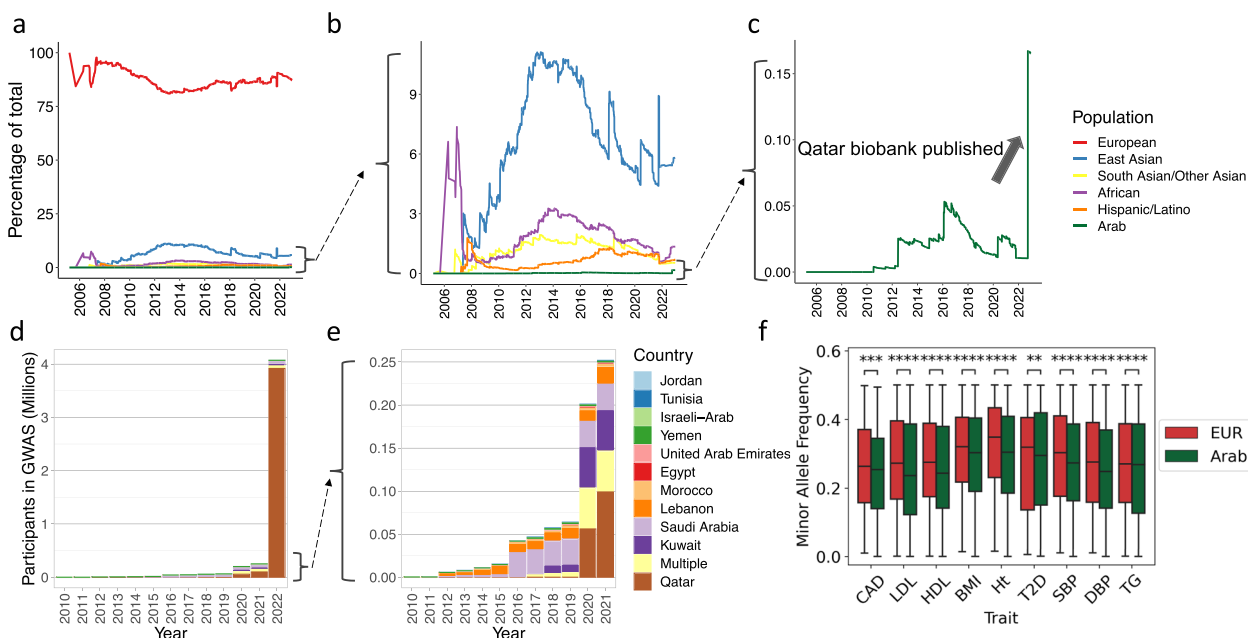
Asian, and Greater Middle Eastern populations making up <1% of the participants in published GWAS studies.

We used the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>) to quantify the representation trend for Arab individuals over 18 years — on aggregate and broken down by country — and find extraordinary underrepresentation in GWAS studies. When we examined Arab populations specifically defined by country of origin (in contrast to the grouping variable “Greater Middle Eastern” that is often used), we found approximately 0.17% of the GWAS Catalog was Arab — meaning about 1 in 600 individuals in published GWAS studies are currently of Arab ancestry (compared to 523 in 600 are of European ancestry). What is more discouraging is that this disparity has not been sufficiently addressed over time (Fig. 1a). When considering only non-European genomes, Arabs are noted to be consistently near 0% since the start of the GWAS Catalog (Fig. 1b). In fact, representation from the majority of countries considered to comprise the Arab world is lacking (Fig. 1d, e). Genomes from Algeria, Iraq, Libya, Oman, Somalia, and Syria are all completely absent.

The Qatar Biobank (QBB) has published data on ~15,000 individuals, and this represents the single largest contribution to Arab genomic representation to date (Fig. 1c) [7]. With even this modest addition in number, the contribution to genomics has been large with the initial GWAS from the QBB identifying 24.6 million previously unknown variants, and numerous novel disease-associated variants predominantly from 5 newly identified non-admixed subclusters within the Qatari population [7, 8]. The addition of QBB data to the GWAS Catalog represents a great advance, but additionally highlights the necessity and urgency of increasing representation from other Arab countries (Fig. 1d, e). While the contributions of the QBB are commendable, Qatar is not representative of all Arabs, for instance the distant North African or Levantine countries, as genetic ancestry varies greatly across the region.

### **Missed opportunities in disease prevention and discovery**

Discoveries from GWAS have begun to move into the clinical space through implementation of polygenic risk scores (PRS). It is now well-established that lack of diversity in reference genomes and published GWAS limit portability and accuracy of PRS prediction in non-European ancestries. The issues begin with SNP array imputation where reference panels are predominantly of European origin, and thus, when linkage disequilibrium (LD) blocks differ across ancestries, variation in diverse populations can be completely obscured<sup>6</sup>. Differences in minor allele frequency (MAF) and effect size of variants



**Fig. 1** Proportional representation of Arabs in genomic studies of common disease. **a** Proportional representation of published GWAS study participants between 2006 and 2022 for individuals of 6 ancestral groups. The y-axis represents the cumulative percentage of the total. **b** Zoomed-in plot of **a** which excludes European-ancestry participants highlighting increase in representation of East Asian, South Asian, African, and Hispanic/Latino-ancestry group participants over time, but a persistent representation near zero for Arabs. **c** Zoomed-in plot of **b** which shows only Arab participants highlighting the recent increase in representation due to the publication of data from the Qatar Biobank (QBB). **d** Participants in published GWAS studies by the Arab country of origin between 2010 and 2022. By 2022, 96.2% of the Arab participants are from Qatar and the remaining 3.8% are from 10 other countries. **e** Zoomed-in plot of **d** which excludes the data of 2022 to highlight the breakdown of participants by the remaining countries. **f** For each of 9 cardiometabolic traits, we compare the minor allele frequency (MAF) of causal variants associated with the trait among individuals of European ancestry in the UK Biobank ( $N=361,194$ ) to their frequency in an indigenous Arab cohort from Saudi Arabia ( $N=1017$ ). Causal variants were obtained from fine mapping results performed using FINEMAP v1.3.1 and SuSiE v0.8.1.0521, available from the Funcane Lab ([www.finucanelab.org/data](http://www.finucanelab.org/data)). Across all 9 traits, the MAF of causal variants was significantly lower in Indigenous Arabs compared to European-ancestry individuals. CAD coronary artery disease, LDL low-density lipoprotein cholesterol, HDL high-density lipoprotein cholesterol, BMI body mass index, Ht height, T2D type 2 diabetes mellitus, SBP systolic blood pressure, DBP diastolic blood pressure, TG triglyceride. Mann–Whitney  $U$  test was used for comparison. “\*\*\*\*\*” denotes  $p < 1.00e-04$ , “\*\*\*\*” denotes  $1.00e-04 < p \leq 1.00e-03$ , and “\*\*\*” denotes  $1.00e-03 < p \leq 1.00e-02$

among individuals of different genetic ancestries also lead to reduced accuracy of prediction — and increase the likelihood of propagating healthcare disparities <sup>6</sup>. When comparing the MAF of causal variants associated with 9 cardiometabolic traits in a large European-ancestry cohort (the UK Biobank) to their frequency in an indigenous Arab population, the MAF is on average 7.6% lower in Arabs ( $p\text{-value}=4.2e-06$ ), explaining part of the reduced performance of PRS that are derived in European populations when they are applied to Arabs (Fig. 1f) [7].

The use of computational methods to improve cross-ethnic PRS performance in genetically diverse populations is encouraging, but not enough to allay real concerns regarding ongoing health disparities. While large biobanks, population cohort studies, and international consortia have allowed for serially larger GWAS, the genetic diversity of these GWAS has not kept pace, despite advocacy from groups like the American Society

of Human Genetics to prioritize portability of scores. Polygenic scores are increasingly being returned to patients, and clinico-genomic models to predict and intervene on risk early in the life course through a precision medicine approach are likely to become the norm. Without the appropriate development or validation of such clinico-genomic models in Arabs, widespread implementation risks propagation of already existing disparities in care.

Beyond the serious impact on public health genomics, the massive underrepresentation of Arabs in genomic studies is a missed opportunity to discover new disease biology. A prior comment in the journal eloquently described the opportunity to identify homozygous loss of function variants and novel candidate genes for recessive disease through more sequencing of highly consanguineous Arab populations [9]. We highlight that also for common disease, homozygous loss of function variants (human knockouts) might

inform interpretation of GWAS. Increasingly, the diversity of GWAS participants, more than the sample size, is advancing the discovery of new loci. For example, in a recent multi-ancestry GWAS of type 2 diabetes, 46% of new loci would not have been identified in a European-ancestry GWAS only [10]. Today, more than 30 million whole genomes have been sequenced which has advanced our understanding of genetic underpinnings of disease, but it is now well understood that in addition to the number of genomes, future opportunities for discovery lie in sequencing a diversity of genomes. First, genomes of different ancestries, when combined, allow leveraging of different LD block structures for better identification of causal variants within a specific genetic region. Second, different causal variants for similar diseases may be identified by studying diverse populations.

The unique ancestral, geographic, and cultural histories of the Arab people offer many opportunities for discovery and improved risk prediction and health for the 450 million Arabs in the world. Future directions and actions to reduce disparities and increase yield of novel variants would include (1) encouraging census bodies to disaggregate Arabs from White individuals within the diaspora, (2) active recruitment of Arabs in genomic databases, and (3) supporting infrastructure and training programs in Arab countries to develop their own biobanks genomic resources to add to the global collective pool of data.

#### Abbreviations

BMI	Body mass index
CAD	Coronary artery disease
CB2	Cannabinoid receptor 2
DBP	Diastolic blood pressure
DKD	Diabetic kidney disease
GWAS	Genome-wide association study
HDL	High-density lipoprotein
Ht	Height
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
MAF	Minor allele frequency
PC	Principal components
PRS	Polygenic risk score
QBB	Qatar Biobank
SBP	Systolic blood pressure
SNP	Single nucleotide polymorphism
T2DM	Type 2 diabetes mellitus
TG	Triglycerides
US	United States

#### Acknowledgements

None.

#### Authors' contributions

RB wrote the manuscript and contributed to study data interpretation. NC analyzed and interpreted the data from the GWAS Catalog and contributed to writing. IS, HK, and XG contributed to data collection and analysis. FSA performed critical review of the manuscript and contributed to study design and oversight of the study. ACF oversaw the study design, data analysis, and manuscript preparation. All authors read and approved the final manuscript.

#### Funding

This study was supported by a sponsored research agreement between King Faisal Specialist Hospital and Research Center and Massachusetts General Hospital. RB is funded by the John S. LaDue Fellowship in Cardiovascular Disease Research at Harvard Medical School. HK, NC, and XG are supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No FCC/1/1976-44-01, FCC/1/1976-45-01, REI/1/5202-01-01, REI/1/5234-01-01, REI/1/4940-01-01, RGC/3/4816-01-01, and REI/1/0018-01-01. IS is supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, South Korea (HI19C1328). ACF is supported by a grant from the National Heart, Lung, and Blood Institute (1K08HL161448).

#### Availability of data and materials

All data are publicly available at <https://www.ebi.ac.uk/gwas/> and <https://www.finucanelab.org/data>.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

All authors have reviewed the text and consent to publication.

##### Competing interests

RB is a medical advisor to Casana Care Inc, unrelated to the present work. ACF is a co-founder of Goodpath and has received a grant from Abbott unrelated to the present work. The remaining authors declare that they have no competing interests.

Received: 22 September 2023 Accepted: 8 November 2023

Published online: 22 November 2023

#### References

- Almarri MA, Haber M, Lootah RA, Hallast P, Al Turki S, Martin HC, Xue Y, Tyler-Smith C. The genomic history of the Middle East. *Cell*. 2021;184(18):4612-4625.e14. <https://doi.org/10.1016/j.cell.2021.07.013>.
- Chakhtoura M, Rahme M, Chamoun N, El-Hajj Fuleihan G. Vitamin D in the Middle East and North Africa. *Bone Reports*. 2018;8:135-46. <https://doi.org/10.1016/j.bonr.2018.03.004>.
- El-Sayed AM, Tracy M, Scarborough P, Galea S. Ethnic inequalities in mortality: the case of Arab-Americans. *PLoS ONE*. 2011;6(12):e29185. <https://doi.org/10.1371/journal.pone.0029185>.
- The World Bank Open Data, Cause of death, by non-communicable disease (% of total)—Middle East & North Africa. (n.d.). Retrieved April 24, 2023, from <https://data.worldbank.org/indicator/SH.DTH.NCOM.ZS?locations=ZQ>
- Patel AP, Wang M, Kartoun U, Ng K, Khera AV. Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: results from the UK Biobank prospective cohort study. *Circulation*. 2021;144(6):410-22. <https://doi.org/10.1161/CIRCULATIONAHA.120.052430>.
- Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-91. <https://doi.org/10.1038/s41588-019-0379-x>.
- Thareja G, Al-Sarraj Y, Belkadi A, Almotawa M, Suhre K, Albagha OME. Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nat Commun*. 2021;12(1):1250. <https://doi.org/10.1038/s41467-021-21381-3>.
- Mbarek H, Ismail SI. Mapping the Arab genome. *Nat Genet*. 2022;54(12):1761-3. <https://doi.org/10.1038/s41588-022-01239-0>.
- Abou Tayoun AN, Rehm HL. Genetic variation in the Middle East—an opportunity to advance the human genetics field. *Genome Medicine*. 2020;12(1):116. <https://doi.org/10.1186/s13073-020-00821-7>.

10. Ken Suzuki, Konstantinos Hatzikotoulas, Lorraine Southam, Henry J. Taylor, Xianyong Yin, Kim M. Lorenz, Ravi Mandla, Alicia Huerta-Chagoya, Nigel W. Rayner, Ozvan Bocher, S. V. Arruda Ana Luiza de, Kyuto Sonehara, Shinichi Namba, Simon S. K. Lee, Michael H. Preuss, Lauren E. Petty, Philip Schroeder, Brett Vanderwerff, Mart Kals, ... Eleftheria Zeggini. (2023). Multi-ancestry genome-wide study in > 2.5 million individuals reveals heterogeneity in mechanistic pathways of type 2 diabetes and complications. *MedRxiv*, 2023.03.31.23287839. <https://doi.org/10.1101/2023.03.31.23287839>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

