wellcome sanger institute

Leading the way

Highlights 2018/19

What we do

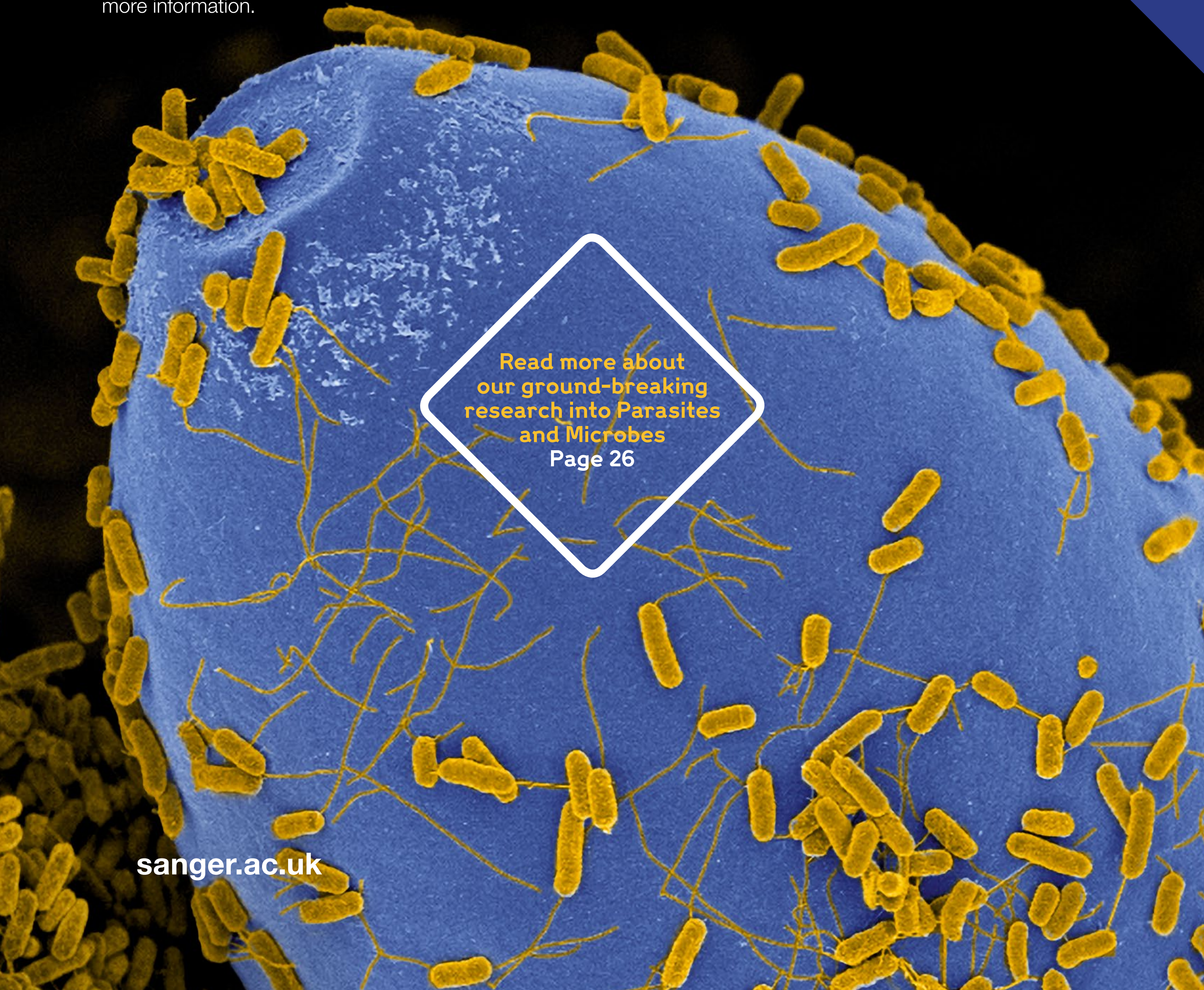
02 Director's Introduction

04 At a Glance

World-class research and innovation that impacts people's lives

Through our cutting-edge infrastructure, scientific independence and innovative ideas we engage in long-term exploratory projects that drive forward science

Wherever you see this dark blue text or yellow, click to find more information.



Read more about our ground-breaking research into Parasites and Microbes
Page 26

08

Our work

10 Cancer, Ageing and Somatic Mutation

16 Cellular Genetics

20 Human Genetics

26 Parasites and Microbes

34 Tree of Life

40

Our approach

42 Scale

44 Innovation

46 Culture

48 Influence

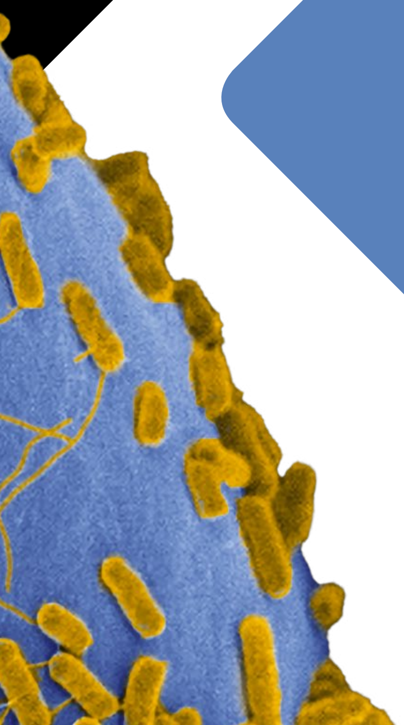
50 Connections

52

Other info

52 Image Credits

53 Institute Information



What we do

Director's Introduction

The past year has been one of great joy and sadness: we celebrated our 25th anniversary and we mourned the death of our first director, John Sulston. Under John's guidance, the Sanger Institute determined one-third of the human reference genome, providing a fundamental gift to life science research worldwide.

It seemed most fitting to seek to make another foundational gift to science as our main celebration. Our 25 Genomes for 25 Years project has delivered high-quality genomes for UK species for which no reference existed. This work is catalysing research and aiding conservation efforts in areas as diverse as golden eagle resettlement and fen raft spider breeding.

There is much that genomics can bring to the ecological issues facing our planet and our communities, and we are proud to have founded a new research programme – The Tree of Life. It will lead the UK-wide collaboration between genomic, conservation and ecology partners to produce reference genomes for all 66,000 species of complex organisms found in the British Isles.

This new work will also serve to strengthen our research into human health. A paradigm shift in sequencing technologies has opened up routine, high-throughput whole-genome sequencing of human genomes. Advances in single-cell technologies, machine learning, and cellular modelling are dissecting health and disease within single cells in individual tissues, over time. Our challenge and goal is to lead the way in extracting the maximum value from this data.

One such project is the UK Biobank Vanguard project, where our sequencing pipeline is delivering 50,000 volunteers' whole-genome sequences to aid genetic exploration of health and disease.

By marrying the wealth of genomic data with electronic health records and other molecular characterisation tools, our researchers have discovered new gene variants associated with osteoarthritis, developed prognostic calculators for cancer, and created an artificial intelligence that predicts new pathogen emergence.

Our established approaches continue to deliver important discoveries. Our malaria researchers applied genome-wide saturation mutagenesis to identify *Plasmodium falciparum*'s essential genes, revealing new therapeutic targets. While our collaboration with the NHS – the Deciphering Developmental Disorders (DDD) project – celebrated its eighth year of delivering diagnoses to children with developmental disorders. After 125 papers and more than 4,500 diagnoses, the project continues to deliver fresh insights into the mechanisms of inheritance, and has identified 49 previously unknown genetic disorders.

As we move into our next quarter century, the promise of genomic science to improve health and conservation is greater than ever, and we are privileged to help deliver the next wave of discovery.

Professor Sir Mike Stratton, Director
Wellcome Sanger Institute

Discover how we have driven forward genomic understanding in 2018

LEICA

At a Glance

In 2018, the Sequencing Centre outputted almost **7,558bn** DNA bases a day

In 2018, we read the genomes of **338** species

The human genome is approximately **3bn** bases long

In 2018, we read the equivalent of one gold-standard (30x) human genome every **17 mins**

In 2018, the Sequencing Centre provided the equivalent of **588** gold-standard (30x) human genomes a week

Centre-based high performance compute cores **20,000**

Usable storage in the Data Centre **55PB**

Total number of compute cores **32,000**

Centre-based cloud-based flexible compute cores **12,000**

Network backbone speed **400GB/sec**

2018 timeline

Jan

Sarah Teichmann awarded Genetics Society's 2018 Mary Lyon Medal

Feb

HDR UK funding for Cambridge hub

Innovations' Adrian Ibrahim chairs new BIA Genomics Advisory Committee

Multidrug resistant malaria spread under the radar for years in Cambodia

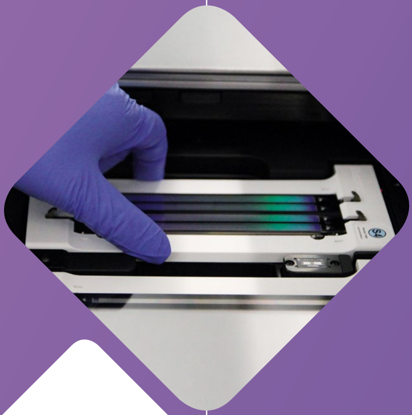
Mar

Institute's founding director, John Sulston, died
Page 47

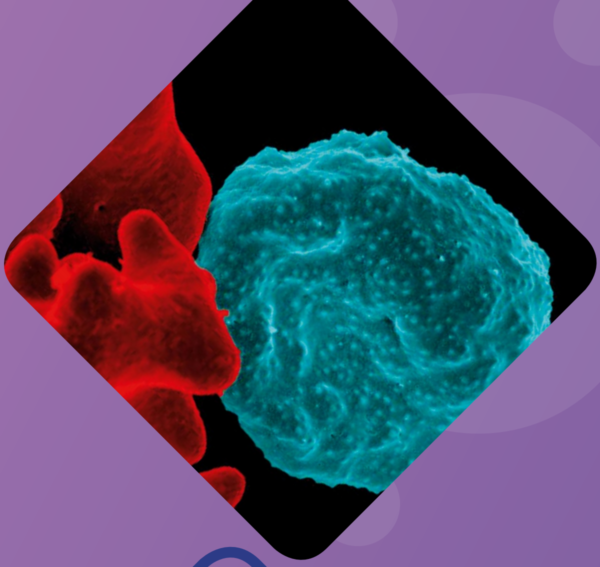
New osteoarthritis genes discovered

Apr

Seeds of kidney cancer sown in adolescence



UK Biobank agreement to sequence 50,000 human genomes
Page 43



May

Malaria screen finds genes essential for life
Page 31

Celgene joins Open Targets

Machine learning flags potential pathogens

Dominic Kwiatkowski joins Fellowship of the Royal Society

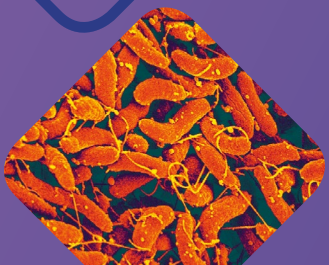
Peter Campbell honoured by EMBO

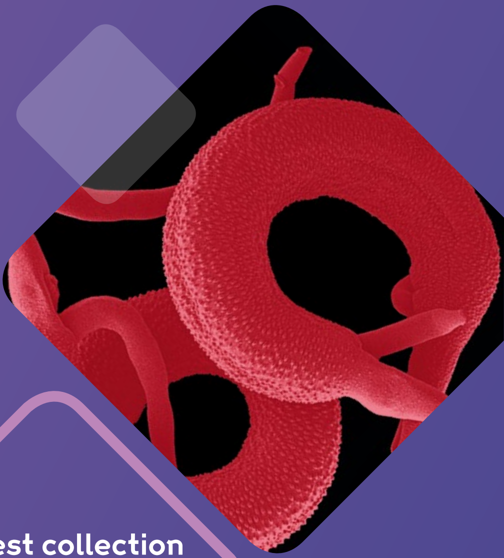
Jun

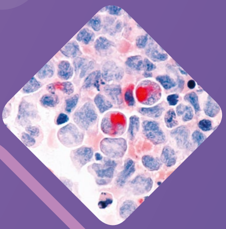
NCTC 3000 genomes reveal antibiotic resistance history

Microbiotica and Genentech sign strategic partnership

Cholera tracked at household level
Page 28







Jul

Blood test could predict leukaemia
Page 11

Genome damage from CRISPR-Cas9 editing more extensive than thought

Aug

Human Cell Atlas boosted by Wellcome funding

Kidney cancer's developmental source revealed

Golden eagle genome sequenced
Page 39

Sep

Skin is a battlefield for mutations
Page 10

Rare genetic diseases more complex than thought

Oct

Personalised prediction of blood cancers

Sanofi joins Open Targets

25 Genomes for 25 Years project
Page 34

Nov

Largest collection of worm reference genomes produced
Page 27

Launch of UK Darwin Tree of Life Project to support Earth BioGenome Project

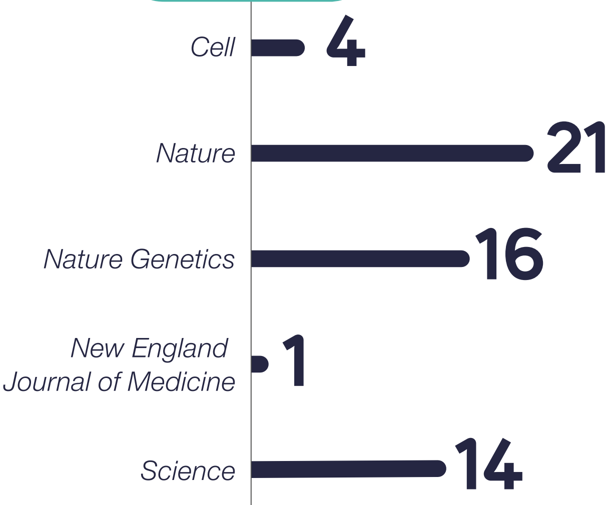
AI-created CRISPR-Cas9 editing prediction tool

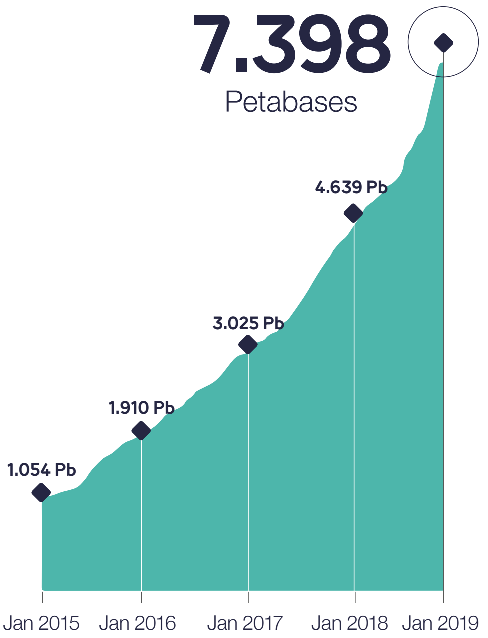
Dec

IT's award winning cloud compute
Page 42

Year in numbers

507
Institute publications in 2018

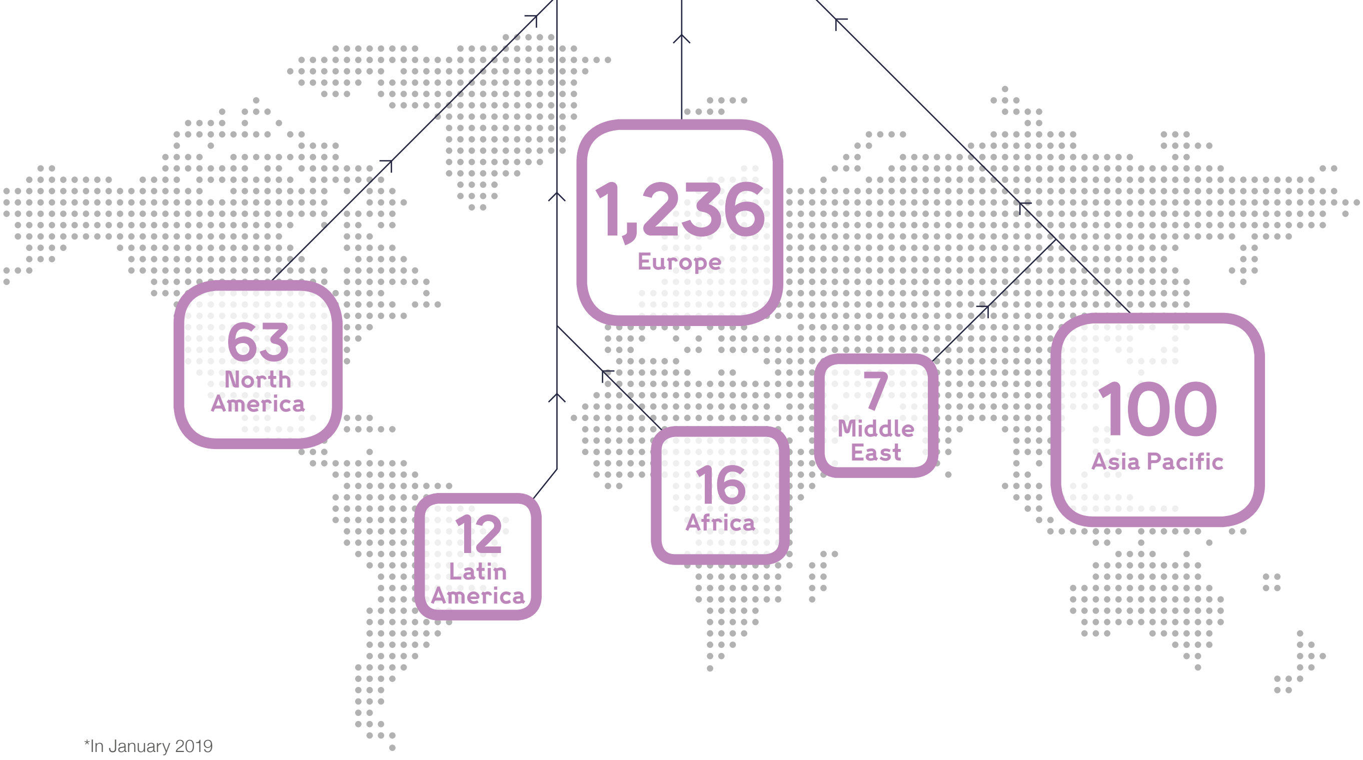




Who published our work?

Where Sanger staff are from*

How much DNA was sequenced?



*In January 2019

Our work

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

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.

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.

20 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.

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.

34 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.

Discover how we tracked cholera house by house
Page 28

Read how Himalayan genomes show how humans adapted to high altitudes
Page 24

Find out how we use single-cell techniques to explore foetal development
Page 19

Cancer, Ageing and Somatic Mutation

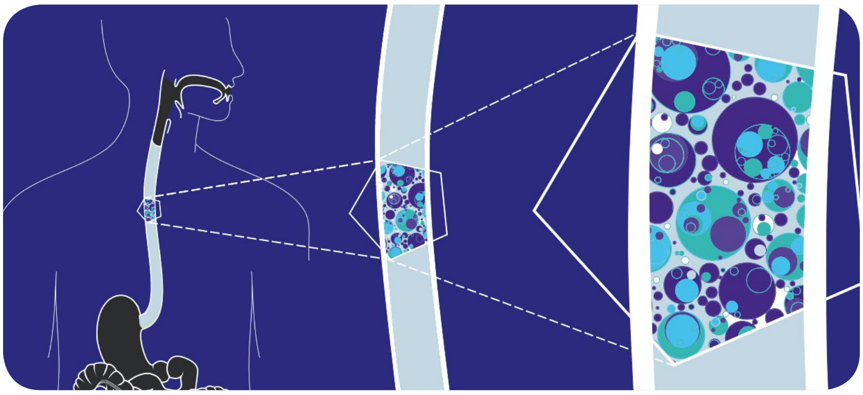
In this section

1 Your body: a Darwinian battlefield
2 Blood test could predict leukaemia
3 One-stop shop for cancer models
4 Cancer calculator gives personalised predictions
5 Genomic microscope sorts lions from lambs
6 'Fast' bone cancer has slow origins
7 Cancer's seeds sown decades earlier
8 Ecological approach reveals hidden stem cell population

1

Your body: a Darwinian battlefield

Research by Sanger Institute cancer researchers, published in *Science* and *Cell Stem Cell*, reveals that our bodies are evolutionary battlefields where only the fittest cells survive.



As our bodies age, our cells acquire an increasing amount of mutations in their genomes, making cancer development ever more likely. While much work has been undertaken to understand the mutations that lead to tumour formation, much less is known about how healthy tissue changes over time. For example, in 2015, Sanger scientists showed that 25 per cent of eyelid cells in people without cancer had at least one cancer-associated mutation.

To address this, the team applied deep-targeted single-cell sequencing to mouse skin samples and human oesophageal tissue to explore how mutational burden develops in healthy cells.

The work revealed that the tissues were made up of a patchwork of clones – many with mutations in cancer-associated genes – all competing for space as they grew. In the mice, cells with a well-documented cancer-driving mutation divided rapidly at first but, after six months, were out-competed by other clones. Even when the mice were exposed to UV light, mimicking the effects of the sun, the clones with cancer mutations didn't survive and progress to tumour formation.

Much to the researchers' surprise, the same was true in oesophageal tissue. Despite being shielded from UV light, a large proportion of the cells carried cancer-associated mutations. Yet tumour formation had not occurred. In fact, by the time we reach middle age, the team discovered, we are likely to have more cells with cancer-associated mutations than not.

These findings have important ramifications when seeking to understand genetic progression to cancer. Mutations in certain genes commonly associated with cancer, for example *NOTCH1*, were more prevalent in healthy cells than tumour cells. Does this mean that genes that are classified as cancer-driving may, in some circumstances, have a protective effect?

While further work is needed to answer such questions, the results show the fundamental importance of considering healthy tissue when studying the origins of cancer.

References

Martincorena I *et al*. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 2015; **348**: 880–886.

Murai K *et al*. Epidermal tissue adapts to restrain progenitors carrying clonal *p53* mutations. *Cell Stem Cell* 2018; **23**: 687–699.e8.

Martincorena I *et al*. Somatic mutant clones colonize the human esophagus with age. *Science* 2018; **362**: 911–917.

2

Blood test could predict leukaemia

Results of a Sanger Institute co-led study in *Nature* show that genomic analysis can identify those at high risk of developing acute myeloid leukaemia (AML) years before symptoms become apparent.

The analysis of blood cells can reveal subtle changes in the genomic landscape and clonal makeup that predict AML development.

Often AML doesn't have any symptoms early on. It is aggressive and usually appears very suddenly in patients who present with the acute complications of bone marrow failure. Yet the roots of the disease form much earlier and identification of these genomic markers would enable more effective monitoring and treatment.

To understand the genomic changes that precede symptoms, Sanger researchers and colleagues at European Bioinformatics Institute (EMBL-EBI) collaborated with the European Prospective Investigation into Cancer and Nutrition (EPIC). They analysed the genomes of blood cells from 124 people with AML that had been taken before they developed the disease and compared the results with those of 676 people who did not go on to develop any form of cancer.

When compared with unaffected people, the genomes of blood cells from healthy people who went on to develop AML contained more numerous genetic mutations in their blood cells, showed greater clonal expansion of these affected cells and showed enrichment of mutations in certain genes. The differences were specific enough to allow the team to accurately predict which people, from a separate group who had also provided blood samples over time, went on to develop AML.

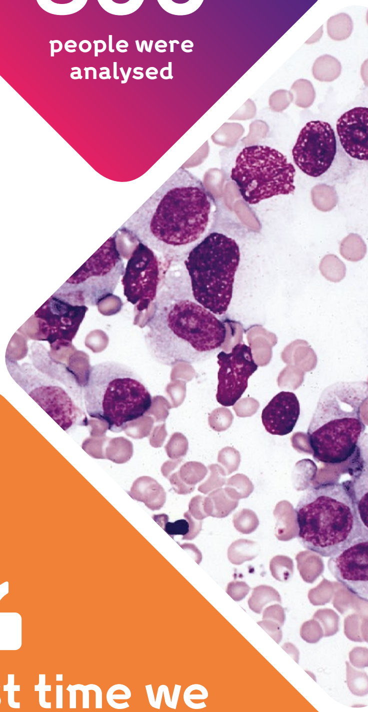
While the subtlety of the differences means that further work is needed to refine the prognostic model, the work provides proof-of-principle that it may be possible to develop tests to identify people at a high risk of developing AML.

Reference
Abelson S *et al*. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 2018; **559**: 400–404.

The blood cell genomes of

800

people were analysed



> **For the first time we can identify people at risk of developing AML many years before they actually develop this life-threatening disease.”**
>
> **Dr George Vassiliou,**
> Wellcome Sanger Institute and the Wellcome-MRC Cambridge Stem Cell Institute

3

One-stop shop for cancer models

A new online resource developed by Sanger Institute scientists is powering worldwide cancer research by speeding the selection and application of cancer cell lines and organoids: Cell Model Passports.

By applying large-scale high-throughput genome sequencing and analysis, researchers in the Sanger Institute's Cancer, Ageing and Somatic Mutations Programme have developed a 'one-stop shop' for information on more than 1,200 cancer models from 43 cancer types in 29 tissues. Through the online hub researchers can access centralised high-quality raw and processed genome sequence data, details of key gene mutations, functional datasets (including drug susceptibility), and details about the original tumour for each cell line or organoid.

Many of the organoids detailed in Cell Model Passports are grown from tumour tissues sent to the Institute from four clinical sites across the UK. This important collaboration is part of the Human Cancer Models Initiative, an international project to generate new cancer cell models.

Being able to study cancer cells in the laboratory – how they behave, and what drugs they respond to – is critical for cancer research. By reducing the time required to select the most appropriate model to study, the work will help to speed research into the genes essential for cancer growth and potential drug targets. In addition, the resource will aid the reproducibility and expansion of cancer research.

Cell Model Passports will be regularly updated with new cell models, and genomic and functional datasets as they are generated.

Reference

van der Meer D *et al*. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Research* 2018; gky872. Published online.

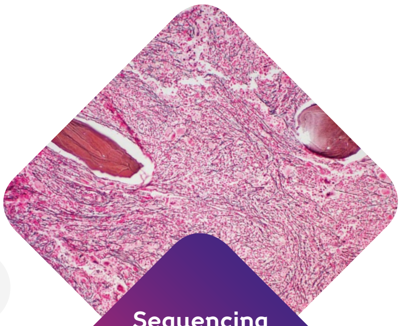
Cell Model Passports is a one-stop shop for information on

1,200

cancer models from 43 cancer types in 29 tissues

4

Cancer calculator gives personalised predictions



Sequencing
69
genes in 2,000 people revealed 8 genomic subtypes of myeloproliferative cancers

In a further proof-of-principle that genomics can power personalised medicine, Sanger Institute researchers have developed an accurate online calculator to predict the outcome of disease for patients with certain types of chronic blood cancer.

Developed in conjunction with the Wellcome-MRC Stem Cell Institute and the University of Cambridge, the freely available resource combines an individual's genetic data with their clinical information to provide greatly enhanced prognoses.

Published in the *New England Journal of Medicine*, the team detailed how sequencing the gene-coding regions of 69 genes associated with cancer in more than 2,000 patients with myeloproliferative neoplasms identified eight genomic subtypes. People within these eight groups – defined for the first time by their underlying causal biology – showed markedly different blood counts, risk of leukaemic transformation and length of time before cancer recurrence.

Taking the analysis even further, the scientists combined 63 clinical and genomic variables to develop personalised predictions of each individual's outcome. This approach successfully produced much more accurate prognoses than current techniques and may help to inform treatment choice.

This work builds on previous work in 2017 by the team, where they used a similar approach to develop a knowledge bank to determine disease outcome and best treatment choice for acute myeloid leukaemia.

References

Grinfield J *et al.* Classification and personalized prognosis in myeloproliferative neoplasms. *New England Journal of Medicine* 2018; **379**: 1416–1430.

Gerstung M *et al.* Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics* 2017; **49**: 332–340.

5

Genomic microscope sorts lions from lambs

For the first time, a genetic marker for osteoblastoma has been discovered.

The discovery by scientists at the Sanger Institute, published in *Nature Communications*, may lead to a diagnostic tool to distinguish between benign osteoblastoma and aggressive osteosarcoma tumours. It could help spare young people from unnecessary and painful treatment.

Osteoblastoma is the most common benign bone tumour, mainly affecting young people aged 10-25 years. It is treated by surgery to remove the tumour and ease symptoms. But its diagnosis is challenging: under the microscope, the tumours can look very similar to osteosarcoma. In contrast to its benign lookalike, osteosarcoma is an aggressive form of bone cancer that requires extensive treatment – sometimes including amputation and intensive chemotherapy.

By conducting whole-genome and transcriptome sequencing of five human osteoblastomas and another form of bone cancer (osteoid osteoma), the researchers discovered rearrangement of the transcription factor gene *FOS* in five of the tumours and rearrangement of *FOSB* (its genetic cousin) in the other. By applying the molecular genomic techniques of FISH (fluorescent in situ hybridization) and immunohistochemistry to another 55 osteoblastoma and osteoid osteoma samples, the genetic changes were found to be ubiquitous and diagnostic.

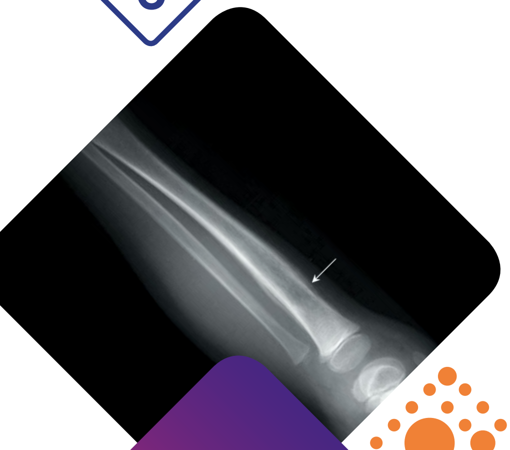
While *FOS* has been implicated in tumour formation before – for example changes in v-fos has been shown to cause osteosarcoma in mice – it is the first time that the transcription factor has been associated with human bone cancer.

Reference

Fittall MW *et al.* Recurrent rearrangements of *FOS* and *FOSB* define osteoblastoma. *Nature Communications* 2018; **9**: 2150.

Osteoblastoma mainly affects young people aged
10-25 years

6



42%

of patients with Ewing Sarcoma showed chromoplexy – formation of loops of DNA that disrupt multiple genes

‘Fast’ bone cancer has slow origins

Ewing sarcoma is thought to be a fast-growing cancer in children, with only harsh chemotherapeutic and surgical treatment options.

Genomic research by the Sanger Institute and the Hospital for Sick Children (SickKids) in Canada, has shown that the cancer actually has slow-growing roots whose detection might enable swifter diagnosis and more successful treatment.

Some tumours are driven not by mutations within individual genes, but by chromosomal rearrangements, leading to gene fusions. Ewing sarcoma is one such cancer, driven by the fusion of the genes *EWSR1* and *ETS*. But how does this gene fusion occur, and could the causative mechanism be important?

By sequencing the entire tumour genomes of 124 children, the collaboration uncovered two key mechanisms at work. The process responsible determined the aggressiveness of the resulting tumour. In addition, these rearrangements occurred many years before the disease is usually detected.

While the majority of children had simple gene fusion of *EWSR1* and *ETS*, 42 per cent had undergone chromoplexy. First identified in prostate cancer, chromoplexy involves the formation of dramatic loops of DNA that disrupt multiple genes.

Those tumours with chromoplexy were found to be more aggressive than those due to simple gene fusion. Developing a genomics-based diagnostic to identify this process might enable tumours to be detected and treated when they are smaller and easier to treat.

Reference
Anderson ND *et al*. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* 2018; **361**: 6405.

7

Cancer’s seeds sown decades earlier

Sanger researchers, together with colleagues at the Francis Crick Institute and the TracerX Renal consortium, have discovered that the first seeds of kidney cancer are sown years, or even decades, before symptoms appear.

To understand how one type of kidney cancer – clear cell Renal Cell Carcinoma (ccRCC) – develops, the team used genomic archaeology techniques to reconstruct the genetic changes that took place. The approach, which they had developed in 2017, is able to trace back the time when each individual mutation appeared in the body, all the way to the first few hours of the developing embryo in the womb.

Reporting in *Cell*, they found that the first seeds of kidney cancer are sown in childhood or adolescence, 40 or 50 years before any cancer develops – and that many of us will be carrying these cancer-associated changes. Yet most people do not develop kidney cancer, and it is hoped that this work will help with mapping the journey of mutation acquisition that is needed for the cells to become malignant.

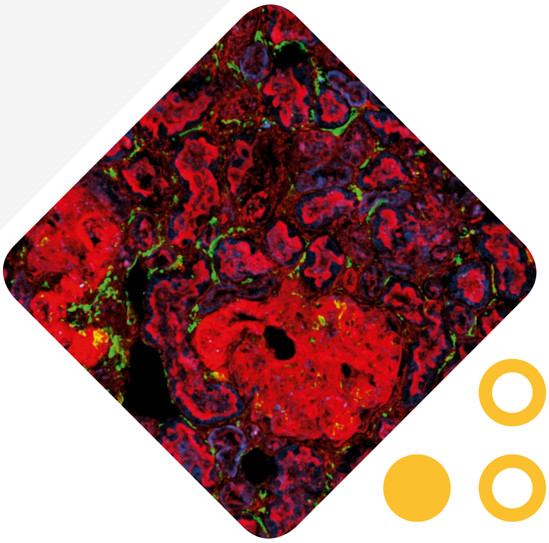
By analysing the whole genomes of 95 biopsies from 33 patients with ccRCC, the researchers found that the first mutation in more than 90 per cent of people was the loss of the short arm of chromosome 3, which removed a number of tumour-suppressing genes. They also found that 35-40 per cent of these people had also gained the long arm of chromosome 5 at the same time, due to a process known as chromothripsis – where chromosomes shatter and are put back together.

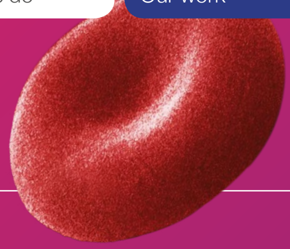
By identifying the key genetic changes over time that drive a kidney cell to cancer, the researchers hope that the knowledge will help to inform early detection of, and intervention for, people at high risk of developing the cancer.

References
Mitchell TJ *et al*. Timing the landmark events in the evolution of clear cell renal cell cancer: TracerX Renal. *Cell* 2018; **173**: 611–623.e17.

Ju YS *et al*. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* 2017; **543**: 714–718.

Stephens PJ *et al*. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011; **144**: 27–40.





8

Ecological approach reveals hidden stem cell population

Sanger researchers used novel techniques to estimate the populations of blood stem cells in healthy adults.

The new approach, detailed in *Nature*, could lead to insights into cancer development, and why some stem cell therapies are more effective than others.

The method was a genomic form of 'capture-recapture' – used by ecologists to study animal populations. The team conducted whole-genome sequencing of 198 blood and bone marrow stem cell colonies from a 59-year-old man to identify and assign sets of mutations unique to each stem cell and its descendants. By searching for these mutations in the rest of the blood, the scientists could work out which blood cells had derived from these stem cells and estimate the total stem cell population.

They found that healthy adults have 10 times more blood stem cells than previously thought – somewhere between 50,000 and 200,000.

The technique is very flexible. Not only can researchers now measure how many stem cells exist, but also they can see how related the cells are to each other and what types of blood cells they produce. The approach can also be applied to study the effects of ageing and disease in other organs.

The team will now use this method to understand stem cell populations in patients with blood cancers. They hope to learn how single cells expand their numbers, outcompete normal stem cells and lead to tumour formation.

Reference
Lee-Six H *et al*. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 2018; **561**: 473–478.

Healthy adults have
10 times
more blood stem cells than previously thought

Cellular Genetics

In this section

1 CRISPR-Cas9 damage worse than thought

2 Machine learning improves CRISPR-Cas9 gene editing

3 Human Cell Atlas bears fruit

4 Immune system evolution reveals its Achilles' heel

5 Immune system secrets unlocked by new CRISPR system

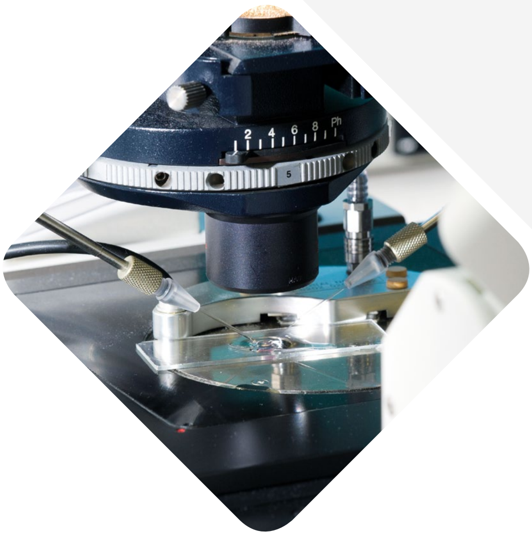
6 How baby placates mum's immune system

1

CRISPR-Cas9 damage worse than thought

Work by Sanger Institute researchers has shown that the widely used CRISPR-Cas9 genome editing method can cause more extensive DNA damage than previously thought.

> We found that changes in CRISPR/Cas9 edited cells have been seriously underestimated before now."
>
> **Professor Allan Bradley,** Director Emeritus at the Wellcome Sanger Institute



They found large deletions and genome rearrangements, in some cases far from the targeted site. These changes could lead to dangerous changes in some cells.

Until now, analysis of CRISPR-Cas9 editing has shown that the damage caused near to the targeted site was low – usually involving just a few DNA bases. However, standard lab-based analysis techniques used to search for this damage require specific sites on the genome to work and sometimes CRISPR can remove them.

To investigate whether or not the process had other unwanted effects, the scientists studied – for the first time – the whole genomes of CRISPR-Cas9 edited cells. They looked at mouse embryonic stem cells, mouse blood-making cells, and human retinal cells. In addition, to overcome the shortcomings of standard lab-based techniques, the team employed long-read sequencing and long-range genotyping for analysis.

Reporting their findings in *Nature Biotechnology*, the researchers discovered that areas of the genome close to the target site had been deleted. In some cases, these were large stretches of DNA, up to several thousand bases long. Further away from the target site, they found complicated rearrangements of DNA, where once-distant stretches of the genome had been fused together.

The team warns that standard tests used to detect unwanted DNA changes may miss important forms of damage caused by CRISPR-Cas9 editing. They counsel that clinical use of the technique in gene editing therapies should undergo specific testing.

Reference

Kosicki M *et al.* Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 2018; **36**: 765–771.

2

Machine learning improves CRISPR-Cas9 gene editing

CRISPR-based gene editing techniques have revolutionised genetic research, by reducing the time and cost of gene knockout experiments.

However, the process is not perfect and can produce off-target mutations that might bias genetic experiments or produce harmful outcomes.

The amount – and location – of these mutations can differ depending on the cell type, even within the same organism. However, this DNA damage is not random and appears to be related to the guide RNA (gRNA) template being used. Armed with this knowledge, Sanger scientists sought to develop a reliable and accurate prediction tool that could help minimise unwanted changes.

Writing in *Nature Biotechnology*, the team outline how they made over 41,000 gRNAs and applied them to cells with a range of genetic backgrounds using different CRISPR reagents. These edited genomes were then sequenced at high coverage to identify off-target changes.

Next, the scientists applied machine learning to the more than 1,000,000,000 mutational outcomes to develop their computational tool. Dubbed FORECasT (favoured outcomes of repair events at Cas9 targets), the resulting system is able to predict the exact mutations resulting from CRISPR-Cas9 genetic editing just from the sequence of the target DNA.

Researchers can use the freely available program to select the most appropriate gRNA, and screen out any damaged cells by using targeted DNA sequencing to look for predicted off-target damage.

Reference
Allen F *et al*. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology* 2019; **37**: 64–72.

CRISPR-Cas9 off-target damage

41,630 gRNAs in synthetic constructs created

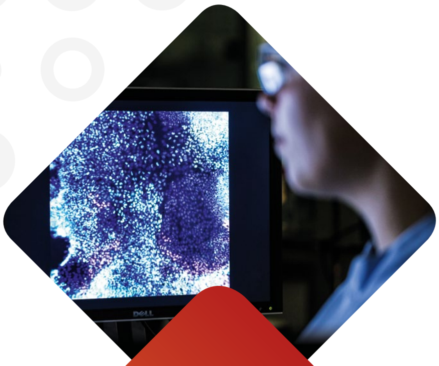
1,000,000,000 mutational outcomes analysed

6,568 human gene targeting gRNAs studied in greater depth

58% of off-target damage are deletions of at least 3 bases

31% of off-target damage occur between repeating sequences of at least 2 bases

3



Gene activity profiles for more than **500,000** individual cells have been released

Human Cell Atlas bears fruit

Scientists around the world are already benefiting from the work of the Human Cell Atlas, the ambitious global collaboration to create a ‘Google Map’ of the human body’s 37 trillion cells.

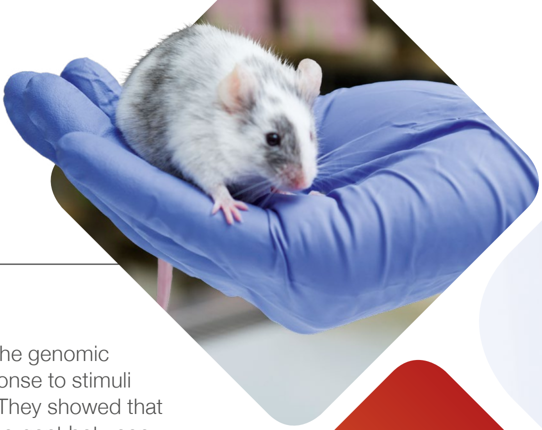
Co-led by the Sanger Institute and the Broad Institute, more than 480 scientists are seeking to discover, and pinpoint the location of, every cell type in the human body at different stages of life from foetus to adult. With backing from funders including Wellcome, the Medical Research Council (MRC), National Institutes for Health, and the Chan-Zuckerberg Initiative the project is also exploring how these cells respond to disease or trauma.

Now the first fruits of this research have been made available online to the scientific community: the single-cell gene activity profiles of more than half a million cells.

The data comes from a number of sources. Sanger scientists, collaborating with the MRC Cancer Unit studied thousands of immune and connective tissue cells from a widely used strain of mouse used to study skin melanoma. Cells were collected over 11 days to understand how they respond and adapt to a tumour over time.

In addition, in conjunction with the Cambridge Repository for Translational Medicine, Sanger researchers generated the first single-cell data from a donated human spleen. They characterised individual cells over time to study the effects of oxygen deprivation on the cells.

The Broad Institute and Harvard University have supplied data from individual immune cells from umbilical cord and adult bone marrow.



4

Immune system evolution reveals its Achilles' heel

To understand more about the immune system, Sanger Institute and European Bioinformatics Institute (EMBL-EBI) researchers worked with collaborators to study the activity of individual immune cells in six different mammal species.

In total, the team characterised the activity of thousands of genes in more than 250,000 individual cells from humans, rhesus macaques, mice, rats, pigs and rabbits. They studied the innate immune system in both immune cells and fibroblasts.

The scientists compared the genomic activity of the cells in response to stimuli that imitated a pathogen. They showed that genes which are highly divergent between species also show a wide variety of activation between cells within an individual tissue. In comparison, genes which were conserved between species appeared to be much more tightly regulated and consistent in their activation between cells in a tissue.

Writing in *Nature*, the researchers hypothesise that the balance between divergent and conserved genes has evolved to ensure that immune responses are proportionate and confined. The divergent genes have evolved to enable immune cells to recognise and attack a particular virus or bacteria. In contrast, the conserved ones regulate the level of response to prevent over-activation of the immune response, which could cause inflammatory diseases.

However, this balance also has an inherent weakness. The conserved genes are more tightly regulated and may have other functions in a cell. These represent an Achilles' heel and are targeted by viruses and bacteria to subvert the immune system.

The team characterised the activity of thousands of genes in more than

250,000

cells from humans, rhesus macaques, mice, rats, pigs and rabbits

Reference

Hagai T *et al*. Gene expression variability across cells and species shapes innate immunity. *Nature* 2018; **563**: 197–202.

5

Immune system secrets unlocked by new CRISPR system

88,000

guide RNA templates were made to knock out 20,000 mouse genes

Sanger researchers and collaborators have provided a much-needed overview of a vital part of the immune system: Th2 cells.

These cells help to coordinate, and limit, how the body's immune system fights off an infection. This understanding could help drive research into new therapies for auto-immune diseases, such as allergies and rheumatoid arthritis.

Reporting in the journal *Cell*, the scientists detail how they systematically switched off every gene in a mouse Th2 cell, by creating a new retroviral CRISPR-Cas9 gene editing system. The novel approach was needed because mouse Th2 cells cannot be edited using standard lentiviral techniques. In total, the team made 88,000 guide RNA templates to ensure that they could knock out all 20,000 mouse genes.

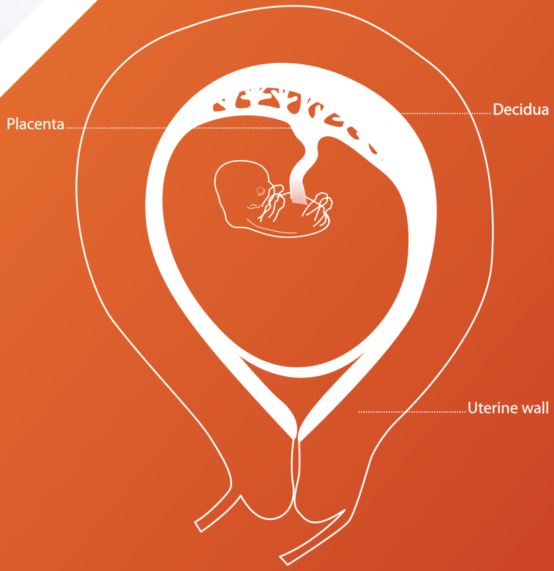
By combining this tool with a new screening protocol, the team was able to study the Th2 cell's genome efficiently and quickly. The result is an atlas of how the cell develops and signals other immune system cells to help direct the body's response.

The researchers identified a number of new genes that regulate Th2 cells' response to infection, providing further avenues of research into auto-immune diseases. In particular, they discovered that a molecule involved in controlling protein production – PPARG – plays an important role.

Until now individual studies have focused on different aspects of the Th2 cells' behaviour, but were difficult to compare or combine. This unbiased genome-wide overview not only confirms many of the previous studies' findings, it also provides a vital map of which genes are switched on and off, and when.

Reference

Henriksson J *et al*. Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. *Cell* 2019; Available online DOI: 10.1016/j.cell.2018.11.044.



6

How baby placates mum's immune system

The first findings from the human developmental cell atlas has shown how foetal cells and maternal cells communicate with each other to prevent the mother's immune system from attacking the baby's placenta. The discoveries will help future research into miscarriages and still births.

The study, published in *Nature*, applied genomic and bioinformatic techniques to approximately 70,000 individual cells taken from the decidua – the junction between the placenta and the uterus – during the first trimester of pregnancy. This is the first time that the maternal-foetal interface has been studied in such detail.

The researchers used DNA and RNA sequencing to identify maternal and foetal cells in the decidua and determine which genes were active in each cell type.

To systematically study the interactions between the maternal and foetal cells and uncover those involved in modifying the mother's immune system, the team first created a publicly available repository of cell-surface and secreted molecules (www.CellPhoneDB.org). They then interrogated the data to identify molecules involved in signalling between cell types.

They also undertook microscopy studies. Combining this data, the scientists discovered how foetal cells entering the mother's uterus lining stimulate it to produce the blood vessels to supply the placenta.

By providing much-needed data on the very first steps of embryo implantation and placenta formation1, researchers can better understand how these processes work and what goes wrong in pre-eclampsia and miscarriage.

Reference
Vento-Tormo R *et al*. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* 2018; **563**: 347–353.

Human Genetics

In this section

1 Visionary study powers NHS' genomics revolution

2 Rare disorders influenced by common variants

3 Recessive genes play minimal role in undiagnosed developmental disorders

4 UK Biobank data reveals new osteoarthritis drug targets

5 Genetic diagnoses in the womb

6 Himalayan genomes show how humans adapted to high altitudes

7 African hepatitis C virus strains evade Western drugs

> “As a result of the DDD project thousands of children with rare neurodevelopmental disorders now have a diagnosis, and their families can use this to inform their decisions about further children and to guide and target appropriate medical care.”
>
> **Dr Jane Hurst,**
> President of the UK Clinical Genetic Society, and Great Ormond Street Hospital

33,500
parents and children

1

Visionary study powers NHS' genomics revolution

In October 2010, the Sanger Institute launched the Deciphering Developmental Disorders (DDD) study, a visionary project to demonstrate the feasibility of using genome sequencing for diagnosis in the clinic – bringing direct benefits to patients.

It has proven so successful that it spurred the UK Government to initiate the 100,000 Genomes Project and to launch the new NHS clinical genomics service, and has founded an analysis company that is powering clinical genomics services worldwide.

The groundbreaking collaboration with NHS services across the UK and Ireland recruited over 14,000 children and their parents for genome sequencing. When they joined, each child had an undiagnosed developmental disorder. So far, combining genomic insights with clinical data has provided diagnoses to 4,500 children. For many families, the diagnosis has pointed to a treatment that may be beneficial. For others, the diagnosis has brought relief, ending a years-long search for the cause of their child's condition.

A diagnosis can help inform family planning decisions, as often the analysis will tell how likely it is that future children may be affected. Many families who received a diagnosis have connected to others with the same condition – finding support groups around the world.

The researchers have also identified 49 completely new disorders, caused by changes in genes previously unconnected to developmental disorders. The study has also powered international research efforts into childhood disorders, resulting in over 125 published research papers.

Every child in the study has had their genome data analysed, but the work is not yet complete. The project is continuing to deliver new insights. As knowledge builds, through DDD and similar studies across the globe, the power to make new discoveries increases. A recent reanalysis of 1,133 children's genomes resulted in an additional 182 diagnoses. Most of those were due to new disorder-associated genes discovered in the years since the initial analysis. The researchers, reporting their findings in *Genetics in Medicine*, estimate that child-parent genome sequencing as a first-line diagnostic test for developmental disorders would diagnose over 50 per cent of patients.

In addition, the Sanger Institute spin-out company, Congenica, has built on the foundations of the DDD study to develop its Sapientia^{TM} clinical genomics analysis platform that is helping to power medical services around the world, including the UK, China and Portugal.

Reference
Wright CF *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine* 2018; **20**: 1216–1223.

8
years

200
NHS consultants

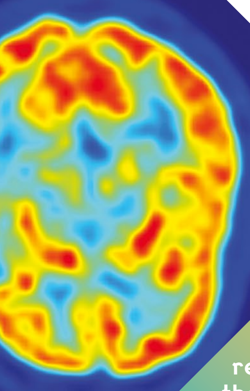
4,500
diagnoses so far

49
completely new disorders

125
research papers

24
NHS Genetics Centres

2



Sanger researchers studied the genomes of nearly

7,000

children

Rare disorders influenced by common variants

Most rare developmental conditions are thought to be caused by damaging mutations in a single gene. However, the symptoms between individuals with the same mutation can differ markedly.

And some individuals who have 'disease-causing' mutations can appear unaffected. To try to understand this phenomenom, Sanger Institute researchers studied the genomes of nearly 7,000 children in the Deciphering Developmental Disorders (DDD) study.

They compared these DNA sequences with those of individuals unaffected by developmental conditions.

The research, published in *Nature*, is the first large-scale study of the role of common genetic variants in rare disorders. Using genome-wide association studies the team tested four million common genetic variants for association with neurodevelopmental disorders.

Surprisingly, the researchers found that these common DNA variations – for example those that increase the risk of schizophrenia – also affected the risk of rare disorders. This is the first time that common genetic variants have been found to have a role in rare developmental disorders.

The team hope their findings will form the basis of further research into how common genetic variation affects both the risks and symptoms of rare diseases.

Reference

Niemi MEK *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* 2018; **562**: 268–271.

3

Recessive genes play minimal role in undiagnosed developmental disorders

Researchers working on the Deciphering Developmental Disorders (DDD) study have discovered that only a small fraction of rare, undiagnosed developmental disorders in the British Isles are caused by recessive gene variants.

The findings, published in *Science*, could help clinicians better predict the risk of a condition affecting any subsequent children, and so help families plan for the future. They also reveal that a large proportion of developmental disorders are due to complex genetic mechanisms that warrant deeper investigation.

The DDD study aims to find diagnoses for children with previously unknown developmental disorders. Sanger Institute researchers analysed the gene-coding DNA of more than 6,000 families to look for recessive causes of their conditions. The team developed a novel bioinformatics approach that allowed the scientists to search in both known and yet-to-be discovered genes.

The researchers estimate that only 5 per cent had inherited a disease-causing gene mutation from both parents, far fewer than previously thought. Separate studies have estimated that half the children in the DDD study have conditions caused by new mutations – mutations that were not inherited from their parents. Together, these findings indicate that, for many patients, more complicated genetic mechanisms may be involved.

The team also identified two recessive genes – *KDM5B* and *EIF3F* – that had not been associated with recessive developmental disorders before.

Dr Hilary Martin, the first author on the paper, is now leading a group in the Sanger Institute Human Genetics programme to identify and explore the genetic complexity of health and disease.

References

Martin HC *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 2018; **362**: 1161–1164.

Deciphering Developmental Disorders study. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* 2017; **542**: 433–438.

Sanger Institute researchers analysed the gene-coding DNA of more than

6,000

families

4

UK Biobank data reveals new osteoarthritis drug targets

In the largest genetic study of osteoarthritis to date, researchers have uncovered 52 new genome variants linked to the condition.

Their findings more than double the number of variants associated with osteoarthritis. Of these, 10 of the variants are in genes that can be targeted by drugs – either already in use for another condition, or in development – accelerating the advance towards new therapies.

Almost 10 million people in the UK have osteoarthritis – a degenerative, painful joint disease with no treatments. The disease indirectly costs the UK economy £14.8 billion per year.

Scientists at the Sanger Institute, GlaxoSmithKline (GSK) and their collaborators used the UK Biobank and arcOGEN resources to analyse the genomes of over 77,000 people with osteoarthritis. They carried out a genome-wide association study of approximately 17.5 million single nucleotide variants in 77,000 people with osteoarthritis and 378,000 healthy individuals. Reporting in *Nature Genetics*, they identified 65 genome variants associated with osteoarthritis, 52 of which were previously unknown.

To understand how these variants contribute to osteoarthritis, the team integrated data from functional genomics and gene expression experiments. They also incorporated transcriptomic characterisation of tissue from osteoarthritis patients undergoing joint replacement surgery, statistical fine-mapping, and evidence from rare human disease and animal models. The variants identified are in genes involved in bone development, collagen formation, or organising the extracellular matrix.

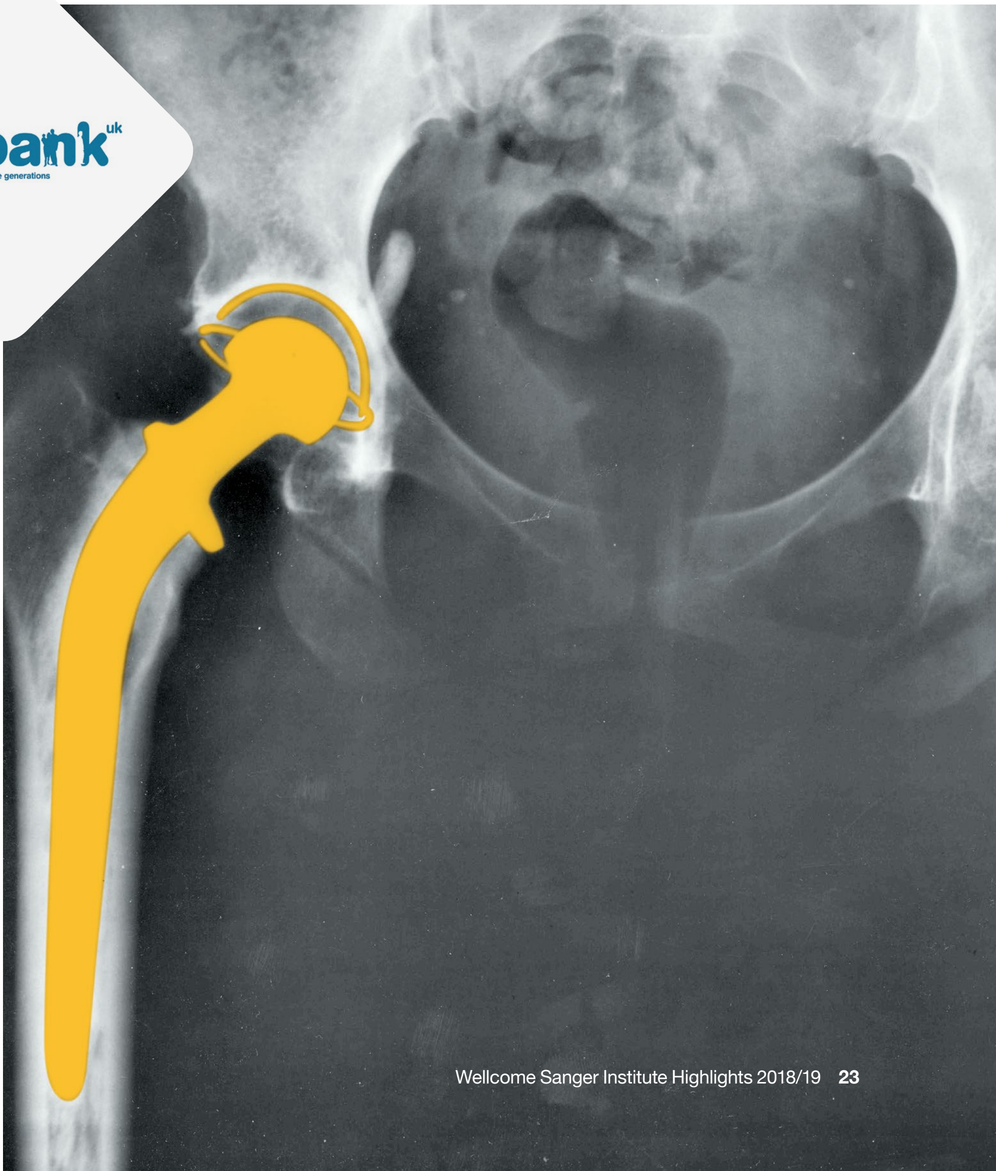
By matching the variants' biological functions to the modes of action of existing or experimental drugs, the researchers identified 10 targetable variants. The findings will play an important role in speeding up drug development.

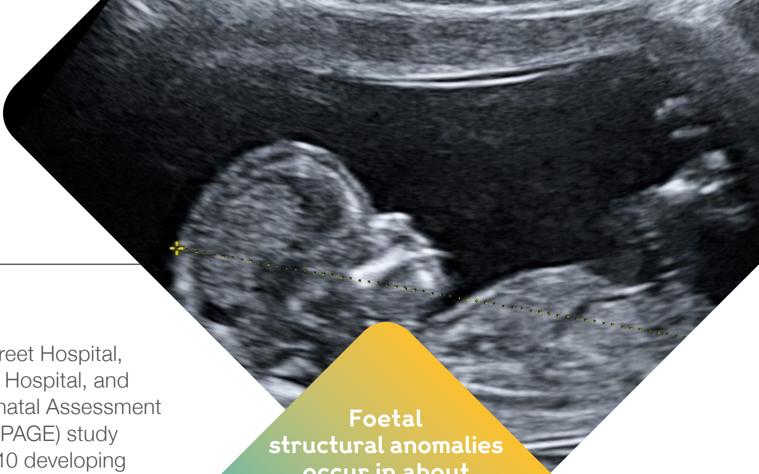
Reference
Tachmazidou I *et al*. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank. *Nature Genetics* 2019; **51**: 230–236.

biobank uk
Improving the health of future generations

52
new genome variants linked to the condition

10 million
people in the UK have osteoarthritis





5

Genetic diagnoses in the womb

Foetal structural anomalies occur in about 3 per cent of pregnancies and are usually detected by ultrasound.

While ultrasound can pick up potential problems, it can't identify their cause. Genetic screening is now routinely used to provide more accurate diagnoses for such abnormalities. This involves looking for additional copies of a chromosome or copy number variation (where sections of a chromosome are duplicated or lost).

However, reading the entire genetic code of the foetus' genes would provide far greater and, potentially, more definitive information on the underlying genetic cause of structural anomalies. So, in the largest study of its kind, researchers at the Sanger Institute, Great Ormond Street Hospital, and Birmingham Women's Hospital, and collaborators from the Prenatal Assessment of Genomes and Exomes (PAGE) study sequenced the genes of 610 developing babies and 1,026 biological parents. Each foetus had abnormalities detected by routine ultrasound – either of the heart, brain, skeleton, or in multiple organs.

Reporting in *The Lancet*, the team were able to provide a genetic diagnosis for nearly 10 per cent of pregnancies. Such diagnoses could be used to provide better information to parents about how their child is likely to be affected, and guide clinical care.

The team focused their analysis on 1,628 genes that have a potential role in developmental disorders. After bioinformatic filtering and prioritisation, 321 genetic variants representing 255 potential diagnoses were selected for review by a multidisciplinary clinical panel. In total, 52 of the 610 pregnancies were diagnosed with a known disorder and a further 24 had a genetic variant of uncertain significance but potential clinical utility.

Foetal structural anomalies occur in about

3%

of pregnancies

The teams hope the method will be widely used, improving diagnoses for families across the UK. The information could provide parents with important information about the outlook for their baby, as well as whether they might face a similar situation in future pregnancies.

Reference
Lord J *et al*. Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* 2019; S0140–6736(18)31940–8.

6

Himalayan genomes show how humans adapted to high altitudes

Researchers from the Sanger Institute and Leiden University have performed the most comprehensive survey to date of genetic variation among people living in the Himalayas.

Their study found that different populations living in the region share a specific ancestral component, suggesting that genetic adaptation to life at high altitude originated once in the region and subsequently spread as populations diverged.

The harsh environment at high altitudes, due to increased ultraviolet radiation, decreased atmospheric pressure and lower levels of oxygen, is inescapable. The team explored the genetic and physiological adaptations of people who have settled in the region, analysing the genomes of 738 individuals from 49 populations in Nepal, Bhutan, North India and Tibet. They looked at 500,000 single nucleotide variants in each individual's genome and compared the data to worldwide population data.

Combining four statistical approaches, they identified regions of the genome that related to adaptations to life at high altitude.

The strongest signals of high-altitude adaptation were near two genes known to be associated with the body's response to low-oxygen environments. They discovered eight additional signals of high-altitude adaptation, five of which have strong biological links to such functions. The team also uncovered the genetic ancestry of the different populations and were able to uncover where they had mixed and diverged over the past 2,000 years.

Their results suggest the presence of a single ancestral population carrying advantageous variants for high-altitude adaptation that separated from populations living in lowland East Asia, and then spread and diverged into different populations across the Himalayan region.

Reference
Arciero E *et al*. Demographic History and Genetic Adaptation in the Himalayan Region Inferred from Genome-Wide SNP Genotypes of 49 Populations. *Molecular Biology and Evolution* 2018; **35**: 1916–1933.

7

African hepatitis C virus strains evade Western drugs

A British-Canadian-Ugandan collaboration has identified three new strains of the hepatitis C virus (HCV) in patients in Uganda, in the largest study of the disease in sub-Saharan Africa.

Genome analysis by the Sanger Institute, the MRC-University of Glasgow Centre for Virus Research and collaborators suggests that these African strains may not be susceptible to antiviral drugs developed to treat HCV infections in Europe and America.

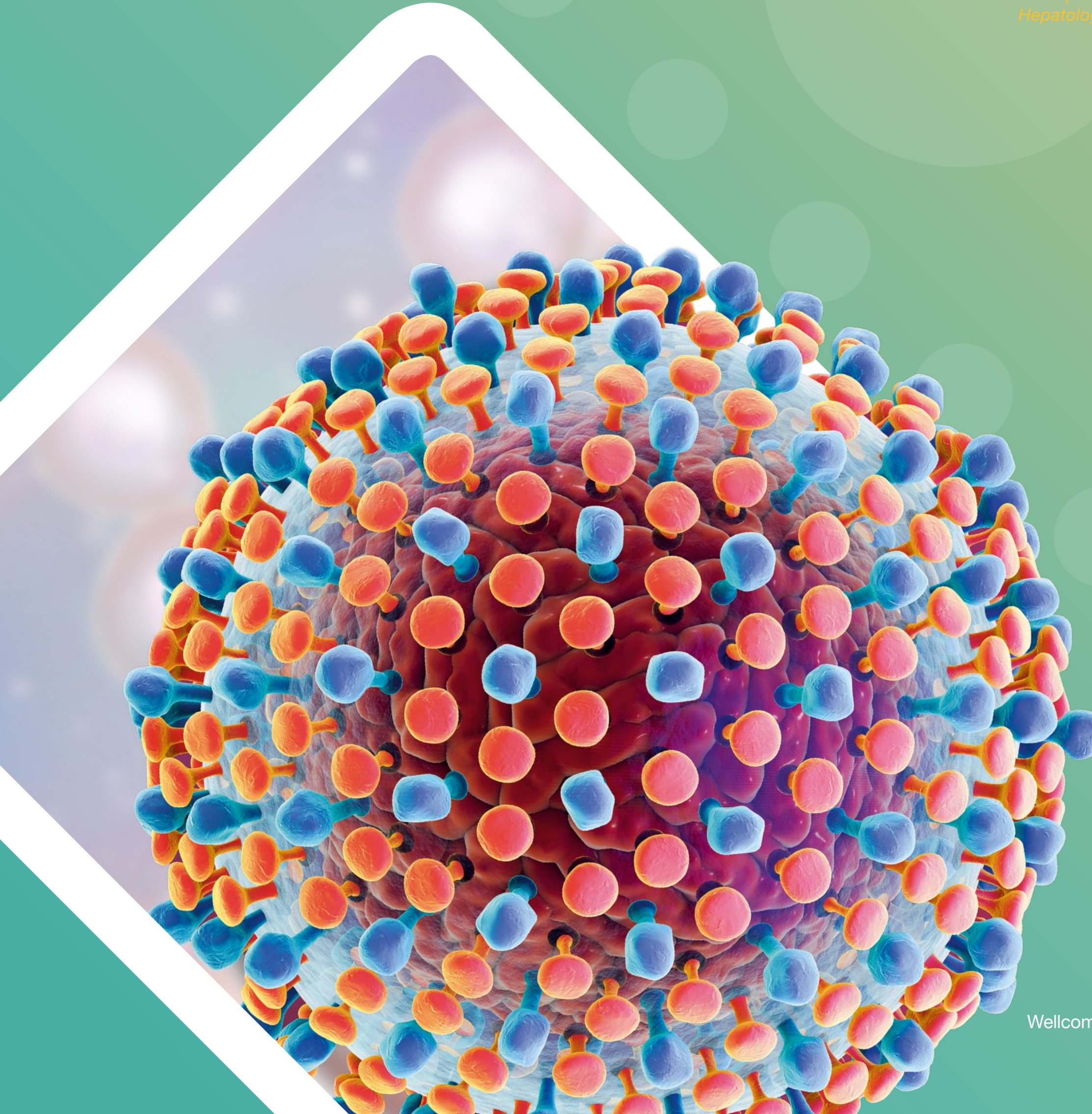
HCV causes hepatitis C – a liver disease that can trigger cirrhosis and liver cancer – and nearly 400,000 people die from it each year. The World Health Organization has set a goal of eliminating the disease by 2030 but, at the moment, an estimated 71 million people around the world have chronic hepatitis C infection. Of these people roughly 10 million live in sub-Saharan Africa.

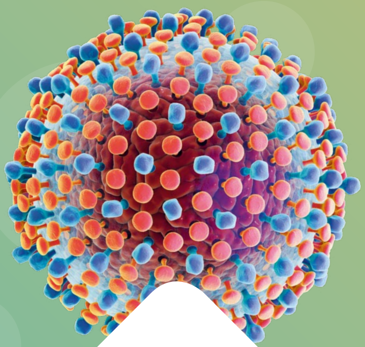
The researchers recruited 7,751 patients from Uganda who were subsequently screened for the virus. They found 20 patients had an undiagnosed hepatitis C infection. Sequencing the virus genome from these individuals, together with two samples from the Democratic Republic of the Congo, revealed the three previously undiscovered strains. The strains have mutations in genes known to be associated with resistance to several commonly used antiviral drugs, highlighting that careful approaches are needed to diagnose and treat HCV effectively in Africa.

Further research to determine the extent of HCV genetic diversity in Africa will aid the development of an effective vaccine and enhance elimination efforts. It will reveal recent and historical transmission patterns to inform public health interventions.

Reference

Davis C *et al*. New highly diverse hepatitis C strains detected in sub-Saharan Africa have unknown susceptibility to direct-acting antiviral treatments. *Hepatology* 2018; doi: 10.1002/hep.30342.





World Health Organization has set a goal of eliminating hepatitis C by

2030

Parasites and Microbes

In this section

1 NCTC 3000 reveals history of antibiotic resistance

2 Largest worm genomes study finds new drug targets

3 Real-time, Europe-wide gonorrhoea monitoring is feasible

4 Tracking cholera – house by house

5 Cholera in a conflict zone

6 How malaria jumped to humans

7 Malaria screen finds genes essential for life

8 Flipping malaria's sex master switch

9 Humans and cows – it's complicated

10 AI identifies emerging pathogens

1

NCTC 3000 reveals history of antibiotic resistance

The National Collection of Type Cultures (NCTC) was set up in 1920 and has become the longest established collection of bacteria in the world.

The first bacteria added to the NCTC was a strain of dysentery-causing *Shigella flexneri* isolated in 1915 from a soldier in the trenches of World War 1. The collection also includes 16 samples deposited by Sir Alexander Fleming, including one from his own nose.

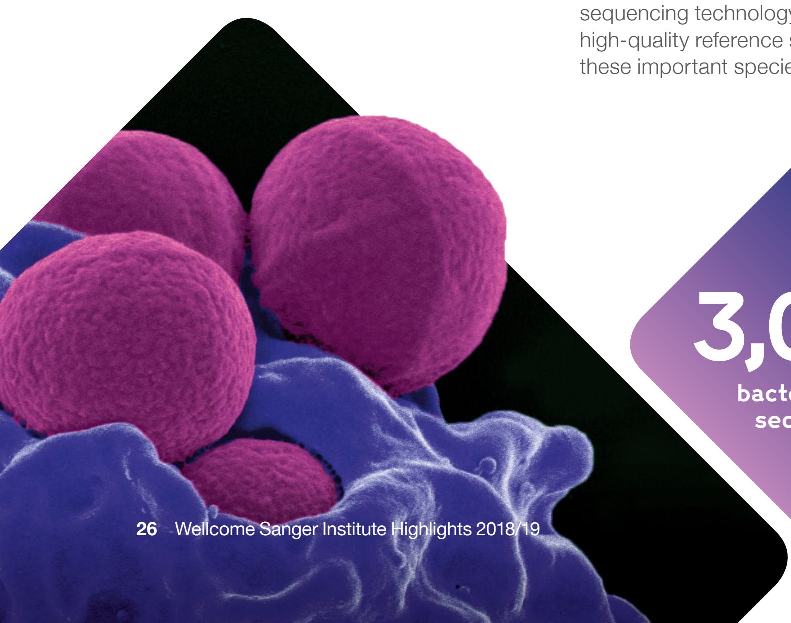
The NCTC contains deadly strains of cholera, plague, MRSA and tuberculosis, making it one of the largest collections of clinically relevant species. Scientists worldwide use it for medical, scientific and veterinary studies.

Now, researchers from the Sanger Institute, European Bioinformatics Institute (EMBL-EBI), Public Health England and Pacific Biosciences (PacBio) have sequenced 3,000 bacteria from the collection. This included sequencing all the 'type strains' of bacteria – those samples which describe a species and are used as a classification standard.

The sequencing was performed using PacBio's single DNA molecule, long-read sequencing technology, resulting in high-quality reference sequences for these important species.

Comparing the historical strains to modern ones enables studies into epidemiology, virulence, prevention, and treatment of infectious diseases. The genome sequences will support further understanding of the bacteria and the diseases they cause. The sequences will also aid in the development of new diagnostics – to rapidly test for infections and identify the source of an outbreak.

Comparing historical and modern strains of bacteria will also aid in tackling the global threat of antibiotic resistance. The genome sequences of historical samples, before the introduction of antibiotics and vaccines, can be compared to those from current strains – allowing researchers to track a species' evolution in response to antibiotics.



3,000

bacteria were sequenced

2

Largest worm genomes study finds new drug targets

A quarter of the world's population are infected with parasitic nematodes (roundworms) or platyhelminths (flatworms).

Infection is rarely lethal, but can cause debilitating pain or chronic diseases such as river blindness and schistosomiasis. To advance research into these neglected tropical diseases, the International Helminth Genomes Consortium, led by Sanger Institute researchers, undertook the largest genomic study of parasitic worms to date.

Reporting in *Nature Genetics*, they compared the genomes of 81 species of roundworms and flatworms, including 45 that had not been sequenced before. Their analysis revealed 800,000 new genes in thousands of new gene families. They found gene families that modulate host immune responses, enable the parasite to penetrate though host tissues (such as the gut), or allow the parasite to feed. The insights will enable researchers to better understand how the worms invade their hosts, and evade immune systems.

At the moment, few drugs are available to treat worm infections. To accelerate the search for new and effective treatments, the team mined the genomic data to look for potential new drug targets. They studied all 1.4 million genes across the parasitic species. Combining this information with the ChEMBL database of bioactive molecules with drug-like properties, they list 40 high-priority drug targets in the worms, and hundreds of possible drugs.

Many of the drug compounds identified by the team are existing treatments for other human illnesses, and have the potential to be re-purposed for deworming. This comparative genomics resource provides a much-needed boost for the research community to understand and combat parasitic worms.

Reference
International Helminth Genomes Consortium. Comparative genomics of the major parasitic worms. *Nature Genetics* 2019; **51**: 163–174.

3

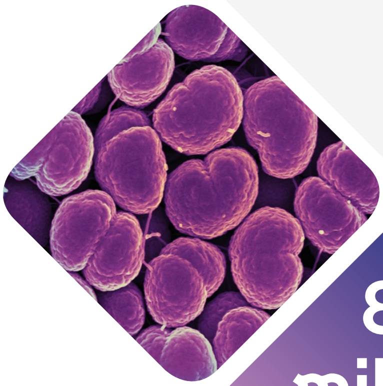
Real-time, Europe-wide gonorrhoea monitoring is feasible

Drug-resistant strains of the *Neisseria gonorrhoeae* bacterium are on the rise. 88 million people worldwide are infected with the sexually transmitted bacteria, which causes gonorrhoea, and increasing numbers of those infections are difficult to treat.

Researchers at the Centre for Genomic Pathogen Surveillance (CGPS) and the Sanger Institute have carried out the first pan-European genome-wide survey of gonorrhoea. Working with the European Centre for Disease Control and collaborators, the team combined sequencing the genomes of 1,054 samples of *N. gonorrhoeae* from 20 countries with the bacteria's phenotypic and epidemiological data. The result is an easily accessible, online map of antibiotic resistance of the bacteria across the continent – PathogenWatch.

Before the samples were sequenced, they were tested locally to identify the strain and any antibiotic sensitivity. Genome sequencing proved to be more accurate than current laboratory techniques in identifying strains resistant to antibiotics, and even highlighted incorrect laboratory results. The findings, reported in *Lancet Infectious Diseases*, will help public health officials monitor emerging resistance and doctors prescribe effective antibiotics.

To aid global efforts to control antibiotic resistance, the team has established a web-based database of *N. gonorrhoeae* genome sequences. The new, open-access resource will support real-time surveillance of gonorrhoea worldwide, enabling researchers to rapidly share, compare and interrogate their data.



88 million
people worldwide are infected by sexually transmitted bacteria

Reference
Harris SR *et al.* Public health surveillance of multidrug-resistant clones of *Neisseria gonorrhoeae* in Europe: a genomic survey. *Lancet Infectious Diseases* 2018; **18**: 758–768.

4

Tracking cholera – house by house

There are three to five million cases of cholera per year around the globe, with the greatest burden in regions where it is endemic.

Reporting in *Nature Genetics*, researchers at the Sanger Institute and collaborators in Bangladesh studied the *Vibrio cholerae* bacterium in Dhaka, where the disease is hyper-endemic. The city experiences two seasonal outbreaks of cholera every year.

Between 2002 and 2005, samples were taken from cholera patients admitted to the Dhaka Hospital of icddr,b in Bangladesh. Over a surveillance period of three weeks, samples were also collected from other members of each patients' household. In total, 303 *V. cholerae* samples were collected from 224 individuals in 103 households.

The team sequenced the genomes of the *V. cholerae* samples. Each sample was mapped to a *V. cholerae* reference genome to determine their genomic similarities. DNA analysis showed that six sublineages of the seventh pandemic *V. cholerae* El Tor (7PET) strain were circulating concurrently in Dhaka during the study. Each sublineage waxed and waned with a different pattern in that time. Non-pandemic strains of the bacteria were also present.

Some people infected were asymptomatic – suggesting an important role for immunity in the spread of the disease. Many people were infected with more than one strain simultaneously.

To determine the diversity of *V. cholerae* within a household, the team linked household contact information to the bacteria's genetic makeup. The data showed a high degree of genetic relatedness between *V. cholerae* isolated from members of the same household. This indicates that infections were due to within-household transmission, or exposure to a common source.

To understand the spread of cholera more widely, the isolates from Dhaka were analysed in the context of a global collection of 513 additional 7PET genomes collected between 1957 and 2014. The results highlight a complex history of cholera in South Asia, and suggests that a connected, regional network of cholera transmission exists.

The work highlights the importance of relevant control strategies. Preventing the spread of cholera within the household could enormously reduce cholera outbreaks. This could have an impact, not only on the individual households, but also on the entire region.

Reference
Domman D *et al.* Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nature Genetics* 2018; **50**: 951–955.

3-5 million
cases of cholera worldwide

The disease has caused over

2,500

deaths in Yemen

5

Cholera in a conflict zone

As well as a brutal civil war, Yemen has faced the worst epidemic of cholera since records began.

The disease has affected over one million people and caused almost 2,500 deaths. In 2018, the United Nations estimated that 16 million of the 29 million people in Yemen lacked access to safe water and basic sanitation.

Researchers from the Sanger Institute, Institut Pasteur and Médecins Sans Frontières worked together to characterise the cholera outbreak during a peak in 2017. Reporting in *Nature* they uncovered that the bacteria was likely introduced into Yemen with the movement of people from East Africa.

The team sequenced the genomes of 42 *Vibrio cholerae* bacterium samples from Yemen, together with 74 cholera samples from South Asia, the Middle East and Eastern and Central Africa. They compared the genomes to a global collection of over 1,000 cholera samples from the current and ongoing pandemic – which is caused by a single lineage of *V. cholerae*, called 7PET.

Previous theories suggested that the two outbreaks of cholera in Yemen in 2016 and 2017 were caused by two different strains of the bacteria – but this study revealed they were caused by the same *V. cholerae* strain. The team also uncovered that, unusually, the Yemeni cholera strain was susceptible to several antibiotics.

They hope their work will enable better understanding of how cholera spreads, as well as inform outbreak prevention.

Reference
Weill FX, *et al.* Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* 2019; **565**: 230–233.

6

How malaria jumped to humans

The *Plasmodium falciparum* parasite causes the most deadly form of malaria in humans, and is responsible for 90 per cent of malaria deaths worldwide.

It belongs to the *Laverania* subgenus of parasites which infects great apes – including humans, chimpanzees and gorillas. *P. falciparum* is the only species of the subgenus to infect humans.

Scientists from the Sanger Institute and their collaborators from the French National Center for Scientific Research (CNRS), French National Research Institute for Sustainable Development (IRD), and the International Centre for Medical Research of Franceville, Gabon, have studied the genomes of *P. falciparum*'s relatives to uncover its evolutionary history – including its jump from infecting gorillas to infecting humans.

The team sequenced the genomes of all known *Laverania* species. Previously, only two other genome sequences were available for these species. A key challenge was obtaining the parasites from great apes. The researchers used blood samples taken as part of routine health checks of chimpanzees and gorillas living in sanctuaries in Gabon. The samples were tiny, containing very few parasites, so they devised strategies to amplify the DNA and obtain good quality genome sequences (see below).

How to get high-quality genomes from miniscule samples

- Deplete the host's DNA from the sample
- Apply parasite cell sorting
- Amplify parasite DNA templates
- Use long-read sequencing (single-molecule real-time technology)
- Use short-read sequencing

Reporting in the journal *Nature Microbiology*, they compared the genome sequences of the *Laverania* species, building a picture of their relatedness, and key events in their evolution. They estimate that *P. falciparum* first emerged 50,000 years ago, and fully diverged as a human-specific parasite 3,000–4,000 years ago. Contrary to previous theories, they estimate that the parasite jumped from gorilla to human more than once during this time.

The team also sought to uncover the genomic changes that allowed *P. falciparum* to jump between species. They detail changes in gene copy number and structural variations in gene families. In particular, they discovered the movement of one cluster of genes that appears to be a significant event that enabled the parasite to infect human red blood cells. Their work forms the basis for new investigations into how malaria parasites adapt to infect humans.

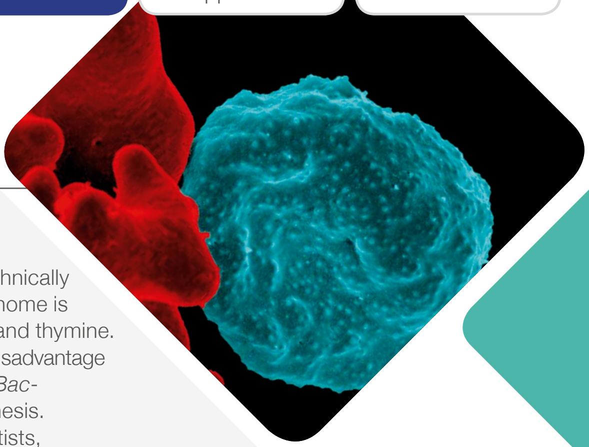
Reference

Otto TD *et al.* Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nature Microbiology* 2018; **3**: 687–697.

***Plasmodium falciparum* became fully human-specific approximately**

3,000–4,000

years ago



7

Malaria screen finds genes essential for life

Malaria is caused by Plasmodium parasites – the most deadly species being *Plasmodium falciparum*. *P. falciparum* is responsible for half of the estimated 200 million malaria infections worldwide.

To understand the parasite better, as well as how it might be tackled, researchers from the University of South Florida (USF) and the Sanger Institute have uncovered which of its genes are essential for survival in humans. Such genes might prove to be important targets for treatments.

The single-celled parasite is technically difficult to study because its genome is rich in the DNA bases adenine and thymine. However, the teams turned this disadvantage into a positive by utilising *piggyBac*-transposon insertional mutagenesis. Developed by the Sanger scientists, this technique disrupts and inactivates genes at random.

The researchers used it to create 38,000 *P. falciparum* mutants that knocked out the action of 87 per cent of the parasite's genes. Each genome was then sequenced with quantitative insertion site sequencing to pinpoint the mutations. Those genes without mutations were considered essential for survival.

Reporting their findings in *Science*, they identified 2,680 essential genes, and ranked their importance. Genes involved in translation, RNA metabolism, and cell cycle control were found to be the most important. 1,000 of the essential genes are conserved across all *Plasmodium* species and have completely unknown functions, suggesting a great deal of important malaria biology is yet to be uncovered.

Genes that are current candidates for drug targets were identified as essential, including genes involved in the proteasome degradation pathway. This is the likely target of the current first choice drug for treating malaria – artemisinin. In contrast, many of the genes that are targets of current vaccines were not shown to be essential. The findings will help guide researchers as they design new treatments.

Reference
Zhang M *et al.* Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. *Science* 2018; **360**: eaap7847.

8

Flipping malaria's sex master switch

The single-celled *Plasmodium* parasites that cause malaria have complex lifecycles.

Researchers at the University of Glasgow and the Sanger Institute previously identified a genetic 'master-switch', known as AP2-G, involved in a key stage of parasite development. They have now created a system to study and manipulate this master switch in the lab – confirming its importance, and opening up new avenues for research into preventing the spread of malaria.

Plasmodium parasites invade their host's red blood cells, and then follow one of two developmental pathways. Some replicate asexually to sustain the infection – causing the symptoms of malaria. Others convert into male and female gametocytes, the sexual stage of development. These gametocytes are taken up by mosquitoes and fuse together to form new parasites that then infect the next person. How gametocytes are made is not well understood, partly because it is difficult to generate high numbers of male and female parasites in laboratory conditions.

AP2-G is necessary for gametocyte production in several *Plasmodium* species – acting as a master switch to trigger gametocyte formation. Reporting in *Nature Microbiology*, the researchers developed a system to increase AP2-G activity in the rodent-infecting malaria parasite, *Plasmodium berghei*. They show that when AP2-G is amplified, it causes the majority of parasites to convert to gametocytes.

This discovery allowed the team to chart the changes in gene activity through gametocyte development. They showed that gender-specific changes occurred within six hours of induction. They also identified several potential targets in AP2-G. If these functions were disrupted by drugs, this could stop the parasite reproducing.

Reference
Kent RS *et al.* Inducible developmental reprogramming redefines commitment to sexual development in the malaria parasite *Plasmodium berghei. Nature Microbiology* 2018; **3**: 1206–1213.

9

Humans and cows – it's complicated

Parasitic bacterial species, like *Staphylococcus aureus*, that can jump between livestock and humans pose a threat to public health and food security.

Usually *S. aureus* is harmless to humans, residing in the nose. But some strains, like MRSA, have evolved resistance to antibiotics, and can prove deadly. The bacteria is also a major burden for the agricultural industry – causing a range of diseases from mastitis to septicaemia in livestock.

Researchers from the University of Edinburgh's Roslin Institute, together with collaborators at the Sanger Institute, have used genome sequencing to understand how this parasitic bacteria can jump from one host species to another. Their findings, published in *Nature Ecology & Evolution*, also show the evolutionary processes that give rise to new, disease-causing strains.

The teams sequenced 800 strains of *S. aureus* from 43 different host species, isolated in 50 countries. They used the data to work out the evolutionary relationships between the species, pinpointing key moments in the bacteria's evolution.

The researchers show that humans were the original hosts of the bacteria. It has made 14 jumps from humans to cows following the domestication of the animals. They estimate 10 strains have jumped back from cows to humans over the last 5,000 – 6,000 years, causing infections in populations around the world. While other animals did swap strains, humans and cows are the main reservoirs.

Each time the bacteria jumps between species, it acquires new genes enabling survival in its new host. The genes are acquired by horizontal (bacteria-to-bacteria) gene transfer, possibly from other species present in the host gut. The team also show evidence of adaptive evolution of the strains to their host environments.

In some cases, the genes acquired during these processes can also confer resistance to commonly used antibiotics. The team found that the antibiotic resistance of human and animal strains of the bacteria varies – reflecting the different way antibiotics are used in medicine and agriculture.

However, the relationship between antibiotic resistance of bacteria in livestock, and those in humans, is complex. To explore this further, the Sanger Institute team undertook a separate study of *Escherichia coli*, in conjunction with researchers at the London School of Hygiene and Tropical Medicine and Addenbrooke's Hospital in Cambridge.

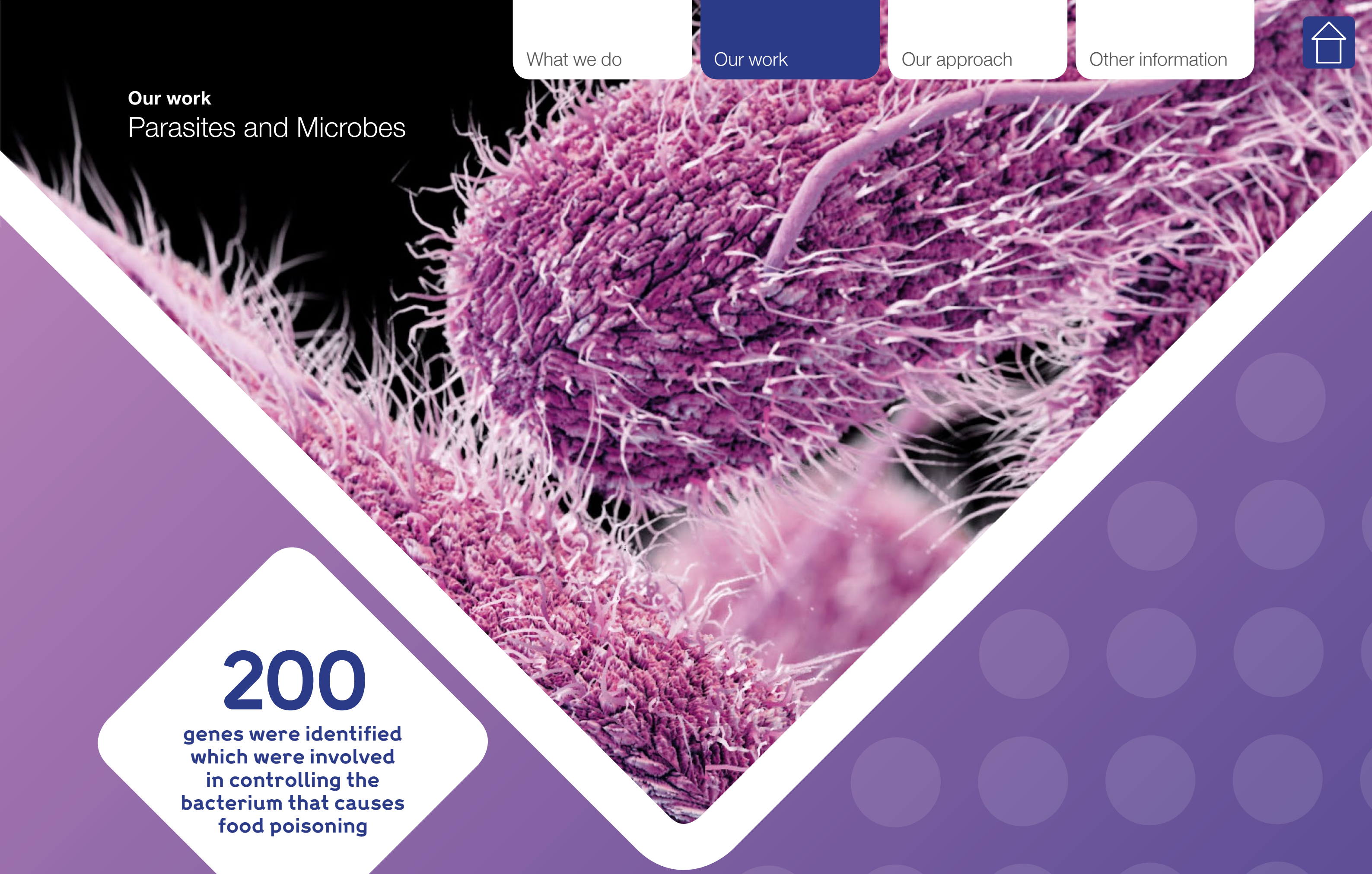
They studied the genomes of 431 *E. coli* samples from livestock and 1,517 from people. All the samples were from the same location: the East of England. They found that antibiotic resistance genes were not shared between the two groups. The results, published in *mBio*, suggests that, at least in the populations sampled, it is unlikely that antibiotic resistance in *E. coli* bacteria is transferred from livestock to people.

References

Richardson EJ *et al*. Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nature Ecology & Evolution* 2018; **2**: 1468–1478.

Ludden C *et al.* One health genomic surveillance of *Escherichia coli* demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. *mBio* 2019; **10**: e02693–18.

> "This study shows how human bacteria can jump to animals and back. This information could help with designing more effective ways to prevent bacterial transfer in farms and better antibiotic practices."
>
> **Professor Julian Parkhill,**
> Wellcome Sanger Institute



200
genes were identified which were involved in controlling the bacterium that causes food poisoning

10

AI identifies emerging pathogens

New, disease-causing bacteria continuously emerge as strains evolve: some pose a major threat, while others may only cause mild symptoms.

Through the widespread use of genome sequencing, researchers are tracking the spread of bacteria around the globe to inform public health interventions.

This work is starting to form the cornerstone of future healthcare initiatives. However, the approach can only monitor the rise of drug resistance or greater virulence in the bacteria, it cannot currently predict when a new epidemic may arise. In order to better understand how bacteria become dangerous, Sanger Institute researchers and colleagues have developed a powerful new machine learning tool.

Reporting in *PLOS Genetics*, the scientists developed and trained a machine learning program using the genome sequences of *Salmonella enterica* strains. They chose sample *Salmonella* strains which live in the gut and cause mild food poisoning, as well as those which escape the gut and cause serious infections, such as typhoid. The novel approach enabled them to identify almost 200 genes involved in controlling whether the bacterium will cause food poisoning or an invasive infection. They identified patterns of genetic changes which were associated with a bacteria evolving to become more dangerous.

The tool was then tested on *Salmonella* strains currently emerging in sub-Saharan Africa. It correctly identified two dangerous strains within a group of commonly circulating infections. The tool also revealed the precise genetic changes that enabled the *Salmonella* strains to adapt to their hosts and become more invasive – providing important information about the bacteria's functions and biology. This information could help researchers design more effective treatments in the future.

The machine learning tool, which is freely available for others, is not limited to *Salmonella*. It could be used to study emerging antibiotic resistance in any bacterium. It could be used in real time, to identify a dangerous strain of bacteria before it spreads and causes an outbreak.

Reference
Wheeler E *et al.* Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLOS Genetics* 2018; **14:** e1007333.

Tree of Life

In this section

1 25 Genomes for 25 Years
2 Biology's moonshot: the Earth BioGenome Project
3 Darwin Tree of Life Project will create UK's genomics 'Domesday Book'
4 Project produces platinum-quality genomes
5 New approach opens up the insect world
6 Bird genomes aid conservation efforts

1

25 Genomes for 25 Years

In 2018, the Sanger Institute turned 25 years old. As part of its anniversary, it chose a uniquely Sanger-inspired way to celebrate.

The Institute sequenced 25 UK species for the first time to provide a genomics resource for the life science research community.

The Institute was founded to sequence the human genome. The international Human Genome Project took 13 years to complete, and the Institute became the largest single contributor. Since then, its scientists have been involved in sequencing important animal species, including the gorilla, mouse, pig and zebrafish. Thousands of microbes, parasites and disease-causing bacteria have also had their genomes decoded by Institute teams.

The 25 new species represent wildlife in the UK, from well-loved iconic species to invasive plants which are threatening biodiversity. In total, 20 of the species were chosen by Institute researchers in collaboration with experts in conservation, natural history and ecology. The final five species were chosen by public vote as part of 'I'm a Scientist Get Me Out of Here'.

Reading the genomes of such a diverse range of species brought interesting and surprising challenges. Some species were difficult to identify. Others have complex cellular structures which made extracting DNA for sequencing a particularly thorny issue.

The sequencing was completed in a little under 10 months, but the work to extract the maximum value from the data continues. The sequence and associated data is being made freely available for researchers worldwide via European Bioinformatics Institute (EMBL-EBI) repositories.

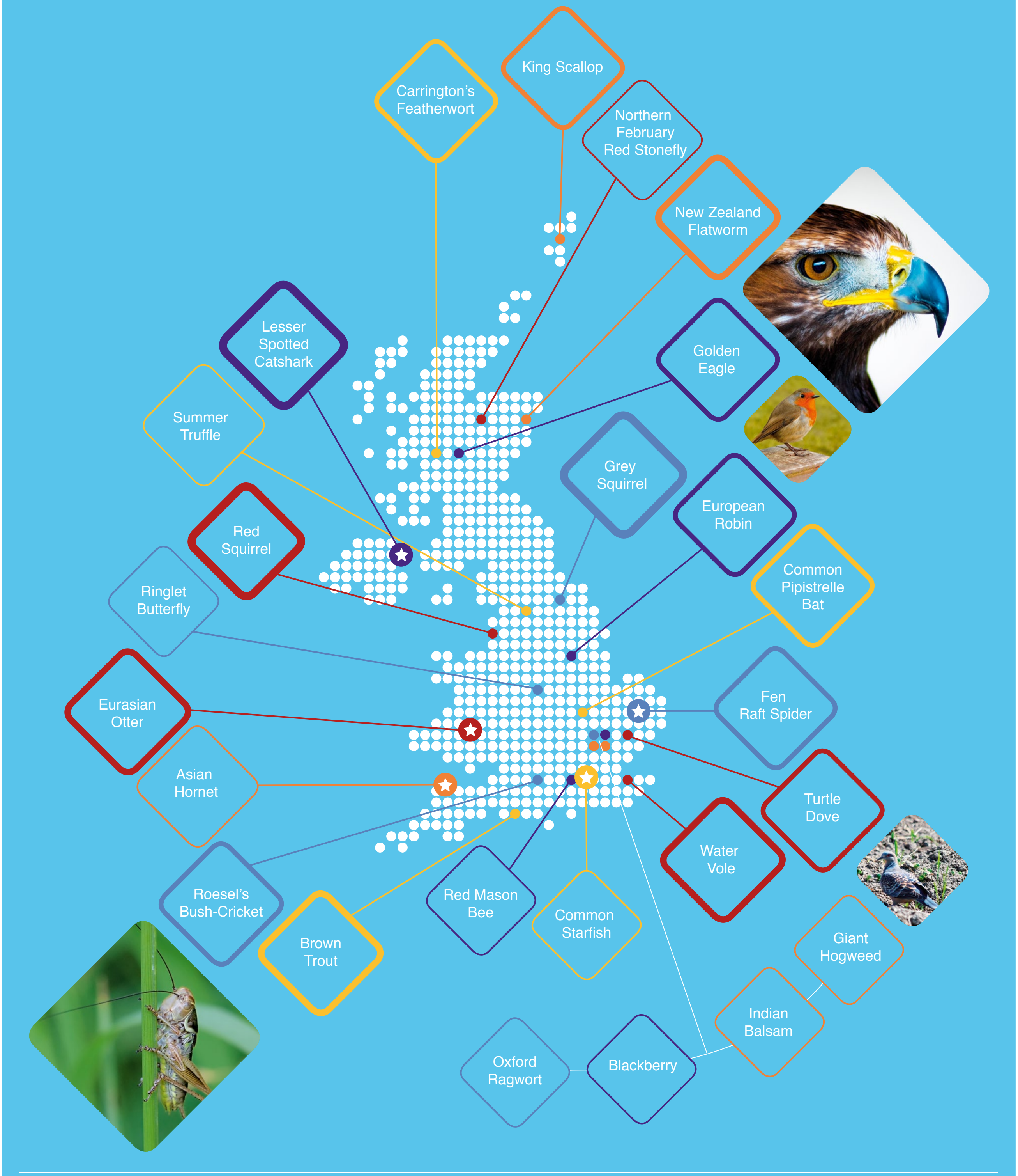
There have already been intriguing insights from the sequences; king scallops are more genetically diverse than humans, and Roesel's bush-crickets have genomes that are four times the size of the human genome.

Many of the species have intriguing adaptations and behaviours that could be understood by future research based on their genome sequences. The bat genome might shed light on why they are resistant to so many diseases, including cancer. While studying the genome of the common starfish could help with research into limb regeneration.

The sequencing of all 25 species was completed in a little under

10 months

25 Genomes for 25 Years of the Sanger Institute



Genome size

◇ <1 GB
◇ 1–1.99 GB
◇ 2–2.99 GB
◇ 3–3.99 GB
◇ >4 GB

Categories

- ● Flourishing – species on the up in the UK
- ● Floundering – endangered and declining species
- ● Dangerous – invasive and harmful species
- ● Iconic – quintessentially British species that we all recognise
- ● Cryptic – species that are out of sight or indistinguishable from others based on looks alone
- ✪ Chosen by the public

2

Biology's moonshot: the Earth BioGenome Project

The Sanger Institute is leading the UK contribution to biology's most ambitious project yet.

The Earth BioGenome Project is a global collaboration of more than 17 institutions with the extraordinary aim of sequencing the genomes of all complex life on the planet: 1.5 million known species of eukaryotes (animals, plants, fungi and protozoa). The work will create a genomic foundation for biology that could help preserve biodiversity and the sustainability of human societies.

Currently, fewer than 3,500, or about 0.2 per cent of all known eukaryotic species, have had their genome sequenced. Just 100 of those are high-quality reference sequences. Described as biology's moonshot, the new sequences will revolutionise the understanding of life on earth. Genome sequences will provide insight into evolution and conservation, and provide new resources for researchers in agricultural and medical fields.

This vast project is expected to take just 10 years to complete and an estimated $4.7 billion which, when accounting for inflation, is less than the cost of the Human Genome Project.

The amount of biological data that will be produced and collected by the project is expected to be on the exascale; more than that accumulated by the current biggest data producers in the world; Twitter, YouTube and astronomy. The data will be stored in public-domain databases and access will be open to all researchers.

When the Human Genome Project began 25 years ago, it was not possible to imagine how it would transform research into human health and disease. The same is true for sequencing all life on earth. The discoveries it could unlock may help to develop new treatments for infectious diseases, help create new bio materials, or generate new approaches to feeding the world.

The vast project is expected to take **10 years** to complete

EARTH BIOGENOME PROJECT
sequencing life for the future of life

The project plans to sequence the genomes of **1.5 million** known species on the planet

> "Globally, more than half of the vertebrate population has been lost in the past 40 years. Using the biological insights we will get from the genomes of all eukaryotic species, we can look to our responsibilities as custodians of life on this planet."
>
> **Professor Sir Mike Stratton,** Director of the Wellcome Sanger Institute

The Sanger Institute is leading the DNA sequencing of all

66,000

eukaryote species in the UK

It is estimated that the Institute's contribution to the project could cost approximately

£100 million

3

Darwin Tree of Life Project will create UK's genomics 'Domesday Book'

As part of the Earth BioGenome Project to sequence all complex life on earth, the Sanger Institute is leading the DNA sequencing of all 66,000 eukaryote species in the UK.

The initiative, named the Darwin Tree of Life Project, brings together the Sanger Institute; Natural History Museum; Royal Botanic Gardens, Kew; Earlham Institute; European Bioinformatics Institute (EMBL-EBI); Edinburgh Genomics, and many others.

To prevent duplication of effort, the Sanger Institute will coordinate the Darwin Project and will liaise with the other initiatives and consortia that are also contributing to the Earth BioGenome Project, including the G10K Vertebrate Genomes Project and the 10,000 Genomes Plant Project.

This vast project is only possible because of recent advances in sequencing technology, not least in terms of speed. Advances in information technology and the analysis of biological data are also important factors, and it is expected that this project will accelerate further progress in both of these areas.

It is estimated that the Institute's contribution to the project could cost approximately £100 million over the first five years, with the entire Darwin Tree of Life Project expected to run for 10 years.

To derive the maximum scientific benefit from the project, the Institute is establishing a new research programme, The Tree of Life. Organised in the same way as the Institute's other research programmes, it will recruit a cohort of Faculty to design and deliver a range of laboratory and computational experiments using the project's data and techniques. Their work will unlock insights into areas such as population genomics, evolution and the emergence of species, and why some species develop certain genetic conditions when others do not.

23,000

species of vertebrate face the threat of extinction

4

Project produces platinum-quality genomes

In September 2018, the Vertebrate Genomes Project, which aims to provide 'platinum-quality' (near error-free and complete) genome sequences of all 66,000 vertebrate species on earth, completed its first 15 reference genomes.

Sanger Institute teams are playing a key role in the project by supplying DNA sequencing and computational analysis to the work.

The international team is part of the Genomes 10K consortium. The 15 new, high-quality reference genomes represent species from all five vertebrate classes; mammals, birds, reptiles, amphibians and fish. All the data is freely available in the Genome Ark database, the project's open-access library of genomes.

The researchers hope that their work will highlight the current threats to survival that so many species face. More than half of the world's vertebrate population has been lost in the past 40 years, and 23,000 species face the threat of extinction. Among them is the kakapo, whose genome was read by the project; there are less than 150 left alive.

A key part of the consortium's work has been to develop protocols and pipelines that can produce 'platinum-quality' genome sequences. The group drew together more than 150 experts from academia, industry, and government, representing over 50 institutions in 12 countries, to compare the major sequencing and analysis technologies.

The experts concluded that single DNA molecule long-read technologies always give the highest quality results and should be coupled with technologies that measure long-range genome interactions to assemble the DNA reads into whole chromosomes. The consortium has found that the common practice of merging an individual's paternal and maternal chromosomes into one genome was causing numerous errors. To overcome this, they now assemble paternal and maternal DNA separately.

5

New approach opens up the insect world

Sanger scientists, collaborating with Pacific Biosciences (PacBio), have developed new laboratory preparation protocols that allow researchers to accurately read the whole genomes of individual mosquitoes.

The technique not only allows fine-grained exploration of mosquito populations, it also enables biologists to obtain *de novo* reference genomes for individual insects and other small species.

Reporting their work in *Genes*, the team detailed how they generated a whole genome sequence from just 100 nanograms of a single mosquito's DNA – the equivalent of half a mosquito. This is a full order of magnitude less than the amount previously needed.

When sequencing a species for the first time, longer reads of DNA are preferable. PacBio's Single Molecule, Real-Time sequencing method produces long stretches of DNA with high accuracy, but it requires approximately 5 micrograms of DNA. For small species this means that researchers have to pool DNA from many organisms. The resulting reads can be difficult to align, increasing the chance of errors and gaps.

By removing two steps from the DNA preparation process – DNA shearing and filtering – the team produced the whole genome sequence of a single *Anopheles coluzzii* mosquito at high quality and with few gaps. As a result, nearly half of the species' previously unplaced DNA fragments have been assigned to their correct chromosomal position.

This advance is an important boost for the Earth BioGenome Project, as it improves the quality of DNA sequencing. It will also allow researchers to use specimens from museums or CryoArks, where using the least amount of material possible is vital. And it will enable researchers to explore full genetic diversity of insects and other arthopods – the most diverse animal group in the Tree of Life – by sequencing individual members of each species.

Reference
Kingan SB *et al.* A high-quality *de novo* genome assembly from a single mosquito using PacBio sequencing. *Genes* 2019; **10**: 62.

> With the golden eagle genome sequence, we will be able to compare the eagles being relocated to southern Scotland to those already in the area to ensure we are creating a genetically diverse population.”
>
> **Dr Rob Ogden,**
> Head of Conservation Genetics
> at the University of Edinburgh
> and a scientific adviser to
> the South of Scotland
> Golden Eagle Project

6

Bird genomes aid conservation efforts

The golden eagle, swiftly followed by the robin and turtle dove, were the first species to have their genomes sequenced as part of the Sanger Institute's 25 Genomes project.

The three birds had not previously been sequenced, and so the resource will be a huge boost to those studying these iconic species.

The golden eagle sequence was completed together with researchers at the University of Edinburgh. It will aid monitoring of existing, reinforced and reintroduced populations of golden eagles, such as those in the South of Scotland Golden Eagle translocation project. Conservation scientists will be able to compare the eagles being relocated to southern Scotland to those already in the area to ensure that a genetically diverse population is being created.

The European robin and turtle dove were sequenced in partnership with the University of Lincoln. The turtle dove is one of the UK's fastest declining bird species – since 1995, 94 per cent of turtle doves have been lost and there are fewer than 5,000 breeding pairs left in the UK. Its genome sequence will help researchers understand the pressures that are affecting the birds – be that disease or a decline in food. It will also aid practical conservation efforts, by enabling conservationists to maximise the genetic diversity of introduced populations.

The European robin genome may harbour the secrets of bird migration. The species live across large swathes of northern Europe and Siberia. Some overwinter in the UK, escaping colder Scandinavian weather, while others live in the UK during the summer, migrating to southern Europe in the winter.

Birds can use the earth's magnetic field for orientation during migratory journeys, and the magnetic compass in birds was first described in a robin. The European robin genome will allow researchers to identify what's driving migration in birds, and understand the variability of migration both within a species and more widely.

Our approach

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

42 Scale

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

44 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.

46 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.

48 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.

50 Connections

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

arm
AtoS
CAVIUM
How can we use mobile technology to transform biological data processing?
#BioDataHack

Read more about our superfast and super-accurate sequencing
Page 43

Discover how we hacked our genomic future
Page 51

Find out about the John Sulston's legacy
Page 47

Scale

In this section

1 Flexible cloud computing wins peer award

2 Institute establishes cancer organoid production line

3 Superfast, super-accurate sequencing

1

Flexible cloud computing wins peer award

The Sanger Institute's High Performance Computing team has been praised by the global IT industry for the way it has pioneered flexible, scalable and secure computing to analyse thousands of human genomes.

The team won *HPCWire*'s 2018 Readers' Choice Award for Best Use of High Performance Computing in the Cloud. The worldwide community cited the creative way that the Institute combines cloud computing with its onsite data centre to store, analyse and share genomic data.

Genomic data, especially when coupled with clinical data, poses many challenges – see box to the right. These concerns can often rule out public cloud storage, yet standard onsite data centre solutions are increasingly unable to cope with the volume and complexity of data being generated. The answer was to put the cloud into the data centre.

To 'put the cloud in a box' the team has deployed a private OpenStack Cloud IT environment within its data centre. This allows scientists to run their own computational experiments, by creating an appropriately sized 'virtual computer' that uses the speed and cost-efficiency of the onsite hardware to deliver its results, while keeping all sensitive data behind formidable firewalls.

Genomic and clinical data challenges

Security
The sensitive and personal nature of human genome and clinical data demands the highest levels of protection

Jurisdiction
Such data must be stored in locations that comply with the EU's General Data Protection Regulation (GDPR) and UK privacy laws

Volume
The scale of the data is vast, porting the Institute's current genomic data to the external cloud would take years

Cost
Storing the Institute's tens of petabytes of data in the cloud would be prohibitively expensive

This Flexible Compute Environment not only provides rapid, reliable results for Sanger scientists, but for worldwide collaborators too. External scientists can upload the software and data needed to run their analysis, which then uses the data centre's vast compute to query its enormous data sets. The results, shorn of any identifying data, are then sent back.

In this way, only small amounts of non-sensitive data pass over the internet, sparing the collaborator's internet bandwidth and computing facility, while also ensuring the highest security.

> "Winning the Readers' Choice Award is especially humbling as it means that our peers worldwide acknowledge the groundbreaking work of our scientific computing and informatics teams."
>
> **Tim Cutts,**
> Head of Scientific Computing
> at the Wellcome
> Sanger Institute

2

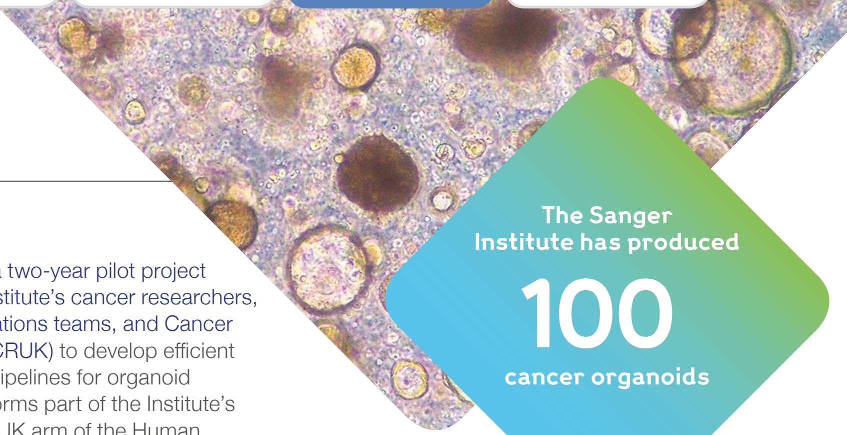
Institute establishes cancer organoid production line

The Sanger Institute has produced 100 cancer organoids – living 3D models that more accurately represent the human tissues tumours grow in – to provide the next generation of cancer models for researchers worldwide.

These 100 cover pancreatic, oesophageal and colon tumours at different stages of development – they were chosen because these cancers currently have few treatment options and poor survival rates.

The work was a two-year pilot project between the Institute's cancer researchers, Scientific Operations teams, and Cancer Research UK (CRUK) to develop efficient networks and pipelines for organoid production. It forms part of the Institute's lead role in the UK arm of the Human Cancer Models Initiative that aims to generate approximately 1,000 new models to empower the study of cancer initiation, development, and progression.

CRUK provided access to its unique clinical network to provide the tumour samples at the time of surgical removal, along with their related clinical data. Working with surgeons and pathologists across the UK, Sanger Institute scientists established an efficient pipeline that covers ethical agreements, transport infrastructure and robust laboratory pipelines to start culturing the tissues at the Institute within 36 hours of surgery.



The Sanger Institute has produced

100

cancer organoids

Sanger's researchers have sequenced the organoids' genomes to create profiles of their mutations, combined this with their medical information, and then screened them against hundreds of drugs to identify their susceptibilities. The models are now available from the ATCC biorepository, and their data is stored on the Cell Model Passports website.

3

Superfast, super-accurate sequencing

When the Sanger Institute was chosen to deliver the UK Biobank Vanguard project to sequence the genomes of 50,000 volunteers, the scale of the task was substantial.

To read each genome to gold-standard level requires each of its three billion bases to be read 30 times over. Therefore approximately 4,500,000,000,000,000 bases of DNA needs to be sequenced – only 10 per cent less than the entire amount of DNA the Institute had read in its first 24 years.

To achieve this goal, the team has dedicated eight of their fleet of 10 NovaSeq sequencing machines to the work. Originally each machine was expected to read approximately 48 gold-standard human genomes per run, at three runs per week. But, through a number of DNA preparation and wash cycle optimisations introduced by the Sequencing R&D team, the amount of high-quality human genomes produced per run is expected to rise even further.

The Institute is delivering approximately 1,200 human genomes to the project each week, and at higher quality and coverage than required. This is the equivalent of one gold-standard human genome every 8.5 minutes, and is expected to become even quicker.

But the benefits are not just for UK Biobank-based research: Sanger researchers can now call upon this same pipeline. By the end of 2019, the majority of the Institute's high-throughput DNA sequencing pipelines, exome sequencing, RNA pipelines and sequencing from the 10X Genomics platform will be carried out solely on its other two NovaSeq machines.

UK Biobank Vanguard project to sequence the genomes of

50,000

volunteers

Innovation

In this section

1 Microbiotica – getting to the guts of disease

2 Spinning out success

3 Sanger-developed technology can stop outbreaks

1

Microbiotica – getting to the guts of disease

The Sanger Institute's microbiome-based spin-out company, Microbiotica, has had a stellar year.

It specialises in analysing the gut flora of patients, using state-of-the-art culturing, sequencing and informatic techniques developed at the Institute. By cataloguing the human gut microbiome at an industrial scale, it has built up an extensive culture collection of human gut bacteria and reference genome database of gut bacterial species.

The company uses these tools to identify the makeup of the bacterial community within individual patients. By applying artificial intelligence to the results, it is discovering microbiome 'signatures' that can be linked to specific diseases and conditions. These are then validated in mice with a humanised microbiome.

Microbiotica's platform has attracted the attention of biopharma world leader Genentech, a member of the Roche Group. The two companies have signed a strategic partnership to discover biomarkers and treatments for inflammatory bowel disease, which could see Microbiotica receive up to $534 million in upfront and milestone payments. In addition, the company is using Genentech's patient samples to further expand its culture collection and reference genome database.

Microbiotica is also collaborating with the University of Adelaide to analyse samples that have successfully treated ulcerative colitis. The aim is to identify the key bacteria involved to develop a bacterial product that will reset the gut's microbiome.

Based on these developments, the company successfully completed its second round of funding, raising a further £4 million from Seventure Partners.

Microbiotica has signed a deal worth up to

$534 million

with Genentech

2

Congenica helped to successfully deliver the UK's

100,000

Genomes Project

Spinning out success

Two established Sanger Institute spin-out companies are making headlines worldwide as they use genomic technologies to advance healthcare.

Genome analysis company, Congenica, is leading the way in personalised medicine. In 2018, it became a key partner to deliver the world's first routine national genomic medicine service for the UK's National Health Service (NHS). Founded on the Institute's Deciphering Developmental Disorders collaboration with the NHS to provide genomics-based diagnoses, the company helped to successfully deliver the UK's 100,000 Genomes Project.

In addition, Congenica is supporting China's 100K Wellness Pioneer Project and is deepening its relations with the country by partnering with Digital China Health Technologies Cooperation Limited to develop a Chinese version of its Sapientia™ decision support platform.

In 2018, therapeutics company Kymab was tipped by US finance and investment magazine Forbes to become a $1 billion valued company. There are only 300 or so 'Unicorns' worldwide.

The company generates antibody-based treatments and vaccines from mice with a humanised immune system whose origins lie in the Sanger Institute. Two of its fully human monoclonal antibodies are undergoing clinical trials.

The first, KY1005, targets a molecule at the heart of a number of T-cell mediated inflammatory diseases, including atopic dermatitis and rejection of transplanted organs. The second, KY1044, is being trialled in collaboration with the pharmaceutical company Roche, to treat solid tumours. It boosts the immune response by reducing the number of Regulatory T cells within tumours and stimulating T effector cells.

3

Sanger-developed technology can stop outbreaks

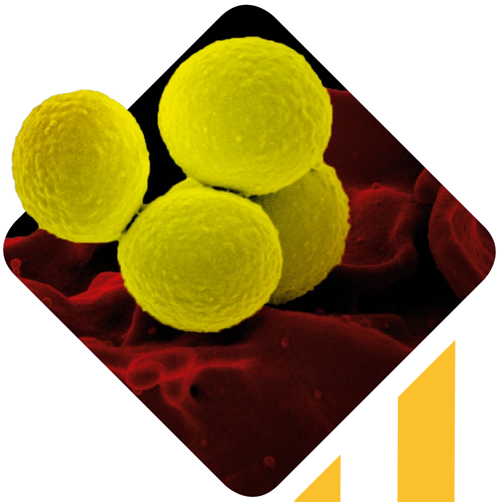
In 2012, Sanger Institute scientists applied genome sequencing to track an MRSA outbreak in a UK hospital's neonatal ward and their insights helped to break the transmission.

The study demonstrated that sequencing bacteria's whole genomes could produce real-time results with a clinical impact. Six years later, Sanger Institute technology has underpinned the formation of a new company based on this approach to deliver a fully automated service that can be deployed in clinical settings worldwide.

Next-Gen Diagnostics offers a comprehensive genomic analysis service to monitor, diagnose and stop infectious outbreaks. Where once diagnoses could take days, now whole-genome sequencing combined with a world-leading data analysis platform can reveal the bacteria's species, strain and relatedness in a matter of hours. In addition, by comparing different patient's bacterial genomes down to the individual DNA base pair, the service flags up emerging patterns of transmission, enabling infection control teams to act to stop the outbreak.

To deliver a high-volume, low-cost and real-time service Next-Gen Diagnostics combines on-site robotic sample preparation and sequencing, with reference genome libraries and artificial intelligence in a single package.

To demonstrate its capabilities, the company is conducting the first-ever trial of prospective whole-genome sequencing of MRSA samples as an infection control system. Funded by Wellcome and the UK Government's Health Innovation Challenge Fund, the study is being run in collaboration with the microbiology and infection control teams at Addenbrooke's Hospital, Cambridge. The technology is also being trialled at Mayo Clinic in the US.



Culture

In this section

1 Promoting equality, diversity and inclusion
2 Committed to our technicians
3 John Sulston's legacy – the Sanger Prize

1

Promoting equality, diversity and inclusion

UK science has a problem – it is biased. But the bias is not only in terms of who conducts the research, but also in terms of how it is conducted, on whom, and on what it is focused.

For example, only 2 per cent of UK professors are black women and female scientists are underrepresented at the highest positions in UK science. Equally, 80 per cent of individuals studied in genome-wide association studies are of Western European descent, while most biomedical animal-based studies have a large bias to male populations.

To resolve these issues, in January 2019, the Sanger Institute joined a new initiative that seeks to change the approach and design of science across academic research, funding and the commercial research sector.

The Equality, Diversity and Inclusion in Science and Health (EDIS) coalition is seeking to create a community of organisations to drive forward evidence-based policies to produce lasting change to the makeup of research teams, scientific focus and funding priorities. Its work will also promote greater diversity and inclusion in volunteer cohorts to ensure that life-changing medical and scientific research benefits all.

By becoming a member, the Institute will provide financial support, share best practice ideas, and help build the evidence base for future change strategies. In 2019, the Institute will attend an EDIS symposium to explore how research and experimental design can overcome biases in the sex of cells and animals studied; the ethnicity and ancestry of participants in genome sequencing projects; and the diversity of participants in clinical trials.

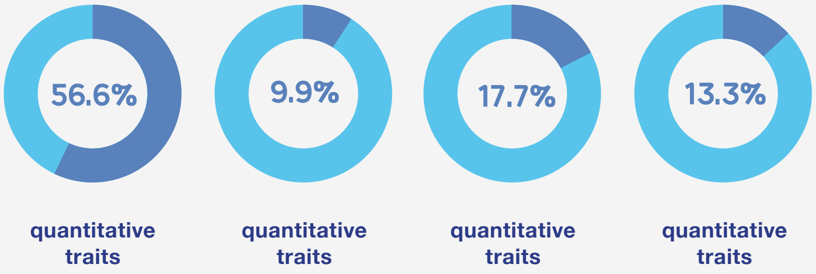
80%

of individuals studied in genome-wide association studies are of Western European descent

In 2017, scientists studied 234 traits in <50,000 mice, across 10 centres

In control mice, sex had an effect on:

In mutant mice, sex modified the effect on:



Reference
Karp NA *et al*. Prevalence of sexual diamorphism in mammalian phenotypes. *Nature Communications* 2017; **8**: 15475.

2

Committed to our technicians

The Sanger Institute's unique ability to conduct large-scale, high-throughput genomic experiments is founded on an army of technicians who support our Faculty research teams.

Their dedication and skill to deliver animal care, computational analysis, DNA sequencing, data storage, and wet-lab experiments powers all our research.

To ensure that our staff are rightly recognised and that the Institute is equipping, empowering and inspiring the next generation of technicians, we are delighted to have signed up to the Technician Commitment.

Supported by the Science Council and Gatsby Charitable Foundation's 'Technicians make it happen' campaign, the Commitment addresses key issues facing technicians working in higher education and research – see box.

Our two-year, technician-led action plan has been developed and will be regularly evaluated by a steering group of more than 25 technicians and managers from across all the Institute's research areas and disciplines. They will facilitate wider networking and support between technicians and create new processes to improve technicians' training, access to professional qualifications, and career pathways. In addition, they will create new ways to celebrate the vital role of technicians in delivering the Sanger Institute's science at scale.

Visibility
Ensure that all technicians are identifiable and that their contribution is visible within and beyond the organisation

Recognition
Ensure that technicians receive the acknowledgement and recognition they deserve at all levels of the organisation. Provide opportunities for professional registration

Career Development
Enable career progression opportunities for technicians through the provision of clear, documented career pathways

Sustainability
Ensure the future sustainability of technical skills across the organisation and that technical expertise is fully utilised

3

John Sulston's legacy – the Sanger Prize

In 2018 the founding director of the Sanger Institute, Sir John Sulston, died. But his legacy lives on in a scheme which reflects many of his qualities.

When Sir John won his Nobel Prize in 2002, he donated his prize money to found a charity that enables undergraduate students from low- and middle-income countries to come and work at the Institute. The annual Sanger Prize awards one student a three-month internship with one of our research groups to practise cutting-edge genomic research – all travel, living and research expenses paid. More than 450 students applied for the 2019 Award.

The first winner, back in 2005, was Anna Protasio, who came from Uruguay and chose to work on the *C. elegans* project. This sparked a research interest in parasite genomics that she pursued first at the University of Cambridge, and then for nine years under Matt Berriman at the Sanger Institute. Anna is now working at the University of Cambridge, but continues her association with the Institute as a lead content developer and educator for Advanced Courses and Scientific Conferences.

In the years since, winners have come from countries including China, India, Mexico and Uganda. Many have been inspired to pursue a career in genomics and are now based in institutes in France, the USA, India, Uganda and Estonia. They are applying genetics and genomics to antibiotic resistance in *Escherichia coli*, drug targets in parasitic flatworms, genome variation across and within species, and the role of epigenetics and gene regulation in neurodevelopment.

More than

450

students applied for the 2019 Award

Influence

In this section

1 Cambridge hub of HDR UK aims to transform diagnoses

2 MPs offer scientists top influencing tips

3 GA4GH Connect – making big data a reality

1

Cambridge hub of HDR UK aims to transform diagnoses

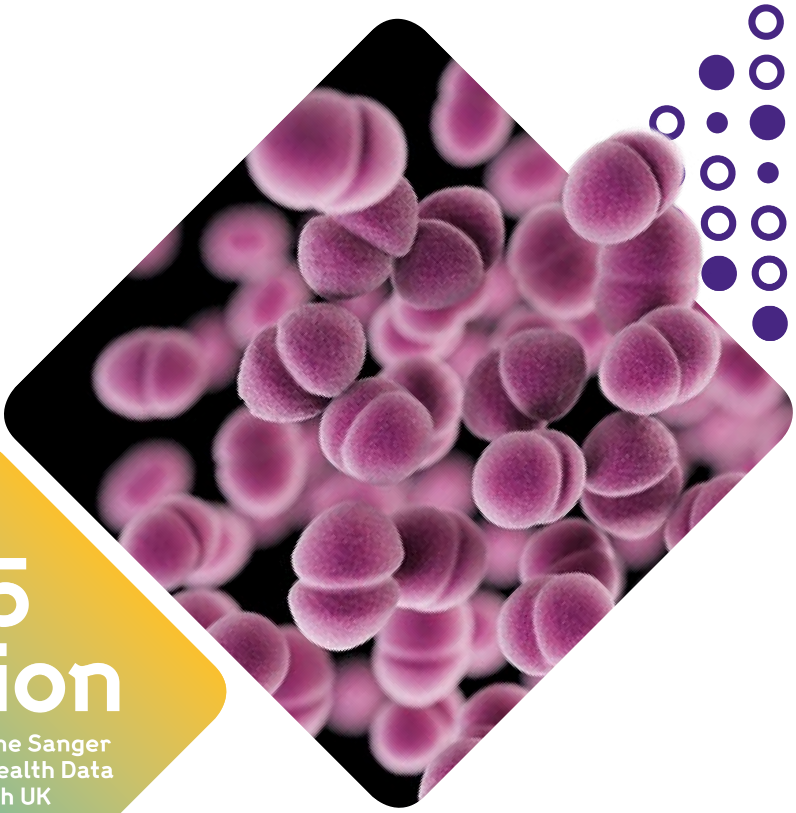
Health Data Research UK (HDR UK) has awarded £5 million to the Sanger Institute, European Bioinformatics Institute (EMBL-EBI) and the University of Cambridge and Cambridge University Hospitals to form HDR UK's Cambridge hub.

The hub is exploring how genomics and single-cell studies can be combined with NHS health data to reveal the molecular underpinnings of disease symptoms and prognoses. This work could shift the paradigm of medical diagnosis from observed physical characteristics to the underlying molecular processes.

The hub has four key areas of activity:

1. Apply genomics, genetics, single cell studies, medical imaging and electronic health records to correlate the symptoms of disparate diseases with shared cellular signalling pathways and molecular processes.
2. Create a knowledge base of the clinical course of rare and extreme diseases. By analysing the health records of individuals with genomic diagnoses made by NHS-based studies, including the Deciphering Developmental Disorders (DDD) project, the Prenatal Assessment of Genomes and Exomes (PAGE) and Genomics England, it may be possible to develop novel therapies and optimise decision making.
3. Monitor the carriage, transmission and evolution of infectious bacteria in the general population to understand how organisms that live harmlessly on large numbers of people can give rise to life-threatening outbreaks and drive antibiotic resistance.
4. Develop IT and data analysis infrastructure to allow seamless and secure gathering of sensitive information, computational exploration, and dissemination of results.

The hub's findings could drive new therapies, optimise treatment strategies, and lay the foundations of truly personalised medicine.



HDRUK
Health Data Research UK

£5 million

awarded to the Sanger Institute by Health Data Research UK

2

MPs offer scientists top influencing tips

Scientific research and knowledge has much to offer the UK's economy, but it should also help guide strategic policy making on issues that impact health and science.

The adage 'Information is dead unless it is read' is never truer than when applied to research findings and decision makers.

Locked away in research papers or conference reports, the vital knowledge MPs and policy makers need may never be seen. Even if the research paper is read, without context or guidance, the reader may not be able to extract its policy implications or recommendations. Equally, researchers can find the complex workings of the UK's Civil Service, Parliament and Government impenetrable.

To bridge this divide, the Sanger Institute's Policy team worked with Connecting Science's Advanced Courses and Scientific Conferences to run a two-day course that brought MPs and scientists together. The course was designed to guide scientists through the policymaking process to help them engage with decision makers.

Current MPs Heidi Allen and Vicky Ford were joined by minister Nicola Blackwood and policy professionals to highlight the most effective ways in which scientists can present their evidence.

A key lesson for the delegates was the need to target honed and brief messages to time-poor MPs. Another insight was the importance of timing and context: the decision makers highlighted the key stages in the legislative process where scientific evidence would have maximum impact.

The conference was a resounding success with many delegates wanting to build their engagement with policy makers.

3

GA4GH Connect – making big data a reality

Big data presents exciting new opportunities for genomic and medical researchers to unlock insights into health and disease.

But the hurdles to exploiting its power are significant: incompatibilities between different approaches to data collection, storage and analyses can make combining and comparing results challenging.

To address this issue, the Sanger Institute was a founding member of the Global Alliance for Global Health (GA4GH) in 2013 and, along with the Broad Institute and the Ontario Institute for Cancer Research, provide the Secretariat of GA4GH. The initiative seeks to enable effective and responsible sharing of health and genomic data by developing a common framework of standards and harmonised approaches.

Significant progress has been made in areas ranging from healthcare and research to patient advocacy and information technology, but the work is far from done. To deliver true interconnectivity, the alliance launched GA4GH Connect – with the aim of enabling responsible sharing of clinical-grade data by 2022.

The work is being carried out by eight alliance work streams, but many of their efforts would not be possible without the administrative and technical support provided by the Sanger Institute. Staff at the Institute help coordinate and guide the workstreams and others contribute to the development of technology standards, policies and guidance in regulation, ethics, and data security.

In this way, the Sanger Institute is playing a key role in shaping the future of genomic and medical research worldwide, in areas as diverse as genomic variation and annotation information, clinical and phenotypic data, ethical data gathering, and researcher identification.

Global Alliance for Genomics & Health

2,000+ subscribers

500+ organisational members

71 countries

Connections

In this section

1 Campus' unique collaboration gains new partners

2 Sanger science at the heart of research excellence

3 Hacking our genomic future

1

Campus' unique collaboration gains new partners

Open Targets is a pioneering pre-competitive public-private partnership between the Wellcome Genome Campus' academic institutes – the Sanger Institute and European Bioinformatics Institute (EMBL-EBI) – and the pharmaceutical industry.

Launched in 2014, the collaboration has five commercial partners: GSK, Biogen, Takeda, Celgene and Sanofi – with the latter two joining in 2018.

The initiative seeks to transform drug discovery by systematically identifying and prioritising drug targets. It combines collaborator's skills and technologies to create a critical mass of expertise that could not exist in one organisation. Large-scale genomic experiments and computational techniques from the public domain are blended with pharmaceutical R&D approaches to identify causal links between targets, pathways and diseases.

Each new partner shares new expertise with the collective. Celgene brings new approaches to the discovery and development of therapies in cancer, immune-inflammatory and other unmet medical needs. While Sanofi's involvement increases the initiative's expertise in immunology, oncology, neurosciences and diabetes.

The freely available Open Targets Platform contains over 28,000 targets, in excess of 3,000,000 associations, covering more than 10,000 diseases. And a new tool – Open Targets Genetics – developed by the Genetics Core Team from the Sanger Institute, allows researchers to explore variant-gene-trait associations from UK Biobank and the GWAS Catalogue.

The new genetics portal enables researchers to browse, visualise and interpret human genetics and genomics data to unravel the biology of human diseases, inform drug repurposing, and predict toxicity effects.

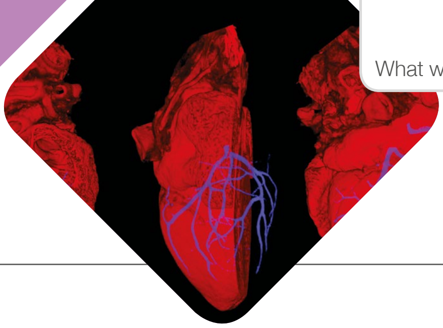
In addition, new Director Ian Dunham, an Honorary Faculty at the Sanger Institute, is seeking to increase the collaboration's use of single-cell sequencing, CRISPR and artificial intelligence.

> "We have built strong partnerships and established powerful research programmes that exploit advances in genetics and genomics to improve drug target identification."
>
> **Ian Dunham,** Director of Open Targets and Honorary Faculty at the Sanger Institute

The Open Targets Platform contains over

28,000

targets



2 Sanger science at the heart of research excellence

In 2019, the British Heart Foundation (BHF) doubled its funding for the Cambridge BHF Centre of Research Excellence to enable five more years of exploration into the underlying biology, diagnosis, prevention and treatment of cardiovascular disease.

The Sanger Institute has extensive involvement with the centre, which brings together teams from the University of Cambridge, the Sanger Institute, Babraham Institute, and MRC Mitochondrial Biology, Biostatistics and Epidemiology Units.

Three of the four key areas of research focus for the Centre are led by Sanger Institute-affiliated scientists:

Cardio-metabolic Medicine
– Associate Faculty Professor Antonio Vidal-Puig

Functional Genomics
– Faculty Professor Nicole Soranzo

Population Sciences
– Faculty Professor John Danesh

Vascular Medicine
– Professor Ziad Mallat, University of Cambridge

The research will be supported by a wide range of Institute researchers. The Functional Genomics work will collaborate with Sarah Teichmann's team and the Human Cell Atlas project, with additional input from Daniel Gaffney's and Gosia Trynka's teams on computational approaches and human induced pluripotent stem cells. The Population and Data Sciences studies will draw on the knowledge of Matt Hurles' team when combining genomic data with medical records.

The Centre will also call on state-of-the-art computational analyses conducted on the Institute's high-performance computers. While the Institute's sequencing, genotyping, single-cell, and genome engineering teams will enable experimental exploration and confirmation of the centre's discoveries.

Through this work, the Sanger Institute will help to drive the development of new clinical products, tools and applications, particularly for pulmonary hypertension, and large artery and small vessel stroke.

3 Hacking our genomic future

For two days in July 2018, the Wellcome Genome Campus Conference Centre became a 'Dragons' Den' of challenge, innovation, and translation of genomic science into medical benefit.

In the UK's first Genomes and Biodata Hackathon, 111 data scientists, genomics researchers, usability experts, patient advocates and entrepreneurs sought to hack the future of healthcare.

The Institute has a strong history of developing and scaling genomic technologies to benefit healthcare. To drive the next wave of innovation, the Sanger Institute's Enterprise and Innovation team worked with the Genome Campus' Connecting Sciences Advanced Courses and Scientific Conferences team to create a heady 48 hours of collaboration and ideation.

The hackathon had a strong entrepreneurial flavour, bringing together academic and industry researchers in 18 multidisciplinary teams. Each group then tackled one of five sponsored challenges, supported by 20 mentors from incubator and accelerator programmes, and Cambridge health technology companies.

After an intense period of analysing vast data sets and resource development with support from mentors, day two saw the teams pitch their viability and value of their solutions to a panel of business leaders, health technology company CEOs and leading researchers.

The event produced a range of solutions; from a mobile, AI-augmented device that allows patients with gut diseases to monitor their own microbiome to a solution that uses known drug-symptom relationships to identify candidates for drug repurposing.

The winning solutions are attracting further interest from industry and it is hoped that a number will be taken forward for further development. In fact, one team presented their idea to the NHS Chair of Pharmacogenomics Sir Munir Pirmohamed.

111
data scientists, genomics researchers, usability experts, patient advocates and entrepreneurs sought to hack the future of healthcare

Image Credits

All images belong to Wellcome Sanger Institute, Genome Research Limited except where stated below:

Outside Front Cover	– *Plasmodium* parasite inside a red blood cell – National Institute of Allergy and Infectious Diseases (NIAID)/Rocky Mountain Laboratory *Neisseria gonorrhoea* bacteria – National Institute of Allergy and Infectious Diseases, National Institutes of Health Cancer cell – Shutterstock
Inside Front Cover	– *E. coli* bacteria on *Trichuris* egg – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited
Page 3	– Child having an injection – Shutterstock Mother and daughter – iStock
Page 5	– Mosquito – Kevin Mackenzie, University of Aberdeen, Wellcome Images Grandmother and daughter – sasint, pixabay.com *Plasmodium* parasite inside a red blood cell – National Institute of Allergy and Infectious Diseases (NIAID)/Rocky Mountain Laboratory *Vibrio cholerae* bacteria – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited
Page 6	– Acute myeloid leukaemia cells – National Cancer Institute Forearm – stevepb, pixabay.com Golden eagle – dawnydawny, pixabay.com *Schistosoma mansoni* worms – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited
Page 9	– Mother and baby – iStock Foetal scan – Dr Wolfgang Moroder, Wikimedia Commons *Vibrio cholerae* bacteria – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited
Page 11	– Mother and daughter – iStock Acute leukaemia – Armed Forces Institute of Pathology (AFIP)
Page 13	– Myelofibrosis Reticulin Stain – Ed Uthman, Wikimedia Commons Vertebral osteoblastoma – Dr Henry Knipe, Radiopedia.org rID 4885
Page 14	– Ewing sarcoma in the tibia of a child – Michael Richardson, Wikimedia Commons Human kidney cortex – Menna Clatworthy, University of Cambridge
Page 15	– Red blood cells – Annie Cavanagh, Wellcome Images
Page 19	– Pregnant woman – iStock
Page 20	– Mother and child – iStock
Page 21	– Sisters – With special thanks for permission from their mother
Page 22	– Brain PET scan – Jens Maus, Wikimedia Commons Child hands and blocks – FeeLoona, pixabay.com
Page 23	– Hip replacement – NIADDK, 9AO4 (Connie Raab-contact); NIH
Page 24	– Foetal scan – Dr Wolfgang Moroder, Wikimedia Commons Mother and baby – iStock
Page 25	– Hepatitis C virus – iStock
Page 26	– MRSA cells – National Institute of Allergy and Infectious Diseases
Page 27	– *Neisseria gonorrhoea* bacteria – National Institute of Allergy and Infectious Diseases, National Institutes of Health *Schistosoma mansoni* worm – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited
Page 28	– *Vibrio cholerae* bacteria – Dave Goulding, Wellcome Sanger Institute, Genome Research Limited Dhaka – rahmatullah77, pixabay.com
Page 29	– Mother and child hands – iStock
Page 30	– Gorilla – Franck Prugnolle (CNRS/IRD/CIRMF)
Page 31	– *Plasmodium* parasite inside a red blood cell – National Institute of Allergy and Infectious Diseases (NIAID)/Rocky Mountain Laboratory *Plasmodium* parasites inside mosquitoes – Claire Sayers, Wellcome Sanger Institute, Genome Research Limited
Page 32	– Cows – Grant Brookes, unsplash.com *Staphylococcus aureus* – Annie Cavanagh, Wellcome Images
Page 33	– *Salmonella* bacteria – CDC/James Archer
Page 34	– Greater horseshoe bat – Marie Jullion, Wikimedia Commons
Page 35	– Roesel's Bush-Cricket – Ryan Hodnett, Wikimedia Commons Golden eagle – dawnydawny, pixabay.com Robin – Allan Cox, unsplash.com Turtle dove – zoosnow, pixabay.com
Page 38	– Canadian Lynx – Eric Kilby, Wikimedia Commons Mosquito – James Gathany, CDC/PHIL
Page 39	– Golden eagle – Kdsphotos, pixabay.com
Page 45	– MRSA cells – National Institutes of Health
Page 48	– *Neisseria gonorrhoeae* – CDC/PHIL
Page 49	– Houses of Parliament – Marcin Nowak, unsplash.com
Page 50	– Medicine pills – rawpixel, unsplash.com
Page 51	– Mouse hearts – NIMRC, MRC

Wellcome Sanger Institute Highlights 2018/19

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, 2019.

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