Biomedicine Hub

Review

Biomed Hub 2023;8:60–71
DOI: 10.1159/000530619

Received: September 23, 2022
Accepted: March 29, 2023
Published online: September 14, 2023

Population Genome Programs across the Middle East and North Africa: Successes, Challenges, and Future Directions

Hagar Ateia[a, b] Pauline Ogrodzki[a] Hannah V. Wilson[b]
Subhashini Ganesan[a, b] Rabih Halwani[c] Ashish Koshy[a] Walid A. Zaher[a, b, d, e]

[a]G42 Healthcare, Masdar City Abu Dhabi, UAE; [b]IROS (Insights Research Organization and Solutions), Abu Dhabi, UAE; [c]Sharjah University, Sharjah, UAE; [d]UAE University, Al Ain, UAE; [e]Khalifa University, Abu Dhabi, UAE

Keywords
Population genome programs · Middle east · Whole-genome sequencing · Precision medicine · Challenges

Abstract
In this review, we discuss the current state of population genome programs (PGPs) conducted in the Middle East and North African (MENA) region. This region has high prevalence of genetic diseases and significant health challenges as well as being a significantly underrepresented population in public genetic databases. The majority of ongoing PGPs represent regions in Europe, North and South America, South Asia, Australia, and Africa, with little to no descriptive information highlighted only on the MENA Region when it comes to genome programs databases, outcomes, or the challenges that MENA region countries may face establishing their own national programs.
This review has identified 6 PGPs currently underway in the MENA region, namely in the Kingdom of Saudi Arabia, Qatar, Egypt, the United Arab Emirates, Bahrain, and Iran. Due to the rapidly growing involvement of the MENA region in national-scale genomic data collection, an increase in representation in public genetic databases is to be expected to occur in the near future. Whilst significant progress is being made in some MENA countries, future initiatives as well as ongoing programs will be facing several challenges related to collaboration, finance, infrastructure and institutional data access, data analysis, sustainability, health records, and biobanks. The review also reiterates the need for ensuring ethical and regulated genomic initiatives which can drive developments in personalized medicine treatments to improve patient prognosis and quality of life.

Introduction

The first human genome sequencing project culminated in 2003 and revolutionized the landscape of genomic and personalized medicine [1]. Multiple population genomic programs have been initiated worldwide in collaborative efforts to increase the diversity in human genetic databases and advance breakthroughs in personalized healthcare [2–7]. Revolution in genomics research started with the very first human genome sequenced and published in the late 21st century, resulting in promising outputs of whole-genome sequencing (WGS) and microarray-based genotyping of millions of human genomes. Ongoing and publicly released population genome programs (PGPs) represent

karger@karger.com
www.karger.com/bmh

Correspondence to:
Hagar Ateia, hagar.ateia@g42.ai
Pauline Ogrodzki, pauline.ogrodzki@g42.ai

regions in Europe, North and South America, South Asia, Australia, and Africa, whereas the Middle East and North African (MENA) regions are significantly underrepresented in public genomic databases (e.g., Trans-Omics for Precision Medicine [TOPMeD] or the Genome Aggregation Database [gnomAD]), despite the initiation of several PGPs [8, 9]. This is evident from the fact that the gnomAD database (Version 3.1) contains data from only 158 Middle Eastern genomes. Since then, resolving medical challenges has been grounded in many European countries as a result of investment decisions in genomic medicine. Potential benefits of genomic medicine and research have become apparent for multiple countries worldwide. Such benefits include, but are not limited to, better understanding of diseases etiology, early diagnosis, drug design, pharmacogenomics, epidemiological studies, and preventive and personalized medicine. Population-specific national genome programs are attracting the attention of multiple countries within the MENA region, with the aim of improving genetic diagnostics and paving the road for implementing large-scale screening and precision medicine into their healthcare systems.

Although there are several million genomes available to date in global public databases [10], there is almost no high-quality single reference genome for the MENA region, or individual countries within the region [9], which is a critical problem that requires attention. The level of confidence and accuracy of genetic diagnosis is highly dependent on the quality of the reference genome. Regional doctors and genetic counselors are currently using a reference genome that is European and Caucasian centric "Genome Reference Consortium Human Build 38 – GRCh38" which does not reflect MENA-specific genetic information [11]. Adoption of precision medicine within the MENA region may benefit significantly from applied reference genome programs as consanguineous marriages resulting in Mendelian and metabolic disorders are highly prevalent, ensuing the spread of genetic diseases within the population.

Therefore, it is imperative to build a new reference genome that accurately represents the genetic diversity, structure, and variants of the worlds' population, by encouraging underrepresented regions such as the MENA region to conduct PGPs, which would be the first step toward precision and personalized medicine. The findings from PGPs have broad implications in the future for advancing prevention and treatment of diseases, thereby improving the healthcare systems, for instance, empowering gene manipulation techniques such as CRISPR-Cas systems, which in turn can significantly help in detection and alteration of the acquisition of disease genes in the affected populations [12].

Further, to have a complete understanding of the role of genetics in diseases, it is imperative to appreciate the link between genetics, environment, and disease. The interplay of intergenerational and transgenerational epigenetics is of special importance in the MENA region which shelters people from different ethnic backgrounds and nationalities and has an increased expat population. The structure of such a population provides an opportunity to appreciate the role of environment in such a diverse population and to study DNA sequence variations and epigenetic inheritance. Therefore, population-wise genomic projects could also increase coverage by extending it to the expat population, as that would serve as valuable information to advance our understanding of the link between genetics, environment, and disease [13, 14].

To address the lack of population genomic information and develop tools and solutions to tackle the high prevalence genetic diseases and significant health challenges in the MENA region, PGPs have been initiated in a number of countries. This review will summarize ongoing MENA-based PGPs and highlight the up-to-date outcomes of each national program, which were available through published literature.

Genomics and Precision Medicine

WGS of an individual's DNA can facilitate diagnosis of a disease, but its potential for guiding treatment has been underestimated for several years. Genomic medicine is the use of genetic information to direct medical care or diagnose a disease based on the genome, and novel technologies such as WGS and whole exome sequencing (WES) have had a huge impact on that [15]. A tailored medicine that will fit every individual based on their own genetic material is an actuality. WGS has been a useful tool in research for identifying new disease-causing mutations, but can it help physicians make better decisions about treatment options? In the 90s, the first draft of the human genome published was expected to have a transformative impact on medicine [16]. Predictions were made about a huge shift in which medicine could be personalized, predictive, and preventive [17]. To many, no such theories were materialized, especially in the Middle East, where there is a drastically continuous increase in recessive Mendelian disorders as a result of consanguineous marriages [18].

The concept of precision medicine is not new [19]. For example, blood typing has been used to guide blood transfusions for more than a century. For the past few decades, this concept has been enhanced by the

Table 1. Tabulation of active PGPs in the MENA region

Country	PGP status	Reference
KSA	Active	SHGP
Qatar	Active	QGP
Egypt	Active	EgyptRef
UAE	Active	EGP
Bahrain	Active	National Genome Center of Bahrain
Iran	Active	Iranome

development of large-scale genomic databases such as the human genome project, powered by powerful molecular tools and methods for processing and analyzing human biological material, namely multi-omics, encompassing genomics, proteomics, metabolomics, and much more [20]. Early examples of successful application of precision medicine are primarily found in oncology and cancer research [21]. This involves sequencing of a patient's genome and of cancer cell genomes to identify therapeutic targets. Furthermore, this approach can lead to the discovery of causative mutations and correct diagnosis in a time-critical clinical setting [22]. The next step in personalized medicine is to broaden research applications to encourage the development of precision medicine tools, test them rigorously, and ultimately apply them to build the evidence base information needed to guide clinical decisions.

PGPs in the MENA Region

The MENA region is comprised of 19 countries including the Kingdom of Saudi Arabia (KSA), Qatar, Egypt, United Arab Emirates (UAE), Bahrain, Iran, Algeria, Israel, Jordan, Iraq, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Syria, Tunisia, and Yemen. To date, PGPs have been initiated in 6 countries, as tabulated in Table 1.

Population genomics initiatives are currently being conducted in 6 MENA countries, namely in the KSA, Qatar, Egypt, UAE, Bahrain, and Iran, of which 4 countries (KSA, Qatar, Egypt, and UAE) have made the majority of information public and hence accessible for this review [23–28]. Data on the national-scale programs were limited for 2 of the 6 countries, namely Bahrain and Iran, and thus these countries are collated together below to be discussed in this review [27, 28].

The 4 countries with ongoing PGPs are summarized in Figure 1 and Table 2 and discussed separately on a country basis in detail below.

The Saudi Human Genome Program

The Saudi Human Genome Program (SHGP) was launched in 2013 by King Abdulaziz City for Science and Technology (KACST). The main goal was to sequence the exomes and/or genomes of 100,000 Saudi nationals within 5 years using next-generation sequencing across seven satellite laboratories [26, 29].

The program was launched to initiate a new era of personalized medicine and diagnosis for hereditary diseases that often result from consanguineous marriages [29]. The main objectives of SHGP are to build a genetic database for Saudi citizens, and leverage information from the SHGP to develop targeted diagnostic, therapeutic, and preventative tools [30]. In addition, focus was given to several diseases which are common in the KSA including diabetes, hearing loss, cardiovascular disease, neurodegeneration, cancer, inherited, and Mendelian disorders. The initiative intends to create an advanced infrastructure in the fields of genomics and bioinformatics, with the goal of enhancing treatment approaches based on the genetic makeup of patients [31].

This program employs next-generation sequencing technologies such as the Illumina platform NovaSeq 6000 (for WGS) and the Thermo Fisher Scientific Life Technologies platforms Ion Proton and S5 XL (for WES). Targeted panel sequencing is also conducted using Sanger sequencing for single DNA fragments, and genotyping arrays for calling a selected set of single nucleotide polymorphism (SNP) variants. These technologies are housed primarily at the SHGP core facility in Riyadh, as well as across 7 satellite facilities located across KSA. The satellite laboratories were established as collaborating institutions with self-adopted timelines, capabilities, pipelines, and projects, which were led by on-site researchers independently from the SHGP headquarter. They are connected to the center computational unit in KACST for the purpose of providing backup and performing secondary analysis tasks [32].

As of April 2022, the SHGP had sequenced more than 56,000 samples and identified 7,500 variants, of which 3,000 are novel causative mutations directly associated with over 1,230 rare genetic disorders [33, 34]. The translational outcomes of this program include genetic counselling, and, in some instances, further genetic testing of families where high-risk underlying genetic mutations have been identified. For example, several genetic counselling services for inherited disorders such as premarital screening (the Saudi Premarital Screening Program [PSP]) and genetic counselling programs for hemoglobinopathies and viral infections have been implemented in the country [35]. Developing applications

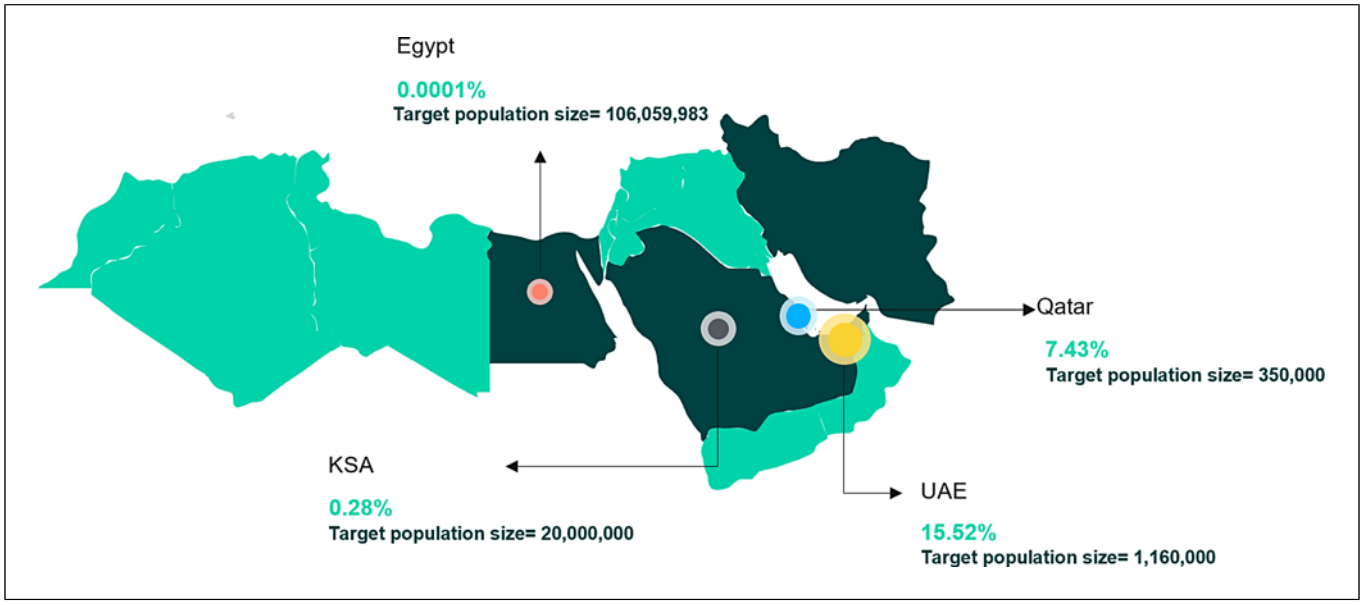

Fig. 1. MENA region population-specific national genome programs map: percentage of sequenced samples/target population size. (Dark green: active national PGPs in the MENA region. Bright green: countries that do not have active PGPs in the MENA region). Target population size refers to the size of the local population (nationals). Population information was accessed on the 11th of May 2022, through www.worldometers.info.

in pharmacogenomics research and personalized medicine are prospective objectives that may also arise from the SHGP. Specifically, via the prediction of potential occurrence of rare genetic syndromes, curative approaches were advanced by proposing suitable medical solutions tailored upon the patient's genetic makeup [33, 36, 37].

The PSP was born from the SHGP and launched in 2019. The PSP involves identifying the genetic carrier status of couples before getting married. Data are compared against known Saudi pathogenic variants associated with autosomal recessive disorders. The PSP also includes genetic counselling to enable couples to make informed decisions and prevent transmission of genetic conditions to future off springs [38]. In addition, the aggregate genetic information offers the opportunity to perform genome-wide association studies. Such studies may shed light upon several polymorphic traits that are associated with diseases and are suspected to add to their penetrance which may gain special importance for endemic conditions like obesity or for diseases like cancer [39, 40].

A national information base was also established at KACST to store data on population variations and make these data available to clinicians in Saudi Arabia to enable future diagnostic and screening efforts [29, 33]. The database is collected from several servers and a large computer hosted by KACST named *SANAM*, an energy-efficient, high-performance computer. The satellite laboratories across the kingdom transfer their database to the central server to manage and analyze the data collected from different resources. The transfer of data is prioritized and scheduled to reduce the required bandwidth, and thus collectively, the central and satellite computer resources as well as the automatic extension with commercial cloud solutions work together like a hybrid multicloud system [29].

The Qatar Genome Programme

The Qatar Genome Programme (QGP) is a population-based program launched in 2015 by the Qatar Foundation to generate a large-scale WGS dataset [41]. The QGP is a longitudinal population-based cohort study which conducts participant follow-ups every 5 years. The program combines WGS data with comprehensive phenotypic information collected by the Qatar Biobank (QBB) [42]. The purpose of collecting these data is to support genomic medicine in the country and the region with the aim of providing unique insights that will enable the development of personalized healthcare in the country. The QGP started by sequencing the whole genomes of 6,045 subjects whose specimens were collected and bio-banked by the QBB. The program included adult participants (over 18 years) who are

Table 2. Active population genome programs in Saudi Arabia, Qatar, Egypt, and UAE as of April 2022 (in chronological order)

	Saudi Arabia	Qatar	Egypt	UAE
Data collection	Ongoing	Ongoing	Ongoing	Ongoing
Year launched	2013	2015	2020	2021
Title	SHGP	QGP	EgyptRef	EGP
Population demographics				
Total country population*	35,780,534	3,183,745	106,059,983	10,117,562
Target population size*	20,000,000	350,000	106,059,983	1,160,000
Program sequencing target	100,000	100,000	110	1,000,000
Sequenced samples	56,000	26,000	110	180,000
Date of sequencing report	2022	2022	2020	2022
Percentage of sequenced samples in target population size, %	0.28	7.43	0.0001	15.52s
Sequencing technology				
Short reads	✓	✓	✓	✓
Long reads		✓	✓	✓
WES	✓	✓		✓
WGS	✓	✓	✓	✓
Technology Platform				
Illumina, Inc	✓	✓	✓	✓
PacBio	✓		✓	
ONT				✓
BGI				✓
10x Genomics, Inc	✓		✓	
Sanger Sequencing	✓			
Genotyping SNP Arrays	✓			
Funding Source				
Governmental	✓	✓	✓	✓
Private Source	✓	✓	✓	✓
Data accessibility and sharing				
Public database			✓	
Private database	✓	✓		✓
Data available with restriction and IRB requests	✓	✓	✓	
Linkage of data				
to Existing biobank		✓		✓
to Health records	✓	✓	✓	✓

EGP, Emirati Genome Program; IRB, Institutional Review Board; QGP, Qatar Genome Programme; SHGP, Saudi Human Genome Program; ONT, Oxford Nanopore Technologies; BGI, Beijing Genomics Institute Genomics; PacBio, Pacific Biosciences; SNP, single nucleotide polymorphism; UAE, United Arab Emirates; WES, whole-exome sequencing; WGS, whole-genome sequencing. *Total country population information was accessed on 11 May, 2022, through www.worldometers.info. *Target population size refers to the size of the local population (nationals).

Qatari nationals or resident in Qatar for at least 15 years. One important goal of QGP is to raise awareness of scientific education by developing a series of educational and training courses [43]. These courses focus on aspects of genomics research relevant to the Qatari population, at both school and university levels.

QGP library construction and sequencing are performed at the Sidra Clinical Genomics Laboratory Sequencing Facility (Sidra Medicine), in partnership with Hamad Medical Corporation (HMC). The Illumina HiSeq X10 platform (Illumina, San Diego, CA, USA) was the technology of choice for the program [44].

The QGP is divided into three phases; Phase 1 concluded in 2018, with the completion of 10,000 whole genomes sequenced. Phase 1 data analysis identified more than 88 million variants, of which 24 million are novel and 23 million are singletons. Data revealed several rare deleterious variants common in the Qatari population, and several variants that seemed to provide protection against diseases. Five non-admixed subgroups were identified in the Qatari population. In addition, hereditary genetic marker associations for 45 clinical traits were also identified. Subsequently, Phase 2 concluded in 2021 with over 25,000 whole genomes

sequences, and Phase 3 was initiated in 2022, with the aim of sequencing 100,000 genomes by 2025.

The QGP has given rise to multiple studies, all of which aim to advance precision medicine and pharmacogenomic development. These studies are supported by collaborative efforts between the Qatar National Research Fund and QGP. The QGP continues to contribute to the Path toward Precision Medicine grant program both administratively and financially. The PPM grant aims to support research that benefits from genomic data in understanding disease phenotype/genotype correlations and translating research findings into medical products to improve patient prognosis. Areas of focus include: immunogenomics/immunotherapeutic, clinical implementation/translation, pharmacogenomics, multi-omics, as well as digital e-solutions and applications [45].

To coordinate national collaborative efforts, the QGP has established partnerships with the leading national healthcare providers such as Sidra Medicine, the Primary Health Care Corporation, and HMC [46]. The QGP has also deployed multiple education initiatives that target the national science curricula, develop educational materials and videos, as well as organize education trips. In addition, QGP initiated 2 graduate programs with partner universities, a masters' program (MSc) and a masters' degree/doctoral degree (MSc/PhD) program in collaboration with Hamad Bin Khalifa University [43].

The QGP has launched several projects including the Q-Chip gene array, pharmacogenomics pilot studies, as well as genomic reports on wellness and general health [46]. These projects are conducted in collaboration with the QBB, the Department of Genetic Medicine at Weill Cornell Medicine-Qatar, the Diagnostic Genomic Division at HMC, as well as the Sidra research team at Sidra Medicine. These collective national efforts aim to translate the outcomes of basic genomic research into high impact deliverables at the clinical care end specific to the Qatari population and the Arab region [47]. To gain insights into the genetic architecture of health and disease-related quantitative traits, first genome-wide association studies (GWASs) of a list of 45 quantitative traits in 6,045 individuals from the Qatari population was conducted [48]. Data revealed multiple associations between Caucasian and Asian GWASs; uncovering differences in allele frequencies (AF) and linkage disequilibrium patterns for replicated loci. Further, novel genetic associations, mostly with variants common in the QGP but rare in other populations, have been identified. Larger GWASs are needed in the MENA region to accurately derive polygenic risk scores optimized for the Middle East.

The Sidra Bioinformatics Core developed a pipeline to perform sequencing analysis for QGP and other internal projects, after the data are received from the clinical genomic laboratory. QGP variant browser was established by the QGP team to provide 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 88,000,000 records with search and filter functionality on the QGP gene variants and its attributes (e.g., allele frequency, homozygosity etc.). The informed consent given by the study participants does not cover posting of participant-level phenotype and genotype data of QBB/QGP in public databases. However, access to QBB/QGP data can be obtained through an established International Organization for Standardization-certified process by submitting a project request according to the agreed regulations which is subject to approval by the QBB Institutional Review Board committee [49].

The Egyptian Genome Reference

The Egyptian Genome Reference (EgyptRef) was launched in 2020 in collaboration with the genetics and systems biology divisions of Lübeck Institute of Experimental Dermatology, Lübeck University, Germany and the Medical Experimental Research Center, Mansoura University, Egypt [24].

The program aimed to investigate population ancestry and ultimately advance personalized medicine healthcare in the country [50]. The program combines long- and short-read WGS technologies (i.e., PacBio, 10x Genomics, and Illumina) which were applied to samples taken from 10 healthy Egyptians (up to the third generation), who were patients of the Mansoura University hospital. The first Egyptian genome was assembled for one Egyptian male and was subsequently used to construct the Egyptian population reference genome from a larger sample set ($N = 109$) by considering genome-wide single nucleotide variants and allele frequency data [50].

Data revealed four major genetic ancestry components in Egyptians and 1,198 Egyptian population-specific variants, 49 of which were novel. EgyptRef findings also demonstrated that haplotypes for disease risk loci previously identified in European cohorts, differ from Egyptian haplotypes which may impact genetic risk within the population, as well as diagnostic accuracy when comparing against current databases [50]. Genotype principal component analysis showed a homogeneous group for which the assembly individual is representative of the Egyptian cohort. The authors report genetic characterization of the Egyptian population ($n = 5,429$ individuals) with respect to 143 other global populations using known variant data. Five datasets were combined to state the following:

1. 929 individuals part of the Human Genome Diversity Project (HGDP) which covers 52 populations representing Africans, Europeans, Western Central and Southern Asians, Oceanians, and Native Americans [51];
2. 2,504 individuals part of the 1000 Genomes Project, which covers 26 populations representing Africans, East and South Asians, Europeans, and Americans [52];
3. 108 individuals part of the study to investigate Qatari descent [53];
4. 478 individuals part of a study that analyzed SNP array-based variant data representing populations from the Arabian Peninsula [54];
5. 1,305 individuals part of a study that analyzed SNP array-based variant data representing 68 African, European, Western, and Southern Asian populations [55].

The data showed that Egyptians were located on the European-African axis and close to Europeans [24]. Their genetic variance spreads to a small degree in the direction of the Asian axis, akin to further individuals from the Middle East. Accordingly, the genetics of Egyptian individuals comprises four distinct ancestry components that sum up to 75% on average. Egyptians have a Middle Eastern, a European/Eurasian, a North African, and an East African component with 27%, 24%, 15%, and 9% relative influence, respectively.

As an example of personalized medicine for Egyptian-specific genetics, EgyptRef outcomes visualized the complete genetic and variant phasing information of the DNA repair-associated Breast Cancer gene 2 (BRCA2) by combining the using of novel technologies such as 10X Genomics. BRCA2 is linked to the progression and treatment of breast cancer and other cancer types [50].

All summary data of the Egyptian genome reference are available at www.egyptiangenome.org, where also variant AF can be queried online. Raw sequencing data and variant data are available at EGA under study ID EGAS00001004303 [50].

The Emirati Genome Program

The Emirati Genome Program (EGP) was launched in the first trimester of 2021, with the goal to generate deep genomics insights that are relevant for the Emirati population. The EGP is the first national-scale population genomics program being conducted in the UAE and where insights will be expected to enable UAE healthcare establishments to improve management of the local population's health and wellness [56]. For instance, the program will aim to provide Emirati individuals that have consensually participated in the program with personalized wellness and lifestyle insights. This effort will create a basis for proactive lifestyle and behavioral changes as well as preventive measures when followed up by selected clinical tests [57].

The EGP is a collaborative program between the Ministry of Health and Prevention, the Department of Health Abu Dhabi, the Dubai Health Authority, and Group 42's subsidiary branch, G42 Healthcare, and is being conducted at the G42 Healthcare Omics Center of Excellence [58, 59]. The program operates in line with the highest ethical and governance standards and ensures the anonymity of all participants and the date of their blood donation [59]. The goal of the EGP is to sequence 1+ million samples and has, to date, collected over 230,000 blood samples and buccal swabs, and sequenced over 214,000 samples in a record time of just 15 months [57].

The EGP samples are processed at the G42 Healthcare Omics Center of Excellence, a ~4,000 m^2 super lab, houses the largest and most advanced genomic technologies in the MENA region including platforms from technology giants Oxford Nanopore Technologies (ONT) [60], Illumina, and Beijing Genomics Institute Genomics (BGI, MGI) [56, 61], with a sequencing capability of over 50,000 whole human genomes per month. The combination of these technologies provides a comprehensive view of the human genome. The Illumina and MGI technologies bring the high-resolution and high-accuracy advantage of short-read sequences, enabling the identification of known and novel variants uniquely associated with the Emirati population [61], whilst the ONT technology brings the ability to access regions that are usually more difficult to reach via short-read sequencing and view larger structural changes in the genome [60]. The EGP sample sequencing data are stored at the G42 Healthcare facility which harbors Artemis supercomputer and has the largest computing resource for computational power and storage in the region [62]. Each genome sequenced at a depth of 90 Gb typically produces 150–750 Gb of raw data depending on the sequencing platform, further processed into various file formats, and quickly adding up to a total of >300 Gb, of which 10–50 Gb of data per genome is to be stored for many years. Therefore, file alignment like CRAM is used to achieve 40–70% compression of data which is an efficient format to store the high-throughput sequencing data of a large number of samples [63].

The EGP samples are stored at the biobank situated at the Omics Centre of Excellence to serve as a national repository of samples for research purposes [64]. The data obtained from the EGP program will be curated for the development and optimization of preventative medicine and early diagnosis national programs, such as in

the existing Premarital and Newborn Screening programs running across the UAE. The data obtained from EGP in collaboration with academia across the country will be used to generate population-scale insights and the identification of disease-specific genomic variations with the objective to enhance the healthcare system and tailor it to the needs of the local population [59]. Both local and global collaborations have been pivotal in launching research programs with a goal to improve healthcare in the UAE. Since the advent of the program, studies have been initiated to gain insights from data collected through the EGP; however, EGP being a relatively young program compared to the other population-wise genome projects in the MENA region does not have much published data or research findings in the public domain. Some of the ongoing projects are listed below:

- The Emirati Reference Genome which aims to create a reference database of variants and structural changes which are both common to the global population and unique to the Emirati population. Characterization of genomic markers will revolutionize healthcare systems, specifically regarding the improvement of treatment efficacy and patient access [57].
- Data mining research programs combining WGS and electronic health record (EHR) data to identify disease-/health-specific mutations and regional prevalence. These initiatives require the involvement of different experts tackling a wide range of topics, from metabolic, cardiovascular, and autoimmune diseases to pharmacogenomics and longevity studies.
- Initiatives in precision medicine across various therapeutic areas such as cancer and rare diseases, facilitated by partnerships between local healthcare entities and EGP stakeholders to integrate sequencing data into patient health records. Such initiatives aim to improve diagnostic and therapeutic decision-making capabilities and provide more accurate patient prognosis [65].
- The first UAE-Israeli joint population genomic research involving the analysis of WGS and EHR data in individuals suffering from vitamin D deficiency [66].

Other Ongoing PGPs and Initiatives in the MENA Region

Kingdom of Bahrain National Genome Center

The Bahraini Genome Center was established as a specialized center for genetic analysis to protect the Kingdom of Bahrain from illness and prevent diseases in current and future generations. The center aims to generate a database of the Bahraini population's DNA to identify opportunities for early diagnosis and treatment. The program also aims to lower the population's risk of contracting diseases [27]. Participation in the Bahraini National Genome Center is open to all citizens above 21 years old, regardless of their health status. Individuals under 21, diagnosed with a rare genetic disease or other genetic diseases, may participate if referred by their physician [67]. The Bahraini National Genome Center has achieved the first phase of its program by collecting and biobanking 6,000 biological samples, including 2,000 samples from people with rare diseases and their families [68]. The program announced the completion of the second phase by the end of the year 2021, although it is unclear what the details of phase 2 entail. Overall, the Bahraini National Genome Center aims to collect 50,000 samples within 5 years and establish a comprehensive national database.

In addition to drug discovery, the Bahrain National Genome Center Program conducted coronavirus genome-related studies and investigated the genome of infected individuals in order to identify and characterize genes related to the severity of the infection. To date, no published data have been made available.

Iran Genome Program

The Iran Genome (Iranome) program database was established in 2019, representing the population of Iran, the second largest population of Middle East. WES was performed on 800 healthy individuals from eight major Iranian ethnic groups. Country-specific aims were identified, such as ancestry/ethnic studies that are applied in the Iranome [28]. Participants aged over 30 years old, whose ancestors were born in Iran and of "pure race" (at least two generations, or four grandparents) were registered in the program anonymously. Collected samples were sequenced using the Illumina paired-end sequencing technology [69]. The main values of the Iranome program are as follows: (a) understand Iranian ethnicities' history and identity, (b) investigate the genetic predisposition or resistance to endemic diseases of the Iranian population, (c) link data with anthropologists', archaeologists', biologists', linguists', and historians' information to bridge the gap between science and humanities, and (d) understand the nature of differences between human populations in Iran and provide a better understanding of differences between ethnicities [70].

In the first phase of the Iranome program, 500 samples were collected from the main Iranian ethnic groups and sequenced. This resulted in a comprehensive catalog of Iranian genomic variations and the construction of the Iranome database, the first public database of AF of

genomic variants in the Iranian population provided by WES. The results identified 1,575,702 variants of which 308,311 were novel (19.6%). Also, by presenting a higher frequency for 37,384 novel or known rare variants, the Iranome database can improve the power of molecular diagnosis. To provide rapid and free access to such data, all the variants identified were made publicly available to the scientific community through a web-based genomic variation browser at the Iranome website [28]. Approximately 70% of the variants identified in the Iranome database were novel or had frequencies less than 1% (rare variants) in public databases. Regarding the prominence of such variants in terms of diagnosis and management of patients suffering from rare Mendelian disorders, the Iranome database offers a comprehensive healthcare resource at the national level by providing population-specific AF of such variants. The data can also be visualized from an ethnic-specific perspective, which can be useful while interpreting the variants identified in patients coming from specific Iranian ethnic groups. This database is an excellent resource for other countries in the MENA region and further due to the geographical spread of various Iranian ethnicities. The database includes: (i) 100 Iranian Persians, useful for Persian populations living in Afghanistan, Tajikistan, the Caucasus, Uzbekistan, Bahrain, Kuwait, and Iraq, (ii) 100 Iranian Kurds, for Kurdish populations living in Iraqi Kurdistan, Turkey, Syria, Armenia, Israel, Georgia, and Lebanon, (iii) 100 Iranian Baluchs, for Baluchi populations living in Pakistan, Oman, Afghanistan, Turkmenistan, Saudi Arabia, and the UAE, and (iv) 100 Iranian Azeris, for populations living in Azerbaijan, Turkey, Russia, and Georgia. Although Iranian Arabs are admixed with other ethnicities in Iran such as Persians, Turks, and Lurs, the genetic variants identified in 100 Iranian Arabs investigated in this study can also be a useful resource for the populations of Arab countries which are geographically close to Iran.

Challenges

PGPs in the MENA region are associated with several challenges. Some of those key challenges include data security and privacy (including data sharing), non-standardization of data collection, and analysis. Efforts must be made to overcome the challenges of accessibility and integration in order to successfully utilize and adopt the interpretation of genomic data into valuable translation of novel findings into healthcare systems.

Open science initiatives, such as the Global Alliance for Genomics and Health, were established to address the importance of applying common standards and methods to use genomic and related data [71]. This is important to ensure the accuracy of data yielded from genome programs.

We recommend 6 key priority steps that should be taken to improve future PGPs and ensure their success. These key steps are related to collaboration, finance, infrastructure and institutional data access, data analysis, sustainability, and health records and biobanks.

Collaboration

For most of the case studies we have presented here, collaborations between local investigators and those from research-intensive nations were critical for the success of population-wide genome projects. Collaborations can provide diverse expertise including competitive research track records, experience in grant writing, administrative support, and the necessary local expertise and knowledge about the target population. Therefore, networking and building long-lasting productive collaborations remains key for investigators to access funding for large-scale genomics research. However, the potential for power imbalance and cultural sensitivity needs to be considered when establishing collaborations with different institutes When capacity building is incorporated, establishing collaborations with research-intensive nations will enhance transfer of knowledge and expertise enabling more genomic research led by investigators in low- and middle-income countries (LMICs). Moreover, data-sharing agreements are important to ensure the interests of the local researcher are respected.

Increasing diversity in genomic studies contributes to more robust findings from replicated results as well as new discoveries, particularly when combined with existing large-scale studies. Developing local research capacity enables contributions to global genomics consortia, as demonstrated in several existing consortia such as the Global Lipids Genetics Consortium [72], the Genetic Investigation of Anthropometric Traits (GIANT) Consortium, Psychiatric Genomics Consortium, and other major initiatives. These have dual and mutual benefits by enabling the discovery of ancestry-specific findings, raising the profile of these findings to a broader audience, and enhancing the careers of local contributing investigators. Participation in global consortia by diverse groups requires trust, which can only be built when all contributors benefit.

Finance

Genomic research is expensive, making it a secondary priority for funding in LMICs. One route toward greater inclusion of underrepresented populations is by leveraging

funding mechanisms from international institutions and those in research-intensive nations. Funders have an opportunity to help address imbalances in global genomics research through their research priorities; dedicated funding calls, such as the "Diversity Centers for Genome Research" by the National Institutes Health in the United States, and also the National Human Genome Research Institute's multiple funding programs supporting genomic research, can be a strategic tool to empower fast progress.

Infrastructure and Institutional Data Access

Many funding calls are exclusively targeted to researchers at institutions in the funder's country. Given the immense and wide-reaching benefits of increasing diversity in genetic research, funders should reconsider such restrictions. In addition to eligibility restrictions, fewer researchers in LMICs have track records competitive for large funding calls due to the limited research capacity, infrastructure, and funding at their local institutions. This predicament makes it very difficult for those researchers to build up large genomic studies without collaborators from research-intensive nations.

To conduct cohort-based genomic research, it is not only critical to access some key infrastructure components but to also align the study with the legal, administrative, and ethical frameworks applicable at the institutional and national levels. A comprehensive understanding of ethical concerns, regulations, and policies could enable researchers to avoid major delays in cross-border shipping of biological samples and to ensure the ability to reuse/share these valuable datasets in future. Most of the studies described above report pre-study consultation with legal experts (often available via their institutions) and implementation of necessary material and data transfer agreements to ensure efficient movement of samples and data. Infrastructure for steps such as sample processing, biobanking, genotyping or sequencing, and computational analysis are often outsourced or accessed via local and international collaborations. However, access to relevant infrastructure at the institutional level could be a major form of capital investment requiring continuous funding for study and future research.

Data Sharing and Analysis

Data sharing is an important aspect of population genomic programs, especially for medical research, that will eventually aid progress toward precision medicine applications in the associated countries. Curation of genomic information obtained from sequencing for public, academic, research, medical, and/or commercial use can have various levels of access. Data sharing needs to be regulated after adequate consideration of ethical and legal issues, identifying stakeholders as well as technical aspects and data security. Established data sharing protocols along with the advancement of genomic technology can pave the way for designing algorithms for disease identification, diagnostic predictions, and suggested treatments using artificial intelligence and machine learning algorithms.

Sustainability

Sustainability is a primary concern for genomic studies as most research requires a sustained funding for 2–3 years for the completion of the research. Further many funding calls do not provide a dedicated capacity-building component which can enhance local research capacity for long-term benefits, such as hiring local students or researchers for training or research positions.

EHRs and Biobanks

The true value of insight generation from data obtained through genomics projects starts with the ability to link the EHRs and additional metadata, clinical, environmental, and self-reported, to genomic information. The power of statistical relevance lies in the cohort size and supporting data made available for analysis. Furthermore, the integration of biobanking capabilities as part of, or underlying, genomics programs is proving to be of much greater value to the future of precision healthcare in the multi-omics and diagnostics sector.

Conclusion

This review has identified 4 ongoing PGPs in the MENA region. Significant progress has been and is continuously being made in establishing high-throughput and high-quality PGPs in the region. The increase in activity in the population-scale genomics program sector allows us to eagerly await and expect research articles to be published in scientific journals in the near future. Subsequently, we expect to see an increase in the representation of genomes of Middle Eastern descent, specific to the MENA region, in public genetic databases. Current and future population-scale genomics projects are advised to take into consideration the various challenges highlighted in this review, with the aim to ensure ethical, regulated, and efficient conduct. Ultimately these initiatives will drive developments in personalized medicine treatments to improve patient prognosis and quality of life.

Acknowledgments

The authors thank all individuals who have participated in the national genome programs initiated in the MENA region aiming for a better healthcare future. The authors also thank the respective programs that have made information publicly available.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The authors received no financial support for the research, authorship, and/or publication of this work.

Author Contributions

Hagar Ateia has made substantial contributions to the conception and design of the work, acquisition of data, and writing the main manuscript. Pauline Ogrodzki has contributed to the conception of the work and revising the manuscript for important intellectual content. Hannah Wilson has contributed to the design of the work, drafting, and revising the manuscript for important intellectual content. Subhashini Ganesan and Rabih Halwani have contributed to the conception and design of the work and revising the manuscript for important intellectual content. Ashish Koshy has contributed to the conception of the work, acquisition of data, and substantively revised the work. Walid Zaher has made substantial contributions to the conception and design of the work, acquisition of data, and revising the work. All authors have approved the final submitted version of the work. All authors have agreed to be personally accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1 Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science. 2003;300(5617):286–90.
2 Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435–44.
3 Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. Cell. 2019;179(4):984–1002 e36.
4 Manolio TA, Bult CJ, Chisholm RL, Deverka PA, Ginsburg GS, Jarvik GP, et al. Genomic medicine year in review: 2019. Am J Hum Genet. 2019;105(6):1072–5.
5 Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, et al. Integrating genomics into healthcare: a global responsibility. Am J Hum Genet. 2019;104(1):13–20.
6 Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96–102.
7 Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-scale whole-genome sequencing of three diverse asian populations in Singapore. Cell. 2019;179(3):736–49 e15.
8 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. Nature. 2020;581(7809):434–43.
9 Abou Tayoun AN, Rehm HL. Genetic variation in the Middle East-an opportunity to advance the human genetics field. Genome Med. 2020;12(1):116.
10 Jones KM, Cook-Deegan R, Rotimi CN, Callier SL, Bentley AR, Stevens H, et al. Complicated legacies: the human genome at 20. Science. 2021;371(6529):564–9.
11 Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016;7:13637.
12 Shehreen S, Chyou TY, Fineran PC, Brown CM. Genome-wide correlation analysis suggests different roles of CRISPR-Cas systems in the acquisition of antibiotic resistance genes in diverse species. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180384.
13 Landolt L, Strauss P, Marti HP. Next generation sequencing: a tool for this generation of nephrologists. EMJ. 2016;1(2):50–7.
14 Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature. 2019;571(7766):489–99.
15 Scott RH, Fowler TA, Caulfield M. Genomic medicine: time for health-care transformation. Lancet. 2019;394(10197):454–6.
16 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
17 Salamanca-Gomez F. [The human genome sequence]. Gac Med Mex. 2001;137(3):267–8.
18 Shamia A, Shaheen R, Sabbagh N, Almoisheer A, Halees A, Alkuraya FS. Revisiting disease genes based on whole-exome sequencing in consanguineous populations. Hum Genet. 2015;134(9):1029–34.
19 Farra N, Manickaraj AK, Ellis J, Mital S. Personalized Medicine in the Genomics Era: highlights from an international symposium on childhood heart disease. Future Cardiol. 2012;8(2):157–60.
20 Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–5.
21 de Bono JS, Ashworth A. Translating cancer research into targeted therapeutics. Nature. 2010;467(7315):543–9.
22 Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, et al. Whole-genome sequencing for optimized patient management. Sci Transl Med. 2011;3(87):87re3.
23 Emirati Genome Program. Available from: https://emiratigenomeprogram.ae/# (accessed April 29, 2022).
24 EgyptRef Egypt genome reference. Available from: https://www.egyptian-genome.org (accessed August 25, 2022).
25 Qatar Genome Programme. Available from: https://qatargenome.org.qa/node/5 (accessed April 28, 2022).
26 Saudi Human Genome Program SHGP. Available from: https://shgp.kacst.edu.sa/index.en.html (accessed April 28, 2022).
27 National Genome Center, Kingdom of Bahrain Ministry of Health. Available from: https://www.moh.gov.bh/GenomeProject?lang=en (accessed April 28, 2022).
28 Iranome. Available from: http://www.iranome.ir/ (accessed April 28, 2022).
29 Project Team S.G. The Saudi Human Genome Program: an oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse. 2015;6(6):22–6.
30 Alrefaei AF, Hawsawi YM, Almaleki D, Alafif T, Alzahrani FA, Bakhrebah MA. Genetic data sharing and artificial intelligence in the era of personalized medicine based on a cross-sectional analysis of the Saudi human genome program. Sci Rep. 2022;12(1):1405.

31 Saudi human genome program SHGP objectives. Available from: https://shgp.kacst.edu.sa/index.en.html#program-objectives (accessed May 13, 2022).
32 Saudi human genome program technologies. Available from: https://shgp.kacst.edu.sa/index.en.html#technologies-bioinformatics (accessed May 12, 2022).
33 Abedalthagafi MS. Precision medicine of monogenic disorders: lessons learned from the Saudi human genome. Front Biosci (Landmark Ed). 2019;24(5):870–89.
34 Saudi Human Genome Program Achievements Available from: https://shgp.kacst.edu.sa/index.en.html#achievements (accessed May 12, 2022).
35 Memish ZA, Saeedi MY. Six-year outcome of the national premarital screening and genetic counseling program for sickle cell disease and beta-thalassemia in Saudi Arabia. Ann Saudi Med. 2011;31(3):229–35.
36 Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet. 2021;53(7):942–8.
37 Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. Revisiting the morbid genome of Mendelian disorders. Genome Biol. 2016;17(1):235.
38 Saudi human genome program "premarital screening program" PSP. Available from: https://screening.shgp.sa/?lang=en (accessed May 14, 2022).
39 Ng SW, Zaghloul S, Ali HI, Harrison G, Popkin BM. The prevalence and trends of overweight, obesity and nutrition-related non-communicable diseases in the Arabian Gulf States. Obes Rev. 2011;12(1):1–13.
40 Shaik AP, Shaik AS, Al-Sheikh YA. Colorectal cancer: a review of the genome-wide association studies in the kingdom of Saudi Arabia. Saudi J Gastroenterol. 2015;21(3):123–8.
41 Mbarek H, Devadoss Gandhi G, Selvaraj S, Al-Muftah W, Badji R, Al-Sarraj Y, et al. Qatar genome: insights on genomics from the Middle East. Hum Mutat. 2022;43(4):499–510.
42 Al Thani A, Fthenou E, Paparrodopoulos S, Al Marri A, Shi Z, Qafoud F, et al. Qatar biobank cohort study: study design and first results. Am J Epidemiol. 2019;188(8):1420–33.
43 Qatar Genome Programme Educational Goals. Available from: https://www.qatargenome.org.qa/genomic-medicine/educational-programmes/graduate-programmes (accessed April 29, 2022).
44 Sidra Medicine. Available from: (https://www.sidra.org) (accessed April 29, 2022).
45 Path towards precision medicine PPM projects. Available from: (https://www.qatargenome.org.qa/research/path-towards-precision-medicine/ppm-projects) (accessed April 29, 2022).
46 Qatar Genome Programme Journey. Available from: https://www.qatargenome.org.qa/about-qgp/qatar-genome/journey (accessed April 29, 2022).
47 Qatar Genome Programme QChip. Available from: https://www.qatargenome.org.qa/translational-genomics/qchip (accessed April 29, 2022).
48 Thareja G, Al-Sarraj Y, Belkadi A, Almotawa M; Qatar Genome Program Research QGPR Consortium, Suhre K, et al. Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. Nat Commun. 2021;12(1):1250.
49 Qatar Biobank/Qatar Genome Programme Data Access Request. Available from: (https://www.qatarbiobank.org.qa/research/how-apply) (accessed April 29, 2022).
50 Wohlers I, Künstner A, Munz M, Olbrich M, Fähnrich A, Calonga-Solís V, et al. An integrated personal and population-based Egyptian genome reference. Nat Commun. 2020;11(1):4719.
51 Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012.
52 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(7571):68–74.
53 Rodriguez-Flores JL, Fakhro K, Agosto-Perez F, Ramstetter MD, Arbiza L, Vincent TL, et al. Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Res. 2016;26(2):151–62.
54 Fernandes V, Brucato N, Ferreira JC, Pedro N, Cavadas B, Ricaut FX, et al. Genome-wide characterization of arabian Peninsula populations: shedding light on the history of a fundamental bridge between continents. Mol Biol Evol. 2019;36(3):575–86.
55 George B. Genotype data for a set of 163 worldwide populations. 2020.
56 DoH and G42 announce Genome Program.
57 Emiratis urged to submit DNA samples for genome project to tackle deadly diseases.
58 Emirati genome program partners with SEHA to ramp up participation across Abu Dhabi.
59 Emirati genome program brochure for healthcare regulators and hospitals.
60 Large-scale population genomics project announced by department of health in Abu Dhabi, using PromethION.
61 MGI participates in world's most comprehensive genome program in Abu Dhabi.
62 G42's Artemis Supercomputer ranks #26 in the world. Available from: https://g42.ai/news/cloud/news-g42s-artemis-supercomputer-ranks-26-in-the-world/ (accessed June 27, 2022).
63 Al Ali A, Kandavel PK, Al Mabrazi H, Carvalho G, Kusuma V, et al. CRAM compression: practical across-technologies considerations for large-scale sequencing projects. bioRxiv. 2022.
64 Inauguration of G42 Omics Center of Excellence. Available from: https://emiratigenomeprogram.ae/news/7/%D8%AA%D8%AF%D8%B4%D9%8A%D9%86_%D9%85%D8%B1%D9%83%D8%B2_%D8%A3%D9%88%D9%85%D9%8A%D9%83%D8%B3_%D9%84%D9%84%D8%AA%D9%85%D9%8A%D8%B2_%D9%84%D9%84%D8%B9%D9%84%D9%88%D9%85_%D8%A7%D9%84%D8%A8%D9%8A%D9%88%D9%84%D9%88%D8%AC%D9%8A%D8%A9 (accessed April 30, 2022).
65 UAE Information_ Research in the field of health: The Emirati Genome Programme.
66 KSM and G42_ First Joint Genome Study Available from: https://www.g42healthcare.ai/latest-update/first-ever-uae-israel-genetic-research-collaboration-with-kahn-sagol-maccabi-ksm/ (accessed April 30, 2022).
67 Bahrain Genome Consent. Available from: https://www.moh.gov.bh/GenomeProject/TakingPart (accessed April 29, 2022).
68 Bahrain Genome Journey_ News. Available from: https://www.moh.gov.bh/GenomeProject/NewsDetails/4825 (accessed April 30, 2022).
69 Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: a catalog of genomic variations in the Iranian population. Hum Mutat. 2019;40(11):1968–84.
70 Banihashemi K. Iranian human genome project: overview of a research process among Iranian ethnicities. Indian J Hum Genet. 2009;15(3):88–92.
71 The global alliance for genomics and health (GA4GH). Available from: https://www.ga4gh.org/ (accessed April 30, 2022).
72 Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514–23.