Wellcome Sanger Institute

Representative Research strategy

Contents

1. What do we want to address?

Existing biases in research

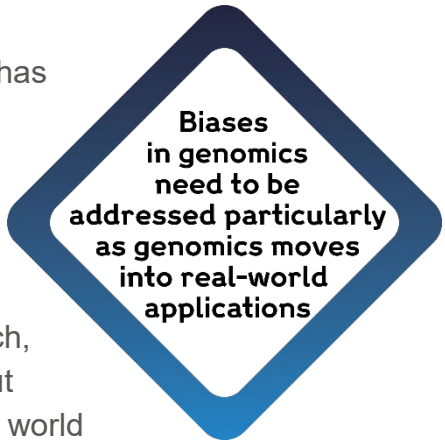
Over the past thirty years, genomics has scaled massively and research studies are using an exponentially larger number of genomes. Globally, countries are recognising the social and economic benefits of enhancing genomic capabilities and are making concerted efforts to increase their genomics capacity. Unfortunately, not all countries currently stand to gain equally from this genomics revolution. High-income countries have scaled their genomics capacity to a much greater extent than their low and middle-income counterparts.

This unequal global access to genomics is exacerbated by the human genomics research field facing traditional challenges recruiting and engaging participants from underrepresented groups. This has resulted in approximately 80% of global genomic datasets being derived from participants of European ancestries, despite Europeans constituting only approximately 10% of the global population. This representation bias has arisen due to numerous interconnected factors, such as the focus on research funding and research infrastructures in global north countries. Access to genomics capabilities and funding for genomics research has led to populations in the global north becoming a default study group. Many of the largest research cohorts that have been sequenced (including UK Biobank, 1000 Genomes Project, UK10K, 100,000 Genomes Project) are predominantly built on European populations. As a result, downstream research that utilises these cohorts have the same Europen bias with the research outputs being most relevant and beneficial to these populations. Furthermore, when we look more closely at these large cohorts, there is often a sampling bias where, for example, research participants are often recruited when they reside close to research recruitment centres resulting in a lack of cultural and socioeconomic diversity within the cohort. This leads to a lack of representation in terms of the global population, but also at a more local level within countries. Within the UK, there are efforts to counter these biases – as seen in the Our Future Health programme, which is actively committing to diversity in their recruitment.

Aside from investment and infrastructure, additional challenges to diversifying research participation include colonialism, historical misuse of biological data, lack of trust, and lack of equitable benefits arising from genomics. Understandably, these contribute to a hesitancy among many communities and populations to participate in research studies. The research community has a responsibility to understand these concerns and work with partners and communities to build trust and develop trustworthy science that can

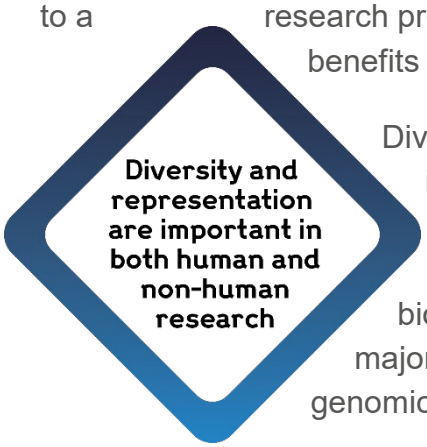
benefit all.

In recent years, a number of countries have established national and international population genomics cohorts. These include national initiatives in England (100,000 Genomes Project), USA (All of Us), Finland (FinnGen) and Uganda (Uganda Genome Resource), as well as continental initiatives in Africa (H3Africa) and Europe (The Genome of Europe/1 million Genomes Project). As the sequencing of these vast cohorts progresses globally, the research community will gain richer insights into global human genetic diversity. The key to unlocking the full potential of these resources lies in data sharing. Ensuring fair and equitable access to, and benefits from, these resources is crucial and must be a priority.



As well as the well-documented ancestral bias, research has historically also been biased on the basis of sex, gender, age, disability, and socioeconomic status, to name but a few. This pervasive lack of diversity and representation, particularly in genomics research, has become an increasingly pressing issue within the research community. Without appropriate representation in research, we can exacerbate social and health inequities throughout society. Moreover, as genomics research moves into real world applications, and is partnered with other big data innovations like AI, it has become an important issue of social justice that the biases in data and scientific outputs are addressed in order to prevent or exacerbate harms in communities already underserved by research.

When considering representation in research we often focus on the demographics of the research participants themselves. However, issues associated with representation and equity extend beyond solely the research participants and into all aspects of the research process including (but not limited to) how we collaborate equitably with global partners, how we conduct scientific analyses, how we fairly credit those who contribute to a research project, and how we share our research outputs to ensure benefits to those communities the research intends to serve.



Diversity and representation are also important in the context of non-human and biodiversity genomics. Biodiversity genomics has historically focused predominantly on European taxa and often excludes, for instance, biodiversity hotspots in the tropics or agricultural pests in the major food-producing countries. Similar to efforts in human genomics research, there are a growing number of projects

sequencing biodiversity across different countries and continents (e.g. Chilean 1000 Genomes Project, European Reference Genome Atlas, African BioGenome Project, Vertebrate Genomes Project, Darwin Tree of Life Project and the Earth BioGenome Project). The insights gained from these comprehensive efforts are pivotal for enhancing our understanding of life on Earth, informing effective conservation strategies, and unlocking the potential benefits of biodiversity for both society and the planet. It is for this reason that this strategy extends beyond human characteristics and health and intentionally attempts to encompass all aspects of Sanger's research, including biodiversity genomics.

Scope of our Representative Research strategy

The Wellcome Sanger Institute is one of the largest producers and holders of genomic data in the world. We conduct innovative and cutting-edge scientific research at a scale that cannot easily be conducted elsewhere, and our unique funding model enables us to think long-term and plan ambitiously. As a world leader in genomics, we have a responsibility to ensure that our science is as diverse, representative, and equitable as possible, and where appropriate, generalisable to wider populations. Moreover, to achieve our organisational mission and to enable Sanger science to bring maximum benefit to all people and our planet, we seek to improve representation and better reflect diversity throughout our science, including human and non-human research.

To formulate the Representative Research strategy, we chose to explore how we can address diversity, representativeness and generalisability throughout the research process and in the context of the whole portfolio of Sanger science - explicitly including both human and non-human research - rather than only the composition of research participants. This reflects the way in which Sanger science is conducted and the rich portfolio of global research, training and capacity strengthening activities that we currently undertake. Sanger researchers are involved in a huge variety of activities and initiatives that aim to improve representation, diversity and equity in genomics, including (but not limited to) involvement in building the human pangenome reference, creating diverse cell atlas maps in the Human Cell Atlas, utilising cohorts of underrepresented communities in research studies (e.g. Genes and Health, Born in Bradford), a comprehensive global training programme by Wellcome Connecting Science, and a vast array of widespread capacity-strengthening activities. These efforts are often researcher-led and we aim to encapsulate all our current efforts and build upon them to maximise the impact and societal benefit from Sanger science.

Although we have chosen to take a broad view, this strategy cannot cover everything. We have limited the scope of the Representative Research strategy to our research, training and capacity strengthening activities, and sought to avoid duplication or redundancy of ongoing efforts within and outside the institute.

A crucial aspect of improving the diversity and representativeness of our research lies in cultivating diverse and representative research teams. While we recognise the importance of a diverse workforce, specific measures to boost equality, diversity and inclusion in our research teams and the wider Sanger workforce are outside of the direct scope of the Representative Research strategy. These commitments are instead detailed in our Equality, Diversity and Inclusion (EDI) strategy.

What is out of scope?

Our Representative Research strategy will not directly address the following topics, however, we do recognise their importance and the intersection of these issues with diversity and representation in science:

- Equality, diversity and inclusion in research teams or the wider Sanger workforce
- Scientific racism
- Potential misuse of research to support racist views

Similarly, while it is important to acknowledge that issues associated with diversity, representation and equity in society feed through to all aspects of lives and livelihoods, they are outside of the direct scope of this Representative Research strategy. We recognise there is a complex interplay of factors leading to underrepresentation and inequity in science and society and that stakeholders across the life sciences sector and more generally across society have a role to play in tackling these issues. However, in this strategy, we describe activity that we will commit to as a research institute to deliver representative research, as well as areas we can influence and advocate to other stakeholders (e.g. to governments and research funders).

Terminology

Diversity, representativeness and generalisability are distinct moral and scientific concepts that are each inherently linked to equitable research practices. We explored how these different terms apply to Sanger science and how we can navigate these challenges in the context of how we conduct our research.

There is no “one size fits all” approach to addressing or defining the terms diversity, representativeness, and generalisability in the context of Sanger research. These terms can mean different things when considered from an institute perspective and at an

individual research study perspective. They can also be in conflict with each other depending on the context they are used in. We note that when discussing these terms with stakeholders, equity was also raised as an important related concept.

In Table 1 below, we outline some illustrative examples that show how these terms may be used in different research contexts:

Producing reference genomes in the Tree of Life programme	Diversity is fundamental to the Tree of Life (ToL) programme. The work of ToL uses DNA sequencing and cellular technologies to investigate the **diversity** of eukaryotic species within the British Isles and to generate high-quality reference genomes for individual species. Each reference genome will be used as a **representative** reference genome for each respective species. Furthermore, while the British Isles is not the most species-rich geographical area, the species within the British Isles **represent** nearly half of all families around the world.
Sequencing and utilising UK Biobank	Sanger sequenced the whole genomes of 500,000 UK Biobank volunteers together with deCODE. UK Biobank is widely used as a research dataset and it has been enriched with this sequencing data. However, the dataset is **not diverse or representative** of the UK population and it demonstrates "healthy volunteer" selection bias. Researchers will need to be mindful of this when **generalising** their research findings.
Utilising the human pangenome	Sanger researchers are contributing to the development of the human pangenome reference genome. The aim of the human pangenome is to create a reference genome that better reflects global genomic **diversity**. A proof of concept based on 47 individuals was published in May 2023 and the goal is to increase this to 350 human genomes later this year. The human pangenome will be generated from ancestrally **diverse** people and aims to present a unified genomic **representation** of the human species.
Project JAGUAR- studying immune cell diversity across	Sanger scientists are working alongside collaborators in Latin America to generate the first high-resolution genetic atlas of immune cells in Latin America. The atlas will highlight factors that impact immune system development and how this is impacted by the environment. The project focuses on the understudied Latin America population and

Latin America	study participants were recruited from seven distinct regions in Latin America including urban Mexico and the Brazilian Amazon rainforest. Research findings will improve the **representation** of Latin American populations in genomics research and better reflect the **diversity** of these populations and their associated environments.
Building a representative Human Cell Atlas	The Human Cell Atlas (HCA) aims to create **diverse** reference maps of all human cells as a basis for understanding human health and diagnosing, monitoring and treating disease. The research consortium is incorporating **diverse** samples to better understand and treat disease for all humans. The samples will represent both sexes and various ages and lifestyles from as many populations as possible from a diversity of environments.

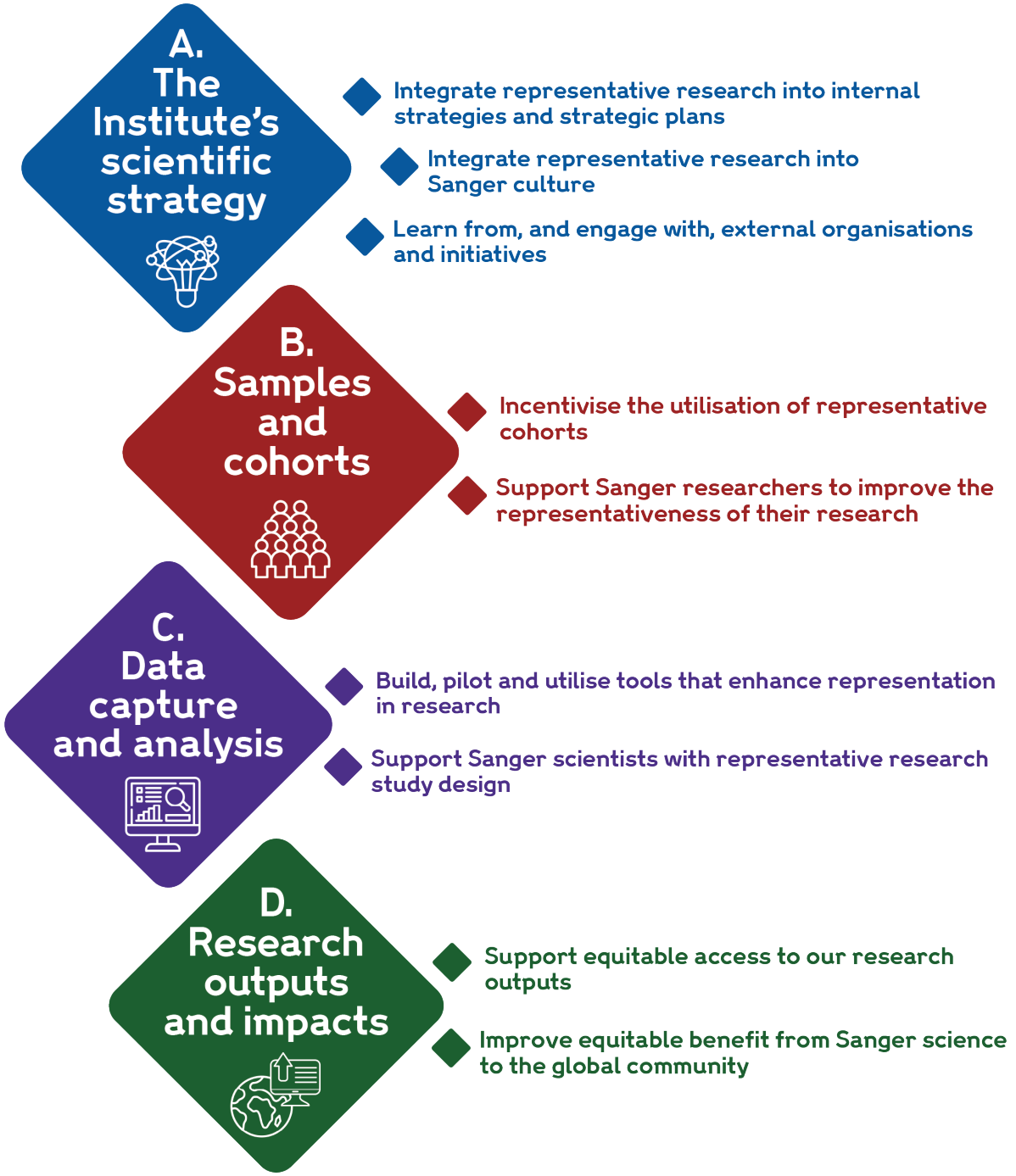
Table 1: Illustrative examples for how the terms '*diverse*', '*representative*' and '*generalisable*' are used in different research contexts.

In developing this strategy, we chose to focus primarily on representation. We felt the concept of representation encapsulates the concepts of diversity, generalisability and equity in context specific ways. This focus on representation allows us to thoughtfully align our strategy in the context of our science by considering the intended beneficiaries of individual research projects. For example, it might not be appropriate to use a geographically diverse set of samples in a study with a regional focus. Moreover, we believe that emphasizing representative research also encourages fair collaborations with local researchers, ensuring that benefits from research projects are shared equitably.

Framing the strategy in terms of representativeness has enabled us to consider what is appropriate, feasible and pragmatic within the context of Sanger science. This approach also allows us to address issues of diversity, generalisability and equity within that context.

2. What are we going to do?

Our 10-year vision is that Sanger science is designed and conducted in an equitable and more representative way and delivers research outputs that benefit our planet and societies around the world. To realise this vision, we will focus on four aspects of the research process that we believe have the greatest impact on the representativeness, diversity and generalisability of Sanger science. The four focus areas are:



A. The Institute's scientific strategy

We aim to weave representative research into the fabric of the Institute. We will:

i. Integrate representative research into internal strategies and strategic plans.

We will embed our commitment to representative research into the design of our scientific strategy and the subsequent delivery of our internal strategic plans. We recognise that broad representation will not be feasible, pragmatic or appropriate for every individual research project we undertake. However, we will seek to embed our commitment to improving representation in the overarching portfolio of Sanger science. This will include embedding representative research in the delivery of our strategies, ambitions and priorities for each of our Scientific Programmes, Wellcome Connecting Science and Management Operations.

ii. Integrate representative research into Sanger culture.

We will ensure representative research remains an important part of our research culture and environment. There are currently a large number of researcher-led activities which support representative research and we will improve how we showcase, support and reward these efforts. We will also explore opportunities to further incentivise and support representative research, for example by incorporating it into our existing relevant training programmes (e.g. Good Research Practice training).

iii. Learn from and engage with external organisations and initiatives.

The research ecosystem as a whole has a collective responsibility to address representation and equity throughout the entire research lifecycle. This includes

individual researchers, research organisations, funders, publishers, governments, industry and healthcare providers. Many stakeholders are already addressing diversity and representation, particularly in the field of genomics (e.g. Genomics England, GA4GH, Earth Biogenome Project, Human Cell Atlas). To drive meaningful change, we need a shared vision across the sector as well as effective collaboration with diverse and global partners. As a leading global genomics institute, we endeavour to leverage our existing networks and build new partnerships (in the global north and global south) to pioneer opportunities to address the lack of diversity and equity in genomics research. In support of this strategy, we will monitor the external landscape, contribute to and lead horizon scanning activities, adopt new tools, develop and implement sector recommendations, or build, pilot and utilise new approaches in our work.

B. Samples and cohorts

We are committed to enhancing representation in the cohorts we use and the samples we study so that our research outputs can benefit all people and our planet. To achieve this, we will:

i. Incentivise the utilisation of representative cohorts.

For human genomics, access to diverse and representative human cohorts will be a key factor in achieving our Representative Research vision. However, Sanger researchers tend to be ‘collators’ rather than ‘collectors’ of samples and often work with third parties (e.g. other researchers, biobanks or existing research cohorts) to access samples. This poses significant limitations on the ability of our research teams to diversify their research and also highlights the dependencies on other stakeholders to address and enable representation in scientific research.

We will explore how we can best utilise new and existing partnerships and cohorts (or indeed subsets of existing cohorts) that may improve the representativeness of Sanger

science. Furthermore, we will advocate for the creation of more diverse and representative cohorts, particularly to other stakeholders who are better positioned to create and maintain large cohorts (e.g. funders and government). Similarly, for biodiversity genomics, we will advocate for and support the generation of reference genomes of biodiversity around the world.

ii. Support Sanger researchers to improve the representativeness of their research.

Sanger research teams are keen to ensure their research has the maximum benefit to global communities and populations, and the planet. However, pressured research environments and heavy workloads can disincentivise representative research. We will therefore support research teams to consider equity and improve representativeness in their research by, for example, facilitating global partnerships (e.g. establishing collaborations and navigating bureaucracy), advocating and boosting compliance support for access and benefit sharing, creating guidance and resources for researchers, and incentivising the consideration of representation at study setup (where possible) via direct and indirect approaches.

C. Data capture and analysis

We will build and use tools, methodologies and practices that enable representative research. Specifically, we will:

i. Build, pilot and utilise tools that enhance representation in research.

There is a recognised need across the sector for tools that enable global populations to benefit fairly from science, and specifically genomics. We will work with relevant organisations and initiatives (e.g. GA4GH, Genomics England, EMBL-EBI, Human Cell Atlas, Earth BioGenome Project and its sub projects etc) to explore, build, pilot and

utilise tools and methodologies that consider the intersectional aspects of representation and can improve equity in genomic science. This may include creating and sharing methods for analyses, building standards and generating ontologies. We will make outputs openly and freely available as appropriate.

ii. Support Sanger scientists with representative research study design.

Designing a research study that prioritises representation is not a trivial task and there is no 'one size fits all' approach to improving representation in research. We will therefore support research teams to improve representation in their science by raising awareness of the importance of representation in research, create and signpost to internal and external support resources, and form recommendations for research teams to consider representation. We will also monitor external developments, be part of collaborative consensus building to address these, and we will seek to implement new recommendations and guidance on issues related to representative research (e.g. on diversity ontologies, sex disaggregation etc).

D. Research outputs and impact

We will improve equitable access to our research outputs, and support positive impact and benefit from Sanger science. Specifically, we will:

i. Support equitable access to our research outputs.

Through the Bermuda Principles and the Fort Lauderdale Agreement, the Wellcome Sanger Institute was founded upon the principle of open science and we make our data, software and research outputs openly available wherever possible. We recognise that although our research outputs are generally *openly* available, there is more that can be done to ensure they are *equitably* accessible. As such, we seek to incorporate representative research into our Open Science statement, build transparency and

address misconceptions throughout our data sharing processes, and explore alignment with FAIR (Findability, Accessibility, Interoperability and Reusability) principles.

ii. Improve equitable benefit from Sanger science to the global community.

Sanger science has a truly global reach and impact - we have collaborated with partners in 178 countries (since 1996) and conducted training and capacity building activities in 130 countries (between 2019-2023). We aim to ensure equitable benefit from Sanger science but recognise that many factors are beyond our direct control (e.g. political environments, infrastructure, resourcing etc). As such, we seek to improve equitable benefit from Sanger science by building upon and raising the visibility of our existing capacity strengthening and global training initiatives; recognising power imbalances in our global partnerships and ensuring that these are not inadvertently exploited during collaborative research endeavours; and building upon our computational infrastructure to enable more equitable access to data and resources from around the globe.

3. What might hinder us?

We face an array of challenges and barriers that may hamper our ability to deliver representative research, including:

Risk	Mitigation
We fail to deliver the Representative Research strategy due to lack of resources and funding.	Where appropriate, we will delegate responsibility for delivering the strategy across relevant teams as part of their Business As Usual (BAU) and will initially prioritise activities that are high impact and low cost (in terms of finances and resources).
Competing priorities hinder engagement from researchers.	We will develop a tailored communications plan to raise awareness of the Representative Research strategy and our ambitions to address representation in Sanger science. We will initially focus on impactful changes that can better support researcher-led initiatives that improve representation in Sanger science.
New and/or enhanced support for addressing representation in research is not utilised by researchers.	We will raise awareness of new and/or enhanced support via a range of internal communication methods (e.g. Fred pages, email, referenced in Good Research Practice training etc).
We fail to develop equity-enhancing tools and/or fail to pinpoint pragmatic approaches to improve representation in research.	We will continue to engage and collaborate with existing partners (e.g. GA4GH, Earth BioGenome Project) and will establish new partnerships that enable us to collectively build, pilot and implement pragmatic equity-enhancing tools and methodologies.

Table 2: Risks that may hamper our ability to deliver the Representative Research strategy.

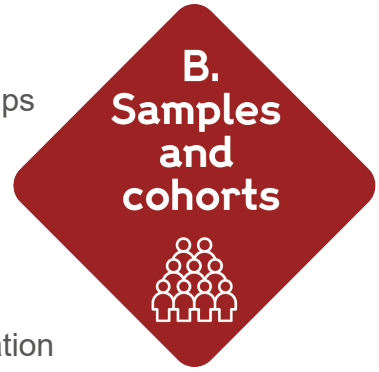
4. What will success look like?

Our 10-year vision is that Sanger science is designed and conducted in an equitable and more representative way and delivers research outputs that benefit our planet and societies around the world. We will regularly evaluate our progress and our measures of success over the next 10 years will include:

The Institute's scientific strategy:

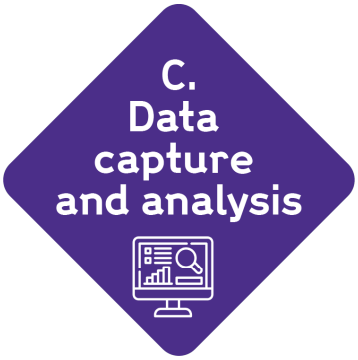
- The delivery of our organisational strategies and strategic plans reflect our commitment to representation, equity, diversity and generalisability in Sanger science.
- Our research and support teams value and foster equity and representation in scientific research.
- Representation, equity and diversity are reflected in the science that we do.
- Representative and equitable research is incorporated in our internal scientific training provisions (e.g. good research practice, research integrity). Training provisions are utilised by everyone who needs them, and we consistently receive positive feedback.
- We have contributed to a network of global partnerships that is driving forward representation, equity and diversity in genomics research.

Samples and cohorts:



- Increased representation of traditionally underrepresented groups in Sanger science.
- Increased utilisation of diverse human cohorts (e.g. national genomics cohorts).
- Enhanced support for research teams in addressing representation and equity in their research (e.g. support for legal, training, capacity strengthening, patient and public involvement).

Data capture and analysis:



- Resources and training on equitable and representative research study design are available to and utilised by everyone who needs them.
- Successful development and utilisation of equity-enhancing genomics tools and methodologies.

Research outputs and impact:

- Improved equitability of global access to research outputs.
- Increased delivery of successful and impactful capacity strengthening activities globally (e.g. in-country bioinformatics training, ‘train the trainer’ courses, mentoring, accessible data pipelines).
- We have contributed to a strengthened genomics research and analysis skills base in the Global South that enables regional implementation and application of genomics knowledge.