


Expanding horizons of science and knowledge

Highlights 2021/22



Lime hawk-moth:
Creating a reference genome for this species
provides the foundations to study chromosome
and species evolution.



We seek to understand life in all its forms to improve our world

What we do

- 4 Director's Introduction
- 6 2021 Timeline
- 8 At a Glance
- 9 Year in Numbers

Our work

- 12 COVID-19
- 16 Cancer, Ageing and Somatic Mutation
- 22 Cellular Genetics
- 28 Human Genetics
- 32 Parasites and Microbes
- 38 Tree of Life

Our approach

- 44 Scale
- 46 Innovation
- 48 Collaboration
- 50 Culture

Other information

- 54 Image Credits
- 55 Institute Information



If you would like to know more, visit:
Genome Notes
[https://www.darwintreeoflife.org/
genomes/genome-notes/](https://www.darwintreeoflife.org/genomes/genome-notes/)

What we do

Director's Introduction

Genomic research has come of age. DNA-based studies are profoundly changing the delivery and application of healthcare worldwide; from delivering near real-time insights into the evolution and spread of a global pandemic to identifying the genetic seeds of adult cancer sown before birth.

So, it is fitting that the Wellcome Sanger Institute, and the Wellcome Genome Campus that supports it, are entering a new phase to build the partnerships, pipelines, and knowledge needed to deliver the benefits of genomic research to all.

The Sanger Institute is founded on the core principles of openness, transparency, collaboration, and capacity building. The power of genomics will only be fully realised through the rapid dissemination of data and discoveries throughout the global scientific community. The COVID-19 pandemic is a clear case in point.

But, rapidly releasing results and openly sharing data is not enough. Genomics will only truly deliver on its potential when all researchers, clinicians, and policymakers can easily access, understand, and apply the information.

We are committed to embedding genomic research, surveillance, and infrastructure around the world. We fund and coordinate the work of the Genomic Alliance for Global Health (GA4GH) to standardise genetic and health data sharing. In partnership with Connecting Science and the University of Cambridge, we nurture the next generation of genomic researchers and clinicians to deliver new insights and healthcare approaches. Through intercontinental collaborations we help governments and researchers build DNA sequencing and analysis pipelines that inform health policy.

Our work requires bold ambition and painstaking attention at scale. We create the partnerships, tools, and approaches needed to understand genetic variation within individuals and communities, and across countries and continents, to tailor healthcare and lifestyle interventions. In partnership with scientists across the globe, we explore the interplay of genomics and the environment in different regions to understand inherent differences in disease prevalence and resistance to infection.

We are pioneering the delivery of reference genomes for all species living in Britain and Ireland to power future research into food security, biodiversity, climate change mitigation, and biomaterials. Our genome assembly pipelines are building a reference base of the world's species that will inform future conservation efforts.

At the other end of the biological spectrum, we are mapping the development, communication, and movement of individual cells over a person's lifetime. This work intimately explores how genetic variation affects people over the course of their life; from the underpinnings of cancer to the reasons why people's immune response to COVID-19 changes with age.

As an Institute, we seek to ensure that everyone enjoys the economic and health benefits that this next wave of genomic research will bring. We are proud of the diversity of knowledge and insights our staff bring from their different lived experiences. But we are aware that these differences may constitute barriers to equal access. To ensure that we draw on the widest spectrum of skills for the benefit of all, we are working to improve our support and inclusion of previously underrepresented communities in our teams.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute



Our work requires bold ambition and painstaking attention at scale. We create the partnerships, tools, and approaches needed to understand genetic variation within individuals and communities, and across countries and continents, to tailor healthcare and lifestyle interventions.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute



2021 Timeline

Sanger Institute joined the Wellcome Leap Global Network

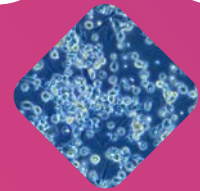


JAN

- Spin-out company Kymab acquired by Sanofi
- Developmental origins of eczema and psoriasis revealed



FEB



- Largest open access resource of malaria genomes released
- Lab-grown mini bile ducts repaired human livers in genetic first



MAR

- Search engine for single cell atlases developed
- COVID-19: Entry factors more prevalent in elderly, men, and smokers



1,400+ viruses identified in human gut; half are new to science



COVID-19: Genomics revealed transmission routes in care homes



APR

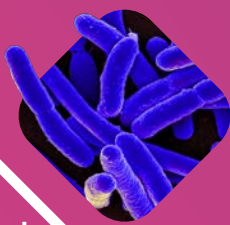
- NanoSeq technology enabled genetic study of any tissue
- Grand finale of inaugural Wellcome Genome StartUp School

Vertebrates Genomes Project released 16 high-quality reference genomes



MAY

- Rise of multi-drug resistant *E. coli* in Norway tracked
- Neurodiversity events highlighted network support for colleagues



Endangered water vole genome produced



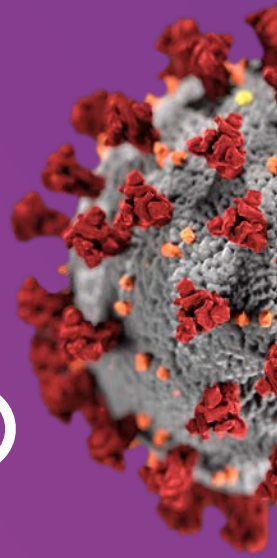
Genomic study provided insights into how bodies develop from a single cell



- New diagnoses for childhood developmental disorders
- Cellular signatures of kidney tumours discovered



JUN



Sanger shortlisted for its outstanding contribution to the pandemic response

Higher mutation rates alone are not to blame for age-related disease

UK Biobank releases health and genomic data of 200,000 participants



How bacteria travel between guts revealed

More than 1 million COVID-19 virus genomes read

COVID-19: Omicron identified and tracked

SpotMalaria genetic surveillance platform tracked drug resistance spread

Reverse mentoring scheme launched to drive cultural change

Organoids provide a sophisticated model of endometrial function



JUL

AUG

SEP

OCT

NOV

DEC



Report shows power of UK genomics
Janet Thornton Fellowship 2021 for career-break postdocs opens

Postdoc Appreciation Week celebrated our postdocs

Genomics combined with mobile data guided Bangladesh's COVID-19 response

Sanger sequencing operations awarded the Papin Prize for COVID-19 work

GA4GH data passport released



Genomics showed how humans adapted to climate change in the Middle East

Mutations linking liver disease with diabetes and obesity discovered



Sanger Excellence Fellowship launched



At a Glance

287

gene-edited cell lines

300,000

constructs cloned
for pooled
CRISPR libraries

324

cell lines differentiated
into neural
progenitor cells

87

organoid models
banked

6,200

flow cytometry
experiments

Cellular

**DNA
sequencing**

Compute

An average of
40,000bn
DNA bases a
day were read

48,000
approx.
total number of
compute cores

Every
3.2 mins
the equivalent
of one gold-standard
(30x) human genome
was read

30,000
approx.
high-performance
compute cores

1,551
species were
sequenced

18,000
approx.
cloud-based
compute cores

Every week,
the equivalent of
3,100
gold-standard
human genomes
were read

84Pb
approx.
of usable
storage

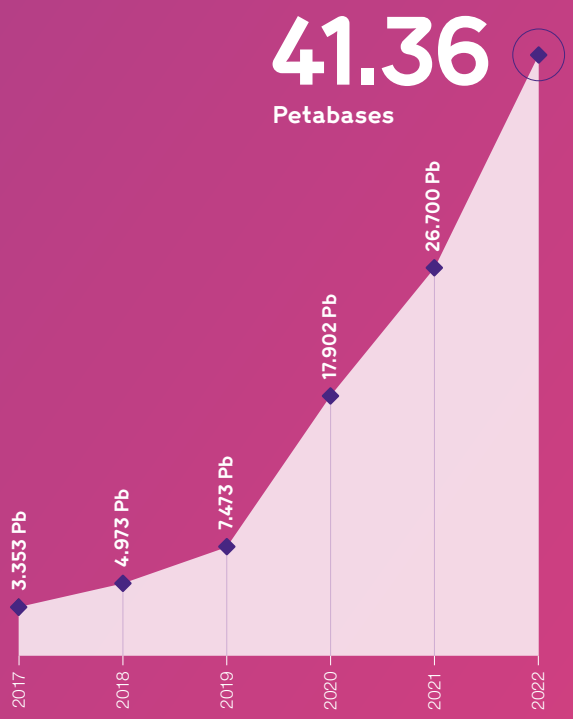
Year in Numbers

559
research articles
in 2021

1,155,653
citations of our
papers and
reviews
to date

9,010
published articles
and reviews
to date

Sanger Institute
authored papers
and reviews are
3.84
times more cited than
the world average*



*(data for last five years 2017-2021). (Field-weighted citation impact (FWCI) metric).



Our work



Wellcome Sanger Institute Data Centre:
Constant innovation by our IT teams, in
partnership with our software developers,
unlock the power of computing to study
biology's intimate secrets at scale.

We conduct ambitious, world-leading science at a scale few can match



COVID-19

We are at the forefront of genomic surveillance: from sequencing SARS-CoV-2 virus genomes at scale to identifying transmission patterns of the infection. Our data and insights have helped to guide the responses of governments to the COVID-19 pandemic.

Page **12**

Cancer, Ageing and Somatic Mutation

We study the genetic changes in normal tissues to better understand their causes and consequences on ageing and disease. We conduct large-scale cellular experiments to discover how mutations affect cancer development.

Page **16**

Cellular Genetics

We map cells in the human body at scale by combining single-cell genomic profiling, 3D imaging, and computational methods. We investigate the dynamic changes that occur within cells, tissues, organs, and organisms during development, health, disease and ageing.

Page **22**

Human Genetics

We combine population-scale genetics and cell-based studies with clinical data to identify and study severe developmental disorders. We study the biology of health and disease in the immune system and blood cells through large-scale cell-based experiments.

Page **28**

Parasites and Microbes

We study the genomics and evolution of disease-causing organisms and the human microbiome. We build networks at scale to help monitor infectious diseases and the effects of health policies worldwide, identifying the drivers of drug, vaccine, and insecticide resistance to guide health planning.

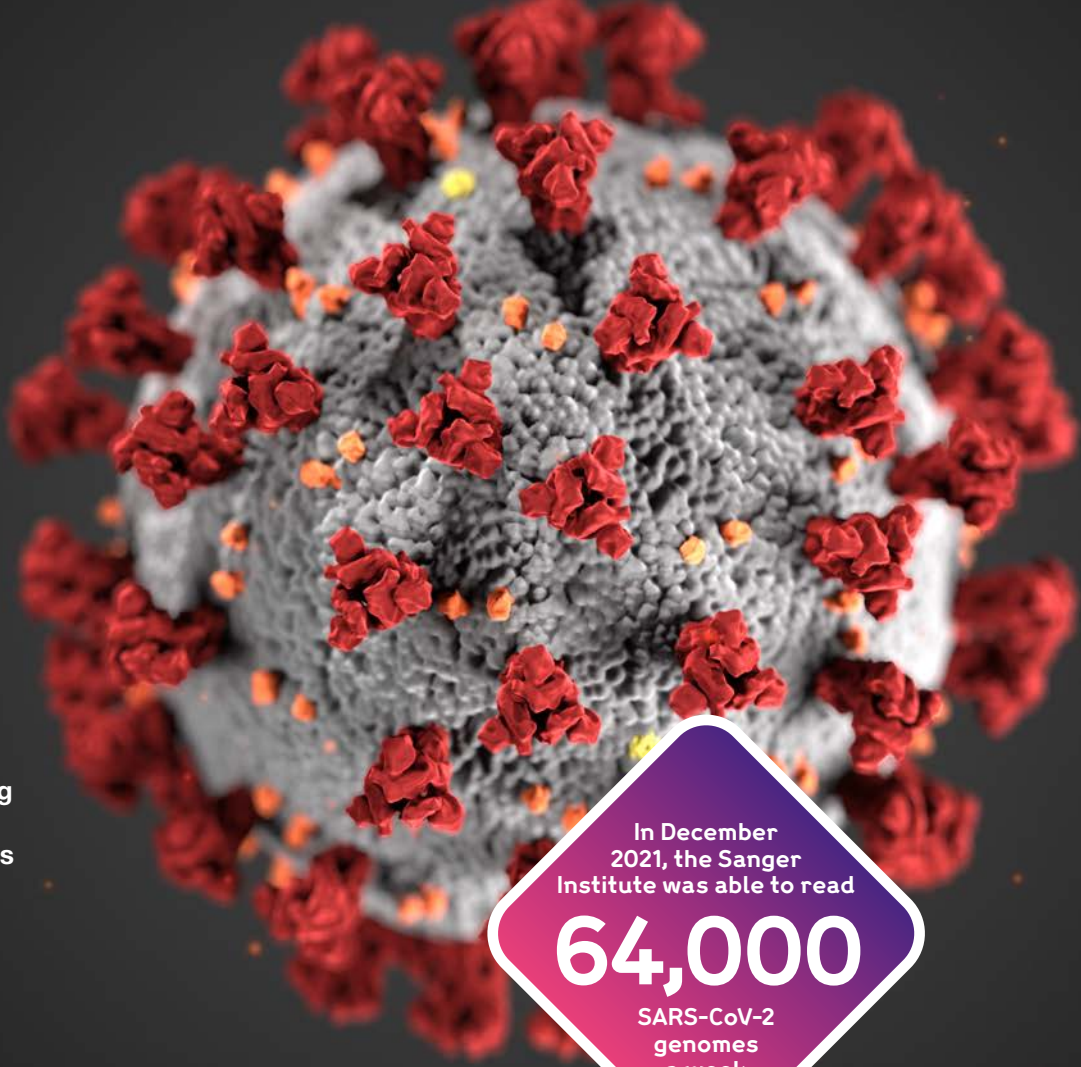
Page **32**

Tree of Life

We are building the library of life. We produce high-quality reference genomes to explore the evolution, function, and interactions of life on Earth. We seek to aid conservation and biodiversity work, and provide the underpinnings of a new way of doing biology.

Page **38**

COVID-19



We are at the forefront of genomic surveillance: from sequencing SARS-CoV-2 virus genomes at scale to identifying transmission patterns of the infection. Our data and insights have helped to guide the responses of governments to the COVID-19 pandemic.

In December 2021, the Sanger Institute was able to read

64,000

SARS-CoV-2 genomes a week

In this section

- 1 Genomic surveillance of COVID-19
- 2 COVID-19 sequencing in Bangladesh informs policies
- 3 New therapeutic targets for COVID-19 found
- 4 How the virus changes within a person
- 5 Why immune responses differ in asymptomatic versus severe COVID-19
- 6 The inside story of the COVID-19 pandemic in England



1 Genomic surveillance of COVID-19

The Sanger Institute has contributed approximately 20 per cent of the world's publicly available SARS-CoV-2 genome sequences. The data are used for identifying and tracking viral variants, tracing COVID transmission in the UK, and aiding public health responses. Freely available around the world, the sequence data are aiding understanding of the virus, its evolution, and its biology.

Over 2021, Sanger Institute teams continued to sequence the genomes of coronavirus samples from the UK's Lighthouse Laboratories, which process postal, drive-through, and walk-in PCR tests from the community.

Capacity has been dramatically increased through major laboratory refurbishments and expansions: from 10,000 samples per week in January, to 64,000 per week by the end of the year. More than 25 million samples have been handled so far.

Alongside increased capacity, efficiencies were improved and turnaround times reduced to enable the rapid detection

of new variants. It takes just a few days from receiving a sample to uploading genome data to public databases for analysis.

Data managers have implemented direct connections with partner organisations, so information can be swiftly and securely transferred. Data are immediately shared with UK Public Health Authorities to aid their responses.

In summer 2021, UKHSA contracted the Sanger Institute to continue sequencing coronavirus genomes, and further increase capacity. The Institute now receives hundreds of thousands of samples a week. UKHSA and Sanger staff have developed an algorithm to select the samples to sequence, ensuring geographic coverage and that priority samples are chosen.

Genomic surveillance remains a key part of the pandemic response. To identify and track the emergence of new variants around the world and to understand how the virus functions, causes disease, and evolves in response to vaccines.



To view the lineages and variants detected in the UK, see <https://covid19.sanger.ac.uk/lineages/raw>

2

COVID-19 sequencing in Bangladesh informs policies

Researchers at the Wellcome Sanger Institute, the University of Bath, and Bangladesh-based Institutes, including the Institute of Epidemiology, Disease Control and Research (IEDCR), icddr,b (formerly the International Centre for Diarrhoeal Disease Research), and the Institute for Developing Science and Health Initiatives, worked together to understand the first wave of SARS-CoV-2 in Bangladesh. The study is the first of its kind and shows the unique advantage of combining mobility and genomic data to help untangle an infectious disease outbreak.

Mutations naturally occur in the SARS-CoV-2 genome as the virus replicates. Sequencing its genetic code allows researchers to create phylogenetic trees that show viral lineages and variants, and to determine how individual samples of the virus are related.

The international team sequenced and analysed 391 SARS-CoV-2 genomes, taken from testing facilities in Bangladesh between March and July 2020, and Bangladeshi samples from GISAID (Global Initiative on Sharing All Influenza Data). An additional 68,000 global genomes were used in the analysis to build the phylogenetic trees.

The researchers found multiple introductions of the virus to Bangladesh, and shifting patterns of lineages across the country. However, during this period most of the virus samples were of three variants – B.1.1, B.1.1.25, and B.1.36.

To investigate the factors that led to country-wide spread of these variants, the consortium examined anonymised population mobility data, collected from Facebook and three mobile phone operators. These data showed a mass migration from Dhaka to all areas of the country at the end of March 2020, ahead of a stay-at-home order coming into effect. These mobility data are consistent with the transmission of SARS-CoV-2 lineages out of Dhaka to the rest of the country during the first wave.

Their analysis led the researchers to work directly with the Bangladeshi Government, and the consortium went on to sequence additional SARS-CoV-2 samples collected

between November 2020 and April 2021. These samples included variants of concern, Alpha B.1.1.7 and Beta B.1.351. The Government of Bangladesh implemented interventions that would limit the spread of the virus through the same channels as observed during the first wave, such as restrictions on inter-city travel. The insights from genomic sequencing data continue to inform national policies.



Reference

Cowley L, et al. *Nature Microbiology* 2021; **6**: 1271–1278.



This study is a great example of what can be achieved when scientists work collectively with public health professionals towards a collective goal.

Professor Nicholas Thomson

Senior author and Head of the Parasites and Microbes Programme and Group Leader at the Wellcome Sanger Institute

3

New therapeutic targets for COVID-19 found

Researchers at the Sanger Institute as part of Open Targets have provided a prioritised list of human host proteins potentially exploited by the SARS-CoV-2 virus. These represent new therapeutically tractable targets for treating COVID-19 disease.

People infected with SARS-CoV-2 experience a range of symptoms – from a mild cough to critical illness. Demographic factors, such as age, account for much of this variation. However, molecular factors, such as genetic mutations, can also have an impact by affecting the amount of proteins in the body.

To identify human proteins that may contribute to the risk of developing severe COVID-19, the team systematically analysed publicly available protein and COVID-19 datasets. This included comparing the genetic and protein profiles of over 30,000 COVID-19 patients and one million healthy individuals. They combined several computational methods to assess the data, and leveraged resources provided by the Open Targets Genetics portal. This

enabled them to identify genetic variations and genetically predicted plasma proteins associated with COVID-19 susceptibility and severity.

Their analysis supported several previously identified proteins as having a role in COVID-19. They also identified new potential drug targets.

In total, nine proteins were found to have an impact on COVID-19 infection and disease severity. Four were ranked as top priorities for potential treatment targets. One, called CD209 or DC-SIGN, is involved in how the virus enters human cells. Two proteins, IL-6R and FAS, could be responsible for the immune overactivation often seen in severe COVID-19. IL-6R has already been targeted in recent clinical trials. The fourth protein, OAS1, which has been reported previously, appeared to reduce susceptibility to COVID-19.

The team also undertook laboratory experiments using cell lines, and showed that CD209 directly interacts with the spike protein of SARS-CoV-2. Knowing more about the human proteins that influence COVID-19 severity opens up research into new treatments.



Open Targets



Reference

Anisul M, et al. *eLife* 2021; **10**: e69719.



Open Targets Genetics portal:

<https://genetics.opentargets.org/>



4 How the virus changes within a person

Detailed analysis of SARS-CoV-2 genomes by researchers at the Sanger Institute has aided understanding of how the virus evolves within an individual.

Genomic analysis has yielded important insights into the origins and transmission of the SARS-CoV-2 virus. While each sample of coronavirus that is sequenced may contain thousands of virus particles, a consensus genome sequence is derived for use in genomic epidemiology. However, deep sequencing data invariably reveal within-host variations of the virus.

It has been suggested that SARS-CoV-2, like related coronaviruses, evolves within an infected host as a quasispecies, with many genomic mutations (within-host variants) arising that may be beneficial for the virus.

To better understand the evolution of SARS-CoV-2, the Sanger team performed Illumina deep sequencing of more than a thousand samples. Each sample was sequenced twice, with separate laboratory preparation steps, to evaluate the quality and reproducibility of the results.

Over 95 per cent of samples showed detectable within-host mutations. Analyses revealed patterns suggesting that mutations were caused by damage to the virus or RNA editing, rather than replication errors.

Within- and between-host diversity of the virus showed strong selection patterns, particularly against mutations that are deleterious to the virus. Additional analysis suggested that SARS-CoV-2 has mutational hotspots within its genome.

The team also analysed mutation frequencies and found that most people were infected by a single lineage, though they identified several putative examples of co-infection with more than one viral variant. These factors make using within-host variations challenging for epidemiological purposes.

The work has aided understanding of how the virus evolves within its human host. Subsequent studies are exploring the implications of within-host evolution in immunocompromised patients, which is a plausible explanation for the emergence of some variants of concern.



Reference
Tonkin-Hill G, et al. *eLife* 2021; **10**: e66857.

5 Why immune responses differ in asymptomatic versus severe COVID-19

In the largest study of its type in the UK, scientists found that people with asymptomatic COVID-19 had raised levels of specific immune cells, whereas people with more serious symptoms had lost these protective cell types and gained inflammatory cells. The research, conducted within the Human Cell Atlas initiative, could help explain serious lung inflammation and blood clotting symptoms.

Symptoms of COVID-19 vary widely; from a mild cough to severe respiratory distress, blood clots and organ failure. Several studies have highlighted a complex

immune response to COVID-19, but the full coordinated immune response and how this differs between symptomatic and asymptomatic patients had not previously been investigated in detail.

To understand how different immune cells responded to the infection, a large team of researchers analysed blood from 130 people with COVID-19. The team performed single-cell sequencing on approximately 800,000 individual immune cells and a detailed analysis of cell surface proteins and antigen receptors.

In those with no symptoms, the researchers found increased levels of B cells that produce antibodies in mucus passages, such as the nose. However, these protective B cells were missing in people with serious symptoms, indicating the importance of an effective antibody-associated immune response in mucus passages.

The team discovered patients with mild to moderate symptoms had high levels of B cells and helper T-cells, which help fight

infection. Those with serious symptoms had lost many of these immune cells. In people who required hospitalisation, there was an uncontrolled increase in specific types of immune cells, including monocytes and killer T-cells – high levels of which can lead to lung inflammation. Those with severe disease also had raised levels of platelet-producing cells, which help blood to clot.

The study highlights the coordinated immune response that contributes to COVID-19 pathogenesis, and reveals discrete cellular components that can be targeted for therapy. All of the data, methods, and analysis are freely accessible.



Reference
Stephenson E, et al. *Nature Medicine* 2021; **27**: 904–916.

6

The inside story of the COVID-19 pandemic in England

A detailed analysis of SARS-CoV-2 genomic surveillance data shows COVID-19 in England as a series of overlapping epidemics. The study has helped researchers understand more about how a new infectious agent spreads and evolves.

In March 2020, the COVID-19 Genomics UK (COG-UK) consortium was established, with the aim of monitoring the spread and evolution of SARS-CoV-2 by sequencing the virus' genome. The Sanger Institute was the sequencing hub of COG-UK, and is now reading more than 60,000 coronavirus genomes per week for the UK Health Security Agency – contributing around one fifth of the world's publicly available SARS-CoV-2 genome sequences.

Researchers at the Sanger Institute and EMBL-EBI analysed SARS-CoV-2 genomic surveillance data from England, collected between September 2020 and June 2021.

They characterised the growth rates and geographic spread of 71 viral lineages, in 315 English local authorities, and they reconstructed how newly emerging variants changed the course of the epidemic.

This analysis revealed a series of sub-epidemics that peaked in early autumn 2020. At the end of 2020, the Alpha (B.1.1.7) variant emerged and spread despite a series of restrictions. Alpha was found to possess a 50 to 60 per cent growth advantage over previous variants. A third, more stringent national lockdown in the UK suppressed Alpha and eliminated nearly all other lineages in early 2021.

Also during early 2021, variants associated with a greater ability to circumvent immunity from vaccination or prior infection continued to appear in the UK. These were characterised by a mutation in the spike protein, E484K. Despite repeated introductions, these variants – including Beta (B.1.351) and Gamma (P.1) – were confined to short-lived local outbreaks.

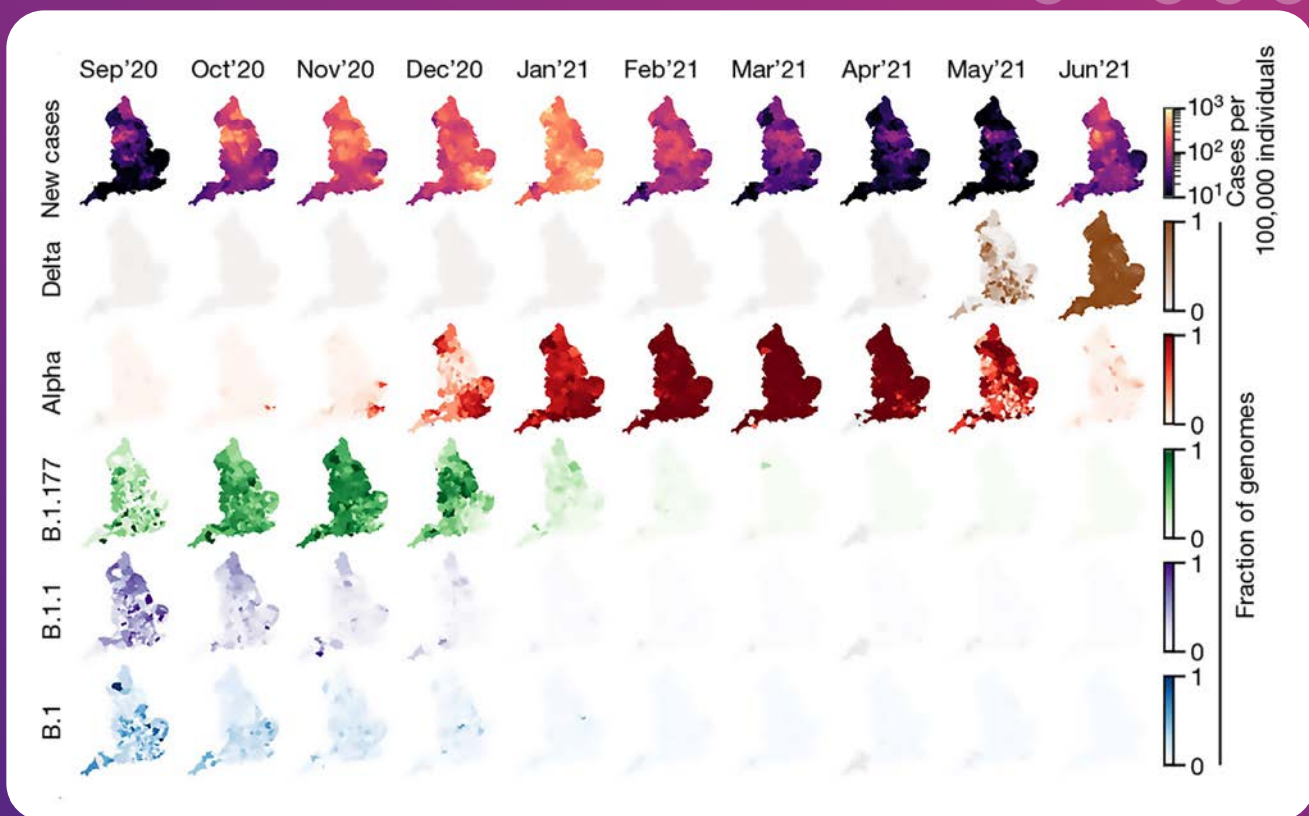
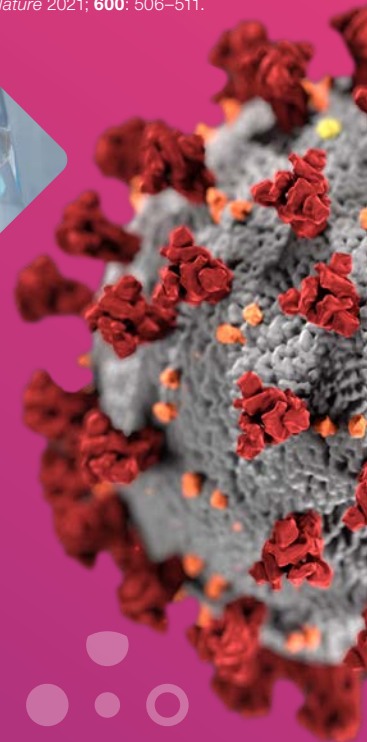
In March 2021, the first samples of Delta (B.1.617.2), which is thought to have originated in India, appeared in sequence data in England. Though Delta contained different mutations to previous variants of concern, it was even more transmissible. Delta had spread to all local authorities and accounted for 98 per cent of viral genomes sequenced by 26 June 2021.

The Sanger Institute continues to deliver large-scale genomic surveillance of SARS-CoV-2 in near real-time, and the data are passed to the UK's public health authorities to inform the pandemic response.



Reference

Vöhringer HS, *et al. Nature* 2021; **600**: 506–511.



Cancer, Ageing and Somatic Mutation

We study the genetic changes in normal tissues to better understand their causes and consequences on ageing and disease. We conduct large-scale cellular experiments to discover how mutations affect cancer development.

The cellular history of a **78 year old** has been retraced all the way back to the first cell division

In this section

- 1 Seeing a lifetime of mutations
- 2 Common threads in cat, dog, and human cancers
- 3 New method to accurately study genetic mutations in any tissue
- 4 New treatment hope for blood cancer patients
- 5 Competition between mutant cells drive out early tumours
- 6 New drug target for resistant bowel cancer
- 7 Global insights into oesophageal cancer
- 8 How the human prostate develops
- 9 How our bodies develop from a single cell
- 10 Understanding 'patchwork' tumours to inform treatments

1

Seeing a lifetime of mutations

New knowledge of how the human body develops from one cell into trillions has been generated by scientists at the Sanger Institute, the University of Cambridge, and their collaborators. The studies are the first to analyse somatic mutations in normal tissues across multiple organs within and between individuals.

The DNA in cells of the human body is continuously damaged, and, although it is mostly repaired, cells steadily acquire genetic mutations throughout life as a result. The presence of the same mutation in different cells often indicates a shared developmental history, and comparing these somatic mutations in individual cells can be used to trace their origins.

Using laser capture microdissection followed by low-input whole-genome sequencing – an approach previously developed at the Institute – the researchers found and analysed somatic mutations across multiple donors and tissues. Reconstructing large-scale phylogenies revealed significant variation between individuals. For example, the two progenitor cells created by the fertilised egg dividing

contributed equally to the body of one individual, but in another donor 93 per cent of their cells were descended from just one original progenitor.

To identify mutation rates and processes across the body, the researchers analysed hundreds of samples from 29 tissue types. Bioinformatic analysis compared the patterns of mutation and showed hallmarks of ubiquitous mutational processes across all of the tissues studied. Other processes were specific to certain tissues, with substantial variability in the mutational landscape between tissues in an individual – giving clues to the causes of DNA damage.

The teams also confirmed much lower mutation rates in immature sperm cells, which are thought to be shielded from mutational processes, and showed that this is likely to be an intrinsic feature of the male germline.

The studies will help to establish baselines of normal development, and how we acquire mutations throughout life. Understanding healthy development and ageing will help to better comprehend the onset of disease.



References
Moore L, et al. *Nature* 2021; **597**: 381–386. and
Coorens T, et al. *Nature* 2021; **597**: 387–392.

2

Common threads in cat, dog, and human cancers

Researchers have revealed genetic similarities between canine and feline hemangiosarcoma and human angiosarcoma. The work could open up avenues for the application of human therapies to companion animals, and shows that cancer clinical trials in pets may have applications for human health.

Angiosarcoma (AS) is a rare, highly aggressive tumour of blood and lymphatic vessels in humans. AS arises from the endothelial cells lining vessel walls and may occur at any site in the body. Known risk factors for AS include radiotherapy, UV light exposure, and chemical exposure.

Angiosarcoma tumours are highly varied, meaning studying their genetics has been challenging. There are significant unmet clinical needs for AS, and studies into new therapies are hampered by a lack of good preclinical models.

Hemangiosarcoma (HSA) is a spontaneous cancer seen in cats and dogs, with similar features to AS. Like AS, it arises from endothelial cells, occurs at varying sites, and it is associated with a poor prognosis.

To investigate the genetic suitability of HSA as a model for AS, the team sequenced 1,000 cancer genes in 41 cases of canine and feline HSA and matched germline tissue.

Deep sequencing of a targeted gene panel revealed recurrently mutated driver genes in canine HSA. These included *TP53*, *PIK3CA*, and *LRP1B* – consistent with previous reports of HSA and AS. The team also found mutated genes *ATRX*, *FAT1*, *GRIN2A*, and *RELN* in canine HSA, which had only previously been seen in human AS. Sequencing of feline cutaneous HSA for the first time also revealed parallels to both canine HSA and human AS.

Similar to AS, a pattern of DNA damage caused by UV was found in a subset of canine HSA's, and both species show differing mutational profiles between tissue sites.

The characterisation of canine and feline HSA shows important parallels to AS, and provides hope that future studies on these cancers will benefit all three species.



Reference

Wong K, et al. *Disease Models and Mechanisms* 2021; **14**: dmm049044.



3

New method to accurately study genetic mutations in any tissue

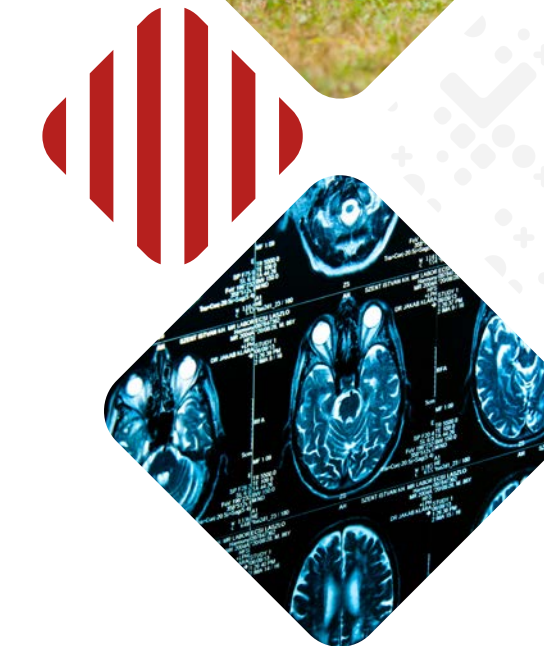
Sanger Institute scientists have developed a new method to accurately sequence DNA. The technique, called nanorate sequencing (NanoSeq), enables the study of DNA changes in any human tissue or cell population. This represents a major advance for research into cancer and ageing, and has important applications for the study of human development and biology.

Somatic mutations that occur as we age drive the development of cancer and may contribute to other diseases. The study of somatic mutations, and their causes, had previously been limited to actively dividing cells and tissues. To overcome this constraint, the team created NanoSeq.

NanoSeq was developed by refining duplex sequencing – an advanced technique with an error rate of about one in every million base pairs of DNA. Analysis of these errors suggested they were caused by the processes used to prepare DNA for sequencing.

The team implemented improvements to laboratory processes and developed bioinformatics methods in order to refine NanoSeq. Over the course of four years, accuracy was improved until they achieved fewer than five errors per billion letters of DNA. This rate enables the detection of somatic mutations in any cell population.

The researchers then used NanoSeq to study somatic mutations in non-dividing cell populations from across the body – something not previously possible. Their analysis of blood cells found a similar number of mutations in slowly dividing stem cells and more rapidly dividing progenitor cells. Analysis of non-dividing muscle cells and neurons revealed that mutations accumulate throughout life in these cells



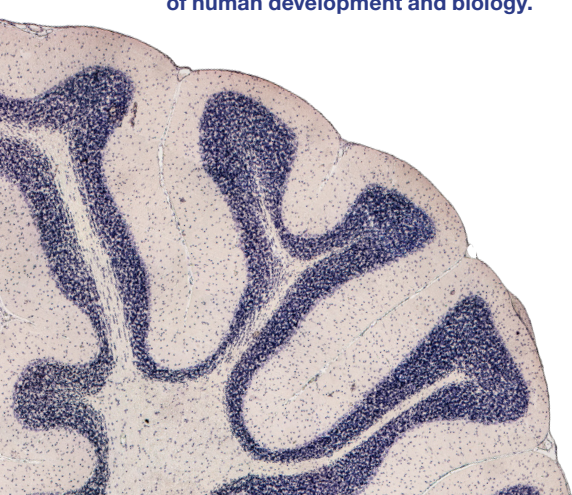
too, at a similar pace. Their work suggests cell division is not the dominant process causing mutations in many body tissues, as had previously been thought.

The ability to reliably detect mutations in single DNA molecules could transform our understanding of somatic mutagenesis, ageing, and cancer. The full bioinformatics pipeline to process and analyse NanoSeq data is freely available to download.



Reference

Abascal F, et al. *Nature* 2021; **593**: 405–410.





4

New treatment hope for blood cancer patients

Sanger Institute researchers have identified a vulnerability in some cases of acute myeloid leukaemia that could be harnessed for targeted treatment of these poor-prognosis cancers.

Acute myeloid leukaemia (AML) is an aggressive blood cancer that affects people of all ages, often requiring months of intensive chemotherapy. It typically develops in cells within the bone marrow, which then crowd out healthy cells, leading to life-threatening infections and bleeding. AML treatments have remained unchanged for decades, and fewer than one in three people survive the cancer.

Through large-scale DNA sequencing analysis, Sanger Institute researchers had previously found that loss-of-function mutations in the *CUX1* gene were seen in several types of cancer, including AML. These mutations are associated with a poor disease prognosis, however the role of *CUX1* in AML development was unclear.

In a new study, the team undertook genome-wide CRISPR/Cas9 screening in human and rodent cells. They showed that a lack of functioning *CUX1* leads to an expansion of certain types of blood stem cells, which are defective in apoptosis – a type of regulated cell death. They found that the loss of *CUX1* causes increased activation of the *CFLAR* gene – which encodes a protein that restrains apoptosis – potentially providing a means for mutated cancer cells to evade cell death and propagate.

The study enables a greater understanding of how the loss-of-function mutation in the *CUX1* gene leads to the development and survival of AML. The researchers have shown that targeting *CFLAR*, or apoptosis evasion pathways that allow cancer cells to continue growing, could lead to new therapies for patients living with aggressive AML.



Reference
Supper E, et al. *Nature Communications* 2021; **12**, 2482.

5

Competition between mutant cells drive out early tumours

Researchers at the Sanger Institute and the University of Cambridge have shown that mutant clones in oesophageal tissue create a highly competitive environment in which early, microscopic tumours struggle to grow. Understanding the mechanisms that prevent the expansion of newly formed tumours will bring insights into the development of cancer.

All cells in the human body accumulate genetic mutations over an individual's lifetime. Previous pioneering research at the Sanger Institute showed that human epithelial tissues contain a patchwork of cell populations with distinct sets of mutations. These 'mutant clones' compete for space to survive. Sequencing the genomes of mutant clones has revealed that most contain DNA mutations that are associated with cancer – yet tumour formation is relatively rare.

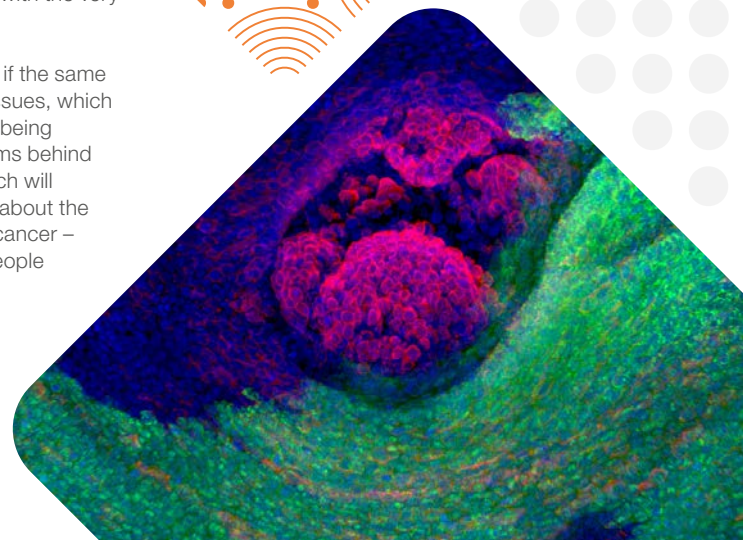
To understand what is preventing the growth of early tumours, researchers used state of the art 3D imaging techniques in mice to visualise microscopic tumours at an earlier stage of development than has previously been possible. The genomes of micro tumour cells, together with the genomes of nearby healthy cells, were sequenced to identify mutations.

They found that the survival of early tumours in mice does not depend solely on the mutations they carry, but also on the mutations within the neighbouring normal tissues. The team identified mutations that can have a tumour suppressive role, independent of the body's immune system. These findings help to explain the relatively low rate of cancers compared with the very high number of mutant cells.

Further studies will look to see if the same interactions occur in human tissues, which mutations lead to cancer cells being successful, and the mechanisms behind this. Current and future research will provide further understanding about the development of oesophageal cancer – a disease that affects 9,200 people each year in the UK.



References
Colom B, et al. *Nature* 2021; **598**: 510–514.
and
Martincorena I, et al. *Science* 2018; **362**: 911-917.





6

New drug target for resistant bowel cancer

Sanger Institute researchers have found that targeting a specific cancer survival gene in colorectal cancer could lead to new treatment options for advanced disease.

Colorectal cancer, or bowel cancer, is the fourth most common cancer in the UK, with around 42,300 people diagnosed each year.

Roughly 10 to 15 per cent of colorectal cancer cases display mismatch repair – deficient (dMMR)/microsatellite instability – where cellular DNA has become unstable. Targeted therapies, chemotherapy, and immunotherapy are used to treat colorectal cancers with dMMR/microsatellite instability, though about half of patients' cancers become resistant to treatments.

As part of the Cancer Dependency Map – a major initiative to provide a detailed rulebook of precision cancer treatments – previous research by Sanger scientists identified the WRN gene as a possible treatment target in cancers with microsatellite instability. WRN encodes for a protein that has important but

poorly understood roles in maintaining genome stability, DNA repair, replication, and telomere maintenance. The team found that WRN is essential for certain cancer cells to survive.

This new study used 60 unique colorectal cancer cell models – laboratory-grown cells derived from patient tumours. It is the largest collection of dMMR/microsatellite instable colorectal cancer cells studied to date. For the first time, they used CRISPR technology to show that treatment-resistant cancer cells still require WRN for survival.

Their work has reinforced WRN as a target for drug development in colorectal cancer – particularly for patients with some of the deadliest forms of disease. It may be important in other cancers with dMMR/microsatellite instability too. If drugs that inhibit WRN can be developed, it would offer new therapy options for people whose cancer has become resistant to existing treatments.



Reference

Picco G, *et al. Cancer Discovery* 2021; **11**: 1923-1937.

7

Global insights into oesophageal cancer

New findings into one of the world's most common cancers have been generated by the Cancer Research UK Grand Challenges Mutographs team, an international collaboration led by the Sanger Institute.

Oesophageal squamous cell carcinoma (ESCC) is the world's eighth most common cancer and the most common type of oesophageal cancer – a disease that affected more than 600,000 people worldwide in 2020. Incidence of the disease varies dramatically around the world.

The Cancer Grand Challenges Mutographs team sought to understand different ESCC incidence rates by studying mutational signatures. These patterns of DNA changes, each with a distinct cause, were first detected and analysed by Sanger Institute scientists in 2013. So far, more than 100 mutational signatures have been identified in human cancers.

The study brought together international experts in cancer epidemiology and mutational signature analysis who studied

552 ESCC genomes from eight countries: Iran, China, Kenya, Tanzania, and Malawi with high incidence rates; Brazil, Japan and the UK with lower incidence rates.

Analyses showed that no mutational signature exists to explain the difference in ESCC incidence globally – the overall mutational profile of the cancer was extremely consistent. However, the team did identify factors that increase the risk of the disease. For example, ESCC was linked to alcohol consumption in Japan, and to opium usage in north Iran.

The researchers also identified a mutational signature present in over 90 per cent of samples, linked to a fault in the APOBEC molecule. This suggests that APOBEC activation is a crucial step in ESCC tumour development.

In future studies, the team plans to sequence and identify mutational signatures in healthy tissues – something that has recently become possible thanks to technological advances.



Reference

Moody S, *et al. Nature Genetics* 2021; **53**: 1553-1563.

Oesophageal cancer has been linked with opium use



8

How the human prostate develops

Leveraging unbiased clock-like mutations, Sanger Institute researchers have defined prostate stem cell dynamics through foetal development, puberty, and ageing. The findings demonstrate that insights into development and maintenance of solid tissue organs in humans can be obtained using spontaneously occurring somatic mutations.

Recent advances in genome sequencing technologies mean it is now possible to detect the somatic mutations that cells accumulate throughout life. Previous research at the Sanger Institute has shown that adult tissues are a patchwork of different cell populations, or clones, and that somatic mutations can be used as a molecular clock to trace cellular history.

To understand mutational patterns and cell dynamics in the prostate, the researchers sequenced whole genomes from 409 microdissections of normal prostate from eight donors.



They used phylogenetic analysis to create ‘family trees’ of cells and mapped their location in the prostate to reconstruct tissue dynamics. The team found that somatic mutations accumulate steadily, at 16 mutations per clone per year. The highest mutation rates occurred in peripheral regions of the prostate, where cancer incidence is highest.

Further analysis showed that glandular subunit structures within the prostate are established during foetal development, by five to ten embryonic cells each. The team showed that structural changes to the prostate during puberty originate from local stem cells, disseminated during early development.

The researchers found only one ‘driver’ mutation – a mutation associated with cancer – in ageing prostate tissue. This mutation drove a clonal expansion larger than any other and was similar to those seen earlier in life, consistent with the theory that prostate cancer is caused by cells reverting to a developmental state.

The work uses observations of somatic mutations to define prostate stem cell dynamics through foetal development, puberty, and ageing – something that is otherwise only achievable in genetically manipulated experimental animal models.



Reference
Grossmann S, *et al. Cell Stem Cell* 2021; **28**: 1262-1274.

9

How our bodies develop from a single cell

Researchers at the Wellcome Sanger Institute, the Wellcome-MRC Cambridge Stem Cell Institute, and the University of Cambridge provide an essential reference of developmental dynamics

Study of human development has been limited primarily to careful microscopy, and a great amount of knowledge is based on model organisms such as mice. In a new approach, blood progenitor cells from human foetal tissue were collected from the Human Developmental Biology Resource. These haematopoietic stem and progenitor cells were grown into 511 single cell-derived colonies. The researchers then sequenced the genomes of the cells to identify somatic mutations.

Somatic mutations can be effectively used as barcodes in lineage-tracing studies, as they are inherited by the progeny of a cell as permanent marks. By analysing the patterns of mutations that have accumulated in cells over time, it is possible to identify relationships between cells and reconstruct the phylogeny of a cell population – in this case tracing a cell’s lineage right back to the first division of the embryo.

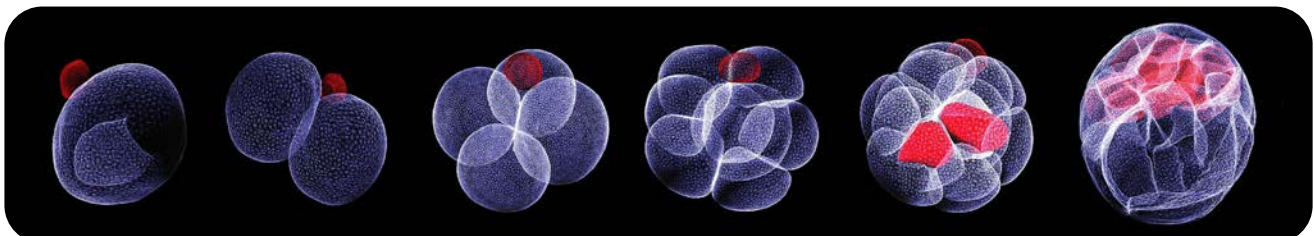
The team found that individual haematopoietic progenitors acquire tens of somatic mutations by 18 weeks after conception. They also timed the ‘decision’ for when cells would become either foetal or extra-embryonic tissue, for example the placenta, occurred between four and 16 cells – much earlier than previously estimated.

There was also evidence that the extra-embryonic mesoderm and the blood cells that deliver oxygen to the foetus in the first trimester of pregnancy arise from a cell layer termed the hypoblast. This is a clear, but previously unknown, difference between human and mouse biology.

These insights provide an essential reference of developmental dynamics for those studying childhood cancers, which often begin *in utero*, as well as rare developmental disorders.



Reference
Spencer Chapman M, *et al. Nature* 2021; **595**: 85–90.



10

Understanding ‘patchwork’ tumours to inform treatments

New research has confirmed that nearly all tumours contain patches of cells with different genetic mutations, and these mutations follow specific patterns in different cancer types. The findings have important implications for treating cancer.

Intra-tumour heterogeneity is a mechanism of treatment resistance in many cancers and so represents an important clinical challenge. Researchers from the international Pan-Cancer Analysis of Whole Genomes Consortium, including Sanger Institute scientists, have characterised the extent, origin, and drivers of genetic mutations within tumours, something that has previously been poorly understood.

The researchers analysed the whole genomes of 2,658 cancer samples from 38 types of cancer, to study intra-tumour heterogeneity. They found that 95 per cent of samples contained at least one subclone, a patch of cells with distinct genetic diversity. This variation is a challenge for doctors, as a treatment that works for one subclone may not be effective against another. In addition, certain subclones can initiate tumour spread or drug resistance.

The researchers’ findings confirm that tumour evolution is likely to be driven by changes that benefit the cancer. This might be the ability to resist a particular treatment or evade the immune system.

The team also showed that the levels and types of genetic changes varied between cancer types, following specific patterns of driver gene mutations, fusions, structural variants, and copy number alterations. Even in the same tumour, the genetic makeup of different subclones varied widely. Their analyses provide detailed insights into tumour evolutionary dynamics.

Understanding this genetic diversity in tumours can be harnessed to predict prognosis, which could help doctors and patients make important treatment decisions. Understanding subclones is also important for clinical trial design.

The findings, data, and methods are openly available, providing a comprehensive resource for studying cancer genomics.



Reference

Dentro S, *et al.* *Cell* 2021; **184**: 2239-2254.

2,658
cancer samples

38
cancer types



Cellular Genetics

We map cells in the human body at scale by combining single-cell genomic profiling, 3D imaging, and computational methods. We investigate the dynamic changes that occur within cells, tissues, organs, and organisms during development, health, disease and ageing.

Studying
115,993
fetal bone marrow cells revealed blood and immune cell development in Down's syndrome

In this section

- 1 Mapping how immune and blood cells form
- 2 Rapid immune response protects children from COVID-19
- 3 Cell signatures of kidney tumours found
- 4 Researchers identify trigger for 'head-to-tail' axis development in human embryo
- 5 Understanding how human microglia vary
- 6 New tool enables unprecedented cellular maps
- 7 Genetic study identifies child's risk of secondary leukaemia
- 8 Origins of Inflammatory Bowel Disease revealed
- 9 Automating the study of millions of cells in action
- 10 Uterus research boosts study of diseases that affect one third of women

1

Mapping how immune and blood cells form

The first comprehensive analysis of how the blood and immune systems develop in prenatal bone marrow has been conducted by scientists at the Sanger Institute and their collaborators within the Human Cell Atlas (HCA) initiative. The researchers found that in the space of a few weeks, numerous blood and immune cell types emerge from developing bone marrow.

A previous HCA study described how the human blood and immune systems begin to develop in the yolk sac – a structure external to the embryo – and foetal liver, a process known as haematopoiesis. But until now, it was unknown how haematopoiesis continued in bone marrow, which produces blood and immune cells throughout life.

The researchers used single-cell RNA technology to analyse developing bone marrow tissue samples, identify the cell types present, and which genes those cells expressed. They assessed over 100,000 individual cells, using multiple methods to study gene activity and protein expression.

The team observed the rapid diversification of cells into the full blood and immune cell repertoire. This diversification occurred over six to seven weeks, early in the second trimester of pregnancy. Compared to foetal liver, there were a large number of B-lymphoid cell types, which are needed to help combat infection. They also found that the bone marrow is the key site of neutrophil emergence – cells that protect against bacteria.

The researchers also studied bone marrow from people with Down syndrome. They identified genome-wide differences in gene expression that may help to shed light on why individuals with Down syndrome are more prone to developing immune disorders and leukaemia.

The study will be an important reference for understanding how the blood and immune systems develop in bone marrow, and how this can go wrong in disorders such as leukaemia, with important implications for diagnoses and treatments.



Reference
Jardine L, *et al. Nature* 2021; **598**: 327–331.

2

Rapid immune response protects children from COVID-19

As part of the Human Cell Atlas initiative, research from the Sanger Institute and University College London identified fundamental differences in the immune response of adults and children, helping to explain why children are much less likely to become seriously ill from SARS-CoV-2.

The innate immune system of children is generally better able to recognise dangerous viruses or bacteria automatically, compared to adults. Adults have a more adaptive immune system containing a huge repertoire of memory cell types, which have been trained through past exposure to respond to a particular threat.

To examine specific differences in response to SARS-CoV-2 infection in children and adults, the researchers collected airway and blood samples from 19 paediatric and 18 adult COVID-19 patients with a range of symptoms, and 41 healthy children and adults.

Sanger scientists undertook single-cell sequencing of 659,217 cells from the samples to assess their states and functioning. Analysis revealed 59 different cell types in the airways and 34 cell types in the blood, including some never previously described.

The team identified several immune system responses and mechanisms that help explain why children are generally protected from severe COVID-19. For example, a stronger innate immune response was found in the airways of children, characterised by the rapid deployment of interferon proteins – which trigger the immune system and help to restrict viral replication. In adults, a less rapid immune response meant the virus was better able to invade other parts of the body where the infection was harder to control.

These insights could contribute to pinpointing the triggers of severe disease in adults, with a view towards risk stratification. It may be possible that higher risk patients could then be considered for pre-emptive treatments.



Reference

Yoshida M, et al. *Nature* 2021; **602**: 321-327.

3

Cell signatures of kidney tumours found

The origins of seven types of kidney cancer, including several rare subtypes, have been identified by Sanger researchers and collaborators. The findings confirm that these cancers derive from specific developmental cells present in the maturing foetus.

Cancer cells may retain patterns of messenger RNA (mRNA) that are characteristic of the cell they arose from. By comparing mRNA patterns of gene expression between cancer and normal cells, it is possible to learn about aspects of a tumour's origin, its differentiation state, or trajectory.

Recent studies have used this approach to identify the origins of some childhood cancers, such as neuroblastoma – a cancer that starts in early nerve cells. The method has relied on single-cell mRNA sequencing to determine an individual cell's activation state and type.

Because this resource intensive technique is not always feasible for rare cancers, the team took a new approach. They used

computational techniques to mine existing datasets, analysing single-cell transcriptome data from Human Cell Atlas datasets that represented healthy cells. They also analysed bulk mRNA data, generated from multiple cells within a tumour, from the International Cancer Genome Consortium and The Cancer Genome Atlas databases. These open, global projects were founded by researchers at the Sanger Institute.

The researchers assessed mRNA signals in 1,300 childhood and adult renal tumours, spanning seven different tumour types. The results confirm, with quantitative evidence, that the entire spectrum of paediatric renal tumours arise from developmental cells.

In contrast, they showed that adult kidney cancers emerge from mature cell types and do not revert to a developmental pattern of gene expression in the majority of cases.

Each cancer type was also found to exhibit unique patterns of gene expression. These patterns could be used to classify tumours – meaning the method holds promise in diagnosing rare cancers.



Reference

Young M, et al. *Nature Communications* 2021; **12**: 3896.



Not only does this computational approach using existing datasets validate previous results... it provides a new way of expanding this research to much larger numbers of tumours... I believe that this approach could act as a blueprint for investigating the behaviour and origins of the entire spectrum of human cancer.

Dr Sam Behjati

A senior author of the study from the Wellcome Sanger Institute

4

Researchers identify trigger for 'head-to-tail' axis development in human embryo

Key molecular events in the developing human embryo between days seven and 14 have been uncovered by scientists, providing greater understanding of critical stages of our development.

The second week of gestation represents a significant stage of embryo development, or embryogenesis. Failure of development during this time is one of the major causes of early pregnancy loss. Understanding more about it will help scientists to uncover how it can go wrong, and take steps towards being able to fix any problems.

The pre-implantation period, before the developing embryo implants into the mother's womb, has been studied extensively in human embryos in the laboratory. On the seventh day, the embryo must implant into the womb to survive and develop. Very little is known about the development of the human embryo once it implants, because it becomes inaccessible for study.

Scientists from the University of Cambridge and the Sanger Institute used a recently developed technique, where embryos are cultured *in vitro* both before and after implantation, but pre-gestation, allowing them to be studied up to day 14 of development.

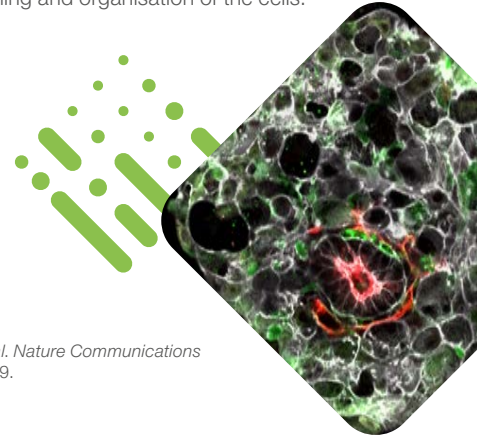
By using single-cell sequencing methodologies and functional experiments, they captured the evolving molecular profile of the developing embryo after implantation. The team revealed the progressive loss of pluripotency, or the ability of embryonic stem cells to differentiate into any cell type, as the fates of various cells are determined.

Their findings provide the first evidence that a group of hypoblast cells outside the embryo sends a message to the embryo that initiates the development of the head-to-tail body axis.

The researchers also revealed that molecular signals involved in the formation of the body axis show similarities to those in animals, despite significant differences in the positioning and organisation of the cells.



Reference
Mole MA, et al. *Nature Communications* 2021; **12**: 3679.



5

Understanding how human microglia vary

Systematic analysis of gene activation in microglia from hundreds of individuals has created a reference of their diversity. The research, by Sanger Institute scientists and their collaborators, gives new insights about cells that are critically important in the central nervous system.

Microglia, the tissue-resident immune cells of the central nervous system, play vital roles in nerve cell functioning and maintaining immune surveillance in the brain. Microglial dysfunction is implicated in a number of neurological disorders, including Alzheimer's.

Single-cell transcriptomic studies, which involve sequencing the RNA of cells to determine gene activation, have been used to assess microglial cells – though only in limited numbers. This has shown that microglial function varies, but conclusions are often not replicated.

To understand how age, sex, pathology, brain anatomy, and common genetic variation influence microglia transcriptomes, the team performed a population-scale study. They profiled gene activity in more than 9,500 microglia from 141 patients undergoing neurosurgery. Comparing the RNA data, they identified four populations of cells with distinct patterns of gene activation. Two of the populations had not been seen previously.

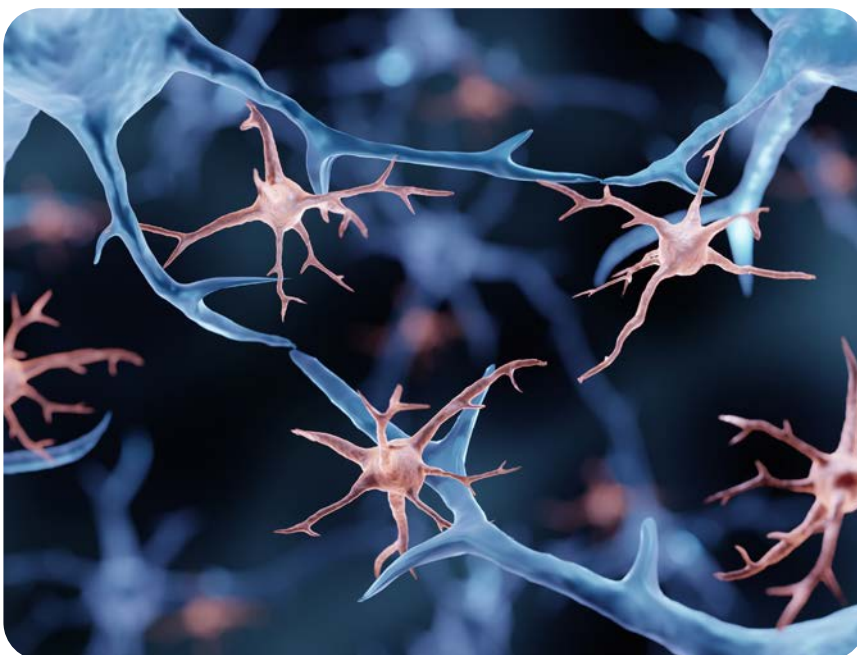
Of the factors they examined, clinical pathology explained more transcriptomic variation than all the other factors combined. All factors except sex – including age, brain region and dominant hemisphere – explained some of the variation.

The team then used expression quantitative trait loci (eQTL) mapping to identify regions of the genome controlling the gene activation levels. This revealed a number of candidate risk genes for several diseases, which had functions in microglia. They followed up on one of the findings using induced pluripotent stem cell models. This functional assessment allowed them to locate a candidate causal genetic variant for Alzheimer's disease near the well-known *BIN1* gene.

The study provides a systematic exploration of microglia diversity and defines a reference dataset of microglial gene activation.



Reference
Young A, et al. *Nature Genetics* 2021; **53**: 861–868.

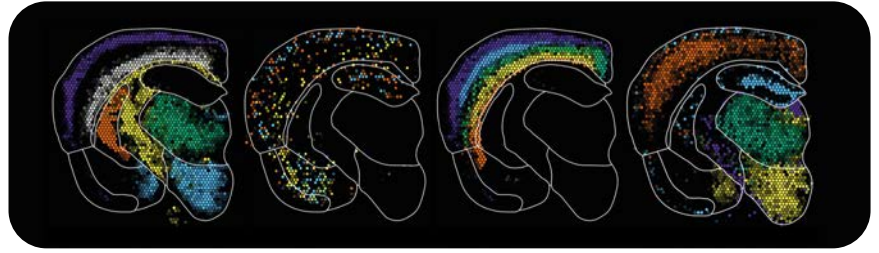


6

New tool enables unprecedented cellular maps

The study of the human body at the single-cell level has received a boost with the creation of a new tool, which will allow researchers to assess not only the function of cells, but also where they are situated within tissues. The tool, called cell2location, has been developed by researchers at the Sanger Institute, the German Cancer Research Centre, and their collaborators.

The human body is made of a myriad of cells. Previously unknown cell types are regularly being discovered by initiatives such as the Human Cell Atlas (HCA), which aims to map every cell type in the human body. This field of research uses single-cell sequencing to analyse the genes that individual cells express, and distinguish the subtle differences in the functions that define them.



Previously, it had not been possible to combine single cell sequencing data with spatial information at scale. This has meant that detailed information on rare cell types, as well as information on the relationships and interactions between cells within tissues, has been missing.

To solve this problem, the researchers created a Bayesian model to combine different data types. They tested the method on three different tissues, and demonstrated that cell2location is a versatile analysis tool for comprehensively mapping tissue architectures.

In the mouse brain, cell2location was able to detect subtle differences in gene activity between cells. This allowed the team to identify rare astrocyte subtypes, not

previously described. Cell2location was also able to map the subtypes, including one that accounted for just 41 cells out of 40,000, to a specific location within the tissue. In the human lymph node and the gut, the researchers spatially mapped and resolved immune cell populations.

Cell2location is already being used as part of the HCA, and the tool and the code behind it are freely available. The richness of cell2location data means it has the potential to one day replace microscope analysis as a technique to analyse biopsies.



Reference

Kleshchevnikov V, et al. *Nature Biotechnology* 2022.

7

Genetic study identifies child's risk of secondary leukaemia

Scientists from the Sanger Institute and the University of Cambridge found that for some children with neuroblastoma – a cancer of immature nerve cells – essential treatment with platinum chemotherapy can sometimes cause changes to the genome that could lead to secondary leukaemia.

Secondary blood cancer is a challenging complication of childhood neuroblastoma treatment. Every year, around 100 children in the UK are diagnosed with neuroblastoma – a highly aggressive disease. The cancer often requires intensive treatment, including several chemotherapy drugs, which can result in side effects such as damage to the DNA of healthy cells. In up to 7 per cent of childhood neuroblastoma survivors, damaged bone marrow cells go on to develop into secondary leukaemia.

In this study, researchers sequenced the whole genomes of bone marrow and blood samples of two children who developed blood cancer following high-risk neuroblastoma treatment. They analysed the genomes to determine the specific DNA mutations causing the leukaemia.

The team found that in both patients the leukaemia had mutations that were caused by neuroblastoma chemotherapy. A wider analysis of 17 children treated for a variety of cancers then identified another child who had undergone neuroblastoma treatment and had developed some of the genetic hallmarks of pre-leukaemia.

In the future, it could be possible to identify the children who have a higher risk of developing secondary leukaemia by sequencing their genome and highlighting any genetic drivers that could be pre-cursors for blood cancer.



Reference

Coorens T, et al. *Blood* 2021; **137**: 2992–2997.



7

per cent of childhood neuroblastoma survivors develop secondary leukaemia

8

Origins of Inflammatory Bowel Disease revealed

A large-scale study has mapped the cells in the human gut from early development through to adulthood. Researchers revealed that Crohn's disease may be caused by activation of developmental pathways. The research helps explain how the gut forms and functions, and it has uncovered potential drug targets for treating inflammatory bowel diseases.

The gut is a complicated tissue made of multiple cell types, and changes enormously during early development. To understand how the gut develops and functions, researchers from the Wellcome Sanger Institute, Newcastle University, University of Cambridge, and their collaborators within the Human Cell Atlas studied almost half a million individual gut cells from developing tissue, and from child and adult donors.

Using cutting edge single-cell genomics and spatial analysis techniques, the team revealed which genes were active in each cell and created a highly extensive Gut Cell Atlas, mapping cells through time and across 12 regions of the intestines.

Their analysis uncovered an immune sensing role for tuft cells – chemosensory cells in the epithelial lining of the intestines – with a subset of tuft cells expressing an antibody receptor. The researchers also described nerve cell populations in the developing enteric nervous system, and predict cell-type-specific expression of genes associated with Hirschsprung's disease.

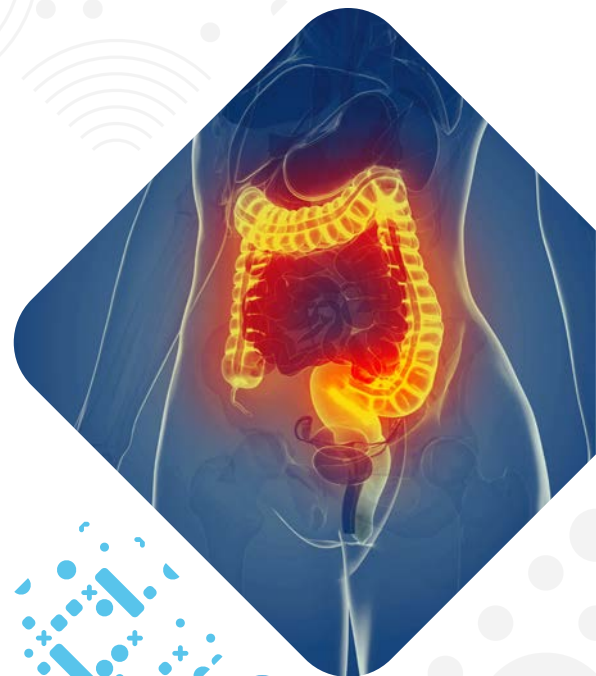
Using a systems approach, the scientists also identified three key cells that attract immune cells to form secondary lymphoid tissue in early human development. They found that this same developmental pathway may cause Crohn's Disease, where immune cells are recruited and retained at the site of inflammation.

The Gut Cell Atlas is already shedding new light on the origins of Crohn's, and the researchers hope that their data – openly available to all – will contribute to future discoveries and the development of new treatments for gut diseases.



Reference

Elmentaite R, *et al.* (2021) *Nature* 2021; **597**: 250–255.



Almost
500,000
cells gut cells
were studied

9

Automating the study of millions of cells in action

Sanger Institute scientists have developed new methods to automate and scale-up single-cell RNA sequencing, providing an unprecedented opportunity for exploring the transcriptomes of millions of individual cells.

Single-cell RNA sequencing (scRNA-seq), first developed in 2009, detects and quantifies messenger RNA molecules in an individual cell – the transcriptome. Used to study cellular states, regulation, responses, and activity, this powerful technique has enabled deeper understanding of how genomes, cells, and organisms function. More recently, methods to sequence RNA in its entirety, rather than in shorter fragments that need to be pieced back together computationally, have become popular.

Existing protocols for such full-length single-cell RNA sequencing produce highly complex datasets containing thousands of distinct genes. Producing full-length transcript data allows researchers to answer questions about different transcripts arising from a single gene, mutations in the DNA code, and the variation seen in immune cell receptors, that cannot be assessed with other methods.

Full-length protocols are also suited to profiling rare cell types, such as those seen during development. However, these methods, whilst they have high sensitivity and specificity, are not cost-effective or scalable.



To overcome these issues, the team optimised two robotic protocols for full-length scRNA-seq. The first protocol is a flexible, customisable, miniaturised, in-house automated protocol with off-the-shelf reagents. It uses small, benchtop robotic equipment that occupies minimal space. The second automated protocol uses commercially available kit and is suitable for larger-scale projects.

The team worked to optimise the methods, resulting in substantially reduced costs and hands-on time, as well as increased reproducibility and workflow efficiency. The methods deliver a throughput of thousands of single cells per day.

The methods have already been used for producing datasets in human lung, placenta, and colon studies aligned to the Human Cell Atlas.



Reference

Mamanova L, *et al.* *Nature Protocols* 2021; **16**: 2886–2915.

10

Uterus research boosts study of diseases that affect one third of women

The most comprehensive cell atlas to date of the human uterus has identified two new epithelial cell states that can be used to distinguish two forms of uterine cancer. Researchers from the Sanger Institute, the University of Cambridge, and their collaborators also identified the genetic pathways that determine two main endometrial cell types.

One in three women will suffer from some form of reproductive disease during their lifetime, including chronic conditions such as endometriosis, and potentially life-threatening uterine cancers. The endometrium – the mucosal lining of the uterus – is challenging to study, partly because it undergoes dynamic, cyclical changes of shedding, regeneration and differentiation.

To better understand the endometrium, the team analysed uterine samples using single-cell and spatial transcriptomics. These advanced sequencing techniques allow researchers to assess which areas of the genome are active in a cell, as well as its location within a tissue – and so determine cell types and functions.

They identified two new cell states, *SOX9+LGR5+* and *SOX9+LGR5-*, associated with uterine cancer. Enrichment of the *SOX9+LGR5+* cell population was associated with later stage cancers that pose a greater threat to the patient, as compared to tumours with higher levels of *SOX9+LGR5-* cells. This finding could help guide treatments.

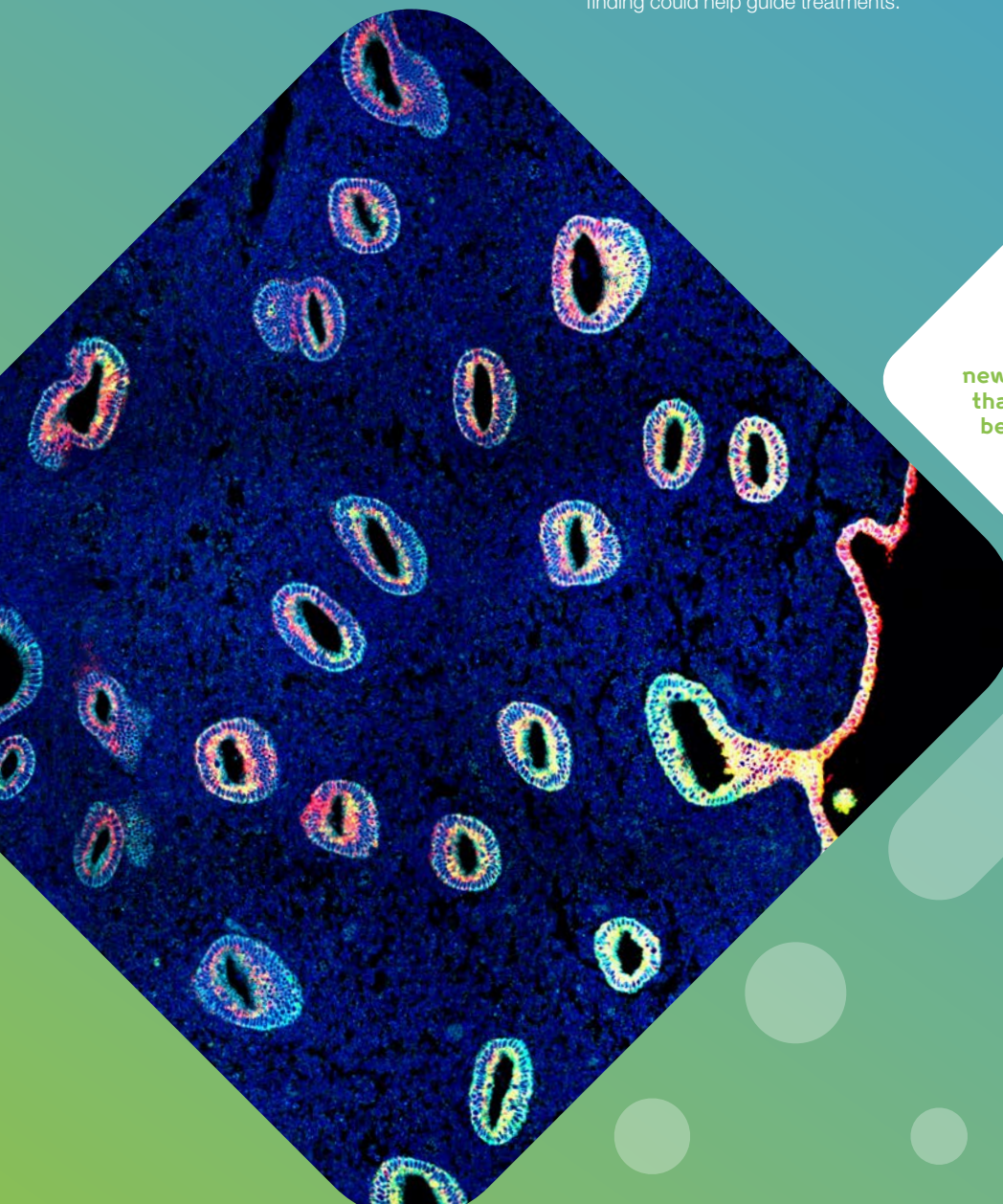
The researchers also used spatial transcriptomic data to identify cell signal pathways – NOTCH and WNT – that differentiate between secretory and ciliated cell types in the endometrium. University of Cambridge researchers then developed endometrial organoids which responded to the ovarian hormones oestrogen and progesterone, mimicking how endometrial tissues function in the body. These organoids provide a sophisticated model to study endometrial function and a blueprint for discovering how tissues are formed in other organs.

Part of the Human Cell Atlas initiative, the study is the first to combine single cell and spatial transcriptomics data of the uterus, providing detailed descriptions of cell types and their location. The data are openly available to scientists worldwide.



Reference

Garcia-Alonso L, *et al. Nature Genetics* 2021; **53**: 1698–1711.



2

new cell states found that can distinguish between 2 uterine cancer forms



Human Genetics

We combine population-scale genetics and cell-based studies with clinical data to identify and study severe developmental disorders. We study the biology of health and disease in the immune system and blood cells through large-scale cell-based experiments.

Analysing
9,858
genomes from DDD study
volunteers has revealed
new diagnoses

In this section

- 1 New diagnosis for childhood disorders
- 2 Genetic history of British Pakistanis mapped for the first time
- 3 Amino acid may predict common diseases risks
- 4 Genetic variants identified that impact immune cells' functioning
- 5 Rescue mutations could shed light on the origins of genetic disorders
- 6 How humans adapted to agriculture and climate change



1 New diagnoses for childhood disorders

Research from the Sanger Institute suggests that non-coding regions of DNA could hold the key to diagnosing developmental disorders in children.

Globally, around 400,000 babies are born every year with new, spontaneous DNA changes – *de novo* mutations – that cause developmental disorders.

De novo mutations in genes that create proteins are a well-established cause of developmental disorders. However, many of the genes linked to these disorders, and the role of non protein-coding DNA, remain unclear.

Most patients with developmental disorders have genetic testing as part of their clinical care, but this only leads to a diagnosis in under half of cases. Genetic testing normally identifies variants present in protein-coding regions of the genome.

Sanger researchers screened the genetic data of 9,858 participants in the Deciphering Developmental Disorders (DDD) study. They examined the untranslated regions (UTRs) of

the genome, adjacent to the protein-coding regions. UTRs regulate the amount, rate, and location of protein production.

The researchers identified six variants in UTRs that impact the gene, *MEF2C* – a gene known to be involved in developmental disorders. Laboratory studies confirmed that these variations cause disease through three distinct loss-of-function mechanisms. None of these variants were previously known to cause developmental disorders, and they account for 23 per cent of diagnoses involving *MEF2C* in the DDD cohort.

These findings enabled clinicians to provide a long-awaited diagnosis to multiple families. A diagnosis for a rare developmental condition can bring new understanding and support, as well as potential new treatment options, ending the 'diagnostic odyssey' that many families face.

This study highlights the importance of UTRs and suggests that they should be included in routine clinical screening for patients without a diagnosis.



Reference

Wright CF, et al. *American Journal of Human Genetics* 2021; **108**: 1083-1094.

2

Genetic history of British Pakistanis mapped for the first time

Analysis of genomic data will help to ensure those of Pakistani ancestry benefit from advances in healthcare.

Most genetic research has been conducted on individuals of European ancestry, but the findings may not translate well to individuals of other ethnicities. In recent years, studies of non-Europeans have started to provide a more complete picture of human genetic diversity and associated medical implications. To investigate the genetic characteristics of the British Pakistani population, researchers at the Wellcome Sanger Institute, the University of Leeds, and the University of Bradford conducted the first fine-scale analysis of genetic diversity within this group.

Researchers analysed genomes from over 4,000 Pakistani-ancestry individuals, together with self-reported information about their family history, to investigate the genetic characteristics of the British Pakistani population.

The team revealed how the genetic structure of the population has been shaped over centuries by the biraderi social system, a practice of marrying within clans.

Though Bradford Pakistani groups were found to be genetically similar to other Pakistani and Indian populations, the study found evidence that the biraderi social system has played an important role in shaping genetic variation. Participants shared the same genetic history until around 2,000 years ago, when they started separating into biraderi groups, including the Pathan, Jatt, Rajput, and Bains.

Genetic differences between biraderi groups were identified that have important implications for the design of medical studies seeking to discover the genetic roots of common illnesses, such as diabetes and heart disease. These data will help to ensure that those of Pakistani ancestry are represented in such research and will be able to benefit from new knowledge and therapies that emerge.



Reference

Arciero E, et al. *Nature Communications* 2021; **12**: 7189.



3

Amino acid may predict common diseases risks

Researchers have identified a link between mitochondrial DNA variants, amino acid fMet, and a range of common, late-onset diseases.

Mitochondria, the organelles that produce energy within cells, can influence the risk of late-onset human diseases, though the reasons for this are poorly understood. Damage or disruption to mitochondria are also thought to play a role in diabetes, heart disease, and depression.

To explore the link between mitochondria and disease, Sanger Institute researchers and their collaborators undertook one of the first population-scale studies of mitochondrial DNA (mtDNA).

Using a hypothesis-free approach, they analysed datasets to look for associations between mtDNA genetic variants and common molecular traits, such as blood cell counts. The first dataset included 5,689 traits in 16,220 healthy donors. There were significant associations between levels of an amino acid, N-formylmethionine (fMet), and mtDNA variants.

They then verified these associations using cellular models and found that fMet modulated cellular activities through several mechanisms. When fMet levels were measured in a cohort of ischaemic stroke patients, they were lower than those in a healthy control group.

The second dataset included data from 11,966 individuals, taken over a 20-year period. The team used this dataset to assess if differences in fMet between individuals were associated with a wider range of diseases. In contrast to ischaemic stroke, higher fMet levels were associated with increased risk of illnesses such as kidney disease and heart failure.

While further study of the molecular mechanisms at work is required, fMet seems to be a promising biomarker that could be used to better predict an individual's risk of developing a wide range of common diseases. The results also highlight the importance of research into mtDNA variants.



Reference

Cai N, et al. *Nature Medicine* 2021; **27**: 1564–1575.



This study has been a true collaborative effort with data from genome-wide association studies, cell lines, and computational analysis combining to offer rich insights into human biology.

Professor Nicole Soranzo

A senior author of the study and Associate Faculty at the Wellcome Sanger Institute

4

Genetic variants identified that impact immune cells' functioning

Scientists have found that certain genetic variants, which alter the binding ability of a protein called PU.1 in white blood cells, are also associated with susceptibility to autoimmune disease. Researchers from the Sanger Institute collaborated with the Josep Carreras Leukaemia Research Institute in Spain, and the MRC London Institute of Medical Sciences for the study.

Differences in disease risk can be influenced by a number of factors, with a substantial component driven by genetics. A previous study at the Sanger Institute called BLUEPRINT used large-scale genome-wide association studies (GWAS) to investigate which genetic differences are linked to changes in blood cells, and if these have a link to disease. BLUEPRINT revealed how variation in numbers of blood cells and their characteristics can affect a person's risk of developing complex diseases, such as heart disease and autoimmune diseases including rheumatoid arthritis, asthma, coeliac disease, and type 1 diabetes.

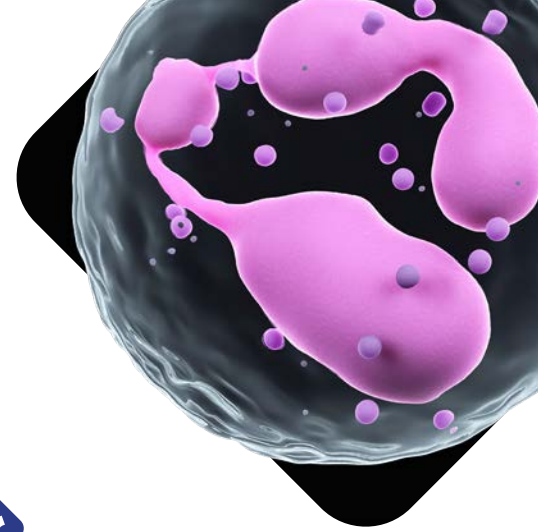
In new research, scientists combined GWAS data from BLUEPRINT with in-depth functional analysis of neutrophils – an abundant type of immune cell that forms the body's first line of defence against infection by pathogens.

The team found that the genetic variants associated with an increased risk of autoimmune disease also have an impact on the binding of the PU.1 'master regulator' protein in neutrophils. Some genetic variants made PU.1 unable to bind to neutrophil DNA, which led to subsequent downstream changes in gene expression and neutrophil behaviour.

While further research is needed to see if this change in the ability of PU.1 to bind directly causes certain autoimmune diseases, this study provides further understanding about the impact of these genetic variants on the cells in the body. In addition, the researchers suggest a list of candidate genes that could hold further information about the genetic causes of autoimmune disease.



Reference
Watt S, et al. *Nature Communications* 2021; **12**: 2298.



[Integrating] large-scale genetic research with functional analysis gives us essential data that widen our understanding of how differences in the human genome and epigenome interact to cause devastating common diseases. Building on this understanding through further research will help inform new avenues for treating these conditions.

Stephen Watt

Lead author and senior staff scientist at the Sanger Institute

5

Rescue mutations could shed light on the origins of genetic disorders

New insights into the ability of DNA to overcome harmful changes have been discovered by scientists at the Sanger Institute, the University of Lausanne, and their collaborators. In each instance they examined in detail a single 'rescue mutation' was responsible for cancelling out another mutation that would have threatened the organism's survival.

The consequence of a genetic mutation can be influenced by the context it is in. For example, a loss of gene function may be tolerated in some cells and lethal in others. The extent to which the effects of mutations are malleable, the structure of modifiers, and the genes involved remain largely unknown.

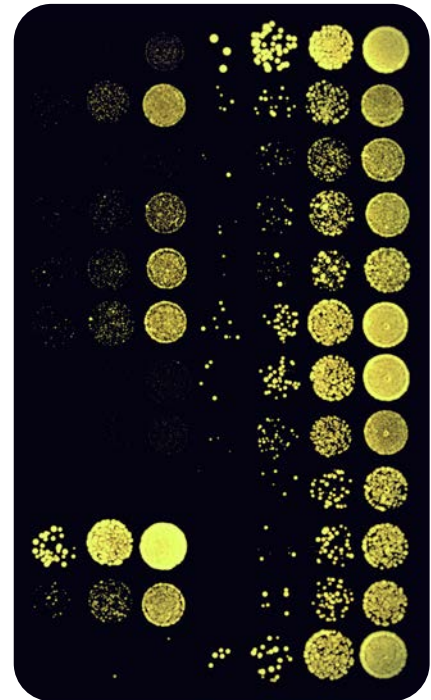
In this study, researchers introduced 1,106 temperature-sensitive mutant alleles from 580 essential genes into 10 genetically diverse wild yeast strains. They assessed the cells to see if natural genetic variation

in the strains would allow the yeast to grow when exposed to an unfavourably high temperature.

They found that fitness defects for 149 of the 580 tested genes – 26 per cent – could be suppressed by genetic variation in at least one yeast strain. Yeast colonies that continued to grow were sequenced at the Sanger Institute. The team used powerful genetic mapping approaches to identify the location of specific mutations that could be suppressing the temperature-sensitive alleles.

The researchers found that suppression was generally driven by gain-of-function of a single, strong modifier gene. These results demonstrate the natural genetic flexibility of cells to fulfil crucial tasks. It also provides important information about how certain DNA variants can suppress the undesirable effects of others.

Work is underway at the Sanger Institute to conduct a similar study in human cells. If the same biological phenomenon is at play, it could provide valuable information about how genetic diseases – including cancer and developmental disorders – arise. It is also possible that 'rescue mutations' might one day help clinicians to treat these conditions.



Reference
Parts L, et al. *Molecular Systems Biology* 2021; **17**: e10138.

6

How humans adapted to agriculture and climate change

Genetic signals indicating a population boom in the Levant that coincided with the transition to agriculture, plus a population crash in Arabia as the region dried up, have been uncovered by Sanger Institute researchers. Their work also provides important data for genomic and health studies in the Middle East – a region that has been understudied until now.

The fortunes of ancient human populations in the Middle East have been greatly influenced by its technological and climatic history. The Levant – a large area in the Eastern Mediterranean region of Western Asia – is regarded as the birthplace of agriculture, whereas Arabia is today dominated by the largest sand desert in the world.

In the first comprehensive population-scale study of Middle Eastern DNA, researchers at the Sanger Institute collected 137 samples from individuals representing eight Middle Eastern populations. The samples were sequenced using linked-read sequencing, a technique that enabled the team to reconstruct the population history of the region in unprecedented detail.



Our study helps to uncover the hidden genetic diversity in the Middle East, which has been largely understudied until now. As well as identifying variants that provide fascinating insights into the lives and adaptation of Middle Eastern ancestors, some of these variants are also important for healthcare in the region today. For example, we detected variants that were beneficial in the past, but today increase the risk of type 2 diabetes in some Arabian groups.

Dr Mohamed Almarri

Lead author of the study from the Wellcome Sanger Institute

The team's analysis revealed that human populations in the Levant experienced massive population growth in the last 15,000 years, which includes the time of the transition to agriculture. But populations in Arabia, who had transitioned to a herder-gatherer lifestyle, experienced a population crash during the aridification of the region. Around 4,000 years ago, Levantine populations also suffered a crash, as the region dried up.

The team identified 23.1 million single nucleotide variations (SNVs) in the genomes they sampled. Of these, 4.8 million SNVs were new variants not previously discovered in other populations. While many were rare, around 370,000 were common, and any of them could hold medical relevance.

These data will be an important resource for the study of genetic health and adaptations, such as type 2 diabetes and lactose tolerance, in Middle Eastern populations, which have been understudied until now. As human diversity and the susceptibility to diseases vary between different populations, this lack of detailed genetic data can exacerbate health inequalities.

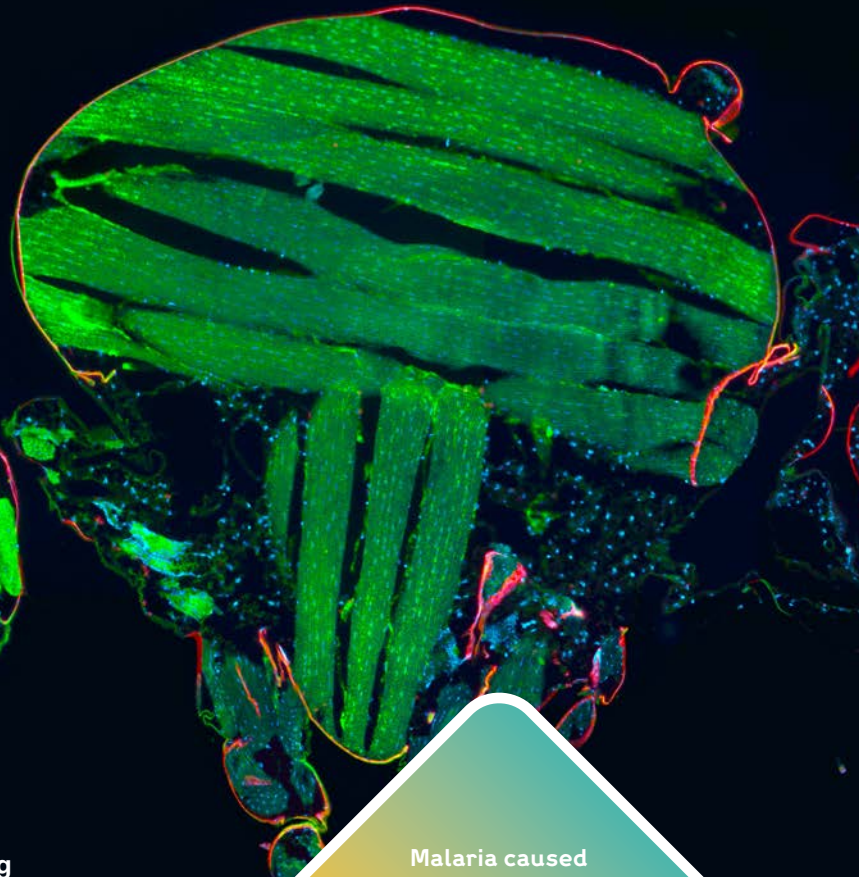


Reference

Almarri M, *et al. Cell* 2021; **184**: 4612-4625.



Parasites and Microbes



We study the genomics and evolution of disease-causing organisms and the human microbiome. We build networks at scale to help monitor infectious diseases and the effects of health policies worldwide, identifying the drivers of drug, vaccine, and insecticide resistance to guide health planning.

Malaria caused
409,000
deaths in 2019 alone

In this section

- 1 Malaria Cell Atlas maps parasite transmission
- 2 Microbiome bacteria adapt to humans via transmission
- 3 Tracking the rise of a multi-drug resistant bacteria
- 4 IBS biomarker could help personalise diet and treatment
- 5 MRSA arose in hedgehogs before use of antibiotics
- 6 Malaria parasites beginning to overcome sickle haemoglobin defence
- 7 Genomic surveillance of important neglected tropical disease
- 8 Vaccine hope for millions in sub-Saharan Africa
- 9 Discovering why a vaccine started to fail in Ireland
- 10 New genetic sequencing method sorts mosquitos for malaria surveillance

1 Malaria Cell Atlas maps parasite transmission

Transmission stages of the *Plasmodium falciparum* malaria parasite life cycle have been mapped by single-cell RNA sequencing for the first time. The Malaria Cell Atlas could lead to opportunities to block the parasite's development and prevent transmission of malaria.

Malaria parasites have a complex life cycle, featuring diverse developmental strategies that are uniquely adapted to specific environments of mosquitos and humans. The transmission stages – as the parasite moves from a mosquito's mid-gut to salivary glands, and then to human skin as the insect bites – have long been considered an attractive target for treatment development. There is an urgent need for new ways to fight malaria, a disease that affected an estimated 229 million people worldwide, and caused 409,000 deaths in 2019 alone.

Over the past four years, Sanger Institute scientists have developed the Malaria Cell Atlas to understand the biology of *Plasmodium* species. The atlas is an open-access data resource for scientists

to explore patterns of gene expression in individual parasites across multiple developmental stages and species. Their latest work adds a new chapter to the atlas, mapping the gene activity of *Plasmodium falciparum* during its transmission stages.

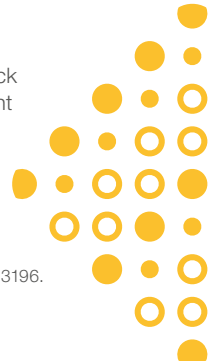
The team isolated different forms of *P. falciparum* during transmission and used single-cell RNA sequencing to illuminate gene usage. They produced data for 1,467 transcriptomes, showing which genes are active at each transmission stage.

Researchers identified essential genes for the formation of sporozoites: the parasite form released into human skin during a mosquito bite. They also compared the data with the related *Plasmodium berghei*, a rodent malaria parasite, identifying common and unique genes for each species.

As well as providing insights into gene function across the transmission cycle, the atlas could lead to new ways to block key stages in the parasite's development and prevent transmission.



Reference
Real E, et al. *Nature Communications* 2021; **12**: 3196.



2

Microbiome bacteria adapt to humans via transmission

Sanger Institute scientists shed new light on the evolution, colonisation, and transmission of beneficial gut bacteria.

Humans are colonised by populations of microorganisms – a microbiome – with the number of beneficial bacteria in the body roughly equivalent to the number of human cells. In the gut, symbiotic bacteria affect the immune system and break down nutrients that human cells cannot, playing an important role in health.

For gut bacterial species to survive, they must transmit from person to person. This includes being able to colonise the gut above a certain abundance to ensure onward transmission and being able to survive in the environment long enough to encounter a susceptible host. The bacteria must then compete with indigenous bacteria for nutrients and replicative niches to colonise.

To explore the genomic and biological adaptations underpinning symbiont transmission, Sanger Institute researchers investigated Firmicutes bacteria – a dominant phylum of intestinal microbiota. Firmicutes can produce spores, which are resistant to external environments and germinate within the intestine to facilitate transmission. Sporulation is a complex process, dependent on hundreds of genes, that results in the death of the original cell.

The researchers analysed the genomes of 1,358 Firmicutes. To determine their ability to produce spores, the team assigned the presence of 66 sporulation-predictive genes to each genome. These genes were identified using their previously developed machine learning model, based on analysis of nearly 700,000 genes.

Combining large-scale genomic analysis with phenotypic validation of human gut Firmicutes, the team observed that the ability to spore had been lost several times in many distinct lineages. This loss was associated with host adaptations such as a reduced genome size and more specialised metabolic capabilities.

Analysis of an additional 9,966 gut metagenomes, from adults around the world, reveals that bacteria no longer capable of sporulation are more abundant within individuals but less prevalent overall in the human population.

Their finding suggests that host adaptation in gut Firmicutes is an evolutionary trade-off between transmission range and colonisation abundance.



Reference

Browne H, *et al.* *Genome Biology* 2021; **22**: 204.



3

Tracking the rise of a multi-drug resistant bacteria

In the largest genomic survey of *E. coli* to date, researchers have tracked the increase of antibiotic resistance genes in the bacterium. The work has implications for controlling the spread of drug-resistant bacteria, which is a significant challenge in healthcare.

Escherichia coli bacteria are commonly found in the gut, where they cause no harm. However, if the bacteria get into the bloodstream, they can cause severe and life-threatening infections. *E. coli* bloodstream infections have been increasing world-wide over the last decade, with multi-drug resistance (MDR) becoming a frequent feature.

To understand the spread of MDR in *E. coli*, researchers from the Sanger Institute and the University of Oslo processed a nationwide catalogue of samples from more than 3,200 patients over 16 years in Norway, in the largest study of its kind.

The team used whole genome sequencing to infer the population structure of *E. coli* bacteria. They investigated the genetic characteristics of the dominant MDR bacterial subpopulations in Norway and compared the results with a longitudinal study from the UK.

The team found that MDR started to increase in the early 2000s in Norway, despite low antibiotic use in the country. MDR *E. coli* seems to be more widely present in the UK, despite having similar policies in place around antibiotic use – showing that longitudinal surveys of *E. coli* epidemiology are not generally comparable.

The researchers also uncovered bacterial lineages, previously not thought to have MDR, that have acquired drug resistance genes. This revealed the increased ability of *E. coli* to share MDR genes between strains.

Observations of emerging multidrug resistant bacteria, and the ability to share resistance genes between strains, have implications for efforts to control MDR and bloodstream infections and highlight the importance of ongoing genomic surveillance.



Reference

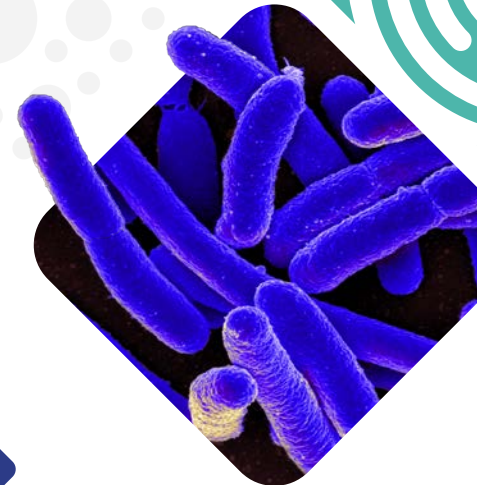
Gladstone RA, *et al.* *Lancet Microbe* 2021; **2**: e331-e341.



This study demonstrates the power arising from a systematic national surveillance of resistant organisms, which both collects and makes the data available for in-depth analyses. Without these in place, it would have been impossible to approach the central research questions formulated in the study and find answers to them.

Professor Jukka Corander

Co-author and Associate Faculty member at the Wellcome Sanger Institute



4

IBS biomarker could help personalise diet and treatment

Researchers have found that specific gut bacteria could be used as biomarkers to identify people with irritable bowel syndrome (IBS) who will benefit from a low FODMAP diet. The findings are helping to shed light on the mechanisms behind IBS and potentially provide new treatment options for those affected.

Irritable bowel syndrome (IBS) is a poorly understood, common, and often debilitating condition. The low FODMAP diet has been shown to improve IBS symptoms for many people, but it is often challenging to follow as it avoids certain fruits, vegetables, milk, and wheat products.

To better understand how the FODMAP diets work and who is more likely to benefit, researchers at the Sanger Institute, the University of Cambridge, and Addenbrooke's Hospital studied the microbiomes of 56 IBS patients. The participants, and anyone they lived with, committed to a low FODMAP diet for four weeks.

The researchers undertook an in-depth genomic, mechanistic, and functional analysis of the participants gut bacteria – microbiome – during the study. The team found that approximately half of the patients had a distinctive and abnormal profile to their gut bacteria at the start of the research. This 'pathogenic' profile was enriched in Firmicutes species of bacteria and genes for amino acid and carbohydrate metabolism but depleted in Bacteroidetes species.

These patients benefitted greatly from a low FODMAP diet – reporting that their symptoms were much improved after four weeks. Their improvement correlated with a dramatic shift in their gut bacteria towards a much more normal, 'healthy' profile. This suggests that the effectiveness of FODMAP may result from the alterations in gut microbiota and the metabolites produced.



IBS is a condition that affects thousands of people every day, yet the way that the gut microbiome and IBS symptoms are linked is still poorly understood. Our research dives deeper into the role of certain gut bacteria in possible IBS symptoms and gives a more precise view of what is going on. By understanding the role of these bacteria, it could lead to new targeted therapies for those living with IBS.

Dr Trevor Lawley

Joint senior author and Group Leader at the Wellcome Sanger Institute



Reference

Vervier K, et al. *Gut* 2021.



5

MRSA arose in hedgehogs before use of antibiotics

Scientists from the Sanger Institute, the University of Cambridge, the University of Copenhagen, and their collaborators have found evidence that the antibiotic-resistant superbug, MRSA, arose in nature long before the use of antibiotics in humans and livestock, which has traditionally been blamed for its emergence.

Methicillin-resistant *Staphylococcus aureus* (MRSA) is an antibiotic-resistant superbug that is considered to be one of the world's greatest threats to human health and a major challenge in livestock farming. The rise of MRSA has been previously linked to the increased use of antibiotics in humans and animals.

Hedgehog surveys from Denmark and Sweden have previously demonstrated a surprisingly high prevalence of MRSA carrying *mecC* (*mecC*-MRSA) – one of the genes responsible for the bacterium's resistance to the antibiotic methicillin. This raises the possibility that evolution of *mecC*-MRSA bacteria was driven by natural selection in wildlife.

To test this theory, the international team examined the distribution of *mecC*-MRSA and other *S. aureus* isolates in hedgehogs in ten European countries and New Zealand. They sequenced the DNA of 244 *S. aureus*

isolates from hedgehogs and 913 from other sources to infer the evolutionary histories, host dynamics, geographical dispersal patterns, and zoonotic potential of *mecC*-MRSA in Europe.

They found MRSA bacteria on 60 per cent of hedgehogs in Denmark and Sweden, and it was present in many other European countries too. To investigate the forces behind the evolution of the bacterium, the team sequenced genes from *Trichophyton erinacei* fungi samples. This fungus also lives on the skin of hedgehogs and produces its own antibiotics. They found evidence that *S. aureus*, which exists side-by-side to the fungus, first developed antibiotic resistance around 200 years ago as a response.

The research emphasises the importance of looking at the entire ecosystem when it comes to assessing antibiotic use and tracking the evolution of antibiotic resistance.



Reference

Larsen J, et al. *Nature* 2022; **602**: 135-141.



6

Malaria parasites beginning to overcome sickle haemoglobin defence

Some malaria parasites in sub-Saharan Africa have genetic variations that allow them to infect people with sickle haemoglobin – a blood condition normally thought to give strong protection against the disease. This new finding provides the clearest evidence to date of an interaction between genetic variants in the malaria parasite and its human host.

Sickle haemoglobin is a benign condition, caused by one abnormal copy of the haemoglobin beta gene. Sickle haemoglobin is commonly found in individuals from sub-Saharan Africa and offers protection against malaria. Malaria remains one of the world's most deadly infectious diseases, and it was responsible for an estimated 627,000 deaths in 2020, most of which were in sub-Saharan Africa.

A Sanger Institute team, together with colleagues from the USA, Mali, Kenya, and The Gambia, sequenced 3,346 *Plasmodium falciparum* malaria parasite genomes from children in The Gambia and Kenya with severe malaria symptoms. They also sequenced the genomes of the individuals infected.

Large-scale sequencing allowed the team to compare thousands of parasite and human genomes. They identified a strong association between sickle haemoglobin in the host and three regions of the parasite genome. Statistical analysis showed this link was not explained by population structure or other covariates. Based on estimation of relative risk, sickle haemoglobin had no apparent protective effect against severe malaria in the presence of the three genetic variations in the parasite. The variations were also found in parasites infecting people without sickle haemoglobin.

The researchers suggest that sickle haemoglobin in humans may have acted as a selective pressure on the parasite, driving it to adapt and leading to a malaria parasite population that can now infect people with sickle haemoglobin, as well as those with normal haemoglobin.

Further research into the function of these genetic variants in malaria parasites and the mechanisms by which they interact with sickle haemoglobin is needed. Understanding this could lead to new ways to protect against and treat malaria.



Reference

Band G, *et al. Nature* 2022; **602**: 106-111.



7

Genomic surveillance of important neglected tropical disease

Sanger Institute researchers have undertaken the first whole-genome analysis of drug treatment effects on the parasitic worm *Schistosoma mansoni*. The work is the largest whole-genome sequencing study of *S. mansoni* to date, and provides a foundation for genomic surveillance of an important neglected tropical disease.

Schistosomiasis is a neglected tropical disease affecting over 240 million people across 78 countries, primarily in sub-Saharan Africa. Caused by the parasitic worm *Schistosoma mansoni*, chronic infection can result in debilitating symptoms and organ damage.

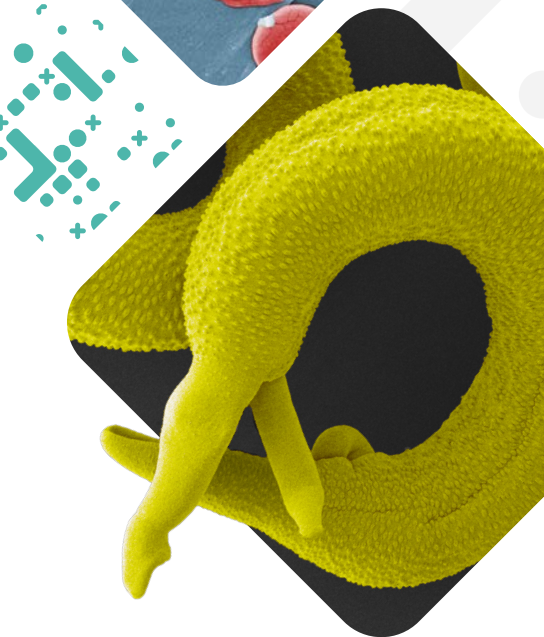
In 2001, the World Health Organization (WHO) endorsed the use of praziquantel in mass drug administration (MDA) campaigns to prevent the disease. In 2019 alone, MDA programmes delivered praziquantel to 105.4

million people across the world. The WHO recently set the ambitious goal of eliminating schistosomiasis in all 78 endemic countries by 2030.

Despite many successes, hotspots of the disease remain, and it is unclear why. The emergence of drug resistance in the parasite is a particular concern and has occurred in related parasitic species that infect animals. Surveillance of reduced praziquantel efficacy currently relies on slide-based counts of the parasite's eggs before and after treatment.

To uncover the impact of MDA on parasite populations, and to gain insights into the genomic diversity, potential gene flow, or drug resistance in *S. mansoni*, the Sanger Institute team performed whole-genome sequencing on parasite samples. The parasites were collected during a 2014 MDA follow-up survey in Uganda, and represented epidemiologically important, high-prevalence regions.

The team found that schistosome populations remain diverse and unstructured after nine years of annual MDA, with no indication of a praziquantel-resistant subpopulation. However, the researchers did identify genomic regions that could indicate gradual adaptation to long-term praziquantel exposure.



Given the limited understanding of how *S. mansoni* adapts, the researchers recommend further genomic surveillance. Accurately measuring changes in parasite populations will be a key part of understanding the current and future impact of MDA, provide advance warning of the emergence of drug resistance, and is a vital part of reaching the ambitious 2030 target of elimination.



Reference

Berger DJ, *et al. Nature Communications* 2021; **12**: 4776.

8

Vaccine hope for millions in sub-Saharan Africa

The first vaccine target for trypanosomes, a family of parasites that cause devastating disease in animals and humans, has been discovered. By targeting a protein on the cell surface of the parasite, *Trypanosoma vivax*, researchers were able to confer long-lasting protection against animal African trypanosomiasis infection in mice.



Caused by *Trypanosoma* parasites, animal African trypanosomiasis (AAT) is a devastating disease affecting livestock in Africa and South America. Animals infected suffer from fever, weakness, lethargy, and anaemia. The resulting weight loss, low fertility, and reduced milk yields have a huge economic impact on the people who depend on these animals, and is linked to poverty in affected areas of sub-Saharan Africa.

In humans, the related parasite species, *Trypanosoma brucei* causes sleeping sickness, and *Trypanosoma cruzi* causes Chagas' disease – potentially life-threatening infections that affect millions. All trypanosome species have developed sophisticated immunoprotective mechanisms to thrive in their hosts. For example, some parasite species constantly change their surface proteins, which prevents host antibodies from recognising the pathogen. Until now, it was thought impossible to vaccinate against trypanosome infection.

In this study, scientists at the Sanger Institute systematically analysed the genome of *T. vivax* to identify 60 cell surface proteins that could be viable vaccine targets. Each protein was produced using mammalian cell lines and then used to vaccinate mice.

One protein, named 'invariant flagellum antigen from *T. vivax*' (IFX), conferred immunity against infection in almost all vaccinated mice for at least 170 days after challenge with *T. vivax* parasites.

The study is the first successful attempt to induce such immunity against a trypanosome parasite. The next step will be to validate the results in cattle and, if successful, begin developing a vaccine for AAT. The findings also raise the possibility of identifying vaccine targets for trypanosome species that infect humans.



Reference

Delphine A, et al. *Nature*; 2021; **595**: 96–100.

9

Discovering why a vaccine started to fail in Ireland

A lineage, or subclade, of the *Streptococcus pneumoniae* bacterium, which is unique to Ireland, has been identified by Sanger researchers. The subclade was associated with half of the vaccine breakthrough infections the team studied, and it has a rare genetic variant that could offer the explanation for its persistence.

Streptococcus pneumoniae is a bacterium often found in the nasopharynx of children. However, it can cause severe invasive pneumococcal disease (IPD) in patients with immunosuppression, chronic diseases, in young children and older adults – particularly in low-income countries. Each year, IPD kills approximately a million children worldwide.

In *S. pneumoniae*, the capsule – a polysaccharide layer surrounding the bacterial cell – is a key virulence factor of the bacterium. It is used to determine the

version, or serotype, of the bacteria and is the target of current pneumococcal conjugate vaccines (PCVs). PCV13 was introduced in Ireland in 2010, and it has significantly reduced IPD in children. However, *S. pneumoniae* serotype 19A remains a significant cause of invasive pneumococcal disease (IPD) in the country – despite it being targeted by the vaccine.

To understand why *S. pneumoniae* serotype 19A persists, a team from the Sanger Institute studied bacterial samples from the Irish reference laboratory taken since 2007. The bacteria were analysed using whole genome sequencing, and the genomes were compared to international collections of 19A, using a standardised nomenclature of Global Pneumococcal Sequencing Clusters (GPSC).

The team found that expansion of specific GPSCs may be associated with vaccine introduction and antimicrobial prescribing policies. They also found a subclade of bacteria that was present in half of the vaccine failure 19A infections. This subclade is unique to Ireland, and the researchers estimated it emerged at the start of the PCV era. All of the samples within the subclade contained a rare *galE* gene variant.

It is possible that changes to the *galE* gene, which is involved in capsule production, may affect bacterial persistence. The research highlights the benefits of using whole genome sequencing in *S. pneumoniae*, and how these data may be useful for informing vaccine strategies.



Reference

Corcoran M, et al. *Vaccine* 2021; **39**: 5064-5073.



1,000

mosquitos can
be identified
at a time



10

New genetic sequencing method sorts mosquitos for malaria surveillance

Sanger Institute scientists have designed a sequencing approach that identifies a mosquito's species, reveals population structure within species, and detects the presence of malaria parasites. The quick, accurate method should help researchers around the world to massively scale up mosquito surveillance, including of their population structure and infection status, as we drive malaria towards elimination.

Anopheles is a diverse genus of mosquitos. There are over 500 described species, including those that carry and transmit malaria parasites. These key disease vectors are found in groups where there is ongoing speciation, as well as hybridisation between species. In order to eliminate malaria, all potential vector species need to be under surveillance.

When identifying a mosquito, typically its morphological features are examined to determine its species group. Targeted genetic analysis, looking at the presence of a few different markers, is then used to uncover the exact species – but identification is not always possible, and the results are not always reliable.

To improve the detection of *Anopheles* species, the Sanger team developed a multi-locus amplicon DNA sequencing approach. This method targets 62 highly variable locations in the *Anopheles* genome, selected based on the whole-genome alignment of representative species and to be phylogenetically informative. The team also selected two genome locations in the malaria parasite to determine if a mosquito is infected or not.

Their focus was on automating and reducing the cost of processing the mosquitos. They optimised a simplified DNA extraction method that does not use pre-existing kits or product purification. The method is high throughput – the DNA from over 1,000 individual mosquitos can be pooled for multiplex sequencing, using a widely-available sequencing platform. The DNA extraction is also non-destructive, so any unexpected results mean that the mosquito can still be assessed using morphological examination if needed.

The researchers tested the approach on 40 mosquito species from around the world, and were able to resolve population structures and dynamics, as well as infection status, in all species.

The end-to-end approach is quick, inexpensive, robust, and accurate, which makes it a promising technique for very large-scale mosquito genetic surveillance and control around the globe.



Reference

Makunin A, et al. *Molecular Ecology Resources* 2022; **22**: 28–44.

Tree of Life



We are building the library of life. We produce high-quality reference genomes to explore the evolution, function, and interactions of life on Earth. We seek to aid conservation and biodiversity work, and provide the underpinnings of a new way of doing biology.

The Darwin Tree of Life project aims to sequence

70,000
species

In this section

- 1 Speed reading every book of life
- 2 Genomics for all
- 3 Sequencing species' genomes at scale will drive era of biological discovery
- 4 Sharing Britain and Ireland's genomic treasure trove

1

Speed reading every book of life

Building on expertise in sequencing novel genomes, new teams, data tools, and pipelines have been established to generate high-quality reference genome assemblies – representing the entire genetic sequence of a species – at an unprecedented scale.

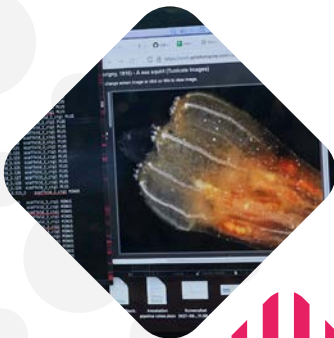
Less than 1 per cent of eukaryote species (animals, plants, fungi, and protists) have had a genome sequenced to date. As part of national and international efforts to sequence the genomes of all complex life on Earth, the Sanger Institute is aiming to sequence the genomes of 70,000 UK species for the first time.

A key part of the work has been optimising DNA sequencing methods. Sanger scientists have assessed approaches, and use a combination of platforms, including long-read technologies from Pacific Biosciences and long-range techniques such as 10X, HiC, and haplotagging to determine a species' genome.

In the next step of the process, bioinformaticians and data scientists quality check and assemble the sequence data into contiguous pieces. Information from long-range techniques is used to scaffold the fragments into chromosome-scale blocks. Sanger researchers have developed automated software tools and pipelines for many of these computationally intense processes, enabling high throughput.

The genome assemblies are assessed for quality and completeness, both manually and using bespoke software. For example, [gEVAL](#) is software that has been designed to identify, visualise, and assess genome assembly issues. Genome curators correct any errors and update the assemblies, improving accuracy. Other tools have been developed to assess genome metadata – for example, [GoaT \(Genomes on a Tree\)](#) was built to present and predict information such as genome sizes and chromosome number.

The raw data and complete, assembled genomes are openly available in public databases, and hundreds of assemblies have been produced so far. The data provide a rich foundation for future research into evolution, biodiversity, conservation, and biotechnology.



2

Genomics for all

Cataloguing, characterising, and sequencing the genomes of all complex life on Earth requires a global community. As part of the Earth BioGenome Project, a justice, equity, diversity, and inclusion committee has been formed to ensure these values are present in biodiversity genomics.

Researchers at the Sanger Institute and Rockefeller University, USA chair the justice, equity, diversity, and inclusion committee of the Earth BioGenome Project (EBP) – a global mission to sequence the genomes of all eukaryote species. The committee was formed following the inaugural Biodiversity Genomics meeting, organised by researchers from the Sanger Institute, Genomes 10k, the EBP, and others.

The committee's first focus was on increasing inclusion and equity and highlighting diversity at the 2021 Biodiversity Genomics meeting. The meeting brought together 5,000 attendees from 103 countries. The organisers intentionally invited diverse speakers, and discussions on equity and inclusion were at the heart of the conference.

The committee ran two new sessions at the meeting – the first focussed on lived experiences of challenges related to equity, diversity, and inclusion. The second focussed on lessons learned in applying actions towards ensuring justice, equity, diversity, and inclusion in genomics research. The meeting also included a code of conduct, and recommendations for accessibility, and it continues to be open to all and free to attend.

The committee also plans to improve the implementation of the FAIR (findability, accessibility, interoperability, and reusability) principles, Nagoya protocol, and other guidelines for scientific data management, equitable benefit sharing, and stewardship in collaborations with indigenous and other populations.

The diversity of people – their varied perspectives, insights, cultural backgrounds, and local knowledge of biodiversity across the world – is critical to meet the challenge of sequencing all life on Earth, and to do so with scientific excellence.



Reference

Biodiversity Genomics 2021

www.darwintreeoflife.org/news_item/biodiversity-genomics-2021-sequencing-genomes-across-the-planet/

3

Sequencing species' genomes at scale will drive era of biological discovery

The Vertebrate Genomes Project (VGP) reached a significant milestone in 2021, completing flagship studies focused on genome assembly quality and standardisation. Scientists at the Sanger Institute led research to produce 16 new, high-quality vertebrate reference genome assemblies, including the Canada lynx, platypus, greater horseshoe bat, zig-zag eel, and Anna's hummingbird. The tools and standards developed are a gold standard in genome assembly.

Chromosome-level reference genome assemblies – representing the entire genetic sequence of a species – underpin functional, comparative, and population studies. The very first genome sequences to be determined, including the human genome, were put together using short DNA sequencing reads. This required tremendous manual effort, software engineering, and cost in projects lasting decades.

Since then, advances in technology have reduced costs and time, but generating a reference genome still resulted in low-quality assemblies, fragmented into thousands of pieces, where many genes were missing or inaccurate. Such errors can require years of manual effort to fix.

To address this, researchers in the VGP developed cost-effective methods for assembling accurate reference genomes.

First, they extensively evaluated multiple approaches on one species, the Anna's hummingbird. The team confirmed that long-read sequencing technologies are essential for maximising data quality. They showed that complex repeat regions of a genome, and haplotype heterozygosity, are major sources of assembly error when not handled correctly.

They deployed the best-performing method on 16 species, refining the techniques. The VGP's approach now combines automated pipelines with streamlined manual curation. The resulting new assemblies correct substantial errors and add missing sequence, in some of the best historical reference genomes.

The genome assemblies also enabled the team to make new discoveries, not possible with previous sequences. The first reference genomes of six bat species, for example, revealed selection and loss of immunity-related genes that may underlie bats' unique tolerance to viral infection.

The international VGP is now working to generate high-quality, complete reference genomes for all 70,000 vertebrate species – enabling a new era of discovery.



Reference

Rhie A, et al. *Nature* 2021; **592**: 737–746.



A high-quality
reference genome of
**Anna's
hummingbird**
has been generated



More than
3,000
species have been
collected so far

4

Sharing Britain and Ireland's genomic treasure trove

The Darwin Tree of Life project aims to sequence the genomes of all 70,000 species of eukaryotic organisms – animals, plants, fungi, and single-celled protists – in Britain and Ireland. The project aims to transform biology, conservation, and biotechnology.

The Darwin Tree of Life project is a collaboration between biodiversity organisations and genomics institutes. It brings together the Sanger Institute, Natural History Museum, Royal Botanic Gardens of Kew and Edinburgh, Marine Biological Association, Earlham Institute, EMBL's European Bioinformatics Institute, and the Universities of Oxford, Edinburgh, and Cambridge, among many others.

The magnitude of the work – involving many organisations and highly diverse species – has presented many challenges. To coordinate efforts, teams have developed data-tracking systems to collate and share sample metadata. The workflow starts as project partners collect, photograph, and identify specimens. A flash-frozen tissue sample is sent to the Sanger Institute for genome sequencing. To aid species identification, researchers also DNA barcode all specimens at mitochondrial, chloroplast, or ribosomal RNA loci before genomic analysis. Over 3,000 species have been collected so far.

The biochemistry of some species has posed additional challenges. For example, the cell walls of plants and mucus in marine invertebrates make extracting long DNA fragments difficult. Laboratory teams are trialling and developing effective procedures for extracting DNA above the 150 Kb size required for sequencing.

Sanger scientists have optimised the sequencing approaches and developed software to automate many of the computational processes required to quality check, assemble, and analyse the data.

Data are openly published; in 2021, 300 genome assemblies were submitted to the European Nucleotide Archive (ENA) and made available to the scientific community worldwide. Genomes are also presented as [Genome Notes](#) in the journal *Wellcome Open Research*, with the aim of promoting the discovery and reuse of the datasets. 71 Genome Notes were published in 2021.



Reference
The Darwin Tree of Life Project Consortium. *Proceedings of the National Academy of Sciences* 2022; **119**: e2115642118.





Chicken of the woods

Laetiporus sulphureus

Genome size: **37.4 Mb**
Chromosomes: **14**

An edible wood-decaying fungus that is found on broad-leaved trees in the UK. It is a key engineer in hollowing out the heartwood of established and ancient trees, creating microhabitats full of biodiversity. Its actions are also an important and beneficial part of trees' natural ageing processes.

<https://wellcomeopenresearch.org/articles/7-83>



European water vole

Arvicola amphibius

Genome size: **2.298 Gb**
Chromosomes: **19**

A small, semi-aquatic mammal that lives on the banks of freshwater habitats and in wetlands throughout the UK. The genome provides a reference for researchers seeking to assess water vole population genetics, to better understand how the species has evolved, and to manage reintroduction efforts.

<https://wellcomeopenresearch.org/articles/6-162>



Bootlace worm

Lineus longissimus

Genome size: **391 Mb**
Chromosomes: **19**

A predatory ribbon worm, renowned as being the world's longest animal. The genome may help uncover new chemical compounds for use in agriculture, biotechnology, and medicine.

<https://wellcomeopenresearch.org/articles/6-272>



Lime hawk-moth

Mimas tiliae

Genome size: **478 Mb**
Chromosomes: **29**

Common throughout southern England, particularly London. It can be found in woodland and in suburban habitats, and it flies in May and June. The genome will help researchers studying chromosome and species evolution.

<https://wellcomeopenresearch.org/articles/6-357/v1>

Our approach



Collaborating on the Tree of Life:
Researchers working in concert, preparing
samples of species gathered by partners
around Britain and Ireland.

Our people are our strength: powering our curiosity, creativity, and culture

Scale

Genomic inquiry requires vast volumes of data, experimental models, and computational power. Our Institute's unique, scalable, and robust infrastructure delivers – both for us and researchers worldwide.

Page **44**

Innovation

To take our research findings to the next level and deliver transformative technologies, we work in collaboration with pharmaceutical industries and funders.

Page **46**

Collaboration

We use the power of the internet and collaboration tools to build genomic research capacity worldwide and facilitate the next wave of discovery.

Page **48**

Culture

The diversity in skills and knowledge that we all bring combine to make the Institute the thriving ideas factory that it is. We support our colleagues to reach their full potential and to help each other thrive in their work. We encourage everyone to benefit from the wide range of creativity and expertise at the Institute by valuing each other's differences in thought, background, and perspective.

Page **50**

Scale



In this section

- 1 500,000 whole human genomes to understand disease
- 2 Prioritising drug targets openly and systematically
- 3 Genomic data access goes global



1 500,000 whole human genomes to understand disease

In a major advance for genomics, the Sanger Institute has sequenced nearly 250,000 whole human genomes, as part of the UK Biobank project. The dataset is the world's largest of its kind and will help researchers to understand the genetic determinants of disease and accelerate drug discovery.

UK Biobank is a large-scale biomedical resource and database containing anonymised genetic, lifestyle, and health information from half a million participants. The project is enabling a better understanding of illnesses including cancer, heart disease, and stroke. More than 28,000 scientists from 98 countries have been approved to use the resource and thousands of studies using the data have been published. The project is a collaboration between government, industry, and charity.

The Sanger Institute undertook a pilot study to whole genome sequence the first 50,000 UK Biobank participants in 2018. Following that success, the Institute was contracted, together with deCODE in Iceland, to sequence the remaining UK Biobank participants. The Institute's world-leading teams and facilities were expanded to deliver this unprecedented, large-scale project. Sequencing was completed on time in early 2022, delivering high-quality genome data to the project.

The expansion at Sanger, specifically of liquid handling capabilities, IT infrastructure, and Illumina technology, increased the Institute's capacity to such an extent that it was subsequently able to take on additional large-scale projects, including sequencing coronavirus genomes to aid the pandemic response.

Five petabytes of data, for 200,000 UK Biobank participants, is already available to researchers – representing the world's largest single release to date of whole human genome sequence data. The additional 300,000 genomes will be made available in early 2023. By combining the genome data with the rich clinical and lifestyle information in UK Biobank, researchers will be able to answer new questions about the genome and its role in health and disease.



2

Prioritising drug targets openly and systematically

Researchers at the Sanger Institute and Open Targets have created an open resource that allows users to easily prioritise genes at disease-associated locations in the genome and assess their potential as drug targets.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits and diseases. However, identifying the causal gene connected to a variant is a major challenge; the variants found in GWAS are often in regulatory regions of the genome, which may affect a number of different genes, and their significance is unknown. Finding the genes that cause disease is crucial, as drugs aimed at genetically validated targets are more likely to progress through clinical trials.

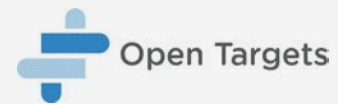
Researchers from Open Targets have previously developed an open resource that formats, harmonises, and aggregates

GWAS results. This includes all 133,441 publicly available human GWAS loci. The team integrated the GWAS genetic data with transcriptomic, proteomic, molecular, and epigenomic data, systematically combining information from over 33,781 studies.

To enable researchers to find genes related to disease, the team have developed a machine-learning model to recognise causal genes from GWAS variants. The model was trained using GWAS loci with high confidence in the causal gene and a set of predictive features, and it was used to identify and score genes related to all GWAS loci. The score reflects the likelihood a gene is causal for a particular trait and allows genes to be ranked by the relative strength of evidence. The resulting prioritised gene list was enriched for known, approved drug targets.

These results are publicly available at genetics.opentargets.org, a web portal that features 72,909,456 genetic variants, 34,659 genes, 4,835 diseases, and 4,900 unique genes with a high score from the machine learning model.

The portal enables users to easily prioritise genes at disease-associated loci and assess their potential as drug targets.



Open Targets is a public-private partnership that uses human genetics and genomics data for systematic drug target identification and prioritisation.



Reference

Mounjtjoy E, *et al. Nature Genetics* 2021; **53**: 1527–1533.



If you would like to know more, visit: <http://genetics.opentargets.org>

3

Genomic data access goes global

Working with the Global Alliance for Global Health (GA4GH), the Sanger Institute has helped to publish the standards, agreements, and processes needed to support open access to genomic and health data worldwide.

Genomic research relies on analysing enormous datasets to find minor differences and, often, no single study contains enough data. To overcome this, researchers pool data from multiple sources around the world to generate the statistical power needed to make new discoveries.

However, differences in regulatory and data protection processes – together with differing clinical and genomic terminology – can create formidable barriers. To remove these obstacles, the Sanger Institute helped to found the Genomic Alliance for Global Health (GA4GH) to standardise genomic data sharing for the benefit of human health.

The consortium, with expert guidance and administrative support from the Sanger Institute, has developed a range of standards and techniques that have been published in a landmark series of ten papers in *Cell Genomics*.

Three of the papers detail processes to standardise data-use requests and researcher identification. These processes work in tandem to enable automatic data access by demonstrating that the scientist, and their research, fulfil the necessary criteria. Two of the papers, on Data Use Ontology and the GA4GH passport, drew on Sanger Institute expertise.

The Data Use Ontology system enables data owners to define the terms of use for each dataset. The GA4GH passports standard then provides a digital identity for individual scientists that details their role and data access permissions, matched to the ontology. In this way, researchers can easily access biomedical datasets – to the level appropriate for their role and permission – across passport-recognising networks. The system is already working successfully for ELIXIR Europe, the US National Institutes of Health, and the Autism Sharing Initiative.



Innovation

In this section

- 1 COSMIC has a stellar future
- 2 Improving health, in practice
- 3 Sanger showcases UK's genomic power

1 COSMIC has a stellar future

The Sanger Institute's world-leading cancer genomics data resource has secured a long-term partnership that will drive its growth and evolution. The deal guarantees that the huge database of knowledge will continue to power global research into precision medicine for cancer.

COSMIC (Catalogue of Somatic Mutations in Cancer) started in 2002 as a spreadsheet to allow Sanger cancer researchers to keep track of all the genes and DNA changes associated with different tumours. Today, it is an expert-curated database used by 20,000 scientists worldwide that covers more than 71 million DNA patterns involved in over 1,500 human cancer types.

To ensure the long-term future of this vital genomic resource and enable ongoing expansion and innovation, COSMIC has secured a multi-year partnership with Qiagen. The deal will provide a stable financial base for the database and its services.

As part of the Institute's commitment to provide open access to vital genomic knowledge, the service will continue to be free and open to the global academic and not-for-profit community. Qiagen will use its sales experience and global customer network to offer licensing opportunities to the commercial sector. Revenues received by the Sanger will be reinvested to support COSMIC's continued growth.

To find out more about COSMIC and its wide range of services, including interactive 3D visualisations and catalogues of drug targets, please visit: <https://cancer.sanger.ac.uk/cosmic>.



If you would like to know more, visit: <https://www.wellcomegenomecampus.org/news/wgc/new-exclusive-partnership-for-cosmic-and-qiagen.article>



Through this partnership, we will further pursue our vision to enable the global cancer research, discovery, and development community with the gold-standard resource that COSMIC represents, whilst reaffirming our commitment to maintain free access to not-for-profit academic users.

Dr Adrian Ibrahim
Head of Technology Transfer and Business Development.

3

Improving health, in practice

With their first cancer-fighting microbiome-based therapy entering clinical trials and their genomic diagnostic platform certified for medical use, two of Sanger's spin outs are changing healthcare.

The Sanger Institute is committed to delivering on the promise of its research to improve diagnosis and treatment. Translating discoveries into real-world tools and therapies requires long-term investment and support, which the Institute provides either by creating a spin-out company, licensable technique, or non-profit consortium.

Microbiotica is a diagnostic and therapeutics company founded on Institute-developed bacterial microbiome libraries and research. It is set to trial its first live microbiome precision medicine for cancer in patients.

The gut microbiome plays a key role in determining which cancer patients respond to Immune Checkpoint Inhibitor (ICI) therapy. Microbiotica identified the first microbiome signature that improved response to the ICI therapy for melanoma and developed a live bacterial therapeutic made up of nine key species from the signature. When used as a co-therapy with ICIs, it showed potent anti-tumour efficacy in mice.

Congenica is a world-leading genomic diagnosis company developed from Sanger's human genetics research. It has received the CE mark for its genomic interpretation software, enabling it to transition from research to clinical use. The company's clinical decision support platform can now provide actionable diagnoses for national health services in the UK and across the EU.

In addition, Congenica has received £2 million to develop clinical decision software that uses pharmacogenetics data to reduce adverse drug reactions. HAPPY (Healthy Ageing Pharmacogenetics and Polypharmacy) will alert GPs to the risks of prescribing specific combinations of medicines for patients aged more than 50 years to enable doctors to investigate less harmful options.



If you would like to know more, visit: <https://microbiotica.com/microbiotica-live-bacterial-therapeutic-mb097-in-development-to-begin-clinical-trials-in-2022-in-immuno-oncology/>

<https://blog.congenica.com/congenica-receives-ce-mark-for-genomic-analysis-software>

<https://blog.congenica.com/ukri-and-ig-fund-congenica-to-reduce-fatal-drug-interactions-by-mapping-genes>

2

Sanger showcases UK's genomic power

The UK genomics sector has massive potential to contribute to the nation's health and wealth. A report by the Sanger Institute, BioIndustry Association, and Medicines Discovery Catapult maps out opportunities.

The report – *Genomics Nation* – maps out the size and shape of the UK's genomics sector, revealing that the country's exceptional scientific strength, world-leading genetic datasets, pioneering collaborations and national health system are driving the next wave of innovation and industry.

The report highlighted the UK's:

- ◆ Strong legacy in genomics – from discovering DNA's double helix, through building the first human reference genome, to sequencing 100,000 genomes
- ◆ World-leading research institutes and academia
- ◆ Unique data resources, such as the UK Biobank (500,000 volunteers) and the upcoming Our Future Health (5 million volunteers)
- ◆ National Health Service (NHS) – the world's largest united healthcare system, delivering genomic innovations to patients

- ◆ Active industrialisation of discoveries through Medicines Discovery Catapult centres
- ◆ Thriving genomics industry and active investor base, working in collaboration with academia and the NHS.

One example of the UK's innovative and connected approach is the clinical services that Sanger spin out company Congenica supplies to the NHS. Developed from the Institute's academic research, it provides the UK Genomic Medicine Service with its clinical decision support platform and genomic data analysis for rare disease cases.

Another is the UK's PCR COVID-19 testing network where the NHS, academia, and industry collaborated to deliver the largest diagnostics project in UK history, supplying a clinically relevant, genomically-driven patient testing system at massive pace and scale. Many positive tests undergo whole genome sequencing and analysis at the Sanger Institute to guide public health decisions.

The report concludes that this alignment of capabilities and collaboration has built the UK's world-leading position and will drive future innovation and investment in genomic research.



If you would like to know more, visit: <https://www.bioindustry.org/news-listing/new-report-reveals-strength-of-the-uks-thriving-genomics-sector.html>

UK Genomics Sector



Collaboration



In this section

- 1 Getting to the guts of continental disease differences
- 2 Fighting the spread of drug-resistant malaria with genetic surveillance
- 3 Open Targets gains Pfizer's drug expertise

1 Getting to the guts of continental disease differences

A collaboration across South America is seeking to understand the differences in people's gut microbes compared to their North American and European counterparts. The consortium hopes its findings will provide insights into the differences in diseases prevalent in Latin America.

The Latinbiota consortium, led by Wellcome Sanger Institute International Fellow Gregorio Iorio, is a continent-wide network of researchers working to characterise the gut microbes of people living in Uruguay, Mexico, Ecuador, Colombia, Brazil, Argentina, and Bolivia.

The bacterial population (microbiome) in a person's gut can affect disease progression and treatment effectiveness. And the main factor that shapes a person's microbiome is their diet which, in turn, is strongly determined by social, cultural, and economic factors. However, the majority of studies on gut microbiomes have been in Europe and North America, where diets are markedly different to those of people in South America.

To understand microbiome variability and its effects on health in Latin American populations, the Latinbiota Consortium is sequencing the DNA of the gut bacteria in 600 people in eight countries. Most of the participants are healthy to allow study of normal variability in healthy individuals and to give a reference of high-quality genomes. Armed with this baseline, the researchers will study diseases of concern in each participating country.

The team are also growing the bacteria in the laboratory to explore further their behaviour. The Sanger Institute helped to create this biobank by training members of Professor Iorio's team to grow the cultures in the anaerobic conditions needed to mirror the gut. It is hoped that this work will help to form the basis of future microbiome-based therapies tailored to South American populations.



2

Fighting the spread of drug-resistant malaria with genetic surveillance

The in-depth genetic surveillance of the malaria parasite across multiple countries in the Greater Mekong Subregion, by Sanger Institute researchers and their collaborators, has tracked regions of drug resistance and helped inform public health decisions in Southeast Asia.

Malaria continues to be a major cause of mortality, and efforts are ongoing to eliminate the *Plasmodium falciparum* parasite that causes the most severe form of disease. Drug resistant strains of *P. falciparum* have repeatedly arisen from Southeast Asia and migrated into Asian and African countries, undoing years of progress in tackling the disease, and costing many lives.

The GenRe-Mekong project involves researchers from the Sanger Institute, the University of Oxford, MalariaGEN, and several partners in the Greater Mekong Subregion. The GenRe-Mekong group has developed and implemented SpotMalaria, a platform for genetic surveillance of malaria to monitor drug resistance.

SpotMalaria only requires small dried blood spot samples from the fingers of people infected. The samples are easy to collect at public health facilities as part of routine treatment, making it possible to obtain data from remote and resource-poor areas.

The malaria parasite DNA is then analysed at Sanger using high-throughput sequencing technologies. The analyses produce a comprehensive set of genotypes including drug resistance markers, species markers, and a genomic barcode. So far, 9,623 blood samples from patients across eight countries in Southeast Asia have been processed. National Malaria Control Programmes are involved throughout, from sampling strategy planning through to joint analyses of results.

In Vietnam and Laos, GenRe-Mekong data have provided knowledge about the spread of drug-resistant strains of malaria into previously unaffected provinces and informed decision-making by public health agencies.

GenRe-Mekong also facilitates data sharing regionally, providing the public health community with valuable insights. The project provides a rich open data resource to benefit the entire malaria research community.

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK



Reference

Jacob C, et al. *eLife* 2021; **10**: e62997.

3

Open Targets gains Pfizer's drug expertise

In 2022, Pfizer joined the Campus' pioneering public-private partnership: Open Targets. The consortium brings together the Sanger Institute, EMBL's European Bioinformatics Institute (EMBL-EBI), and pharmaceutical companies to speed drug development and delivery through the application of big data.

The Open Targets initiative launched in 2014 to address a key challenge in drug development: that the majority of all compounds entering clinical trials fail to become licensed medicines due to lack of effectiveness or patient safety issues. The root of this failure is often a lack of understanding of the drug's biological target.

Open Targets addresses this problem by marrying the expertise of academic and commercial researchers to create a critical mass of knowledge and data that could not exist in any one organisation. The consortium combines genetic and pharmaceutical information with genomic data to systematically identify small molecules that are most suited to a disease's biological targets. In this way, the partners can prioritise the drugs that are most likely to succeed, reducing the time and cost of development.

Pfizer will contribute its unique expertise in oncology, immunology, and metabolic disorders and add to the pharmaceutical R&D approaches and information provided by partners GSK, Bristol Myers Squibb, and Sanofi. The Sanger Institute provides data and analysis methods from large-scale genomic experiments, together with computational techniques developed by EMBL-EBI.

Open Targets openly shares its experimental data and findings to benefit the broader scientific community through its online Open Targets Platform <https://platform.opentargets.org/>. The freely available online tool integrates publicly available data to reveal associations between drug targets and diseases and uses annotation information to contextualise those associations. The platform contains over 60,600 targets, more than 12,500 drugs and small molecules, and almost 10 million associations, covering more than 18,000 diseases.

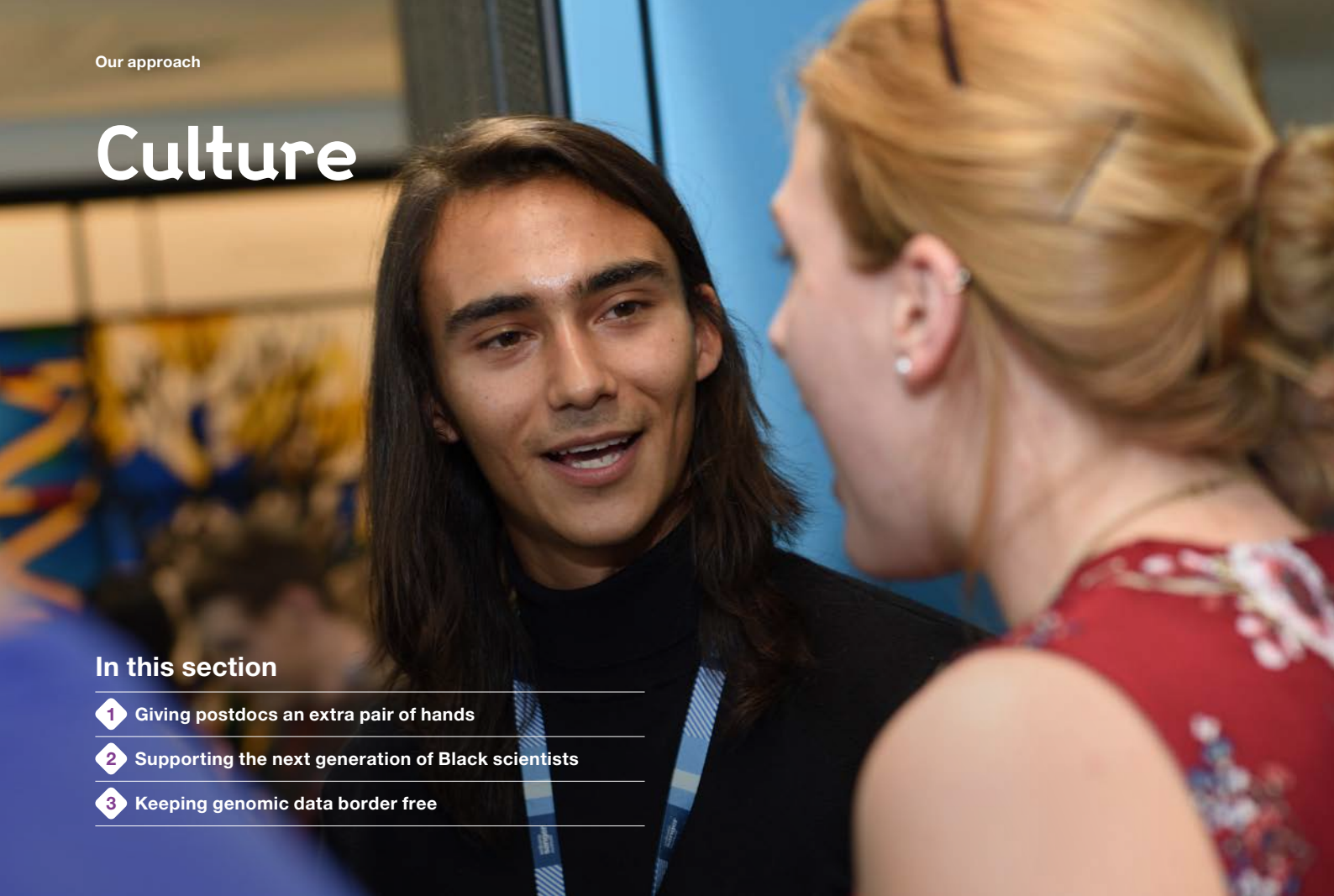


Pfizer's emphasis on the use of human genetics and genomics in drug discovery is a perfect fit for the partnership.

Dr Ian Dunham

Director of Open Targets

Culture



In this section

- 1 Giving postdocs an extra pair of hands
- 2 Supporting the next generation of Black scientists
- 3 Keeping genomic data border free



1 Giving postdocs an extra pair of hands

Postdoctoral fellows with caring responsibilities during the pandemic have found it more difficult to recover their science. An innovative partnering scheme provides a literal second pair of hands.

The Institute has a large and diverse postdoctoral fellow community of approximately 120 scientists across all five research programmes and scientific operations. Comprising more than 24 nationalities, 65 per cent of postdocs are non-British, and one quarter live alone.

To support its postdocs during the pandemic, and to enable them to recover their science after restrictions lifted, the Institute surveyed their needs. Key negative impacts were time away from the laboratory, lockdown isolation, and damage to career plans.

In response, the Institute gave all postdoctoral fellows a six-month funding extension, offered tailored resilience and mental health webinars, and provided bespoke career coaching. But, the survey revealed another need that required a more innovative solution.

Approximately one-third of the Institute's postdocs have carer responsibilities, and the burdens of homeschooling or caring for frail family members had disproportionately affected their science. In total, 61 per cent of carers indicated that their work was delayed by more than six months, compared with 25 per cent of non-carers.

To help, the Institute is providing postdoctoral fellows with carer responsibilities a second pair of skilled hands to conduct experiments. Through its post-pandemic research support scheme, Sanger is paying for technicians from the Institute's Scientific Operations division to work part-time on secondment for qualifying postdocs.

Both the postdoctoral fellow and volunteer technician gain from the scheme. The postdoctoral fellow is able to make up for lost time, while the technician is able to develop their skills and experience.

2

Supporting the next generation of Black scientists

To nurture talent from the widest spectrum of experience, the Sanger Institute has launched a Fellowship to support the training and career development of early-stage Black heritage researchers who have studied at a UK institution.

The Sanger Excellence Fellowship builds on the Institute's commitment to greater equality, diversity, and inclusion within its organisation and science. The three-year funded Fellowship seeks to address a critical imbalance in science's attraction and retention of people from Black heritage backgrounds.

The Sanger Institute recognises that persistent racial inequalities disadvantage people from Black backgrounds in all walks of life, including academia. People from Black backgrounds are under-represented in higher education and in the biological sciences, compared to the wider UK population.

According to the education consultancy Leading Routes, over a three-year period, just 1.2 per cent of nearly 20,000 studentships awarded by UK research councils were given to Black or Black Mixed students.

The Sanger Institute developed the Fellowship in close collaboration with internal and external experts, including senior Black academics and Black-led community groups. It seeks to bring in excellence that may otherwise be lost to science, and catalyse change in academia. The aim is to provide fellows with training, mentorship, and support to drive their careers forward and support a more diverse pipeline of future talent. The fellow will work within the Institute or Wellcome Connecting Science.

To ensure that the widest possible pool of talent can apply, all applicants receive 1-2-1 guidance on developing their application and proposed science from Sanger Faculty. And all unsuccessful candidates receive feedback and the opportunity to be mentored and receive career development support to help them find future success.



It is our hope that the new Fellowship will help to level the playing field by providing a clear career development pathway, training, mentorship, and support to ensure that Black researchers and academics can achieve their full potential.

Dr Saher Ahmed
Head of Equality,
Diversity and Inclusion
at the Wellcome
Sanger Institute



The Sanger Institute is committed to
open-access data sharing

3

Keeping genomic data border free

The Sanger Institute has partnered with the NHS Confederation to help the EU ensure that health and genomic data are openly shared across Europe and the UK.

The Institute is committed to open access and data sharing and to ensure genomic data is freely available across international borders for the research community. Future improvements in healthcare and drug development depend on the collective power of large datasets drawn from cohorts spanning many countries.

However, each country handles health and genomic data differently, with their own legal and cultural restrictions. To help navigate these issues, while respecting all contributors' needs, the Sanger Institute is supporting the [Joint Action Towards the European Health Data Space initiative](#).

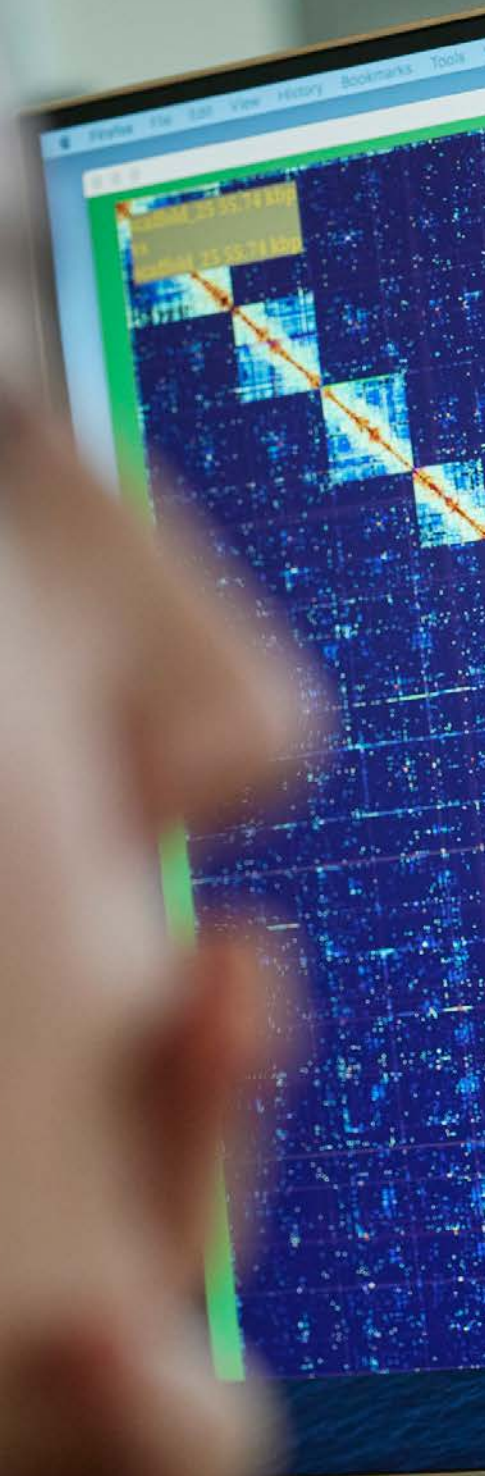
Working with the NHS Confederation to provide insights into the needs of UK-based researchers, the NHS and UK patients, the Sanger Institute is supplying its expertise in two areas. First, it is providing legal recommendations on the structures and processes needed to ensure compliance with GDPR regulations and other legislation. Second, it is supporting the project's citizen engagement work in the UK to gather patients' perspectives and respect cultural sensitivities.

The project aims to ensure that EU citizens, communities, and companies will benefit from secure and seamless access to health data regardless of where it is stored. In particular, it focuses on developing and promoting concepts for the use of health data to benefit public health and health research and innovation in Europe. Its findings will inform the European Commission's legal framework for the European Health Data Space.



To take part in the consultation and have your say on how your health data should be used, please visit: <https://ourhealthydata.eu/en>

We lay the foundations for future science



Careful curation:
Creating the tools, techniques, and networks needed to generate high-quality reference genomes to power the study of the genomics of all life on Earth.

Image Credits

All images in this Annual Highlights document belong to the Wellcome Sanger Institute or have been sourced from AdobeStock or iStock except where stated below:

Contents

Pages 4-5 Scientist in laboratory – ThisisEngineering RAEng, Unsplash

Timeline

Page 6 Wellcome Leap Logo – Wellcome Leap
Kymab Logo – Kymab
MalariaGEN Logo – MalariaGEN
E. coli – National Institute of Allergy and Infectious Diseases (NIAID)
Cells dividing – K. Hardy CC BY 4.0
DDD Logo – Deciphering Developmental Disorders project

Page 7 Camels crossing desert – Jeff Jewiss, Unsplash
UK Biobank Logo – UK Biobank
SARS-CoV-2 Virus – Alissa Eckert, MSMI; Dan Higgins, MAMS

COVID-19 Programme

Page 12 SARS-CoV-2 Virus – Alissa Eckert, MSMI; Dan Higgins, MAMS

Page 13 Open Targets Logo – Open Targets

Page 15 SARS-CoV-2 Virus – Alissa Eckert, MSMI; Dan Higgins, MAMS
England virus transmission maps – Vöhringer HS, *et al. Nature* 2021; **600**: 506–511.

Cancer, Ageing and Somatic Mutation Programme

Page 16 Cancer cells – Anne Weston, Francis Crick Institute

Page 18 Oesophageal cells – Dr Bartomeu Colom, Wellcome Sanger Institute

Page 20 Prostate cancer cell – Anne Weston, Francis Crick Institute
Developing embryo – Dr M. Zernicka-Goetz, Gurdon Institute

Page 21 Breast cancer cells – Annie Cavanagh CC BY-NC 4.0

Cellular Genetics Programme

Page 22 Red blood cells – David Gregory and Debbie Marshall CC BY 4.0

Page 23 Kidney cells – Kenny Roberts, Bayraktar Lab, Wellcome Sanger Institute

Page 24 Developing embryo cells – Mole MA, *et al. Nature Communications* 2021; **12**: 3679.

Page 25 Cell2location cell maps – Omer Bayraktar, Wellcome Sanger Institute

Page 27 Uterine endometrial cells – Kenny Roberts, Wellcome Sanger Institute

Human Genetics Programme

Page 28 Child and building blocks – Pixabay
Hands and building blocks – FeeLoona, Pixabay
DDD Logo – Deciphering Developmental Disorders project

Page 30 Yeast cultures – Jolanda van Leeuwen

Page 31 Camels crossing desert – Jeff Jewiss, Unsplash
Landscape – Rabah Al Shammary, Unsplash

Parasites and Microbes Programme

Page 32 Fluorescent mosquito – Gianmarco Raddi, Wellcome Sanger Institute

Page 33 *E. coli* – National Institute of Allergy and Infectious Diseases (NIAID)

Page 35 Sickle cell – Janice Haney Carr / CDC

Page 36 *Streptococcus pneumoniae* – Debbie Marshall CC BY 4.0

Page 37 *Anopheles gambiae* mosquito – James Gathany / CDC

Tree of Life Programme

Page 38 Insect in sample vial – David Lavene / Wellcome Sanger Institute
Bioinformatic work – David Lavene / Wellcome Sanger Institute

Page 40 Four photos of laboratory work in the Tree of Life Programme – David Lavene / Wellcome Sanger Institute
Darwin Tree of Life Logo – Darwin Tree of Life project

Page 41 Bootlace worm – Darwin Tree of Life project

Our Approach Section

Pages 42-43 Sanger scientists in laboratory – David Lavene / Wellcome Sanger Institute

Page 44 UK Biobank Logo – UK Biobank

Page 45 Open Targets Logo – Open Targets

Page 46 COSMIC Logo – COSMIC

Page 47 Microbiotica Logo – Microbiotica
Congenica Logo – Congenica

Page 48 Sanger scientists discussing results – David Lavene / Wellcome Sanger Institute

Page 49 MalariaGEN Logo – MalariaGEN
Sanofi Logo – Sanofi
GSK Logo – GSK
Open Targets Logo – Open Targets
Bristol Myers Squibb Logo – Bristol Myers Squibb
Pfizer Logo – Pfizer

Page 50 Sanger researchers sharing ideas – David Lavene / Wellcome Sanger Institute

Page 52 Sanger bioinformatician at work – David Lavene / Wellcome Sanger Institute

Wellcome Sanger Institute Highlights 2021/22

The Wellcome Sanger Institute is operated by Genome Research Limited, a charity registered in England with number 1021457 and a company registered in England with number 2742969, whose registered office is 215 Euston Road, London NW1 2BE.

First published by the Wellcome Sanger Institute, 2022.

This is an open-access publication and, with the exception of images and illustrations, the content may, unless otherwise stated, be reproduced free of charge in any format or medium, subject to the following conditions: content must be reproduced accurately; content must not be used in a misleading context; the Wellcome Sanger Institute must be attributed as the original author and the title of the document specified in the attribution.

Printed by Kingfisher Press
on FSC® certified paper.

Kingfisher Press is an EMAS certified company
and its Environmental Management System
is certified to ISO 14001.

This document is printed on Soporset Offset,
a paper containing 100% virgin fibre sourced
from well-managed, responsible, FSC® certified
forests and other controlled sources.

Designed and produced by **MadeNoise**

We use information from genome sequences to advance understanding of health and disease

Hibernating hedgehog:

An international collaboration has discovered that the MRSA superbug arose on the skin of hedgehogs in response to a natural fungus, long before the use of antibiotics.

Wellcome Sanger Institute
Tel: +44 (0)1223 834244

sanger.ac.uk

