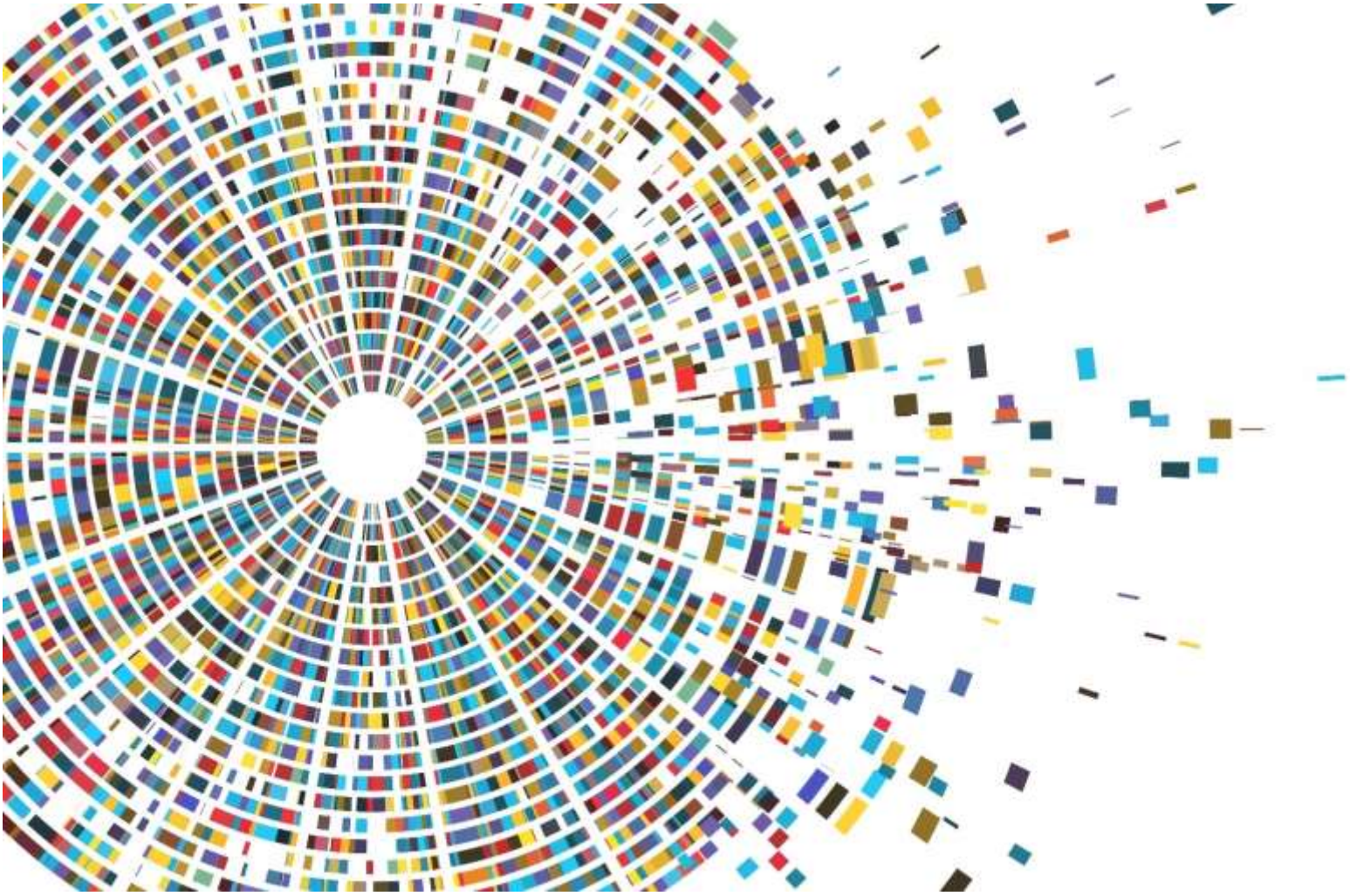
Public Policy Projects
Public Policy Institute

State of the Globe: Bringing the benefits of genome sequencing to the world

Roundtable Five

Diversity in Genomic Data



This report collated the thoughts and actions addressed during the fifth roundtable of this series. The conducted discussion operated under the Chatham House rule.

The genus of this Programme of work was to explore how to bring the benefits of genome sequencing to the world and an important factor in this was to seek to explore how the lack of diversity in the dataset has the potential to exacerbate the health inequalities that we know exist and that the pandemic has so clearly highlighted. If genomics can provide significant benefits but the dataset is not fully representative of our global population and yet findings are inherently focused on the populations that are most represented, how can we break this vicious cycle?

Put another way, if genomics and genome sequencing is the best opportunity to appropriately prepare for the next pandemic (and indeed other transmissible diseases) as well as optimise the protection of population and individual health, if genomic data continues to fail in reflecting the diversity in the global population, will it fail in bringing those benefits equally to everyone? At the heart of this programme of work is a point which comes out time and time again – equity, both in terms of access and return.

There were several interrelated issues that the fifth roundtable in this series of work sought to explore:

- What are the inherent barriers that continue to prohibit greater diversity in the data set and how can we overcome them?
- How is greater diversity of participation encouraged, not just from those contributing their data, but also to make sure there is greater diversity in the scientific and research community? How can the tools to enable equal access to data regardless of where in the world research is based be provided?
- How do we ensure that the benefits of a more diverse data set actually benefit those who are participating in them? It can surely no longer be acceptable that benefits are derived from one data source and yet those participants never benefit directly. How can this be overcome?

What are the inherent barriers that continue to prohibit greater diversity in the data set and how can we overcome them?

The need to have representation across genomic data sets is widely recognised. But the real question is, how is this achieved? Many of the groups that are underrepresented are groups that have historically been wronged; where their data has been misused or used unethically meaning that their data has been taken and used and yet such communities have never reaped the benefits or rewards of their contributions. The levels of engagement amongst those communities are therefore incredibly low. The real question, therefore, is: how do you engage those communities and bring those communities in so that they will participate, that they want to participate, and they see the value of participating?

The issue of cost was also identified as a barrier to greater diversity of data, both in terms of how much it costs to generate the data but also to analyse it. There is limited infrastructure and training across the continent of Africa, for example, and whilst that has changed enormously over the H3 Africa period it is still not up to scratch with global big data communities. This is picked up in more detail later in the report. Genomics Thailand also identified that there were cost disadvantages of doing

genome sequencing locally as the cost of reagents and instruments, including tax, are about 2 or 3 times higher in countries like Thailand as opposed to western countries. As they were collecting less samples at the beginning there was a much higher cost per sequencing which makes it more difficult to scale up. The roundtable considered a few examples of specific initiatives happening now that were seeking to address these barriers.

In Canada the 'Silent Genome Project' has been ongoing for several years. This has the goal of building a variant database for genetic diseases in children. However, the bulk of the project is focused on developing a framework for engagement within Canada's indigenous communities: engaging those who are disengaged, ensuring that there is appropriate governance, ownership and use of the acquired data. The other focal point has been to engage and involve indigenous researchers as part of the research teams. This enables the researchers to act as part of the engagement with the communities. They are bringing in training and really ensuring that it's not just a dominant Western, or colonial, approach to research; it really is a true engagement. There are challenges though – it has taken 2 to 3 years for the Silent Genome Project to build the framework and governance structure before taking the first sample.

Australia's National Centre for Indigenous Genomics has a specific focus on First Nations people and in addition, Australia is setting up the Centre for Population Genomics which will be focusing on gathering a truly diverse and representative sample of the population in Australia. Some of the ethical and technical challenges around how to approach this are explored in more detail later in this report.

In the United Kingdom (UK), Genomics England has just begun its Data Diversity Initiative. This is looking at Genomics England's datasets to identify the gaps in ancestry and ethnic group origins that need to be filled to improve diagnostic reporting. Genomics England's dataset uses the NHS UK Census Codes to assess diversity of the dataset. The initial analysis demonstrates that only about 10 per cent of the individuals in the Genomics England dataset are of non-white background. That is a very small proportion of the dataset and obviously is not representative of the makeup of the UK population. The Data Diversity Initiative is therefore looking to plug these gaps through partnering with communities and programmes as a way of trying to reach out and bring this diversity in the genetic data so that is it consistent with the UK population to improve diagnostics. There are challenges about why communities are reticent to participate.

Genomics England have some excellent examples of community engagement which are relevant to these discussions and agreed that it would be terrific to get community engagement in the governance absolutely from the beginning, including in what they want to research. For example, it is established that within the Caribbean community in the UK myeloma is a particular concern because it happens at a higher rate but there is not a good enough reference genome for people of Caribbean origin to do the work to enable this to be explored in more detail so there is a bit of a "chicken and egg" problem. It is their aim to propose greater collaboration around the extraordinary diasporas of African populations in the UK, for example the very large populations of Nigerians, Somalis and Kenyans who live in the UK and have access to their life-long NHS records. These are communities who are reluctant to engage – they have fears and concerns with respect to genomics but the one thing that has been established is that they all want to benefit their communities back in their country of heritage. This could be important as we would then have a very large population willing to contribute their genomes

and the associated health data that goes along with it that is so crucial to interpretation and they want that information to be used in Africa.

If such an approach could be implemented this could have the additional benefit of creating specific point of care diagnostics or pharmacogenomic testing within individual countries in Africa which are able to use this data to directly benefit their own populations. This way the data would be available to African scientists to use from the communities in the UK who want to make that contribution. This is an interesting concept that requires further consideration.

There are other organisations, such as the Global Genomic Medicine Coalition (G2MC) and the work of H3 Africa, which are working with emerging economies and regions that have limited researchers to build genomic capacity. Building this capacity and training of scientists and experts in these regions is critical. "Helicopter Research" (which has been going on for many years) which merely seeks to go into a region and take samples to increase the numbers in another jurisdiction's database is the wrong approach. Countries and communities need to be empowered to find their own solutions and use their data for their own public health and policy decisions.

How is greater diversity of participation encouraged, not just from those contributing their data, but also to make sure there is greater diversity in the scientific and research community? How can the tools to enable equal access to data regardless of where in the world research is based be provided?

It was articulated a number of times that we need an ecosystem change if we are to truly drive greater diversity in genomics. Change will not simply be achieved by creating a more diverse database; there is a global need to change both the collective mindset regarding diversity and the ecosystem within which research is conducted. This needs to encompass a range of areas including groups such as these and their dialogues, peer reviews, review of funding grants, selection of cohorts and editorial boards as well as building capacity locally.

It was also suggested that there was a need, globally, to define exactly what we mean by "equity". If that definition could appreciate that equity means a benefit for **all** not just for **some** that would enable a new dialogue to begin. The question was posed as to whether there should be a global framework to come up with policies in this regard? And importantly how this would be enforced.

It was identified that Africa is more than ready to contribute and augment the number of African genomes and bring about African diversity to the database that the global community can benefit from. However, resources need to be invested in Africa, in that process empowering Africans to lead not just the generation of the data but also in terms of making the science and the data available from an African perspective and in the context of Africa. It is important to empower those with that context and that historical background so that the data can be translated in a way from which the whole global community can benefit. Africa now has the infrastructure – there are laboratories in Africa that have the latest technology in terms of Next Generation Sequencing platforms. That has been witnessed through COVID-19 and there is no reason why African countries should not be given the opportunity to contribute and contribute significantly.

In terms of creating greater diversity within the scientific and research community it was identified that African scientists will, notwithstanding the significant progress that has been made over the H3Africa period, enter into any collaboration on the backfoot. This has been evidenced when H3 Africa has tried to get African scientists involved in IHGC or G2MC – there is hesitation from them feeling they will not enter on an equitable footing. These collaborations (and others) need to recognize the importance of the African scientists who bring the context and their work and shouldn't be middle authors or come in on a different footing to others just because they don't have the compute resource or haven't learned specific skills. Collaborations having a capacity development component to their collaboration would certainly help.

In encouraging diversity of participation, particularly with respect to the research community, the BRAIN Initiative, part of the National Institutes of Health (a funder of research programmes) have recently included a Plan for Enhancing Diverse Perspectives (PEDP) as part of their application process for funding. This means that not only will there be an outline of the science and the science design in a research application but also that the applicants will have to demonstrate how their programme will enhance diverse perspectives – will there be training? Will there be engagement with different types of communities or organisations around the world? How is the project infrastructure present to ensure there are career enhancing opportunities, particularly for investigators from diverse backgrounds? The BRAIN Initiative does not dictate how the PEDP plans will be developed but this will be part of any application and part of the scored review criteria. This is very new for NIH but it will be part of their funding decisions in an attempt to drive greater diversity.

The issues with respect to increasing diversity were also picked up by Illumina who outlined their commitment to building capacity with Next Generation Sequencing and data infrastructure in Africa. As an example, in January 2021 Illumina had placed about 5,000 instruments in Africa, but this had grown to more than 17,000 by July in response to the pandemic. They are considering how these instruments will be utilised in the future (and how the global community can assist) so that they can be repurposed to build capacity. Illumina has also been generating data, helping H3 Africa with the development of the Infinium Assay which is specific to the African population. That was started in 2014/2015 and has been used to generate datasets and Illumina are committed to continuing this work with H3 Africa. Illumina also have a philanthropic programme that has partners in Africa – samples are sent to the Illumina lab in San Diego and that information is also sent to ClinVar (more than 208,000 data submissions have been made). Illumina expressed a commitment to greater collaboration.

The Sanger has recently done some research examining who was accessing their dataset (as an open-access institute (ie access is free) this is important). The dataset is primarily Caucasian white European (they do also have some projects in Central Asia, South East Asia, parts of Africa and South America) and it is primarily accessed by people in Europe, the US and a smaller number from Australia. Over the course of 5 years there have only been 5 applications from across the entirety of Africa.

The Sanger have worked with several African partners to try to understand some of the reasons for this. Interestingly the barrier was not the largely Caucasian nature of the data (which was Sanger's initial thought). African partners were interested in the data, but they didn't feel they could access it or it "wasn't for them" – there was an assumption access would only be given to other European

organisations. There were also some technical issues about being able to download the datasets (a challenge) and lack of experience of handling data access agreements. Some individuals or organisations don't have well established legal teams or processes and this was a barrier. The Sanger are planning several seminars to try to counter some of these challenges and they are also working with GA4GH to try to improve the data access process, an important factor if they are to allow access to data more widely without compromising the integrity of the data access process.

Previous roundtables have considered the role that cloud-based technology can play in enabling access to data anywhere in the world as opposed to having to have the significant infrastructure created at cost everywhere. There are good examples of this – Genomics England, for example, identified that they make data available by allowing access to it via their platform hosted by Amazon Web Services so there is no need to download the data.

How do we ensure that the benefits of a more diverse data set benefit those who are participating in them? It can surely no longer be acceptable that benefits are derived from one data source and yet those participants never benefit directly. How can this be overcome?

If there is to be a global framework to focus on what we mean by "equity" this would need to look long term and think about how the benefit of data sharing is returned to the communities and researchers that are making these contributions. That must be not only the benefit of sharing the data but also the translation of that data and the outcomes of that data. It was identified that through the COVID-19 pandemic great data sharing across the globe was seen but when it came to sharing the benefit of that data sharing, that was not returned equally. There have been issues about protection of intellectual property, for example. Could a Global Framework or Global Partnership avoid this from happening again? Could it help establish how data is co-created in partnership and indeed how we co-share and co-benefit from everything that genomics can give the world?

Of course, there are challenges – cost and lack of equity in scientific collaboration has already been mentioned. But there is also scepticism, particularly when it comes to commercial entities seeking to collaborate, for example in relation to the 3 million Genomes Project in Africa – there were concerns that there could be strings attached to any sort of collaboration and yet if there is a desire to convert some of this data to health benefits then some sort of commercial partnership is going to be required. If concerns remain about historic exploitation will this hamper the use of data?

Part of the answer to this is about how we communicate what we mean by "commercialisation". This needs to be explained far better – by asking someone if they agree to allowing a company to convert your data into something useful; something that can benefit you or your community that is likely to get a more positive reaction than asking them if they mind their data being commercialised.

Do we need to change the language and talk about (and really mean) commercial **benefits** partnerships? Again, this comes back to education and communication – science communication, where we are better at communicating with participants and their communities about the importance of global diversity within genomics – for example, if we have a good African reference panel that can be included in other reference panels, that can go a long way to informing a particular response to a particular patient. But we need to have the data to create the panels in the first place. And these

benefits are important to the participants themselves but also to governments and to those who are going to be funding the future programmes.

If we are to bring this ecosystem together that is going to require funding. Is it realistic to expect Governments to attribute funding to a genomics programme now when the benefits may not materialize for some time? This is broader than pandemic preparedness – it is more about what we can use this genome sequencing capacity for. Funding will need to come from a variety of sources and there is a role for profit making organisations to play in building capacity and resource. But we need to create the environment for that conversation to happen and for all organisations to feel that they are being treated fairly.

In Thailand the national Genomics Thailand Initiative is initially focusing on the research, the collection of data and samples and carrying out the genome sequencing without immediate benefits or short-term outcomes. But it is a hard sell, especially in developing countries like Thailand. The focus must be on collecting the genome sequences to integrate into clinical benefits. However, once this begins, for example in relation to hereditary cancer, the researchers have discovered that about 40 per cent of the sequences are novel and it wasn't possible to even interpret their meaning. As more and more data is collected it is possible to start identifying what is common in the local population and further, when potential pathogenic variants were initially identified, it was established that half of the variants in Thailand are not present in ClinVar (even in common genes like BRCA 1 and 2 which is the most abundant in the ClinVar database). This recognises that important discoveries are being made but there is also a need for greater data sharing and collaboration to enable progress to be made more quickly.

Similar issues were identified with respect to the experience in Australia, but this also focused on some of the challenges about how we attribute ethnicity (and people's issues about how they identify their ethnicity) and what it means in terms of genetics. This is a complex area with ethical and technical challenges, particularly around building variant allele frequency databases for different ethnicities and use of reference genomes. It is now possible to build on graph versions of genomes where understandings of different haplotypes of different ethnicities can be added in and then have bioinformatic methods that can alter the way that mapping is done depending on these haplotypes. This next level of the bioinformatics challenge is going to be interesting and the work of GA4GH for standards of the reference sequence – enabling us to define exactly what reference genome is being used and how - is going to be important if we are going to be able to standardize the way that we say how we are using the genome and therefore how results can be interpreted between populations.

Once variants have been interpreted it would be ideal if they could be shared in real time and in context such that ClinVar starts to be a diverse and rich database – that is going to be important for clinical application. In addition, we need to consider how we ensure equity of access to clinical testing from diverse populations which may be from more remote or rural areas or countries.

Genomics England is also partnering with academics and companies such as Illumina to develop the tools they need from an analytical point of view to enhance the way that genetic data is processed, including addressing the "dark matter" issue – the missing DNA that we see in non-Europeans vs the reference sequence. There are gaps in the datasets that are widely used for interpreting and

prioritising variants found in these genomes and whilst the Data Diversity Initiative is primarily focused on the Genomics England dataset a key part of what needs to be done is make the information and tools available to the wider clinical communities and the resources, they will develop available to others so that they can support genomics in other populations.

It was agreed that collaboration between scientists is important. But questions were also raised about collaboration between those whose data is in the dataset between different countries. If we were able to put people together and allowing them to be a joint area of governance with a population in one country with a population in their country of heritage, that would be enormously empowering for everyone.

We know that genomic data is most meaningful when it has health data connected to it. That health data almost invariably can't leave country of origin, so we really must get to this idea of federated discovery, access and analysis of the data leaving the data in the country of origin. There are examples within rare disease communities where action is being driven by the families using, for example, Facebook, but this is not transferable or scalable. These interoperability issues need to be fixed at population/country/health system level.

There is a massive push in Europe around the European Health Data Space and the activities around that – it is going to be critical to think about how that data gets incorporated and how the genomics data sits into the overall picture so that the benefits of diversity can be brought to bear. A working group has just been established and endorsed by the national representatives of the One Million Genomes Initiative to generate a reference cohort which will be representative of each of the countries. The investment that will be made available will be reflective of the size of the country and needs to be diverse and not just from a disease cohort. That will then be combined by virtue of the federated data infrastructure that will be in place meaning there will be a massive resource in the future that is very diverse and not just a European population.

A further area that was raised related to incentives for participation. In the UK the Our Future Health Programme aims to be the UK's largest ever health research programme with a target of 5 million participants, designed to help people live healthier lives for longer through the discover and testing of more effective approaches to prevention, earlier detection, and treatment of diseases, by combining multiple sources of health and health-relevant data including genetic data. A question was posed as to what could be done with these individuals to give them something back for participating - could it be possible to move towards, with consent, a way in which large pharmaceutical or cosmetic companies could be linked with individuals so that when they are, for example, doing online shopping they might be able to pick out things that are either likely to have a better outcome for them or could highlight side effects? The issue of side-effects - trying to find a way to avoid 5-10 per cent of people ending up in hospital because of side-effects - was highlighted as a specific concern. It was acknowledged that there are social science challenges here, but it was suggested that there are real opportunities if the logistics, the cloud, and the companies could be connected with those individuals who consent to such an approach.

Technology will not be the barrier here – whatever the challenges are they will be human challenges; trust and inclusion and respect, incentive, and purpose and all those things that create a sense of humanity. The pandemic has been a powerful mobiliser, but it has also shown us all the disconnects.

Conclusion

On the surface, discussing the need to encourage and achieve greater diversity in genomic data could easily lead one to assume that the debate revolves entirely around the diversity of the data. Roundtable 5 served to showcase that this is far from true. There are many underlying factors which impede the achievement of greater diversity within genomic data.

There were a number of themes which kept emerging, the first of which was *community engagement.* The discussion highlighted some fantastic examples of engagement with various which also provided the opportunity to begin to explore in more detail how to bring benefit such communities. It was made clear that high levels of fear and concern remain surrounding genomics but what is key is remembering that countries and communities need to be empowered and engaged to find their own solutions and use their data for their own public health and policy decisions.

The discussion then evolved and focused upon *capacity building*. This encompassed a whole discussion not only on the 'bricks and mortar' but the fundamentals to providing a whole genome ecosystem. H3Africa for example are working with emerging economies within the region to build such frameworks, from training and mentoring scientists to discussing financial models for buildings and technology. *Training and mentoring* was another key aspect that was discussed. This addressed questions such as, how was can we ensure that this global ecosystem is not western dominated? And how can we allow emerging economies and minority communities to truly engage through training and mentoring?

As with other sessions, a significant topic of discussion during this session centred around *equity.* What do we mean by the word equity, how is it defined, both in terms of access and return? The global community must look at the long-term plan regarding how to best deliver the benefits of genome sequencing to all parts of the global fairly and without anyone being left behind. *Equity* must be defined as benefit for **all** and not simply for some. Africa has the capacity and is beginning to build the infrastructure to support such advancements and developed economies should support their empowerment to achieve such goals. In turn, this means that there is no reason why African nations cannot contribute to the global genomic ecosystem significantly. Collaboration and respect are also critical elements if we are to drive the ecosystem changes that were identified.

The discussion of roundtable 5 served as a stark reminder that if such issues, such as diversity, are to be addressed within genomics; it is not appropriate or sufficient to simply look at each issue in isolation. The addressed issues must be examined and considered holistically because – as was showcase in this roundtable – they are all intertwined.

There is more work to be done on these interconnected issues. PPP is keen to drive a greater focus on what we mean by equity and whether or how some sort of global framework or partnership is required which could better bring together all the elements of the genomics ecosystem.

Upcoming roundtable

Roundtable 6 will take place on 2 September 2021 and will focus on the economic benefits of having a genomics programme. Register for roundtable 6 HERE