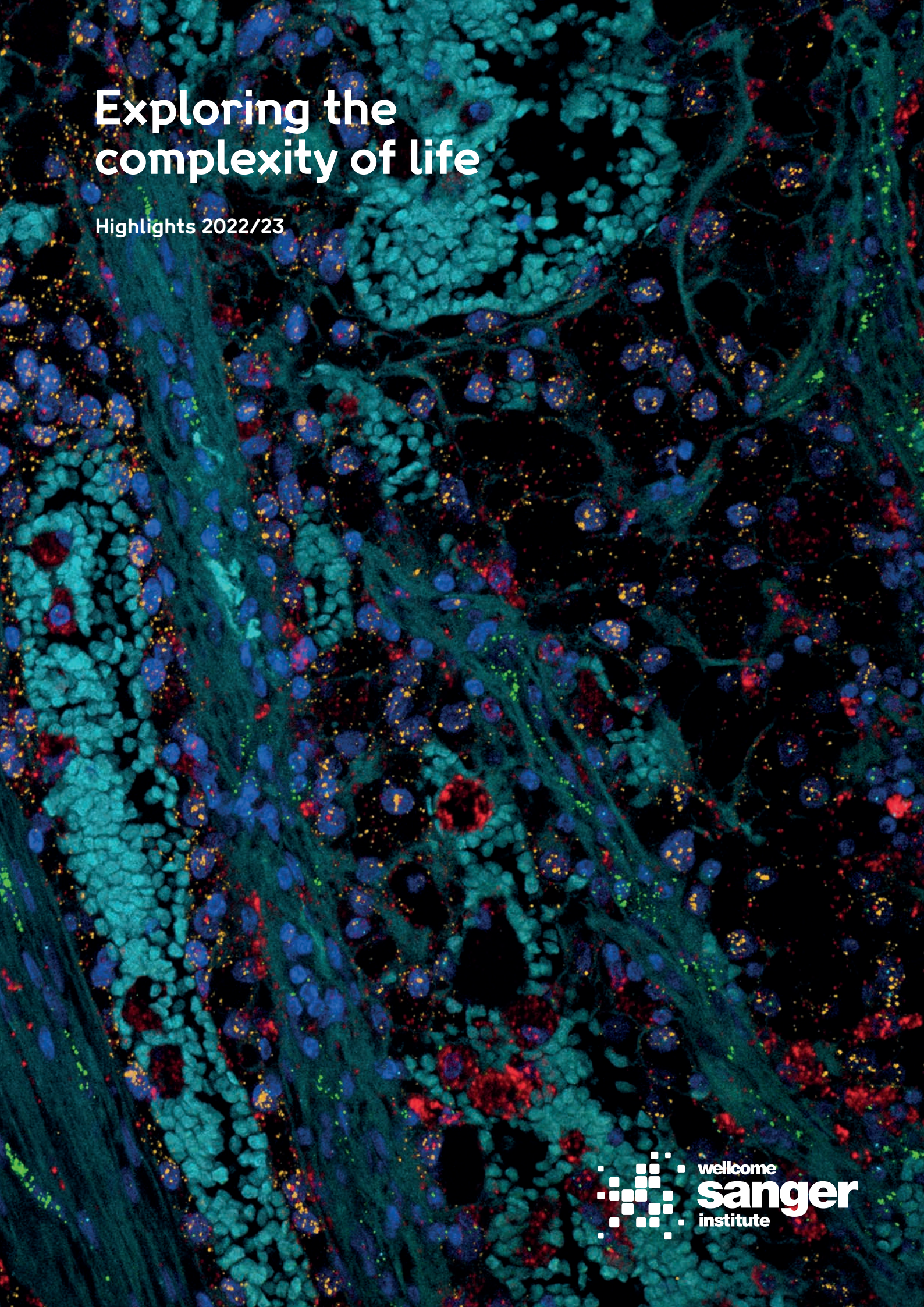


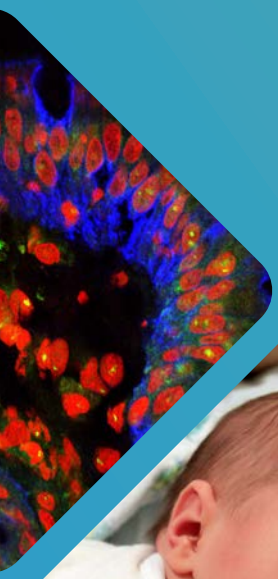
Exploring the complexity of life

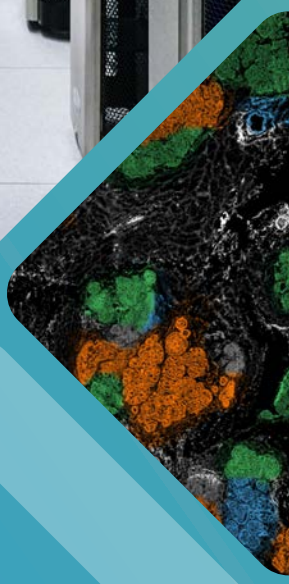
Highlights 2022/23

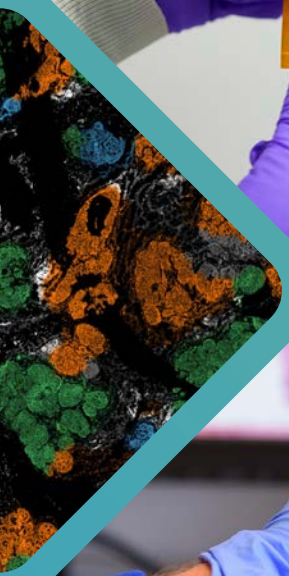
wellcome sanger institute

Contents

We push the frontiers of biology through genomic science







What we do

Our work

Our approach

Other information

Building a brighter, fairer future

As the cloud of COVID-19 restrictions lifts around the world the value of truly global genomic research is clear. Yet inequalities in scientific capacity abound: from administrative support and genomic resources, through technological knowledge and digital infrastructure, to data sharing and reward. The need for research equity has never been more pressing.

We are proud of the long-term partnerships that the Sanger Institute has fostered to share the benefits of infectious disease surveillance and improve clinical understanding of inherited and complex conditions. After 18 years, the DECIPHER project continues to deliver vital insights into developmental disorders (page 40). The Global Pneumococcal Sequencing Project has provided unparalleled insights into a bacterium responsible for hundreds of thousands of infant deaths each year (page 27). While the Human Cell Atlas consortium is laying the foundations for single-cell exploration of health and disease for years to come (page 21).

But we know we can do more. Over the past year, our Policy team has worked with our international partners, research equity experts, social scientists, funding organisations and journal editors to develop guidelines that build equity into every stage of our research projects' life cycles (page 46).

Here in the UK, equity is also of vital importance. The Sanger Institute is founded on the diversity of thought and experience of all its staff. Yet marginalised groups are often underrepresented in research, especially at more senior levels. Over the past 10 years, our Equality, Diversity and Inclusion team have built networks of support, changed policies and lowered barriers to entry to create an environment where the greatest diversity of researchers and staff can be themselves and thrive (page 43).

As part of this work, we are delighted to welcome our inaugural Sanger Excellence Research Fellows (page 42). The Fellowships are specifically designed to support the training and career development of scientists from Black heritage backgrounds. Our first three fellows will be powering the Institute's research into how the respiratory metagenome develops in early life, the timelines and trajectories of blood cancers, and how proteins work together to create and sustain eukaryotic life.

As our scientists ramp up their processes and pipelines to deliver the reference genomes for all eukaryotic life in the UK, our Tree of Life programme celebrated generating its 500th species' genome sequence (page 35). While these genomes provide vital insights to conservationists seeking to retain and promote biodiversity, their impact will spread into all areas of human health and disease. By analysing the genomes of 16 different species of mammal, Sanger researchers have found that lifespan is inversely proportion to the rate of somatic mutation (page 13).

Lastly, the COVID-19 cloud has yielded two exciting silver linings. Drawing on our expertise built up by delivering the UK's COVID genomic monitoring, we launched the Genomic Surveillance Unit (page 38). This new venture will power new networks of global infectious disease monitoring networks and data sharing.

Here in the UK, we are studying the treasure trove of respiratory samples gathered during nationwide testing for COVID-19 to create a comprehensive picture of viruses and bacteria circulating among people in the UK (page 31). The Respiratory Virus and Microbiome Initiative will provide unparalleled insights into the symbiotic and opportunistic relationships between respiratory viruses and bacteria to guide future therapies and monitoring.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute

[The Institute] has worked with international partners, research equity experts, social scientists, and funding organisations to [seek to] build equity into every stage of our research.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute

2022 Timeline

JAN

- Leverhulme Centre for the Holobiont launched to study organisms and microbiomes
- PhD to explore structural inequalities in genomic research created with Wellcome Connecting Science

wellcome connecting science

MRSA arose in hedgehogs before the use of antibiotics

FEB

- New drug combinations found for resistant cancers
- Digital Sequence Information network promotes fair genomic data sharing

Genomics shows COVID-19 travel restriction effectiveness

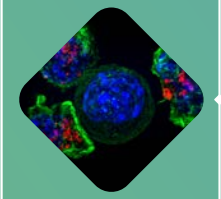
MAR

DNA variations associated with immune-mediated diseases mapped

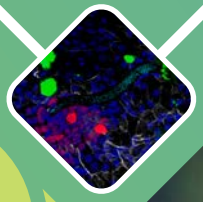
- Complete human genome sequence published
- Evolutionary pressures on genes associated with childlessness

APR

- Mutations across animal kingdom shed light on ageing
- New bacteria linked with Inflammatory Bowel Disease in mice
- Stonewall Workplace Equality Index survey launched



Mini guts used to study whipworm infections



MAY

Mistakes in father's sperm causes children's increased mutation rates

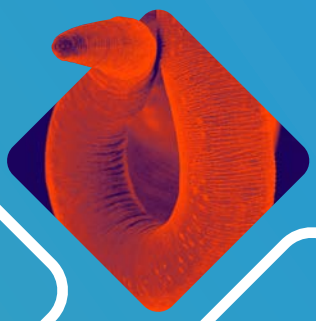
- Human immune system's development from early life to adulthood mapped
- Open Targets – drug targets revealed by tracking T-cell activation

Open Targets

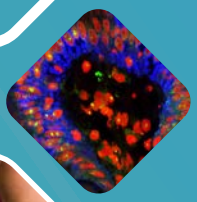
JUN

- How DNA mutations affect lifelong blood cell production revealed
- Cellular secrets of ageing unlocked

European badger reference genome published



Global Genomic Surveillance Unit launched

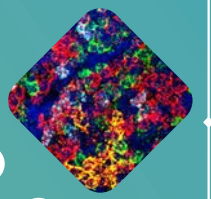


Biodiversity Genomics Europe launched to tackle biodiversity crisis

Genetic model predicts COVID-19 patients' risk of sepsis

JUL

- Cells involved in human sex determination mapped
- First genetic map of one of humans' oldest parasites
- World's largest database for predicting cancer treatment response released



AUG

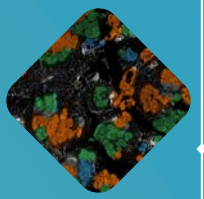
- Changes in genes involved in Crohn's disease identified
- Nuffield Research Placement students contribute to Sanger Science

SEP

- Open-access software uses genomic imbalances to find cancer cells

OCT

- Three inaugural Wellcome Sanger Excellence Fellows announced
- Parasitic worm's drug resistance genes mapped



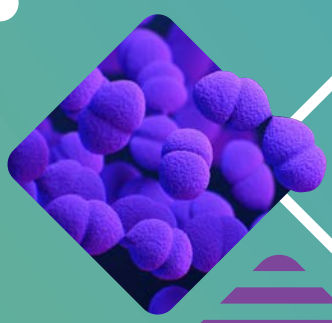
NOV

- Breast cancer spread uncovered by new molecular microscopy
- Archaeology and field walking tours of Campus extension works

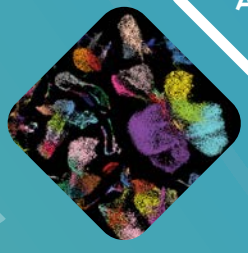
DEC

- Reference genomes to help protect Britain's wild and ancient apples
- Healthy UK newborns not colonised by multi-drug resistant hospital bacteria

Genomics empowers vaccine makers to tackle shapeshifting bacteria



Killer whale genome published



Developmental Lung Cell Atlas uncovers 144 cell states

At a Glance

Cellular

196
gene-edited cell lines

190,000
constructs cloned for pooled CRISPR libraries

81
CRISPR screens

106
organoid models banked

15,000
flow cytometry experiments

DNA sequencing

An average of
11,000bn
DNA bases a day were read

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

2,074
species sequenced on short-read machines

976
species sequenced on long-read machines

Compute

28,500
approx. total number of compute cores

15,000
approx. high performance compute cores in Sanger HPC clusters

13,500
Approx. cores as part of Sanger's private cloud environments

80PB
approx. useable storage

Year in Numbers

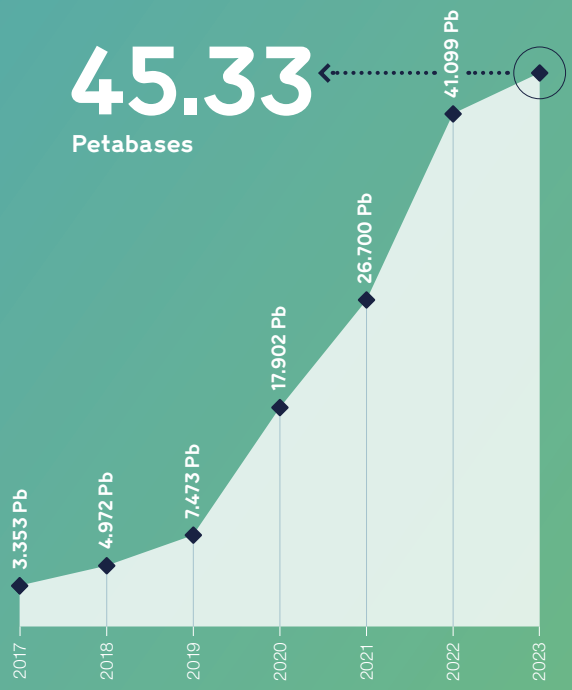
544
articles and reviews in 2022

1,293,476
citations of our papers and reviews (1996-2022)

9,584
published articles and reviews (1996-2022)

Sanger Institute authored papers and reviews are
4.14
times more cited than the world average*

**data for the past five years 2018-2022.* (Field-weighted citation impact (FWCI) metric)



Cumulative total of DNA sequenced by the Sanger Institute

Our research in numbers

Where Sanger staff are from*

How much DNA was sequenced?

Our work

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.

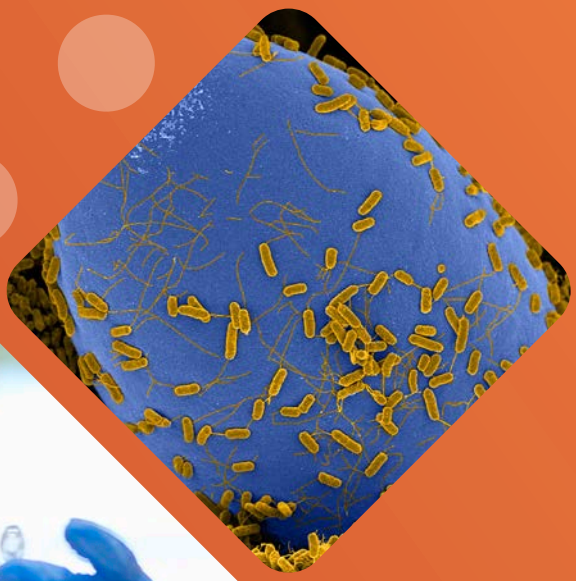
Page 12

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 18

We lead & support major national & global research collaborations



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 22

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 26

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 32

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.

In this section

1

Immune cells acquire genomic scars in a lifetime defending against infection

Wellcome Sanger Institute researchers have genetically sequenced immune cells in a higher resolution than ever before, giving insight into how and when they accumulate mutations. The findings could aid understanding of ageing, immune disorders and cancer.

Cells of the adaptive immune system, including memory B and T cells, undergo programmed genetic mutation to generate diverse receptors or antibodies – enabling the cells to fight a wide range of pathogens. The programmed mutation, if off-target, is known to cause some cancers, but an understanding of the mutational processes of immune cells has been missing.

To quantify and compare the genomic landscapes of these lymphocyte cells, researchers from the Sanger Institute and the University of York created a new protocol to grow and sequence colonies of B and T cells from a single cell.

The team then sequenced genomes from 717 immune cells from seven donors. They found that the number and patterns of DNA mutations in B and T cells varied widely from cell to cell – much more than between individuals. This suggests that mutations acquired in different areas of the body, caused by infections and inflammation, could play a larger role in disease than inherited variation.

Analysis of the patterns of mutation, known as mutational signatures, showed that some of the genetic variation was caused by the accumulation of mutations with age. Some variation was caused by DNA damage from sunlight, which lymphocytes are exposed to as they pass through the skin, surveying for infections.

Other variation was caused by off-target damage from the normal genetic enhancement of antibodies and antigen receptors. It is this collateral damage that can lead to lymphomas. It will be important to understand why these mutations cause some cells to become cancerous.

This high-resolution research gives in-depth insights into the mutational landscape of healthy immune cells and the processes that cause mutations as we age.

Reference
Machado H.E. *et al. Nature* 2022; **608:** 724-732.

2

Theory of ageing confirmed across the animal kingdom

The first study to compare the accumulation of genetic mutations across animal species has answered decades-old questions about the role of these DNA changes in ageing and cancer. Sanger Institute researchers uncovered that, despite huge variations in lifespan and body size, different animal species end their natural life with similar numbers of genetic changes.

Genetic changes, or somatic mutations, accumulate in all cells throughout life. Most are harmless, but some mutations can start a cell on the path to cancer or impair its normal functioning.

Since the 1950s, researchers have speculated that somatic mutations play a role in ageing, though it has not been possible to observe these until recently. Another long-standing question is Peto's paradox – the observation that cancer incidence is independent of body size; despite having many more cells, larger animals are not at higher risk of the disease.

To explore these theories, the Sanger team measured somatic mutations in 16 species, including human, mouse, lion, giraffe and the long-lived, cancer-resistant naked mole-rat.

The researchers used and optimised recently developed methods including low-input DNA sequencing protocols, and laser capture microdissection, to sequence the whole genomes of 208 intestinal crypts from the 16 species. A bioinformatics pipeline was developed to identify the somatic mutations.

Analysis of the somatic mutation patterns showed that these were caused by similar, internal mechanisms in all the species.

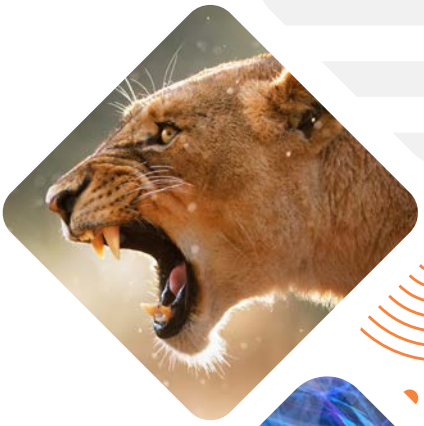
There was no association between somatic mutation rate and body mass, and so the search for an answer to Peto's paradox continues.

However, the team did find that each animal acquired a similar number of somatic mutations over its life, independent of the body mass or lifespan. Those with shorter lifespans had faster rates of somatic mutations, and those with longer lifespans had slower mutation rates. This inverse correlation confirms the long-standing prediction of the somatic mutation theory of ageing.

With the recent advances in DNA sequencing technologies, we can finally investigate the roles that somatic mutations play in ageing and in multiple diseases. That this diverse range of mammals end their lives with a similar number of mutations in their cells is an exciting and intriguing discovery.

Dr Inigo Martincorena
Group Leader, Wellcome Sanger Institute

Reference
Cagan A *et al. Nature* 2022; **604:** 517-524.



3

Mutation pathway to bowel cancer discovered

Researchers have found that inherited mutations in the *MUTYH* gene, which repairs oxidative DNA damage, are linked to an increased mutation rate in healthy tissues. This understanding could lead to new ways to prevent and treat bowel cancer.

Bowel cancer is the fourth most common cancer in the UK, with around 42,300 people diagnosed each year. The causes of the disease are complex, with inherited genetic mutations known to play a role.

Inherited mutations in the *MUTYH* gene, which repairs DNA damaged by oxidative stress, lead to a syndrome known as MUTYH-associated polyposis (MAP) and an elevated risk of early-onset bowel cancer. However, the mechanistic progression from inherited mutations to the development of bowel cancer is not fully understood.

Sanger Institute scientists and their collaborators from Cardiff University sequenced the genomes of intestine and blood cells from 10 MAP patients. They characterised the mutation rates and mutational processes in these healthy tissues at a near single-cell resolution.

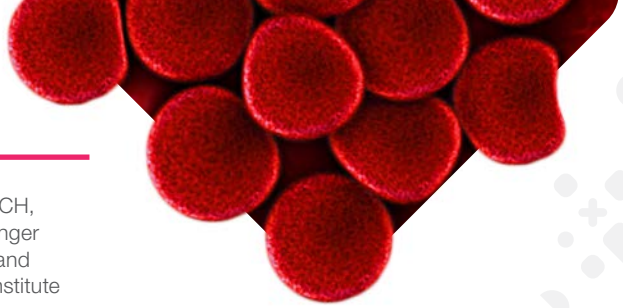
The team identified two distinctive patterns of mutations, or mutational signatures, previously seen in individuals with bowel cancer not linked to an inherited *MUTYH* gene mutation. This finding adds to the current understanding of cellular changes linked to the development of this disease.

The team also found a considerably increased rate of mutation in healthy intestinal cells. There was a correlated increased mutation rate in circulating blood, suggesting it may be possible to develop a blood test to stratify bowel cancer risk in the future.

Their work supports the recent finding that normal cells can handle more mutations than previously thought, and that increased mutation burden alone is not responsible for the ageing process. However, it is likely that increased mutation rates in normal intestinal cells, throughout life, lead to an increased risk of cancer.

References
Robinson P.S. *et al. Nature Communications* 2022; **13:** 3949.

Robinson P.S. *et al. Nature Genetics* 2022; **53:** 1434-1442.



4

Understanding blood cell production, ageing and cancer

Genetic variants have been identified that increase the likelihood of developing clonal haematopoiesis, a risk factor for multiple conditions including heart disease and blood cancer. Researchers have also uncovered how genetic mutations hijack the production of blood cells in different periods of life, and how these changes relate to ageing and the development of age-related diseases.

All human cells acquire genetic changes, or somatic mutations, throughout life, with a specific subset of mutations driving cells to multiply. This is common in blood stem cells and results in the growth of clones – populations of blood cells with identical mutations. The process, termed clonal haematopoiesis (CH), becomes ubiquitous with age. Although symptomless, CH is a risk factor for developing blood cancer and other age-related conditions.

To understand the mechanisms of CH, researchers from the Wellcome Sanger Institute, the Universities of Bristol and Cambridge, the Health Research Institute of Asturias in Spain and AstraZeneca, analysed genetic and health data from 200,453 people. These individuals are part of UK Biobank, a large-scale biomedical database containing information from half a million UK participants.

The team investigated the genetic, health and lifestyle factors linked to CH in the participants, using genome-wide association methods, among others. This in-depth analysis found 14 genes associated with CH, ten of which had not previously been identified. The genes implicate several mechanisms involved in CH: DNA damage repair *(PARP1, ATM, CHEK2)*, stem cell migration *(CD164)* and oncogenesis *(SETBP1)*.

They found that CH accelerated the process of ageing and influenced the risk of developing atrial fibrillation. They also clearly established that smoking is one of the strongest risk factors for developing CH. The work is a step change in understanding how CH increases the risk of disease.

In a related study, scientists at the Sanger Institute, the Cambridge Stem Cell Institute and EMBL's European Bioinformatics Institute (EMBL-EBI) tracked nearly 700 blood cell clones from 385 individuals. They found most clones expanded at a stable exponential rate, influenced by the nature of the mutated gene. Mathematical modelling uncovered that clone behaviour changed dramatically with age, depending on the identity of the mutated gene. For example, clones with mutations in splicing genes *U2AF1* and *SRSF2* expanded exclusively later in life and exhibited some of the fastest growth.

The age-dependent clonal behaviours mirror the frequency of emergence of different types of blood cancers. The study represents the first time that the lifelong impact of genetic mutations on cell growth dynamics has been explored.

References

Kar S.P. *et al. Nature Genetics* 2022; **54:** 1155-1166.

Febre M.A. *et al. Nature* 2022; **606:** 335-342.

5

Bowel cancer mutations that impact immunotherapy identified

A new library documenting hundreds of genetic mutations in bowel cancer has been created by researchers from the Wellcome Sanger Institute, Open Targets and their collaborators. The teams used CRISPR gene-editing technology to understand the role of thousands of mutations and found around 300 genetic changes that are also seen clinically in tumours. Some of the mutations change how receptive cancer cells are to the body's immune response. Their findings can help to explain why some cancers do not respond to immunotherapies and highlight potential pathways that could be drug targets in the future.

Immunotherapy treatments help the body's immune system fight off cancer. While they are not suitable for all cancer types, they have greatly improved treatment options for some types of bowel and blood cancers. However, immunotherapy is not effective in some patients, and it is unclear why.

To understand more, the team used systematic CRISPR Cas-9 base editing to study mutations that influence how sensitive bowel cancer cells are to cytokine interferon-γ (IFNγ). IFNγ signalling underpins the body's responses to infection, inflammation and anti-tumour immunity.

The team used deep mutagenesis on bowel cancer cell lines to identify mutations that affect how the cancer cells respond to IFNγ. While some of the mutations were previously known, their significance was unclear. To validate the findings, the researchers used tumour organoids, which mimic a tumour in the body, to see the direct impact of the different mutations on the development and progression of cancer in the presence of immune cells. They found 300 mutations that alter IFN-γ activity.

As a result, the team has created a library of different mutations that affect the sensitivity of bowel cancer cells to immune cell attack. The resource may help explain why some patients do not respond to immunotherapy.

Being able to see the impact of different genetic mutations on immunotherapy could also lead to more targeted treatment approaches in the future, where the patients who will benefit the most can be identified. It also highlights mutations that could be targeted in drug development, leading to new therapies.

Reference

Coelho M.A. *et al. Cancer Cell* 2023; **41:** 288-303.e6

6

Cellular secrets of ageing unlocked

Research has uncovered how genetic changes that accumulate slowly in blood stem cells throughout life are likely to be responsible for the dramatic change in blood production after the age of 70. The study, by scientists at the Sanger Institute and the Wellcome-MRC Cambridge Stem Cell Institute, suggests a new theory of ageing.

Ageing is likely to be caused by the accumulation of multiple types of damage to cells over time. One theory is that the build-up of acquired genetic changes, or somatic mutations, causes cells to progressively lose capacity to function. However, it is unclear how such gradual accumulation of molecular damage could translate into the abrupt deterioration in organ function after the age of 70.

To investigate the ageing process, the team studied the production of blood cells – haematopoiesis – in 10 individuals aged 0 to 81 years. They sequenced the whole genomes of 3,579 blood progenitor cells and identified the somatic mutations. They used the data to reconstruct 'family trees' of each person's blood stem cells, showing for the first time an unbiased view of the relationships among blood cells, and how these change across the lifespan.

The team found that blood production in adults under 65 came from 20,000 to 200,000 stem cells. In contrast, haematopoiesis in individuals aged over 70 was from 10 to 20 stem cell populations, which contributed as much as half of all blood production. These highly active stem cells had multiplied over time, due to the steady introduction of 'driver mutations'. The driver mutations caused the growth of functionally altered clones over decades.

This model explains the dramatic and inevitable shift to reduced diversity of blood cell populations after the age of 70. Which clones become dominant varies from person to person, and so the model also explains the variation seen in disease risk and other characteristics in older adults.

We've shown, for the first time, how steadily accumulating mutations throughout life lead to a catastrophic and inevitable change in blood cell populations after the age of 70 … We know this can increase cancer risk, but it could also be contributing to other functional changes associated with ageing.

Dr Peter Campbell
Head of Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute

Reference
Mitchell E. *et al. Nature* 2022; **606:** 343-350.

7

Protective mutation impairs oesophagus tumour growth

Researchers from the Wellcome Sanger Institute have found that mutations in the *NOTCH1* gene, prevalent in the human oesophagus after the age of 50, may have a protective effect. Their work highlights potential new ways to prevent or treat cancer.

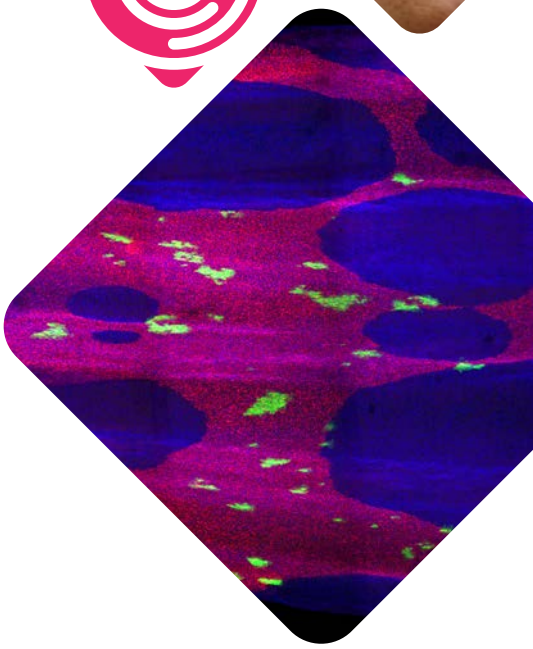
Over time, all cells in the body acquire genetic mutations, and while the majority of these do not affect how a cell functions, some give an advantage that allows cells to grow at a faster rate. Sometimes, mutations cause uncontrollable growth leading to cancer.

By middle age, the human oesophagus evolves into a patchwork of mutated cells. While the majority of these mutations do not lead to cancer, if tumours do form, they can be hard to treat as symptoms appear when the cancer has started to spread.

Oesophageal cancer is the sixth leading cause of cancer deaths globally, with more than 500,000 new cases diagnosed annually and an overall five-year survival rate below 30 per cent. While mutations in genes such as *TP53* are found in almost all oesophageal squamous cell carcinoma tumours, mutations in the *NOTCH1* gene are relatively rare in these cancers despite being very common in normal oesophagus tissue, suggesting *NOTCH1* mutants may protect against cancer.

To understand the role of *NOTCH1*, Sanger Institute researchers analysed the DNA of samples from normal oesophageal epithelium from middle-aged and elderly donors. They found frequent *NOTCH1* mutations affecting both copies of the gene and inactivating its function.

Using mice, the team found that cells containing a *Notch1* mutation spread throughout the tissue but then reverted to near normal behaviour. When tumours were formed by treating mice with a chemical from tobacco, *Notch1* mutants spread over the normal oesophagus but slowed tumour growth. In addition, a NOTCH1 blocking antibody reduced tumour growth.



NOTCH1 inhibitors are in clinical development for certain types of cancer, and understanding more about how *NOTCH1* mutations function could help uncover new ways to prevent tumours.

Reference
Abby E. *et al. Nature Genetics* 2023; **55:** 232-245.

8

Origins of germ cell tumours unravelled

Scientists from the Wellcome Sanger Institute, Cambridge University Hospitals NHS Foundation Trust and collaborators have detailed the origins of germ cell tumours. They found that even though these tumours appear at different ages and can contain multiple cell types, their mutational origins can often be traced back to a genetic event that happened by chance in the womb. Their findings also reveal possible avenues for future therapies.

Malignant germ cell tumours can appear at any age and are one of the most common cancers in young men. Germ cell tumours derive from a type of cell found in the embryo and recapitulate a variety of cell types, including muscle, placenta or teeth and hair. The cell types within the tumour have implications for prognosis.

The researchers analysed how the tissues develop and contrasted them with how healthy tissues grow, something that has not previously been possible.

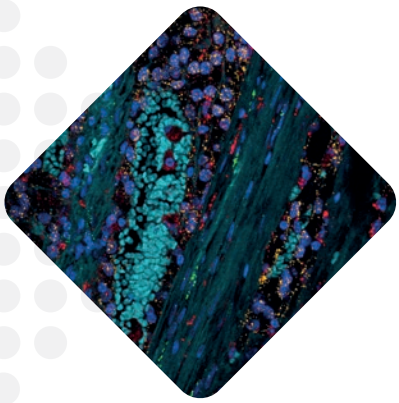
The team examined 547 microdissections of tumour samples from 15 individuals. By applying in-depth genetic sequencing techniques, they were able to study the DNA and RNA of all the different tissues within the tumours at an unprecedented resolution.

The results revealed common genomic features of the tissues, including the retention of foetal developmental patterns of gene activation, specific activation patterns on chromosome 12 and a sequence of whole genome duplication followed by diversification. The similarities in how the tumours created tissues, such as cartilage or muscle, to how tissue is created in a growing embryo may help in the development of new treatments.

The team was able to trace the mutational origin of the tumours all the way back to the beginning of their development in the womb. The initial mutation appears to happen by chance.

The researchers also identified different mutational signatures, or patterns, in tumours of young children compared with tumour samples taken from older children. These could be used as a future biomarker that allows healthcare professionals to identify which course of chemotherapy is the most appropriate.

Reference
Oliver T.R.W. *et al. Nature Communications* 2022; **13:** 4272.



I'm optimistic that targeting *IL1B* macrophages may provide us with a way to treat renal cell carcinomas without resorting to surgery. This will be particularly important for patients with VHL disease because we should be able to prevent tumours forming in the first place.

Dr Thomas Mitchell
Wellcome Sanger Institute
and University of Cambridge

9

New hope for kidney cancer treatment using existing drugs

The most comprehensive study of kidney cancer at the single-cell level has discovered a potential drug target to treat renal cell carcinoma, a cancer with a high mortality rate that is hard to detect. Researchers from the Wellcome Sanger Institute, the University of Cambridge and Cambridge University Hospitals identified a type of immune cell crucial to tumour development. These cells are a promising therapeutic target, as they are the focus of existing drugs that prevent lung cancer.

Renal cell carcinoma (RCC) is the seventh most common cancer in the UK, with three quarters of cases and the majority of deaths caused by clear cell renal cell carcinoma (ccRCC). The disease has a 50 per cent mortality rate, partially because most patients show no symptoms until the cancer is at a late stage.

To better understand the complex multi-cellular ecosystem of the tumour microenvironment, the team studied over 270,000 single cells and 100 micro-dissections from 12 patients with kidney tumours. Samples were taken from different parts of the tumour, which is well known to be heterogeneous, as well as normal kidney tissue. These samples were analysed using single-cell RNA sequencing and spatial transcriptomics to map the exact location of specific cells within tissues.

This analysis highlighted a particular type of immune cell, a macrophage expressing the gene *IL1B*, as abundant at the fringes of tumours.

The researchers are already planning clinical trials to test whether targeting *IL1B* macrophages is an effective treatment for RCC and could prevent secondary cancer for those with RCC predisposition syndromes, such as Von Hippel-Lindau (VHL) syndrome.

The fact that existing drugs targeting this pathway are proven to be effective in preventing some lung cancers offers hope that these trials may deliver promising results.

Reference
Li R. *et al. Cancer Cell* 2022; **12:** 1583-1599.e10.

10

Breast cancer spread uncovered by new molecular microscopy

New technology can trace which populations of breast cancer cells are responsible for the spread of the disease. The method was created by a team from the Wellcome Sanger Institute, EMBL's European Bioinformatics Institute (EMBL-EBI), the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), the Science for Life Laboratory in Sweden and collaborators. In the future, this approach could be used to see how treatments influence the cancer at not only the genetic level, but also any impact on how the tumour interacts with the immune system and the environment around it.

Breast cancer typically begins when genetic mutations in a cell cause it to grow uncontrollably. Over time, the tumour becomes a patchwork of cells, called clones, each with different mutations. Each clone may have a different reaction to treatments, or a different ability to metastasise. Which mutations occur is influenced by the tumour's microenvironment, the cells that surround it and the individual's immune system. The mechanisms driving this have remained elusive.

To give a complete view of breast cancer microanatomy, evolution and microenvironments, the team developed a new method centred on base-specific in situ sequencing (BaSISS) technology. It used hundreds of thousands of fluorescent molecular probes to target cellular DNA and RNA and scan large pieces of tissue using multiplexed fluorescence microscopy. This single-cell approach was combined with whole-genome sequencing. The multiple layers of data were linked by dedicated algorithms, resulting in detailed quantitative maps of multiple genetic clones.

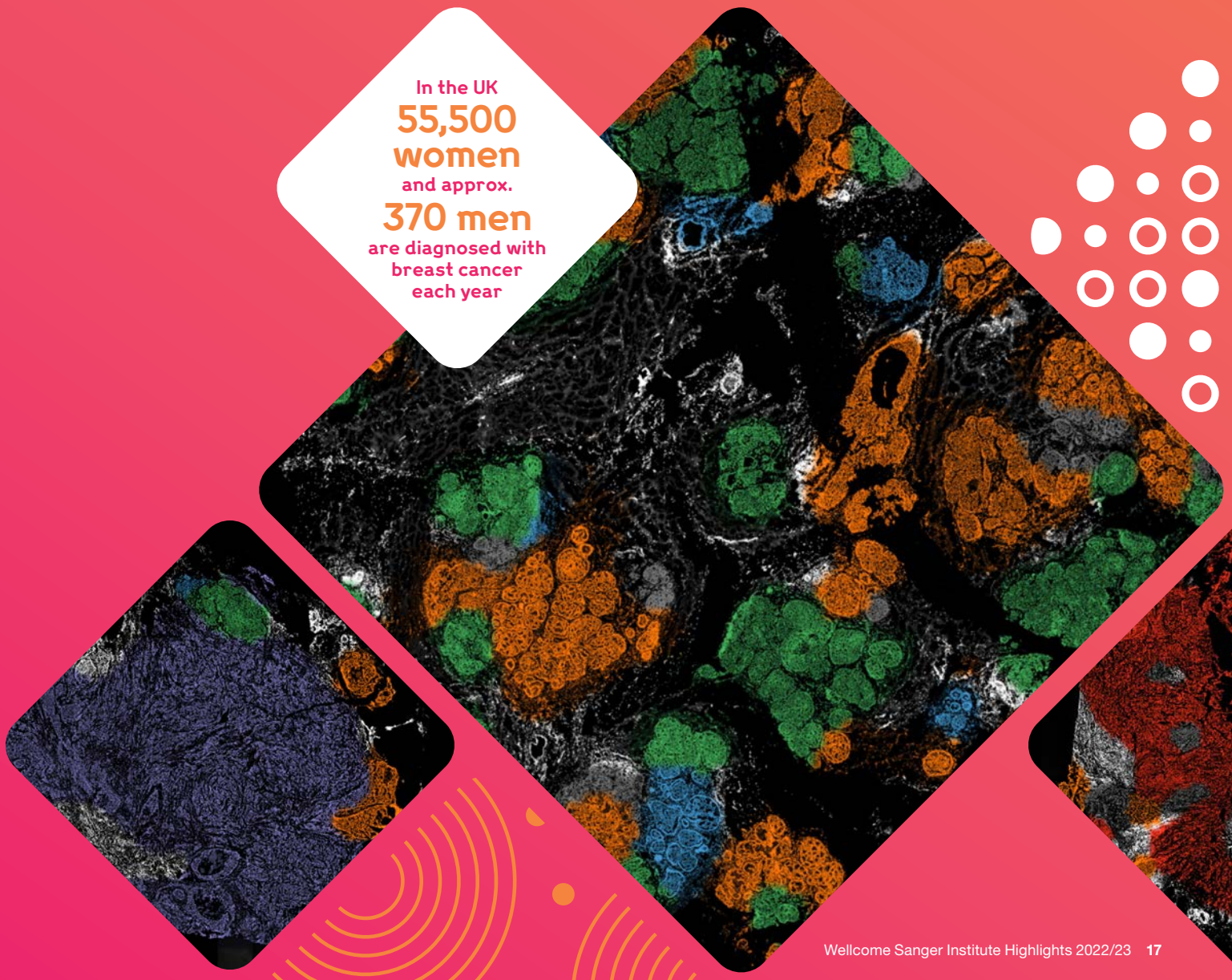
They used the data to study tumour evolution, uncovering specific patterns of clone growth across multiple stages of breast cancer development. They showed that genetic clones behave differently depending on where in the breast they started, suggesting that it is not only genetics that influence how cancers grow, but also their location.

The new technology combines multiple techniques and expertise, bringing together different approaches to give a complete view of cancer that has not been previously possible. It could be used to help answer some of the big questions in cancer, such as why some cancer cells spread, how treatment resistance is formed and why some therapies fail.

Reference

Lomakin A. *et al. Nature* 2022; **611:** 594-602.

In the UK
55,500 women
and approx.
370 men
are diagnosed with breast cancer each year



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.

In this section

1

A new way of identifying cancer cells

A new method of separating cancer cells from non-cancer cells has been developed by researchers at the Wellcome Sanger Institute, in a boost for those working to better understand cancer biology using single-cell mRNA sequencing. The method is openly available as a software package for researchers across the world to use.

Single-cell mRNA analysis of cancer cells is one of the leading-edge techniques being used to better understand cancer biology. The data generated can be used to try to disrupt cancers with drugs, or study how cancers arise.

A fundamental step in analysing single-cell data is separating cancer and non-cancer cells, but this is not always an easy task. Currently, the best method of doing this is to measure the average gene expression, or activation of cells in the sample, with higher or lower expression used to distinguish cancer cells from healthy cells. But this method can lead to false conclusions.

In this study, a team at the Sanger Institute performed whole-genome sequencing and single-cell mRNA sequencing on samples collected by Great Ormond Street Hospital.

By locating imbalances of different gene versions in these data, which indicate copy number changes in the genome, the team was able to identify cancer cells more reliably than with previous methods. This approach will primarily be useful for validating new cancer cell types and better understanding the microenvironment of tumour tissue.

The software, alleleIntegrator, is openly available for researchers around the world to apply to their own data, advancing the effectiveness of single-cell sequencing to understand cancer.

Reference

Trinh M.K. *et al. Communications Biology* 2022; **5:** 884.

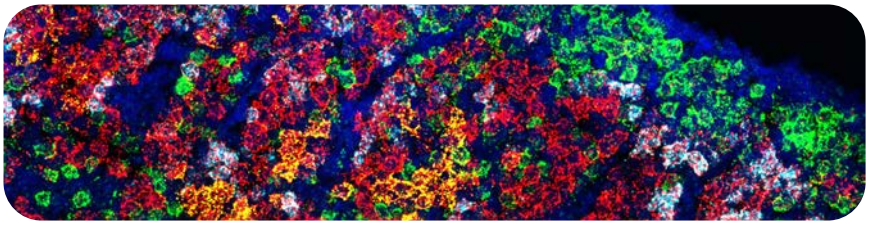
alleleIntegrator
sanger.ac.uk/tool/alleleintegrator

2

Cell map of sex determination could help fertility treatments

The first large-scale cellular map of human gonadal development has been created by researchers at the Wellcome Sanger Institute, as part of the Human Cell Atlas initiative to map all cell types in the human body. The research identifies new cell types, including those that express the 'sex determination' gene, which initiates the process that decides whether an individual will become phenotypically male or female. The map will help improve the growth of reproductive cells in fertility treatments and understanding of reproductive conditions.

Human embryo and fetal samples were obtained from the MRC and Wellcome-funded Human Developmental Biology Resource (HDBR).



The gonads play a key role in human development, as they determine biological sex before maturing into ovaries or testes. The early weeks of development are extremely dynamic, with cell types appearing and disappearing rapidly as their purpose is fulfilled – making it challenging to study. Most knowledge of gonadal development comes from mice, though it has been uncertain how much of this can be translated to humans.

To characterise the route that cells take to become either a testis or an ovary, researchers at the Sanger Institute created the first large-scale cellular map of gonadal development in both sexes. Around half a million cells from human gonadal tissue covering weeks six to 21 of pregnancy were analysed. The team used single-cell sequencing and multi-omics together with spatial techniques to disentangle the cellular and molecular programmes that mediate human gonadal development.

The team also generated a similar map in mice to understand where human and mouse biologies differ. Their analysis showed patterns of gene expression that are unique to humans and found the cell type that is the first to express the 'sex determination' gene and initiate the process to decide whether the gonad will become a testis or an ovary. Named Early Supporting Gonadal Cells (ESGCs), they peaked around six weeks after conception. ESGCs are present in both humans and mice, but their gene expression pattern is different in the two species.

The comprehensive cellular maps provide a unique resource to study gonadal function relevant to understanding infertility, differences in sex development and gonadal pathologies.

Reference

Garcia-Alonso L. *et al. Nature* 2022; **607:** 540-547.

3

Immune cell characteristics mapped across life

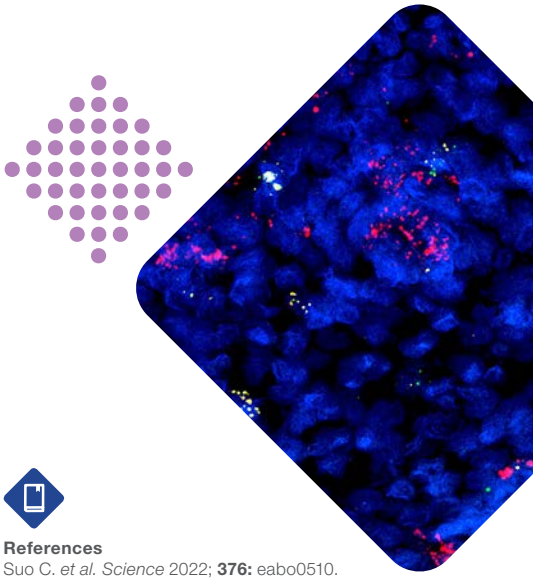
As part of the international Human Cell Atlas consortium, Sanger researchers have created open-access atlases of immune cells in the human body. Their work focuses on the early development of the immune system and the localisation of immune cells across tissues. They also studied immune cells in multiple tissues from adults, providing a framework for predicting cell types and insights into immunological memory.

The Human Cell Atlas (HCA) aims to map every cell type in the human body to understand health and disease. To create an atlas of the developing human immune system, Sanger Institute researchers used spatial transcriptomics and single-cell RNA sequencing to map the exact location of specific cells within developing tissues across nine organs. The team identified a new type of B cell and distinctive T cells that appear in the early stages of life. Their work is the first time the entire immune system has been mapped as a distributed network, and its development through time and space has been reconstructed.

The team also analysed immune cells from 16 tissues of 12 adult organ donors. Over one million cells were studied in total. They developed a database and algorithm that automatically classifies cells, called CellTypist, to handle the large volume and variation of immune cells. CellTypist identified around 100 distinct cell types.

The researchers then created a cross-tissue immune cell atlas to reveal the relationship between immune cells in different tissues. Tissue-resident immune cells are understudied, compared to those circulating in the blood. They found similarities in some families of immune cells, such as macrophages, and differences in others. For example, some memory T cells show unique features depending on which tissue they are in.

The atlases are freely available to others and may help to interpret and inform future studies. Knowing more about immune cells in tissues at different stages of life could help research into vaccines or anti-cancer treatments.



References

Suo C. *et al. Science* 2022; **376:** eabo0510.

Domínguez Conde C. *et al. Science* 2022; **376:** eabl5197.

CellTypist is publicly available, with user-friendly documentation at **celltypist.org**

Human developmental tissue samples were obtained from the MRC-Wellcome Trust–funded Human Developmental Biology Resource (HDBR). Tissue was obtained from deceased organ donors through the Cambridge Biorepository for Translational Medicine (CBTM) and from Columbia University.

4

First map of immune system wiring

Researchers have detailed the first full connectivity map of the human immune system, showing how cells communicate with each other. As well as a new understanding of the immune system, the work could lead to novel immunotherapies to treat cancer, infectious diseases and autoimmune conditions.

An in-depth understanding of the interactions between immune cells is vital for developing new treatments for disease.

Cell-surface proteins regulate immune cell activity by enabling cell signalling and structural adhesion. Such proteins are vital for homeostatic and disease processes – from tumour surveillance, to autoimmunity and infection control. Yet a full map of immune cell interactions had not previously been available.

To systematically survey interactions between immune cells, the Sanger Institute team developed a method for testing binary interactions of all possible protein pairings. Their method, the scalable arrayed multi-valent extracellular interaction screen (SAVEXIS), makes it possible to screen hundreds of thousands of interactions using minute amounts of protein.

They then created a near-complete set of 630 human immune cell surface proteins and screened them all against each other individually. The screen confirmed previously identified protein interactions and discovered new ones, increasing the number of known interactions in the human immune system by 20 per cent to over 100 interactions.

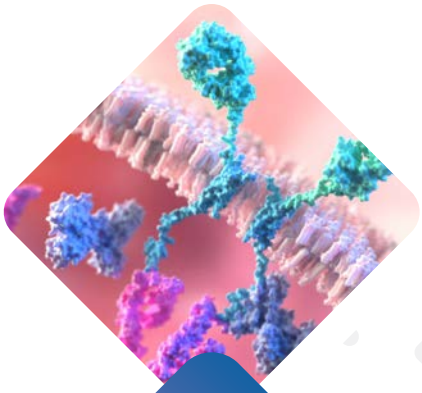
The team validated the new interactions and determined their biophysical parameters. Using multiplex microscopy, they linked the receptor interactions to functions in the immune system.

Combining their data with published data on cell protein expression, the researchers also built a mathematical model that predicts cellular connectivity from basic principles.

Their study represents the culmination of over two decades of work, and the result is a high-confidence and quantitative map of the 'wiring' that connects human immune cells. Freely available, the atlas of immune cell connections will enhance research into existing immunotherapies and provide the basis for developing new ones.

Reference
Shilts J. *et al. Nature* 2022; **608:** 397-404.

Atlas of immune interactions
sanger.ac.uk/tool/immune-interaction/immune-interaction



Researchers screened **630** human immune cell surface proteins

> This is a huge step in understanding the inner workings of the immune system and will hopefully be utilised by researchers all around the world to help develop new therapies that work with the body's defence mechanisms.
>
> **Jarrod Shilts**
> First author from the Wellcome Sanger Institute

5

Study sheds light on why immunodeficiency affects only one identical twin

Scientists have long queried the causes of immune disorders in only one of two identical twins. New research from the Wellcome Sanger Institute and the Josep Carreras Leukaemia Research Institute in Spain found the answer lies in both alterations in immune cell-cell communication and the epigenome, the host of biological processes that regulate how our genes function.

Common variable immunodeficiency (CVID) encompasses a range of immune disorders caused by a reduced ability to produce protective antibodies, which leaves the individual vulnerable to persistent or repeated infection. These individuals usually have low levels of immunoglobulins or antibodies, due to problems with the B cells that create them.

Though identical twins have the same genes, most will be born with a small number of genetic and epigenetic differences, and the number of variations will increase over their lifetime. But where one twin experiences a health problem that their sibling does not, in most cases genetic differences alone cannot explain why this has occurred.

To understand this, the researchers generated single-cell data to investigate epigenetic factors involved in CVID. Samples were taken from a pair of identical twins, with only one having CVID, as well as a wider group of CVID patients and healthy individuals.

Analysis of the identical twin participants found that not only did the sibling with CVID have fewer B cells, but that B cell defects resulted in epigenetic problems with DNA methylation, chromatin accessibility and transcriptional defects in memory B cells themselves. In addition, researchers found massive defects in the cell-to-cell communication required for the immune system to function normally.

The single-cell multi-omics datasets the team produced are publicly available through the Human Cell Atlas and give insight into future diagnosis and treatments of CVID patients.

Reference
Rodríguez-Ubreva J. *et al. Nature Communications* 2022; **13:** 1779.

6

Lung atlases uncover new cell types and immune defences

As part of the Human Cell Atlas initiative to map every cell type in the human body, researchers at the Wellcome Sanger Institute, EMBL's European Bioinformatics Institute (EMBL-EBI) and the Gurdon Institute at the University of Cambridge have examined which genes are activated in different stages of lung development, one cell at a time. They combined this with spatial technologies, which pinpoint the exact location of cells, to create the Developmental Lung Cell Atlas, showing how the respiratory system comes into being.

The origin of many diseases is during development, yet how the human lung forms in early life is underexplored. The early weeks of development are extremely dynamic, with cell types appearing and disappearing rapidly as the organ-building processes move between stages.

To understand the development of the lung, the team combined single-cell sequencing of early-stage cells with spatial technologies to generate an in-depth dataset of lung development. This resource describes which cell types are present in the developing lung architecture, and how these are regulated.

The team identified 144 cell types, such as intermediate and transitional cell types, including a subtype that could be linked to the development of human small cell lung cancer later in life.

The team used the atlas to make predictions about how lung cells develop, especially which genes are the key players driving this process. They then used organoid models to validate the emerging hypotheses, demonstrating that the atlas can be used to accurately predict the stages and cells involved in tissue development.

An additional study used the dataset to investigate neonatal lung disease, uncovering insights about the causal mechanisms and developing new organoid models to aid further research.

The freely available resource acts as a guidebook for healthy lung development and can be used as a baseline to investigate how lung diseases originate.

Sanger Institute researchers have also created the most comprehensive adult lung cell atlas to date, revealing 11 new cell types and detailed insights into an immune process involved in fighting respiratory infections.

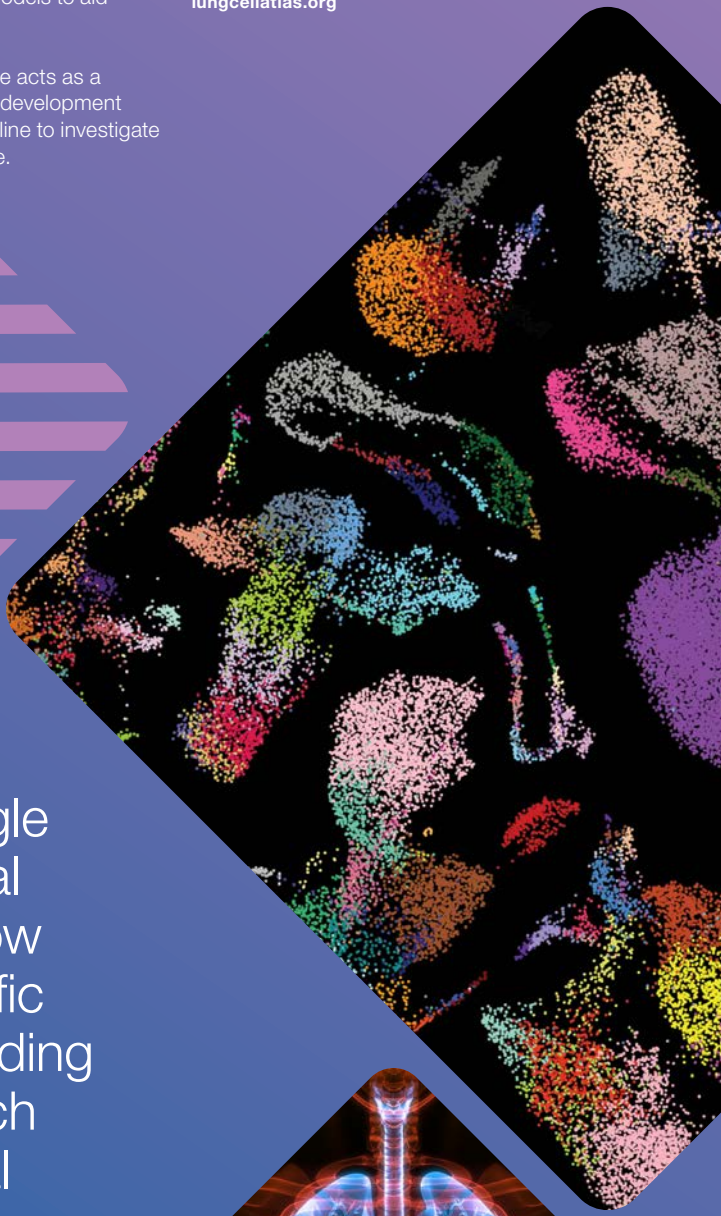
References

He P. *et al. Cell* 2022; **185:** 4841-4680.E25.

Lim K. *et al. Cell Stem Cell* 2022; **30:** 20-37.E9.

Angelidis I. *et al. Nature Communications* 2022; **10:** 963.

Lung Cell Atlas
lungcellatlas.org



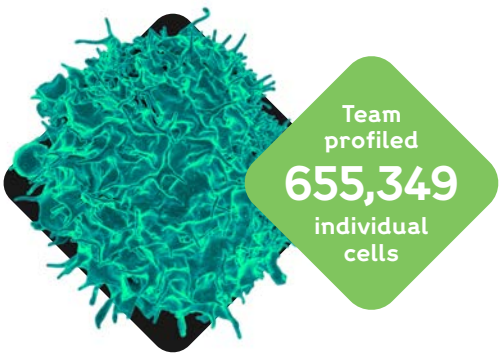
> Our research builds on the single cell era and, by including spatial data, we have begun to see how lung cells interact in their specific microenvironments. Understanding how lung cells interact with each other in a healthy lung is crucial if we hope to identify where something has gone wrong to cause disease.

Dr Kerstin Meyer
Principal Staff Scientist, Wellcome Sanger Institute

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.

In this section



1

Tracking T-cell activation reveals immune disease drug targets

In a first-of-its-kind experiment, 127 genes have been linked to immune diseases, providing newfound insights into the sequence and timing of gene activity during T cell activation, a key process in regulating the body's immune response. The study, led by researchers from Open Targets, provides key information to guide the development of new therapies for immune diseases such as rheumatoid arthritis, type-1 diabetes and Crohn's disease.

T cell activation is the first step in the immune system's response to infection. The cells also play a role in autoimmune diseases, as T cell malfunction can cause severe immune deficiencies. Mapping T cell activation at a molecular level is crucial to understanding where it can go wrong, and at which points therapeutic interventions can influence the process.

Researchers at the Wellcome Sanger Institute and GSK profiled 655,349 individual cells using single-cell RNA sequencing technology from 119 individuals. They identified 38 distinct cell subtypes and mapped the timing of gene activity for each cell subtype in the T cell activation process. They identified genes regulated by variations in DNA at different time points during T cell activation.

By comparing their data with known genetic variants predisposing to development of 12 immune diseases, they found 127 genes associated with those conditions. Some of these only manifested at specific time points giving the first-of-its-kind insight into regulation of T cell activation in immune disease.

By cataloguing the genes involved in T cell activation, this work provides the first step to a deeper understanding of immune processes, and how they go awry in diseases. The approach and data generated in this study can also be used to identify genes involved in other disorders. Open Targets aims to provide genetic evidence for emerging new drug targets, increasing the success of future therapies. Through close collaborations with industry partners, these data directly informed potential new pharmaceutical targets.

Reference
Soskic B. *et al. Nature Genetics* 2022; **54:** 817-826.

2

Sepsis and COVID-19 patients most at risk predicted with genetic model

A new model for understanding which patients with sepsis, COVID-19 and influenza are more likely to suffer poor outcomes has been developed by researchers at the Wellcome Sanger Institute, the University of Oxford, Queen Mary University, Imperial College and their collaborators.

Sepsis is caused by a dysregulated immune response to infection or injury, leading to an estimated 11 million deaths per year globally. But it has been difficult to predict who will get sepsis, who will recover and who will have poor and sometimes fatal outcomes, such as organ dysfunction. Timely identification of patients at risk of sepsis is important, as it could help physicians manage clinical care. Despite hundreds of clinical trials aimed at improving sepsis outcomes, there are currently no targeted treatments.

To change this, the researchers developed a model to understand which patients with sepsis are more likely to have particular immune responses and potentially poor outcomes.

The team used RNA-seq technologies to study blood samples from 1,655 sepsis patients, to identify which genes were activated, or expressed. The resulting data were then combined with existing data from sepsis patients and healthy individuals.

Analysis identified patterns of gene expression signifying an inappropriate immune response, allowing researchers to predict clinical outcomes from the expression pattern of just 19 genes.

To gauge whether the 19-gene model could also be applied to other diseases, a machine-learning framework was developed to test it on sepsis, SARS-CoV-2 and influenza. The model was able to successfully predict an individual's likelihood of poor outcomes for all three diseases.

Their method enables early identification of those with a dysfunctional immune response, bringing precision medicine techniques to infection. The next steps will be to develop biomarker-led clinical trials, with the goal of targeting therapies at those who would benefit most.

Reference
Cano-Gomez E. *et al. Science Translational Medicine* 2022; **14:** eabq4433

3

Genetic risk factors for heart disease vary between populations

Researchers at the Wellcome Sanger Institute have shown that genetic risk factors for heart disease vary between populations with different ancestries. Their findings could help improve risk prediction and help guide the use of preventive therapies in the clinic. The study contributes to the increasing representation of individuals of diverse ancestry and varying socio-economic status in research studies and aims to help decrease health disparities.

Individuals with South Asian ancestry have a higher risk of coronary artery disease (CAD) compared to individuals with European ancestries. However, the genetic basis of CAD risk is not well characterised in South Asian ancestry populations, because genome-wide association studies have been mostly carried out in those with European ancestry.

Common genetic variation is an important determinant of CAD and related risk factors such as blood pressure, lipids and body mass index (BMI). The genetic component of disease risk can be used to identify underlying disease genes and pathways, to estimate the effects of risk factors, to improve risk prediction through the application of polygenic scores and prioritise different prevention strategies or drug targets.

To determine how the genetic determinants of cardiometabolic traits are shared by European and South Asian ancestry populations, the team performed a comparative analysis. They studied data from the Genes & Health (G&H) cohort, a community-based study of 22,490 British Pakistani and Bangladeshi individuals who have donated health and genetic information. This unique cohort represents an understudied and clinically vulnerable population with high levels of socioeconomic deprivation.

The team applied new approaches to the transferability of genomic risk loci across populations, performed ancestry-specific and trans-ancestry Mendelian randomisation analysis and investigated the transportability of polygenic scores for CAD and its risk factors.

For lipids and blood pressure, they found that causal genetic variants at published loci are widely shared with European ancestry populations. The prediction accuracy of polygenic scores for these traits was similar between G&H and European ancestry samples. However, the predictive performance of BMI and CAD polygenic scores was reduced, and CAD also had fewer transferable loci.

The team optimised the analysis for South Asian ancestry individuals, and this showed an improvement in risk reclassification when combined with a clinical risk score. Adding polygenic scores to conventional risk factors in the prediction of CAD in primary care could improve the efficient use of preventive interventions, such as lipid-lowering medications.

The team's new approach can serve as methodological standards for this type of work. Their investigation contributes to the increasing representation of individuals of diverse ancestry and varying socio-economic status in research studies, with the hope that this will help to decrease health disparities.

Reference
Huang Q.Q. *et al. Nature Communcations* 2022; **13:** 4664.

4

Study zeroes in on genes involved in Crohn's disease

An international consortium, led by researchers from the Wellcome Sanger Institute and the Broad Institute, identified genetic variants in 10 genes that elevate a person's susceptibility to Crohn's disease, a form of Inflammatory Bowel Disease. The study is the largest to date to focus on rare genetic variants associated with Crohn's disease.

Crohn's disease is a debilitating condition characterised by chronic inflammation of the gastrointestinal tract. The causes of the disease are poorly understood, but it is believed to be triggered by a hyperactive immune response against gut bacteria in genetically susceptible individuals. Though drugs are available that improve symptoms for many patients, there is no cure and relapsing bouts of severe illness are common.

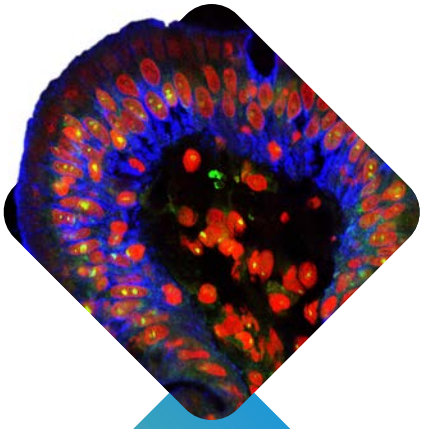
Genome-wide association studies (GWAS) have previously identified hundreds of genetic loci associated with Crohn's disease, but robust identification of the genes dysregulated by these variants has been challenging.

To complement GWAS studies and find actionable drug targets for Crohn's disease, the team performed exome sequencing on around 30,000 patients with Crohn's. These data were compared to exome sequences from around 80,000 individuals without the condition.

They found 10 genes directly implicated in Crohn's disease. This included genetic variation within six genes in regions of the genome that had not been previously connected to Crohn's disease. Several of these genes are known to play important roles in a type of stem cell in the gut called mesenchymal cells, suggesting that disruption of these cells contributes to the initiation and maintenance of intestinal inflammation.

The research highlights the causal role of mesenchymal cells in intestinal inflammation, helping to zero in on the genetic roots of inflammatory bowel disease and providing better data to help develop the next generation of treatments.

Reference
Sazonovs A. *et al. Nature Genetics* 2022; **54**: 1275-1283.



Researchers found **10** genes by analysing 120,000+ genomes

5

Increased mutations can be traced back to mistakes in father's sperm

Researchers have traced the cause of increased numbers of genetic mutations in children to a higher rate of random mutations in sperm cells of the biological father, associated with rare genetic defects in DNA repair or chemotherapy.

The average number of new genetic mutations, generating single-nucleotide variants (SNVs) in the genome, is estimated to be 60–70 per human genome per generation. However, little is known about individuals with hypermutation, where an unusually large numbers of mutations are present. Hypermutation is rare, but it increases the risk of a child having a rare genetic disorder.

To investigate, scientists from the Wellcome Sanger Institute and their collaborators analysed genome-wide sequences of 21,879 families with rare genetic diseases.

The team identified 12 children with between two to seven times more mutations than the general population. By analysing the patterns of mutations, known as mutational signatures, the team were able to trace the cause of the mutations. For eight of these children, the excess mutations could be linked to increased mutations in the sperm of the biological father. Two of the fathers had rare recessive genetic variants that impaired DNA repair mechanisms. Four had been treated with certain types of chemotherapy earlier in life.

However, most fathers and all mothers who had received chemotherapy prior to conceiving a child did not have children with a notable excess of mutations.

Overall, the results suggest that sperm and egg cells are well protected from mutagenic effects, hypermutation is rare, the number of excess mutations is relatively modest and most individuals with a hypermutated genome will not have a genetic disease.

This study exemplifies the value of linking nationwide genetic data and routine clinical records in secure, anonymised and trustworthy ways to provide unique insights into unanticipated, but important, questions.

Reference
Kaplanis J. *et al. Nature* 2022; **605:** 503-508.

6

Evolutionary pressures on genes associated with childlessness

A study by Wellcome Sanger Institute researchers suggests genetic variants that damage the genome are associated with reduced reproductive success and an increased likelihood of not having children. However, this genetic link may play a very minor role in the overall likelihood of being childless – less than 1 per cent – when compared to more influential factors such as sociodemographic factors and choice.

Some genetic variants, caused by mutation, can damage the genome and lead to neurodevelopmental disorders.

The damaging genetic variants that are most likely to be linked to disease are in the genes that exhibit a marked depletion of damaging genetic variation in the general population.

Sanger Institute researchers investigated the evolutionary processes, such as reproductive success, that might be leading to the depletion of damaging genetic variation in this subset of genes, with the hope of finding novel genetic causes of neurodevelopmental disorders. The team did not initially set out to study the factors that influence childlessness.

The researchers used data from more than 340,000 UK Biobank participants to calculate how much damaging genetic variation each person carried across their entire genome – their 'genetic burden'.

They found that men with the highest genetic burden had on average fewer children, and that increasing genetic burden was associated with a higher chance of being childless, more so in men than women.

When investigating the reasons behind this, the team found that a higher genetic burden was not associated with infertility and was unlikely to increase the risk of health conditions that could impact on reproductive success.

However, the researchers did find that men and women with a higher genetic burden were more likely to have mental health disorders or lower household incomes and were less likely to perform well on cognitive tests or have a university degree.

Men with a higher genetic burden were less likely to have a partner at home compared to women. The team concluded that the effect of damaging genetic variants on cognitive and behavioural traits may reduce these men's chances of finding a partner with whom to have children. Their results suggest that these damaging genetic variants are likely being removed from the general population under negative selection, partly in a manner that is consistent with Darwin's theory of sexual selection.

However, the genetic association with childlessness cannot be used to predict who will or will not remain childless, as it explains less than 1 per cent of the overall likelihood of childlessness. Other factors, such as infertility, socioeconomic status and choice are much more likely to be associated with not having children.

Reference
Gardner E.J. *et al. Nature* 2022; **603:** 858-863.

It's important to emphasise that we have not found a 'gene for childlessness', as that implies a strong, causal effect of genetic variation on whether or not someone will have children. Instead we have shown that people with damaged genomes, particularly men, are slightly more likely to be childless. This is probably due to the effect of damaging genetic variants on cognitive and behavioural traits, which make these men less likely to find a partner to have children with.

Dr Eugene Gardner
First author previously from the Wellcome Sanger Institute, now based at the MRC Epidemiology Unit at the University of Cambridge

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.

In this section

1

Pathogen survey reveals pockets of resistance

Unparalleled insights into the secret life of *Streptococcus pneumoniae*, the bacterium responsible for hundreds of thousands of infant deaths each year, have been revealed by researchers at the Sanger Institute, the University of Oxford, the Shoklo Malaria Research Unit and Imperial College London.

S. pneumoniae causes diseases ranging from ear infections through to pneumonia, septicaemia and meningitis. Around nine million people are infected every year, with elderly adults and children particularly susceptible. More than 300,000 children die from pneumococcal infection each year.

Not all individuals colonised by *S. pneumoniae* will become ill – the bacterium is carried without symptoms by up to 60 per cent of children and by a smaller percentage of adults. Children are generally colonised by common strains, allowing the immune system to recognise and guard against them. Adults tend to be colonised by rarer strains that may not have been encountered before.

Researchers set out to document the diversity of *S. pneumoniae* in children and adults. Almost 4,000 samples, collected from infants and mothers over a two-year period, were whole genome sequenced to test whether the full diversity of *S. pneumoniae* lineages present could be mapped.

The team showed that deep sequencing has unsurpassed sensitivity for detecting colonisation with multiple strains, doubling the rate at which highly invasive serotype 1 bacteria were detected compared with gold-standard methods.

The greater resolution identified an elevated rate of transmission from mothers to their children in the first year of the child's life. The team also studied treatment data and found that infants were at an elevated risk of both the acquisition and persistent colonisation of a multidrug-resistant bacterium after antimicrobial treatment.

These results highlight the benefits of deep sequencing for the genomic surveillance of bacterial pathogens.

Reference

Tonkin-Hill G. *et al. Nature Microbiology* 2022; **7:** 1791-1804.

2

Genomics empowers vaccine makers to tackle shapeshifting bacteria

A pioneering genomic surveillance study has provided the clearest picture yet of the arms race between *Streptococcus pneumoniae* and the vaccines designed to protect against the most dominant types. A bacterial strain called GPSC10 was found to be a particular threat, due to its increased virulence, ability to transform its structure to evade vaccines and its resistance to several common antibiotics.

Since 2000, a series of pneumococcal vaccines, called PCV-13, have been deployed that have targeted up to 13 *S. pneumoniae* serotypes, responsible for most disease in infants, resulting in a reduction in disease worldwide. However, there are more than 100 distinct serotypes. Knowing which serotypes to target with the vaccines, and what the likely impact will be on disease and the wider pneumococcal population, is vital when designing effective global vaccination strategies.

Scientists from the Wellcome Sanger Institute performed whole-genome sequencing on 510 samples of *S. pneumoniae* serotype 24F from Europe. 24F has been on the rise across the world. To provide a global comparison, an international collection of other *S. pneumoniae* genomes was added from the Global Pneumococcal Sequencing project database.

Analysis showed that 24F was present in many countries largely due to the spread of three strains: GPSC10, GPSC16 and GPSC206. One strain in particular, GPSC10, was responsible for the rapid increase in 24F in France around four years following the introduction of PCV-13. It was found to have high disease potential and be resistant to multiple antibiotics. Perhaps the biggest concern arising from the study was GPSC10's ability to express 17 different serotypes, only six of which are included in current vaccines.

The findings demonstrate the value of genomic surveillance to inform vaccine design and highlight the challenge posed by 'shapeshifting' strains like GPSC10.

Reference
Lo S.W. *et al. Lancet Microbe* 2022; **3:** e735-e753.

3

Healthy newborns in the UK are not colonised by drug-resistant hospital bugs

Researchers from the Wellcome Sanger Institute, the University of Helsinki, the University of Oslo and collaborators analysed the gut bacteria of more than 300 newborn babies from the UK Baby Biome study. They mapped the bacterial strains present and their interactions, unravelling the gut colonisation process. The research shows that the babies' gut bacteria did not contain multi-drug resistant strains of common hospital pathogens, such as *Enterococcus faecalis* and *Klebsiella pneumoniae*.

The human gut contains thousands of bacterial species and other microorganisms, known as the microbiome. Newborn babies are colonised very quickly, with the majority of bacteria being beneficial. However some species, including *E. coli*, have the potential to cause infections if they enter the bloodstream. Multi-drug resistance (MDR) has recently become a frequent feature of *E. coli* infections and is also seen in *Klebsiella pneumoniae* and *Enterococcus faecalis*, which can cause symptoms from urinary tract infections to meningitis.

The team analysed newborn microbiomes through faecal samples in a cohort of 300 babies, sampled over time. They used new, high-resolution metagenomics, combined with high-precision genomic reference libraries to assemble single genomes and identify the bacterial strains present in each sample. They found strong inter- and intra-species competition dynamics in the gut colonisation process, but also synergistic relationships among several *Klebsiella* species. In the future, it might be possible to enhance the microbiome to outcompete drug-resistant or pathogenic strains.

Multi-drug resistant strains of common pathogenic bacteria were not found in healthy newborn babies, despite some of them staying at the hospital for multiple days.

The researchers also combined the high-resolution colonisation profiles with data from previous work, such as the Baby Biome study, to estimate the relative virulence of different strains of *E. coli*. This is the first time that it has been possible to derive a population-based estimate of how virulent a particular strain of *E. coli* is, allowing researchers to focus on understanding and preventing the most immediate threats.

Reference
Mäklin T. *et al. Nature Communications* 2022; **13:** 7417.

Our finding that babies leave the hospital mostly without multi-drug resistant bacteria shows that these strains cannot colonise our guts at all stages of life. Understanding what protects healthy babies against this could lead to a new way of treating infections.

Dr Trevor Lawley
Group Leader, Wellcome Sanger Institute

4

A new way to evaluate malaria drug resistance

Wellcome Sanger Institute researchers have developed a new way to better understand the complex interplay of drug resistance and overall fitness in malaria parasites. To accelerate the elimination of malaria, new drugs will be needed to counter the current wave of emerging multidrug resistant parasites. The ability to experimentally observe how drug exposure interacts with particular genotypes to change overall parasite fitness will allow researchers to better predict how new treatments will perform.

The repeated emergence of antimalarial drug resistance in *Plasmodium falciparum*, the parasite which causes the most deadly form of malaria, is an ongoing problem for malaria control. Malaria kills an estimated 619,000 people per year, mostly in sub-Saharan Africa. *P. falciparum* has continually evolved to evade treatments, including the latest front-line drug, artemisinin. The emergence of this resistance is consistently seen in certain regions of the world and has been linked to the genetic makeup of the parasite populations there.

Next-generation DNA sequencing has been used to identify variations in genes associated with resistance. However, there is still a need for sensitive and accurate laboratory techniques to quantify the impact of drug resistance acquisition on parasite fitness.

The Sanger Institute team has developed a scalable method to provide this information. Their method uses CRISPR/Cas9 genome-editing to insert a unique DNA barcode sequence into a non-essential parasite gene *(pfrh3)* in *P. falciparum* strains. The strains were then pooled and assessed using competitive growth assays under different conditions and drug exposures. Their relative proportions were quantified using a single PCR, followed by next-generation sequencing.

The team's experiments confirmed the impact of known mutations associated with drug resistance. They also showed that the method can be a powerful approach for tracking artemisinin response, as it can identify an artemisinin-resistant strain within a mix of multiple parasite lines.

Their approach will help other researchers understand the interactions between drug resistance and parasite fitness, factors in the evolution of resistance to artemisinin, how drug resistance spreads and the impact of parasite genetic backgrounds on transmission.

An estimated **619,000** people die from malaria each year

Reference
Carrasquilla M. *et al. mBio* 2022; **13:** e0093722.

5

Newly identified bacteria in mice impact on Inflammatory Bowel Disease research

Researchers from the Wellcome Sanger Institute, the Hudson Institute of Medical Research, Australia, the University of Cambridge and collaborators have discovered two new strains of bacteria in the microbiome of mice that cause Inflammatory Bowel Disease (IBD) symptoms. Their study shows that the bacteria are commonly found in mice that are used to study IBD and could impact on the results of research into the condition.

Inflammatory Bowel Disease (IBD) is a chronic condition that impacts 6.8 million people globally each year. It occurs when there is longstanding inflammation of the gastrointestinal tract and can cause debilitating symptoms.

While the exact cause of IBD is unknown, it has been suggested that in some individuals the immune system reacts to the naturally occurring bacteria in the gut, highlighting the importance of studying the microbiome in understanding and treating this condition.

Research into IBD relies heavily on using the 'dextran sodium sulphate model' mouse, yet knowledge of the mouse microbiome remains limited. It is important to understand the mouse microbiota, its impact on disease and its variation across the world, in case this influences study results.

To identify and characterise bacteria in the mouse gut, the team analysed the physical characteristics and the microbiome of 579 genetically identical mice. Their large-scale data-driven approach identified two new bacteria that were driving weight loss and intestinal inflammation in mice: *Duncaniella muricolitica* and *Alistipes okayasuensis*.

By using data from previous work that catalogued 26,640 mouse microbiome bacteria, they found that both new species are common in mouse colonies around the world.

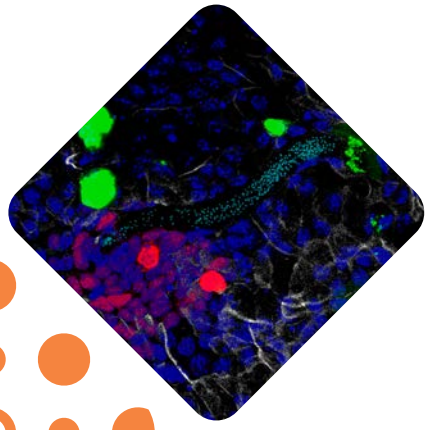
As these bacteria cause IBD symptoms, they impact on the outcomes of mouse model studies of IBD. The researchers suggest the presence of these bacteria should be noted in designing and interpreting IBD studies in mice.

References
Forster S.C. *et al. Nature Microbiology* 2022; **7:** 590-599.

Beresford-Jones B.S. *et al. Cell Host and Microbe* 2022; **1:** 124-138.e8.

6

‘Mini guts’ used to study whipworm



Worldwide up to
500 million
people are affected by whipworm

By designing a new organoid ‘mini gut’ model, scientists from the Wellcome Sanger Institute have identified what happens in the body in the early stages of infection with *Trichuris trichiura*, the parasitic whipworm. The research could help in the design of future treatments and vaccines for trichuriasis, a neglected tropical disease.

Trichuriasis, caused by infection with the whipworm parasite *Trichuris trichiura*, affects up to 500 million people worldwide. Chronic whipworm infections can cause a range of debilitating gastrointestinal issues, nutritional deficiencies and are linked to delays in physical and cognitive development, especially in children. There is no vaccine, and treatments are not fully effective. Until now, it has been unclear how the worm larvae invade mammalian epithelial cells and establish multicellular infections in the gut.

To study early infection, the team developed ‘mini gut’ organoids. These are derived from mouse gut cells but grown in the lab. Mice are naturally infected by a related whipworm species and so represent a good model for host parasite interactions, which are challenging to study in people.

This is the first time that this organoid model has been used to study multicellular parasites and paves the way for similar research using human whipworm.

The team also used single-cell RNA sequencing on the cells from whipworm-infected mice to uncover the body’s response to the worm tunnelling into the gut lining. They found that infection results in cell damage and an expansion of a specific type of immune cells, potentially triggering the host immune response.

Their study unravels the gut invasion by whipworms, showing how the larvae initially degrade mucus layers to access epithelial cells, before becoming intracellular in multiple, live, dividing cells.

This knowledge of specific host-parasite interactions could be used to develop new treatments or vaccines.

Reference
Duque-Correa M.A. *et al. Nature Communications* 2022; **13:** 1725.

7

First genetic map of one of humans’ oldest parasites

Researchers have published the largest genomic analysis of the whipworm *Trichuris trichiura*, giving the most in-depth evolutionary insight about this human parasite to date. The research, the result of an international collaboration including scientists from the Wellcome Sanger Institute, the University of Copenhagen and others, gives a global set of genetic data to base new disease-fighting approaches on.

Caused by infection with the parasitic whipworm, *Trichuris trichiura*, trichuriasis disease is common in tropical and subtropical regions. However, parasite eggs found in fossilised remains show that the worm was once globally distributed.

To study the parasite’s evolution, the researchers performed whole-genome DNA sequencing of worms collected from both humans and primates around the world, in the first population genomics study of *T. trichiura*. The genomes of modern worms were compared to genome data from ancient samples obtained from archaeological dig sites, dating up to a thousand years old, primarily from latrines in Viking settlements in Denmark.

These samples are the oldest parasitic worm samples from which whole-genome sequencing data have been generated and are suggested to be the oldest eukaryotic pathogens, providing a unique insight into the parasites of humans’ past.

A comparison of past and recent parasites showed how populations of geographically distributed worms are related and the likely influence of human migration in the global spread of modern-day parasites. The team also confirmed that the worm has a zoonotic reservoir in primates, with the potential to transmit between primates and humans, which is important information for managing the parasite.

The study provides a global set of genetic data that could be used to inform strategies to reduce the spread of disease and help track treatment resistance.

Reference
Doyle S.R. *et al. Nature Communications* 2022; **13:** 3888.

8

Drug resistance genes mapped in parasitic worms to tackle livestock loss

Researchers at the Wellcome Sanger Institute, the University of Glasgow and the Moredun Research Institute have mapped the genes linked to drug resistance of the parasitic worm, *Haemonchus contortus*, for the first time. The team identified genetic variants leading to resistance to three of the most important drugs used for parasitic worm control, pinpointing how drug resistance emerges and providing crucial data for the tracking of resistant variants in the field. The research lays the foundations for understanding how drug resistance arises and, most importantly, how it can be controlled.

Parasitic worms, or helminths, infect people, livestock and pets, with over a billion people and countless animals requiring drug treatment to control infections each year. Left untreated, infections can result in elephantiasis or river blindness in humans and result in significant production loss and death in livestock.

Global efforts to limit the impacts of helminths are underway, but widespread resistance to anthelmintic drugs, particularly in worms that infect livestock, means that in many places certain drugs are ineffective. This adds to the huge costs of helminth infections – already responsible for annual livestock production losses of €686 million in Europe alone.

Although drug resistance is not yet widespread in human-infective helminths, it is a growing threat, particularly in low- and middle-income countries.

To uncover the genetic basis of drug resistance in the barber's pole worm, *Haemonchus contortus*, the researchers crossed drug-susceptible and multidrug-resistant strains of the parasite, followed by drug treatment using either fenbendazole, levamisole or ivermectin.

Whole-genome sequencing of the parasites was undertaken at the Sanger Institute. Analysis of the parasite genomes before and after drug treatment, together with genomic data from parasites sampled around the world, uncovered a small number of new and known genes implicated in drug resistance.

The fact that so few genes are involved in drug resistance is promising, as it will allow rapid development of new tools to detect and track resistant strains. Work is already underway to create rapid diagnostic tests and genomic sequencing methods for use in the field.

Reference
Doyle S.R. *et al.* *Cell Reports* 2022; **41:** 111522.

9

How antibiotic resistance spreads in the human microbiome

Wellcome Sanger Institute scientists have identified mobile genetic elements that transfer antibiotic resistance between phyla of bacteria in the human gut. Their work establishes the antibiotic resistance genes dissemination network between bacteria that aid human health in the gut, and pathogens.

Humans are colonised by microbial communities dominated by bacteria from the Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria phyla, constituting the microbiome. These play essential roles in regulating the immune response and digestion and providing protection against colonisation by pathogens. Antibiotic treatment, while intended to eliminate pathogens, can also eradicate the harmless, commensal bacteria, vastly altering the microbiome. However, commensal species with intrinsic or acquired antibiotic resistance are protected from elimination.

Following antibiotic treatments, antibiotic resistance genes may accumulate in commensal bacteria, acting as a reservoir from which they may be transferred on mobile genetic elements (MGEs) or by bacteriophage transduction to other species, including pathogens, via horizontal gene transfer.

The nature of horizontal gene transfer events between human gut commensals and pathogens has remained poorly characterised until now. To understand the process better, across the entire human microbiome, the Sanger Institute team systematically compared the genomes of 1,354 cultured commensal strains from 540 species to 45,403 pathogen strains from 12 species. This utilised the team's previous work, which demonstrated that the vast majority of the human gastrointestinal bacteria can be cultured, as well as publicly available genome sequence information.

They computationally identified 64,188 MGE-mediated antibiotic resistance gene transfer events between the two groups. These were consolidated and found to represent 5,931 MGEs, 15 of which were predicted to have crossed different bacterial phyla while also occurring in animal and environmental microbiomes.

The team then validated their findings experimentally and showed that these MGEs can mobilise from commensals *Dorea longicatena* and *Hungatella hathewayi* to pathogen *Klebsiella oxytoca*, crossing phyla simultaneously.

The study highlights the need to focus on tracking, managing and limiting antibiotic resistance gene associated MGE dissemination.

Reference
Forster S.C. *et al.* *Nature Communications* 2022; **13:** 1445.

46,757
bacterial strains studied





10

Human respiratory viruses will be tracked in new initiative

Building on work to sequence SARS-CoV-2, a new, world-leading initiative at the Wellcome Sanger Institute will lay the groundwork for large-scale genomic surveillance of respiratory viruses including influenza virus, respiratory syncytial virus (RSV), adenovirus and rhinovirus. The initiative aims to develop the capability for routine genomic surveillance of respiratory viruses and ultimately contribute to global efforts to establish pathogen genomics for routine public health and research. At the same time, the work will address some of the gaps in our basic knowledge about respiratory infection and health and will monitor for emerging pathogens.

Building on the technology and methodology that have been developed at the Sanger Institute for genomic surveillance of SARS-CoV-2 during the pandemic, scientists will initially establish combined genome sequencing of SARS-CoV-2, influenza, respiratory syncytial virus (RSV) and other common respiratory viruses in a single test. The ultimate goal is to determine all of the species and genes – including viral, bacterial and fungal species – present in a single nose swab sample, using a metagenomic approach.

The techniques will be developed using residual material from diagnostic COVID-19 swab samples. The results will give a baseline of respiratory virus dynamics in the UK and will generate an extensive viral genome dataset that will be made publicly available.

The team will also develop the methodology required to enable a routine surveillance programme for respiratory pathogens, with the capability to provide actionable data to inform public health decisions. This represents a world-leading initiative, as routine genomic surveillance of common respiratory viruses do not currently exist.

Alongside genomic surveillance, researchers at the Sanger Institute and collaborators will study the data to better understand the transmission and evolution of respiratory viruses, as well as seeking to identify new viruses and potential pandemic threats. Better understanding of which pathogen strains are in circulation will help to generate new vaccines and ensure existing ones are likely to be protective.

Scientists will also undertake work into the dynamics of the respiratory microbiome: all of the organisms that reside in our nose, throat and lungs. The team will study how the microbiome changes during infection, and how this relates to illness severity. The metagenomic data generated will enable researchers to track antimicrobial resistance, for example in methicillin-resistant *Staphylococcus aureus* (MRSA).

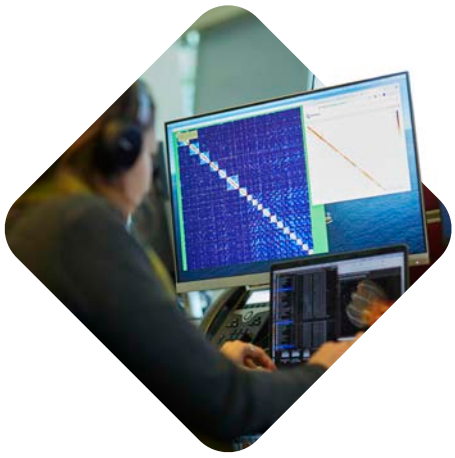
The data, protocols and methods generated by this project will be made freely and publicly available. The aim is to scale up and adapt the platforms, technologies and methodology for use around the world.

Respiratory Virus and Microbiome Initiative
sanger.ac.uk/group/respiratory-virus-and-microbiome-initiative

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.

In this section



1

Europe's drive to reverse biodiversity loss

Sanger Institute scientists have joined European experts in Biodiversity Genomics Europe, a project to tackle the biodiversity crisis using genetic data. The new pan-European consortium brings together experts in DNA barcoding and genome sequencing to form the foundations for using DNA-based techniques to protect biodiversity.

One in four species are currently threatened with extinction, putting livelihoods, food supplies and essential water and nutrient cycles at risk. Yet an estimated 80 per cent of the world's species have not been discovered. Even for described species, telling them apart is often difficult. Interactions within and among species, and between species and their environment create a hugely complex picture.

Biodiversity Genomics Europe (BGE) is a new €20 million initiative funded by Horizon Europe and UK Research and Innovation that brings together 33 organisations across 20 countries. Sanger scientists helped steer the successful funding application and continue to sit on the BGE coordination team. The aim is to map the interdependencies between species, predict how individuals and groups may respond to environmental change and reverse biodiversity loss.

The BGE project also links to global initiatives. One of these, BIOSCAN, aims to monitor ecosystems using DNA barcodes – short sequences of DNA code that can be used to quickly identify a species.

BGE also links to the European Reference Genome Atlas (ERGA) and the global Earth BioGenome Project (EBP), which aims to sequence genomes for the entire diversity of eukaryotic life. The Sanger Institute is one of five European sites contributing to the ERGA and the EBP through DNA sequencing to produce hundreds of high-quality reference genomes of at-risk European species.

Together, these ambitious global projects will form the foundation for a future bio-surveillance system across the planet.

Biodiversity Genomics Europe
biodiversitygenomics.eu

2

New orca genome

Researchers from the Wellcome Sanger Institute and the Norwegian University of Science and Technology have provided a high-quality, chromosome-level reference genome for the apex ocean predator *Orcinus orca*, the killer whale.

The killer whale, *Orcinus orca*, is the largest and most widely geographically distributed species of dolphin. Killer whales are found from the Arctic to the Antarctic, though in the greatest densities at high latitudes. In these areas, they have evolved into different 'ecotypes'. The best studied of these are fish-eating and mammal-eating ecotypes of the North Pacific; these populations differ in diet, behaviour and social structure.

Previous studies have shown significant genetic differentiation between these two ecotypes. The new high-quality genome sequence will aid studies into the divergence of the two populations, which remains a research focus.

Genomic studies are also providing insights into the fitness and health of populations through the estimation of mutation load and inbreeding. For example, previous studies had found a small population of killer whales in UK waters are not able to reproduce, consistent with high inbreeding. However, other populations of killer whales in UK waters are surviving, including groups that migrate each summer from Iceland to Scotland, where they hunt seals close to shore. Genomic resources are key to monitoring their health, in terms of inbreeding.

To improve the existing draft genome assembly of the killer whale, the researchers obtained a blood sample – taken during health checks of a rescued animal. The sample was used to generate Hi-C data and single molecules of DNA suitable for long-read sequencing. Together, these data were used to create a high-quality genome sequence of the animal.

The resulting genome assembly has provided the backbone for several studies which have contributed to the understanding of wild killer whale populations. It will usher in a new phase of genomics research, including identifying structural variants and phasing of population genomic data.

Reference genome will aid population studies

Reference
Foote A. *et al. Wellcome Open Research* 2022; **7:** 250.

3

New tool to assemble genomes faster

Wellcome Sanger Institute researchers and their collaborators have created a new, fast, reliable and accurate tool for constructing chromosome-scale scaffold maps. The tool is now being used routinely by the Darwin Tree of Life project to assemble the genome sequences of thousands of species for the first time. The tool consistently outperforms other state-of-the-art scaffolding tools in both accuracy and contiguity across a wide range of species and genome sizes. The open-source tool has been designed to be robust, scalable and easy to use.

1000s of genome assemblies supported by YaHS

The rapid revolution of long-read, single-molecule DNA sequencing technologies in read length, base accuracy and per-base cost is driving a golden age for *de novo* genome assembly. Capitalising on these advances, several genome sequencing projects have been launched in the past few years. These include the Darwin Tree of Life Project, in which the Sanger Institute has a leading role. The initiative aims to assemble high-quality, chromosome-scale genomes for all animals, plants, fungi and protists species in Britain and Ireland.

Despite the technological advances, the assembly of high-quality genomes with long-read sequencing data alone is not yet possible. Long-distance linkage information is also needed. These data are used to create scaffolds on which to assemble the long-read data into chromosome-scale sequences. The linkage data often come from Hi-C maps – a sequencing-based proximity assay that provides contact information between pairs of loci in the genome.

Several scaffolding tools have previously been developed for the construction of chromosome-scale assembly with Hi-C data, though each has its own limitations, and the results are affected by various factors.

The team at the Sanger Institute and the University of Cambridge created a new scaffolding tool, called YaHS, to construct chromosome-scale scaffolds utilising Hi-C data. The tool differs from others in its novel method for building the contact matrix. The team tested it against other recently published Hi-C scaffolding tools, and found that YaHS generated genome assemblies of higher accuracy and contiguity, and it was more robust to assembly errors, across a wide range of species.

The tool was constructed to be fast, reliable and accurate and is already in use routinely by the Darwin Tree of Life project and others.

Reference
Zhou C. *et al. Bioinformatics* 2023; **39:** btac808

YaHS: yet another Hi-C scaffolding tool
github.com/sanger-tol/yahs

4

New genomes to help protect Britain's wild and ancient apples

The full genetic codes of Britain's only native wild apple, the European crab-apple, and four heritage edible apple varieties have been sequenced by Wellcome Sanger Institute scientists. These gold-standard genome assemblies will allow researchers to better understand the UK's apple heritage and will help prevent wild apples being hybridised out of existence by their domestic relatives.

The European wild or crab apple, *Malus sylvestris*, has been used by humans for millennia, evidenced by the fruit being found at Neolithic archaeological sites across the continent. It is still important today, as one of the main contributors to the domesticated apple, *M. domestica*.

M. domestica is a hybrid species, and its complex history is slowly being unravelled using genetic methods. As well as *M. sylvestris*, its ancestors include varieties from Central Asia and potentially Siberia, as it was traded along the Silk Road.

Apples are largely self-incompatible, and offspring grown from seed do not resemble the mother tree. To preserve desirable characteristics such as taste and disease resistance, trees are propagated using grafting, a technique that has been in use for centuries.

M. domestica continues to hybridise with wild apple trees, causing concern that the genetic integrity of *M. sylvestris* might be eroded in the long run. Morphological identification of hybrid trees is difficult, and so access to the full genomes of both *M. sylvestris* and *M. domestica* will facilitate the development of genome-wide species-specific markers, enabling the reliable assessment of levels of introgression in the European wild apple.

In order to better understand UK apple heritage, the team also produced DNA sequences for over 40 additional British and Irish apple varieties. Using comparative genomics, they investigated how the varieties are related, and how they had been moved around Britain and Ireland.

The increasing availability of full genome sequences of different apple varieties and wild relatives contribute to understanding their origins. It will also accelerate the use of wild species and genetically distinct genotypes of *M. domestica* for apple improvement, including for health benefits and resilience to disease and climate change.

Genomic information also allows us to peer into the past history of important crops, to look to the future of agricultural science and address the conservation issues of today.

Professor Mark Blaxter
Head of Tree of Life,
Wellcome Sanger Institute

References

Ruhsam M. *et al. Wellcome Open Research* 2022; **7:** 296.

Könyves K. *et al. Wellcome Open Research* 2022; **7:** 297.

3
day virtual conference

145
talks

2,140
people registered

5

Biodiversity Genomics 2022

In October 2022, researchers around the world gathered online at the virtual Biodiversity Genomics conference, hosted and facilitated by the Tree of Life programme at the Sanger Institute.

Attendees heard from genomics researchers across the globe, covering topics ranging from the conservation of Scotland's red squirrels and New Zealand's alpine parrots to new genome sequencing initiatives set up from Canada to Colombia. Much of this work is carried out in lockstep with the Earth BioGenome Project (EBP), an initiative aimed at sequencing the genomes of every species of plant, animal, fungus and protist on Earth.

The EBP aims to create a new foundation for biology to drive solutions for preserving biodiversity and sustaining human societies. Teams of scientists located all around the world, grouped regionally or via specific taxa, are contributing to this effort. To create, expand and unite these scientific communities, EBP and associated projects joined the online event.

The conference was live on three days, consisting of two 14-hour days moving around the globe. Talks covered comparative and evolutionary genomics, conservation and bioeconomy and genome assembly, annotation and analysis. There were also sessions on justice, equity, diversity and inclusion issues, including early careers and indigenous sovereignty. Inclusivity was at the heart of the event, which was free and open to all.

The final day included wider discussions and specific symposia for global projects. A total of 2,140 people registered for the 145 talks.

The talks from Biodiversity Genomics 2022: Science Day are freely available online: **youtube.com/playlist?list=PLsOubX-LMwUsh1brHKWwoxBYZx6egEn9Z**

6

Celebrating 500 genomes: from mistletoe to mackerel

Wellcome Sanger Institute staff have generated reference-quality genomes for 500 eukaryotic species, the majority of which have not previously been sequenced. The species have been collected, their DNA extracted and sequenced and their genomes assembled and released to public databases for researchers to freely and openly use in scientific initiatives across the globe.

The 500 genomes have been generated working in partnership with several projects, including the Darwin Tree of Life project, the Aquatic Symbiosis Genomics project, the European Reference Genome Atlas and the Vertebrate Genomes Project. The genomes also contribute to the world-wide effort to determine the DNA sequence of all complex life on Earth – the Earth BioGenome Project – which will create a new foundation for biology and drive solutions for preserving biodiversity and sustaining human societies.

Over the past three years, new pipelines encompassing data collection, sample management and processing, sequencing, quality control, genome assembly and data analysis have been set up at the Sanger Institute. Lepidoptera (butterflies and moths) have proved relatively easy to collect and process through these pipelines, and over 200 Lepidoptera genomes have now been publicly released.

Comparative genomics studies on arthropods are already utilising the data. For example, a first-of-its-kind analysis of homeobox genes, which encode transcription factors with essential roles in patterning and cell fate in developing embryos across Lepidoptera. The study shows the potential of newly generated genome assemblies in understanding evolution.

Other genomes have been trickier, but the team has made progress and has now extracted and sequenced DNA from all branches of the tree of life. Over 2022, the first fungi, cnidarians (jellyfish, sea anemones etc.), tunicates (sea squirts), molluscs and vascular plants have been completed.

All of the genomes are uploaded publicly onto the European Nucleotide Archive (ENA): ebi.ac.uk/ena and subsequent analysis is published in Wellcome Open Research wellcomeopenresearch.org/gateways/treeoflife.

Reference
Mulhair P.O. *et al. Genome Research* 2023; **33:** 32-44.

Our approach

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 38

Impact

We strive to carry out the most insightful research and help scientists and doctors to use our discoveries to develop new techniques, tests and therapies.

Page 40

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 42

We incubate the next generation of pioneers in genome research

Innovation

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

Page 44

Influencing Policy

We advise all levels of government both in the UK and across the world on the role, impact and importance of genomic and life science research.

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

Scale

In this section

1

Making global genomic disease surveillance a reality

The Sanger Institute's Genomic Surveillance Unit works with partners around the world to accelerate the use of genome sequencing to tackle infectious diseases, including SARS-CoV-2 and malaria.

Launched in 2022, the Unit seeks to help research partners around the world develop their genomic surveillance capabilities to monitor and understand disease-causing microbes in circulation. This information could save lives, reduce infectious disease health burden through informed public health interventions and contribute to global pandemic preparedness.

The value of genomics in disease surveillance and health interventions was proven over the past two years by the ongoing COVID-19 pandemic, where sequencing became an invaluable tool to track the virus' transmission and emergence of new variants.

The Genomic Surveillance Unit has been created to embed this approach worldwide. It works with partners to implement scientific pipelines, from sample collection through to the delivery of actionable data.

In particular, the team focuses on strengthening capacity for the generation and sharing of genomic data, developing easy-to-use tools and methods and creating accessible resources that provide actionable information for public health.

The goal is to implement systems that can track the spread and evolution of pathogens and monitor for new disease outbreaks globally.

MalariaGEN and COVID-19 surveillance at the Sanger Institute are now supported and delivered by the Unit. The move strengthens support for operational research and surveillance delivery and embeds these initiatives at the heart of the Unit's work.

Worldwide infectious disease genomic surveillance networks will improve global health delivery and pandemic preparedness greatly. By making genomics more accessible and easier to scale, the Unit aims to make genomic surveillance a practical and cost-effective infectious disease control tool for all.

> [Our] goal is to leverage experience and knowledge to enable … partners to obtain the data and actionable information to answer their pressing public health questions.

John Sillitoe
Director of the Genomic Surveillance Unit

MalariaGEN and COVID-19 surveillance supported by the Unit

2

Major new cancer map will aid precision medicine

Researchers from the Wellcome Sanger Institute and the Children's Medical Research Institute (CMRI), Australia have completed the world's largest protein map for cancer. It includes data for 8,498 proteins, in 949 cancer cell lines, representing 40 types of cancer. The work will enhance ongoing efforts to predict the response of an individual cancer to different treatments, helping to enable precision medicine in the clinic. These data will also inform the development of new treatments.

Proteins are responsible for most of the functions of life, including the behaviour of cancer cells, and how they respond to treatment. For some cancers, measuring the quantities of specific proteins can help guide treatment choices for a patient. But a comprehensive picture of all the proteins present in cancer cells has not been available until now.

In the first study at such scale, the CMRI team developed a high-throughput workflow using mass spectrometry to measure the amounts of 8,498 proteins in 949 cancer cell lines, representing 40 cancer types. The workflow was designed to also include biopsy samples, meaning it will be possible to use in the clinic in the future.

The cancer cell lines were grown by Sanger scientists, who previously deeply characterised them by measuring gene activation and drug responses and used whole-genome CRISPR-Cas9 screens to identify their vulnerabilities. The Sanger team expanded pharmacological screens of anti-cancer drugs during this new study and have now measured the cells' responses to combinations of 650 drugs.

Data analysis by CMRI and Sanger scientists comprehensively characterised the proteins in the cell lines, uncovering previously unseen patterns of activation. They also developed a new deep-learning technique to integrate the protein data with drug response data and gene essentiality screens. This method was then used to predict the response of the cancer cells to treatment. The results also pinpoint vulnerabilities in cancer cells that provide opportunities for developing new treatments.

The database, unprecedented in size, is a major resource now freely available for cancer researchers worldwide. It is also a significant step towards using this type of data to help clinicians choose the best treatment for individual patients.

949 cell lines studied

40 cancer types covered

8,498 different proteins measured

Reference

Gonçalves E. *et al. Cancer Cell* 2022; **40:** 835-849.e8

All data are publicly available at **cellmodelpassports.sanger.ac.uk**

3

World's most ambitious human genome sequence project delivered

In just three and a half years, the Sanger Institute sequenced 243,633 human genomes for the UK Biobank project. These data join those provided by project partners and will lay the foundations for medical and genomic discovery worldwide for many years to come.

In 2019, the Sanger Institute started work on the most ambitious human genome sequencing project in the world. In 2022, it delivered nearly 250,000 'gold standard' human whole-genome sequences and more than 20 petabytes of data for the UK Biobank biomedical database.

To ensure that each genome sequence was correct, every genome was read 30 times. Because each genome is made up of 3.05 billion pairs of DNA, it meant that the Sanger Institute's DNA sequencing teams read in the order of 21 quadrillion (21×10^{15}) letters of DNA. To deliver at this scale and speed required all the teams to redesign their processes – from sample handling and preparation, through DNA sequencing to data storage and analysis.

The whole-genome data are now in the UK Biobank resource, with each sequence linked to anonymised medical information. Together with the sequences from project partners, the resource is the largest human genome sequence database in the world, containing in-depth health information from 500,000 UK participants. More than 28,000 researchers from 98 countries use it to study disease risk, taking into account lifestyle, genetics and health factors.

The additional genomic information allows researchers to look for links between the genetic code and health using data that did not exist before – including in non-coding regions of the genome. Two hundred thousand sequences are already available to approved researchers, and this data sharing has enabled a wide range of studies and new discoveries across cancer, diabetes and heart disease. The whole full 500,000 human genome sequences will be made widely available towards the end of 2023.

Approx **21 quadrillion** (21×10^{15}) letters of DNA were read

More than **28,000** researchers from 98 countries use the UK Biobank

Impact

In this section

1

DECIPHER delivers for 18 years … and counting

Since 2004, a ground-breaking academic-medical collaboration has been shedding light on the genomic basis of developmental disorders in children. It laid the foundations for global genomic data sharing and enables clinicians to explore the underlying genomics of neurological conditions.

The Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) was launched soon after the publication of the draft human genome. Founded by researchers at the Sanger Institute and Addenbrooke's Hospital in Cambridge, it was one of the first times the reference genome was used in clinical practise: combining patients' genetic data with their medical information.

The database enabled clinicians and researchers to explore the effects of natural deletions and repeats of sections of DNA within the genome on health and disease. It is one of the world's key global genomic resources powering research into a wide range of medical and developmental fields, enabling the development of new diagnostic tests and improving genetic counselling.

Today the resource contains:

- more than 45,000 open-access patient records
- more than 184,000 physical observations
- more than 46,000 open-access records of copy number variations
- and more than 12,000 open-access sequence variants.

Since its inception, the resource has facilitated more than 2,600 peer-reviewed scientific publications and is now a collaboration between 302 projects across the globe. It has also processed more than 4,800 collaboration requests.

In early 2023, the database followed in the footsteps of previous Sanger Institute-supported foundational resources – such as Pfam and GENCODE – and is now hosted and supported by its Campus collaborator, EMBL's European Bioinformatics Institute (EMBL-EBI).

DECIPHER database
deciphergenomics.org

Facilitated **2,600+** peer-reviewed scientific publications

Supports **302** projects worldwide

Processed **4,800+** collaboration requests

2

Startup School incentivises Innovation on Campus

Since its inception in 2020, three cohorts of Wellcome Genome Campus Startup School trainees have gained the entrepreneurial skills needed to deliver genomic research into real-world products and services from genomics business leaders, investors and entrepreneurs. Early-stage ideas developed in the school are being progressed by the Innovation team.

Three of the key goals of the Sanger Institute are to grow the next generation of genome scientists, identify and explore new frontiers of genome science and lead the way in developing, applying and implementing genomics technologies. The Wellcome Genome Campus Startup School, run by the Innovation team, delivers on these goals by developing entrepreneurial scientists whose ideas and skills might otherwise remain untapped.

In a Campus-wide survey conducted in 2022, only 30 per cent of scientists who responded stated that their next role would be in academic research. To enable Sanger scientists to explore how they could use their genomic skills in commercial and entrepreneurial settings, the Innovation team developed the Startup School.

Each year, 24 researchers, from the Sanger Institute and EMBL-EBI, take part in teaching sessions and workshops delivered by leaders from research institutes, companies and investors. Designed as an online learning experience, it provides participants with the tools, insights and genomics sector connections needed to develop their research ideas into a successful product, service or spin-out company.

So far, roughly 20 per cent of the ideas from Startup School have been taken forward with support from the Innovation team. Three researchers from the School have used the skills they acquired to successfully win Technology Translation funding from the Sanger Institute, and most of the others are using the acquired entrepreneurial skills to broaden their career opportunities. For example, one participant has become the Principal Scientist at a Microbiome startup, and another is now working at a commercial multiomics innovation hub.

24
Campus scientists take part in the Startup School

STARTUP SCHOOL

3

Transatlantic collaboration powers Alzheimer's research

Using a human iPS cell line created at the Sanger Institute to enable scientists to develop new models for disease research, The Jackson Laboratory has created a range of resources for developing new therapeutics for neurodegenerative disorders.

The Sanger Institute has a long history of successful partnerships with The Jackson Laboratory, known as JAX, to provide the global research community with much-needed mouse and human cell lines at scale to power studies into health and disease.

In 2022, JAX – an independent, non-profit biomedical research institute based in the United States – launched the HUMAN iPS CELLS portal to provide a comprehensive resource of research models for scientists worldwide studying the genes involved in neurodegenerative diseases. It is available through JAX's repository, in partnership with the US National Institute for Health (NIH) and the Chan-Zuckerberg Initiative (CZI).

The cell lines present in the portal are derived from the Kolf2_C1 line, which was created at the Wellcome Sanger Institute as a line to be further engineered on demand to support multiple downstream projects. The Technology Translation team at Sanger negotiated and obtained the rights necessary to transfer the cell and enable JAX to engineer this cell line to eliminate a likely pathogenic mutation and make it useful for the study of neurodegenerative diseases.

The lines have been specially engineered by JAX for the NIH's Center for Alzheimer's and Related Dementias (CARD) program for the study of gene variants linked to four specific neurodegenerative disorders:

- Alzheimer's disease
- Parkinson's disease
- ALS (Amyotrophic Lateral Sclerosis)
- Frontotemporal Dementia.

> It is exciting to see that the Sanger Institute's science has been translated into real-world resources to tackle diseases that are difficult to study, such as Alzheimer's. This is only possible due to the 'science at scale' that the Institute delivers.

Mariya Chhatriwala
Wellcome Sanger Institute

Culture

In this section

1

First three Sanger Excellence Postdoctoral Fellows start at the Institute

The Sanger Institute welcomed three exceptional inaugural Fellows to power research in cancer, infectious diseases and the Tree of Life. The Fellowship scheme is designed to support the training and career development of scientists from Black heritage backgrounds and builds on the Institute's commitment to greater equality, diversity and inclusion.

The Fellowship is an annual offering that provides fully funded three-year positions. It is open exclusively to early-stage researchers who have an undergraduate degree and a PhD (or equivalent research experience) from a UK institution and are from a Black heritage background. The programme seeks to provide Black researchers with a clear career development pathway with training and mentorship that supports them to achieve their full potential.

The scheme was developed in response to the Sanger Institute's recognition that persistent racial inequalities disadvantage people from Black backgrounds in all walks of life, and that talent and excellence is lost along the academic pipeline.

A wide network of people and institutes from throughout the UK helped to design and develop this programme, creating a community of support that can be accessed at any point by the Excellence Fellows. The community also provides mentoring sessions to unsuccessful applicants to support them as they progress their careers.

> We must all do our part to break the 'glass ceiling' of career progression for Black scientists. Diversity helps our science to thrive, and we must constantly listen, learn, and evolve to ensure that everyone is supported personally and professionally.

Professor Sir Mike Stratton
Director of the Wellcome Sanger Institute

The Fellows are:

Dr Oumie Kuyateh, who is studying how the respiratory metagenome develops in early life in response to environmental factors. Her work could inform the diagnosis and treatment of respiratory infections such as pneumonia and meningitis.

Dr Ore Francis, who is using sequence predictions of proteins in eukaryotes to gain insights into how they work together to create and sustain life.

Dr Kudzai Nyamondo, who is exploring the timelines and trajectories of blood cancers, as a jointly funded Sanger Institute and Cancer Research UK Excellence Fellow. Her work seeks to create opportunities for novel early detection and intervention strategies.

2

Ten years of growing and supporting diversity

Since 2012, the Campus-wide Equality in Science initiative has delivered significant changes to enable the greatest spectrum of cultures, knowledge and working styles to thrive. To build an inclusive and supportive workplace that benefits everyone, the Equality, Diversity and Inclusion team take an intersectional approach to continually remove barriers to entry and career progression.

Family-related caring duties have often been a major barrier to career progression. On Campus, the Parent and Carers network provides practical support. The Institute has created generous family-friendly policies, including enhanced maternity and shared parental leave provision, along with additional paid leave for carers. In 2022, paid paternity leave was increased, and support was extended to colleagues with fertility issues or who have experienced pregnancy loss.

To overcome disadvantages due to parental leave and career breaks, the Institute's recruitment processes explicitly take these into account. The annual Janet Thornton Fellowship, launched in 2014, directly supports researchers looking to return to science after a break.

The LGBTQ+ network supports raising awareness of culture and working practices with regard to LGBTQ+ inclusion. It provides physical and virtual forums of support and helps to guide future workplace improvements. As part of this, the Institute took part in its first Stonewall Workplace Equality Index last year.

Neurodiversity can be a hidden area of inequality. To support our community with physical or unseen differences, the Institute's Ability Working Group rolled out the Hidden Disabilities Sunflower lanyard scheme across Campus in 2022. It provides a discrete way to ask for a little more understanding or help. The Institute's work to raise awareness around neurodiversity has been recognised with the ADHD Foundation's Neurodiversity friendly award.

Our thriving Race Equity Network promotes race equity, inclusion and cultural diversity. The Institute continued to roll-out its race equity strategy, including inaugural programmes of reverse mentoring for senior leaders and an anti-racism development programme to increase awareness and understanding. In addition, the first cohort of Sanger Excellence Fellows are now building their genomic research careers at the Institute.

3

Empowering staff to own their careers

The strength of the Sanger Institute lies in the expertise of its diverse research and administrative community. To support our technicians and technical experts to reach their full potential, the Institute is proud to be part of the Technician Commitment.

The Institute is a member of the Research Institute Technician Group which aims to identify ways to empower technicians and technical experts. At the Group's 2021 Symposium, career development and networking were identified as clear focus areas for all nine participating institutes. In response, Sanger has developed a series of open-access laboratory and career days, workshops and mentoring schemes to support and inspire.

In July 2022, the Institute hosted a four-day careers event with talks, workshops and networking opportunities. Building on the event's success, a full-day follow-up session centred around networking was held a month later. In addition, regular one-to-one CV writing workshops provide staff with practical support.

To build on this work, a series of open laboratory days has been developed to showcase the diversity of research techniques, opportunities and genomic careers at the Institute. The tours allow staff to build connections across teams, identify skill sets to develop and discover different technologies.

Two technicians took part in the Herschel programme for Women in Technical Leadership in 2022. Named after one of the earliest technicians at the turn of the 19th century, the career development programme focuses on navigating leadership challenges in research organisations. On successfully completing the scheme, the staff gave inspirational presentations, and 15 Sanger staff are now on the scheme.

Five Institute colleagues have completed their Professional Registration and are providing one-to-one mentoring to help their peers do the same. The scheme helps colleagues apply to the Science Council's Chartered Scientist (CSci), Registered Scientist (RSci) and Registered Science Technician (RSciTech) programmes. As an added benefit, the mentors gain new skills themselves, further powering their career development.

Innovation

In this section

1

Piecing together cancer clues

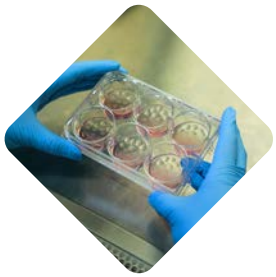
Mosaic is the latest spin-out company to come from the Sanger Institute's cancer research programme. It combines large-scale genomic knowledge with artificial intelligence to deliver a world-class drug discovery platform that can reveal new cancer treatments.

Many approaches to treating cancers are based on their tissue of origin; however, this does not take into account the cancer's underlying genetic roots and is prone to failure. Mosaic aims to disrupt this traditional approach by focusing on the DNA changes that enable a cancer cell to grow uncontrollably. These changes are the cancer's Achilles heel: they drive cell division but often create weaknesses that can be targeted.

With translational support from the Institute's Innovation team, the spin-out companys' capabilities are founded on expertise and tools developed within the Cancer, Ageing and Somatic Mutations programme at the Sanger Institute for more than a decade. The result is a unique dual technology platform that maps DNA changes across the entire genome in thousands of cancer models and combines the data with advanced machine-learning and statistical algorithms to selectively identify and exploit the vulnerabilities of different types of cancers.

The unique end-to-end capability of Mosaic's platform gives it a deep knowledge of how to integrate functional and genomic datasets to deliver innovative new therapies.

MOSAIC^{TX}



We aim to change the traditional way of treating cancer by combining our genetic knowledge with state-of-the-art technologies to select the best targets for drug development by identifying which individual patients will respond.

Mathew Garnett
Wellcome Sanger Institute

2

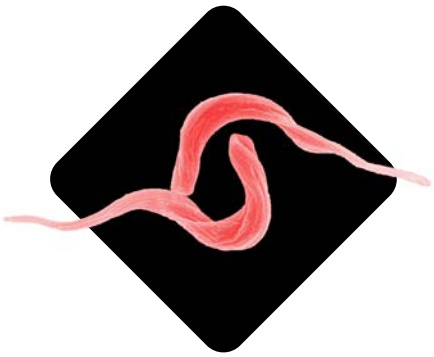
Funding future health

Over the past 10 years, the Sanger Institute's Translation Committee has funded successful initiatives that resulted in spin-out companies and technologies that have transformed drug development and microbiome-based therapies. In 2022, the Committee funded six additional innovative projects in disease research, infection surveillance and healthcare.

The Translation Committee Fund nurtures research by Sanger scientists to a level where it can be developed into real-world applications either as a licensable product or services or as a spin-out company. Previous successes include the Microbiotica gut bacterial platform, Mosaic's computational cancer research platform and unlocking prominent charitable funding from the Bill and Melinda Gates Foundation for the SpotMalaria test.

The 10th anniversary of the Fund saw a number of projects reach significant development milestones to attract investment. One key project that was validated was a new technology to create personalised CRISPR libraries that could enable tailored treatments for patients. Also a digital technology that could provide personalised prognoses for blood cancer progressed towards achieving regulatory registration.

The Translation Committee funded the creation of proof-of-concept data for the Sanger Institute's sixth and latest spin-out company, Mosaic (see page 44). The Committee also secured further funding to allow a new Trypanosoma vaccine to undergo field trials.



In 2022, there were a record number of applications to, and awards from, the Fund. The new projects come from all areas of the Institute's research:

- Developing a new *Haemophilus influenzae* vaccine for young children to further reduce the risk of developing meningitis or pneumonia.
- Creating a more precise method to create safe and effective cell-based therapies by exploring how mature human cells can be reprogrammed into stem cells in a more efficient way.
- Adapting high-throughput DNA sequencing approaches into methods suitable for portable low-cost, low-throughput machines.
- Using machine intelligence to understand the rules of gene regulation to generate new cell models of health and disease.
- Mapping the developmental state of the cells present in different tissues of the body.
- Powering new medical treatments by developing new ways to efficiently differentiate stem cells into three cell lines that have UK clinical approval.

For more about the Fund's impact over the past decade, please see below.

Translation Committee Fund Impact 2012–2022

2022 Call – Largest call in recent years
Demonstrating an increased engagement with innovation and rising number of opportunities to translate Sanger science for health and biodiversity benefit.

Influencing Policy

In this section

1

Embedding research equity at every level

Sanger Institute researchers strive to put equity and capacity building at the heart of their international research collaborations. The Policy Team has been working with ethics experts and project partners in low- and middle- income countries to support our scientists to work equitably with collaborators.

To ensure that all aspects of research collaboration were included, the Policy Team mapped out the entire lifecycle of a research project: from inception, funding requests and contract negotiation, through delivery and sharing data to outputs and publication of scientific papers. From this, the team conducted an extensive literature search to identify the key issues for equity and leading ethics researchers and social scientists in the field.

Based on discussions with these experts, the team drew up and refined draft principles for equity in research collaborations. These provided a framework for discussion in two roundtables organised in conjunction with the Centre for Science and Policy (CSaP) at the University of Cambridge.

The first roundtable was held virtually to allow the widest possible participation by genomic research partners in low- and middle- income countries. The need for full engagement with collaborating partners at project inception was highlighted to ensure that priorities aligned with theirs to ensure the greatest impacts and benefits for all.

The need for strengthening research capacity to support collaborators' long-term goals was pinpointed. Sharing expertise and resources to grow administrative support, skills and technical infrastructure for future projects was a strong theme.

The second roundtable drew together leading UK-based researchers who had established collaborations in low- and middle- income countries, alongside ethicists and people working in research journals and funding bodies. Highlighted issues included the need for time in a research project to address equity and ways to incentivise equitable practice.

These insights are being used to develop a set of principles and guidelines to support Sanger scientists embed equity throughout their research projects.

2

Protecting frictionless but fair global genomic data sharing

The Sanger Institute Policy Team has worked with researchers, governments and global organisations to ensure that global genomic databases remain openly available to the scientific community. A major step towards enshrining this principle was made at the UN Biodiversity Conference COP15.

Today, most genomic research is powered by globally available open-access genomic databases. Without them, the rapid development of COVID-19 vaccines would not have been possible. However, there is concern that not all countries that openly share their data are fairly rewarded.

The United Nations developed the Nagoya Protocol that required bilateral agreements between countries sharing physical genetic resources. However, it did not cover digital sequence information (DSI) available in open-access global databases. To address this, the UN tasked the Convention on Biological Diversity (CBD) to develop a DSI protocol for approval at COP15.

The Sanger Policy Team engaged with the protocol's development at many levels to ensure fair benefit sharing that does not hinder scientific research. They answered calls for evidence, met with, Department for the Environment, Food and Rural Affairs (Defra) officials (the UK's negotiators), participated in a CBD-commissioned Informal Advisory Group and worked with the international DSI Scientific Network to present the scientific community's view.

The debate culminated at COP15 with an agreed statement that DSI benefit sharing should 'not hinder research and innovation' and 'be consistent with open access to data'. Governments will now develop a multilateral mechanism to collect and share benefits via a global fund.

The Policy Team will continue to engage with developments. One key detail to be determined is when benefit sharing is triggered. If it is set at point of access, this could lead to additional bureaucratic burden for scientists, but if it is set later (such as when a commercial product is created), then frictionless sharing could continue.

3

Making sure red tape didn't bind scientists' hands

In 2022, the UK Government proposed changes to data protection law and regulations to reduce bureaucracy. To ensure that the proposals did not negatively impact national and international research collaboration, the Sanger Institute worked closely with Wellcome, the NHS Confederation and the research community to guide the Government's decisions.

The UK Government's Department of Culture, Media and Sport (DCMS) released a 150-page set of proposals to clarify and streamline UK General Data Protection Regulation (GDPR) and Data Law. It sought to consolidate all the rules relating to data protection and research into one place. As part of this, it proposed: a new statutory definition of, and lawful basis for conducting, scientific research; lowering the threshold to accessing patient data; and embedding the principle of broad consent to data use.

These changes risked greatly increasing red tape for UK researchers who would need to develop new positions and create new processes to accommodate them. They also introduced ideas that could reduce public trust in the use of patient data. These ideas also moved the UK's GDPR out of alignment with Europe, threatening the favourable adequacy decision that had been previously agreed and preventing the free flow of data into and out of the European Union.

As part of the Institute's previous work in this area, it had established a UK advisory group of data protection experts, research regulators and academics in conjunction with the NHS Confederation. In 2022, Wellcome approached the Institute to use the group as a sounding board in the run up to a roundtable it was hosting with the DCMS. The outcomes of the workshop directly influenced the focus and agenda of the meeting, which the Sanger Policy Team attended.

Within the meeting, there was broad agreement that many of the changes were unnecessary and endangered the adequacy agreement and public trust. The issues raised by Sanger and others have been taken on board by the Government, and the latest proposals are much more likely to support adequacy and European collaborations.

Collaboration

In this section

1

Human genome is finally complete

Sanger scientists have collaborated to create the first complete, gapless reference human genome sequence, two decades after the first draft was published. The work was carried out by the Telomere to Telomere (T2T) consortium, with the Institute contributing its expertise to analyse and refine the updated genome.

The first human reference genome, published in 2003, covered roughly 92 per cent of the genome and laid the foundations for much of the world's genomic research into human health and disease. However, the remaining 8 per cent, made up of the repetitive ends of chromosomes (telomeres) and dense middle sections (centromeres), were too complex to resolve. Yet having a complete DNA sequence is critical for understanding the full spectrum of variation in the human genome and its contribution to health and disease.

T2T consortium researchers took advantage of laboratory, DNA sequencing and computational techniques developed over the past 20 years to generate the completed genome. They used a special cell line that has two identical copies of each chromosome to provide the DNA for sequencing and used short- and long-read methods to map the missing regions.

Sanger researchers helped to design a new error-checking strategy that revealed small errors and structural misassemblies. To correct these errors, the team designed a new repeat-aware strategy that made accurate assembly corrections in large repeats, fixing 51 per cent of the existing errors.

The complete genome sequence, including more accurate maps for five chromosome arms, will significantly add to the research community's knowledge of how chromosomes segregate and divide.

T2T consortium were also able to discover more than two million additional variants in the genome, providing vital information about changes in 622 medically relevant genes.

Many research groups are already using this information in their research.

For the first time ever, we are now able to analyse a human genome in its entirety and produce many more at this level of quality.

Dr Kerstin Howe
Wellcome Sanger Institute

Reference
McCartney A.M., *et al. Nature Methods* 2022; **19:** 687-695.

If you would like to know more, visit:
nature.com/articles/s41592-022-01440-3

2

Sanger pioneers at scale

The Sanger Institute explores and supports new fields of genomic research through experimentation and analysis at vast scale. In 2022, we welcomed new faculty whose diversity of ideas and experience will lay the foundations for innovations in cancer, cellular genomics, human health, infectious disease and ecology.

The Institute's faculty conduct ambitious, curiosity-led research that can only be carried out using our large-scale, high-throughput genomic experimentation, DNA sequencing and analysis resources. Over the past year, we have welcomed seven new faculty to explore new fields of genomic research and provide the methodologies and tools the global research community needs to exploit them.

In the Cancer, Ageing and Somatic Mutation programme, Jyoti Nangalia uses genomics to explore the life history of cancer development to inform diagnosis and prevention. Her colleague Raheleh Rahbari interrogates the genomes of people who are predisposed to developing cancer to understand the triggers and preventers of disease.

In the Cellular Genetics programme, Muzlifah Hannifa is building a comprehensive atlas of the cellular and molecular programmes at work in human development. She will use these findings to explore how these systems are co-opted in tumour formation and immune-mediated diseases.

In the Human Genetics programme, Ben Lehner is seeking to make biology programmable by marrying experimentation at scale with artificial intelligence to unlock the building blocks of protein structure and function.

In the Parasites and Microbes programme, Josie Bryant is developing genomic tools to explore the hidden microbial world of our lungs in health and disease. Her insights could help explain how co-infections and antibiotic resistance develop.

Joana Meier has joined our Tree of Life programme to explore how new species are formed, offering insights into the interplay of ecology and development. Her colleague Kamil Jaron is exploiting the outputs of the Darwin Tree of Life project to understand the genomics of why reproductive strategies occur in nature.

7
new group leaders across all 5 research programmes

3

Genentech adds expertise to the Open Targets consortium

The Roche Group company has joined the Wellcome Genome Campus' pioneering public-private partnership to drive the delivery of safe and effective medicines.

Founded in 2014, the Open Targets consortium addresses all aspects of human health and disease, with a particular focus on immunology and inflammation, oncology and neurodegeneration research. It draws together the skills and knowledge of not-for-profit research institutes and pharmaceutical companies to transform drug discovery.

Genentech will work pre-competitively with the collaboration's academic and commercial partners to create and use big data in innovative experimental and informatics projects.

The collaboration provides a unique environment that creates a critical mass of expertise that could not be built elsewhere. It blends the genomic knowledge of the Sanger Institute and EMBL's European Bioinformatics Institute (EMBL-EBI) with the research approaches of Bristol Myers Squibb, Genentech, GSK, Pfizer and Sanofi to deliver insights that no single organisation could.

Clinical trials for new drugs are much more likely to succeed if the therapies are backed by genetic evidence. The partnership combines large-scale genomic experiments and computational techniques with knowledge to identify causal links between targets, pathways and diseases. This knowledge enables researchers to pinpoint potential drugs that are most suited to a disease's biological targets, reducing the time and cost of drug development.

The partnership is committed to rapid publication to openly share experimental data, informatic methods and other knowledge generated by the consortium with the broader scientific community.

Open Targets

> Open Targets brings a team of multidisciplinary scientists together to work in partnership on projects that no one partner could do alone … learning from [Genentech's] approaches will strengthen our exploration of new targets and biology that can lead to novel, safe and effective medicines.

Gosia Trynka
Experimental Science Director of Open Targets and Group Leader at the Wellcome Sanger Institute

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:

Cover

Kidney cancer sample – Omer Bayraktar and Kenny Roberts, Wellcome Sanger Institute

Contents

Page 2
Villus in small intestine – S. Schuller
Data Centre server wires – Dan Ross / Wellcome Sanger Institute

Page 3
First collection plate of insects from BIOSCAN at the Sanger Institute – Lyndall Pereira / Wellcome Sanger Institute
Tree of Life sample preparation – David Lavene / Wellcome Sanger Institute

What we do

Page 5
Postdoctoral Fellows – Phil Mynott / Wellcome Sanger Institute

2022 Timeline

Page 6
Wellcome Connecting Science Logo – Wellcome Connecting Science
T-cells – Alex Ritter, Jennifer Lippincott Schwartz and Gillian Griffiths, National Institutes of Health
Mini guts used to study whipworm infections – M. Duque-Correa, D. Goulding, F. Rodgers, *et al.* (2022)
Open Targets Logo – Open Targets

Page 7
Cells involved in human sex determination – Cecilia Icoresi Mazzeo / Wellcome Sanger Institute
Streptococcus – Meredith Newlove, CDC/ Antibiotic Resistance Coordination and Strategy Unit
Villus in small intestine – S. Schuller
Tree of Life Bioinformatician – David Lavene / Wellcome Sanger Institute
Scanning electron microscope image of the whipworm *Trichuris trichiura* – Dave Goulding / Wellcome Sanger Institute
Developmental lung cell atlas – Peng He, *et al. Cell* 2022. DOI: 10.1016/j.cell.2022.11.005

Our Work Contents

Page 10
Tree of Life DNA sequencing photos – David Lavene / Wellcome Sanger Institute
Trichuris trichiura egg and bacteria – Dave Goulding / Wellcome Sanger Institute

Cancer, Ageing and Somatic Mutation

Page 15
Notch1 mutant cells in ageing mouse oseophagus – Emilie Abby / Wellcome Sanger Institute

Page 16
Kidney cancer sample – Omer Bayraktar and Kenny Roberts, Wellcome Sanger Institute

Cellular Genetics

Page 18
Developing B Cells in prenatal gut – Chenqu Suo, Sophie Pritchard, Nadav Yayon / Wellcome Sanger Institute

Page 19
Cells involved in human sex determination – Cecilia Icoresi Mazzeo / Wellcome Sanger Institute
Developing B Cells in prenatal gut – Chenqu Suo, Sophie Pritchard, Nadav Yayon / Wellcome Sanger Institute

Page 21
Developmental Lung Cell Atlas – Peng He, Kyungtae Lim, Dawei Sun, *et al.*

Human Genetics

Page 22
T cell – NIAID

Page 24
Villus in small intestine – S. Schuller

Parasites and Microbes

Page 26
Trichuris trichiura egg and bacteria – Dave Goulding / Wellcome Sanger Institute

Page 27
Streptococcus – Meredith Newlove, CDC/ Antibiotic Resistance Coordination and Strategy Unit

Page 29
Mini guts used to study whipworm infections – M. Duque-Correa, D. Goulding, F. Rodgers, *et al.* (2022)
Scanning electron microscope image of the whipworm *Trichuris trichiura* – Dave Goulding / Wellcome Sanger Institute

Page 31
Respiratory sample processing photos – Dan Ross / Wellcome Sanger Institute

Tree of Life

Page 32
Tree of Life bioinformatician – David Lavene / Wellcome Sanger Institute
Tree of Life butterfly photo – Luke Lythgoe / Wellcome Sanger Institute

Page 35
Tree of Life photos – Luke Lythgoe / Wellcome Sanger Institute

Our Approach Contents

Page 36
Postdoctoral Fellow – Phil Mynott / Wellcome Sanger Insitute

Scale

Page 38
DNA sequencing – Dan Ross / Wellcome Sanger Institute

Impact

Page 40
Baby music – thedanw / Pixabay

Culture

Page 42
Postdoctoral Fellows – Phil Mynott / Wellcome Sanger Institute

Page 43
Lanyard – Hidden Disabilities
Long-read DNA sequencing – David Lavene / Wellcome Sanger Institute

Innovation

Page 45
Trypanosoma vivax – Dave Goulding / Wellcome Sanger Institute

Influencing Policy

Page 46
Houses of Parliament – Paul Buffington, Unsplash

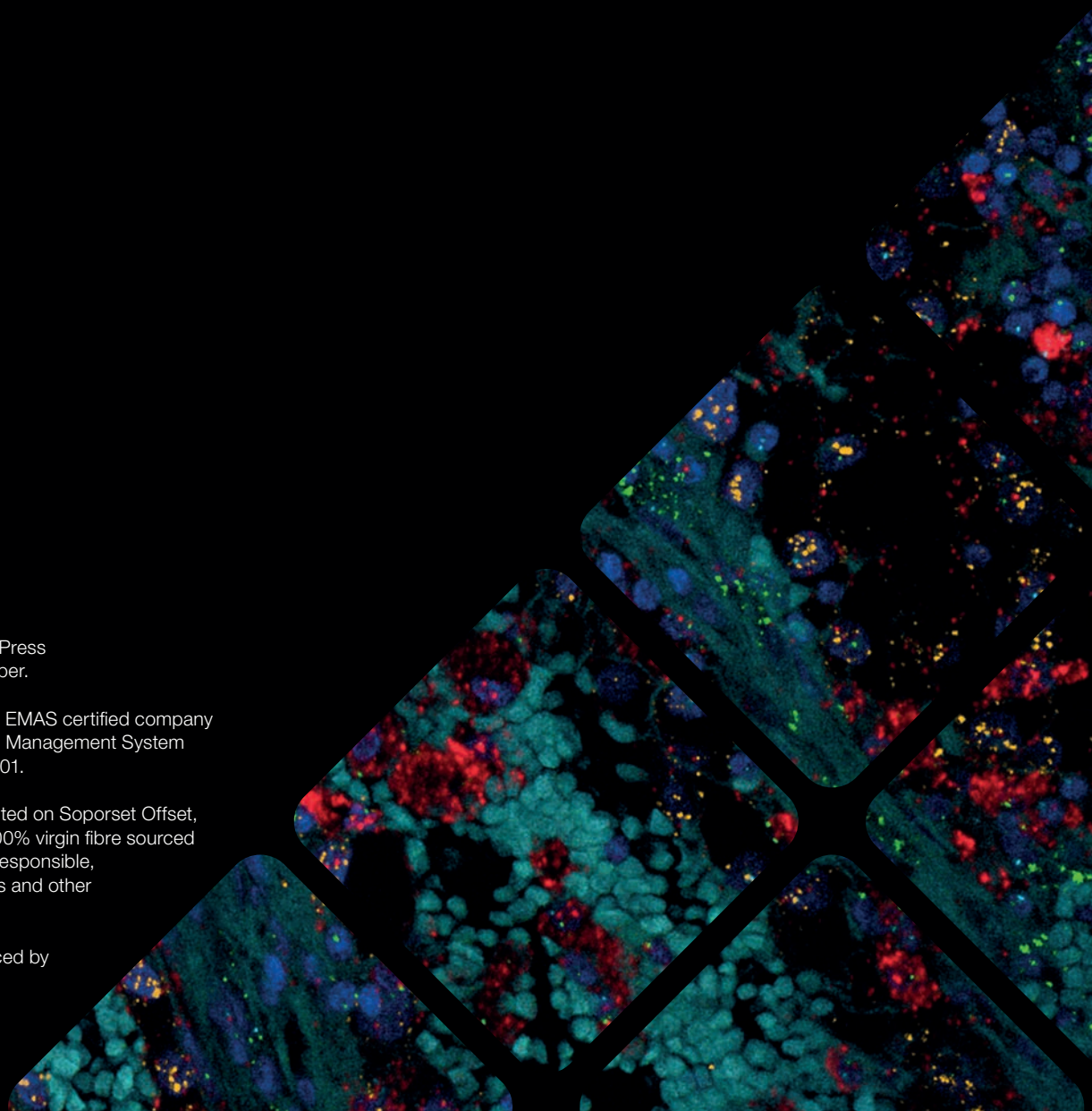
Collaboration

Page 49
Open Targets Logo – Open Targets

Wellcome Sanger Institute Highlights 2022/23

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, 2023.



Printed by Kingfisher Press
on FSC® certified paper.

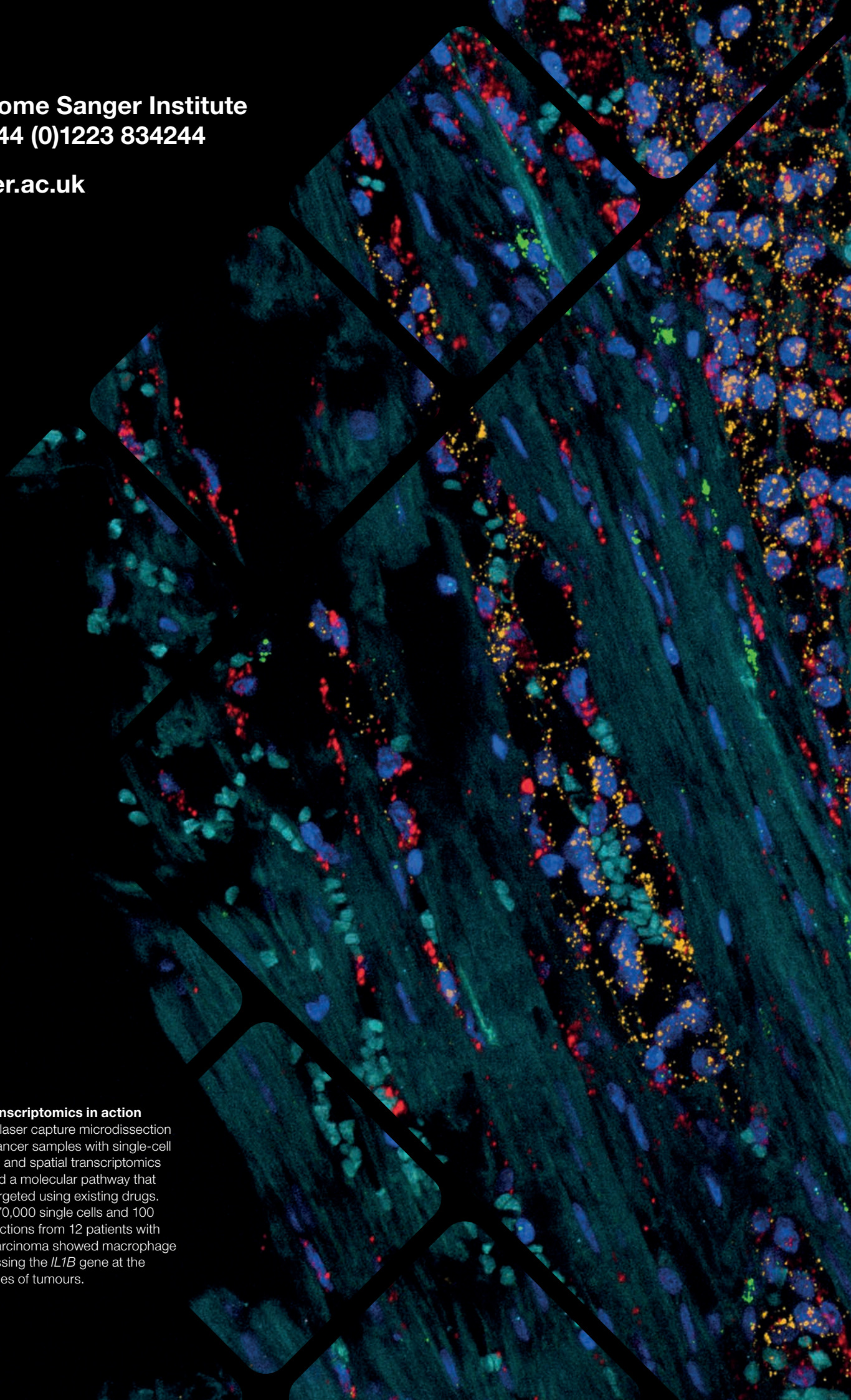
Kingfisher Press is an EMAS certified company and its Environmental Management System is certified to ISO 14001.

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

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

sanger.ac.uk



Spatial transcriptomics in action
Combining laser capture microdissection of kidney cancer samples with single-cell sequencing and spatial transcriptomics has revealed a molecular pathway that could be targeted using existing drugs. Studying 270,000 single cells and 100 micro dissections from 12 patients with renal cell carcinoma showed macrophage cells expressing the *IL1B* gene at the leading edges of tumours.