wellcome

Data and diversity in genomics

Landscaping report

October 2024

Image credit: Jack Cole/Wellcome

Table of contents

Forward

In 2023, Wellcome commissioned IQVIA to conduct landscaping analysis on the state of diversity in the global genomics field. Following extensive research and interviews, we present here the findings on current representativeness of human genomic datasets and opportunities for funders to make a positive impact.

Genomics projects and human genomics datasets worldwide are heavily skewed towards populations with European ancestry, usually based in the Global North and lacking information linked to the socio-demographic status of the communities where the genomics data is coming from. This has detrimental consequences for biological understanding, clinical insight, and for equitable and inclusive scientific practice. The lack of diversity in global human genomic data reflects a complex set of inter-related factors including, but not limited to, structural barriers and differing levels of trust and research engagement across communities and populations.

Wellcome has a history of funding genomics and genome-related research, from large-scale investments in the Human Genome Project and the Wellcome Sanger Institute to supporting an array of research teams, resources and projects to tackle different scientific challenges. Wellcome is a global funder, which necessitates a global perspective, as different regions will encounter different challenges, and present different opportunities.

This report details current state of human genomic datasets globally and proposes distinct levels of “genomic maturity” in different geographical regions. Additionally, it summarises opportunities via which stakeholders can ensure that genomic datasets are more representative of the global populations. We believe that the content of this report will be of interest to a diverse range of audiences, including researchers, funders as well as policy experts.

Ekin Bolukbasi
Technology Manager at Wellcome

Executive summary

Definition of diversity in genomic research

Diversity in genomic research is defined as targeting underrepresented populations, analysing sub-population data, cultivating a workforce with different backgrounds and cultural origins, forming international collaborative networks and reaching to communities beyond cultural reach

Genomic diversity archetypes

Three archetypes were defined based on level of addressing diversity:

- High maturity (e.g., United States (US), European Union (EU)): large database size, focusing on chronic disease etiology research and precision medicine
- Medium maturity (e.g., Japan, Taiwan): medium database size, focusing on regional understanding of genetic factors for disease etiology
- Low maturity (e.g., Brazil, Uganda): low database size, focusing on expanding databases to understand chronic disease etiology

Key opportunities for enhancing data diversity globally

- Encourage timely community engagement and socioeconomic data collection
- Support training on bioinformatics and genetic counselling
- Fund genomic sequencing in low maturity regions

The objectives of this project were:

1. Mapping what Wellcome has supported in the genomic data diversity space to date
2. Mapping the nature and representativeness of global genomic datasets and their impact on communities
3. Mapping opportunities to enhance the current genomic data diversity landscape
4. Generate recommendations and case studies to support prioritising opportunities

Wellcome sought to understand and map genomic sample diversity and assess opportunities to enhance it globally.

IQVIA addressed this objective by:

1. Developing a long list of 440 global initiatives, alongside Wellcome
2. Prioritising up to 200 relevant initiatives for further desk research
3. Sending a survey to prioritised initiatives, which received 55 responses
4. Conducting 45 minute qualitative interviews with 27 key initiative representatives across the globe

Methodology and sample sizes

Overview of diversity in genomic research

IQVIA followed a sequential approach to build a diversity profile for 198 initiatives through desk research and survey

Activity

Consolidated long list of global genomics initiatives

440 initiatives identified in a consolidated long list from Wellcome and IQVIA

Focused long list enhanced with in-depth desk research

Up to 200 initiatives relevant to Wellcome's objectives were prioritised for further desk research by agreeing on exclusion criteria**

Genomic diversity assessment survey

55 unique responses to the diversity assessment survey were received after three weeks of outreach & follow-up

Qualitative interviews

45-minute online discussions with key initiative representatives (N=27)

Considerations

Use of existing data in IQVIA genomics database and Wellcome's list of funded initiatives to develop long list of genomics initiatives

Exclusion criteria (non-exhaustive):

- 47 commercial initiatives with restricted data access policies
- 67 focused therapy area initiatives
- 64 technology/analysis methodology studies
- 28 other reasons (e.g., initiative ended, focused on policy)
- 8 non-human and archaeology studies
- 28 non-geographical interest (e.g., China)

Factors contributing to the response rate included:

- 4 initiatives answered for more than 1 initiative in the focused long list
- 1 initiative was not able to complete survey on time
- 1 initiative provided duplicated responses
- 14 initiatives were not interested/not contactable

Overview of the geographical coverage of the initiatives and initiative type reviewed in this research

Europe

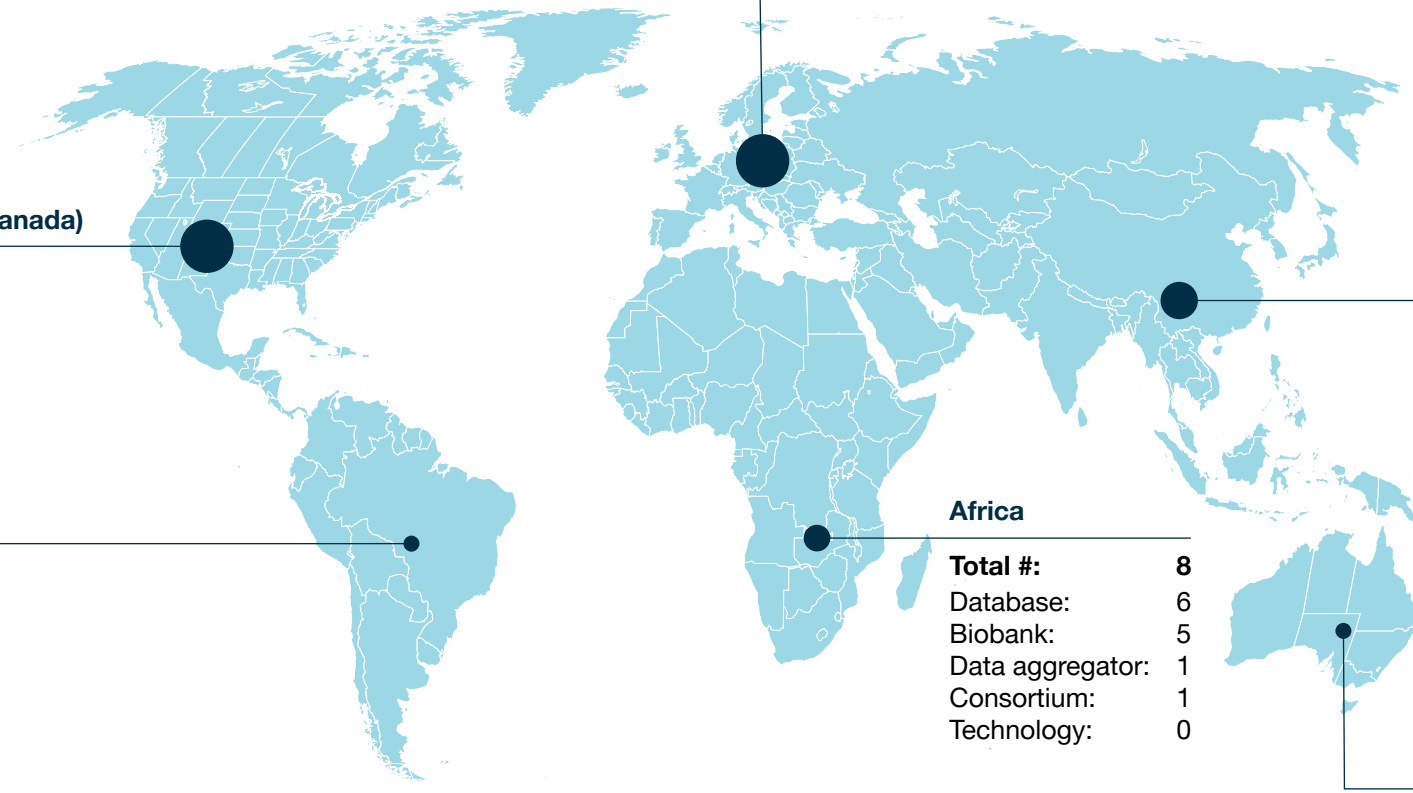
Total #:	**52**
Database:	31
Biobank:	21
Data aggregator:	10
Consortium:	9
Technology:	9

North America (US, Canada)

Total #:	**54**
Database:	36
Biobank:	16
Data aggregator:	13
Consortium:	8
Technology:	7

Latin America

Total #:	**8**
Database:	6
Biobank:	5
Data aggregator:	1
Consortium:	1
Technology:	0



Africa

Total #:	**8**
Database:	6
Biobank:	5
Data aggregator:	1
Consortium:	1
Technology:	0

Global (≥2 regions)

Total #:	**28**
Database:	13
Biobank:	1
Data aggregator:	12
Consortium:	13
Technology:	7

Asia & Middle East

Total #:	**35**
Database:	30
Biobank:	18
Data aggregator:	1
Consortium:	5
Technology:	5

Oceania

Total #:	**5**
Database:	2
Biobank:	2
Data aggregator:	1
Consortium:	2
Technology:	1

Key takeaways

- Biobanks are limited in Africa and Asia (considering the size of the population)
- Data aggregation limited in a lot of underserved geographies
- Few regional consortiums in Africa and Latin America (seems to mostly be happening on a Global level), may impact the utilisation of data into local health systems/policies

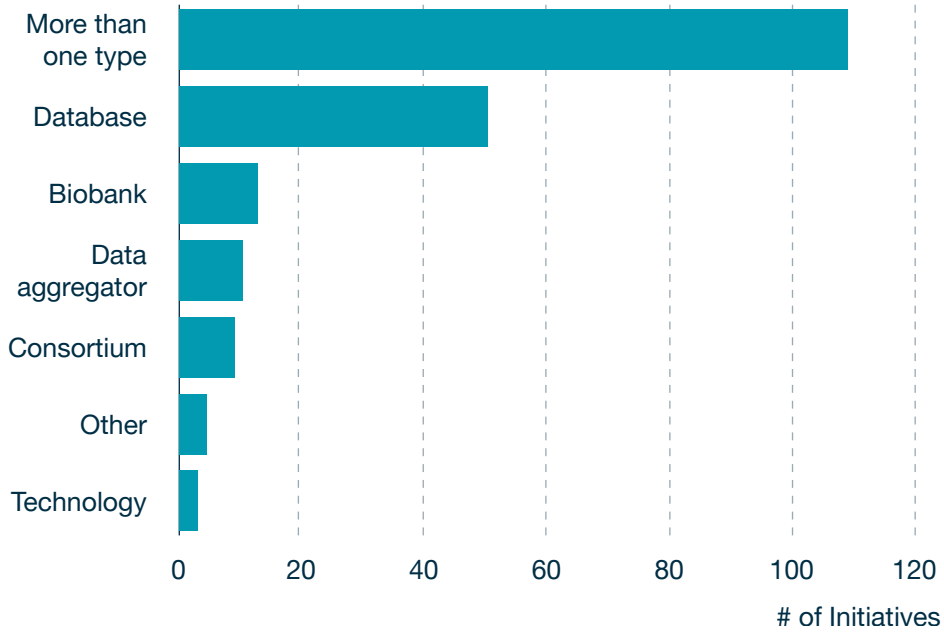
Opportunity

- Improving sharing and cross collaboration within regions

Research caveat: above representation is based on and limited to the initiatives included in this research and does not cover the full global spectrum of initiatives

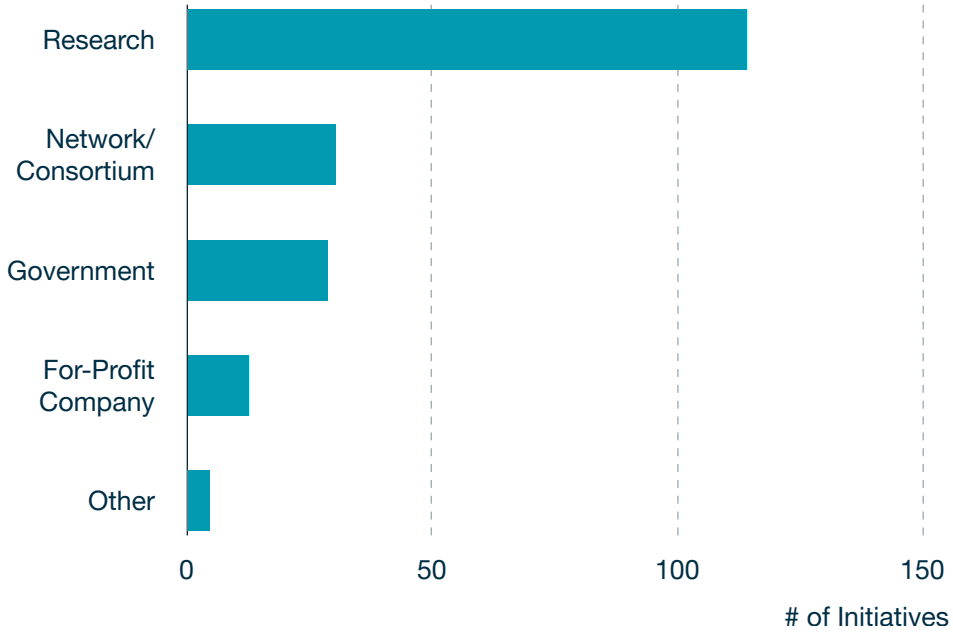
The focused long list of initiatives included a range of initiative types, led by different types of organisations

Type of Initiatives



- Most initiatives can be classified as more than one type
- Database is most common representing 51 initiatives

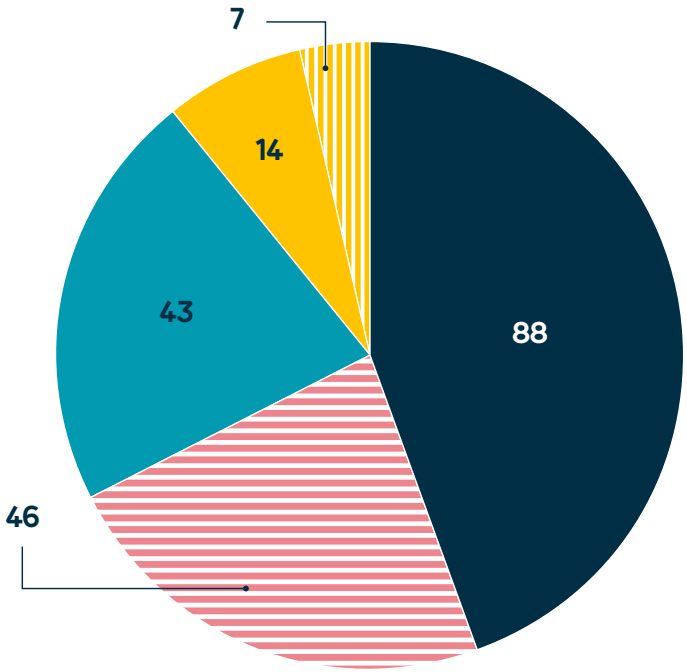
Type of Organisation



The majority of initiatives are led by research organisations

Regarding initiative impact, initiatives generally capture at least one type of genomic data and have favourable data access policies

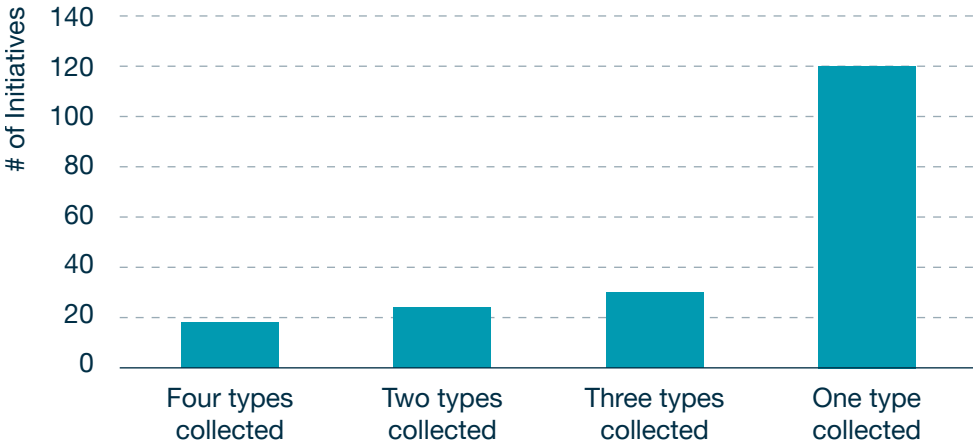
Data Access



At least 138 initiatives allow data access for research, including publicly available datasets, availability to researchers via application or mixed (e.g., subset of data is available)

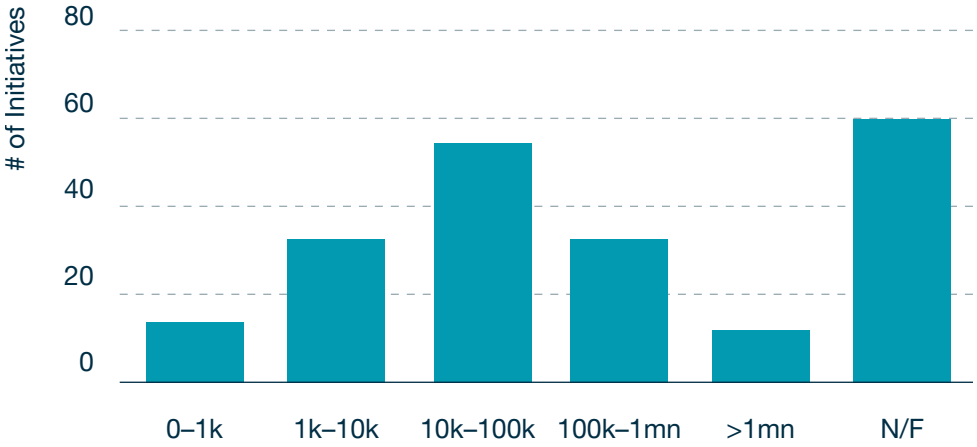
*WGS: whole genome sequencing **WES: whole exome sequencing

Genomic Data Type Collected



80 initiatives have WGS*, 56 initiatives have WES**, 35 collect biological samples

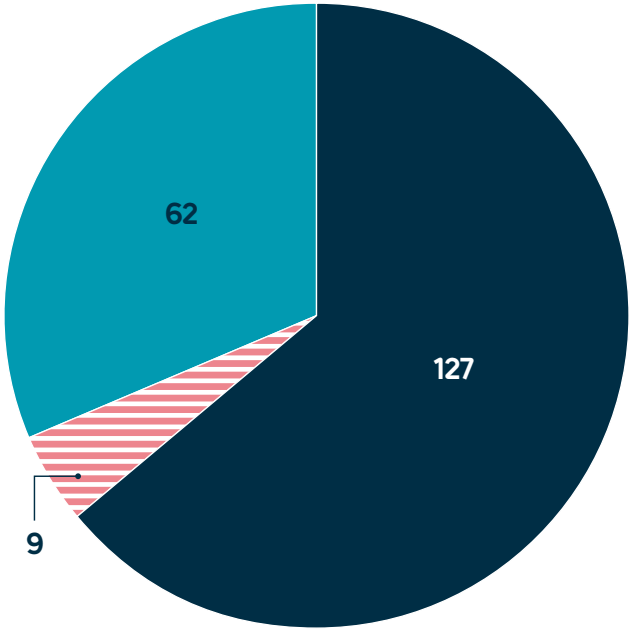
Cohort size



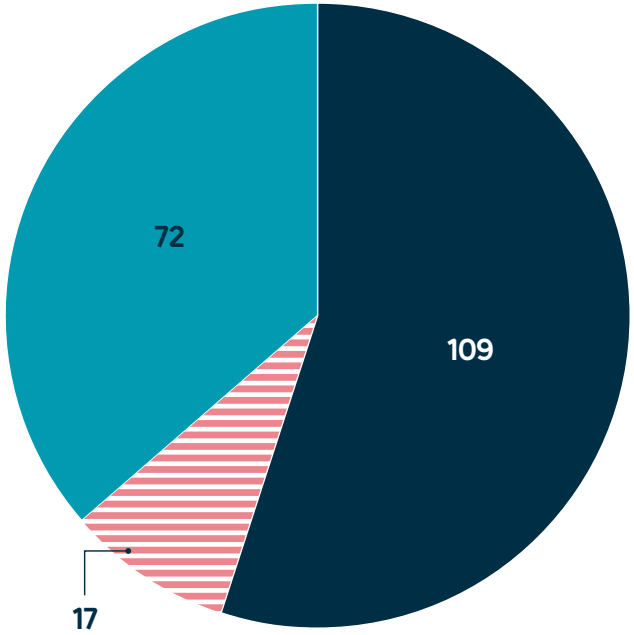
10k–100k participants is the most common category of cohort size

Regarding diversity value, variables relating to demographics and health information are more commonly captured

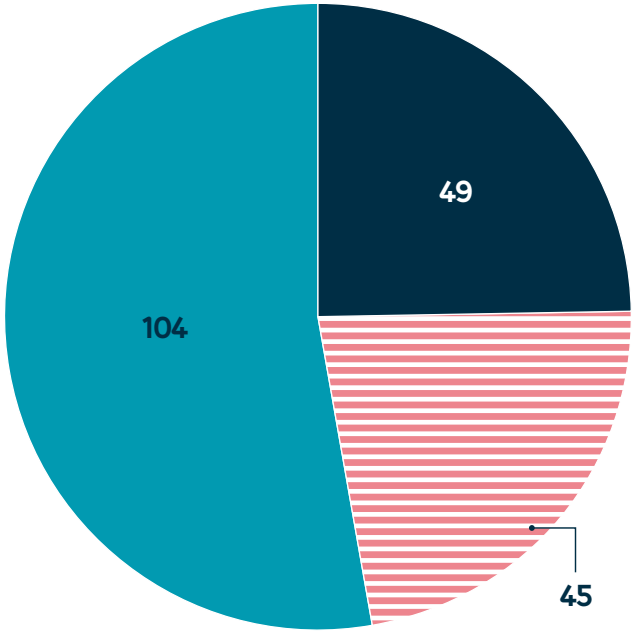
Availability of demographics



Availability of health information



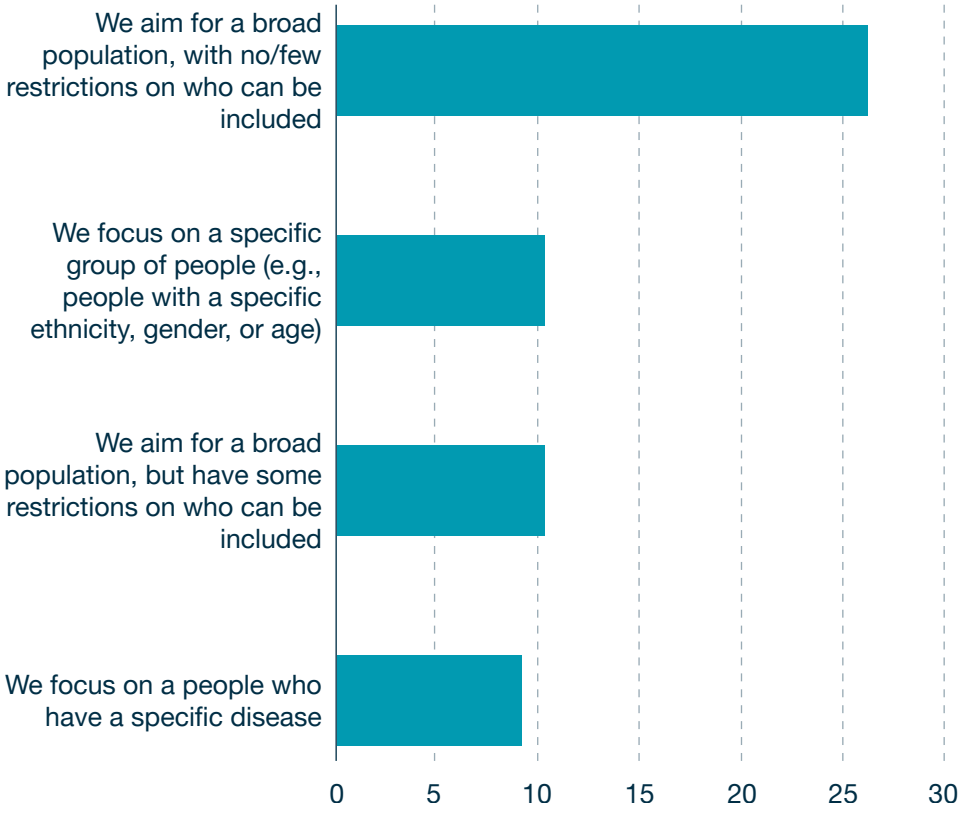
Availability of socio-economic information



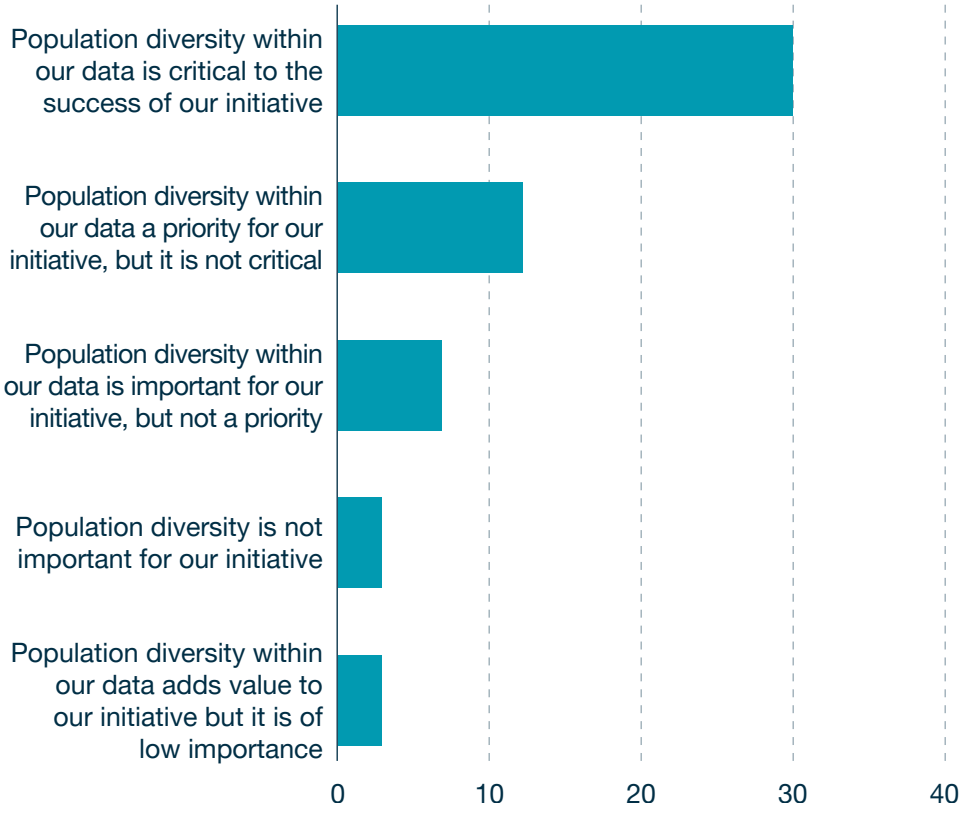
Demographic and health information are captured by many initiatives, but socioeconomic information is limited, with more initiatives reporting non-capture of socioeconomic information.

55 initiatives responded to our survey, most stating that they aim for broad representation and view population diversity as key to their success

Population of Interest (n=55)

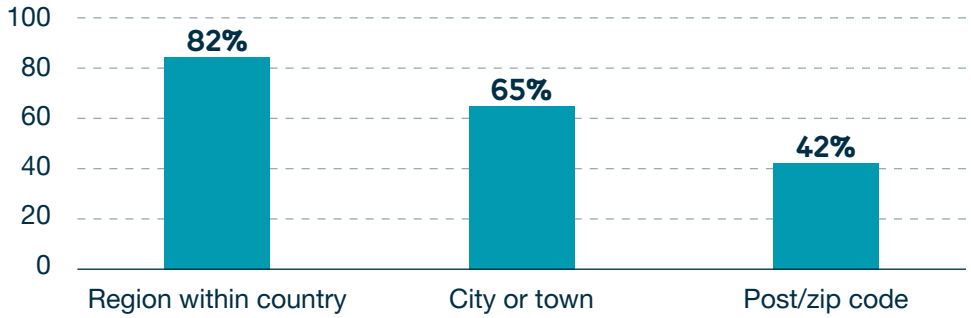


Level of Importance (n=55)



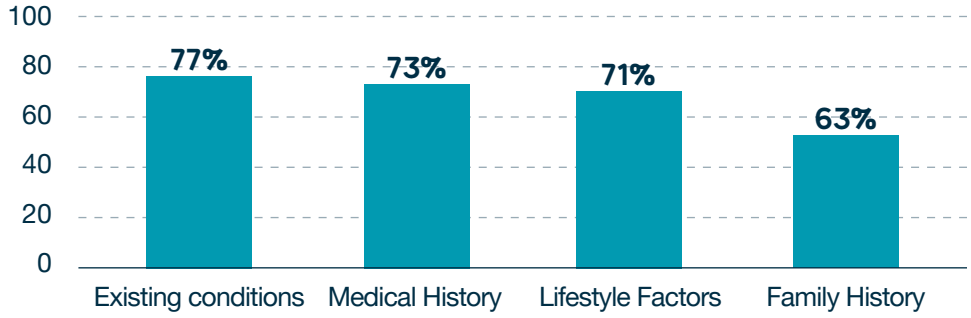
In survey respondents, region, age, gender and ethnicity are most commonly collected, whereas data on education, occupation, income and housing status are collected much less

95% collect geographical information



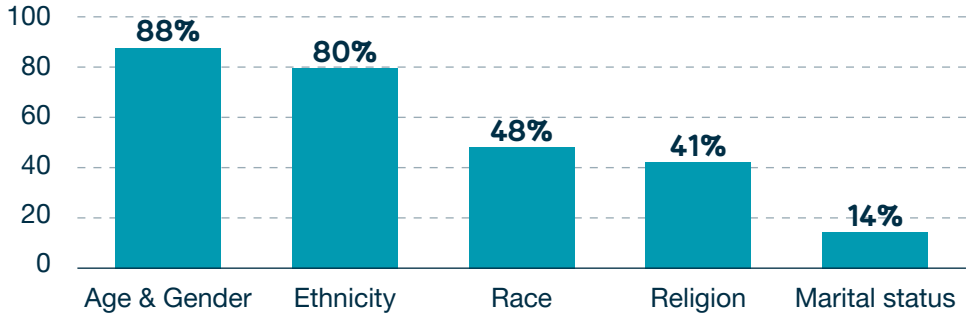
Mainly in North America and EU

82% collect health information



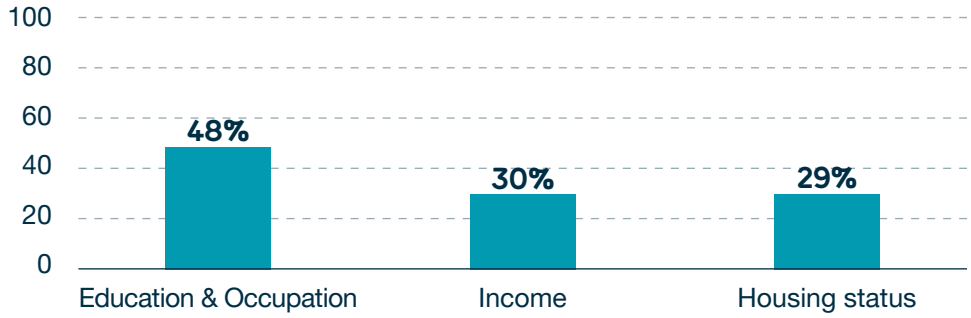
Mainly in North America, EU

95% collect demographical information



Mainly in North America, EU and Asia

51% collect socioeconomic information



Mainly in North America and EU

Socioeconomic data should be collected more regularly to inform downstream policy, funding and clinical decision making.

IQVIA then spoke with representatives from 27 initiatives from four regions to gain deeper insight into the matter of diversity in genomics and their efforts to achieve it. Responses from these interviews have been anonymised.

Diversity extends beyond recruiting a representative sample into purposeful data analysis and expansion of voices and perspectives in the organisations

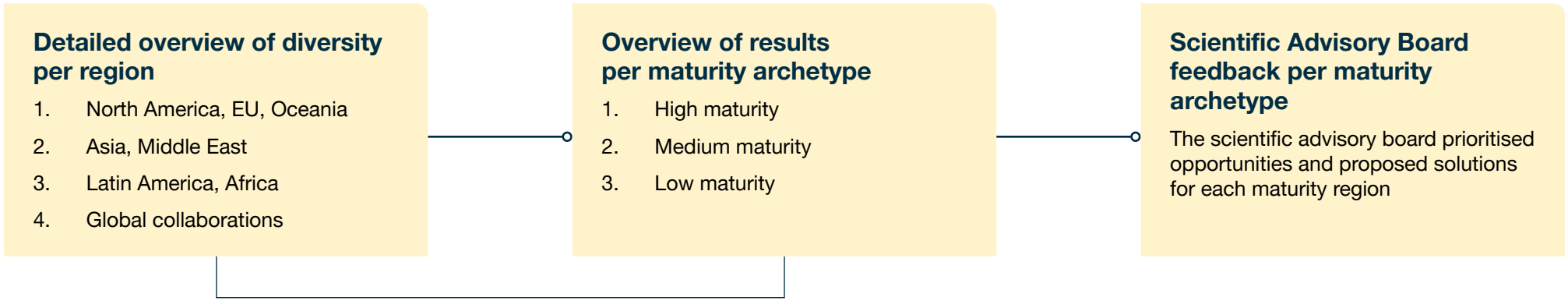
Different layers of Diversity in Genomic data

Diversity in recruitment	Recruiting data from underrepresented populations and understanding/overcoming cultural barriers
Diversity lens in data analysis	Building in-house capability or leveraging external experts to analyse data from sub-populations / use novel data analysis methods to address data missingness and identify subcontinental differences
Diversity in workforce	Having a workforce that understands diversity to the ground, expanding voices and perspectives and having funding continuity to retain trained staff
Diversity in collaborations	Consortiums try to ensure the organisations involved are diverse with different perspectives, and link with groups outside of cultural reach, broadening views and participation

Insights by geographical region

This report includes a breakdown of diversity insights by geographical region, as well as the opportunities for funders by 'genomic maturity archetype'

Due to differences between countries in the same geographical region in terms of their level of maturity in genomic research and diversity efforts or challenges, the initiatives interviewed have been clustered based on three genomic maturity archetypes.



Regional analysis: North America, EU, Oceania

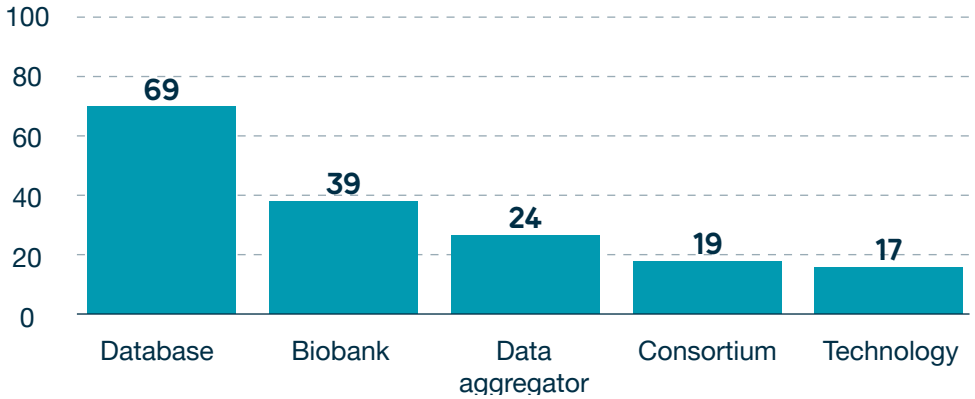
Key metrics and work in genomics: North America, Europe, Oceania

Key metrics:

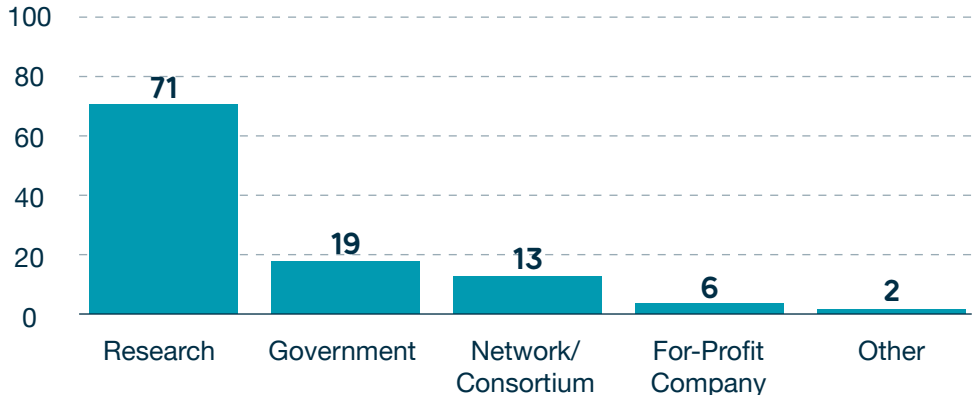
of initiatives

IQVIA database	Surveyed	Interviewed
111	26	9

Types of initiatives and organisations (111)



N=111

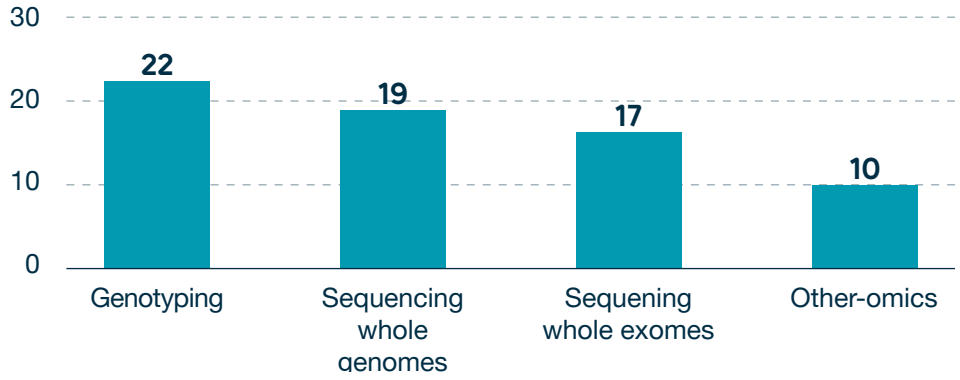


N=111

Key metrics and work in genomics: North America, Europe, Oceania

Key metrics:

Types of genetic data collected (26)



N=26

Work in Genomics

Mission and goals:

Initiatives in North America/Europe/Australia aim to:

- broaden reach to less represented populations
- support etiological studies of chronic diseases (either specific or general disease focus)
- aspiring to enable healthcare systems to provide precision medicine and personalised healthcare

Established capacity and infrastructure:

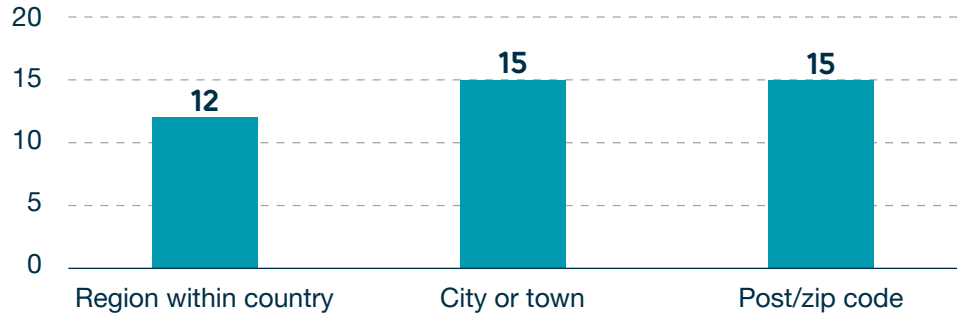
- Databases in US and EU already contain a large amount of data (55,000 in Canada, 200,000 in Estonia, 1,000,000 in US Veterans' database) and there are projects in the horizon to collect data on EU level (e.g., "Genome of Europe" aiming to sequence ~500,000 samples)
- Consortiums in US and EU work towards realising cross-border access to genomic information and increasing the public's trust in this area
- Some databases/biobanks are already working on one data system that contains data from different sources (demographics, genetics, health records, etc), while others are now starting to build this technical cloud-based infrastructure

Note: Oceania (n=1 Australia) is grouped with North American and Europe, as they collaborate on basis of its maturity in the diversity space (advanced infrastructure and acknowledgement from other initiatives), language and the large population with EU ancestry.

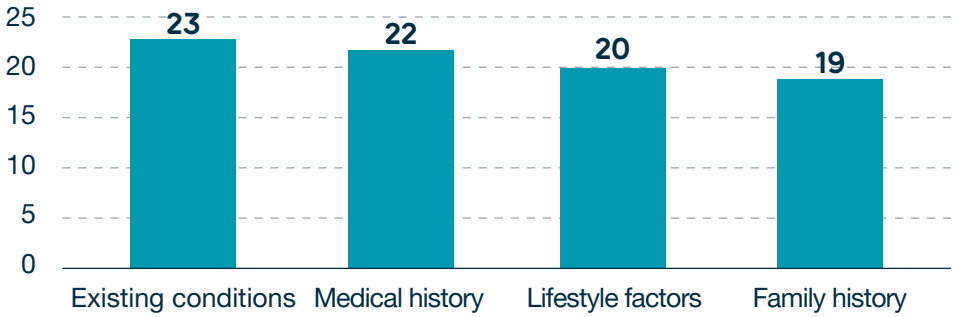
Diversity data collected: North America, Europe, Oceania

Types of data on diversity collected by the 26 survey respondents from these regions.

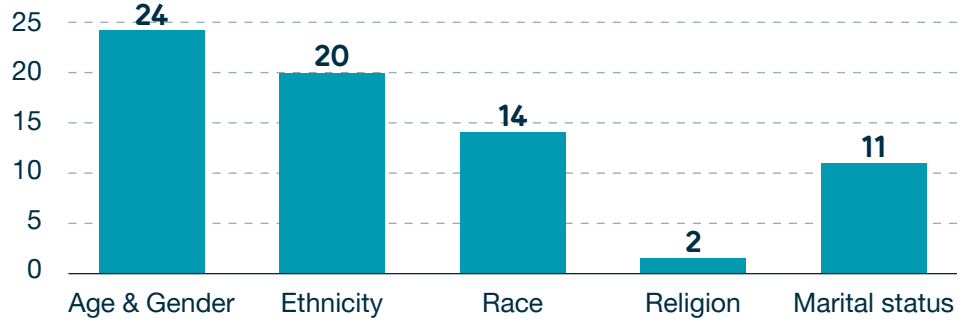
23 collect geographical information



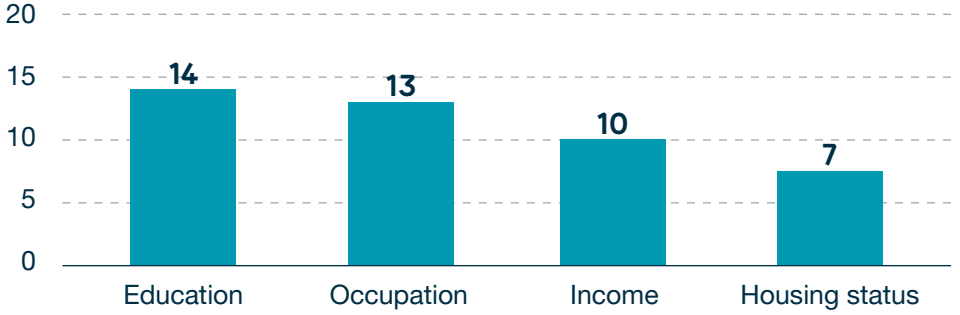
23 collect health information



8 collect demographical information



14 collect socioeconomic information



Almost all initiatives surveyed collect geographical, demographical and health information, while socioeconomic information (education, income etc.) is only collected by ~50% of the initiatives.

Efforts to increase diversity in initiatives: North America, Europe, Oceania

Expanded target population

Shifting focus to add underrepresented target groups missing from original scope, e.g., Native Americans in the United States and Canada, indigenous populations in Australia, or women and Hispanics in US Veterans database.

Alternative recruitment methods

Such as moving vans to reach rural areas, community engagement events, targeted advertisement and communications, strategically chosen sample collection sites, etc.

External Collaborations

Advisory boards are engaged to increase access to diverse populations, such as Indigenous advisory boards (e.g., Australia), Participant advisory boards (e.g., Canada), local communities (e.g., US, EU), or social scientists (Estonia).

Specific focus on diversity

Setting up “initiatives within the initiative” or targeted goals which are focused on addressing specific research questions relating to diversity.

Novel data analysis methods

Developing novel statistical methods to address data missingness and its impact on fairness and equity in genomics.

Example from North America

After having reached their initial target of participants, one North American initiative are now trying to expand their database with more females, Hispanics and native populations. They’re doing this by partnering with veteran groups, joining specific events, making targeted communications and running focus groups.

Example from Australia

One Australian initiative is collaborating with indigenous-led genomic initiatives to leverage an indigenous advisory board to understand and work through disagreements with traditional communities and support their participation in genomic research.

Example from the UK

One UK initiative has launched a diversity-focused project to improve understanding of genomic diversity by reviewing, stimulating, and conducting research into diversity and its impacts on scientific, clinical, and health system outcomes. It aims to facilitate the whole genome sequencing of over 15,000 participants from minority communities.

Challenges towards increasing diversity: North America, Europe, Oceania

Participant recruitment and retainment

Targeting specific groups for recruitment necessitates ongoing community engagement, which can be very time-consuming and expensive.

Data privacy, storage and analysis

Building a secure, online data storage analysis platform needs time and significant organisational planning. Linking genomic data with other health data may render respondents identifiable. In-house analysis expertise on diversity is needed to analyse sub-populations.

Legal challenges

In multi-country initiatives, the different legal environments and beliefs or maturity amongst countries can delay the process of building common infrastructure and systems.

Return on investment for participants

Participants may feel the initiative has not delivered what was promised, such as precision or personalised medicine, which is complex when it is not clear who should implement the scientific findings from the biobank analyses.

Understand indigenous communities

Both for North America and Australia it is important for the initiatives to understand diverse communities and their processes, which can be challenging, since complications arise in navigating cultural differences.

Example from the UK

One initiative is very focused on consent practices, as many population groups tend to withdraw their consent to show they are unsatisfied and expect to be engaged in a conversation around what is happening with their data (lack of trust).

Example from North America

Working and managing diversity in large amounts of genomic data leads to significant issues with data storage and analysis. The creation of a large cloud-based infrastructure solution is needed, which requires significant organisational planning, expertise and money.

Example from North America

Faced challenges in understanding indigenous communities, their cultures, traditions, values and ways of doing things. They propose that indigenous data should be controlled/collected by indigenous people.

Regional analysis: Asia and Middle East

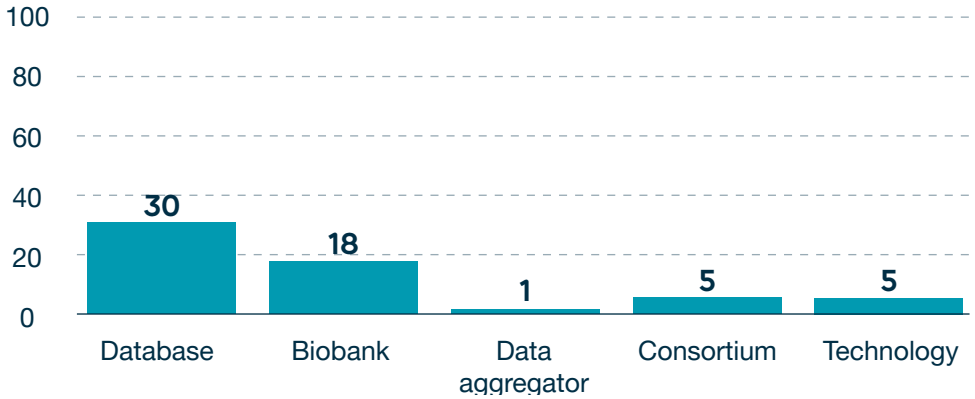
Key metrics and work in genomics: Asia and Middle East

Key metrics:

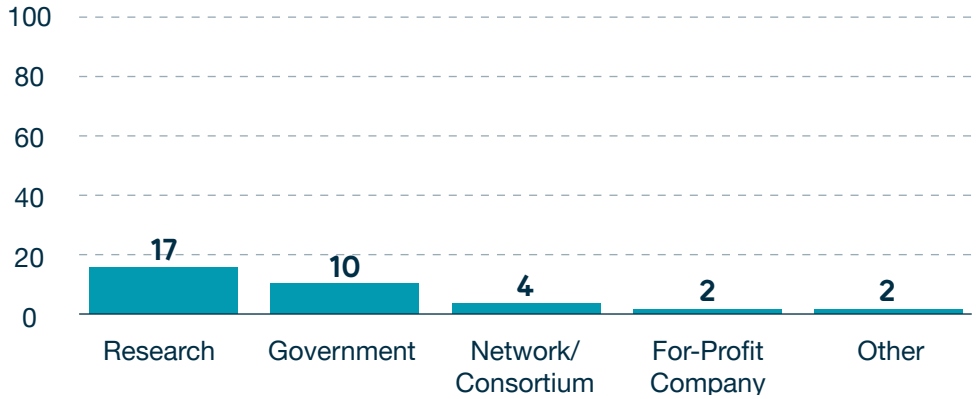
of initiatives

IQVIA database	Surveyed	Interviewed
35	8	6

Types of initiatives and organisations (35)



N=35

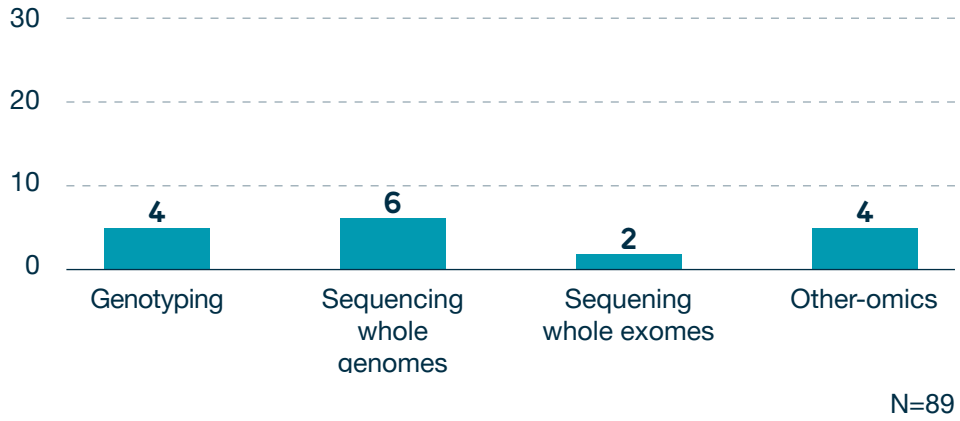


N=35

Key metrics and work in genomics: Asia and Middle East

Key metrics:

Types of genetic data collected (8)



Work in Genomics

Mission and goals:

Initiatives in Asia and Middle East are mainly Databases/Biobanks aiming to:

- enhance understanding of genetic factors involved in chronic disease etiology and progression in each country
- enable healthcare systems to provide better preventative information and health outcomes to the population

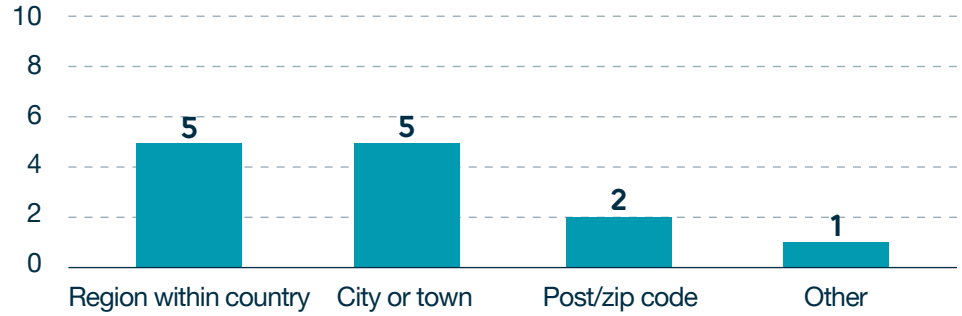
Established capacity and infrastructure:

- There is a higher proportion of government-funded initiatives in this region compared to the other regions, while networks, consortiums and data aggregators are low in number
- Databases and biobanks in Asia and Middle East contain a relatively large amount of data, (e.g., 260,000 for BioBank Japan, 60,000 for Qatar biobank), while there are other, more recent biobanks that are now trying to expand (India's LoCARPoN with 8,584 datapoints, Genomics Thailand with 50,000 samples so far)

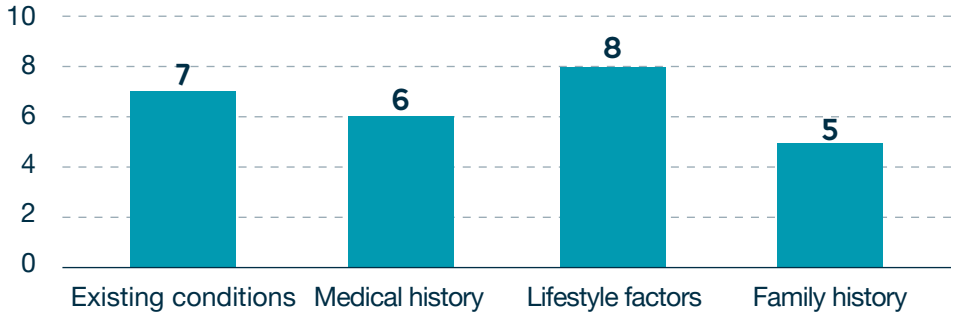
Diversity data collected: Asia and Middle East

Types of data on diversity collected by the eight survey respondents from these regions.

8 collect geographical information

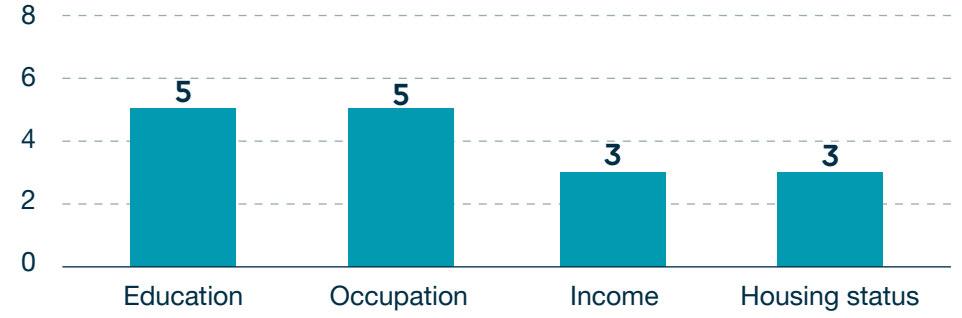


8 collect health information



8 collect demographical information

5 collect socioeconomic information



All initiatives surveyed collect geographical, demographical and health information, while socioeconomic information (education, income etc.) is collected by 5/8 of the initiatives.

Efforts to increase diversity in initiatives: Asia and Middle East

Strategic sample collection sites

Initiatives try to spread the sample collection sites as much as possible across the region to increase catchment area and respondent diversity.

Example from Asia

One Asian Biobank set up 40 sample collection sites across the country to ensure sampling was representative of the population. They encourage citizens to participate by explaining that their contribution supports the health of future generations.

Collaborations

Collaborations with multiple hospitals country-wide has helped some to spread their geographic reach and to grow. Others have collaborated with European initiatives who support with data analysis and sharing of expertise.

Example from Asia

One Asian initiative has placed emphasis on forming collaborations with international organisations to gain advice on best practices for genomic data collection and analysis and gain insights into other genomic projects in the public health sector.

Supporting participation

Some initiatives offer in-home sample and data collection to ease participation. Transportation to and from the study site has also been explored in India.

Example from India

To overcome the fact that Indian people are reluctant to participate in genomic research, the researchers from one initiative visited each home in the community to recruit respondents and collected initial blood samples.

Return of investment for participants

To encourage participation, some initiatives have offered free health check ups to participants as well as medical referral if results are abnormal. Others have encouraged participation by explaining that the outcomes of research will support efforts to improve national health

Example from the Middle East

One Middle Eastern biobank strives to bring value back to participants by offering free health check-ups once a year, and if any results are found to not be within the normal range, they offer a direct medical referral process for participants.

Challenges towards increasing diversity: Asia and Middle East

Limited opportunities for international cooperations

Due to strict data access laws and regulations in parts of Asia and the Middle East.

Limited expertise

Lack of in-house expertise in bioinformatics and genetic counselling and lack of accredited training programs lead to a lack of experienced researchers.

Limited funding

It is expensive to build and maintain a large and diverse data set, particularly considering that materials and be more expensive in these locations. Funding has been a considerable challenge to expansion. The additional expense of reaching remote populations also can hinder efforts for diversity.

Lack of engagement

Difficult to engage with certain population groups due to concerns and scepticism surrounding genetic research. For tribal populations, there are additional requirements for gaining permission and abiding by local laws.

Storage and analytical difficulties

Storage capacity can be a challenge, particularly for biological samples. It can be difficult to analyse hugely diverse data.

Example from Asia

One Asian Biobank, due to the stringent legislation in the country they operate in, cannot share patients' data and health information globally, which largely hinders international collaborations and data sharing opportunities with other consortiums or biobanks.

Example from the Middle East

Permission to access the data in one Middle Eastern biobank can be granted, but the process is lengthy and bureaucratic due to local regulations. The initiative would like their data to be utilised internationally so they now release annual analyses.

Example from Asia

One Asian initiative underlined that a significant challenge is the lack of bioinformaticians, who are crucial for developing software tools and methods for understanding and analysing complex genetic data.

Example from Asia

One Asian biobank proposed that a vital challenge towards increasing diversity in genomic research is that sequencing genomic data locally can cost more than $1,000 per person.

Regional analysis: Latin America and Africa

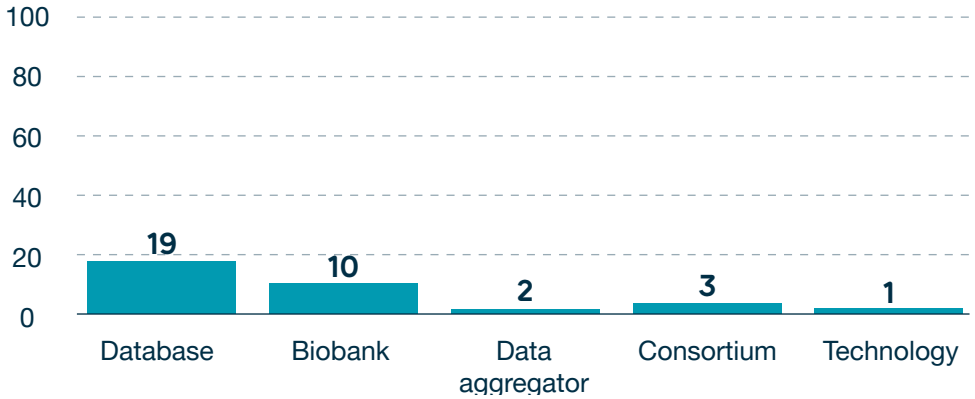
Key metrics and work in genomics: Latin America and Africa

Key metrics:

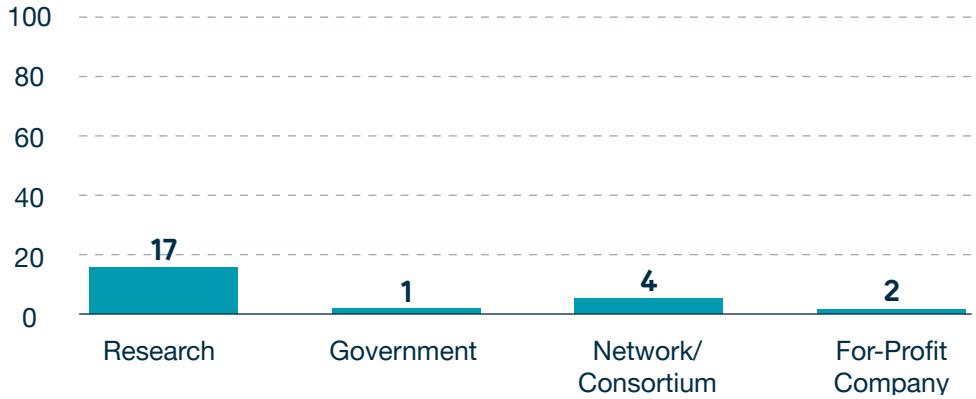
of initiatives

IQVIA database	Surveyed	Interviewed
24	9	7

Types of initiatives and organisations (24)



N=24

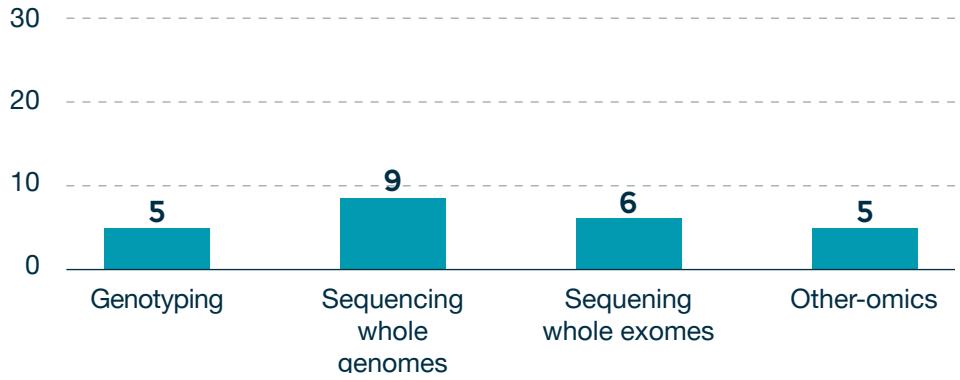


N=24

Key metrics and work in genomics: Latin America and Africa

Key metrics:

Types of genetic data collected (9)



N=9

Work in Genomics

Mission and goals:

Initiatives in Latin America and Africa aim to:

- build a genomic database that represents their native population
- broaden reach to less represented populations with unique genetic make-up
- enhance understanding of how genetic factors influence morbidity and mortality

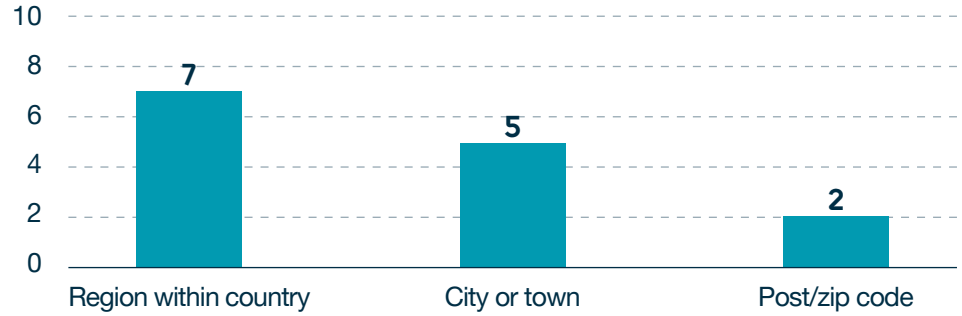
Established capacity and infrastructure:

- Compared to US and Europe, databases in Africa and Latin America generally have smaller amounts of data
- Databases and biobanks have focused on building local capacity, such as labs and data collection sites, for the collection and pre-processing of biological samples
- Many databases and biobanks appear to be outsourcing the genotyping and whole genome sequencing of biological samples, often to the Global North, since they lack the technology or expertise locally
- Many databases and biobanks underline that they hope to upskill local talent and build adequate infrastructures to undertake genomic research in-house

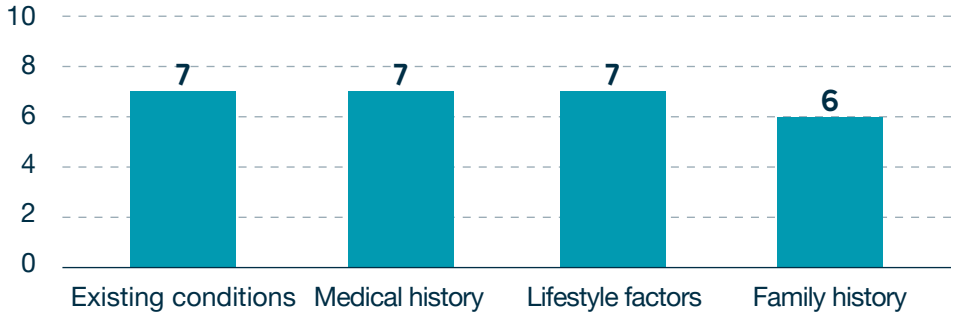
Diversity data collected: Latin America and Africa

Types of data on diversity collected by the nine survey respondents from these regions.

8 collect geographical information

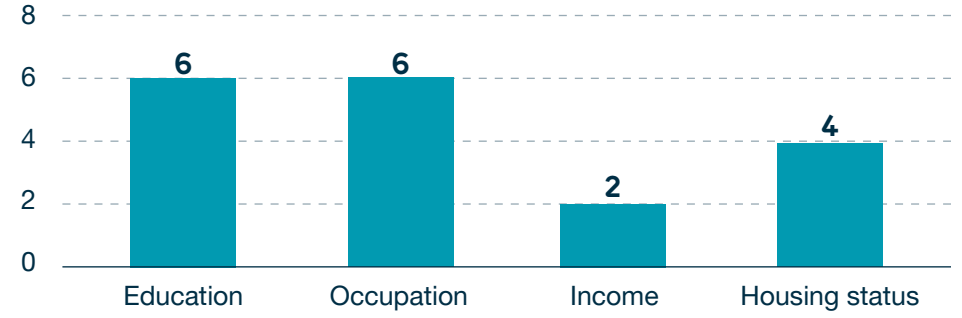


8 collect health information



9 collect demographical information

6 collect socioeconomic information



Almost all initiatives surveyed collect geographical, demographical and health information, while socioeconomic is collected by 6/9 of the initiatives.

Efforts to increase diversity in initiatives: Latin America and Africa

Direct recruitment methods

Tactics such as door-to-door, community engagement workshops, strategically chosen sample collection sites and collaborating with local stakeholders.

Capacity and infrastructure building

To increase diversity both in terms of recruitment and workforce, all initiatives have placed efforts in training local researchers to collect, handle and analyse data, and establishing/collaborating with research centres in various regions for collecting and sequencing biological samples.

Shift to 'for-profit' model

To overcome considerable funding challenges, some companies have developed a business model in which data is collected and access is granted (at a cost) to researchers – this allows for consistent funding to expand their representative dataset.

Provide value to local community

Some initiatives aim to directly relate genomic data to predominant health challenges in their region. Some aim to give genetic test results back to participants (for-profit).

Underlined ethical genomic research practices

To reduce participants' distrust, initiatives focus on ethical genomic research practices by ensuring that documentation is written in simple and local language, and populations comprehend the benefits of genomics for them and future generations.

Example from Africa

One African initiative strengthens local scientific capabilities by training 5 Masters and 2 PhD level researchers in advanced genomic data handling. Additionally, the initiative has established six centres across African countries, led by local experts, to enhance research infrastructure for collecting and processing biological samples.

Example from Africa

One African initiative partners and sells genomic data to pharmaceutical companies that carry out research on next generation cancer drugs that are effective in diverse populations. By becoming a for-profit, they can overcome funding hurdles to build a database that is representative of the African population.

Example from Africa

One initiative emphasises the value of genomic research in Africa for public health and clinical diagnosis. The initiative guarantees that informed consent is written in sensitive language and ensures ethical data handling by informing participants of genomic research implications. They collaborate with Research Ethics Committees to minimize re-identification fears.

Challenges towards increasing diversity: Latin America and Africa

Population reluctancy

Participants' distrust and fear of exploitation poses a challenge for sample collection. Governments' reluctance and sensitivity to discrimination further impedes participant recruitment and retainment.

Example from Africa

Due to the historical backdrop of colonialism, citizens have expressed heightened scepticism and fear that genetic information might be exploited by the global North. Governmental hesitation, driven by concerns of discrimination or exclusion, has contributed to an overall sense of distrust within population.

Lack of local infrastructure

Lack of in-house sequencing infrastructure and long-term storage facilities leads to dependence on international collaborations for processing of genetic samples.

Example from Africa

In Africa, researchers partner with international biobanks for genomic research on African populations due to the lack of local sequencing infrastructure. Outsourcing genomic research poses challenges, including high costs and logistical complexities, hindering the development of local researchers and limiting the capacity for additional analyses.

Lack of local expertise

Shortage of trained and experienced biogenetics/bioinformatics impacts local capacity for data analysis. Trained staff may also leave the country to work elsewhere once funding runs out, leading to local brain drain.

Analytical challenges

Analytical methods developed for European populations pose difficulties for the diverse genetic mix in Latin American and African populations. Population heterogeneity and lack of reference genomes create challenges in genetic associations and data analysis.

Example from Latin America

One study found that analytical methods tailored for European populations, lacking the unique genetic mixture present in Latin American populations, pose statistical challenges. Securing funding for training analysts in novel statistical approaches is deemed essential for handling complex and diverse data effectively.

Challenges towards increasing diversity: Latin America and Africa (continued)

Logistical issues

Shipping samples within Africa and Latin America poses significant challenges due to the cost, time, geographical distance, and inadequate infrastructure for proper sample handling, such as long transportation times, lack of crushed ice. etc.

Example from Africa

One organisation seeks to enlist and gather data from individuals from numerous ethno-linguistic groups in Uganda, yet faces various logistical challenges such as the geographical distance between data collection sites and analysis facilities, and ensuring the timely transportation of samples to their central processing lab.

Data privacy and storage

Guarded data practices impede collaboration and data sharing efforts. Building a secure, online cloud-based research environment needs time and money. Not all researchers in Latin America and Africa have access to high performance computer cluster.

Low national priority

Genomic research is a low priority for some governments who may be sceptical about genomic data, have more pressing health or funding priorities or be reluctant to engage with the international research community. As a result, it can be difficult to get the funding needed for infrastructure building and to get legislation changes needed for conducting genomic research.

Example from Latin America

Restrictions of government funding and legislation made it impossible to build a national biobank within the country this organisation operates in. After years of frustrations, the academic founders of this organisation decided to switch to a “for-profit” business model so they could fund ongoing research. Private sector clients (mostly pharma) will purchase exclusive access to the data and fund data collection. After a set period, this data base will be released for academic use.

Detailed insight: global collaborations

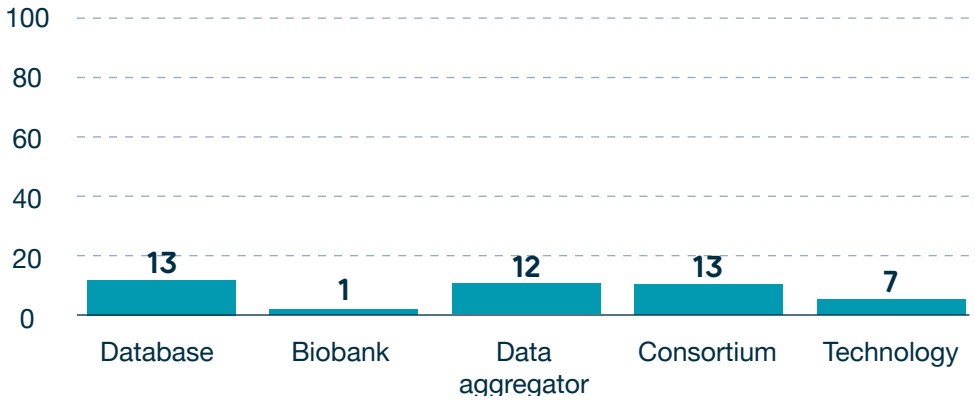
Key metrics and work in genomics: global collaborations

Key metrics:

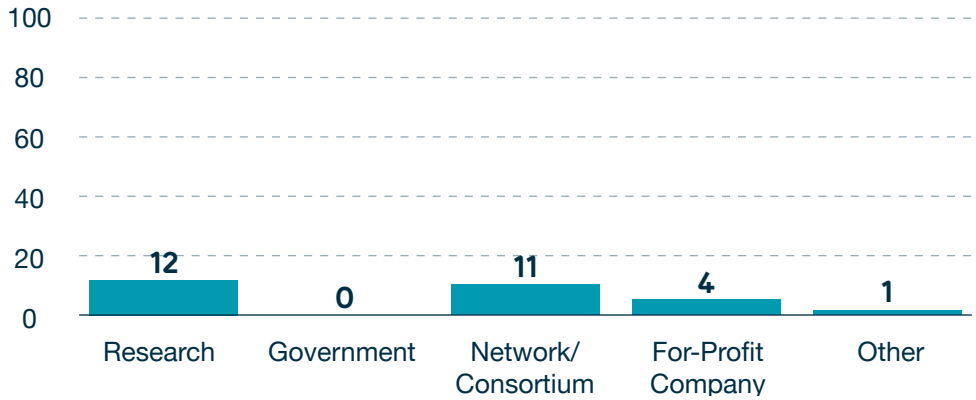
of initiatives

IQVIA database	Surveyed	Interviewed
28	9	5

Types of initiatives and organisations (28)



N=28

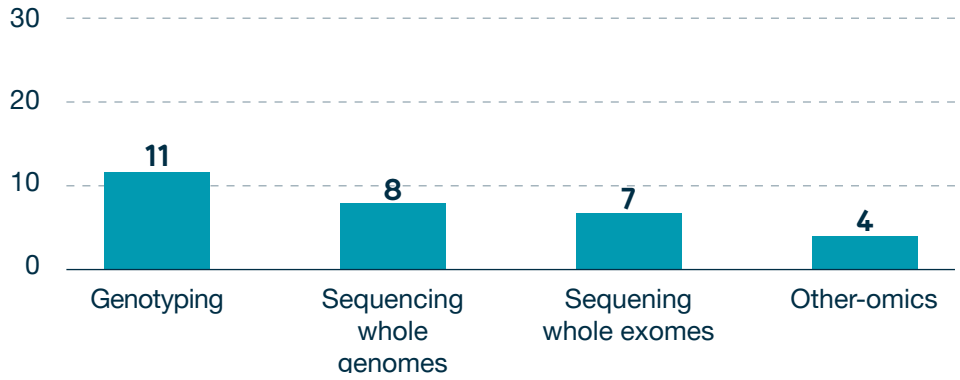


N=28

Key metrics and work in genomics: global collaborations

Key metrics:

Types of genetic data collected (12)



N=12

Work in Genomics

Mission and goals:

The global initiatives interviewed aim to:

- help understand the genetic mechanisms underlying complex disease and work with pharmaceutical companies to improve drug discovery, via science and technology
- create standards and policies for other organisations to follow and make genomic data more consistent and usable across the globe
- provide a personal discovery tool for individuals to gain advance knowledge of their DNA identity

Established capacity and infrastructure:

- Global consortiums are collecting data from primary studies across the worlds to uncover how common genetic variants contribute to diseases across the clinical spectrum
- They work towards a collaborative research environment by educating organisations, initiatives and researchers, encouraging cohorts to interact with each other and providing early career training
- Policies and best practices have been developed and are being followed by several regional initiatives
- Many databases and biobanks underline that they hope to upskill local talent and build adequate infrastructures to undertake genomic research in-house

Efforts to increase diversity in initiatives: global collaborations

Size and geographies of cohorts

Global databases prioritise large cohorts and try to ensure cohorts held in low- and middle-income countries are included in the system, aiming to enhance diversity.

Fine-tuning technologies based on localities

Initiatives developing risk profiling technologies work to understand target populations and convince medical teams that their technologies are appropriate for their population, ensuring sufficient validation has been carried out.

Global events and upskilling

Global consortiums hold webinars and conferences to engage new with diverse cohorts and allow initiatives to interact with each other. They also provide training to upskill and mentor researchers, as they consider them the future of diversity.

Funding for sustainability of efforts

Global consortiums provide financial resources for research, infrastructure development and workforce for project sustainability, including funding researchers from low- and middle-income countries to attend global conferences.

Reestablishing trust

Global consortiums focus on reestablishing trust that is lacking due to historical disconnects among regions by underling the value of diversity for researchers, contributors and the global community and connecting local communities in low- and middle-income countries with wider initiatives.

Example

One organisation holds bi-annual conferences to reach broader scientific community in the realm of genomic research and scheduling monthly webinars in two different time zones, allowing for real-time interaction of cohorts and promoting the consortium's function. Also supporting early career investigators, providing mentorship and access to esteemed researchers globally.

Example

One organisation funds efforts for local genomic data collection and storage, by building infrastructure and capacity in low- and middle-income countries. Moreover, it underlines the need for funding continuity at the local level to prevent trained researchers from moving away at the end of the project.

Example

One consortium puts significant effort into forming relationships and trust with researchers and institutes in middle- and low-income countries by underlining the value of cross-country collaborations in genomic research and encouraging them to share their genomic data with the consortium.

Challenges towards increasing diversity: global collaborations

Lack of trust

Trust is a critical factors in global collaborations, especially around data colonialism. Building strong relationships with researchers in low- and middle-income countries is challenging.

Research biases and limited data on ancestry

There are relatively few cohorts with sufficient genetic data and sufficiently high-quality medical data. African ancestry data is lacking, and strong socioeconomic biases exist in many cohorts.

Logistical challenges

The different time zones, religious holidays and other factors pose challenges in collaborating and holding live events across different regions.

Limited funding

Genomic research in regions with constrained infrastructure incurs substantial expenses, requiring pivotal funding for community access and benefit. Additionally, the high-cost of genome sequencing hinders broad benefits of genomic data in research and business.

Data access and sharing

Data sharing structures from other researchers are often lacking and access agreements are often not signed off. Access agreements being so individualised between countries makes it often easier to work with European or North American cohorts.

Example

While one international organisation maintains effective interactions in North America, Europe, Africa, and Latin America, it encounters logistical challenges in coordinating with Southeast Asia and other regions. Challenges stem from varying time zones, diverse religious holidays, and hemispheric differences, impacting seamless coordination and communication.

Example

One organisation underlined lack of funding as a substantial challenge in increasing diversity in genomic research, as carrying out whole genome sequencing is very expensive especially in developing countries that often need to transfer data for analysis and storage.

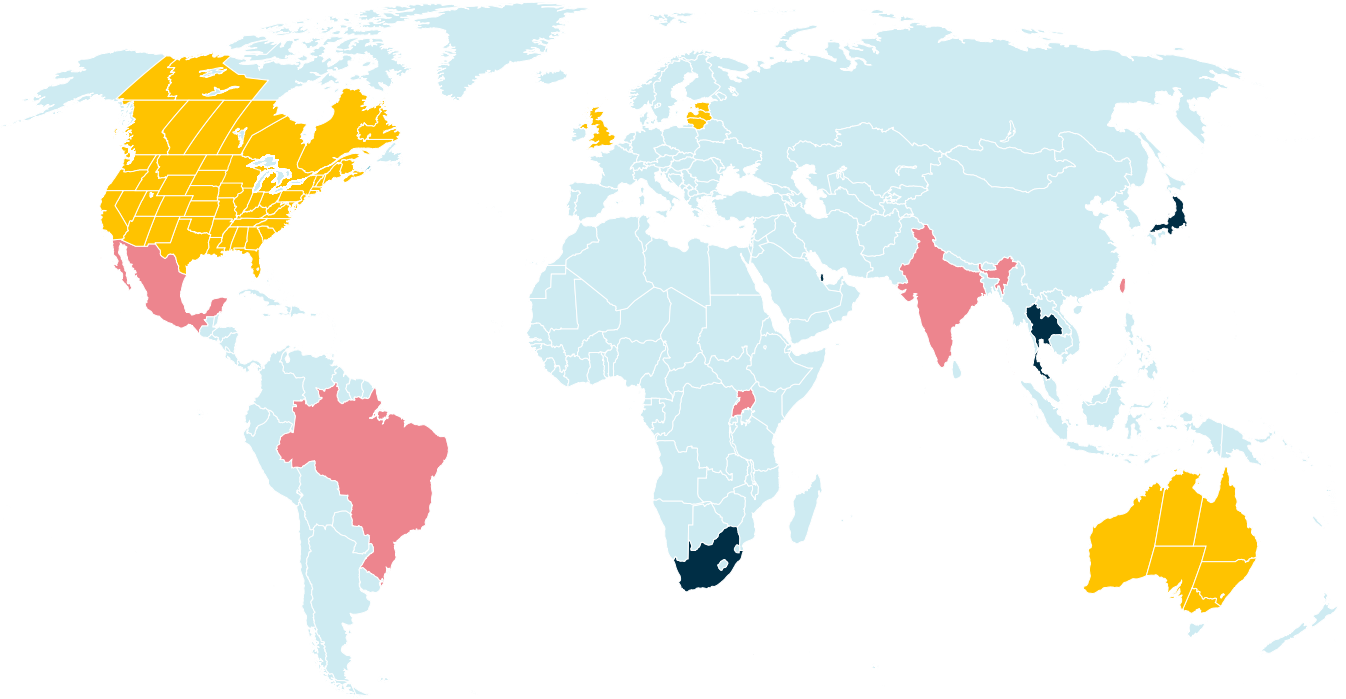
Example

One organisation highlighted that inconsistent data sharing structures among researchers and unapproved access agreements are common challenges towards increasing diversity. Additionally, financial constraints, such as substantial fees for data access, further complicate the process, making collaboration with European or North American cohorts more accessible due to standardised access arrangements.

Opportunity per genomic maturity archetype

There is high regional variation in terms of maturity of genomics research, as well as different approaches and challenges for diversity per region

Maturity in genomics research



Low Medium High

Key takeaways

- High maturity regions are focusing on expanding datasets into minority populations
- Medium and low maturity regions are working on building representative databases

Research caveat: above representation is based on and limited to the initiatives included in this research and does not cover the full global spectrum of initiatives

There is high regional variation in terms of maturity of genomics research, as well as different approaches and challenges for diversity per region (continued)

High maturity regions (USA, Canada, UK, Estonia, Australia)

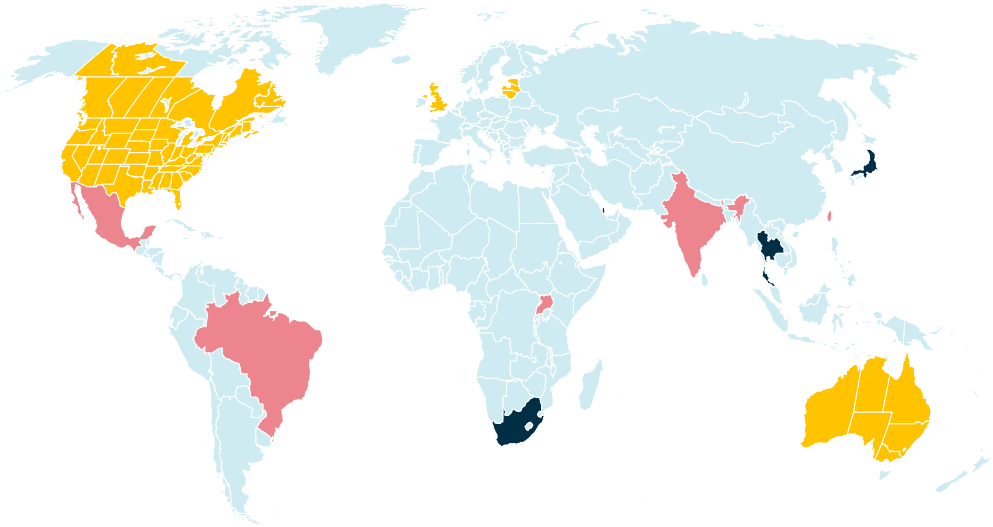
- Types of initiatives: databases, biobanks, data aggregators, consortiums, technology
- Average cohort size: large (~579,000)
- Objectives: support etiological studies of chronic diseases / ultimately contribute to precision medicine and personalized healthcare
- Diversity efforts: expand into minority populations, make external collaborations, work on novel data analysis methods
- Top challenges for diversity: Engaging with minority populations, IT secure infrastructure, cross-regional collaborations

Medium maturity regions (Japan, Taiwan, Thailand, Qatar, Hong Kong, South Africa)

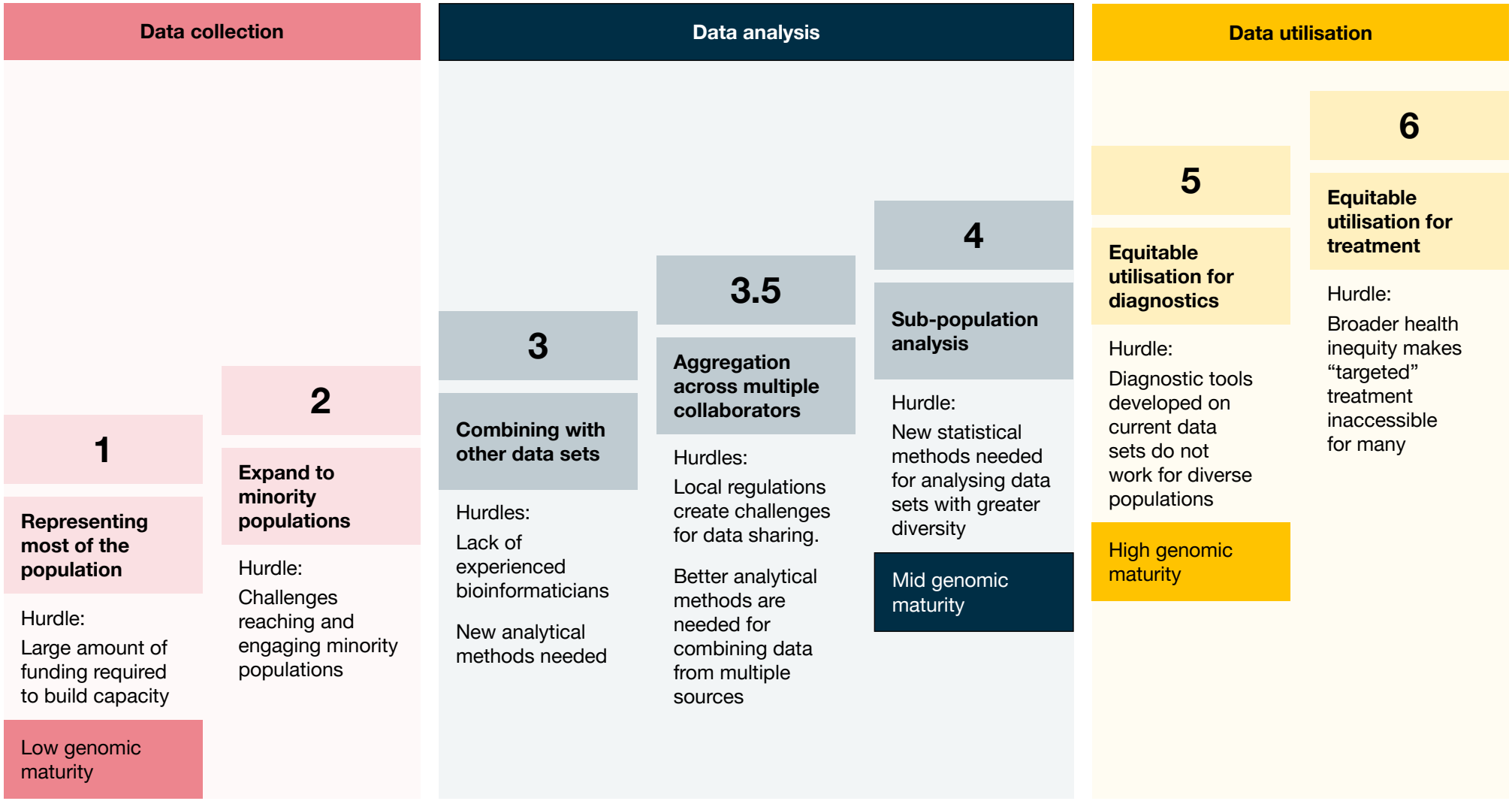
- Types of initiatives: mainly databases, biobanks
- Average cohort size: medium (~70,000)
- Objectives: regional understanding of genetic factors associated with chronic disease etiology
- Diversity efforts: expand databases, overcome logistical barriers, improve data analysis, overcome legal barriers to international collaborations.
- Top challenges for diversity: Legal / regulatory restrictions to share data internationally, cost of genetic sequencing.
- Top challenges for diversity: Engaging with minority populations, IT secure infrastructure, cross-regional collaborations

Low maturity regions (Brazil, Mexico, Uganda, India)

- Types of initiatives: mainly databases, biobanks
- Average cohort size: low (~13,000)
- Objectives: include less represented populations, enhance regional understanding of genetic factors associated with chronic disease etiology
- Diversity efforts: expand databases, use alternative recruitment methods, build capacity and infrastructure.
- Top challenges for diversity: Logistics, respondent / government trust, project funding, brain drain, analytical challenges



The path to full and equitable utilisation of genomics data has several key steps and many hurdles to overcome



A PESTLE Analysis provides an overview of the challenges in genomic diversity in low-and high maturity regions

	P	E	S	T	L	E
	Political	**Economic**	**Social**	**Technological**	**Legal**	**Environmental**
High maturity regions	Lack of trust from low/middle income country organisations, due to data colonialism	Targeted recruitment can be very costly and reward does not always match investment	Cultural, language and funding challenges to engage, recruit and analyse minority populations	• Building a secure, online space for large amount of data to be stored and processed is difficult • Expertise is needed to analyse minority populations' genomic data	Cross-regional collaborations can be restricted by the different laws and legislations around data privacy	The slow healthcare progress towards precision medicine development can make participants feel they did not gain what was promised
Low maturity regions	Governments' reluctance to fund/ support genomic research due to amplified sensitivity to discrimination and low prioritisation of genomics	High funding needs to overcome logistical challenges (e.g., research sites, long-term sample storage facilities, sample transport)	Due to colonialism, people can be reluctant to participate, in fear of data exploitation by the Global North	• Analysis methods on Caucasian datasets not applicable to all sub-populations • Lack of in-house gene sequencing infrastructure • Lack of local expertise	• In African regions, different data privacy and sharing policy standards exist in each country • Lack of policies that protect the privacy of genetic information	• Researchers have become guarded with data, thus impeding data sharing efforts • Challenging to give back to communities, e.g., unable to offer high-cost drugs

Opportunities for funders to positively impact data and diversity in human genomics

After challenges and opportunities were identified, the Scientific Advisory Board of external experts was engaged in order to categorise opportunities for funders and other influential organisations by priority level.

For high maturity regions, efforts should focus on minority and community groups engagement, alternative recruitment methods and respondent return on investment

High Maturity Archetype

Higher ranked opportunities	Solutions/recommendations
Community efforts	
Support efforts to collaborate with minority populations (indigenous advisory boards, local community outreach plans)	• Consider thoughtful public/participant engagement plans • Engage with Patient Advisory Groups so they can see value of this type of research. Fund conference attendance and specific research that drive community engagement • Support research that connects with local leaders or include within winning criteria
Support community engagement to develop perspectives on use and storage of genetic data	Create funding programmes that seek to assess community views on consent to data usage (to take pressure off individuals) and support community empowerment – ideally led by Principal Investigators from these communities
Recruitment	
Support with alternative recruitment methods to reach diverse populations (e.g., mobile blood collection vehicles to reach distant, rural areas)	• Funders could request Principal Investigators to identify country-specific challenges and support research groups who understand and target these diverse groups • Fund a network or work with existing network to encourage bulk blood collection
Community efforts	
Support in the provision of individual outputs to provide patients with a “return of investment” – responsibility for sharing genomic insight back with original country plus health data	• Funding social programs of work for academics to develop individual understanding on research purposes • Behavioural change initiatives addressed to doctors and pharmacists to understand their existing challenges and motivators to change engagement with diverse groups

For high maturity regions, data storage and analysis are the second priority

High Maturity Archetype

Lower ranked opportunities

Solutions/recommendations

IT infrastructure

Support in the development of IT infrastructure for secure, online platforms to store and analyse data which will support data sharing

- Increase investment in key platforms and cloud-based databases
- Develop programmes of work that consider existing infrastructure and drive innovative approaches to store and analyse the data, and allow information to “follow” the individual patient
- Ensure consistency/harmonization across tools and tech to allow analysis and future use

Data analysis

Support with advancements in compiling and analysing data from multiple sources (e.g. linking genomic data to health records)

- Fund development of open-source software, not tied to sequencing instrument
- Leverage artificial intelligence and machine learning for researchers to provide rapid feedback on tools, e.g., polygenic risk scores not working for a specific population

Opportunity

Support and funding for individual researchers who are at the forefront of developing technical advancements in the genomics area.

Legal issues are also important, but outside of the remit of most funding organisations

High Maturity Archetype

Lowest ranked opportunities

Data sharing

Support in addressing legal challenges for data sharing

Support in translation efforts to enable use of genomic data to support diagnosis and care

For mid maturity regions, the most important opportunity is to support regional teams with education and training

Mid Maturity Archetype

Higher ranked opportunities

Education and training

Support regional teams with educating and training researchers – bioinformatics, database development and data interpretation to drive use of data.

Solutions/recommendations

- Capacity building for human genomics research for both wet-lab and data analysis; grant calls for locally-led genomic data generation/analyses, pairing them with expertise on databases and interpretation, with train-the-trainer models to build regional expertise for further training and scale
- Offer funding for “low-hanging fruit” initiatives, e.g., country-led initiatives to build local genomic data repositories, well-characterised biobanks
- Acknowledgement and credit to data contributors and data owners to be built into such collaborative projects, potentially explore benefit-sharing mechanisms

For mid maturity regions, increasing participant engagement and addressing recruitment/sample logistical challenges is also important

Mid Maturity Archetype

Lower ranked opportunities

Solutions/recommendations

Participant engagement

Support citizen engagement programmes to educate participants on the value of genomic research and understand their challenges to participating

- Survey to understand participant awareness and challenges in being part of genomic research
- Clear communications and advocacy on the importance of increasing local data generation to ultimately address local needs, equitable and cost-effective access to fit-for-purpose diagnostics, therapeutics and vaccines that serve the local populations

Recruitment

Support with alternative recruitment methods to reach diverse populations (e.g., mobile blood collection vehicles to reach distant, rural areas)

Address sample collection, transportation and logistical challenges across rural/urban communities and geographical spread – support projects on mapping and optimising population and diagnostic / sequencing facility networks

For mid maturity regions addressing IT and legal challenges could foster diversity

Mid Maturity Archetype

Lowest ranked opportunities

IT infrastructure

Support in the development of IT infrastructure for secure, online platforms to store and analyse data which will support data sharing

Support in development of IT infrastructure to allow linkage of electronic medical records or paper-based records to include genetic information

Legal challenges

Support in addressing legal challenges for data sharing and allow international collaboration

For low maturity regions, efforts should focus on growth and retention of local talent and local infrastructure development

Low Maturity Archetype

Higher ranked opportunities	Solutions/recommendations
Local infrastructure	
Support in infrastructure building (sample storage, sequencing capabilities, required materials and technologies)	• Support infrastructure to undertake genetic research locally • Support biobanking and sample storage (long-term funding) • Assist in harmonising data dictionaries to certify consistency in genomic research practices
Support with growth and development of local genomic research teams to promote long-term research capacity	Promote collaborations across different initiatives e.g., exchange fellowships
Talent growth and retainment	
Support with the retention of talent and expertise to prevent local “brain drain”	• Provide fellowships for people to return to their native countries and contribute to local research • Provide grants to inspire feeling of reassurance to local researchers
Specific support for educating and training in bioinformatics, database development and data interpretation to drive local use of data	• Fund fellowships for researchers to train local workforce on cutting-edge skills in bioinformatics, artificial intelligence and machine learning, large language models, etc • Organise courses, offer graduate studies, and enable mobility for up-skilling epidemiologists and genetic counselors

For low maturity regions, IT infrastructure and local engagement come as second priority

Low Maturity Archetype

Lower ranked opportunities	Solutions/recommendations
IT infrastructure	
Support in the development of IT infrastructure for secure, online platforms to store and analyse data which will support data sharing	Support individuals working on developing these platforms
Engagement of local community	
Support efforts to engage with local communities.	• Help in improving genetic literacy by establishing short courses • Employ professionals (e.g., behavioral scientists) to optimally engage the public
Support efforts to “give back” to the participants/community to support in strengthening of local health systems	• Aid in building genetic counseling program in West Africa, to advise community on peculiar genetic variant mutations • Provide fellowships to local community to build capacity in anthropology, medical sociology and genetic counseling disciplines

For low maturity regions, support on novel analytical approaches and alternative recruitment methods important, but less of a priority

Low Maturity Archetype

Lowest ranked opportunities

Novel analyses

Support advancements in analytical capabilities to develop novel approaches to handling data sets with higher diversity

Recruitment

Support with alternative recruitment methods to reach diverse populations (e.g., mobile blood collection vehicles to reach distant, rural areas)

Opportunity

- Novel data analysis methods are important for this maturity archetype, however it is considered as a “second step” that can follow once the most critical areas, like developing local infrastructure and growing/retaining talent, have been achieved.
- Alternative recruitment methods to reach diverse populations are less important for this archetype, as diversity is already a core component of its population mixture.

For global collaborations, efforts should focus on data analysis, low- and middle-income country collaborations, and the set-up of regional centres

Higher ranked opportunities	Solutions/recommendations
Data analysis	
Support with the development and sharing of novel techniques for the analysis and integration of large and diverse data sets	• Fund a goal-driven program to develop local and regional solutions that are cost effective and capable of scaling up. For example, targeted sequencing to develop custom-genotype arrays relevant to future research and clinical care • Offer fellowships and exchange visits of local talents from low- and middle-income countries to learn data analysis in advanced environments/labs • Fund exchange visits of experts from leading labs to go to from low- and middle-income countries and give lectures, offer advice, stimulate local students
Low- and middle-income country collaborations	
Encourage collaborations with global initiatives and local research teams in low- and middle-income countries	• Offer fellowships and exchange visits from low- and middle-income countries scientists to leading labs/groups of global initiatives • Collaborate with major philanthropic and industry funders (and governments) to fund local capacity building in low- and middle-income countries
Empower initiatives from low- and middle-income countries within global initiatives to overcome history of data colonialism and to shift existing power imbalance	• Fund the biobanking/genotyping and sequencing of samples, with special attention to minorities and enslaved populations to learn more about their roots, ultimately enlarging the number of samples from them • Return insights from genomic sequencing to minority groups and low- and middle-income countries
Regional centers	
Support global initiatives with the set up of regional centres which help to foster more convenient collaborations	• Provide training and technical support (including computation and storage) to support local/national resources in the wider region • Fund genotyping, sequencing and data analysis efforts for data from minorities/ low- and middle-income countries to be included in global efforts • Consider forming consortiums with private / public partnerships with local ownership

For global collaborations, knowledge and skill sharing, and data harmonisation come as second priority

Lower ranked opportunities

Knowledge sharing

Support efforts for global knowledge and skills sharing (consortiums, webinars), with a particular focus on engaging low- and middle-income countries

Solutions/recommendations

- Organise sponsored meetings/workshops/webinars with special focus on low-and middle-income countries and minoritised groups
- Offer fellowships and training networks programs to support training of scientists in low-and middle-income countries across a wide range of topics about genomics and genome analysis. Training might offer two tracks: a technological track (data production/analysis) and an implementation track (data usage in local clinics and society)

Data harmonisation

Support discussion and processes to enable harmonisation and convergence on data analysis and storage and other areas to drive international collaboration

- Offer fellowships and exchange of local talents from low-and middle-income countries and minoritised groups to learn data analysis in advanced environments/top labs
- Offer funding for exchange visits of experts from the leading labs to go to low-and middle-income countries and give lectures, offer advice, stimulate local students
- Fund the biobanking/genotyping/sequencing of samples in low-and middle-income countries

For global collaborations legal challenges and cloud-based solutions are deemed important

Lowest ranked opportunities

Data access policies

Support with overcoming data access policies so that lower maturity regions are able to collaborate with global initiatives

Legal challenges

Support with the development of a cloud-based technological infrastructure for large scale data storage and analysis

Opportunity

- Addressing legal challenges in data access and developing cloud-based technological infrastructure are pivotal considerations in addressing the multifaceted matter of diversity in genomic studies. The use and open access to digital sequence information is under consideration in the UN Convention on Biological Diversity.
- Harmonisation would be critical for data analysis and downstream use of genomic results in healthcare: consider supporting the creation of a charter that outlines privileges and responsibilities of signatories.

Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, infectious disease, and climate and health.

Wellcome Trust, 215 Euston Road, London NW1 2BE, United Kingdom
T +44 (0)20 7611 8888, E contact@wellcome.org, wellcome.org