+ Our Future Health

In partnership with NHS

Processing and releasing genetic data at Our Future Health

Ben Elsworth

Lead Data Engineer

ben.elsworth@ourfuturehealth.org.uk

+
Our
Future
Health

Forthcoming series of Charity Researcher Webinars

- We are pleased to announce a series of webinars for researchers organised in collaboration with our Charity Partners.
- For further details, please email partnerevents@ourfuturehealth.org.uk to register an interest
- Here are the dates for the upcoming webinars, which will feature several of our Early Adopters.

1. British Heart Foundation & Our Future Health – Wednesday, April 2, 9-11 am
2. Cancer Research UK & Our Future Health – Monday, May 12, 3-5 pm
3. Asthma and Lung UK & Our Future Health – Tuesday, June 10, 12-2 pm
4. Affiliate Charity & Our Future Health Webinars – Monday, June 16, 10-12 pm and Thursday, June 19, 2-4 pm

Agenda

1. Programme Overview
2. Genetic Data
3. Where We Are
4. The Future

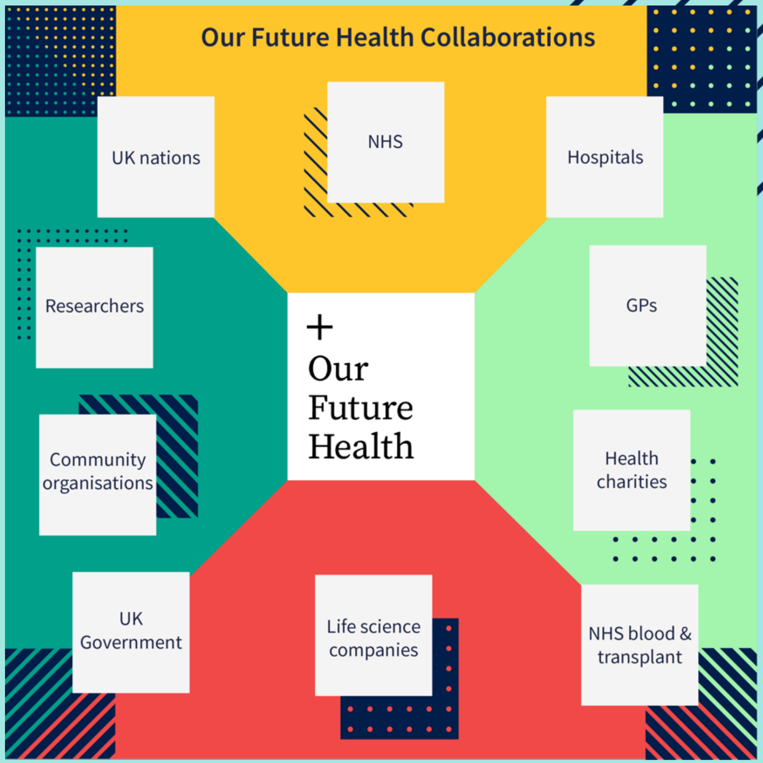
+
Our
Future
Health

Our Future Health

+ Our Future Health

A strong collaboration across the life sciences sector and health system

- Our Future Health is designed to harness the power of collaboration so we can collectively help people live healthier lives for longer.
- We’re a not-for-profit organisation combining support from industry, charities and government to build a world-leading health research programme



Funders

- UKRI
- Cancer Research UK
- LifeArc
- Asthma + Lung UK
- British Heart Foundation
- Alnylam Pharmaceuticals
- AstraZeneca
- Boehringer Ingelheim
- Biogen
- MSD
- illumina
- Johnson & Johnson Innovation
- novo nordisk
- Pfizer
- GSK
- RANDOX
- RGC Regeneron Genetics Center
- NOVARTIS
- Roche
- AMGEN Oncology
- ThermoFisher Scientific

Affiliate charities

- Alzheimer’s Society
- Kidney Research UK
- Stroke Association
- Chest Heart & Stroke Scotland
- Glaucoma UK
- debra The Butterfly Skin Charity
- Versus Arthritis
- Prostate Cancer Research
- Blood cancer UK
- Breast Cancer Now The research & care charity
- Fight for Sight
- Parkinson’s UK
- Macular Society Beating Macular Disease
- Royal Osteoporosis Society Better bone health for everybody
- Action Against AMD
- Prostate Cancer UK
- Diabetes UK Know diabetes. Fight diabetes.
- Brain Tumour Research
- Pancreatic Cancer UK
- Alzheimer’s Research UK For a cure

+
Our
Future
Health

Central aims of the programme

- Build a new prospective cohort in the UK with the mission to help everyone live longer and healthier lives through the **discovery and testing of more effective approaches to prevention, earlier detection, and treatment of diseases**.
- Recruit 5 million volunteers from the UK adult population to create a unique resource which will **catalyse aetiological and translational research for both common and rare diseases**.
- **Facilitate translational research** by providing **access to data and stored biosamples**, and by **enabling recontact studies/trials with participants selectively invited** based on demographics, phenotypes, and disease risks.
- Provide **personal disease risk information for participants**, based on genetic and non-genetic information, and deliver **population health insights**, facilitating insights into the health of the whole population and the assessment of the **impact of interventions/ policies at the individual level**.

Recruitment

+ Our Future Health

Age 18+

UK resident

Our Volunteers

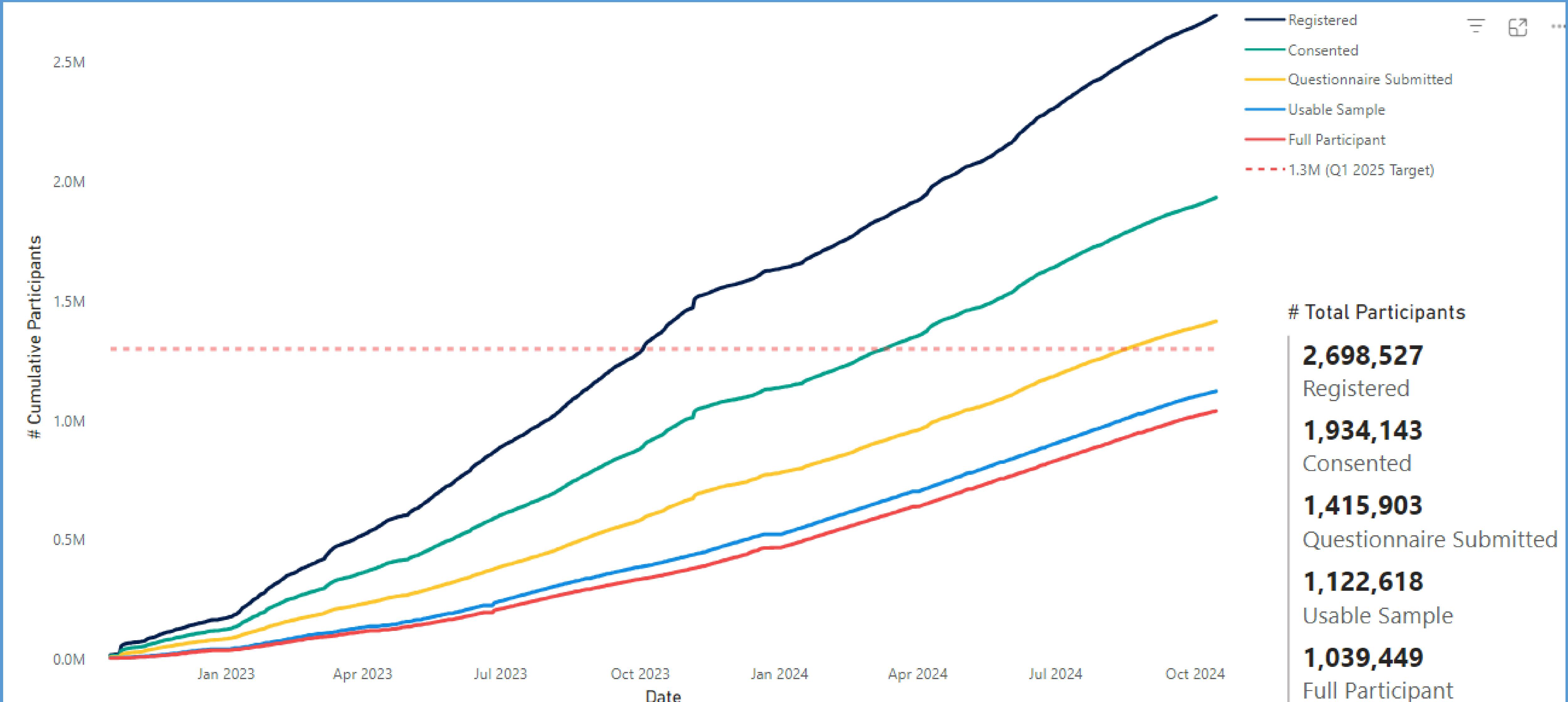
- We’re monitoring recruitment to ensure our participants are reflective of the UK population by **age, sex, ethnicity and deprivation.**
- Our Future Health will be the largest cohort of:
  - Multi-ethnicities in the UK
  - Non-European ancestry in the world
  - Young people and working age populations
  - Over 60s

+
Our
Future
Health

What our participants donate to Our Future Health

- Complete a baseline questionnaire (demographics & household, work & education, lifestyle, family history, personal health history)
- Physical measurements taken at clinic appointment (height, weight, waist circumference, blood pressure, heart rate)
- A blood sample for genotyping and for long-term storage of plasma, buffy coat, and DNA
- DNA is genotyped using a custom array that captures 700k up-to-date variants of interest for optimal research insights
- Provide permission for data linkage to health-related datasets (e.g., primary care, secondary care, cancer, death)
- Provide permission to be recontacted for further surveys/ biospecimens/assessments and to be invited to additional studies/trials

Recruitment to date



Total Participants

2,698,527
Registered

1,934,143
Consented

1,415,903
Questionnaire Submitted

1,122,618
Usable Sample

1,039,449
Full Participant

+
Our
Future
Health

13 participants
May 2021

The Our Future Health custom genotype array types 700,000 variants for aetiologic and translational research

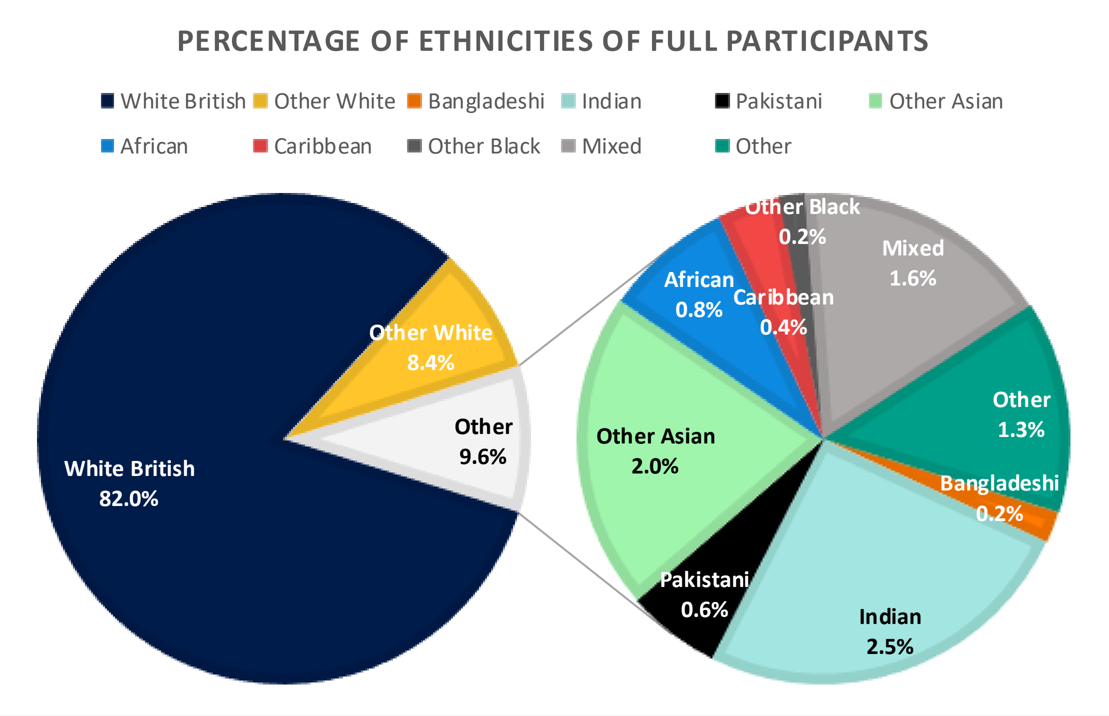
- GWAS backbone is **optimised for multi-ethnic imputation capability**
- Design incorporates **up to date sets of disease- and phenotype-associated genetic variants** from the Genotype Assay Working Group including polygenic risk score (PRS), GWAS catalog, ACMG, and ClinVar sets
- The array is also designed to **predict blood types and assess pharmacogenetic variants** to further enhance healthcare insights.
- We have built an integrated genotyping workflow and have contracted with Genomics to provide imputed data, polygenic and integrated risk scores.
- Further **concordance analyses** will enable any additional array optimisation and a final assessment of accuracy relative to a gold standards (WGS, blood type serology)

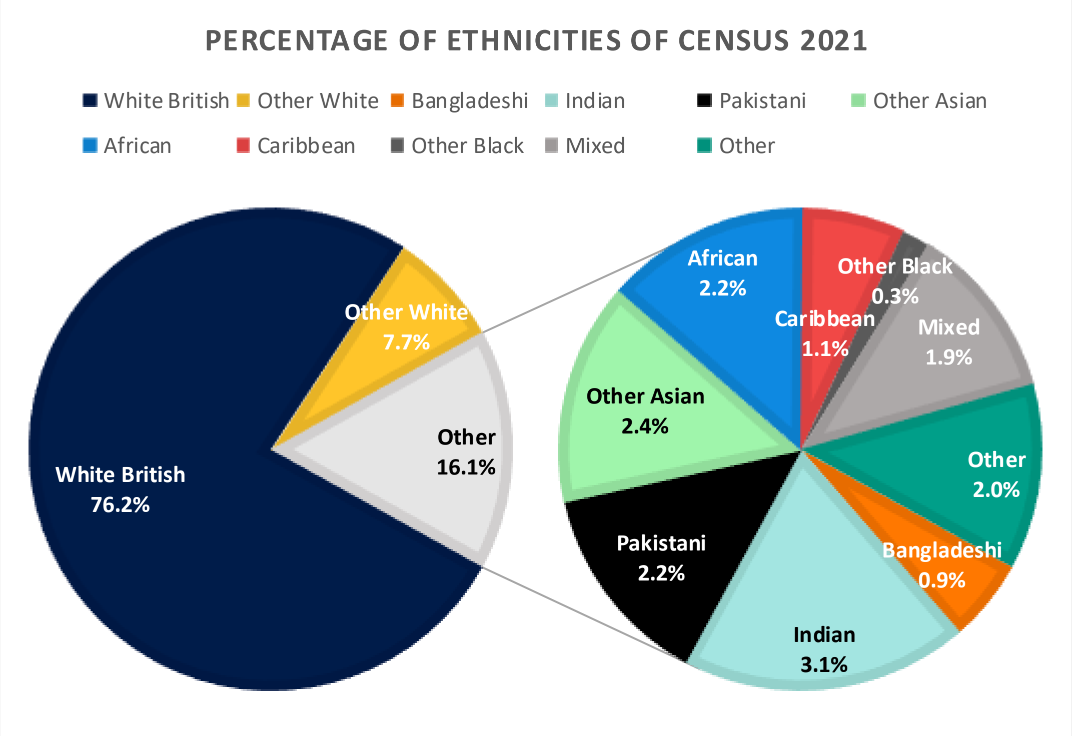
biocentre illumina eurofins GENOMICS

+
Our
Future
Health

Progress to Date

Demographics - 1,035,446 full participants

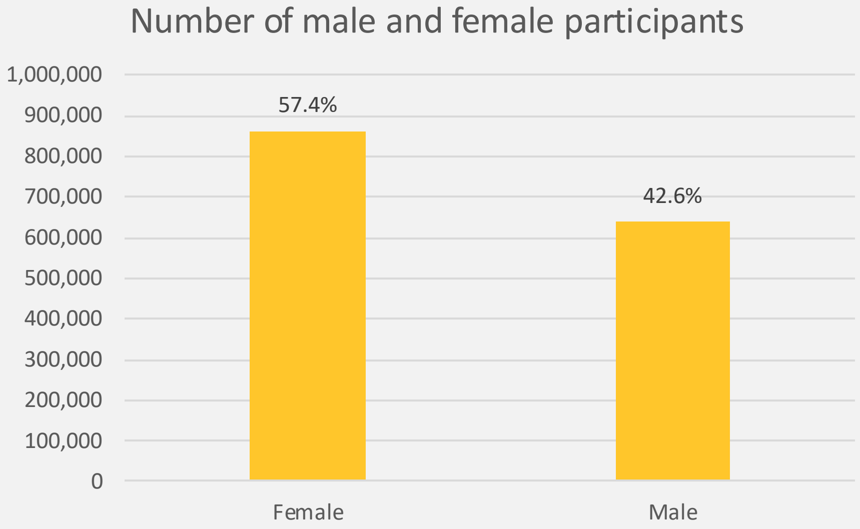

PERCENTAGE OF ETHNICITIES OF FULL PARTICIPANTS

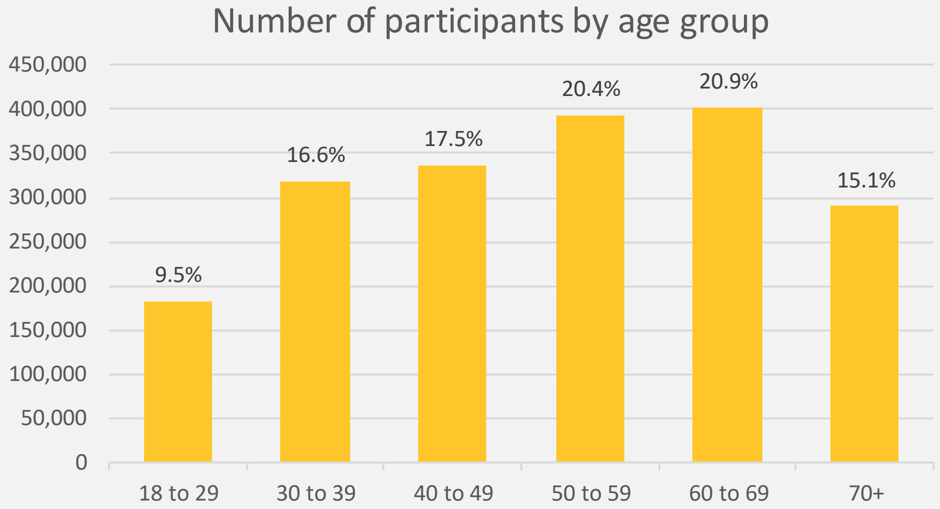

PERCENTAGE OF ETHNICITIES OF CENSUS 2021

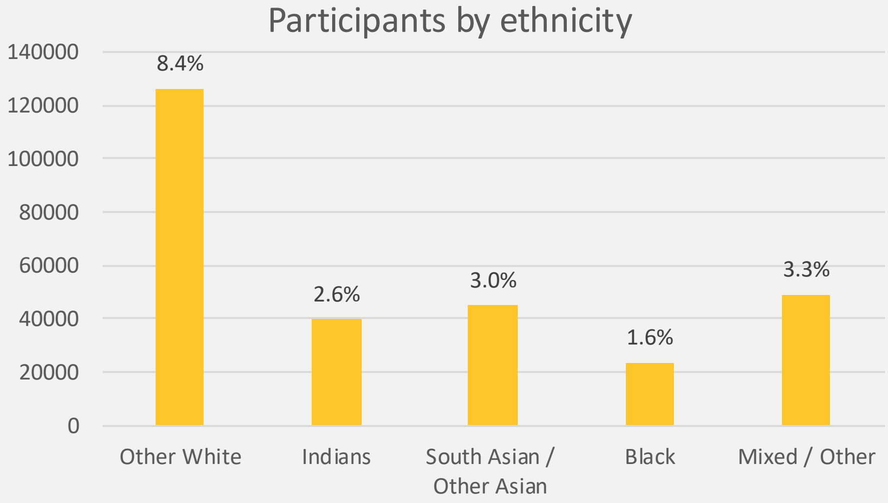
Data up to 13 October 2024

+
Our
Future
Health

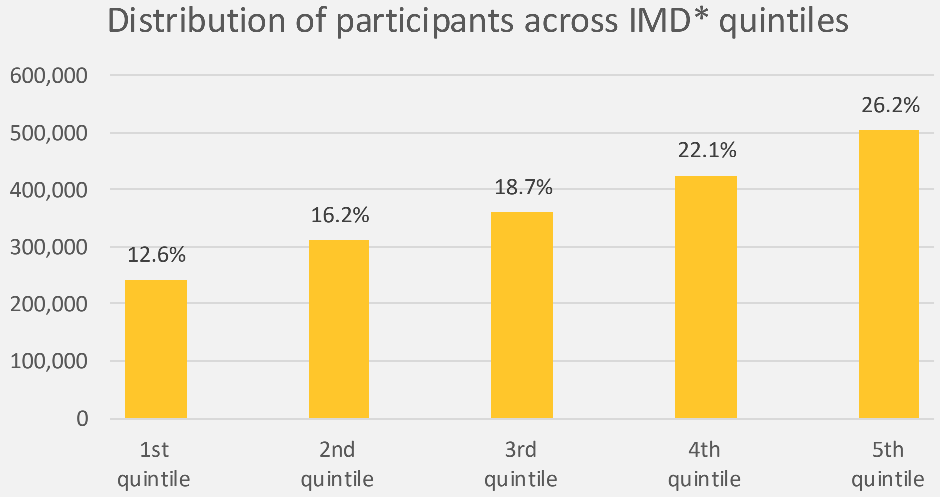
Demographics –1,925,136 consented participants







*1,219,122 (81.1%) White British participants



*Index of Multiple Deprivation - 1st quintile is most deprived; 5th is least deprived

Data on sex and ethnicity are shown only for participants who started the questionnaire

Data up to 13 October 2024

Our Future Health

A platform for research

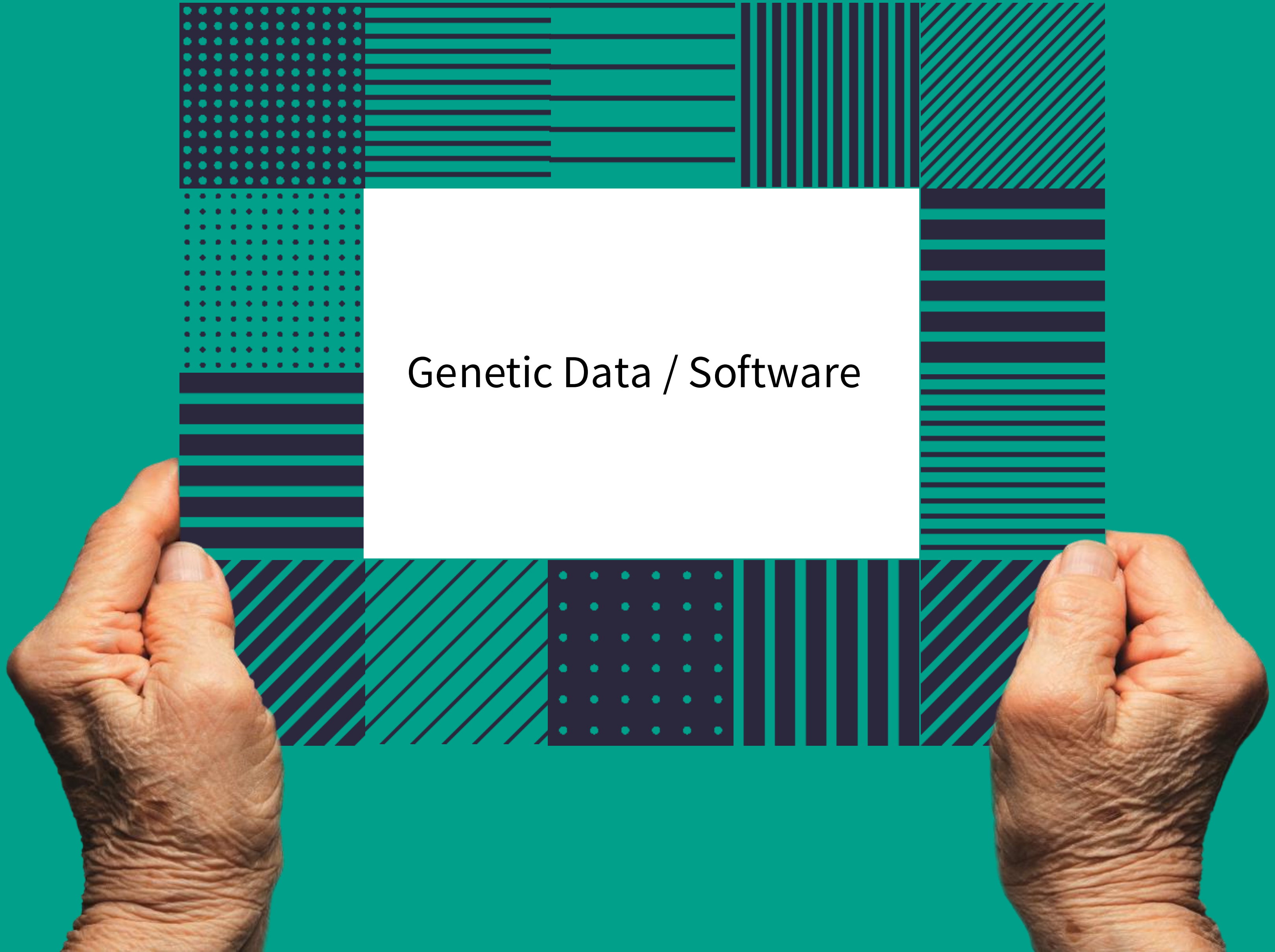
What data *do* we have and what data *will* we have on our volunteers

+ Our Future Health

Secure, de-identified data available to approved researchers in our Trusted Research Environment

1,414,260	participants with baseline health questionnaires
651,050	participants with genotype array data
1,151,453	participants with NHS-E secondary care data
1,025,498	participants with clinic measurements data
5,128	participants with imputed data *(for testing purposes only)*

+
Our
Future
Health



Genetic Data / Software

Genetic data is…

Varied and esoteric:

- Raw genotype data contains **7** different file types, **3** of which are specific to genetics
- Release data contains **25** different file types, **12** of which are specific to genetics

Relatively large:

- 710k genotyped / 160M imputed variants per person
- 4 different values for each genotype
- 1.8 trillion data points (P9 = 650k samples x 710k variants x 4 data points)

Often redundant:

- Raw genotype data for each participant contains the exact same data for 9 out of 10 columns
- 50 TB of raw data = 2 TB of release data

In demand:

- Researchers are very keen to get access ASAP.

Genetic software is often…

Frequently file based:

- Transformations from one file type into another

Domain specific:

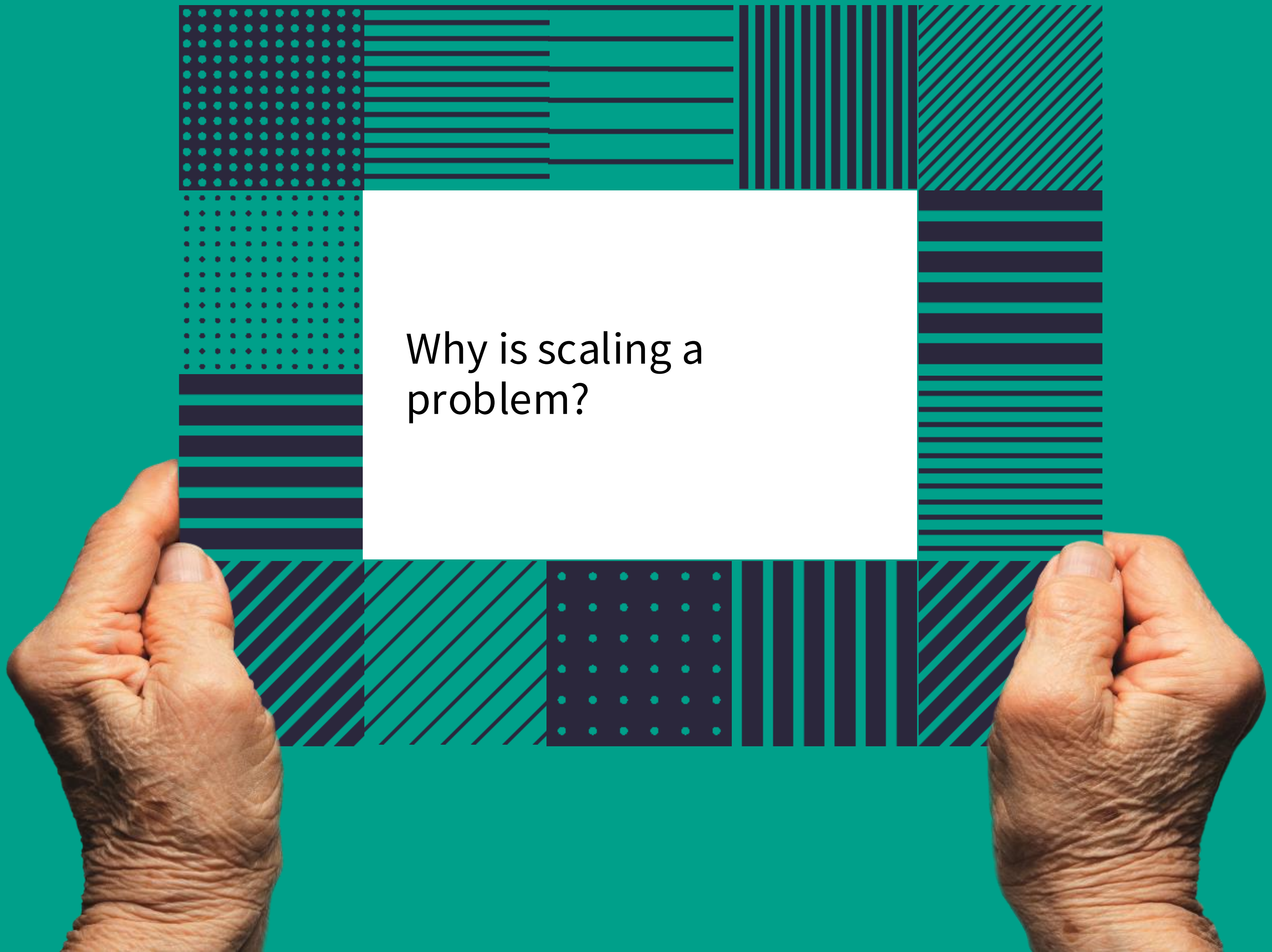
- Little uptake outside the bioinformatics community

Not cloud native:

- Historically run using on premise compute

Not production ready / optimized:

- Single threaded
- Rigid format expectations

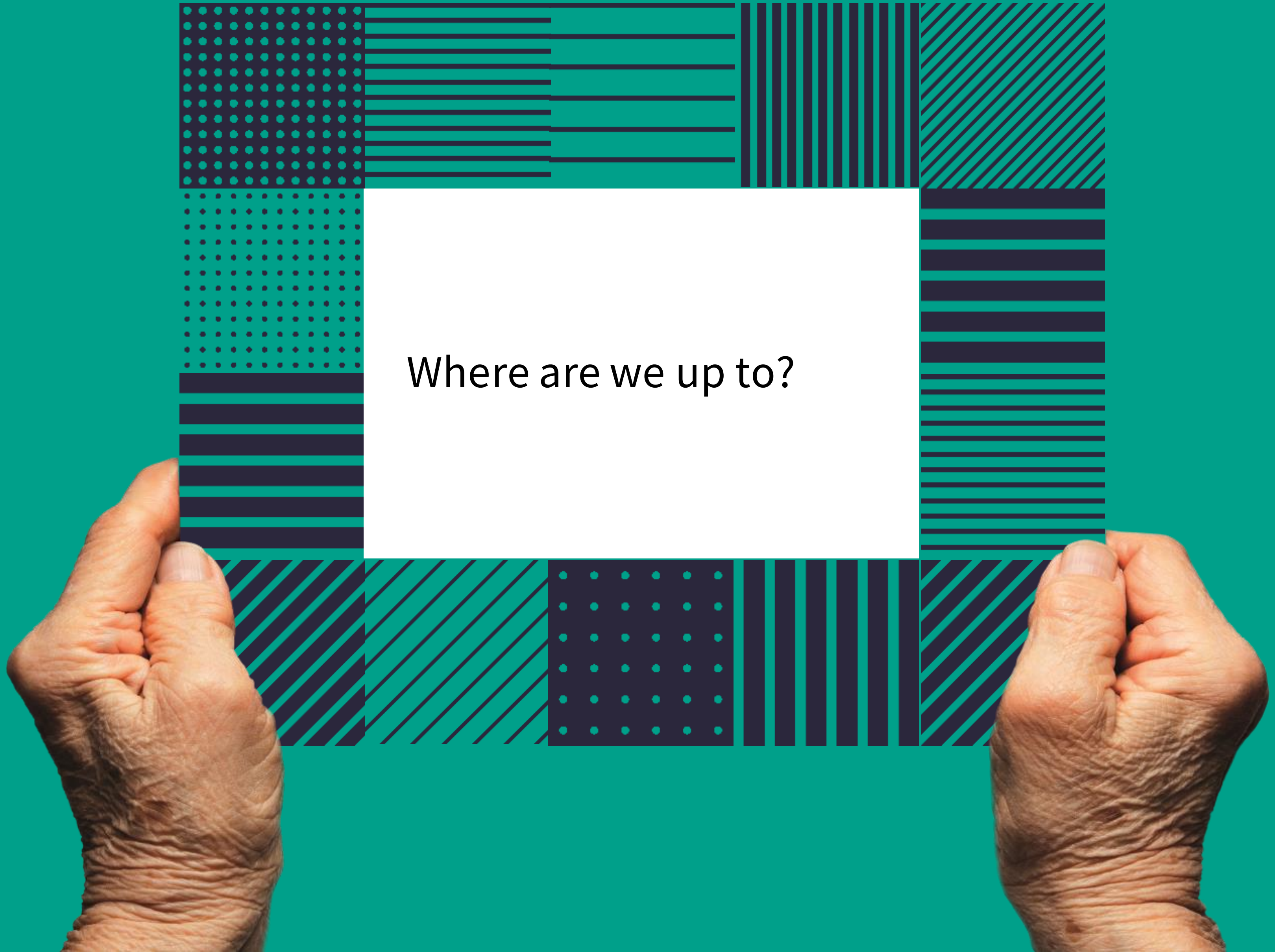


Why is scaling a problem?

Why we can’t use standard methods

Domain specific files and tooling don’t lend themselves towards traditional methods.

- To add or remove a sample requires reading and writing all data every time
- This gets slower as the cohort grows
- Works for now, but alternative is required long term



Where are we up to?

Overall

- To meet the quarterly release cycle, and withdrawal requirements, we have (probably) **built the most efficient and advanced genetic data processing and release pipelines in the world**, in record time.

- This requires **specific skills and expertise from across various domains**, e.g., engineering, bioinformatics, statistical genetics, product, delivery and management,

- **Compared to other similar initiatives, the squad is very small**. However, we are fortunate to have invaluable assistance from the wider organization.

- To date, we have **met all deliverables**, but this has not been without a cost, and we still deal with large amounts of technical and data debt.

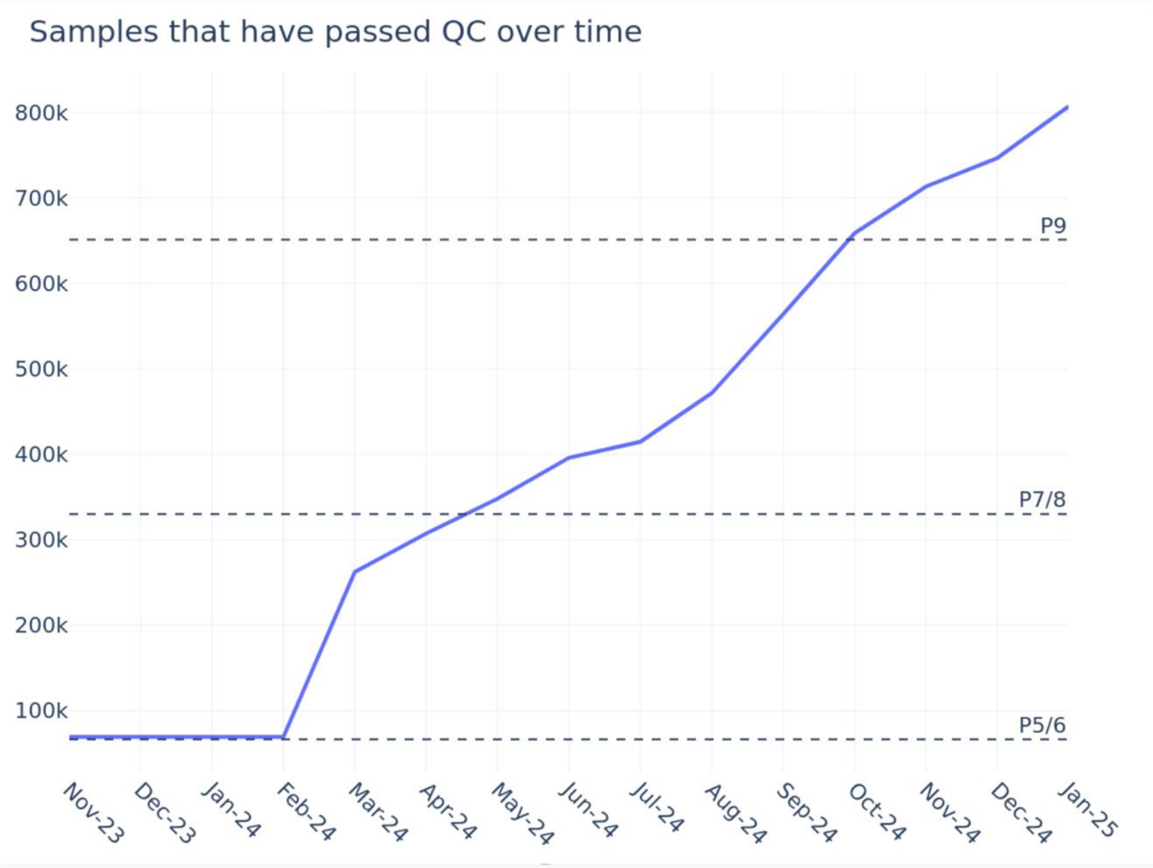
Successes

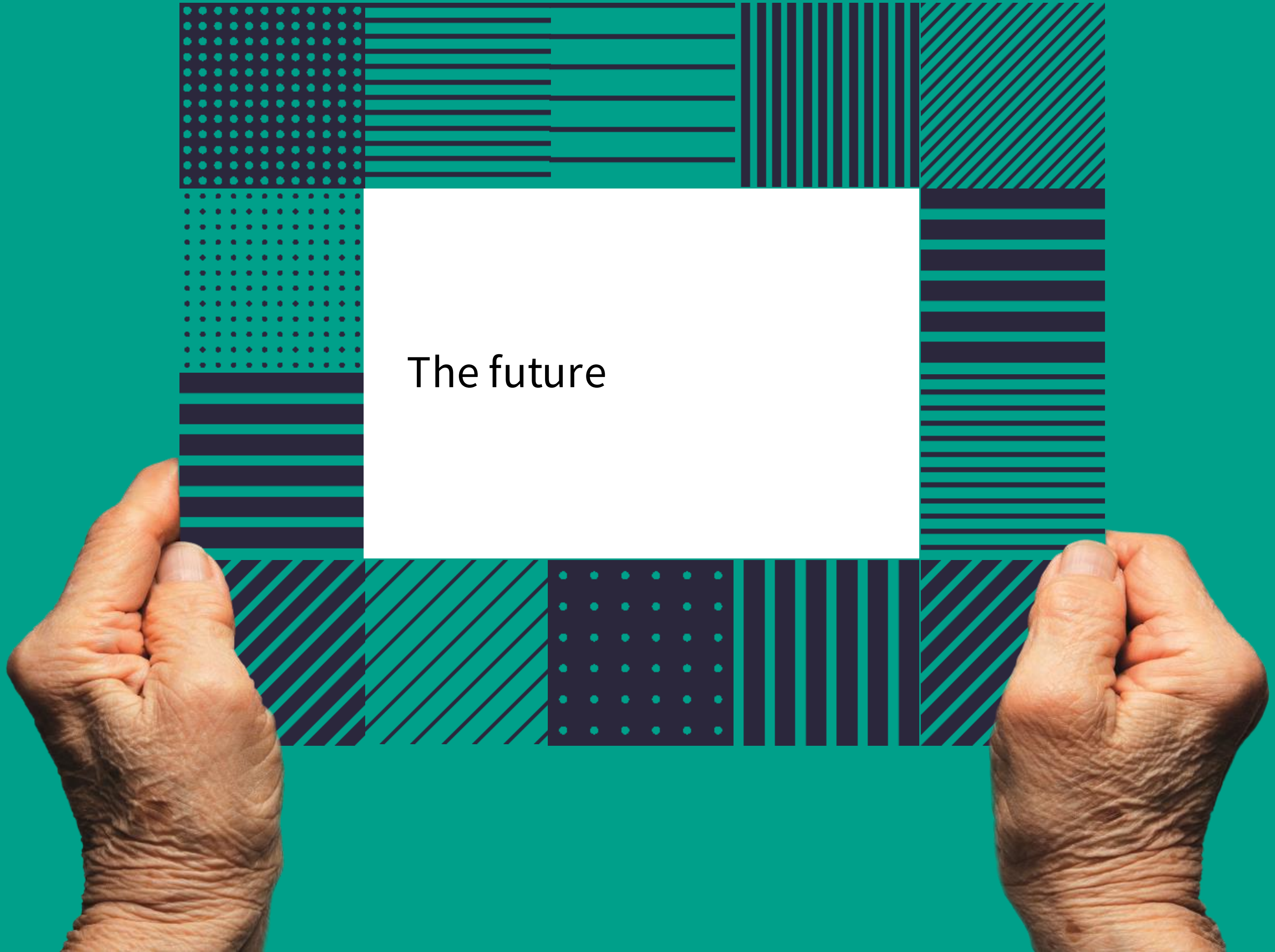
1. Received, QC'd and tracked genotype data for over 1 million participants (2 million total).

2. Built robust and scalable pipeline that can produce genotype release data for millions of participants in under 24 hours

3. Within 12 months of receiving data we were able to create a genotype data release for 660k participants

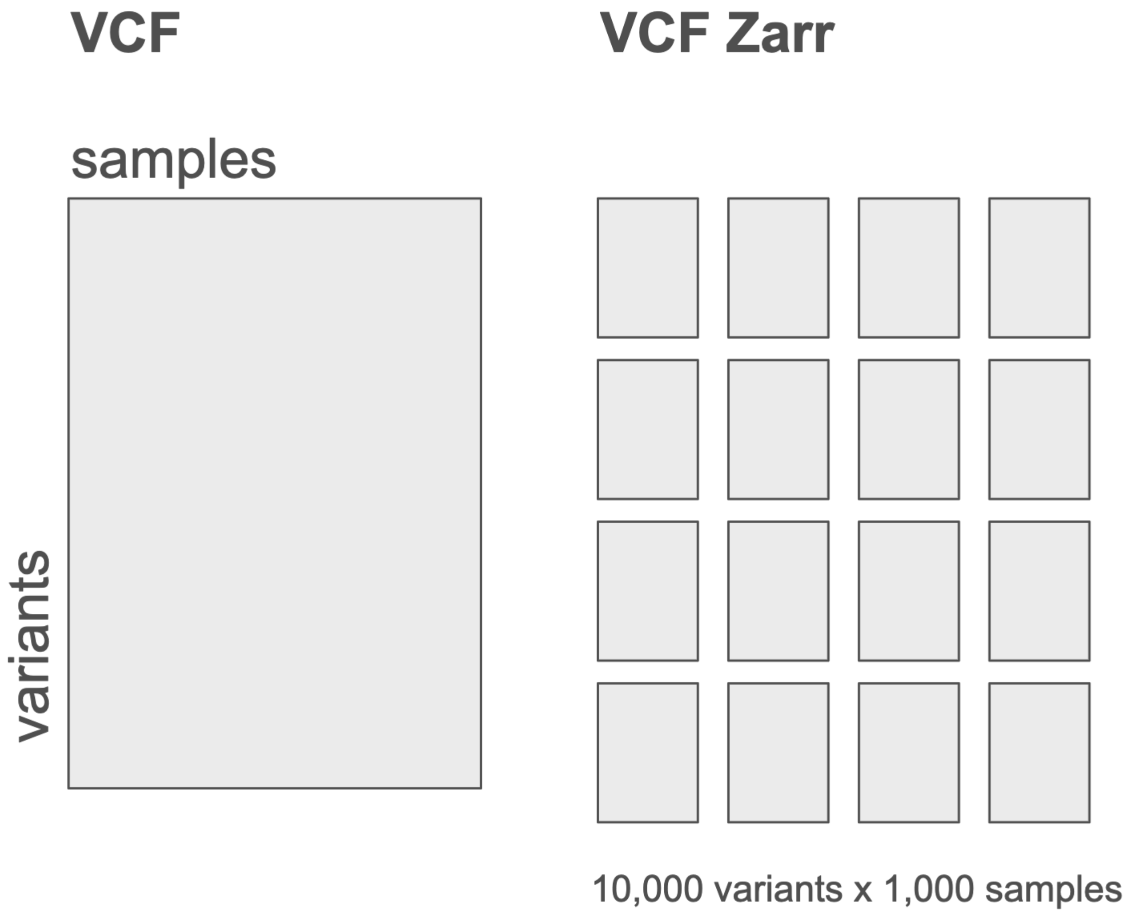
4. Set up imputation data flow via Genomics Ltd and UK Biobank and released first data





The future

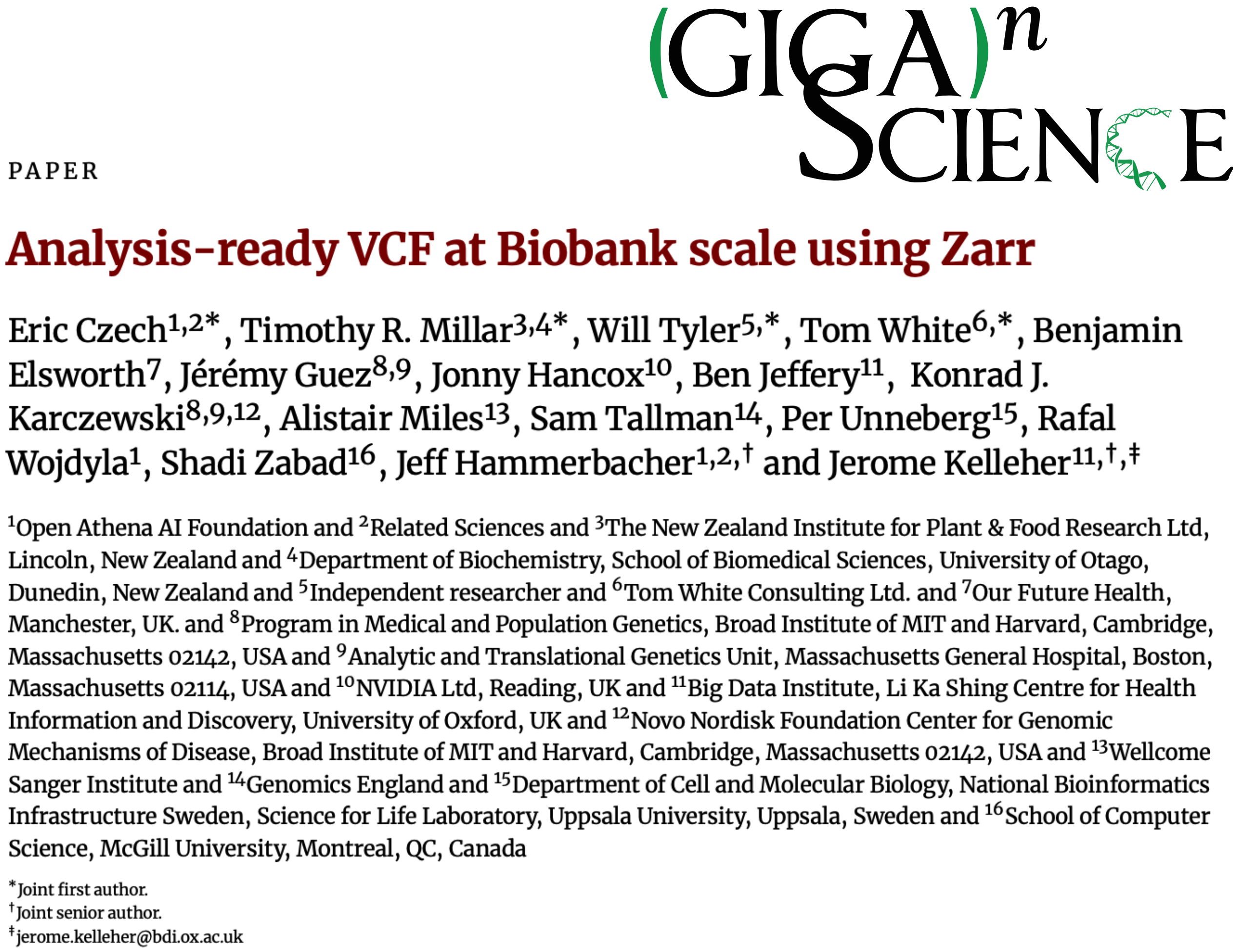
VCF vs Zarr



VCFs are monolithic two dimensional objects that can only be searched efficiently by row

Zarr chunks the data by sample and more:

- Natively supports cloud object stores
- Single thread analytic performance is better than traditional methods
- Parallelisable



(GIGA)^n SCIENCE

PAPER

Analysis-ready VCF at Biobank scale using Zarr

Eric Czech[1,2*], Timothy R. Millar[3,4*], Will Tyler[5,*], Tom White[6,*], Benjamin Elsworth[7], Jérémy Guez[8,9], Jonny Hancox[10], Ben Jeffery[11], Konrad J. Karczewski[8,9,12], Alistair Miles[13], Sam Tallman[14], Per Unneberg[15], Rafal Wojdyla[1], Shadi Zabad[16], Jeff Hammerbacher[1,2,†] and Jerome Kelleher[11,†,‡]

[1]Open Athena AI Foundation and [2]Related Sciences and [3]The New Zealand Institute for Plant & Food Research Ltd, Lincoln, New Zealand and [4]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand and [5]Independent researcher and [6]Tom White Consulting Ltd. and [7]Our Future Health, Manchester, UK. and [8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA and [9]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA and [10]NVIDIA Ltd, Reading, UK and [11]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK and [12]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA and [13]Wellcome Sanger Institute and [14]Genomics England and [15]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden and [16]School of Computer Science, McGill University, Montreal, QC, Canada

[*]Joint first author.
[†]Joint senior author.
[‡]jerome.kelleher@bdi.ox.ac.uk

Collab with groups around the world:

- Genomics England
- Broad / MIT
- Big Data Institute
- Novo Nordisk
- NVIDIA

+
Our
Future
Health

+

Thanks.
Any questions?