Impacting life on a global scale

Highlights 2020/21

wellcome sanger institute

COVID-19 social distancing circles in a public park in Heidelberg Neckarwiese. The Sanger Institute is the central sequencing hub for the COVID-19 Genomics UK (COG-UK) consortium

What we do

Director’s Introduction 2

At a Glance 4

Year in Numbers 7

Our work

COVID-19 National Genomic Surveillance 10

Cancer, Ageing and Somatic Mutation 16

Cellular Genetics 20

Human Genetics 26

Parasites and Microbes 30

Tree of Life 36

Our collaborative research and innovations improve people's lives around the world

Our approach

Scale 42

Innovation 44

Culture 46

Influence 48

Connections 50

Other information

Image Credits 52

Institute Information 53

Read more about how the Sanger Institute is collaborating in the UK's COVID-19 National Genomic Surveillance at: **www.sanger.ac.uk/science/covid-19-science/**

> “The Sanger Institute continued to deliver fresh insights and create innovations… delivered by scientists working at kitchen tables or in laboratories under strict COVID-19 compliance rules. That the Institute has continued to deliver world-class science is a testament to their dedication.”

Professor Sir Mike Stratton, Director,
Wellcome Sanger Institute

Director’s Introduction

This past year has demonstrated – like no other before it – that no country is an island; we are a global community interconnected by international travel, communication channels, ideas… and disease. When a person in one corner of the world develops a cough, the world can rapidly catch a fever. Scientific discovery, and the work of this Institute in particular, is no different.

Now, more than ever, collaborative global networks of science are key to delivering the genomic, health and epidemiological data at scale needed to combat the world’s ills. Yet vital data may be hidden behind passwords, firewalls and international laws, stifling international genomic inquiry. For these reasons, I am delighted that the Wellcome Sanger Institute is contributing to a wide range of initiatives dedicated to providing open-access data that is equitable to all contributors and sensitive to the diverse cultural needs of participants. From the Wellcome LEAP initiative, through the visionary work of International Common Disease Alliance and Earth BioGenome Project, to the databases of the Human Cell Atlas initiative and COSMIC, our researchers are defining and delivering the foundations for future worldwide research.

Equally, I am proud of the work of the Institute’s staff who have proved that a pandemic cannot stop science, but that science may stop a pandemic. Early in 2020 the Sanger joined the global effort against COVID-19 by providing knowledge, sequencing capacity and funding to the work of the COVID-19 Genomics UK (COG-UK) consortium. Their contribution is truly the Institute in microcosm. From administrative staff in Human Resources and Health and Safety, through Finance and Logistics, to laboratory technicians and bioinformaticians more than 300 staff from all areas of the organisation collaborated to deliver the largest SARS-CoV-2 virus sequencing operation in the world. And the work is still ongoing, our scientists are delivering process improvements and new insights that are benefiting the global genomic surveillance community.

This, in itself, would be a major achievement for any organisation, yet the ingenuity, determination and curiosity of our teams meant that the Institute continued to deliver fresh insights and create innovations in fields as diverse as cellular competition in cancer and the roots of developmental disease, to the genomic secrets of scallops and inherited resistance to malaria. Many of these insights were delivered by scientists working at kitchen tables or in laboratories under strict COVID-19 compliance rules. That the Institute has continued to deliver world-class science is a testament to their dedication.

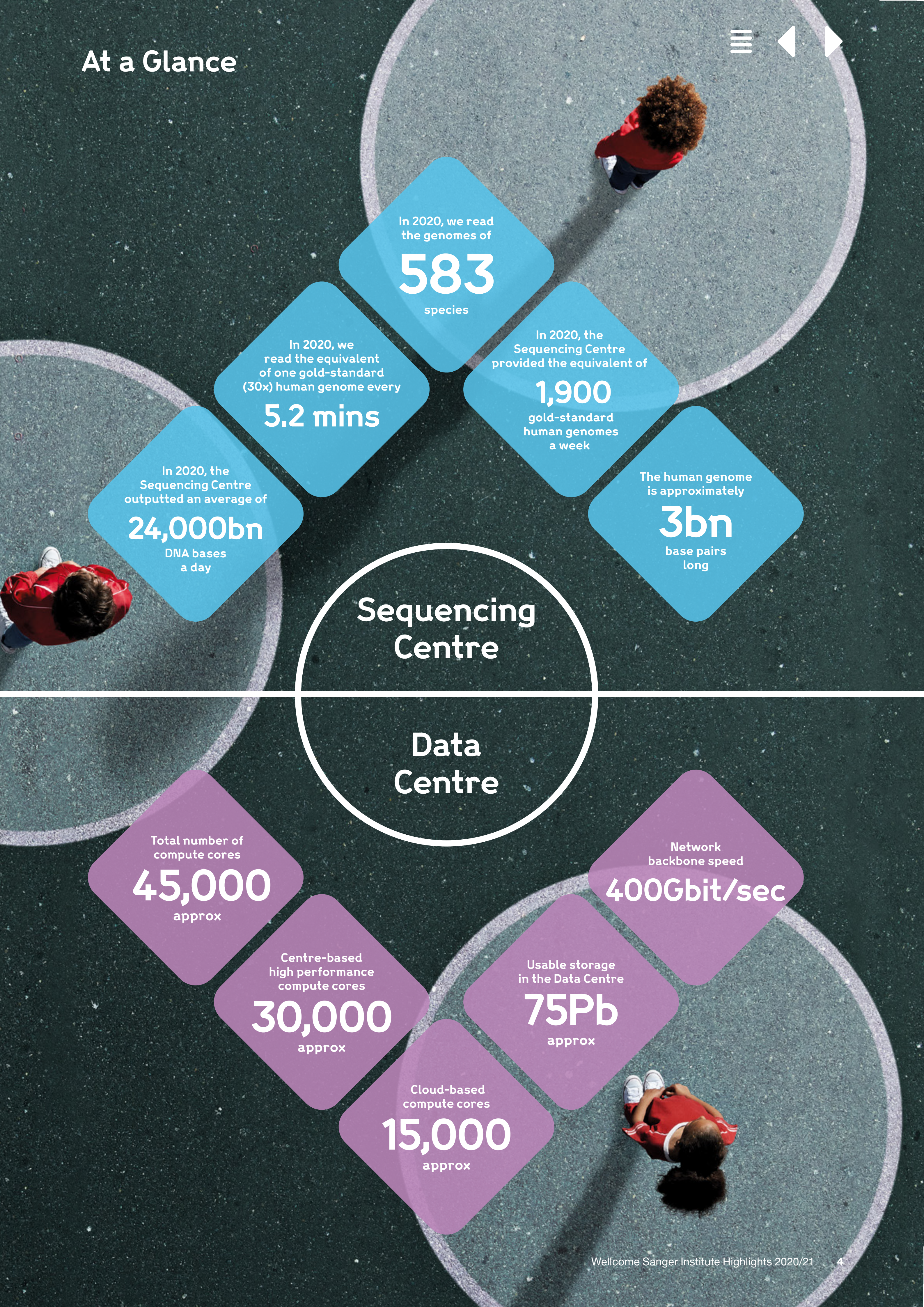
However, our staff are not machines, able to blot out the world and its woes. The Institute’s strength lies in people’s creativity and diversity of experience, ideas and perspectives. From isolated PhD students living away from home to busy families juggling homeschooling and caring for elderly parents, Sanger colleagues faced many challenges to their physical and mental wellbeing. In the midst of this maelstrom, I am delighted that the ‘Sanger spirit’, first developed during the Human Genome Project, came to the fore with teams not just delivering their science, but finding innovative ways to support and encourage each other with understanding, tolerance and good humour.

As an Institute we have been able to underpin these efforts by providing funding and online support channels to enable flexibility in working hours and project delivery timings, and supply mental health and career development needs. Through extended funding for individuals and scientific projects, combined digital town halls and training channels, and supplemented by online Family Cabaret events and virtual coffee meetings, I am proud that we have been able to support our most valuable resource – our people.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute

Find out more about the effort to sequence the genome of the SARS-CoV-2 virus at the Wellcome Sanger Institute.

At a Glance



Sequencing Centre

In 2020, we read the genomes of **583** species

In 2020, we read the equivalent of one gold-standard (30x) human genome every **5.2 mins**

In 2020, the Sequencing Centre provided the equivalent of **1,900** gold-standard human genomes a week

In 2020, the Sequencing Centre outputted an average of **24,000bn** DNA bases a day

The human genome is approximately **3bn** base pairs long

Data Centre

Total number of compute cores **45,000** approx

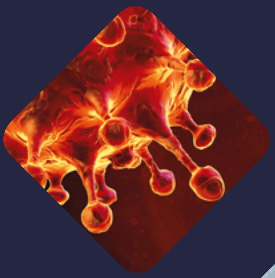
Network backbone speed **400Gbit/sec**

Centre-based high performance compute cores **30,000** approx

Usable storage in the Data Centre **75Pb** approx

Cloud-based compute cores **15,000** approx

COVID-19 timeline



First upload of SARS-CoV-2 sequence data to the CLIMB database (Medical Research Council database)

Mar

Biological risk assessment conducted

Surveillance ethics formulated

Wellcome Genome Campus temporarily closes to all but critical research activities

Sanger Institute helps to found and fund the COVID-19 Genomics UK (COG-UK) Consortium to provide genomic data and help to inform Public Health Agencies decision making

Sanger's SARS-CoV-2 virus sequencing effort begins – to provide data for analysis by COG-UK partners

1st UK virus samples received

Sample runs conducted on Illumina MiSeqs

UK enters first national lockdown

Apr

Biological risk assessment approval from Health and Safety Executive

Surveillance ethics approval secured

Relationship managers put in place to coordinate with NHS and Public Health England (PHE) sites

Sequencing moves to high-throughput Illumina NovaSeqs

10,000 samples received in one month

May

300+ volunteers involved with Sanger's SARS-CoV-2 sequencing pipeline

Refrigeration container units (reefers) installed to store approximately 7 million samples

Process to combine sequence to PHE metadata developed

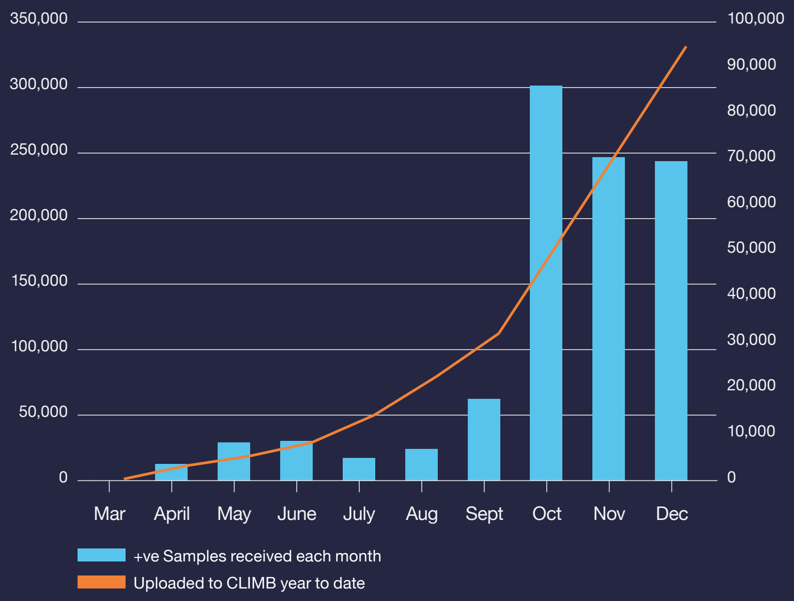
Jun

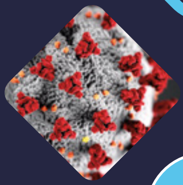
Focus moves to supporting COG-UK by sequencing Lighthouse Laboratories samples to enable background surveillance

Sequencing and handling protocols refined to ramp up volumes and lower turnaround times

Enhanced Laboratory Information Management Systems developed to streamline processing and optimise turnaround times

Number of positive samples received and number uploaded to CLIMB year to date





Jul

Robotics upgraded and deployed to cherry pick samples to speed pipeline

Sample destruction process scoped out

Aug

Sample destruction process agreed

Plan developed for new lab to expand sample processing capacity with new robotics

Formal recruitment to Sanger's SARS-CoV-2 sequencing pipeline initiated to release volunteers to their original projects

Sep

Sanger begins planning for ongoing 7-days-a-week sequencing pipeline for 2021 to support COVID-19 monitoring

Sample destruction process commenced

Design of new lab space completed

Oct

Tailed primer process implemented to shorten sample preparation time

Build of new lab space begins

Nov

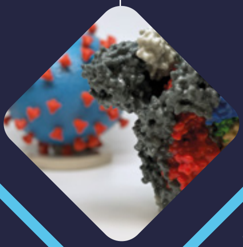
Sanger demonstrates value of genomic surveillance to test and trace officials to guide future strategy

Sanger-funded scientists/ technicians embedded in Lighthouse Laboratories to speed sample identification and transfer

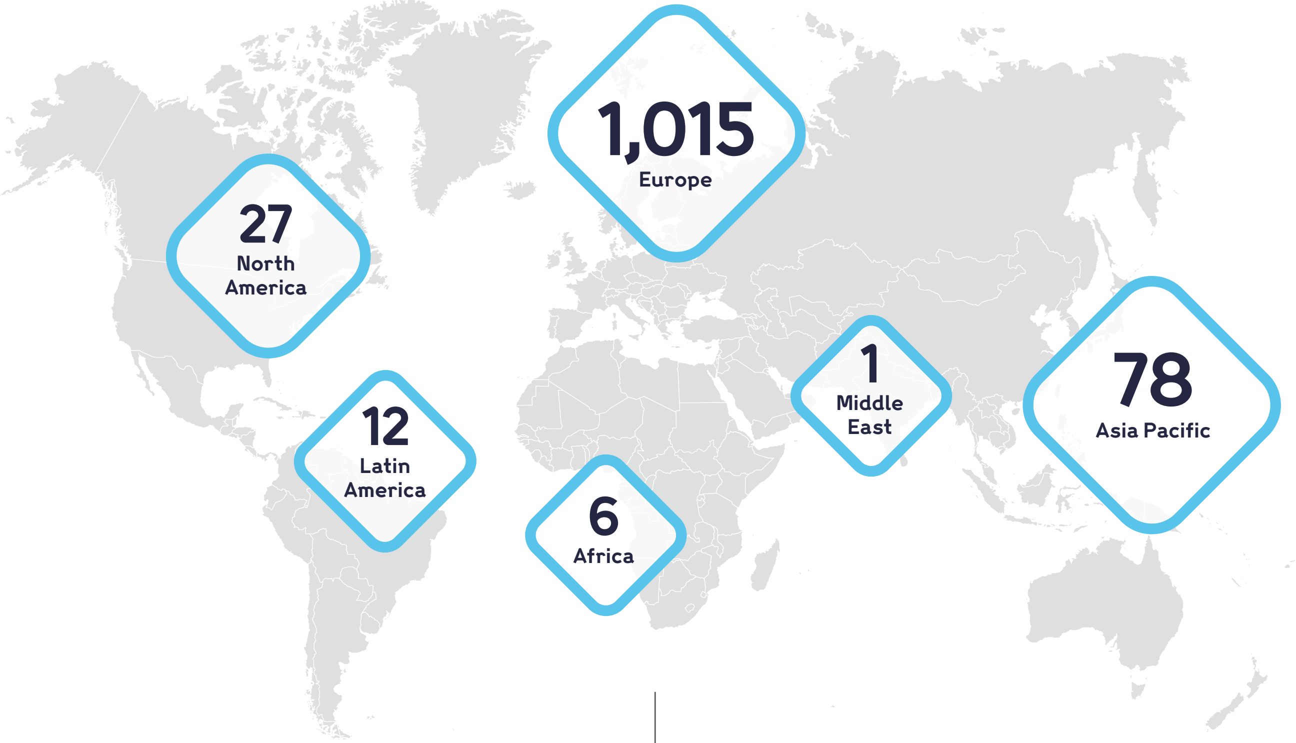
Dec

Data and analysis guide UK Government to initiate a second national lockdown

Sanger contributes to sequencing and analysis that reveals rapid spread of B.1.1.7 variant



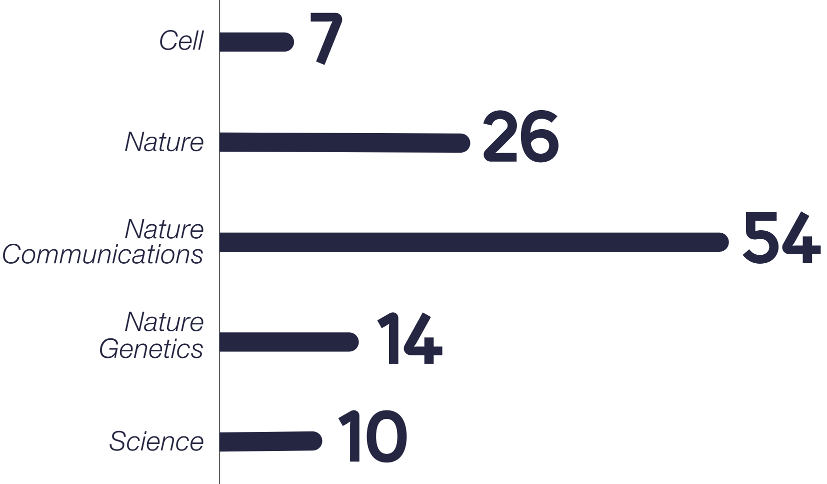
Year in Numbers



Where Sanger staff are from

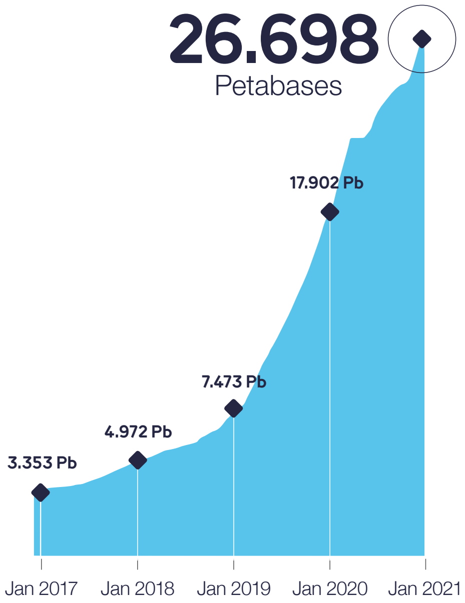
In January 2021

Who published our work?



568 research articles in 2020

How much DNA was sequenced?



What we do helps us to understand and improve all life – wherever that may be

With secured funding from Wellcome, we are able to strategically focus our work in key research fields:

COVID-19 National Genomic Surveillance

Tracking COVID-19 by sequencing thousands of virus genomes to inform public health measures and to monitor for new variants. Exploring how SARS-CoV-2 interacts with cells, and developing new diagnostics.

Page 10

Cancer, Ageing and Somatic Mutation

Provides leadership in data aggregation and informatics innovation, develops high-throughput cellular models of cancer for genome-wide functional screens and drug testing, and explores somatic mutation's role in clonal evolution, ageing and development.

Page 16

Cellular Genetics

Explores human gene function by studying the impact of genome variation on cell biology. Large-scale systematic screens are used to discover the impact of naturally-occurring and engineered genome mutations in human iPS cells, their differentiated derivatives, and other cell types.

Page 20

The Sanger Institute is helping to understand symbiosis in coral reefs as part of the Aquatic Symbiosis Project. Read more at: **www.sanger.ac.uk/collaboration/aquatic-symbiosis-genomics-project/**

Human Genetics

Applies genomics to population-scale studies to identify the causal variants and pathways involved in human disease and their effects on cell biology. It also models developmental disorders to explore which physical aspects might be reversible.

Page 26

Parasites and Microbes

Investigates the common underpinning mechanisms of evolution, infection and resistance to therapy in bacteria and parasites. It also explores the genetics of host response to infection and the role of the microbiota in health and disease.

Page 30

Tree of Life

Investigates the diversity of complex organisms found in the UK through sequencing and cellular technologies. It also compares and contrasts species' genome sequences to unlock evolutionary insights.

Page 36

COVID-19 National Genomic Surveillance

Sanger's genomic surveillance is a 7 days a week operation

In this section

1 Sequencing COVID-19 at the Sanger Institute
2 How Sanger helped uncover a new COVID-19 variant
3 COVID-19 entry points identified
4 New COVID-19 saliva testing strategy could translate into easy-to-use home device
5 How to scale up by condensing down
6 How SARS-CoV-2 genome sequencing is helping fight the pandemic

1

Sequencing COVID-19 at the Sanger Institute

The Sanger Institute's large-scale genome sequencing capabilities have been deployed to work on coronavirus. Sequence data are being used, in real-time, to inform public health measures and help save lives. The information is being used to identify routes of transmission and track the virus as it evolves and spreads. Researchers are also monitoring the virus for the emergence of new variants and strains, their properties, and to see if the virus is mutating to escape vaccines.

At the start of the pandemic, the Sanger Institute, together with UK Public Health Agencies, academic partners and NHS organisations across the country, formed the COVID-19 Genomics UK Consortium (COG-UK) to sequence and analyse virus genomes.

One of the aims is to use the sequence data to trace the virus as it spreads. Tracking the virus within a hospital, town, country or across the world is possible because genomes mutate. Letters in the genome sequence change as organisms replicate. Individual virus sequences can be placed on a phylogenetic tree, much like a family tree. Researchers can use this data to determine the relatedness of different viruses. The analysis can help identify chains of transmission, super spreading events and fast-growing variants.

As of April 2021

200,000+

virus genomes read

The viral genome data from COG-UK are also combined with clinical and epidemiological datasets to help guide UK public health interventions and policies. Specifically, the consortium provides regular reports to SAGE (Scientific Advisory Group for Emergencies) to inform UK Government policy. Additionally, COG-UK is monitoring for new variants, any change to the biology or properties of the virus, and potential vaccine escape. Data and analysis are immediately made freely available, and information is passed directly to public health authorities and others who need it.

The Institute has set up large-scale genomic surveillance for many diseases, with global collaborations that monitor antibiotic resistance, malaria, and a range of other pathogens. The data are used to inform public health responses around the world. Setting up sequencing for coronavirus was genomic surveillance at a scale and speed never seen before.

Logistics, software, laboratory, technical and scientific teams worked together to rapidly set up, validate and run processes for handling and sequencing coronavirus samples. Thousands of samples are received on site every day from the laboratories that are undertaking COVID-19 tests in communities across the country. These samples consist of the residues of diagnostic swab tests – both positive and negative test samples are received, and the positive ones are picked out for processing and sequencing. Teams have handled tens of millions of samples, and sequenced over hundreds of thousands. Much of the process has been automated, and the Institute has capacity to sequence 20,000 virus samples a week.

Initial funding for the work, totalling £20 million, was contributed by the Sanger Institute, the Department of Health and Social Care, and UK Research and Innovation. Subsequent investment has enabled the Institute to help develop a national real-time genomic surveillance system of COVID-19 to help tackle the pandemic.

If you would like to know more, please see our online articles at www.sanger.ac.uk:

UK launches whole genome sequence alliance to map spread of coronavirus

Wellcome Sanger Institute and COG-UK receive £12.2m UK Government investment for COVID-19 real-time genomic surveillance system

The race to sequence SARS-CoV-2

300+

staff involved

More than

20 million

samples have been handled by the Sanger Institute

2

How Sanger helped uncover a new COVID-19 variant

In December 2020, rapid analyses by teams at the Sanger Institute and the European Bioinformatics Institute (EMBL-EBI), alongside investigations by researchers at Public Health England, Imperial College London, the University of Edinburgh and others, built conclusive evidence that a new coronavirus variant, termed B.1.1.7, is more transmissible. It was found to be largely responsible for the increase in cases seen at the end of 2020 in the UK. These analyses formed the scientific basis for advice that changed Government policy at the time.

Genomic surveillance of COVID in the UK is the highest in the world. As the sequencing hub of the COVID-19 Genomics UK Consortium (COG-UK), Sanger scientists and technicians are constantly sequencing SARS-CoV-2 genomes, taken from across the country.

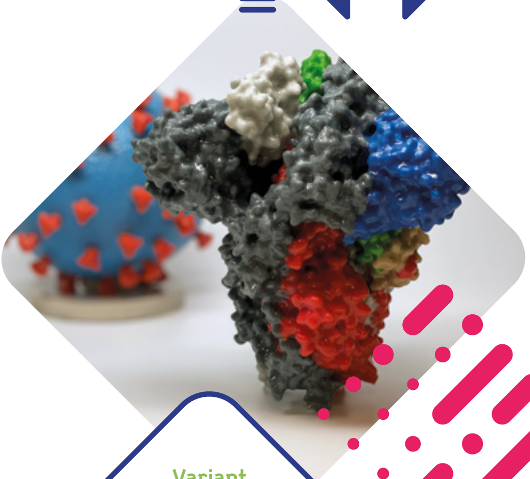
Sanger Institute and EMBL-EBI researchers used the genomic surveillance data alongside daily incidence data to infer the infection dynamics of the B.1.1.7 lineage in November and December. The variant was present in London and the South East of England, and its frequency was increasing both locally and nationally.

During this time, national restrictions were in place, and most other lineages of the virus decreased. However, they saw B.1.1.7 was proliferating despite public health control measures. Statistical analysis showed that the proliferation was most likely due a biological advantage specific to the new lineage, rather than a general failure of viral containment or any artefact in the data.

The variant differs from previous ones with 23 new genomic mutations. Unusually, all of these were seen at the same time; until then the virus had accumulated mutations at a rate of 1–2 per month.

Seventeen of the DNA changes are in parts of the virus' genome that make the proteins it is built from. Several of these are in the spike protein, which is essential for the virus to invade human cells, and are the possible cause of its increased transmissibility.

Their finding that B.1.1.7 is about 50 per cent more transmissible than previous variants, was immediately published openly, and shared with relevant public health agencies. This led to changes in UK Government policy, and has alerted the world to the threat of B.1.1.7.



Variant
B.1.1.7
found to be more transmissible

References
Vöhringer H, *et al* on behalf of the COVID-19 Genomics UK Consortium. Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown. **virological.org**

Rambaut A, *et al* on behalf of the COVID-19 Genomics UK Consortium. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. **virological.org**

3

COVID-19 entry points identified

Researchers used data from the Human Cell Atlas to identify two types of cell in the nose as likely initial infection points for the SARS-CoV-2 virus, along with cell types found in the intestine and eye.

Scientists are applying the tools and data generated by the global Human Cell Atlas initiative, co-founded at the Sanger Institute, to understand COVID-19 transmission and entry into cells.

As part of the Human Cell Atlas Lung Biological Network, Sanger Institute researchers worked with colleagues at the University Medical Centre Groningen, University Cote d'Azur and CNRS, Nice to analyse multiple single-cell RNA sequencing datasets. They studied more than 20 types of tissue from donors unaffected by COVID-19 to see which cell types were actively producing two proteins key for SARS-CoV-2 virus entry.

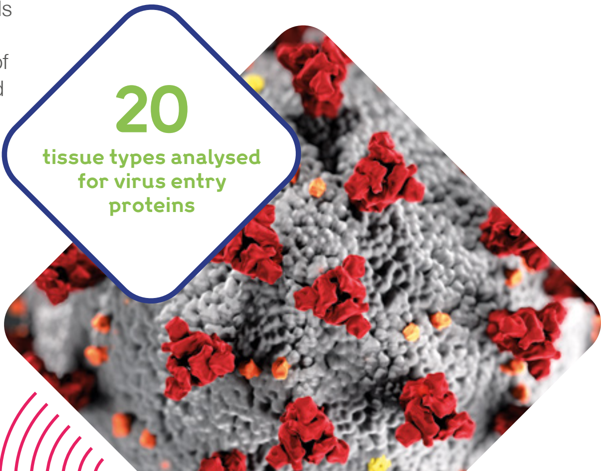
The two proteins – the receptor ACE2 and protease TMPRSS2 – are found together, in the greatest concentrations, in the nose's mucus-producing goblet cells and ciliated cells. This makes these cells the most likely initial infection route, and their location on the surface inside the nose makes them highly accessible to the virus. The cells may also act as reservoirs for virus dissemination within and between individuals.

ACE2 and TMPRSS2 were also found in cells in the cornea of the eye and the lining of the intestine, suggesting other possible routes of infection. Follow-up studies have discovered that the proteins are expressed in specific types of lung cells. The comprehensive datasets are freely available via www.covid19cellatlas.org, offering an important resource for ongoing COVID-19 research.

References
Sungnak W, *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature Medicine* 2020; **26:** 681–687.

Ziegler CGK, *et al.* SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* 2020; **181:** 1016-1035.e19

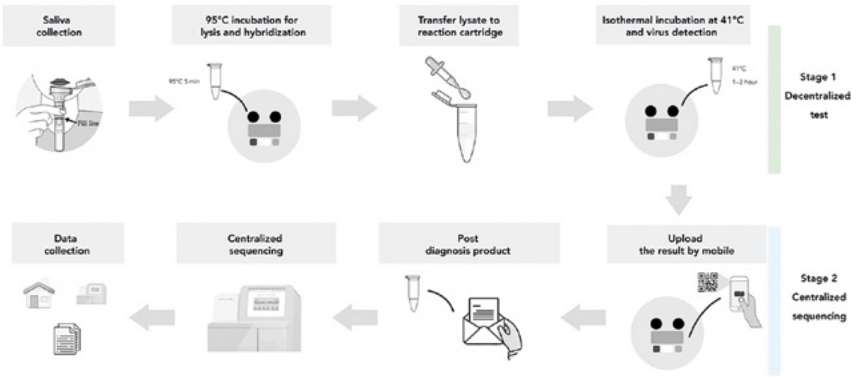
20
tissue types analysed for virus entry proteins



4

New COVID-19 saliva testing strategy could translate into easy-to-use home device

A new two-stage saliva testing strategy for SARS-CoV-2 virus infection that is simpler and easier to use than nasal swab tests has been developed by scientists at the Sanger Institute and their collaborators.



Using samples of synthetic viral RNA mixed with saliva from healthy individuals, the team have developed a new testing strategy for COVID-19. The next phase of the project is further research and development using patient samples. Should the results be validated, it is hoped the equipment required to perform the test could be incorporated into a portable device for home use.

In the first stage of the testing strategy, viral RNA in saliva is amplified using a Nucleic Acid Sequence-Based Amplification (NASBA) isothermal reaction. This amplifies RNA of interest to detectable levels, using a combination of chemicals and a constant temperature. A positive or negative result for SARS-CoV-2 is given within two hours.

The presence of around 50 copies of SARS-CoV-2 viral RNA can be detected with the method, which is comparable sensitivity to the current PCR test used in the UK's public testing strategy.

The second stage of the testing strategy aims to validate the initial result by genetic sequencing of the sample. When the user completes the test, a DNA barcode is added to any viral RNA present in the saliva, allowing the sample to be identified at a later point. The sealed sample would then be posted to a sequencing facility, where tens of thousands of samples could be sequenced simultaneously in around 12 hours to confirm if SARS-CoV-2 is present.

This two-stage process would reduce the risk of incorrect results, and could potentially be scaled up to facilitate population-level testing, which is likely to be needed as the pandemic continues.

Compared to nasal-swab PCR testing, the new testing strategy has the benefit of easier sample collection and a simpler laboratory process. This raises the prospect of developing a home-testing device that is cheap, easy-to-use and scalable.

Reference
Wu Q, *et al.* INSIGHT: A population-scale COVID-19 testing strategy combining point-of-care diagnosis with centralized high-throughput sequencing. *Science Advances* 2020; **7:** eabe5054.

If you would like to know more visit
https://www.insight-covid19.org/

5

How to scale up by condensing down

By refining and developing laboratory protocols, Sanger scientists have increased the Institute's capacity to sequence SARS-CoV-2 genomes – providing more data to researchers studying the virus, worldwide.

The global emergence and spread of SARS-CoV-2 required scientists to create sequencing methods that reliably and rapidly generate high-quality genome data at the lowest cost. The open-access ARTIC protocol, developed by a coalition of researchers, has become the mainstay for SARS-CoV-2 sequencing, and is used worldwide. Sanger researchers and technicians quickly adopted this method, and dovetailed it with in-house sequencing protocols.

To sequence SARS-CoV-2 genomes, genetic material from the virus, RNA, is extracted from sample swabs. It then undergoes a reverse transcriptase reaction to generate DNA. PCR reactions are used to amplify the DNA so that it reaches high enough volumes for the sequencing machines to read.

To be able to sequence large numbers of samples, they are pooled together before loading onto the sequencer. This is enabled by the addition of a unique DNA 'barcode' to each sample during PCR so that it can be separated out, computationally, at the end of the process.

Sanger Institute scientists have refined this process since it was first introduced. Research and Development teams constructed and improved the protocols and were able to reduce the number of steps needed in the laboratory by condensing multiple processes into one. Fewer steps in the process means it is now quicker, a reduced variety of liquid handlers are required, and fewer reagents are needed. It also means that more SARS-CoV-2 samples can be sequenced, substantially increasing capacity at the Institute.

The Sanger Institute's COG-UK Protocols are available at:
https://www.sanger.ac.uk/tool/covid-19-genomics-uk-cog-uk-protocols/

1,000
lineages of SARS-CoV-2 entered the UK in early 2020

6

How SARS-CoV-2 genome sequencing is helping fight the pandemic

The COVID-19 Genomics UK (COG-UK) consortium delivers large-scale and rapid whole-genome virus sequencing to the NHS and public health authorities. As the consortium's central sequencing hub, the Sanger Institute is reading the genomes of hundreds of thousands of virus samples. All data and analysis by COG-UK members is rapidly published and freely available. Below are just some of the ways that the data have been used by public health agencies, academic partners and others to help fight the coronavirus pandemic.

Understanding COVID-19 introduction to the UK

Much early work of COG-UK researchers focused on understanding the fine-scale genetic lineage structure of the SARS-CoV-2 virus, to trace its evolution and transmission. Researchers used the virus genome sequence data, generated by the Sanger Institute and others, to analyse the first wave of infections in the UK. Results showed that over 1,000 lineages of SARS-CoV-2 were introduced in early 2020. Together with a study of introductions into Scotland, the work highlighted that European travel was the main route for COVID-19 into the UK.

References

du Plessis L, *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* 2021; **371:** 708–712.

Lycett SJ, *et al.* Epidemic waves of COVID-19 in Scotland: a genomic perspective on the impact of the introduction and relaxation of lockdown on SARS-CoV-2 COG-UK Report.

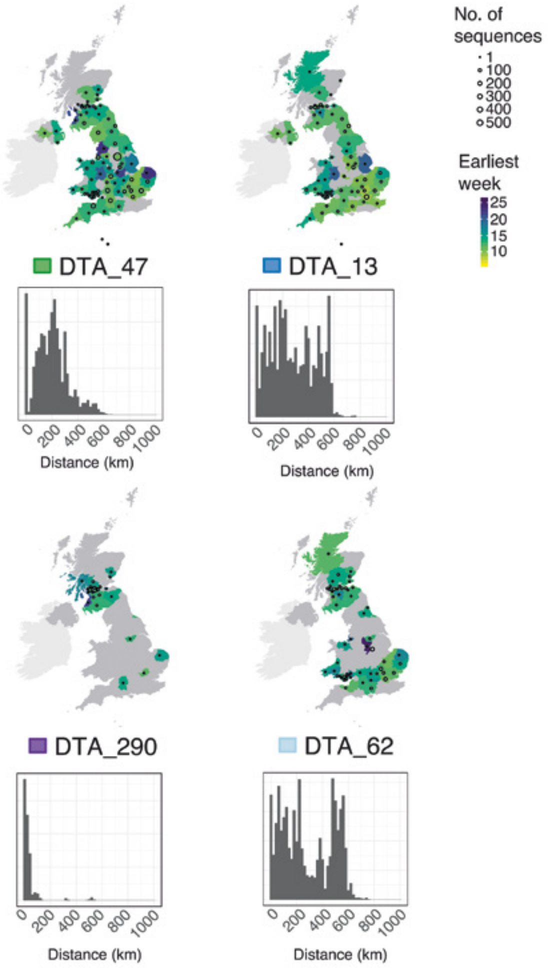


Illustration of the diverse spatial range distributions of UK transmission lineages. Colours represent the week of the first detected genome in the transmission lineage in each location. Circles show the number of sampled genomes per location. The bar charts show the distribution of geographic distances for all sequence pairs within each lineage.

Finding and stopping hidden chains of transmission

By combining viral genome data with people's movements and interactions, COG-UK researchers have generated evidence on SARS-CoV-2 transmission patterns in different settings. A study in a Cambridge hospital uncovered a≈previously hidden route of virus transmission – patient transport. As soon as this was identified, arrangements were changed and the outbreak was stopped.

References

Meredith LW, *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infectious Diseases* 2020; **20:** 1263–1272.

Connor T, *et al.* Welsh SARS-CoV-2 Genomic Insights, October 2020.

Monitoring for new variants

Researchers are also using the genome data to study genetic mutations as they evolve. While most new mutations don't affect the virus's function, some do. Those that affect the viral spike protein that it uses to enter human cells are of particular interest. COG-UK researchers have rapidly identified and assessed new mutations and variants of the virus, alerting UK authorities and the world to the threat they pose.

References

UK Government Press Release: PHE investigating a novel variant of COVID-19

BMJ News: Covid-19: The E484K mutation and the risks it poses

PHE Document: Investigation of novel SARS-COV-2 variant. Variant of Concern 202012/01

Vöhringer H, *et al.* Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown. **Virological.org**

Supporting other research projects

Collaboration both within and beyond the consortium is vital. Within the UK, there are partnerships with GenOMICC (helping to understand COVID-19 critical illness), the UK Coronavirus Immunology Consortium, the G2P-UK (Genotype to Phenotype) National Virology Consortium, REACT (real-time assessment of community transmission of coronavirus), the Office for National Statistics, and the UK COVID-19 wastewater consortium.

Partnerships with COG-UK

GenOMICC – https://genomicc.org/

UK Coronavirus Immunology Consortium – https://www.uk-cic.org/

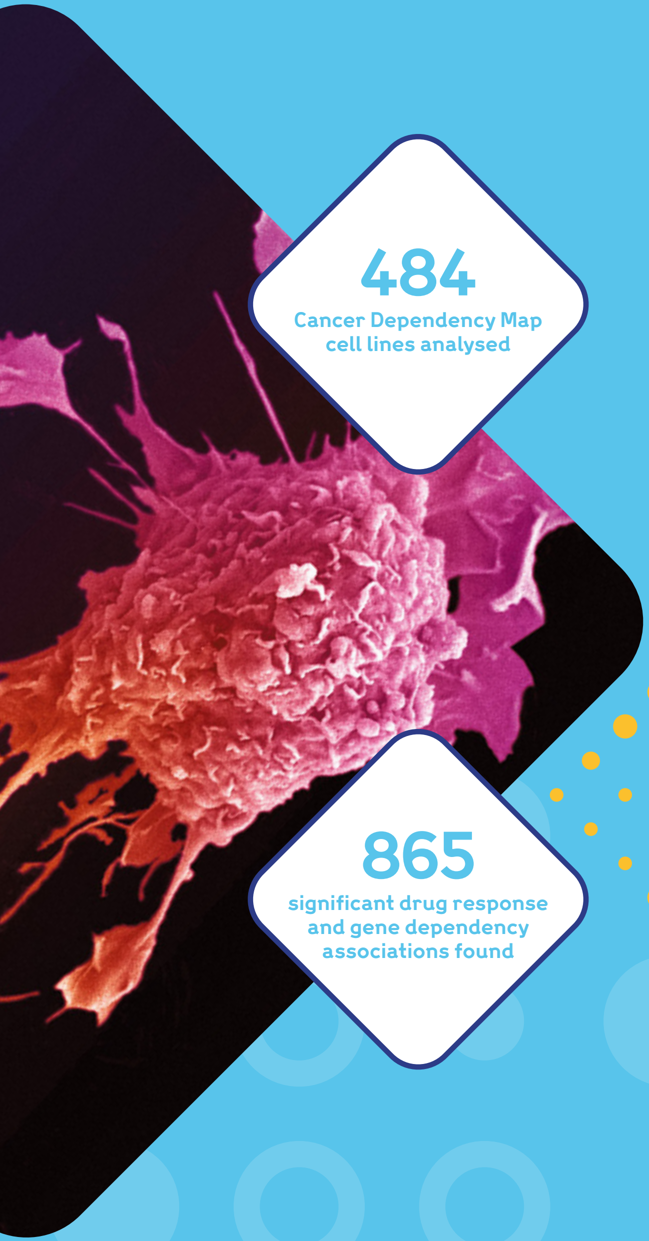
G2P-UK National Virology Consortium – https://www.ukri.org/news/national-consortium-to-study-threats-of-new-sars-cov-2-variants/

REACT – https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/

Office for National Statistics – https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases

UK COVID-19 wastewater consortium – https://www.cogconsortium.uk/news_item/what-can-wastewater-tell-us-about-covid-19/

Cancer, Ageing and Somatic Mutation



484
Cancer Dependency Map cell lines analysed

865
significant drug response and gene dependency associations found

In this section

1 Drug screens and CRISPR combine for cancer treatment insights
2 Cancer-driving mutations colonise the uterus in early life
3 How competition prevents cancer
4 AI catches cancerous genetic changes
5 COSMIC updates drive precision oncology
6 Analysis of CRISPR-Cas9 screens identifies cancer ‘survival genes’ in melanoma

1

Drug screens and CRISPR combine for cancer treatment insights

By combining large-scale pharmacological and CRISPR screens, researchers have brought unparalleled insights into how hundreds of cancer treatments work. The research brings us a step closer to precision cancer medicine.

To help address low success rates in cancer drug development, scientists at the Sanger Institute, EMBL’s European Bioinformatics Institute and AstraZeneca aimed to improve the understanding of how anti-cancer drugs work – something that is not always known, or easy to uncover.

2

Cancer-driving mutations colonise the uterus in early life

Scientists from the Sanger Institute and the University of Cambridge have discovered that cells in healthy tissue from the inner lining of the uterus carry cancer-driving mutations. The team found these mutations often occur early in life.

The rate, pattern, causes and consequences of DNA mutations in cancer cells are well studied. Yet there is limited understanding of the mutations that accumulate in healthy tissues over a lifetime. How these somatic mutations relate to the development of cancer is also unclear.

To understand the processes underlying somatic mutations, the team studied endometrial tissue. They used laser-capture microdissection to isolate 292 histologically normal endometrial glands from 28 women. Using a protocol developed at the Sanger Institute designed for the small amounts of DNA from such samples, each gland was whole-genome sequenced.

They found the glands are clonal populations of cells, each descended from a single progenitor stem cell. Analyses revealed that many endometrial glands carry ‘driver’ mutations in known cancer genes. Using the recent Pan-Cancer Analysis of Whole Genomes dataset as a comparison, they showed that the number of mutations in the healthy cells is much lower than those found in endometrial cancer cells.

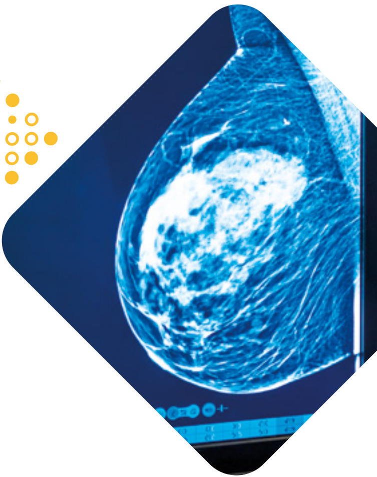
The team exploited data from the Cancer Dependency Map project. This huge resource, created by the Sanger Institute, contains hundreds of deeply characterised cancer cell models (including cell lines) representing patient tumours. The team compared two data sets from the cells:

- CRISPR-Cas9 screens, where each gene has been turned off one-by-one in each cell line. This gives a measure of how critical a gene is for a particular cancer's survival.
- Pharmacological data, where each cell line has been screened against hundreds of anti-cancer drugs (both licenced and in development).

The team measured how drug sensitivity corresponded to CRISPR knock-out of drug targets by searching for associations between the two datasets. In 484 cell lines they identified 865 significant associations between drug response and gene dependency. Further analysis, including of protein–protein interaction networks, enabled the scientists to identify the molecular pathways that approximately half of the 397 drugs tested were affecting. They also estimated drug on-target and off-target activity, specificity, potency and toxicity.

The study revealed a previously unknown association between *MCL1* and *MARCH5* genes in breast cancer cell lines. This will help researchers to understand how MCL1 protein inhibitor treatments work, and when they will be effective.

This enhanced picture of the biological mechanisms of drug responses, and the genomic context in which they happen, will help researchers understand why a drug works on one patient's cancer but not another. Such knowledge is vital for guiding drug combination therapies and combatting resistance to cancer drugs.



Reference
Gonçalves E, *et al.* Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Molecular Systems Biology* 2020; **16:** e904.

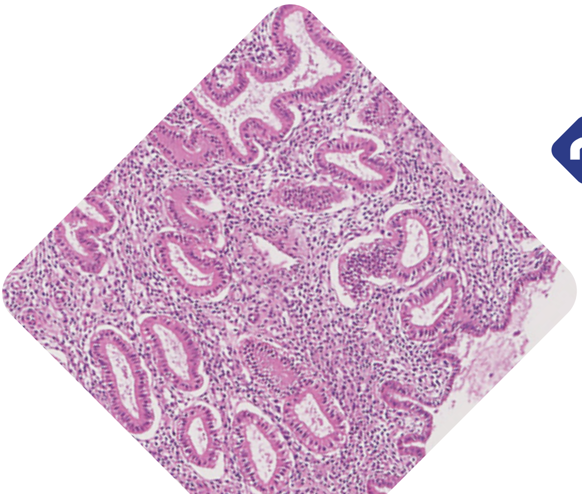
Cancer Dependency Map project
https://depmap.sanger.ac.uk/

To characterise the evolutionary history of the cells, the team constructed genetic family trees of individual endometrial glands. This showed that many driver mutations occur early in life. The team estimate that the conversion from a normal cell with driver mutations, to symptomatic malignancy, is extremely rare.

The results suggest that many cancers are initiated during childhood, and the evolution to disease takes place over a lifetime. Together with parallel studies at the Institute into other types of healthy cells and tissues, the work is revealing the earliest stages of cancer and the landscape of somatic mutations in normal human cells.

Reference
Moore L, *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* 2020; **580:** 640–646.

Pan-Cancer Analysis of Whole Genomes
https://dcc.icgc.org/pcawg



3

How competition prevents cancer

The growth of mutant cells that could lead to cancer is often kept in check by neighbouring cells, researchers from the Sanger Institute and their collaborators have found. The team discovered that competition between equally-matched cells in the oesophagus of mice acted as a brake on one another's growth.

During ageing, cells acquire DNA mutations as they replicate. By middle age, normal human tissues, including the lining of the oesophagus, have become a patchwork of mutant clones.

Some mutations, particularly those affecting known cancer genes such as *TP53* and *NOTCH1*, confer a competitive advantage to cells. These mutant clones multiply more rapidly than normal cells. Yet the vast majority do not go on to form cancers. The processes that underpin mutational selection in normal tissues remain poorly understood, despite their relevance in ageing and cancer.

The team's previous studies have shown that cell clones competing against each other helps to prevent cancer development. To investigate the mechanisms underlying this effect, the team studied the oesophageal epithelium tissue of mice.

The researchers combined genetic lineage tracing and ultradeep sequencing to observe the evolving mutational landscape of the tissue. The team found many mutant clones with multiple cancer-driving gene mutations, including within *Notch1*, *Notch2* and *Trp53* – genes which are also relevant in humans. They showed that a cell's fate depends on the mutations it carries and the nature of its neighbouring cells. Expanding mutant clones that bumped up against cells of similar 'fitness' lost their competitive advantage and reverted towards the balanced growth of normal tissue.

Understanding clonal competition opens up the possibility of designing therapies that can cut the competitive advantage of potentially dangerous clones. By removing these clones or preventing them from becoming dominant, the risk of cancer could be reduced.

Reference
Colom B, *et al.* Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nature Genetics* 2020; **52:** 604–614.

> In tissues like the oesophagus, which is full of cells carrying cancer-driving mutations, it is actually mutant clones cancelling each other out that helps to keep the tissue healthy."
>
> **Dr Bartomeu Colom,**
> first author of the study from the Wellcome Sanger Institute

4

AI catches cancerous genetic changes

EMBL-EBI and Sanger Institute researchers developed an artificial intelligence (AI) algorithm that uses computer vision to distinguish between healthy and cancerous tissues. The ability to visually detect the effects of gene mutations and activity changes could help improve cancer diagnosis, prognosis, and treatment.

Cancer diagnosis and prognosis are usually based on examination of cancer tissue under the microscope and supplemented to analysis of genetic changes in cancer cells. Both approaches are essential to understand and treat cancer, but they are usually seen as separate tools in the clinic.

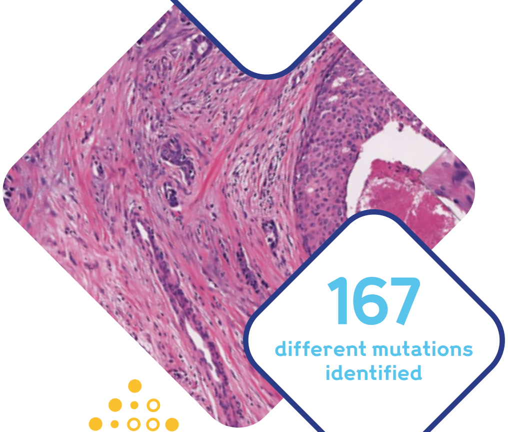
Computer vision algorithms are a form of artificial intelligence that search for, and recognise, specific features in images. The researchers repurposed an algorithm developed by Google – originally used to classify everyday objects, such as lemons, sunglasses and radiators – to distinguish various cancer types from healthy tissue.

Previous studies have used similar methods to analyse images from selected cancer types with certain molecular alterations. The team generalised the approach on an unprecedented scale: they trained the algorithm by showing it 17,355 histopathology slide images from The Cancer Genome Atlas project. The slides covered 28 cancer types and were correlated with the tumours' gene mutations, gene activity and survival data.

After training, the algorithm was able to detect changes in the appearance of tumour cells and tissues associated with specific genetic mutations. In total, it recognised the patterns of 167 different mutations and thousands of gene activity changes.

This approach offers a promising new way to access a tumour's genomic makeup without using sequencing technologies. Using microscopy slides to read the molecular features, cell composition, and estimate survival associated with individual tumours could provide a swift and cost-effective tool to enable clinicians to tailor treatments to a patient's specific needs.

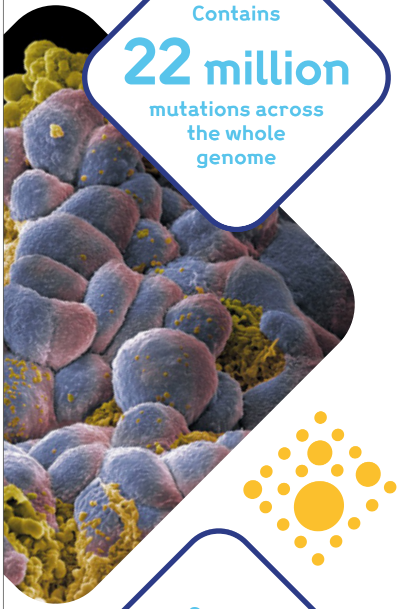
AI trained using
17,355
cancer histopathology slide images



167
different mutations identified

Reference
Yu Fu, *et al*. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* 2020; **1:** 800–810.

5



Contains
22 million
mutations across the whole genome

Spans
1,460
types of cancer

COSMIC updates drive precision oncology

New updates to the Catalogue of Somatic Mutations in Cancer (COSMIC) – Cancer Mutation Census and Actionability – are part of COSMIC's push to enable precision oncology, where a patient's tumour genetics are used to determine their treatments.

COSMIC, launched in 2004, is the world's most comprehensive knowledge centre for cancer mutations. Manually curated by a team of experts at the Sanger Institute, it now includes data from more than 22 million mutations across the genome, spanning 1,460 types of cancer.

The Cancer Mutation Census system integrates biological, biochemical and population information from multiple sources, allowing users to discover and understand which mutations are driving different types of cancer, and which are only a result of the disease. Uniquely, the Cancer Mutation Census doesn't rely solely on AI and Machine Learning algorithms to score mutations. Experts curate the database and calculate biological properties, bringing together multiple different perspectives on what constitutes a driver mutation.

The online tool combines COSMIC's comprehensive resources with the biological features that define driver mutations. It integrates data on multiple different types of mutation seen in cancer genomes, as well as data on gene expression, gene function and 3D protein structure. In addition, the interpretations are fully transparent, allowing researchers to see the biological reasoning and the evidence base used to classify mutations.

The COSMIC Mutation Actionability in Precision Oncology product (Actionability) aims to indicate the availability of drugs that target mutations in cancer. Drugs that target mutations are represented at all stages of drug development, through safety and clinical phases to market and repurposing.

As part of COSMIC's ongoing commitment to enable global research, the updates are included as standard for all licence holders, and free for users at academic and non-profit organisations.

If you would like to know more visit
https://cancer.sanger.ac.uk/cosmic

6

> By marrying bioinformatic techniques with laboratory experimentation, we have been able to identify a new set of potential therapeutic targets for melanoma."

David Adams,
Group Leader the Wellcome Sanger Institute

Analysis of CRISPR-Cas9 screens identifies cancer 'survival genes' in melanoma

New treatment targets for skin cancer have been uncovered by Sanger Institute researchers. Several of the targets have existing drugs available, suggesting these could benefit melanoma patients.

A hallmark of skin cancer are mutations in genes that affect a specific biochemical reaction in cells – the mitogen-activated protein kinase (MAPK) pathway. Targeting this pathway with treatments has transformed care for patients with metastatic melanoma, and can lead to tumour regression. However, melanoma cells eventually acquire resistance to these targeted treatments, and disease relapse is common.

To identify additional treatment targets for melanoma, Sanger scientists studied data produced by teams at the Broad Institute in the US, as part of the Cancer Dependency Map initiative.

The data result from comprehensive CRISPR-Cas9 screens of cancer cell lines grown in the laboratory, which represent patient tumours. In the genome-wide CRISPR-Cas9 screens, each of the approximately 20,000 genes in the genome were individually disrupted in each cell line. The subsequent growth of the cancer cells is monitored, and 'survival genes' – genetic dependencies – of each cell line are identified.

The team analysed data from 28 melanoma cell lines and 313 cell lines of other tumours to identify fitness genes that affect growth. The researchers found 33 genes that were specific to the survival of melanoma cells. This set includes established melanoma fitness genes in the MAPK pathway, as well as new genes not previously associated with melanoma growth. The team verified their findings in laboratory tests.

Several of the new genes encode proteins that can be targeted using available drugs. Some of these proteins are in the MAPK pathway, suggesting further targeting of this would be beneficial.

These data provide a resource of genetic dependencies in melanoma, which could be explored as potential drug targets.

Reference
Christodoulou E, *et al.* Analysis of CRISPR-Cas9 screens identifies genetic dependencies in melanoma. *Pigment Cell & Melanoma Research* 2020; **34:** 122–131.

28
melanoma cell lines and 313 other tumour cell lines analysed

Identified
33
genes necessary for melanoma cell survival

Approximately
20,000
genes were individually disrupted in each cell line

Cellular Genetics

Sequencing and machine learning was used to analyse

500,000

cells to build the Skin Cell Atlas

In this section

1 AI uses gene activity to find new cell types

2 Cell-sifting software reduces costs and errors

3 Comparing visualisation tools for single-cell RNAseq data

4 Psoriatic arthritis discovery could aid targeted treatments

5 Genomics can guide child cancer care

6 Exploring heart health, cell by cell

7 Cell atlas shows origins of eczema and psoriasis

8 Immune defence of the central nervous system uncovered

9 Gut study reveals link to developing Crohn's Disease

10 New research on brain structure highlights cells linked to Alzheimer's

1

AI uses gene activity to find new cell types

Researchers from the Sanger Institute and EMBL's European Bioinformatics Institute (EMBL-EBI) have created a new method to identify different types of cells within a tissue or organ. The machine learning method can replicate the challenging and time-consuming manual, expert annotations that are normally used, and can characterise newly discovered cell types.

Most cells in the human body have the same genome, yet they perform a wide variety of functions – for example, as blood, skin, or nerve cells. In addition, cells of the same type can exist in a number of different states as they develop. In the past, researchers used visible features or the activity of a handful of

2

Cell-sifting software reduces costs and errors

A new computational method developed by Sanger researchers successfully sorts single-cell RNA data from multiple people in a mixed sample. This approach could help to accelerate research across areas of medicine from transplants to obstetrics to malaria.

Single-cell RNA sequencing (scRNAseq) is the bedrock of cell atlas studies. It identifies which genes are switched on in an individual cell, revealing cell types and their functions. This information can reveal how genetic variants in different people affect which genes are expressed during diseases or in response to drugs.

Pooling cells from a number of different people into a scRNAseq experiment helps to reduce costs and errors, and reveals how genetic differences affect gene expression. However, for this approach to be successful, it is essential that the resulting data can be sifted and assigned to each individual donor. Until recently, this could only happen if the researchers had reference information about each individual's genotype.

The new method, Souporcell, developed by Sanger researchers and their colleagues, allows researchers to assign accurately complex scRNAseq data to different donors without the need for any additional information. To achieve this, the method clusters cells together using the genetic variants detected within the scRNAseq data itself. In addition, Souporcell estimates the amount of background RNA from dead cells, genes to sort cells into their type and state. But now, single-cell RNA sequencing (scRNA-seq) gives researchers a picture of all the individual genes active in a single cell – and their levels of activity.

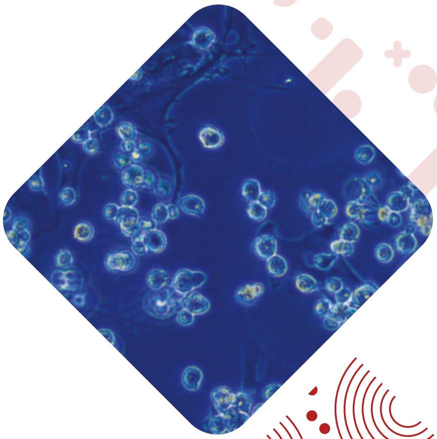
A cell's pattern of gene expression represents its function, and allows scientists to classify or 'cluster' that cell with others that are similar. Until now, annotating cells from scRNA-seq data has required time-consuming human intervention, with automated methods unable to identify cell types or states that had not been previously annotated by experts.

To accelerate single-cell and Human Cell Atlas research, the researchers applied machine learning to these challenges. The tool the team developed – the Single Cell Clustering Assessment Framework (SCCAF) – starts by using a clustering algorithm to group cells from a sample together, based on their gene expression patterns. Each cell cluster is then split into two sets:

- a 'training set' –which a classifying algorithm works on to distinguish cell clusters and make predictions
- and a 'testing set' – which the system then tests its predictions on to assesses how accurate its classifications are.

The tool goes through round after round of refinement until its predictions are 95 per cent accurate when applied to the testing set.

The method has been shown to be highly reliable and fast. It reproduces and refines existing cell type classifications, and helps to reveal new cell types and states from unannotated samples.



Reference
Miao Z, *et al.* Putative cell discovery from single-cell gene expression data. *Nature Methods* 2020; **17:** 621–628.

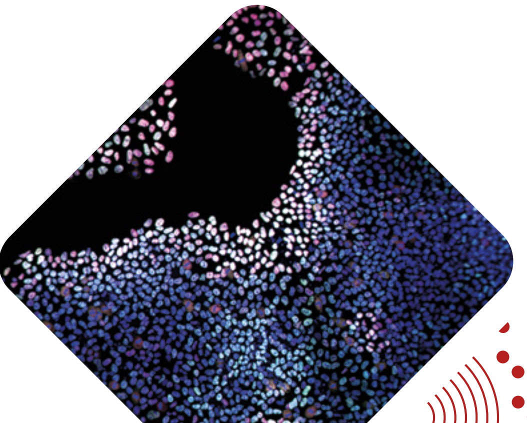
allowing researchers to remove this 'noise', so that only the live cells' data is left.

The method will aid research into cellular environments where cells from more than one person or organism are present, for example:

- transplants, which contain cells from the patient and the organ donor
- placenta, where both mother and baby cells are present
- malaria, where a person's blood may be carrying a number of different parasite strains.

Souporcell achieved high accuracy across a range of experiments, outperforming other methods. The software is free, and openly available for the global research community to use.

Reference
Heaton H, *et al.* (2020) Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nature Methods* 2020; **17:** 615–620.



3

Comparing visualisation tools for single-cell RNAseq data

Sanger researchers have benchmarked and compared tools that allow scientists to visualise single cell RNAseq (scRNAseq) datasets. The work will help scientists share their data in formats that others can quickly and easily explore, and so accelerate science.

The size and volume of scRNAseq data have increased exponentially over the last decade, providing information about the genes and regions of the genome active within an individual cell. When these datasets were first available they included data from a single cell, now, data from millions of cells are combined.

It is not always possible to share scRNAseq data in a scientific report or an article for publication, due to the large number of variables they contain. Several visualisation tools have been created to enable researchers to share data, analyse and interact with these complex datasets. The team reviewed several current tools, and benchmarked those that allow a user to visualise data on the web and share it with others.

They assessed the computational memory and time required to prepare datasets for sharing, testing them with increasing numbers of cells. They also reviewed the user experience and features available in the web interfaces. Their results show that each tool has particular advantages and disadvantages; none stood out as best in all categories.

To address the problem of format compatibility, the team developed a user-friendly software package. This freely available software allows users to easily convert their own scRNAseq datasets into the different formats needed for input into the tools. Through this they have opened up new avenues of scientific discovery and understanding.

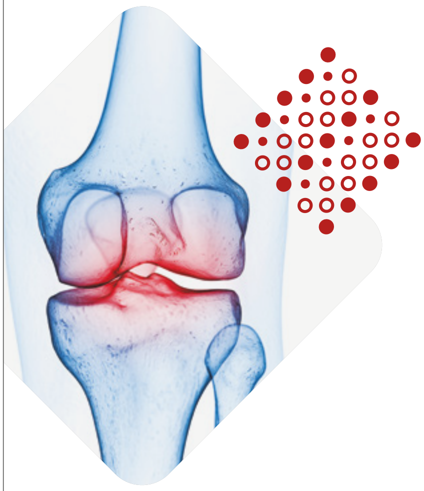
Reference
Cakir B, *et al.* Comparison of visualization tools for single-cell RNAseq data. *NAR Genomics and Bioinformatics* 2020; **2:** lqaa052.

An online training course for using scRNAseq data is available at: **https://www.sanger.ac.uk/tool/scrna-seq-analysis-course/**

4

Psoriatic arthritis discovery could aid targeted treatments

Analysing immune cells' genetic activity has revealed there may be a single trigger for psoriatic arthritis, offering hope for developing a targeted treatment.



Psoriatic arthritis is a long-term, debilitating immune condition where joints become swollen, painful and, in severe cases, permanently damaged. While some treatments are available, there is currently no cure. The cause of the disease, which can affect people at any age, is not understood.

To investigate the basis of psoriatic arthritis, Sanger researchers teamed up with scientists at the University of Oxford. They analysed thousands of immune cells, including T cells, from knee joints and blood of patients with psoriatic arthritis.

Using single-cell RNA sequencing, the scientists determined which genes were active in the cells, creating the largest single cell dataset from psoriatic arthritis joints to date. The team found that T cells in the joints were activated and contributing to inflammation.

The researchers also amplified and sequenced RNA from receptor genes. Their analysis showed that many T cells in the joint were clones, each with the same T cell receptor. The results suggest that there could be a single chemical trigger which binds to the receptor and induces a single T cell to move into the joint and reproduce.

Using machine learning technology, the team then compared T cell receptors from different patients. They showed that these different receptors could be recognising the same chemical. This molecular signal is likely to be the trigger for psoriatic arthritis.

The next stage of the work will be to identify the molecule which is triggering the immune cells. Finding it could allow researchers to develop a targeted treatment, or create therapies to prevent psoriatic arthritis.

Reference
Penkava F, *et al.* Single-cell sequencing reveals a clonal expansion of pro-inflammatory synovial CD8 T cells expressing tissue homing receptors in psoriatic arthritis. *Nature Communications* 2020; **11:** 4767.

5

Genomics can guide child cancer care

Sanger researchers have shown that a rare childhood cancer – bilateral neuroblastoma – can be due to two tumours that arise independently of each other. The discovery has uncovered the roots of neuroblastoma and shows that genomics can guide treatment decisions.

Neuroblastoma is a rare, highly aggressive cancer, affecting about 100 children each year in the UK. It develops from specialised nerve cells – neuroblasts – left behind from a baby's development in the womb. It most commonly occurs in one of the adrenal glands above the kidneys but, occasionally, bilateral tumours occur.

To understand the genetic histories of how these tumours develop, Sanger scientists worked with researchers at the University of Cambridge, Great Ormond Street Hospital for Children NHS Trust and others to study the genomes of tumours from multiple sites in two children with the condition.

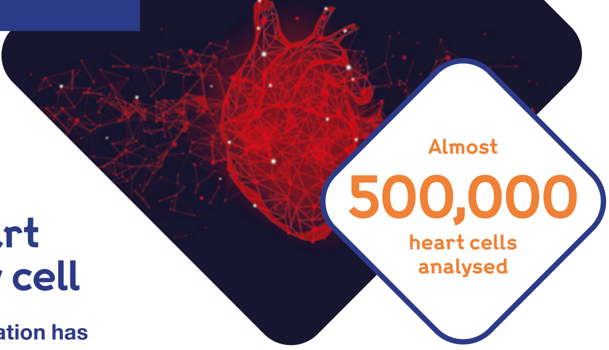
Using advanced analysis techniques, the team extensively sequenced and compared the tumour mutations with those in healthy cells to identify the somatic mutations (DNA changes acquired over a lifetime) the tumour cells had accumulated as they grew. This application of evolutionary genomics showed that the neuroblastoma tumours in an individual arose independently at the very earliest stages of life, within a few cell divisions after fertilisation.

Understanding whether tumours within a patient arise independently from each other is vital for clinicians when deciding the best treatment options. A tumour that has spread is more aggressive – and requires more intensive treatment – than one that has stayed in its original location. Knowing that two tumours may be distinct means that clinicians can consider less intensive treatments, with fewer side effects.

Reference
Coorens THH, *et al.* Lineage-independent tumors in bilateral neuroblastoma. *New England Journal of Medicine* 2020; **383:** 1860–1865.

> We are in a world now where genome sequencing tumours is a part of healthcare. Using genomics to analyse a tumour's origins can give us detailed insight into what we are dealing with and how to tackle it."
>
> **Dr Sam Behjati,**
> lead author of the study from the Wellcome Sanger Institute and Addenbrooke's Hospital, Cambridge

100
UK children develop neuroblastoma each year



6

Exploring heart health, cell by cell

An international collaboration has unlocked the secrets of the heart's form and function by creating the most detailed Heart Cell Atlas. This map will help understand heart disease and guide personalised medicine.

Heart disease is a leading cause of death worldwide, killing millions each year. To understand heart disease it is vital to learn exactly what is happening in the individual cells that coordinate to pump blood around the body 100,000 times a day.

To build the human Heart Cell Atlas, researchers from the Sanger Institute collaborated with scientists from the Max Delbrück Center for Molecular Medicine, Germany, Harvard Medical School and Imperial College London. By combining cutting edge single-cell technologies with artificial intelligence methods, the team analysed nearly half a million individual cells from six different areas of the heart from 14 organ donors.

The work, part of the Human Cell Atlas initiative, revealed which genes were active in each heart cell for the first time. In total, the researchers identified 11 different cell types, and described more than 62 different cell states at a level of detail not seen before.

The atlas also revealed the intricate network of blood vessels in the heart, and how each area of the heart had specific sets of cells; highlighting their separate developmental origins and potentially different responses to treatment. Freely available online to researchers globally, this map will allow exploration of the healthy heart in more depth than ever before. It will also help to accelerate understanding of heart disease and guide improvements in treatments and personalised medicines.

> Openly available to researchers worldwide, the Heart Cell Atlas is a fantastic resource, which will lead to new understanding of heart health and disease, new treatments and potentially even finding ways of regenerating damaged heart tissue."
>
> **Dr Sarah Teichmann,**
> a senior author from the Wellcome Sanger Institute
> and co-chair of the Human Cell Atlas Organising Committee

Reference
Litviňuková M, *et al.* Cells of the adult human heart. *Nature* 2020; DOI: 10.1038/s41586-020-2797-4.

Read more about Heart Cell Atlas
www.heartcellatlas.org

7

Cell atlas shows origins of eczema and psoriasis

Mapping how healthy adult skin develops, cell by cell, has revealed clues to the underlying cause of inflammatory skin disease. The new atlas of skin cells shows that cellular processes from development are re-activated in cells from patients with inflammatory skin disease. The findings point to new drug targets for conditions including eczema and psoriasis.

Atopic eczema and psoriasis are chronic, painful skin conditions caused by an overactive immune system. Their cause is unknown and while treatments can help with symptoms, there is no cure.

To understand how skin forms and how this relates to health and disease, researchers from the Sanger Institute, Newcastle University and Kings College London studied cells from developing skin, healthy adult skin, eczema and psoriasis patients.

Using single-cell sequencing technology and machine learning, the team analysed more than 500,000 cells to identify the genes switched on in each. This allowed the researchers to assess each cell's functions, including their molecular communication signals.

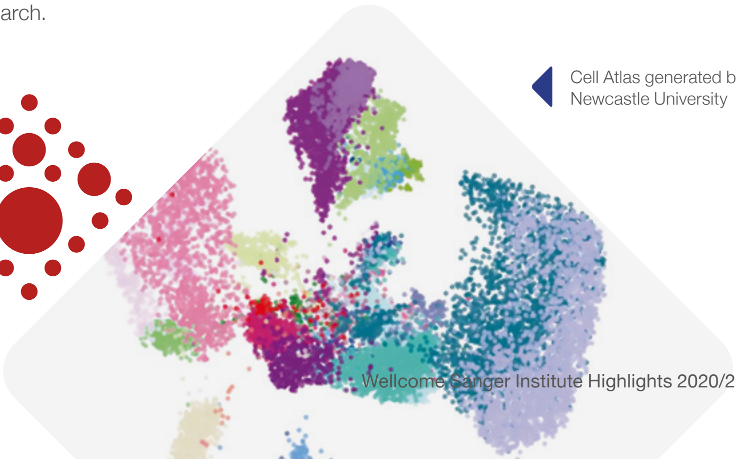
The team discovered that skin cells affected by eczema and psoriasis were sending a signal normally only seen when healthy skin is first developing. This signal summons immune cells to form a protective layer in the developing skin, but in adults it could over activate the body's immune cells and cause disease.

This understanding offers new targets for treatment development. It is possible that other inflammatory conditions, such as rheumatoid arthritis or inflammatory bowel disease, could be triggered in the same way, and so the work opens up new approaches for research.

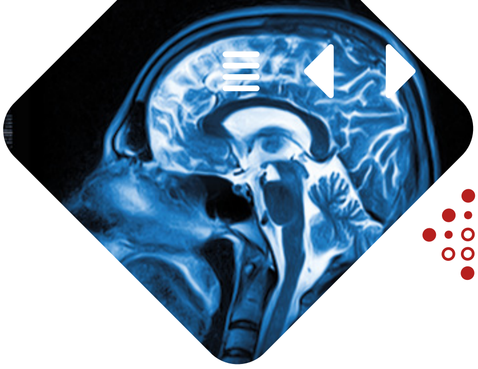
The atlas also reveals how healthy skin tissue develops, and the cell types present in adult skin. This knowledge could provide the foundations for regenerative medicine, for example, helping researchers to grow skin to treat burns.

The work was carried out as part of the Human Cell Atlas initiative and the data are freely available to the global research community.

Reference
Reynolds G, *et al.* Developmental cell programs are co-opted in inflammatory skin disease. *Science* 2021; **371:** eaba6500.



Cell Atlas generated by Newcastle University



8

Immune defence of the central nervous system uncovered

A study by researchers at the University of Cambridge, the National Institutes of Health (NIH), US, and the Sanger Institute has uncovered how the brain is uniquely protected against invading pathogens by immune cells primed for defence in the gut.

The central nervous system, which includes the brain and spinal cord, has historically been thought of as an immune-privileged site, with no access for immune cells. It is surrounded by a barrier made of three layers of watertight, impermeable tissue – the meninges. But recent data have shown that the meninges themselves contain a diverse population of immune cells.

The team studied the details of these cells in mice. They found that the meninges are home to immune cells known as plasma cells, which secrete antibodies. These cells are positioned next to large blood vessels running within the meninges, allowing them to secrete antibodies to defend the perimeter of the brain.

The researchers found that the antibodies were Immunoglobulin A (IgA), which are usually made in the gut lining or in the lining of the nose or lungs – mucosal surfaces that interact with the outside environment.

Sequencing the relevant genes in both the meningeal and gut plasma cells confirmed that meningeal IgA cells originate in the intestine. Removing these cells left mice vulnerable to infections in the brain, showing that the plasma cells and the IgA they produce are essential for defending the central nervous system.

The team confirmed the presence of IgA cells in human meninges by analysing samples that were removed during surgery, showing that this defence system is likely to play an important role in defending people from infections of the central nervous system – meningitis and encephalitis.

> The exact way in which the brain protects itself from infection, beyond the physical barrier of the meninges, has been something of a mystery, but to find that an important line of defence starts in the gut was quite a surprise.”
>
> **Professor Menna Clatworthy,**
> Associate Faculty at the Wellcome Sanger Institute

Reference
Fitzpatrick, Z *et al*. Gut-educated IgA plasma cells defend the meningeal venous sinuses. *Nature* 2020; **587:** 472–476.

9

Gut study reveals link to developing Crohn's Disease

The very early stages of gut development have been mapped in incredible detail by researchers from the Sanger Institute and University of Cambridge. They discovered specific cell functions in the developing gut that appear to be reactivated in the gut of children with Crohn's Disease. The findings offer potential for understanding Crohn's and other gut diseases, and for creating new treatments.

Crohn's Disease is a type of Inflammatory Bowel Disease affecting around one in every 650 people in the UK. It has increased dramatically in recent decades, especially in children. Symptoms can be aggressive, and include lifelong abdominal pain, diarrhoea and fatigue. There is no cure, the cause is not understood, and treatments are often ineffective.

The researchers used cutting-edge single-cell RNA sequencing technology to identify active genes in individual cells of the developing human gut lining, six to ten weeks after conception. They compared this with the epithelium from the guts of children with Crohn's Disease, and revealed that some of the cellular pathways active in the epithelium of the foetal gut appear to be reactivated in Crohn's Disease.

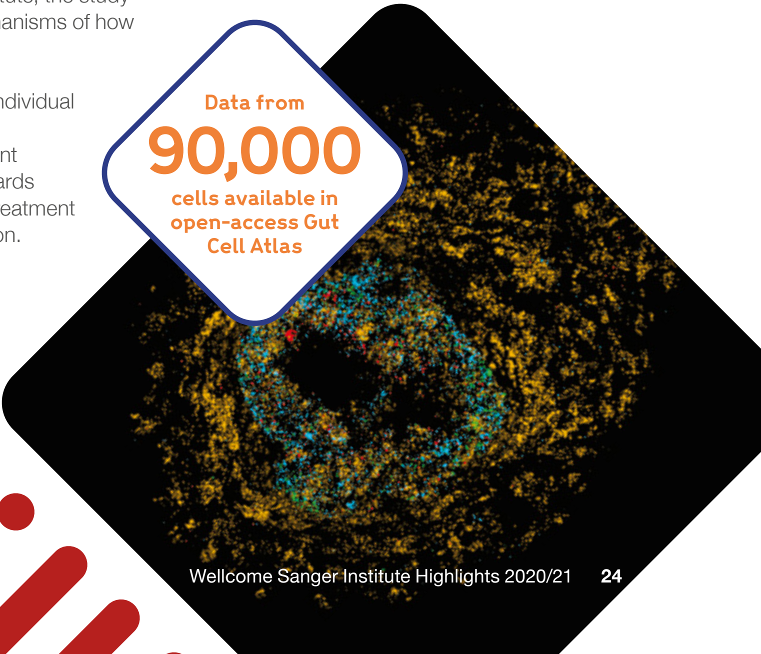
Part of the global Human Cell Atlas initiative to map every cell type in the human body, which was co-founded by the Wellcome Sanger Institute, the study also reveals intricate mechanisms of how the gut develops.

The data from all 90,000 individual cells are openly available.
The results are an important resource for research towards better management and treatment of this devastating condition.

Reference
Elmentaite R, *et al*. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. *Developmental Cell* 2020; **55:** 771–783.E5.

Gut Cell Atlas
www.gutcellatlas.org

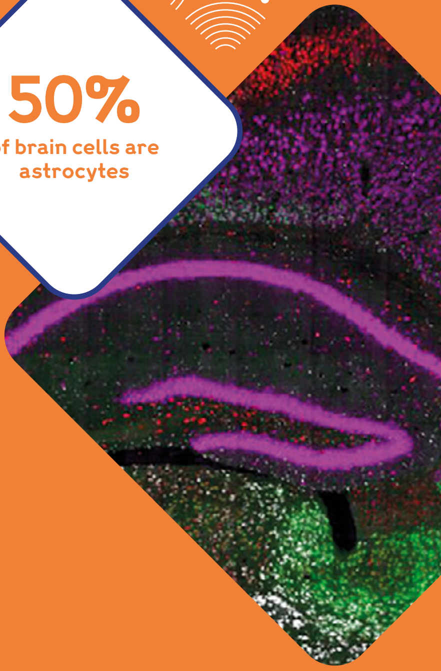
Data from
90,000
cells available in open-access Gut Cell Atlas



10

> This discovery redefines our view of the structure of the mammalian brain… If you want to properly understand how our brains work, you have to consider how astrocytes are organised and what role they play.”

Dr Omer Bayraktar,
Group Leader at the Wellcome Sanger Institute



50%
of brain cells are astrocytes

New research on brain structure highlights cells linked to Alzheimer’s

Sanger Institute researchers and their collaborators have uncovered previously unseen arrangements of cells, redefining the known structure of the mammalian brain.

Star-shaped astrocyte cells have a wide range of roles in brain development, function and in supporting nerve cells. They are also implicated in diseases, including multiple sclerosis and Alzheimer’s. Yet the organisation and subtypes of astrocytes are not well understood.

Researchers at the Sanger Institute and the Wellcome-MRC Cambridge Stem Cell Institute took a new approach to study brain cells in detail. They first identified genes of interest in the brain, using publicly available datasets. They then assessed gene activity for 46 genes in single astrocyte cells, taken from thin slices of the brain. The automated method, LaSTmap, allows single cells to be studied in bulk. The results give the most detailed picture of cell types and their locations in the brain to date.

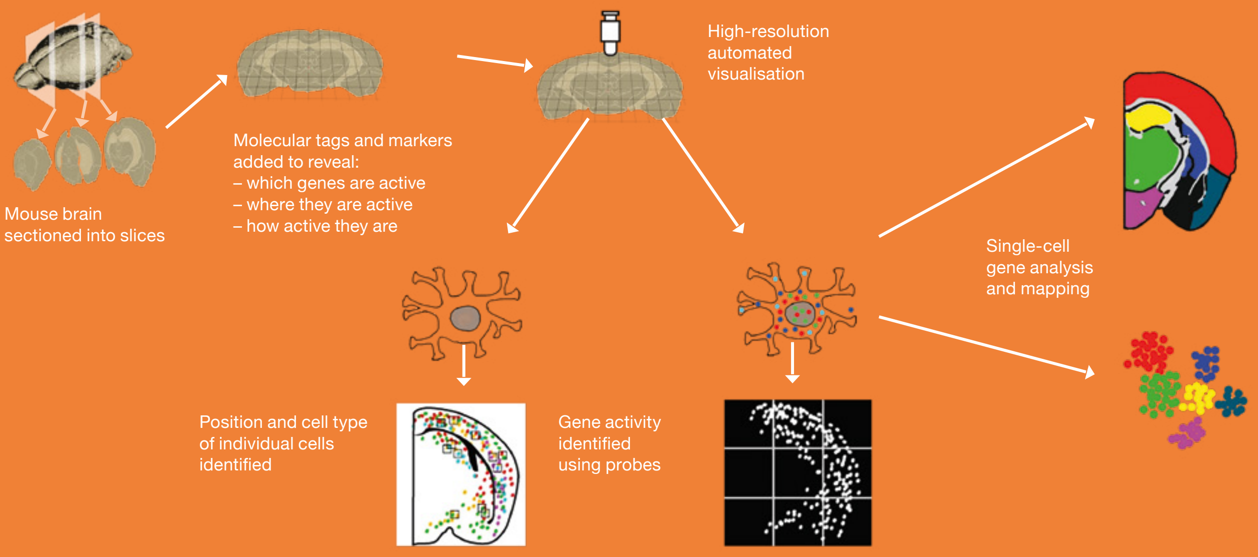
The team discovered that astrocytes are not uniform, as had been previously thought. Similarly to nerve cells in the brain, they have distinct types depending on their location. Also like nerve cells, they form six layers in the cerebral cortex – though the boundaries are not identical.

This work was mostly done in mice. The team showed that astrocyte layers are established in the cortex of new-born mice and mostly persisted in adult mice. They were able to use LaSTmap on human samples too, including archival ones. They saw that the same layers of astrocytes in mice, are present in the human cortex too.

The findings will shape the study of neurological and neurodegenerative diseases, where understanding of the roles of specific cell types is crucial.

Reference
Bayraktar OA, *et al.* Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. *Nature Neuroscience* 2020; **23:** 500–509.

Discovering the location and activity of every cell in the brain



Mouse brain sectioned into slices

Molecular tags and markers added to reveal:
– which genes are active
– where they are active
– how active they are

High-resolution automated visualisation

Single-cell gene analysis and mapping

Position and cell type of individual cells identified

Gene activity identified using probes

Human Genetics

Open-access DECIPHER database contains

165,732

clinical and symptom observations

In this section

1 Missing pieces of human genetics and history discovered

2 Invaders and traders leave little genetic legacy

3 How IBD raises cancer risk

4 Global studies pave way to predict blood disease risk

5 How open-access data DECIPHERs rare disease

1

Missing pieces of human genetics and history discovered

Two studies have provided the most comprehensive analysis of human genetic diversity to date. The work identified new genetic variation, added missing pieces to the reference human genome sequence, and gave deep insights into our evolutionary past.

As part of the Human Genome Diversity Project, scientists at the Sanger Institute and the University of Cambridge sequenced and analysed the genomes of 929 people from 54 geographically, linguistically and culturally diverse populations around the globe. This was the first time that the latest high-quality sequencing technology has been applied to such a large and diverse set of people.

The team identified 67.3 million single-nucleotide polymorphisms, 8.8 million small insertions or deletions, and 40,736 copy number variants in the genomes. They found millions of previously unknown DNA variations that are exclusive to a geographical region. Though most were rare, they included some previously unknown variations that are common in certain African and Oceanian populations.

2

Invaders and traders leave little genetic legacy

The Near East is often described as the cradle of civilisation. It was a crossroads for the ancient world's empires, and invasions over centuries caused enormous changes in cultures, religions and languages in the region. But what genetic impact did this have on the population? Sanger researchers studied ancient skeletons to find out.

Over the centuries, the Near East has been ruled by the Egyptians, Babylonians, Assyrians, Persians, Greeks, Romans, Crusaders, Arabs, and Ottomans. Historical records and archaeological findings show that most of these had permanent cultural effects on the local population, including changes to religion and languages.

929
people

54
geographically, linguistically and culturally diverse populations

67.3m
single-nucleotide polymorphisms

40,736
copy number variants in the genomes

8.8m
small insertions or deletions

Such variations may influence the susceptibility of different populations to disease, but as medical genetics studies have predominantly been conducted in populations of European ancestry, the implications of most of the variants is not known. Identifying them is a first step towards fully expanding the study of genomics to under-represented populations.

The researchers also identified 126,018 structural variations in the genomes, affecting larger regions of DNA. Structural variations are more difficult to detect than smaller changes, and are not well studied in large datasets. They found variations in medically-important genes in populations from Papua New Guinea that were inherited from Denisovan ancestors. Other variations affect digestion and the immune system.

This analysis added new regions of sequence to the human reference genome, the world standard for all human genetics studies, which is nevertheless incomplete.

Studying the evolutionary history of humankind, the researchers compared the genomes to archaic sequences. They found evidence that the Neanderthal ancestry of modern humans can be explained by just one major 'mixing event', most likely involving several Neanderthal individuals coming into contact with modern humans shortly after the latter had expanded out of Africa. In contrast, several different sets of DNA segments inherited from Denisovans were identified in people from Oceania and East Asia, suggesting at least two distinct mixing events.

The discovery of small amounts of Neanderthal DNA in west African people, most likely reflecting later genetic backflow into Africa from Eurasia, further highlights how human genetic history is characterised by multiple layers of complexity. Until recently, it was thought that only people outside sub-Saharan Africa had Neanderthal DNA.

The data, freely available, is the most detailed representation of the genetic diversity of worldwide populations to date.

References

Almarri MA, *et al*. Population structure, stratification, and introgression of human structural variation. *Cell* 2020; **182:** 189–199.e15.

Bergström A, *et al*. Insights into human genetic variation and population history from 929 diverse genomes. *Science* 2020; **367:** eaay5012.

84,729,690
mutations studied

81
mutational signatures identified

However, previous genetic research by Sanger scientists has shown that present-day people in Lebanon are mainly descended from local people in the Bronze Age (2100–1500 BCE), with very few lasting traces of even the 11th–13th Century invasions of Crusaders.

To understand more, researchers at the Sanger Institute, University of Birmingham and French Institute of the Near East in Lebanon, studied the DNA of 19 people from four archaeological excavation sites in Beirut. Working closely with archaeologists, the researchers transferred the bones to a dedicated ancient DNA extraction laboratory in Estonia before sequencing and analysing them at the Sanger Institute.

Using recent advances in technologies to study the ancient and damaged DNA, the team analysed genomes covering the time period 800 BCE–200 CE and, by combining with previous ancient and modern data, created an eight-point time line across the millennia.

While the invasions and conquests may have been revolutionary for the elite rulers, the team only detected lasting genetic changes in the local people from three time periods:

- the beginning of the Iron Age (about 1,000 BCE)
- the arrival of Alexander the Great (beginning 330 BCE)
- the domination of the Ottoman Empire (1516 CE).

References

Haber M, *et al.* A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years. *American Journal of Human Genetics* 2020; **107:** 149–157.

Haber M, *et al.* A transient pulse of genetic admixture from the Crusaders in the Near East identified from ancient genome sequences. *American Journal of Human Genetics* 2019; **104:** 977–984.

Haber M, *et al.* Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences. *American Journal of Human Genetics* 2017; **101:** 274–282.

3

How IBD raises cancer risk

By applying techniques developed to study colon cancer, Sanger researchers explored how inflammatory bowel disease increases cancer risk.

Inflammatory bowel disease (IBD) – primarily ulcerative colitis and Crohn's disease – affects approximately 6.8 million people worldwide. It is thought that inflammation occurs as a result of an inappropriate immune response to gut microbes, but the causes of the disease remain unknown. Not only is the condition highly disruptive to a person's life, it also increases their risk of developing gastrointestinal cancers compared with the general population.

To investigate the genetic changes underpinning IBD, Sanger scientists used laser-capture microdissection to separate out 446 individual crypts – the tiny cavities that make up colon tissue – from 46 patient samples. These were whole-genome sequenced and compared with 412 crypts from 41 individuals without the condition.

They discovered that rate of DNA change within IBD-affected colon cells was more than double that of healthy colon cells, increasing the risk of developing cancer.

The study uncovered evidence of an evolutionary process whereby mutations in particular genes are under positive selection pressure. Some of these positively selected mutations were enriched in genes associated with colorectal cancers, shedding further light on the link between IBD and cancer. The researchers also detected evidence of positive selection of mutations in genes associated with immune system regulation in the gut and the ability of the cells to fend off the bacteria resident in the colon.

These discoveries provide valuable insights into the microevolution taking place within the body, and how it affects the development of IBD and colorectal cancers.

Reference

Olafsson S, *et al*. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* 2020; **182:** 672–684.e11.

6.8m

people worldwide affected by IBD

4

Global studies pave way to predict blood disease risk

Two studies from Sanger Institute researchers and colleagues from over 100 international research institutions have identified more than 7,000 regions of the human genome that control blood cell characteristics and our risk of developing blood disorders.

The scientists analysed genomic and health data from hundreds of thousands of volunteers taking part in the UK Biobank, Blood Cell Consortium and other global studies. Comparing the DNA sequences and health measurements revealed how genetic variations affect the physical characteristics of blood cells.

Blood cells play an essential role in health, including in the immune response. Blood disorders, such as anaemia, haemophilia and blood cancers are a significant global health burden, and blood cells play an important role in complex, common conditions, such as asthma and cardiovascular disease.

The analysis is the largest Genome-Wide Association Study (GWAS) of blood cells to date, and included people of European, East Asian and African American ancestry. The researchers discovered 7,193 distinct genetic regions associated with 29 blood cell characteristics affecting blood cell generation, development and maintenance.

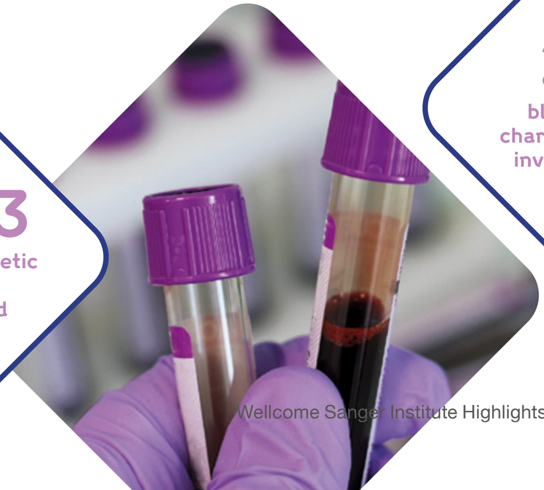
The work highlights the power of large international datasets to uncover variants with small effects. When the effects of these hundreds of small variants are combined into a polygenic score, they can predict predisposition to complex diseases. In future, these scores could be used in healthcare to help predict personal risk of developing blood disorders. The scientists also showed that common variants with small effects may have a role in blood disorders previously thought to be caused by a change to a single gene.

This research also emphasises the importance of studying diverse populations to get the full genetic picture of health and disease. They highlight that studying populations of European ancestry alone misses variations with impacts on health and disease, as well as the differing effects of genetic variations in different populations.

References

Vuckovic D, *et al*. The polygenic and monogenic basis of blood traits and diseases. *Cell* 2020; **182:** 1214–1231.e11.

Chen M-H, *et al*. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* 2020; **182:** 1198–1213.e14.



7,193

distinct genetic regions identified

29

blood cell characteristics investigated

5

> After the human genome was sequenced in 2004, we created DECIPHER to put that data to use. Our aim remains the same as when we first launched – to understand the significance of genetic changes in health and disease."

Dr Helen Firth,
Honorary Faculty at the Wellcome Sanger Institute and consultant clinical geneticist at Addenbrooke's Hospital

38,000
patient details

250+
contributing centres worldwide

165,732
clinical and symptom observations

How open-access data DECIPHERs rare disease

DECIPHER is one of the world's largest and most comprehensive rare diseases database, bringing emerging knowledge of human genetics and genomics to the forefront of clinical diagnosis and treatment. Powered by open-access data, its latest updates allow researchers and clinicians to input, share and interpret all types of genomic variants across the whole genome.

Data from DECIPHER has been used in thousands of studies. One of the latest identified 28 new genes associated with developmental disorders. The work could enable 500 families affected by rare conditions to receive a diagnosis.

Rare conditions affect 1 in 17 people in the UK. The vast majority having an underlying genetic cause, but often the precise DNA mutation is unknown, and unique to the individual. Finding the cause of rare genetic diseases is important. Such a diagnosis may bring understanding, support from families with the same genetic variation, improved care and, for some, treatments. The genetic cause of a disease is also the basis of research into new treatments.

The DECIPHER database was created by Sanger researchers and clinical collaborators in 2004 to identify the links between genetic variants and rare diseases, and interpretation of their meaning. More than 250 centres across the world have added details from over 38,000 patients. The data includes 165,732 clinical and symptom observations, and details of tens of thousands of genetic variations.

A recent update enables users to input, view and share genomic variants, of any type, in any region of the genome. Previously, users could view changes of a single letter of genetic code, and changes in number of copies of a gene. Now, additional variants have been included, which can also be grouped to better reflect a condition's genetic complexity:

- regions of repeats
- regions of the genome that have been inverted or inserted
- chromosome variants, including regions inherited in an atypical way.

In an international study, data from DECIPHER was combined with datasets from Radboud University Medical Center and GeneDx. Sanger researchers and their collaborators analysed genetic sequence data from 31,058 children affected by rare genetic conditions, together with sequence data from their unaffected parents. The large scale gave the team enough statistical power to search for previously undiscovered mutations.

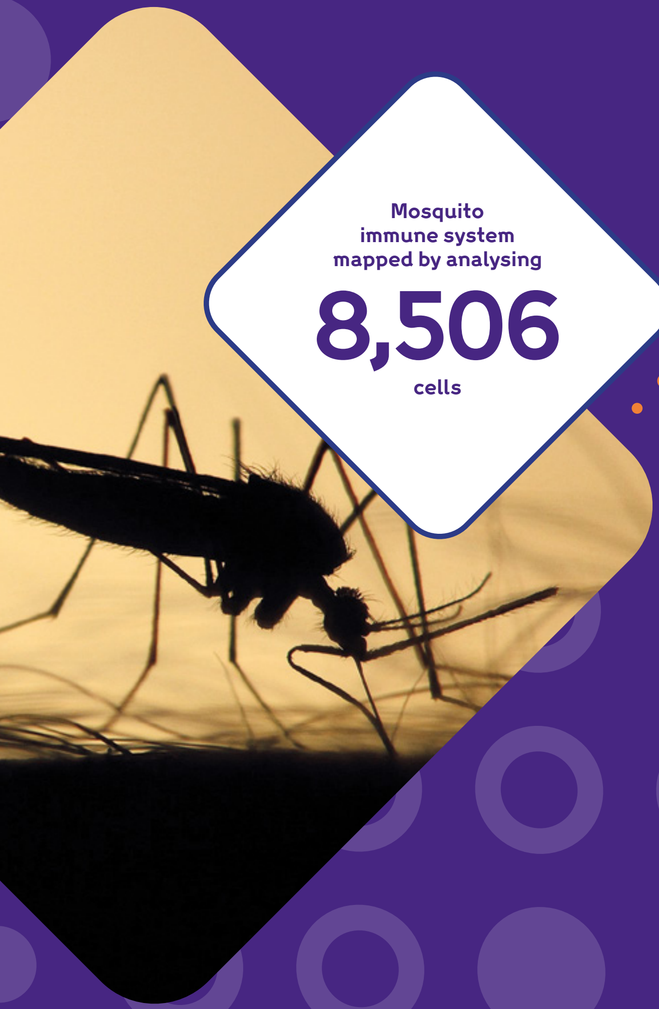
The researchers identified 28 genes newly-associated to rare developmental conditions, enabling diagnoses for around 500 families.

The researchers also applied statistical modelling to the data to estimate that approximately 1,000 more development disorder-associated genes remain undiscovered. They calculate that a 10-fold increase in data is needed to find them all.

The work demonstrates the value of combining genomic and healthcare data to gain new insights that improve the lives of patients. The researchers stress that open-access to anonymised patient data is vital to understand rare conditions and help families living with them.

Reference
Kaplanis J, *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 2020; **586:** 757–762.

Parasites and Microbes



Mosquito immune system mapped by analysing

8,506

cells

In this section

1 Mapping the mosquito immune system cell by cell

2 How a genetic gift protects against malaria

3 Genomic surveillance tracks comeback of neglected tropical disease

4 How to build a parasitic worm

5 Silently incubating a killer inside the nose

6 Superbug's stealthy spread revealed

7 Getting to the guts of a delicate balance

8 Building a global picture of Black Fever

9 How genomics changed Argentina's health policy

1

Mapping the mosquito immune system cell by cell

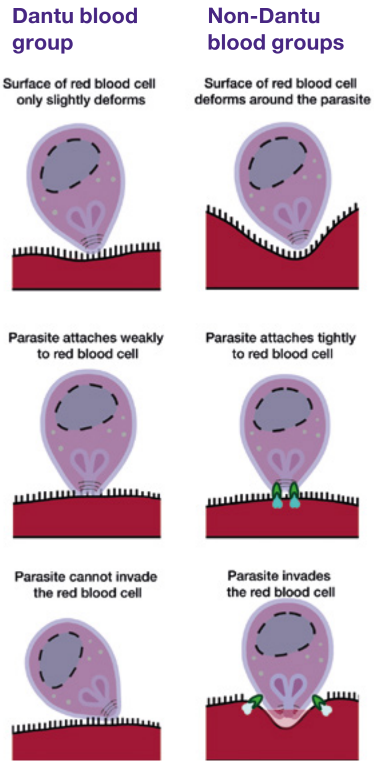
Using cutting-edge single-cell techniques, an international collaboration has created the first cell atlas of mosquito immune cells, to understand how mosquitoes fight malaria and other infections. One of the cell types researchers uncovered may help to limit malaria infections.

Malaria affects more than 200 million people worldwide, and killed an estimated 405,000 people in 2018 alone, mostly children under five. Caused by *Plasmodium* parasites, it is spread via the bites of female *Anopheles* mosquitoes.

2

How a genetic gift protects against malaria

An African-UK collaboration has found that red blood cells in people who inherit the rare Dantu blood variant have a higher surface tension that prevents them from being invaded by the world's deadliest malaria parasite, *Plasmodium falciparum*.


Dantu blood group
Non-Dantu blood groups
Surface of red blood cell only slightly deforms
Surface of red blood cell deforms around the parasite
Parasite attaches weakly to red blood cell
Parasite attaches tightly to red blood cell
Parasite cannot invade the red blood cell
Parasite invades the red blood cell

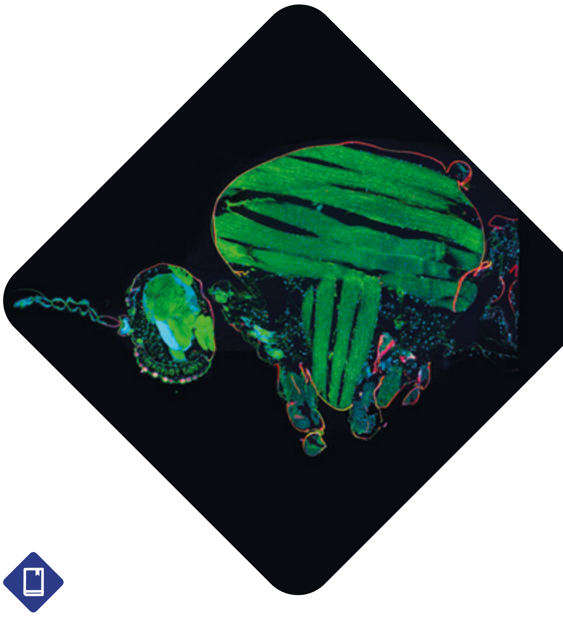
The mosquito immune system controls how the insect can tolerate or transmit pathogens, yet little is known about the exact cell types involved. In this first in-depth study of mosquito immune cells, the team of researchers studied *Anopheles gambiae*, which transmits malaria, and *Aedes aegypti*, which carries the viruses that causes Dengue, Chikungunya and Zika infections.

Researchers from the Sanger Institute, Umeå University, Sweden and the National Institutes of Health (NIH), US, analysed 8,506 immune cells to see which genes were active in each cell, and identify molecular markers for the unique cell types. They discovered twice as many types of immune cell than had been seen previously, and used the markers to find and quantify these cells in the mosquitoes.

One new cell type – a megacyte – appeared to switch on further immune responses to the *Plasmodium* parasite. This is the first time a specific mosquito cell type has been implicated in regulating malaria infection and could be vital in understanding how the mosquito immune system successfully fights the parasite.

They also followed how *Anopheles* immune cells reacted to infection with the *Plasmodium* parasite, and showed that specific types of immune cell – granulocytes – proliferated and matured to respond to it.

The atlas identifies cellular events that underpin mosquito immunity to malaria infection, and offers new opportunities for research that seeks to break the chain of transmission between mosquitoes and people. It will also be a valuable resource for research into other mosquito-borne diseases.



Reference
Raddi G, *et al*. Mosquito cellular immunity at single cell resolution. *Science* 2020; **369:** 1128–1132.

In 2017, researchers discovered that the rare Dantu blood variant, which is found regularly only in parts of East Africa, provides some degree of protection against severe malaria. To find out why, scientists at the Sanger Institute, the University of Cambridge and the KEMRI-Wellcome Trust Research Programme, Kenya, collected red blood cell samples from 42 healthy children in Kilifi, Kenya. The children either had either zero, one, or two copies of the Dantu gene.

The researchers then observed the ability of *P. falciparum* parasites to invade the cells in the laboratory, using multiple tools, including time-lapse video microscopy to identify the specific step at which invasion was impaired.

Analysis of the characteristics of the red blood cell samples indicated that the Dantu variant created cells with a higher surface tension – like a drum with a tighter skin. At a certain tension, malaria parasites were no longer able to enter the cell, halting their life cycle and preventing their ability to multiply in the blood.

Because the surface tension of all human red blood cells increases as they age, it may be possible to design drugs that imitate this natural process to prevent malaria infection or reduce its severity for anyone, regardless of whether they have the Dantu variant or not.

References
Kariuki SN, *et al*. Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature* 2020; **585:** 579–583.

Leffler EM, *et al*. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* 2017; **356:** 6343.

3

Genomic surveillance tracks comeback of neglected tropical disease

Genome analysis has revealed how the bacterium that causes yaws re-emerged in Papua New Guinea following an azithromycin-based mass drug administration campaign. The findings are helping to guide the World Health Organization's global yaws eradication strategy.

Yaws, caused by the bacterium *Treponema pallidum* subspecies *pertenue* (*TPP*), can cause chronic disfigurement and disability. Despite global efforts, yaws remains common in tropical areas in some of the world's poorest countries, affecting millions of children. The World Health Organization is carrying out campaigns to eradicate yaws using mass drug administration (MDA) of the antibiotic, azithromycin.

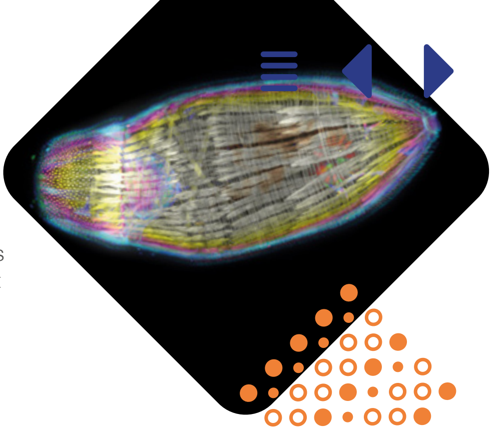
In 2013, 83 per cent of the population on Lihir Island, Papua New Guinea, were given azithromycin and the MDA campaign was initially successful. But after two years, cases of the disease starting increasing. Molecular testing showed the bacteria were of a single type, however it was unclear if the re-emergence had a single source. A small proportion of the bacteria were found to be resistant to azithromycin, the first time any such resistance had been seen.

Sanger Institute researchers worked with colleagues in the UK, Spain and Papua New Guinea, to sequence the genomes of bacteria from 20 swab samples taken during the follow up of the MDA campaign in Lihir.

Comparing the DNA sequences of the *TPP* bacteria, the team constructed phylogenetic 'family' trees to map their evolution. They found that the re-emergence was caused by at least three *TPP* lineages – most likely from latent infections in people, without symptoms, who didn't receive the treatment. They found that the azithromycin resistance had evolved once, as it was only present in one of the lineages.

The findings have important implications for disease control. The researchers recommend maximising MDA population coverage to reduce the number of people who are missed, plus intensive post-MDA surveillance to detect 'the last yaws cases' and enable swift detection and treatment of azithromycin resistance.

Reference
Beale MA, *et al*. Yaws re-emergence and bacterial drug resistance selection after mass administration of azithromycin: a genomic epidemiology investigation. *The Lancet Microbe* 2020; **1:** e263–e271.



4

How to build a parasitic worm

Sanger Institute researchers have developed the first cell atlas of an important life stage of *Schistosoma mansoni*, a parasitic worm that affects millions of people each year, primarily in sub-Saharan Africa. The atlas provides an instruction manual that will enable research into new vaccines and treatments.

S. mansoni has a complex life cycle: its larvae emerge from snails into rivers and lakes, where they can pass through the skin into humans. Once inside a person, the parasite develops into adult worms that live in the blood vessels and gut. These worms reproduce and release eggs that are passed out of the body to infect snails once more.

However, some eggs remain trapped in the body, causing schistosomiasis. This neglected tropical disease often leads to the inability to work, organ damage and death.

To better understand how the parasite adapts to live in humans, Sanger researchers characterised all of the cells in the parasite at the stage after it has entered the body. They used single-cell RNA sequencing to identify the active genes in each cell from two-day old larvae.

The team was able to find and validate genes that were specifically active in different parasite tissues, for example muscle, oesophageal, and gut. The researchers also identified 13 distinct cell types, including previously unknown types in the nervous and parenchymal systems.

By making individual fluorescent probes for the genes active in different cell types, scientists at the Morgridge Institute for Research, US, confirmed the position of the discovered cells within whole parasites under the microscope.

In other studies, the Sanger team used RNA sequencing on groups of cells from different developmental stages of the parasite to understand how gene activity changes over time during the life cycle.

The researchers hope that this deeper understanding of the parasite's developmental biology will help to expose vulnerabilities that could be targeted by new treatments.

References

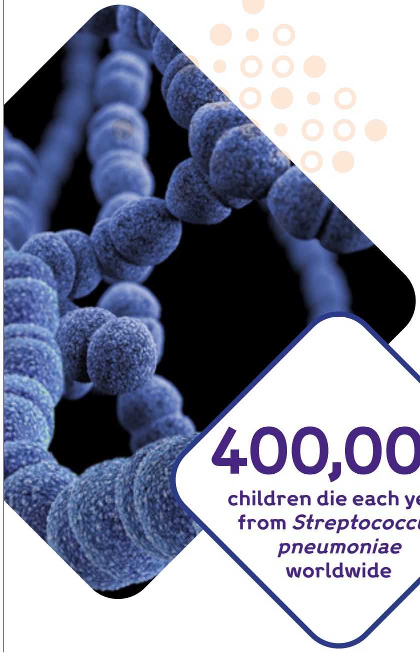
Diaz Soria CL, *et al*. Single-cell atlas of the first intra-mammalian developmental stage of the human parasite *Schistosoma mansoni*. *Nature Communications* 2020; **11:** 6411.

Wangwiwatsin A, *et al*. Transcriptome of the parasitic flatworm *Schistosoma mansoni* during intra-mammalian development. *PLoS Negl Trop Dis* 2020; **14:** e0007743.

5

Silently incubating a killer inside the nose

Genomic analysis of streptococci bacteria living in the noses and throats of new-borns reveals rapid microevolution that could give rise to antibiotic resistance and greater invasiveness.



400,000

children die each year from *Streptococcus pneumoniae* worldwide

Streptococcus pneumoniae is a bacterium that usually lives harmlessly in the nose and throat. However, in susceptible individuals, it can invade the body and cause fatal illnesses such as meningitis and sepsis. As a result, the bacterium kills 400,000 children a year.

Asymptomatic carriage of *S. pneumoniae* is common, and drives its evolution, transmission and pathogenesis, yet little is known about the genomic changes during these periods of natural colonisation. To shed light on these changes during natural infection, Sanger scientists worked with researchers at the Medical Research Council (MRC) Unit The Gambia.

The team collected samples sequentially over one year from 98 new-borns in The Gambia. The bacteria is present in up to 97 per cent of infants in the region.

Whole-genome sequencing and analysis of the *S. pneumoniae* samples revealed that the genetic mutation rates of the bacteria within an individual were forty-fold faster than those observed over longer timescales, and were driving high genetic diversity. The team discovered that multiple variants of different strains coexisted within the new-borns – these were either co-transmitted, acquired independently, or evolved over the period of carriage.

In addition, the evolution included genes which control antibiotic resistance, immune evasion and adhesion to human cells– suggesting that within-host microevolution is not only rapid, but also adaptive to the host.

The team hope that their work, carried out as part of the Global Pneumococcal Sequencing Project, will enable the research community to find genomic changes that could be targeted to reduce the bacteria's carriage and evolution.

Reference

Chaguza C, *et al*. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. *Nature Communications* 2020; **11:** 3442.

6

Superbug's stealthy spread revealed

For the first time, researchers from the Centre for Genomic Pathogen Surveillance and their collaborators have tracked how hospital superbugs use plasmids to spread antibiotic resistance. Their results demonstrate that it is vital to monitor superbugs' genomes and plasmids to effectively intervene in, and control, outbreaks.

The hospital superbug, *Klebsiella pneumoniae* can become resistant to last-line antibiotics called carbapenems, and is considered a critical threat in the World Health Organization's list of priority pathogens. *K. pneumoniae* evades antibiotics by acquiring 'carbapenemase' antibiotic resistance genes, which code for an enzyme that 'chews up' the drug.

In *K. pneumoniae*, carbapenemase genes are usually found on plasmids – small circular pieces of DNA, outside of the bacterial chromosome. Plasmids can jump between different strains and species of bacteria, enabling antibiotic resistance to rapidly spread.

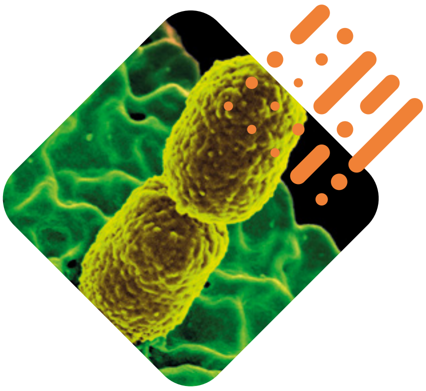
To read the genetic code of the bacteria's plasmids, researchers from the Centre for Genomic Pathogen Surveillance, based at the Sanger Institute and the University of Oxford, and collaborators, conducted long-read genome sequencing on 79 *K. pneumoniae* samples from hospital patients, taken from a Europe-wide survey.

Building on their earlier work to map the spread of *K. pneumoniae* in European hospitals in 2019, they analysed the full plasmid sequences alongside 1,717 previously short-read sequenced bacteria samples from 244 hospitals in 32 countries, to discover how antibiotic resistance genes are spreading.

The team uncovered three pathways for the spread of the genes:

- one plasmid jumping between multiple strains
- multiple plasmids spreading among multiple strains
- multiple plasmids spreading within one strain of *K. pneumoniae*.

These insights provide critical understanding for controlling outbreaks of antibiotic resistant infections. For example, if there is the possibility that a plasmid might jump into other bacterial strains or species, then they need to be monitored. The findings demonstrate that future monitoring approaches need to incorporate plasmids to allow interventions to be tailored – either to control the dominant plasmid, control the dominant strain, or in complicated situations, control both.



Reference
David S, *et al*. Integrated chromosomal and plasmid sequence analyses reveal diverse modes of carbapenemase gene spread among *Klebsiella pneumoniae. Proceedings of the National Academy of Sciences* 2020; **117:** 25043–25054.

7

Getting to the guts of a delicate balance

Sanger Institute researchers have created the first detailed atlas of immune and bacterial cells in the human colon. It offers new insights into how the body balances beneficial bacteria with preventing disease.

The gut microbiome is a complex ecosystem of millions of microbes that is essential for human health. Made up of mostly bacteria, it plays important roles in digestion, regulating the immune system and protecting against disease.

The gut also has a rich community of immune cells that prevents unwanted bacteria invading the body. Imbalances between the gut bacteria and immune cells can contribute to autoimmune diseases, such as ulcerative colitis and Crohn's disease, yet there is little detailed information on how they coexist in the gut. To shed light on this complex interaction the team studied three different parts of the healthy colon from organ donors, simultaneously analysing the immune cells and the bacterial microbiome from each area.

The researchers sequenced the active genes of 41,000 individual immune cells, to identify cell type and specific genes that were switched on in different cell populations in each location. They also identified the bacterial species present in the same colon regions.

The study revealed different immune niches, and changes in both the bacteria and the immune cells throughout the colon, with a broader range of bacteria further down the colon. The team also found that regulatory immune cells, which dampen down an immune response, moved from the lymph nodes to the colon.

These insights point to a way that the intestine tolerates or even welcomes the microbiome, and provides a foundation for future studies into digestive diseases, including irritable bowel disease and colorectal cancer.

Reference
James KR, *et al*. Distinct microbial and immune niches of the human colon. *Nature Immunology* 2020; **21:** 343–353.

Active genes of
41,000
individual immune cells sequenced

8

Building a global picture of Black Fever

Visceral leishmaniasis, known as Black Fever in Hindi, is a potentially fatal neglected tropical disease that affects approximately 90,000 people a year. Genomic sequencing and analysis has provided new understanding of the evolution of the parasites responsible and a valuable genetic resource for future surveillance.

Leishmania are a diverse group of single-celled parasites transmitted by sand flies that cause a range of clinical conditions. The most serious, visceral leishmaniasis, can be fatal. The disease is caused by *Leishmania donovani* and *L. infantum* species – together termed the *L. donovani* species complex. Both are widespread, with *L. donovani* concentrated in the Indian subcontinent and East Africa, *L. infantum* in the Mediterranean and Middle East, and both in China.

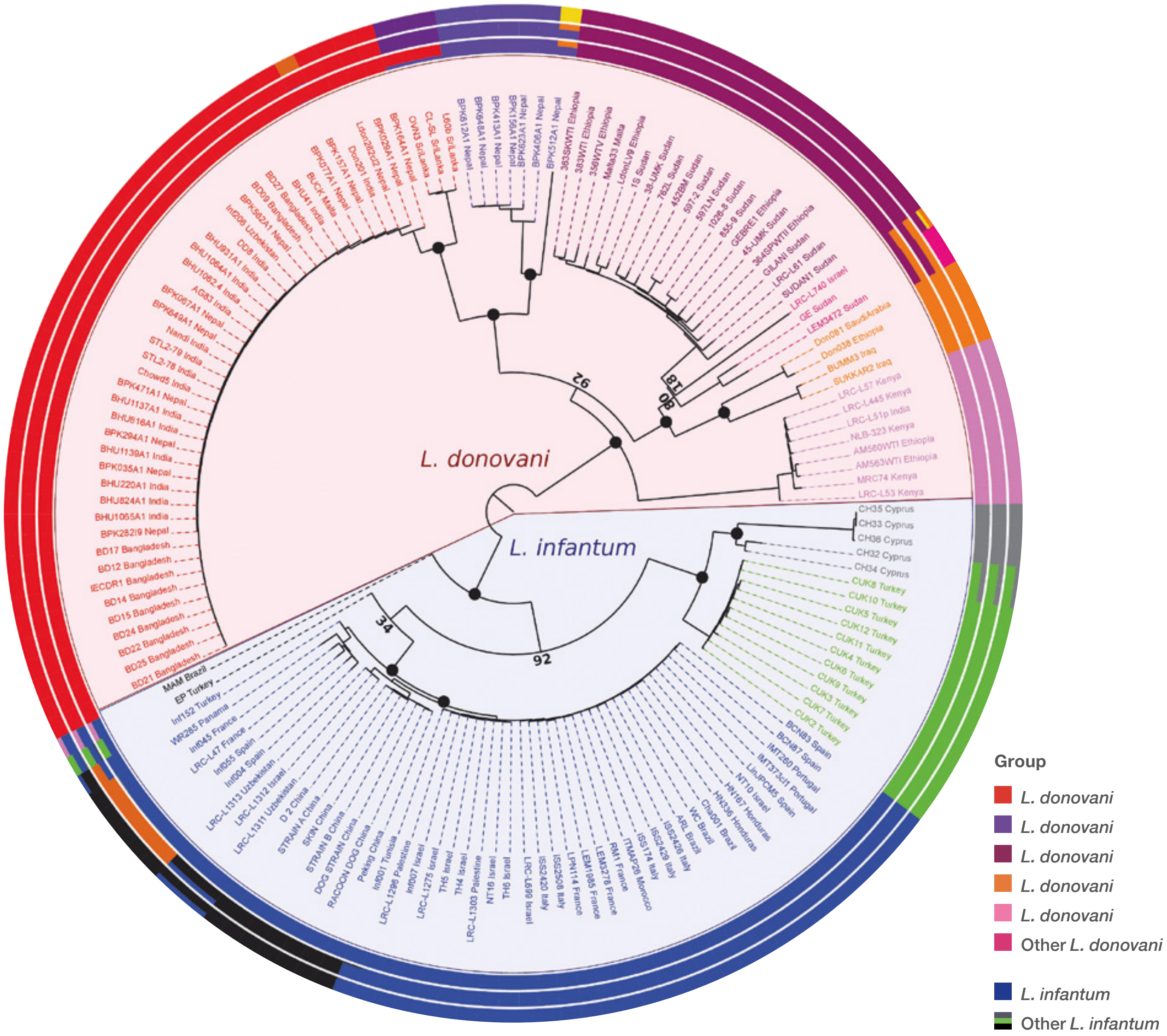
To understand more about its genetic diversity and evolution, Sanger researchers undertook whole-genome sequencing of isolates from across the globe.

The genome data revealed *L. infantum* mostly form a single group, with little genomic diversity. In contrast, *L. donovani* are in five diverse groups. The team characterised the unusual chromosomal makeup of the parasites, and identified extensive structural variation in *L. donovani* species, including in known and suspected drug resistance regions of the genome.

Groups of hybrid parasite were also identified. Hybrids are relatively rare, and it is not clear how the parasite mates or exchanges genetic information. But this ability means that drug-resistance, or more deadly strains, could evolve. To characterise such hybrid parasites in more detail, the Sanger team worked with researchers at the Addis Ababa University in Ethiopia to sequence the genomes of parasites from natural infections in the country.

Their results reveal a complex pattern of mating and inbreeding of the parasite, that has likely shaped its epidemiology, and brings a fuller understanding of the nature of genetic recombination in natural populations of *Leishmania*.

Reference
Franssen SU, *et al.* Global genome diversity of the *Leishmania donovani* complex. *eLife* 2020; **9:** e51243.



9

"

7PET cholera can cause massive epidemics. To control cholera epidemics efficiently, it is vital that we can distinguish and understand the differences between the local, endemic *V. cholerae* that coexist alongside 7PET during major outbreaks."

Matthew Dorman,
first author on the study from the Wellcome Sanger Institute

How genomics changed Argentina's health policy

Using whole genome sequencing, Sanger researchers and their collaborators have mapped the evolution of epidemic and endemic strains of cholera-causing bacteria in Argentina. The work influenced Argentine health policy, where the national alert surveillance system now uses whole-genome sequencing to distinguish between pandemic and non-pandemic bacteria.

Cholera, caused by strains of *Vibrio cholerae* bacteria, is endemic in a large number of countries world wide, and kills nearly 100,000 people a year. Since the 1800s, there have been seven cholera pandemics around the globe, causing millions of deaths. The current pandemic, which began in the 1960s, is caused by a single lineage of *V. cholerae*, called 7PET.

Sanger scientists sequenced the genomes of a unique set of historical *V. cholerae* samples, held at INEI-ANLIS 'Dr. Carlos G. Malbrán', the national reference laboratory of Argentina. Their analysis confirmed that the 1992 outbreak of cholera in Argentina was caused by one introduction of 7PET *V. cholerae* bacteria, originally introduced into Peru. The bacteria then evolved very little during the six years of the epidemic.

In contrast, the multiple endemic strains of *V. cholerae*, which were circulating at the same time, showed wide genetic diversity. Previous work by the team in 2017 had shown that while endemic strains can make people ill, they seem to lack the potential to spread quickly and cause an epidemic.

These findings demonstrate that genomic surveillance can help guide public health interventions, and public health authorities in Argentina have now changed their national alert system to distinguish between pandemic 7PET lineage and local *V. cholerae* lineages using whole-genome sequencing.

Reference
Dorman MJ, *et al.* Genomics of the Argentinian cholera epidemic elucidate the contrasting dynamics of epidemic and endemic *Vibrio cholerae. Nature Communications* 2020; **11:** 4918.

Tree of Life

60,000

UK species' genomes will be sequenced by the Darwin Tree of Life Project

In this section

1 Finding the code for aquatic cohabitation

2 Otter genome to help species' future

3 Genome of malaria relative determined

4 King scallop contains gene riches

5 Sequencing all life on Earth

1

Finding the code for aquatic cohabitation

The Aquatic Symbiosis Project will empower researchers to answer important questions about the ecology and evolution of species who thrive together.

Despite the complexity of symbiotic relationships, they have evolved independently countless times. But little is known about how symbiotic partners adapt to one another over time, how resilient these partnerships are and how they respond to disruption. To answer these questions, the Sanger Institute has partnered with the Gordon and Betty Moore Foundation, US, to deliver the Aquatic Symbiosis Genomics Project.

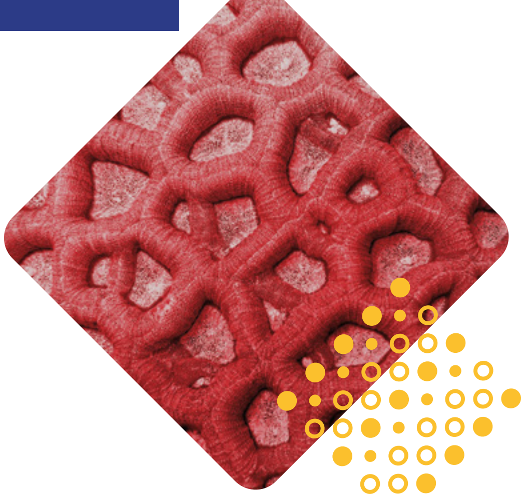
Based in the Tree of Life Programme at the Sanger Institute, the Project will call upon Sanger scientists' expertise in genome sequencing and sequence assembly to create gold-standard genome sequences for approximately 1,000 marine and freshwater species that form around 500 symbiotic partnerships. Once completed, they will be made freely available via the European Nucleotide Archive (ENA), GenBank and the DNA Databank of Japan.

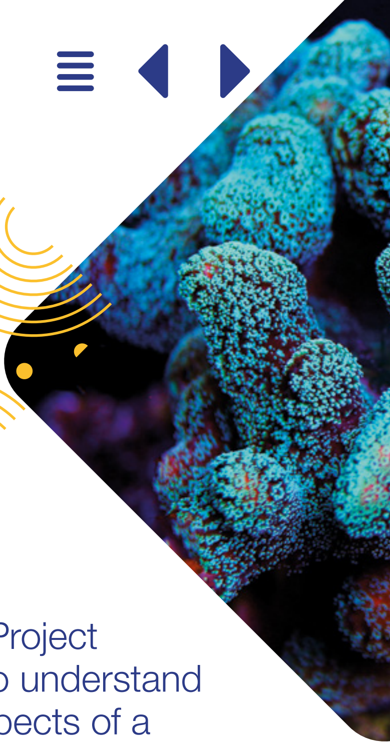
2

Otter genome to help species' future

The otter genome sequence will unlock a wealth of data stored in DNA archives to help understand the species' response to environmental changes.

Sanger scientists collaborated with researchers from the Cardiff University Otter Project to produce the first high-quality Eurasian otter genome using a combination of the latest sequencing technologies. The freely available, open-access reference genome will provide a strong foundation for researchers seeking to understand how pollutants and environmental changes affect aquatic species.





The results will allow scientists to investigate the evolutionary trajectory of symbiotic organisms, their place on the tree of life, and how they live together. The information will also be used to address some of the most urgent conservation challenges in our oceans, rivers and lakes, at a time when biodiversity is being lost at an alarming rate.

This includes phenomena such as coral bleaching, where higher ocean temperatures lead to corals losing their symbiotic algal partner after which they become sick or die. Mass bleaching events threaten the future of reef ecosystems worldwide. A greater understanding of the impact of this threat may allow researchers to develop strategies that could help reefs to survive.

> The Aquatic Symbiosis Genomics Project offers a unique opportunity for us to understand the origins, biology and future prospects of a huge range of symbioses. The visionary funding from the Gordon and Betty Moore Foundation, and the enthusiastic and open collaboration of our partners across the globe, promises to yield exciting new insights into this major part of Earth's biodiversity.
>
> **Professor Mark Blaxter,**
> Programme lead for the Tree of Life Programme

In the 1970s, accumulation of pollutants such as DDT and dieldrin in the environment caused a dramatic crash in British otter populations, which fell by 80–94 per cent. Contaminant levels have gradually declined since a ban on many of the worst pollutants, allowing otters to return to rivers from which they had been missing for decades.

But threats remain, with potentially harmful chemicals still widely used in pesticides, which can find their way into rivers. The implications of these chemicals for Eurasian otters are not known, but the species remains at risk and is listed as 'Near Threatened' on the The International Union for Conservation of Nature Red List of Threatened Species.

The new reference genome will help researchers at the Cardiff University Otter Project as they seek to understand the 1970s population crash, discover how otters interact with their environment and monitor for new threats. They will use the reference genome to analyse the wealth of data stored in the Project's DNA archives to better understand the biology of otters, how they have responded to changes in their environment, and inform ongoing conservation efforts.

This research will also provide insights into what is happening to other members of the otter's ecosystem who experience the same changes in environment and pollution; such as fish, birds, insects and bacteria.

Reference
Mead D, *et al.* The genome sequence of the Eurasian river otter, *Lutra lutra* Linnaeus 1758 [version 1; peer review: 2 approved]. *Wellcome Open Research* 2020, **5:** 33.

Eurasian otters are on The International Union for Conservation of Nature Red List of Threatened Species

3

Genome of malaria relative determined

Sanger scientists have generated a genome sequence for *Hepatocystis* – the first time this data has ever been produced and analysed.

Hepatocystis parasites are single-celled organisms, closely related to the *Plasmodium* species which cause malaria. They infect monkeys, bats and squirrels. Unlike *Plasmodium*, they don't cause disease in their mammalian hosts. They are transmitted from one host to the next by biting midges, not by mosquitoes.

The Sanger team worked with researchers at Duke University, US and the Crick Institute in London to determine the genome sequence, using DNA sequencing reads from the blood of a naturally infected Ugandan red colobus monkey.

Using comparative genomics, the researchers traced the evolutionary history of *Hepatocystis*, and confirmed that they are descended from *Plasmodium*. The team analysed areas of the genome that may be involved in the replication stages of the parasite's life cycle – as it is the replication of *Plasmodium* in the red blood cells of its host that causes disease.

By analysing the transcriptome – regions of the genome that are active – they strengthened the evidence that *Hepatocystis* does not replicate in the blood. They also discovered that the genes involved in interaction with red blood cells in *Plasmodium* have been lost in *Hepatocystis* as they evolved.

The team also found that genes which are active when the parasite is in the insect are rapidly evolving. This highlighted the equivalent genes in *Plasmodium* that might be critical for understanding interaction between malaria parasites and mosquitoes.

The researchers hope that understanding the difference between the two sets of species will enable them to understand more about how malaria causes disease.

Reference
Aunin E, *et al*. Genomic and transcriptomic evidence for descent from *Plasmodium* and loss of blood schizogony in *Hepatocystis* parasites from naturally infected red colobus monkeys. *PLoS Pathogens* 2020; **16:** e1008717.

Read the full story at:
www.sangerinstitute.blog/2020/08/03/genomes-within-genomes/

4

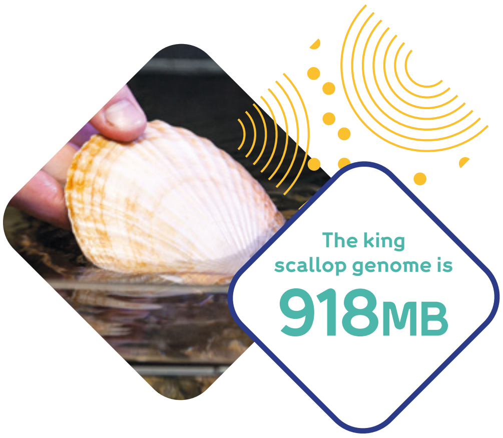
King scallop contains gene riches

Working with colleagues at the Natural History Museum, Sanger scientists sequenced and assembled the genome of *Pecten maximus*, the king scallop. The genome and its annotated gene set provide a high-quality platform for studies on shell biomineralisation, pigmentation, vision, and resistance to algal toxins.

The king scallop is a bivalve mollusc, found in shallow marine waters of Europe and West Africa. It is an important species both economically – worth millions to the fishing industry – and ecologically, where it is a key part of food webs, and cycles nutrients during filter feeding.

The Sanger Institute R&D and DNA Pipeline teams used a range of state-of-the-art technologies to sequence and assemble the scallop genome. The finished sequence, 918MB, is the most contiguous of all published bivalve genomes to date. To identify the location of genes, the team also sequenced the active areas of the genome from the same scallop sample. Using software to combine this data with previous studies of genome activity, and then to compare the results with similar species, they predict it has 67,741 genes – relatively high when compared to related species.

The researchers also assessed the ability of the king scallop to resist potent neurotoxins. Such toxins, for example domoic acid, are produced by phytoplankton, and can accumulate in filter feeders with seemingly few ill-effects. However, if eaten by mammals higher up the food chain, the toxins can cause severe illness and sometimes death.



The king scallop genome is **918MB**

The team discovered that specific genetic mutations within the sodium channel gene, *Neuron Navigator 1* (*Nav1*) could be the key to the scallop's resistance to domoic acid's affects. The finding could help with research into Alzheimer's and Parkinson's Disease, as domoic acid mimics the effects of the neurotransmitter glutamate, which has been linked to the conditions.

Freely available to researchers world wide, the genome sequence will also aid research into the evolution of colour and vision, including into the unique eye structure.

Reference
Kenny NJ, *et al*. The gene-rich genome of the scallop *Pecten maximus*. *GigaScience* (2020) **9:** giaa037.

5

> I think one of our biggest achievements has to be that we're now properly up and running, despite the disruption of coronavirus. The support from our colleagues in sequencing operations has been amazing."

Caroline Howard,
Scientific Manager for the Tree of Life Programme

Plenary talks attended by
1,000+
people

Each workshop had
500–600
attendees

Sequencing all life on Earth

Between 5–9 October 2020, researchers around the world gathered online at the virtual Biodiversity Genomics conference, organised and supported by the Tree of Life Programme at the Sanger Institute.

Using a mix of workshops, plenaries, virtual poster sessions and chat rooms, the meeting celebrated achievements in genome sequencing the eukaryotic tree of life, explored current challenges and their likely solutions, and looked forward to the application of genomics across the globe.

Powerful advances in genome sequencing technology, informatics, automation, and artificial intelligence mean it is now possible to sequence the genomes of all 1.5 million known eukaryotes (animals, plants, protozoa and fungi). These techniques will also allow researchers to discover the remaining estimated 10–15 million species currently hidden from science.

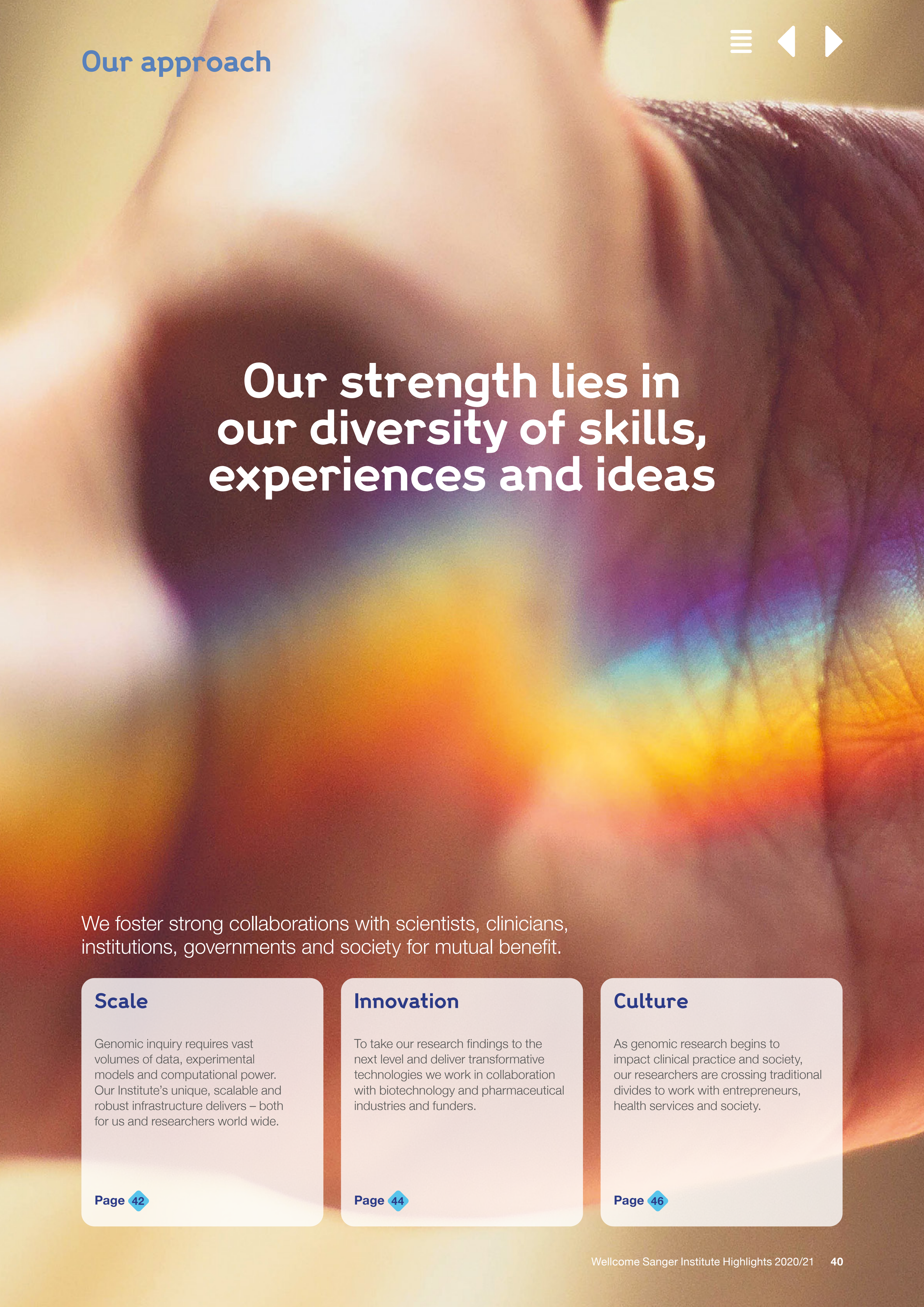
Less than 1 per cent of known eukaryotes have been sequenced so far. Sequencing all eukaryotic life in the next 10 years is the goal of the Earth BioGenome Project – and the conference gave scientists an opportunity share their latest progress.

Specific projects with different geographical or taxonomical focus also featured at the meeting. Sanger Institute speakers presented work on the Darwin Tree of Life Project which aims to sequence the genomes of 60,000 UK eukaryotes.

Sessions covering the latest technologies in genome sequencing, assembly and annotation were included and the application of genomics to conservation, agriculture and economics were discussed. Plenary talks attracted 1,000 live attendees, and most science workshop sessions had 500-600 live attendees. A total of over 5,200 individuals logged in to the conference.

View Conference views on YouTube:
www.youtube.com/channel/UCA4F7J0Z472T0c5IqJAxe2w

5,200+
individuals logged in to the conference



Our strength lies in our diversity of skills, experiences and ideas

We foster strong collaborations with scientists, clinicians, institutions, governments and society for mutual benefit.

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 world wide.

Page 42

Innovation

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

Page 44

Culture

As genomic research begins to impact clinical practice and society, our researchers are crossing traditional divides to work with entrepreneurs, health services and society.

Page 46

Influence

By leading global initiatives and facilitating cross-cutting partnerships we seek to lay the foundations for a strong and vital future of genomic research, data sharing and clinical application.

Page 48

Connections

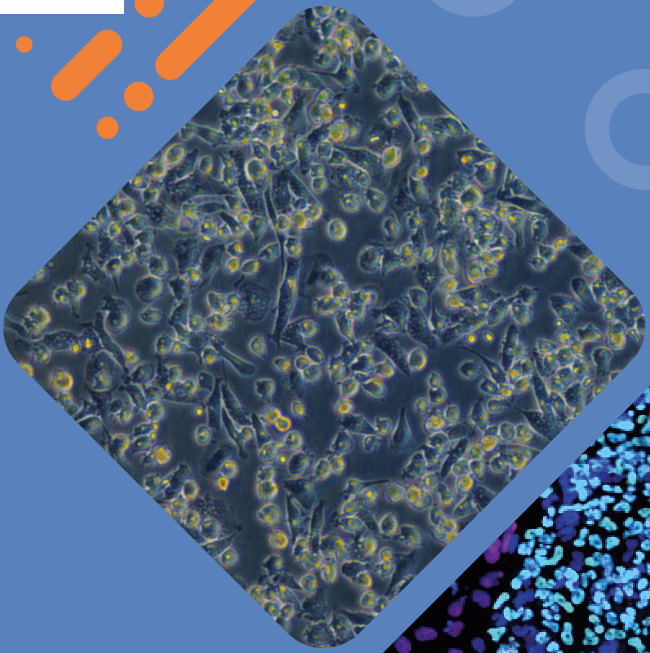
We use the power of the internet and collaboration tools to build genomic research capacity worldwide and facilitate the next wave of discovery.

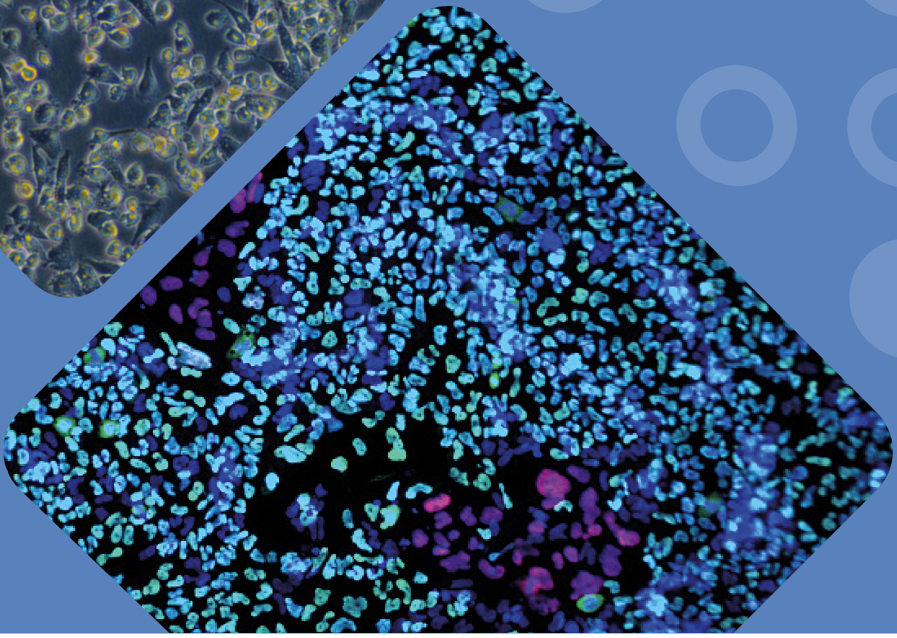
Page 50

The Sanger Institute is committed to empowering and developing its staff at all levels.

Read more at:
www.sanger.ac.uk/about/careers/life-at-sanger/

Scale





In this section

1 Finding single cells in oceans of data

2 Global partnership to fight a global killer

1

Finding single cells in oceans of data

Sanger scientists have created a simple search engine that enables anyone to find relevant single cell data in seconds, on a normal computer.

The Human Cell Atlas, along with many cancer, parasite and human genetics projects, are sequencing and mapping millions of individual cells from a wide range of species. This wealth of data offers great opportunities to unlock new understanding in development, health, disease and evolution – but only if scientists can find the specific cells they are looking for.

The volume and complexity of the data stored in open-access cell atlases has meant that identifying relevant cells has been the preserve of bioinformaticians, supported by significant computer processing power. To democratise this data for the global research community, Sanger bioinformaticians have developed scfind, a freely available software tool that analyses multiple cell atlases in just a few seconds, using natural language input, on a standard computer.

To enable fast and efficient searches without specialist hardware scfind uses a two-step strategy to compress data approximately 100-fold. Efficient decompression means that queries that used to take days now take seconds.

Scfind can also be used to identify new genetic markers that are associated with, or define, a cell type. The scientists found that scfind is a more accurate and precise method to do this, compared with manually curated databases or other computational methods available.

Finally, to make scfind more user friendly, it incorporates techniques from natural language processing to allow for arbitrary queries, opening access to the data still further.

> To ensure that large single-cell datasets can be accessed by a wide range of users, we developed a tool that can function like a search engine – allowing users to input any query and find relevant cell types.”

Dr Martin Hemberg,
Former Group Leader at the Wellcome Sanger Institute

Reference
Lee JTH, *et al.* Fast searches of large collections of single-cell data using scfind. *Nature Methods* 2021. https://doi.org/10.1038/s41592-021-01076-9.

scfind is freely available at:
https://scfind.sanger.ac.uk

2

Global partnership to fight a global killer

The Malaria Genetic Epidemiology Network (MalariaGEN) gathered 7,113 malaria parasite samples from 28 endemic countries, which were genome sequenced. The result is the world's largest open genomic data resource on the parasite's evolution and drug resistance.

Founded in 2005, MalariaGEN is a scientific network that connects researchers and clinicians in malaria-endemic countries with cutting-edge DNA sequencing technologies and tools for genomic analysis. It now has partners in 39 countries, each leading their own studies into different aspects of malaria biology and epidemiology, with the common goal of finding ways to improve malaria control and elimination strategies.

As part of the network's goal to build high-quality data resources for malaria research and surveillance, 49 partner studies at 73 locations in Africa, Asia, South America and Oceania, contributed 7,113 samples of *Plasmodium falciparum* for genome sequencing. At the Sanger Institute, each sample was analysed for over 3 million genetic variants and the data were carefully curated before being returned to partners for use in their own research.

The data are also available in an open-access resource for the wider scientific community, which is the world's largest resource of genomic data on malaria parasite evolution and drug resistance. It provides benchmark data on parasite genome variation that is needed in the search for new drugs and vaccines, and in the development of surveillance tools for malaria control and elimination.

MalariaGEN was created to provide equitable access and credit for the global malaria research community. One of its core principles is to provide clear attribution and recognition of all the groups that have contributed. In this dataset, each sample is listed against the partner study that it belongs to, with a description of the scientific aims of the study and the local investigators that led the work.

Reference

MalariaGEN and multiple co-authors. (2021) An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples [version 1; peer review: awaiting peer review]. *Wellcome Open Research.* DOI: 10.12688/wellcomeopenres.16168.1.

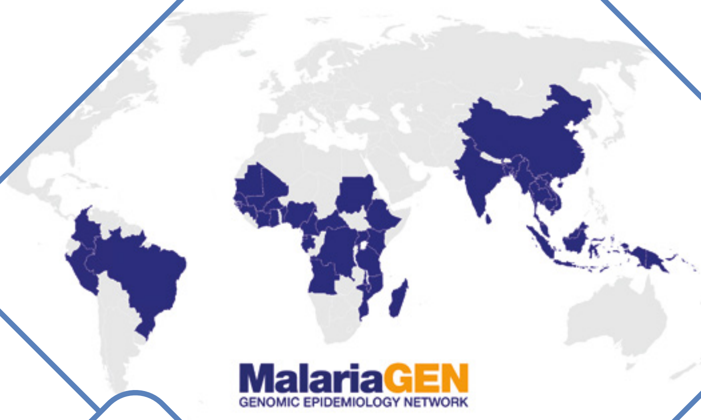
> We have created a data resource that is 'analysis ready' for anyone to use, including those without specialist genetics training. Like the Human Genome Project was a resource for the analyses of human genome sequence data, we hope this will be one of the main resources for malaria research."

Dr Richard Pearson,
Co-author from the Wellcome Sanger Institute

> Over time, this openly available resource will facilitate research into the malaria parasite's evolutionary processes, which will ultimately inform effective and sustainable malaria control and elimination strategies."

Professor Abdoulaye Djimde,
CAMES Professor of Parasitology and Mycology
and Honorary Faculty at the Wellcome Sanger Institute

73
locations in Africa, Asia, South America and Oceania



49
partner studies

Innovation

In this section

1 Supporting the genomics and biodata founders of the future (virtually)

2 Innovative Sanger spin-out proves its worth

3 Targeting each cancer's unique weaknesses

1

Supporting the genomics and biodata founders of the future (virtually)

The Entrepreneurship & Innovation team has launched a brand new initiative to inspire and to help develop and nurture the next generation of entrepreneur scientists.

The Sanger Institute is an ideas factory that generates bold solutions to benefit human health and society through innovation in genomics and biodata. Many of its discoveries are driving innovative academic genomic research, companies and healthcare programmes worldwide. World-class science requires commercial support to fully realise its potential to improve medical care or diagnosis and needs to be translated from the laboratory into commercial offerings or spinout ventures.

To enable Sanger researchers and technicians to embrace the value of entrepreneurship in research, the Entrepreneurship & Innovation team has developed a virtual Startup School composed of 10 two-hour sessions covering the key strategic, technical, and commercial steps to create a successful genomic/biodata offering.

This virtual programme is the first pre-acceleration learning initiative that the Wellcome Genome Campus has hosted and brings together individuals from across different institutes, research programmes and disciplines. The course draws on the Campus' exceptional pool of expert mentors, including thought leaders from companies based at the BioData Innovation Centre.

The first cohort of 24 participants from the Sanger Institute and EMBL's European Bioinformatics Institute joined the online mini boot camp in November 2020. Through a mixture of teamwork challenges, coaching sessions and talks from industry insiders, the scientists have been taken out of their comfort zones and given a taste of life as an entrepreneur.

The Startup School team hopes that this course will furnish its participants with tools, insights and healthcare sector connections to successfully take their ideas forward.

Find out more at:
https://bit.ly/WellcomeGenomeStartupSchool

> It's a mini boot camp which aims to give you the tools needed to build your own venture with meaningful insights to help consolidate your plans."
>
> **Dr Julia Wilson,**
> Associate Director, Wellcome Sanger Institute

STARTUP SCHOOL

2

Innovative Sanger spin-out proves its worth

Kymab, a Sanger Institute company, has been acquired by Sanofi to deliver antibody treatments to combat immune diseases and immunology-based therapeutics for cancers.

In 2010, the Sanger Institute created a spin-out enterprise to produce antibody and vaccine treatments based on genome engineering developed by the Institute's former Director, Professor Allan Bradley. Using initial funding from the Wellcome Trust Investment Division, the company recreated the entire diversity of the human immune system's B lymphocyte component in a humanised mouse.

Over the past 10 years, Kymab's researchers have worked with Heptares, Novo Nordisk and the MD Anderson Cancer Center to explore how immunological processes affect health and disease in the fields of autoimmunity, cancer, blood-related and infectious diseases. The team also collaborated with the Bill and Melinda Gates Foundation to test many of the vaccines the Foundation is developing for low- and middle-income countries.

In January 2021, Sanofi acquired the company based on the positive results of a Phase 2a randomised, double-blinded, placebo-controlled study of one of its fully human monoclonal antibody treatments – KY1005. After 12 weeks, KY1005 produced clinically meaningful improvements in 88 adults with moderate to severe atopic dermatitis whose disease could not be adequately controlled with topical corticosteroids, while also being tolerable and safe.

The potential of KY1005 to benefit health is not limited to atopic dermatitis. It has a novel biological action; binding to OX40-Ligand, a key regulator of T-cell activation. For this reason, it could play a role in treating a wide variety of immune-mediated diseases and inflammatory disorders.

The $1.1 billion deal will ensure that Kymab's pipeline of novel therapies will continue to be developed and, hopefully, delivered to patients.

3

Targeting each cancer's unique weaknesses

A new spin-out company from the Sanger Institute aims to disrupt the traditional paradigm of cancer treatment by generating precision healthcare tailored to the genetic makeup of a patient's tumour cells, not based on its location in the body.

Until recently, most diagnoses and treatments for cancer have been made on the basis of the tissue in which the tumour was found. However, the results of this approach have been mixed.

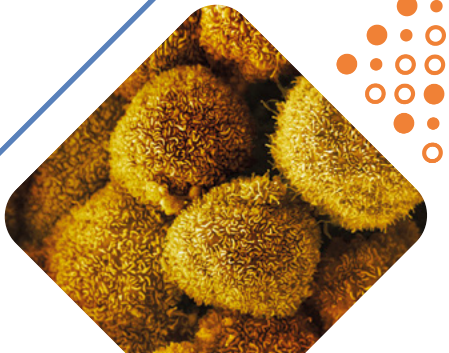
Over the past 20 years, genomic research has revealed that a tumour's weaknesses are associated with the genetic changes its cells have accumulated over time. To translate this knowledge into new precisely targeted treatments, MOSAIC Therapeutics has been created.

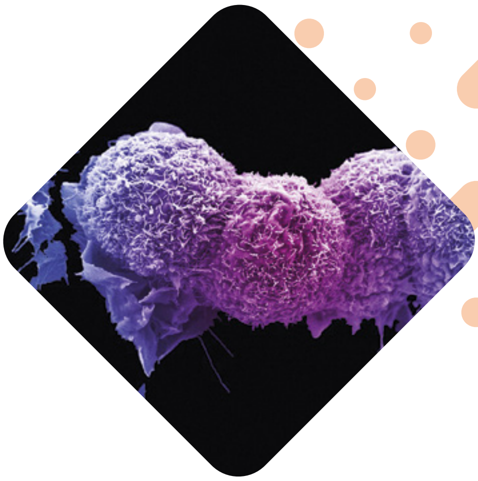
The company's unique approach is rooted in the science of the cancer itself. Changes in the DNA of a cancer cell enable it to grow unchecked, but at the cost of creating weaknesses that can be targeted. MOSAIC Therapeutics will apply advanced computational methods and next-generation cancer models to find new synthetic-lethal cancer targets for drug development.

Born out of the Cancer, Ageing and Somatic Mutation Programme's pioneering science and anchored around a proprietary database, MOSAIC's main mission is to tackle cancers with substantial unmet need.

The company has active programmes across a variety of cancer indications and disease biology fields, and it will look to strategically partner with pharma companies to drive forward its research.

MOSAIC Therapeutics uses organoid tumour systems, genetic tools and massive datasets, combined with the power of genomics and AI, to identify and exploit cancer's vulnerabilities. Using these platforms, the team will be able to select the best targets for drug development coupled with an understanding of which individual patients are most likely to respond.





Culture

In this section

1 Diversity is the Sanger Institute's strength
2 Supporting mental health during the pandemic

1 Diversity is the Sanger Institute's strength

In 2020, the Institute received the Athena SWAN Silver Award, joined Stonewall Diversity Champions, and supported homeworking staff during the pandemic to enable all its people to thrive in challenging circumstances.

The Sanger Institute's science is founded on the diversity of skills, perspectives and experiences of its researchers and support staff. It is committed to developing an organisational culture of mutual respect and support that provides inclusive, nurturing experiences for its students, scientists, technicians and administrators.

In April 2020, the Institute's work to address gender inequality was rewarded with the Athena SWAN Silver Award. The award recognised the support the Institute has put in place – from introducing best-practice employment processes to providing grants and initiatives for parents, carers and those returning to science from a career break.

These measures include the Stop-the-Clock initiative that extends fixed-term contracts for Postdoctoral Fellows and PhD students when they take maternity or shared parental leave. Other schemes are helping parents to thrive with family friendly meeting times, flexible working and a subsidised holiday club. In addition, career breaks are now taken into account during recruitment and promotion.

The COVID-19 pandemic has provided new challenges that the Institute has overcome to support its staff. Over the past year, the Campus' LGBT+ network has developed online support mechanisms to provide networking and peer-to-peer mentoring opportunities, targeted workshops and talks. To drive continued development, the Institute has joined the Stonewall Diversity Champions programme.

The pandemic also brought concern for Postdoctoral Fellows whose research was delayed. To provide certainty and support the Sanger Institute has extended their contracts – both for those funded directly by the Institute and those funded by third parties.

> Our diverse, interdisciplinary community encompasses a broad range of global perspectives, expertise and experiences. We value all of our members and the skills they bring, and are committed to creating a working environment where difference is valued and welcomed, and everyone can reach their full potential."
>
> **Saher Ahmed,**
> Head of Equality, Diversity and Inclusion at the Sanger Institute

2

Supporting mental health during the pandemic

Working under COVID-19 restrictions has presented mental health challenges for Sanger Institute staff juggling work with family responsibilities or facing increased isolation. To provide valuable support, the Health and Safety and Human Resources teams have employed a range of creative solutions – from wellness apps and online training to virtual family variety shows and gardening sessions.

When the UK Government introduced COVID-19 restrictions in March 2020 and the Wellcome Genome Campus closed to all but essential science, Sanger staff faced a new work environment. Whether it was working at a kitchen table in a busy household of homeschoolers, supporting the COVID-19 sequencing effort on Campus while adhering to social distancing rules, or working alone at home many hundreds of miles from family, everyone had new mental health challenges to face.

To support staff both practically and emotionally, the Health and Safety and Human Resources teams rapidly adapted their Wellness@Work, mental health first aid, solo worker, and parent and carer support approaches. To ensure that all staff are nurtured, the Institute has sent out monthly surveys to identify unmet need and assess the effectiveness of its initiatives.

Professional career and skills development training moved to online delivery and the Institute partnered with LinkedIN learning to provide remote access to all areas of personal development. Parents could join regular online meetings for mutual support and expert advice. While isolated workers have been able to encourage each other through the solo-living network.

Personal mental health needs were addressed through virtual mental health drop-in sessions with a counsellor, and every staff member has been able to download the Headspace mobile app for free.

There have also been opportunities to laugh and enjoy time together: the Institute organised a virtual family variety show featuring a comedy magician and children's storyteller, and an online gardening workshop for those who had discovered the joy of having a garden.

Influence

In this section

1 Showcasing Sanger science – in person and virtually

2 Sanger's scientific impact recognised

3 Working together for the world's common (diseases) good

1

Showcasing Sanger science – in person and virtually

Sanger Institute researchers shared their science at the 2020 and 2021 AAAS meetings (American Association for the Advancement of Science) – the world's largest multidisciplinary scientific gathering.

In 2020, Sanger scientists presented two short talks and took part in panel sessions. The first talk explored what can be learnt from species that are resistant to cancer and how this can be applied to help prevent cancer in people. The second revealed how a patient's response to sepsis treatment is linked to their gene activity – potentially providing a new way to deliver targeted, effective treatments.

Sanger researchers also participated in two panel discussions. The first detailed how the Human Cell Atlas initiative, co-chaired by the Sanger Institute, coordinates international efforts to define the activity of the ~20,000 genes in each of the bodies' 37 trillion cells. The results, freely available in open-access databases, will help power future research into human development, biology, health, and disease.

The second panel discussed how the Earth BioGenome Project is sequencing the genomes of all plants, animals, fungi and protists on Earth. The Sanger Institute is contributing to this work by coordinating the Darwin Tree of Life project to sequence, and openly release, the genome data of all such species in the UK.

In 2021, the meeting took place online, with the Institute contributing to two virtual sessions. The first centred on the Cancer Dependency Map collaboration, which uses large-scale CRISPR gene editing to uncover critical cancer survival genes for future treatment development.

The second session saw representatives of the parasites and microbes programme discuss the use of genomics in tracking and defeating infectious diseases. Topics included: using genomics to trace the spread of cholera and break chains of transmission; genomic surveillance of malaria across Africa; and the UK's national SARS-CoV-2 genomic surveillance efforts to identify new variants that were involved in the COVID-19 pandemic.

Read more on the Sanger Institute Blog:
https://sangerinstitute.blog/

Cancer across the animal kingdom and Sepsis and precision medicine – is it all in your genes?

2

Sanger's scientific impact recognised

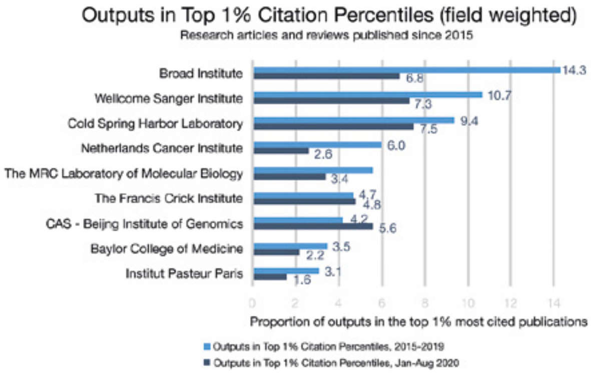
The vital contribution of the Sanger Institute's foundational science to global genomics and bioinformatics research has been revealed in Clarivate's 2020 Highly Cited Researchers report.

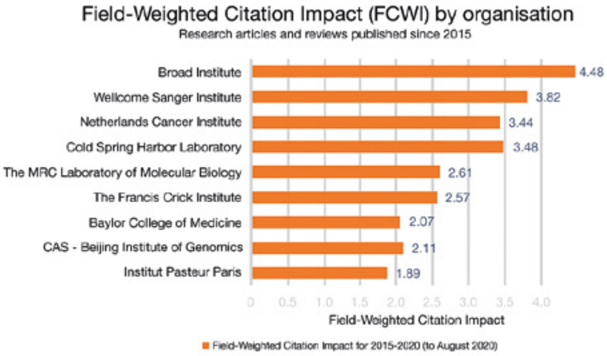
The annual report identified 6,389 researchers from more than 60 countries whose research papers are the most highly cited in their respective fields over the past decade. The researchers were selected from 21 science, social science and cross-field categories based on the rate at which others cited their work between 2009–2019.

Overall, the UK ranks third in the world with 514 of the most highly cited researchers. The Sanger Institute is affiliated with 25 of these scientists, making it the only UK research institute to feature in the short list. The other UK organisations listed are the University of Oxford (52 researchers), the University of Cambridge (46 researchers), University College London (41 researchers), King's College London (27 researchers) and Imperial College London (26 researchers).

The report confirms the Institute's own impact assessment work – see the charts below – which it is using to understand the contribution its science makes to driving global research and to guide future research endeavours.

On average, 10.7 per cent of the Institute's articles and views published since 2015 have been among the top 1 per cent of the world's most cited publications, even when variations in citation rates across different research are taken into account. In addition, the organisation's publications have been cited 3.82 times more when compared with the global average for similar publications.





3

Working together for the world's common (diseases) good

Sanger Institute researchers are part of the International Common Disease Alliance (ICDA) that has proposed a range of global actions and international collaborations to accelerate progress from genetic maps to biological mechanisms to medical treatments.

Over the past two decades, the worldwide research community has made significant progress in using genetics to systematically understand the biological basis of common diseases. So far, more than 70,000 associations between specific genetic regions and human diseases have been discovered, and many new disease genes and mechanisms have been identified. However, important barriers remain before these discoveries can be translated into biological insights and new therapies.

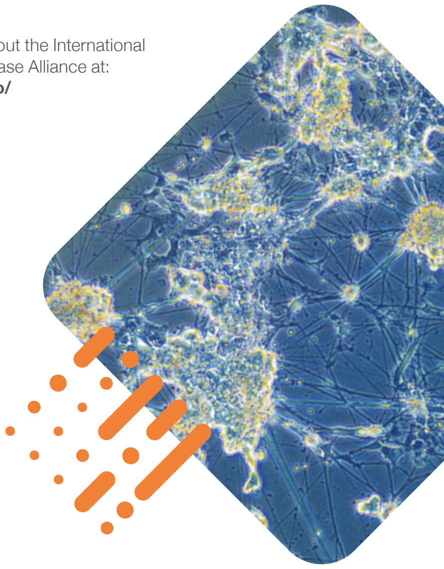
Launched in September 2019, the ICDA aims to improve the prevention, diagnosis, and treatment of common diseases, including diabetes, schizophrenia, Alzheimer's and heart disease. The alliance seeks to encourage and support the global scientific community to focus and coordinate their research for maximum effect.

A number of Sanger faculty leaders from the Human Genetics and Cellular Genetics Research Programmes and the Open Targets and HDR-UK associate research programmes, are founding members of the ICDA and contribute to its work. Now, the ICDA brings together more than 150 researchers from academia, medicine, pharmaceutical companies, technology companies, and biomedical funders around the world.

To enable researchers to fully understand the relationship between genetic variation and disease processes in common conditions, the ICDA has identified a range of difficulties that need to be addressed. These include legal, procedural and scientific barriers to connecting genetic variants to the target genes, cell types, and biological pathways in which they act.

The group has drawn up a white paper of specific actions to overcome these difficulties, including international collaborations, data resources, critical infrastructure, policy needs, and equity considerations. It aims to help jump-start priority projects, policies and activities over the next 5–10 years to help deliver biological insights and clinical benefits from genetic research.

Read more about the International Common Disease Alliance at:
www.icda.bio/



Connections

In this section

1 Supporting fledgling scientists in low- middle-income countries

2 International Fellows generate global genomic excellence

3 Seminar series shares Sanger science with the world

1

Supporting fledgling scientists in low- and middle-income countries

A Sanger Institute supported podcast – Your Digital Mentor – aims to help early career genomics researchers grow their science in low- and middle-income countries.

Genomic discoveries and bioinformatics resources are revolutionising healthcare across Europe, the US and the Far East. However, for the benefits to be reaped across the globe, there is a pressing need for greater expertise in low- and middle-income countries. Yet a lack of available mentors in these regions may be holding talented researchers back.

To provide these early career researchers with the knowledge and experience they need to succeed, the Sanger Institute, Connecting Science's Advanced Courses and Scientific Courses, and Social Entrepreneurship to Spur Health collaborated to support a dedicated mentorship podcast. The result is the Your Digital Mentor series: available free on SoundCloud, YouTube, Apple Podcasts and Spotify.

The series is made up of 12 episodes, each focused on a specific topic relevant to building a fruitful scientific career. Topics include finding a mentor, building supportive relationships, networking tips and decolonising science.

To provide truly valuable advice, the podcast features honest discussions and real stories from expert guests across the world about their experiences as mentors or mentees. It is hoped that these conversations between researchers, public health professionals and clinicians will help fledgling scientists and generate greater awareness of the value of mentorship.

2

International Fellows generate global genomic excellence

The Sanger Institute's International Fellows programme nurtures early career scientists by supporting collaborations in developing countries with access to Sanger's infrastructure and resources; raising Fellows' visibility and standing around the world.

The Sanger Institute is a global hub of genomic science, playing a key role in facilitating continental and intercontinental networks of genomic discovery and research capacity. Building on the initial success of the Sanger Institute's International Fellows programme, the Institute is now supporting six Fellows based in Africa, Latin America, and Southeast Asia.

Founded in 2011, the programme first supported Professor Sam Kariuki in Kenya and Dr Abdoulaye Djimdé in Mali. They leveraged their access to Sanger resources to establish and build genomic research capacity in non-typhoidal *Salmonella* and antimalarial drug resistance respectively. Both are continuing their research collaborations with the Institute as Honorary Fellows.

Carla Daniela Robles-Espinoza
Pablo Tsukayama
Gregorio Iraola
Alfred Amamabua -Ngwa
Annette Nakimuli
Claire Chewapreecha

The Institute has now widened its focus to provide six four-year Fellowships for early career stage scientists in low- or middle-income countries to address health issues in their region. In particular, the programme aims to help researchers in junior faculty positions to develop collaborative projects and programmes at the national, continental and global level.

Currently, International Fellows are embedded in four Sanger Research Programmes: Cancer, Ageing and Somatic Mutation, Cellular Genetics, Human Genetics, and Parasites and Microbes.

- **Alfred Amamabua-Ngwa** – The Gambia – Using CRISPR to explore drug resistance in malaria and genome sequencing to monitor parasite and mosquito evolution.
- **Claire Chewapreecha** – Thailand – Studying the genetic basis of melioidosis, caused by the bacterium *Burkholderia pseudomallei*.
- **Gregorio Iraola** – Uruguay – Using sequencing, bioinformatics and culturing to investigate the evolutionary patterns of disease-causing bacteria in humans and animals.
- **Annette Nakimuli** – Uganda – Identifying genetic risk factors for pre-eclampsia and associated placental dysfunction disorders, and placental response to infection.
- **Carla Daniela Robles-Espinoza** – Mexico – identifying genomic drivers and potential therapeutic targets for acral lentiginous melanoma, and exploring how environmental factors affect the development of liver and lung cancer.
- **Pablo Tsukayama** – Peru – monitoring the transmission of antibiotic resistance between humans, animals and water systems near places such as hospitals and poultry markets.

3

Seminar series shares Sanger science with the world

The global COVID-19 pandemic has thrown up a number of barriers to scientific discussion and knowledge sharing. To overcome them, the Sanger Institute has been providing a freely accessible series of online seminars.

Scientific progress and innovation is founded on the free exchange of ideas and discoveries, with scientific conferences providing a vital melting pot for such research discussion. However, the global pandemic has removed researchers' ability to travel and mingle both nationally and internationally, drastically curtailing this vital flow of knowledge. To overcome these physical barriers and provide a globally accessible space for scientific discourse, Sanger Faculty are delivering their seminars virtually.

The monthly series of freely available and open virtual seminars started in July 2020, showcasing how Sanger scientists are tackling some of the greatest challenges in human health and disease. So far the topics covered have ranged from using genomic approaches to map all cell types in the human body to understanding how cancer develops, and from tracking the evolution and spread of global diseases to sequencing the genomes of all species on the tree of life.

The talks, along with their online Q&A sessions, are available on the Institute's website and YouTube channel, enabling researchers to revisit them at any time.

Image Credits

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

Page 6	– 3D printed models of SARS-CoV-2 virus and spike protein – NIH
Page 6	– Computer rending of SARS-CoV-2 virus – CDC, Alissa Ecker/MS and Dan Higgins/MAMS
Page 10	– Computer rending of SARS-CoV-2 virus – CDC, Alissa Ecker/MS and Dan Higgins/MAMS
Page 12	– 3D printed models of SARS-CoV-2 virus and spike protein – NIH
Page 14	– Man undergoing COVID-19 nasal swab testing – Lukasmilan, Pixabay
Page 14	– Geographic distribution of SARS-CoV-2 lineages over time – *Science* 2021; **371:** 708-712, CC BY-SA 4.0
Page 16	– Lung cancer cells – Anne Weston, Francis Crick Institute
Page 17	– Endometrium histology slide – Dr Luiza Moore, Wellcome Sanger Institute
Page 18	– Histology slide – The Cancer Genome Atlas project
Page 18	– Breast cancer cells – Annie Cavanagh
Page 23	– Skin Cell Atlas visualisation – Newcastle University
Page 24	– Gut cells – Kenny Roberts and Sophie Pritchard, Wellcome Sanger Institute
Page 27	– Skeleton – Directorate General of Antiquities (Lebanon)
Page 29	– Baby feet cradled in adult hands – Michael Fallon, Unsplash
Page 30	– Mosquito – Emphyrio, Pixabay
Page 31	– Fluorescent image of mosquito – Gianmarco Raddi, Wellcome Sanger Institute
Page 32	– *Schistosoma mansoni* – Jayhun Lee, Postdoctoral Fellow, Newmark Lab, Morgridge Institute for Research
Page 32	– *Streptococcus pneumoniae* – Dan Higgins – Medical Illustrator, CDC
Page 33	– *Klebsiella pneumoniae* – National Institute of Allergy and Infectious Diseases (NIAID)
Page 34	– Phylogeny chart of *L. donovani* and *L. infantum* samples – *eLife* 2020; **9:** e51243
Page 36	– Leaf – Josch13, Pixabay
Page 38	– Ugandan Red Colobus Monkey – Charles J Sharp, Sharp Photography, CC BY-SA 4.0
Page 40	– Hand with rainbow light – cm_dasilva, Pixabay
Page 45	– Pancreatic cancer cells – Anne Weston, Francis Crick Institute
Page 45	– Lung cancer cells – Anne Weston, Francis Crick Institute
Page 50	– Podcast mike – Dan LeFebvre, Unsplash

Wellcome Sanger Institute Highlights 2020/21

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, 2021.

FSC®
www.fsc.org
MIX
Paper from responsible sources
FSC® C001785

Printed by Park Communications
on FSC® certified paper.

Park 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
CONRAN DESIGN GROUP

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

sanger.ac.uk