

Highlights  
2016/17



Through our ability to conduct research at scale, we are able to engage in bold and long-term exploratory projects that are designed to influence and empower medical science globally


# Contents


## **02 What we do**

03 Director's introduction

04 Institute in numbers

## **08 Our work**

 **10** Cancer, Ageing and Somatic Mutation

 **16** Cellular Genetics

 **20** Human Genetics

 **24** Infection Genomics

 **28** Malaria

## **30 Our approach**

 **32** Scale

 **34** Innovation

 **36** Culture

 **38** Influence

 **40** Connections

## **42 Other information**

42 Image credits

43 Institute information



Wherever you see this icon, click to find more information on the Sanger Institute website.

# What we do

Today at the Sanger Institute genomic scientists face unparalleled opportunities – and challenges – offered by the confluence of powerful new technologies, paradigm shifts in understanding, daring scientific ambition and global cooperation.

More than two decades ago the first reference human genome was born from the scientific community coalescing to pool knowledge, technologies and funding at a scale never before seen. The result was truly historic – the publication of humankind's 'book of life'. But, just as in nature, this sequence is not static: continual challenge has developed and refined it many times. The same is true of the Sanger Institute.

Running through this Institute Highlights are two entwined narratives – the ever-changing nature of genomes resulting from interplay between organisms

and their environment, and our evolving research facilitated by global dialogue between scientists, data sources and institutions.

The dialogue between genomes and their environment is continually shaping human and pathogen genomes, creating a flux that impacts every aspect of disease and health. From the rise of drug resistance in malaria to finding the drivers of cancer, and from the sources of rare developmental disorders to healthy bacterial mixes in the gut microbiome, genomic research plays a pivotal role in understanding health and disease. The research outlined within the following pages is providing new knowledge and insight to inform approaches to diagnosis, treatment, disease prevention and health promotion.

The only way to explore such a vast world of connections is through inclusive and equitable

research dialogue: between scientists and institutes, industry and academe, investigators and healthcare, and researchers and governments. This Institute was established to facilitate such collaborative working – first as part of the Human Genome Project and now by laying the foundations for sustainable global networks for genomic research.



**Mike Stratton, Director**  
Wellcome Trust Sanger Institute

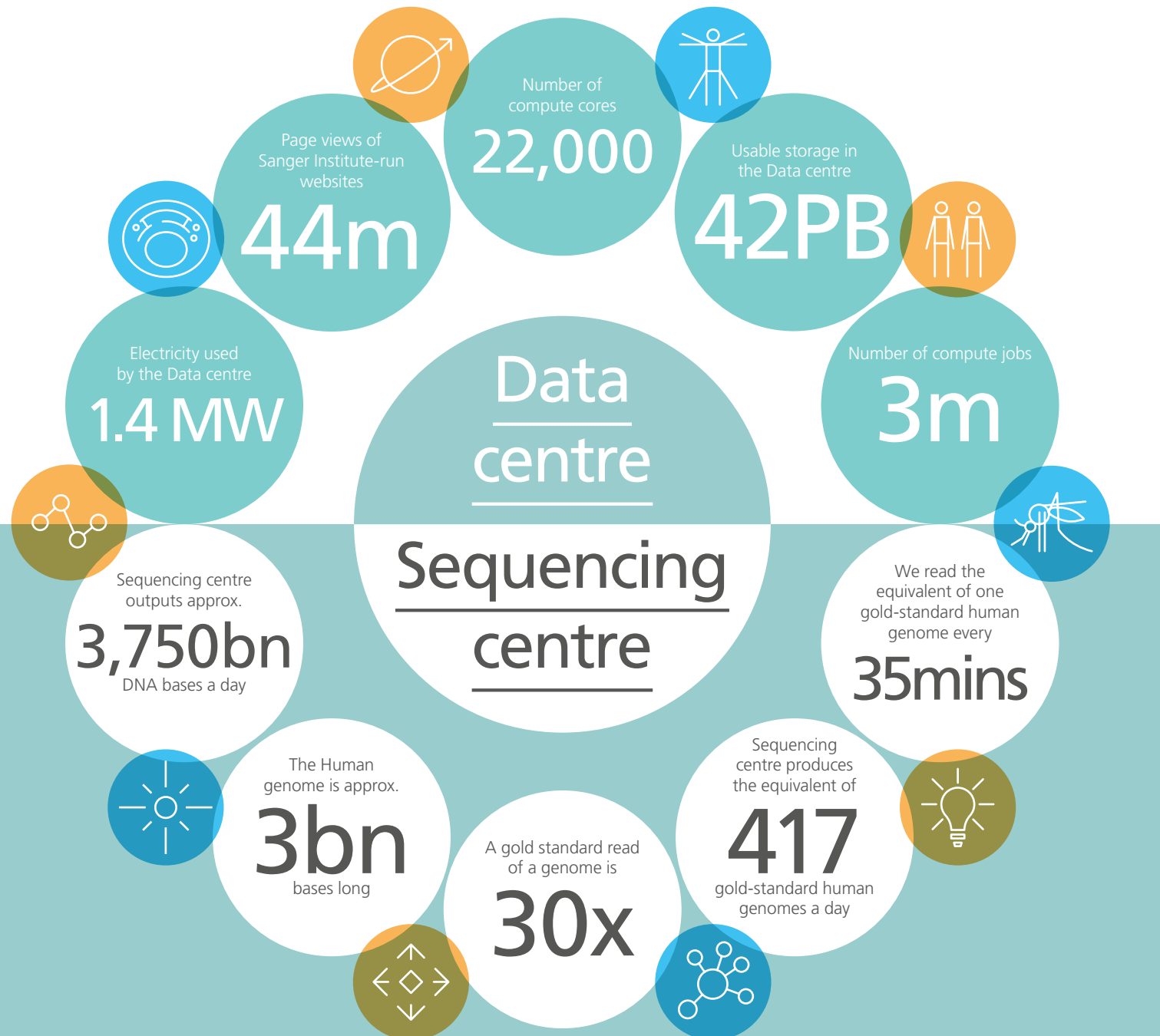
What we do

Our work

Our approach

Other information

## 2016 at a glance

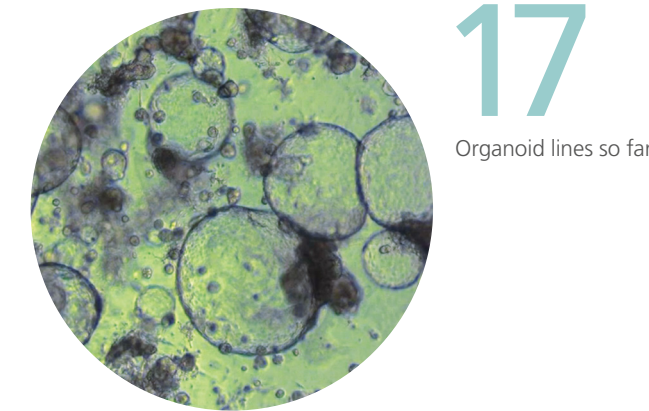


All facts and figures gathered in December 2016

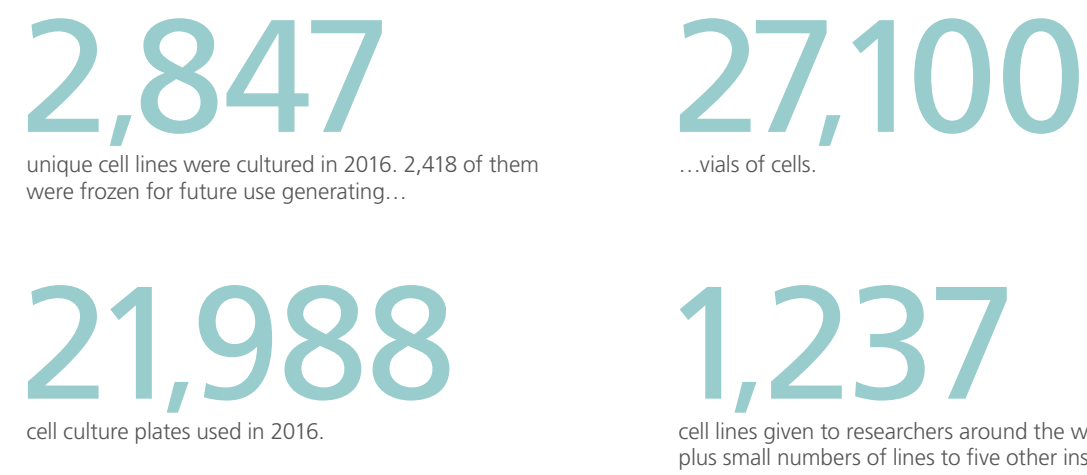
## Publications



## Organoids



## Cellular Generation and Phenotyping

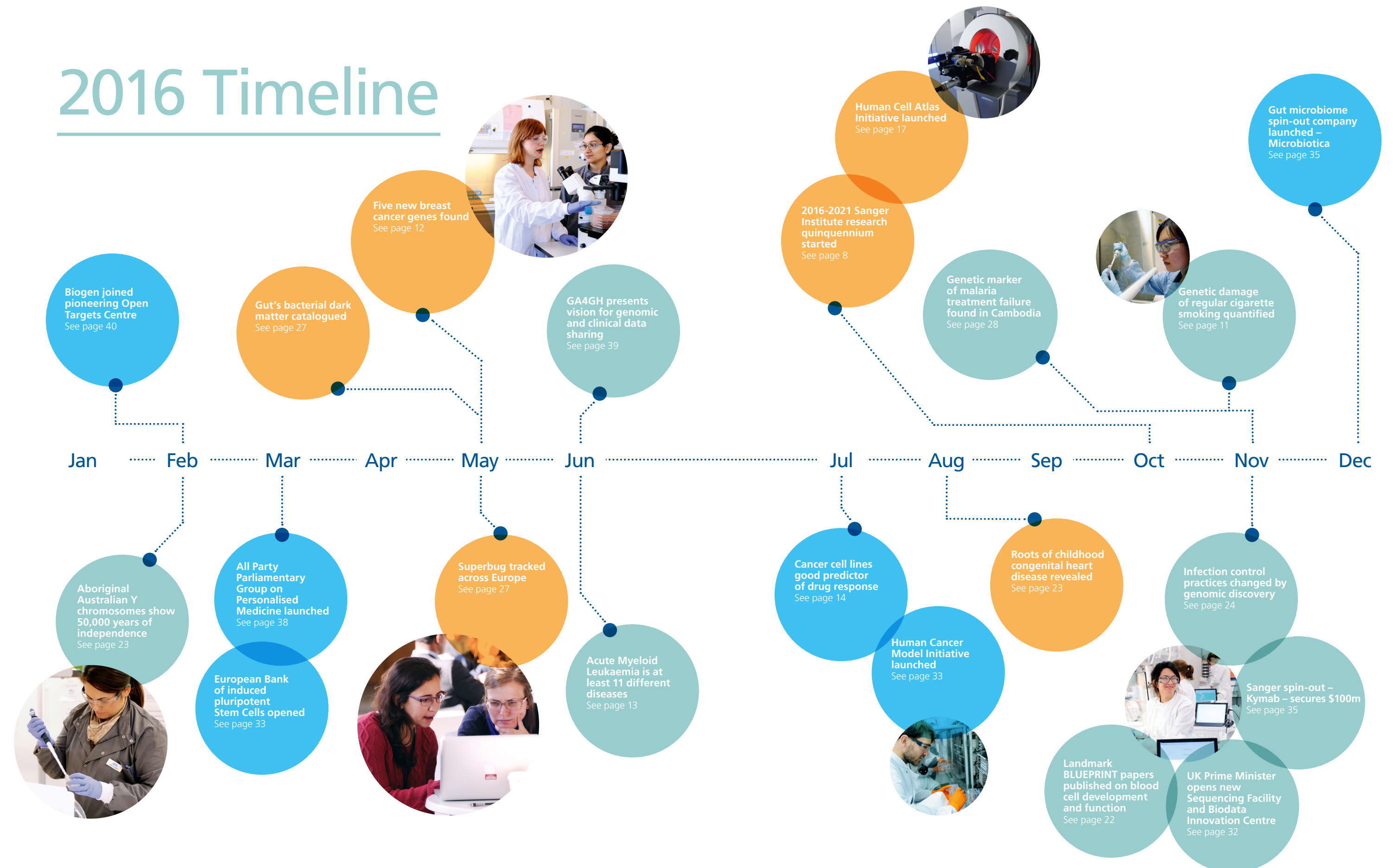


### Cell lines shared with partner organisations



ECACC and EBISC act as distribution hubs for our induced pluripotent stem cell lines and send them to researchers around the world.

## 2016 Timeline



# Our work



## 10 Cancer, Ageing and Somatic Mutation Programme

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 Programme

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 Programme

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.



## 24 Infection Genomics Programme

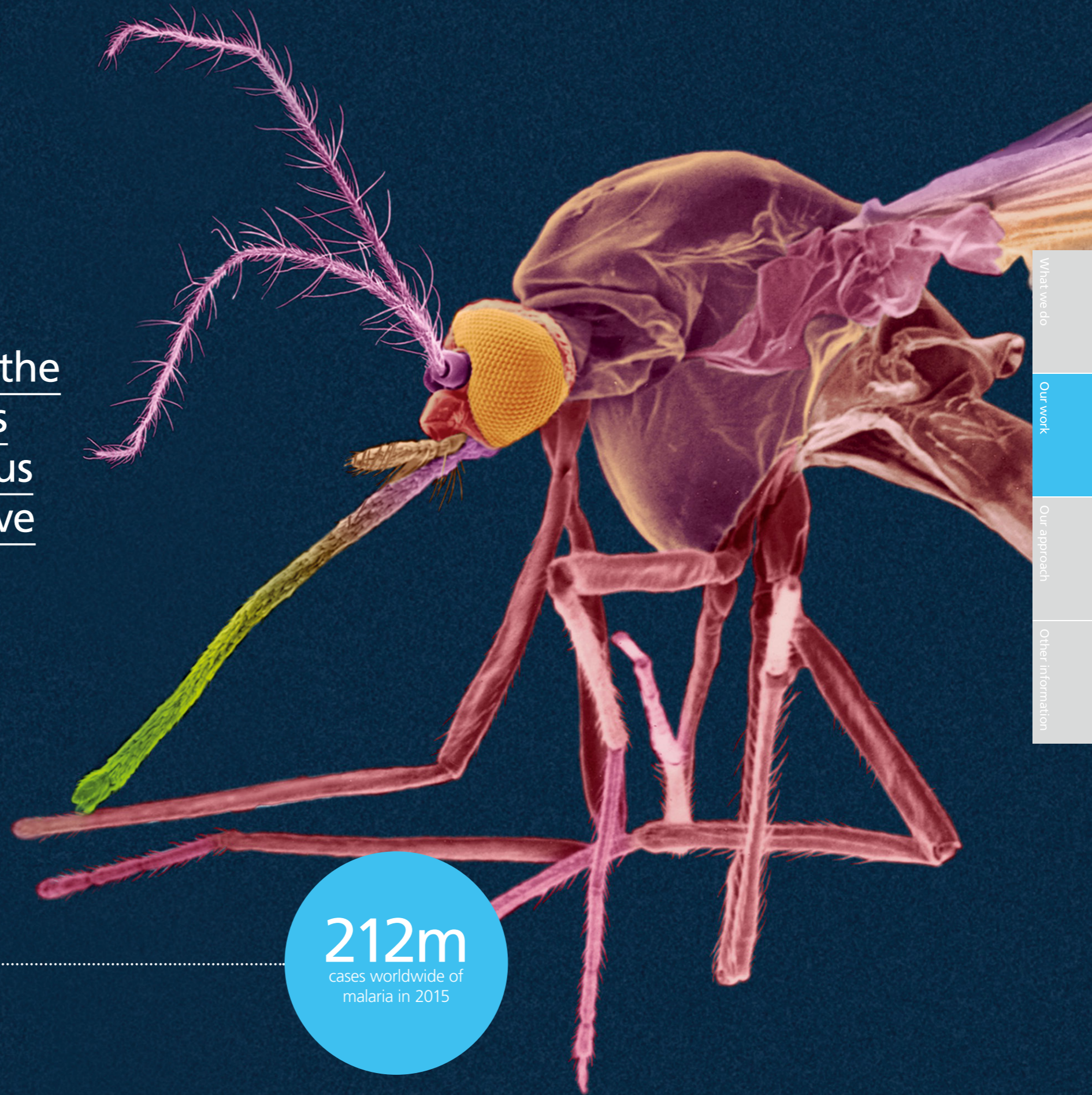
Investigates the common underpinning mechanisms of evolution, infection and resistance to therapy in viruses, bacteria and parasites. It also explores the genetics of host response to infection and the role of the microbiota in health and disease.



## 28 Malaria Programme

Integrates genomic, genetic and proteomic approaches to develop and enhance high-throughput tools and technologies to study specific biological problems relevant for malaria control and to understand the fundamental science of the human host, the mosquito vector and the *Plasmodium* pathogen.

With secured funding from Wellcome for the next five years we aim to focus our work in five key research programmes



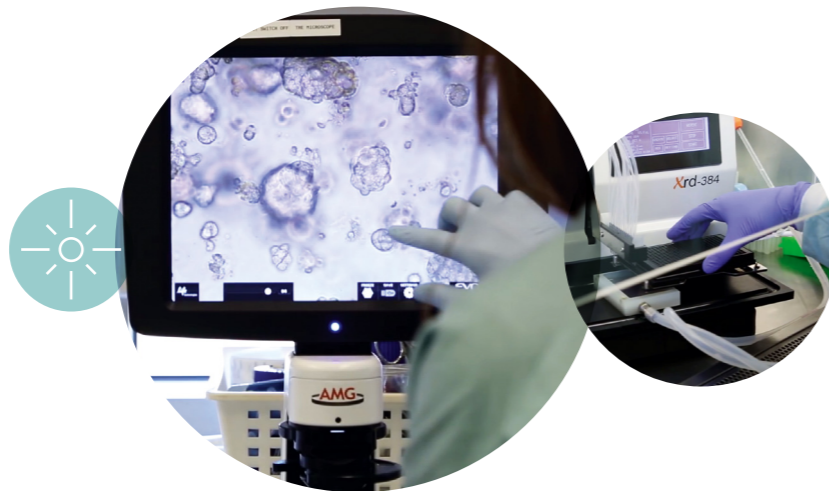
212m cases worldwide of malaria in 2015

- What we do
- Our work
- Our approach
- Other information



# Cracking cancer's conundrums

The Cancer, Ageing and Somatic Mutation Programme has sequenced and analysed cancer genomes at scale to reveal complexity and surprising commonality. Sanger researchers have begun to tease apart the multiple mechanisms disrupting the genome and driving cancer – offering insights that could inform new diagnostics and treatments.



## In this section

- 11 Mapping DNA damage in the body
- 12 Hidden root of skin cancer found
- 12 Breast cancer: the next step
- 13 Genomic microscope gives better diagnoses
- 14 Cell lines point to best drugs
- 15 Cancer in 3D: organoids

For more than a decade, Sanger researchers have been at the forefront of large-scale cancer genome projects that have generated an ever-lengthening catalogue of genes linked to various cancers. These efforts have revealed a daunting diversity and complexity to cancer's genomic landscape: mutations in many genes can cause cancer, particular types of cancer can have multiple genetic causes, and cancers are constantly evolving within individual patients.

To meet these challenges, the Cancer, Ageing and Somatic Mutation Programme employs a range of scientific approaches. At the international level the Programme leads and contributes to many worldwide collaborations that generate, share and interpret cancer genome data.

Through computational genomics Sanger scientists pioneer innovative methods of data analysis and pattern recognition to provide important new insights into cancer cell biology. At the laboratory bench Sanger research groups work with international partners to develop novel cancer cell models that could transform *in vitro* studies of cancer and the development of therapies.

**“All cancers are due to mutations that occur in all of us in the DNA of our cells during the course of our lifetimes. Finding these mutations is crucial to understanding the causes of cancer and to developing improved therapies.”**

**Professor Sir Mike Stratton**  
Director of the Sanger Institute

## Mutational signatures: new insights into DNA damage

In recent years, Sanger researchers have made great progress in understanding not only what has gone wrong in cancer cells, but also how. By analysing the altered genetic landscapes of many thousands of cancers, they have identified dozens of distinctive patterns of DNA damage. Each of these ‘mutational signatures’ is the result of a distinct DNA-damaging process. This work is shedding light on the molecular mechanisms that compromise the integrity of the genome – knowledge that could underpin new ways to treat or prevent cancer.

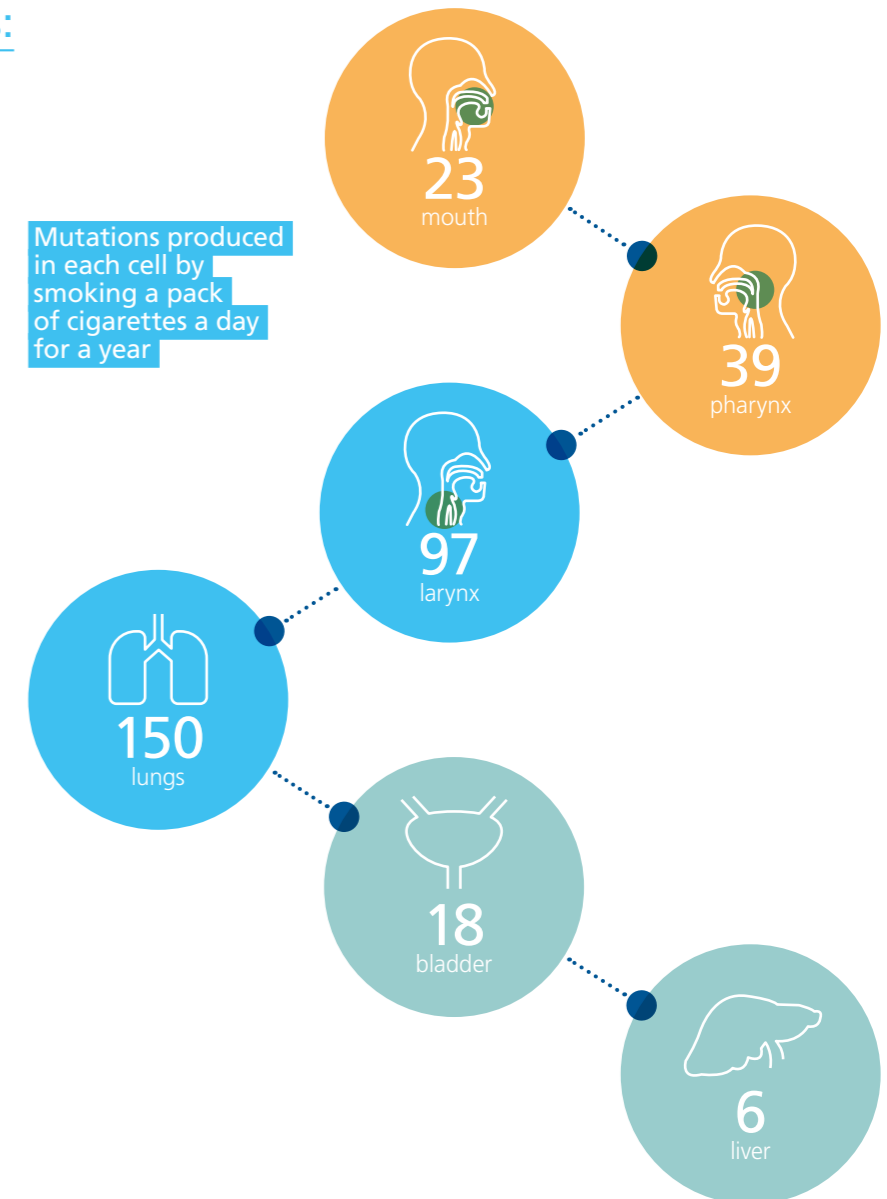
In 2010, Sanger researchers identified distinctive genetic changes in the genome of a smoker's lung cancer, revealing how the constituent chemicals in tobacco smoke damage DNA in different ways. Extending this approach to other forms of cancer, Sanger scientists systematically analysed data from multiple cancer genomes, identifying a wide range of mutational signatures linked to different DNA-damaging processes. For example, some types of mutation are linked to abnormal activation of an antiviral defence mechanism (the APOBEC system), while others seem to reflect the action of a cellular ‘clock-like’ mechanism, leading to the steady accumulation of mutations over time.

### Signatures show extent of smoking's genetic damage

In 2016 Sanger scientists returned to smoking-associated cancer by leading a study of more than 5,000 genomes of smoking-linked cancers. Computational analysis, published in *Science*, revealed that smoking is associated with multiple mutational signatures, suggesting that tobacco smoke damages DNA and drives cancer through a diversity of mechanisms.<sup>1</sup>

One common type of damage is chemical modification of DNA bases by compounds in tobacco smoke, which was seen only in cancers of tissues exposed to tobacco smoke, such as lung, larynx and mouth. In cancers of tissues not directly exposed to tobacco smoke, such as pancreas and kidney, smoking affected other cellular processes that introduce mutations into

Mutations produced in each cell by smoking a pack of cigarettes a day for a year



DNA, including APOBEC and clock-like mechanisms. The analysis also revealed that, on average, smokers consuming a pack of cigarettes a day accumulated an extra 150 mutations in every lung cell each year, 23 mutations in each mouth cell and six mutations in each liver cell.

Like tobacco smoke, ionising radiation can damage DNA and increase the risk of cancer. Each year, a small number of patients receiving radiotherapy develop cancers because they are exposed to radiation. In work published in *Nature Communications*, Sanger scientists compared the genomes of 12 such

radiation-induced cancers with those not linked to radiation exposure identifying two mutational signatures specifically associated with exposure to ionising radiation.<sup>2</sup>

Radiation-induced cancers typically carried approximately 200 small deletions (of up to 100 base pairs) as well as highly unusual short stretches of ‘flipped’ DNA (balanced inversions). These signatures will help clinicians to determine whether cancers are linked to radiation exposure, and enable researchers to explore the possibility of specific types of treatment.



## Redheads carry a cancerous legacy

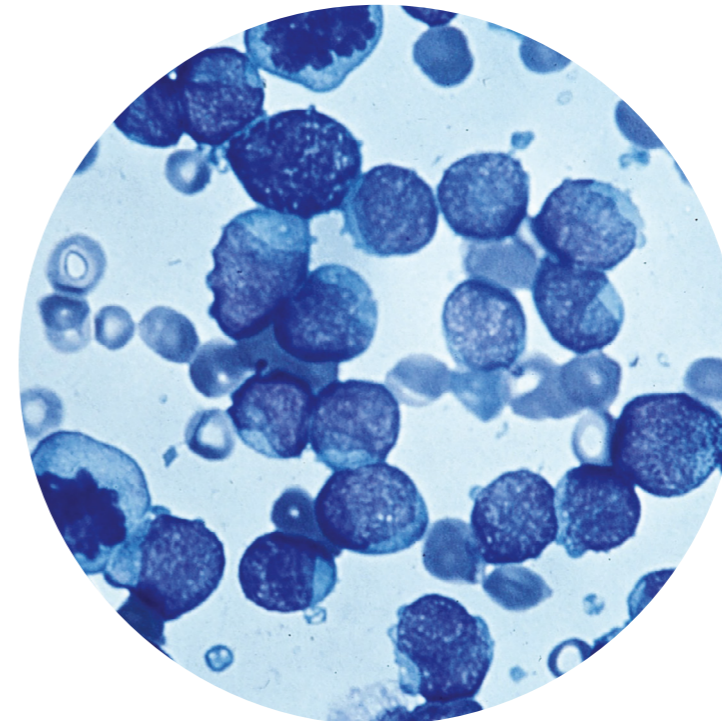
UV radiation is a particular danger to people with fair skin, red hair and freckles, who carry two copies of a variant of the melanocortin 1 receptor (*MC1R*) gene. Such people are unable to produce a protective skin pigment and are at increased risk of melanoma.

By analysing public databases of melanoma cancer genomes, Sanger researchers reported in *Nature Communications* that the tumours of patients with the *MC1R* variant contained markedly more mutations – equivalent to an extra 21 years of sun exposure.<sup>3</sup> Moreover, even people with just one copy of the *MC1R* variant – who are not redheads – showed high levels of DNA damage, and are at significantly increased risk of melanoma.

Furthermore, as well as the telltale mutational signature associated with solar UV exposure, people with one or two copies of the *MC1R* variant also showed increased levels of other mutational signatures. Hence the *MC1R* variant may be affecting cancer risk through multiple mutational processes in skin cells.

**“This is one of the first examples of a common genetic profile having a large impact on a cancer genome and could help better identify people at higher risk of developing skin cancer.”**

**Dr David Adams**  
Group leader at the Sanger Institute



## Acute myeloid leukaemia

New genomic classifications based on specific combinations of mutations offer a way to discern a person's likely prognosis and response to therapy

A Sanger-led collaboration reported in *The New England Journal of Medicine* its analysis of 111 genes implicated in AML in more than 1,500 patients taking part in clinical trials of intensive therapy.<sup>6</sup> The study identified more than 5,000 cancer-causing mutations affecting 76 regions of the genome, with most patients having at least two mutations. Crucially, the team was able to link the presence of certain combinations of mutations to key clinical factors, such as prognosis and response to treatment. In total the team distinguished 11 different types of AML, some corresponding to existing patient subgroups but also new subtypes that reflect novel mechanisms by which AML may arise.

Although the new classification needs to be verified, it has the potential to offer an important new way to classify AML patients according to their likely prognosis and response to therapy. The findings are also a key step towards more efficient clinical trials in which participants are recruited on the basis of their cancer's genomic landscape and its likelihood of response to the treatment.

As AML illustrates, a single type of cancer may have multiple genetic causes. Equally, the same genetic abnormality may contribute to more than one type of cancer. This offers the possibility that targeted treatments could be effectively deployed across multiple types of cancer that share cancer-causing mutations.

**“We have shown that AML is an umbrella term for a group of at least 11 different types of leukaemia. We can now start to decode these genetics to shape clinical trials and develop diagnostics.”**

**Dr Peter Campbell**  
Head of the Cancer, Ageing and Somatic Mutation Programme

## Breast cancer research enters new era

Mutational signatures have also generated new insight into the origins of breast cancer, one of the most intensively studied cancers. Sanger researchers led the most comprehensive analysis yet of breast cancer genomes, revealing a host of processes disrupting genomic integrity.<sup>4</sup>

In *Nature*, the international collaboration reported the results of analysing whole-genome sequences from 560 cancer genomes. While previous sequencing efforts have typically concentrated on finding DNA variations in the exome (the protein-coding parts of the genome), this study included non-coding regions to explore the role of gene activity control in triggering cancerous cell growth. By surveying the entire genomic landscape of breast cancer, the team was also able to search for signatures of specific genomic rearrangements associated with the disease.

The study added five new members to the list of protein-coding genes implicated in breast cancer – now up to 93. It also identified 12 base substitution and six rearrangement mutational signatures associated with breast cancer. Three of the rearrangement signatures were associated with defective DNA repair mechanisms – one linked to the common *BRCA1* gene, one to *BRCA2* and one of unknown origin.

The study revealed key details of the mutational processes leading to breast cancer and the critical genes they affect. It suggests that most of the genes involved in breast cancer have been identified, and that fusion genes and mutations in non-coding regions are likely to play only a minor role.

In associated work, published in *Nature Communications*, Sanger scientists have led an analysis examining how the multiple mutational processes implicated in breast cancer affect different regions of the genome.<sup>5</sup> Notably, mutations do not occur

**“In the future, we'd like to be able to profile individual cancer genomes so that we can identify the treatment most likely to be successful for a woman or man diagnosed with breast cancer. It is a step closer to personalised healthcare for cancer.”**

**Dr Serena Nik-Zainal**  
Group leader at the Sanger Institute

at random across the genome, but show an association with particular genomic landmarks, such as those involved in DNA replication, transcription and folding of DNA into higher-order chromatin structures. Furthermore, each mutational signature shows its own distinct relationship with these genomic landmarks.



## Whole-genome profiling

Visualising breast cancer genomes in this way may lead to more effective, personalised treatment decisions

By linking mutational signatures to cellular processes, the study provides new insight into the molecular mechanisms by which these mutational processes alter the genome. More generally, the study shows that mutation is not a random process proceeding at a steady pace – different mutational processes preferentially affect particular genomic sites and act at different rates.

## Personalising treatment: developing genomic diagnosis

A key goal of cancer genomics is to create new tools that help clinicians and benefit patients. In particular it is hoped that a deeper understanding of the genetic origins of individual cancers will lead to more targeted drug therapies aimed at specific molecular abnormalities, and diagnostics that will enable clinicians to tailor treatment to the specific genomic landscape of a patient's cancer. Sanger researchers are in the vanguard of this work.

For decades, cancers have been characterised according to their position in the body and appearance under the microscope. Cancer genomics offers an alternative approach, with individual cancers classified according to the genetic changes that have rendered them cancerous.

An international collaboration led by Sanger scientists has taken an important step in this direction for the blood cancer acute myeloid leukaemia (AML). Genome-sequencing projects have identified multiple genes causing AML, making it a useful testbed for genetic classification systems.



## Personalising treatment: marrying genomics to successful therapies

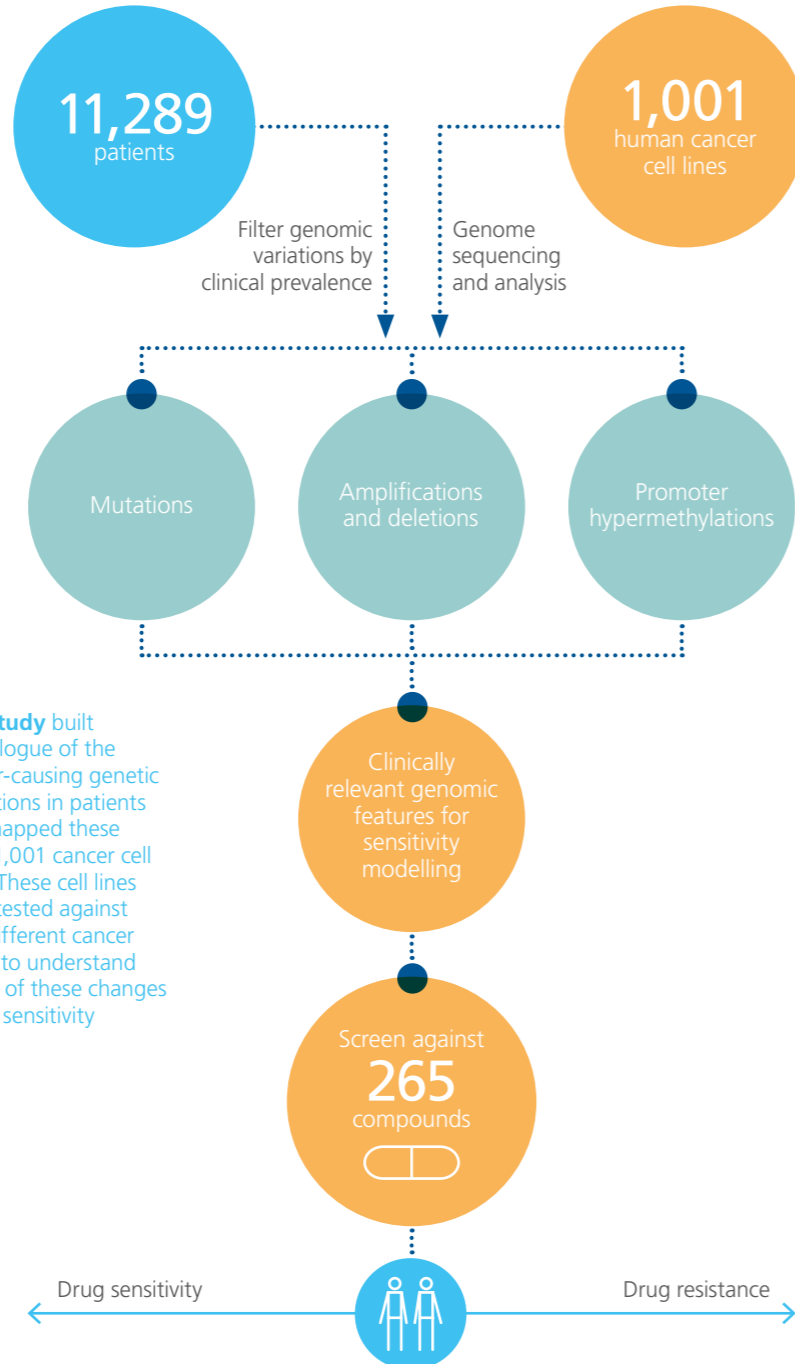
The principle that shared mutations may mean that different cancers could be treated with the same drugs has underpinned a major programme of work led by Sanger researchers in collaboration with Massachusetts General Hospital. High-throughput technologies were used to assess the impact of hundreds of cancer drugs – licensed and in development – on a wide range of genetically characterised cancer cell lines, in order to identify genetic signatures correlating with sensitivity to particular drug treatments. This programme identified several correlations between particular genetic aberrations and drug responsiveness, with the potential to extend the use of certain drug treatments.

In 2016, a large-scale international collaboration led by Sanger researchers extended this work by comparing the genetic changes seen in cancer cell lines with those present in cancer patients. The study, reported in *Cell*, hunted for clinically relevant genetic changes in some 11,000 cancers spanning 29 tumour types.<sup>7</sup> The presence of these mutations was then assessed in 1,001 cancer cell lines treated with 265 anti-cancer compounds, to identify combinations of mutations predictive of drug sensitivity.

**“Our research shows that cancer cell lines do capture the molecular alterations found in tumours, and so can be predictive of how a tumour will respond to a drug. This means the cell lines could tell us much more about how a tumour is likely to respond to a new drug before we try to test it in patients.”**

**Dr Ultan McDermott**  
Group leader at the Sanger Institute

### Mapping the landscape of drug interactions in cancer



The study discovered that the genetic changes identified in patients were typically also present in cancer cell lines, confirming that the latter are a useful tool in cancer drug development. Cancer cell lines are easy to work with, and researchers can now be more confident that findings will be relevant to patients.

Crucially, the work also generated much more information on combinations of mutations affecting drug responsiveness. As well as shedding light on cancer cell biology, this information will also support more efficient clinical trials by matching patients to drugs to which they are likely to respond.



## Organoids: Modelling cancer in 3D

Cancer cell lines are useful tools in researching cancer, but they do have their limitations when compared with cancers seen in patients. Cell lines do not show the same three-dimensional arrangement of cells, which can be important to the properties of cancers, and they consist of a single type of cancer cell, whereas native cancers are typically made up of multiple cancer cell variants.

Researchers at the Hubrecht Institute in the Netherlands have developed a way to culture gut stem cells so that they produce clumps of organised gut tissue, or ‘organoids’. Organoids can be produced from both normal tissue and from the tumours of colorectal cancer patients, creating a ‘living biobank’.

In a 2015 *Cell* paper, Sanger scientists, working with their Hubrecht colleagues, reported using high-throughput screens to assess how cancer organoids respond to anti-cancer drugs.<sup>8</sup> Importantly, they were able to explore how the presence of heterogeneous populations of cancer cells – as seen in patients – affects sensitivity to drugs. This technology raises the tantalising prospect that organoids could be cultured from individual patients and then used to screen potential drug treatments. However, the speed and efficiency of organoid culturing will need to be enhanced for this to become a clinical reality.

In 2016, the foundation Hubrecht Organoid Technology and the Sanger Institute teamed up with the US National Cancer Institute and Cancer Research UK to establish a partnership to promote the use of organoid cancer cell models. The Human Cancer Models Initiative will make available approximately 100 cancer cell models grown directly from patient samples, preserving three-dimensional tissue structure and cancer cell diversity.

The cancer models will be genetically characterised and linked to anonymised clinical information from donors. As well as providing more life-like models for research, the new resource will also encourage work on standardised systems, accelerating cancer characterisation and testing of therapies.



### References

- Alexandrov LB *et al. Science*. 2016; 354: 618–622.
- Behjati S *et al. Nat Commun*. 2016; 7: 12605.
- Robles-Espinoza CD *et al. Nat Commun*. 2016; 7: 12064.
- Nik-Zainal S *et al. Nature*. 2016; 534: 47–54.
- Morganella S *et al. Nat Commun*. 2016; 7: 11383.
- Papaemmanuil E *et al. N Engl J Med*. 2016; 374: 2209–21.
- Iorio F *et al. Cell*. 2016; 166: 740–54.
- van de Wetering M *et al. Cell*. 2015; 161: 933–45.







# Mapping 37.2 trillion cells – one cell at a time

In 2016, the Cellular Genetics programme began to lay the foundations for one of the most ambitious research initiatives ever conceived – the International Human Cell Atlas. The global collaboration seeks to use genomics to understand the biology of every cell type in the human body. To achieve this aim, Sanger scientists are creating the laboratory techniques and computational tools needed to categorise the building blocks of life.

A complete picture of health and disease calls for a full understanding of how cells perform their specific roles and work in concert. Yet, surprisingly, it is not even certain how many different types of cell exist among the 37.2 trillion cells in the human body. Sanger researchers are working with collaborators around the world to develop methods to probe individual cells in unprecedented detail. This work will pave the way to deliver the most intricate map of all – a cell atlas of the human body.

## In this section

16 Single cells reveal immunotherapy target

17 TraCeR tracks T-cell response

17 Creating a cellular encyclopaedia of the human body

18 Induced pluripotent stem cells: test bed and future therapy?

19 Single-cell method ties epigenetics to expression levels

19 Zebrafish show how blood cells mature

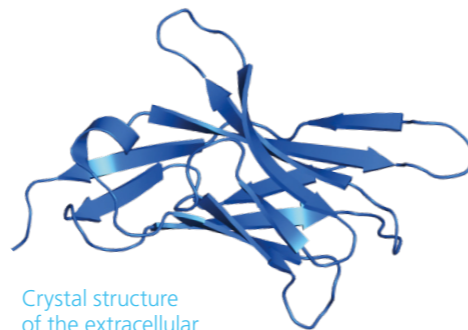
19 Computational tool enhances single-cell analysis

## Single-cell study reveals immunotherapy target

Cells have traditionally been classified according to their appearance or expression of surface molecules. More recently, techniques such as RNA-seq – sequencing of RNA transcripts – have allowed researchers to use gene expression profiles instead. This approach is now being applied to individual cells, using single-cell RNA-sequencing (scRNA-seq).

In 2016, a Sanger-led study used scRNA-seq to investigate a recently defined class of immune cell – innate lymphoid cells – and discovered a potential therapeutic target for asthma and autoimmune diseases.<sup>1</sup>

Innate lymphoid cells play a key role in orchestrating immune response. So far three classes of the cells have been defined, but little is known about their origins or how they acquire their distinct functions.



Crystal structure of the extracellular domain of murine PD-1

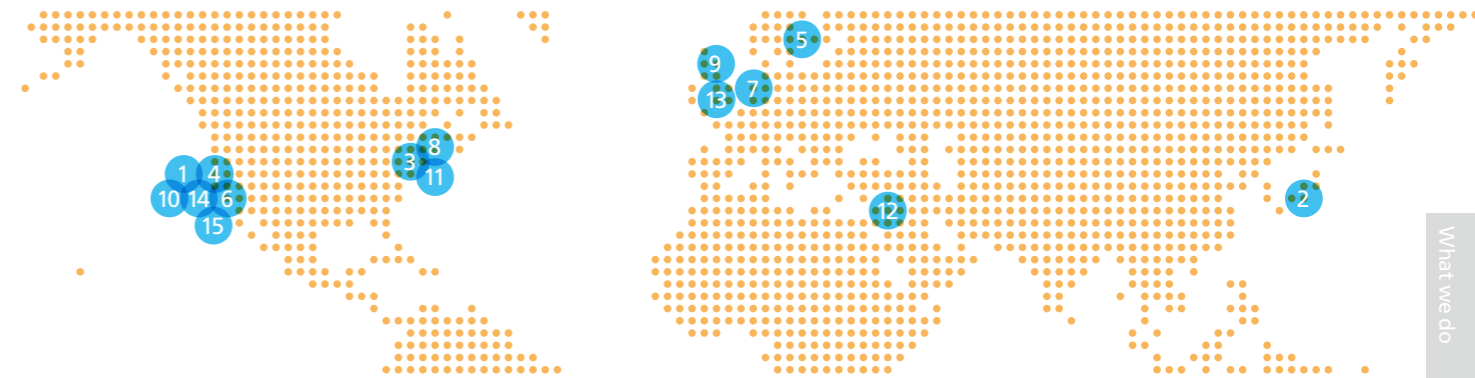
Sanger researchers used scRNA-seq to characterise gene expression in the progenitor cells that give rise to innate lymphoid cells, identifying different stages in the cells' development and pathways by which they differentiate. Significantly, the scientists discovered that the progenitor cells displayed a molecule known as PD-1 on their cell surfaces, and that this same molecule was expressed at high levels in activated innate lymphoid cells.

In experimental models, removing PD-1-expressing innate lymphoid cells reduced cytokine production during infection and blocked allergen-induced inflammatory responses in the lung. Targeting PD-1 could therefore be a way to modulate immune responses – particularly as drugs acting on PD-1 have already been developed for cancer immunotherapy.<sup>2</sup>

**“This study helps us understand the biology of the immune system in ways that were impossible previously... Not only is this useful for asthma and other inflammatory diseases, it could also help us understand what is happening during PD-1 cancer treatment.”**

**Dr Yong Yu**  
Postdoctoral fellow at the Sanger Institute

## International Human Cell Atlas (IHCA) 2016 International Organising Committee



### Key

1. Dr Cori Bargmann, Chan Zuckerberg Initiative, US
2. Dr Piero Carninci and Dr Jay W. Shin, RIKEN Center for Life Science Technologies, Japan
3. Dr Nir Hacohen, Massachusetts General Hospital, US
4. Prof Arnold Kriegstein, University of California-San Francisco, US
5. Prof Sten Linnarsson, Karolinska Institute, Sweden
6. Prof Gary Nolan, Stanford School of Medicine, US
7. Prof Alexander van Oudenaarden and Prof Hans Clevers, Hubrecht Institute, Netherlands
8. Prof Dana Pe'er, Sloan Kettering Institute, US
9. Prof Chris Ponting, University of Edinburgh and the Sanger Institute, UK
10. Prof Steve Quake, Stanford University and the Chan Zuckerberg biohub, US
11. Prof Aviv Regev and Prof Eric Lander, Broad Institute of MIT and Harvard, US
12. Prof Euhd Shapiro and Dr Ido Amit, Weizmann Institute of Science, Israel
13. Dr Sarah Teichmann, Prof Michael Stratton and Dr Peter Campbell, Wellcome Trust Sanger Institute, UK
14. Prof Jonathan Weissman, University of California San Francisco and Howard Hughes Medical Institute, US
15. Prof Barbara Wold, California Institute of Technology, US

## The International Human Cell Atlas

The drive to understand the role of cells in health and disease – and the development of powerful new tools to study and manipulate them – is underpinning one of biology's most ambitious initiatives, the International Human Cell Atlas.

Conceived by researchers at the Sanger Institute and the Broad Institute in the US, the International Human Cell Atlas has been acknowledged as a key project by the Chan Zuckerberg Initiative and has the support of the Wellcome Trust and other influential bodies. The ultimate goal is to identify and characterise all the cell types of the human body – creating a resource that would underpin biomedical research for generations.

The International Human Cell Atlas has been envisioned as a global collaborative enterprise. Pilot studies have been launched in four key areas – the immune system, nervous system, epithelial cells and cancer – and the initiative's leaders are developing a longer-term roadmap for the delivery of the project.

## Discovering how T cells respond to infection

Sanger scientists have been able to track the behaviour of populations of related T cells during infection by combining single-cell studies with innovative bioinformatics. The T-cell receptor consists of two subunits encoded by separate genes.

These genes are enormously diverse, so it is highly unlikely that two T cells will share the same pair of subunits unless they are derived from the same ancestor. Sanger researchers developed a computational tool, TraCeR, that analyses scRNA-seq data to identify paired T-cell receptor sequences from individual cells, which act as a tag of related cells.<sup>2</sup>

To demonstrate its potential, the team used TraCeR to track T-cell clones before, during and after infection of mice with *Salmonella*. The scRNA-seq data revealed

how individual cells within each clone respond at each stage of infection, shedding light on their functional roles. In the future this approach could be used to examine the contribution of T-cell clones to immune responses during infection, after vaccination or in autoimmune disease.

**“This kind of breakthrough work can only be done using single-cell measurements. This new tool gives us a new approach to the study of T cells, and opens up new opportunities to explore immune responses in disease, vaccination, cancer and autoimmunity.”**

**Dr Sarah Teichmann**  
Head of Cellular Genetics and Group leader at the Sanger Institute



## Induced pluripotent stem cells: test bed and future therapy?

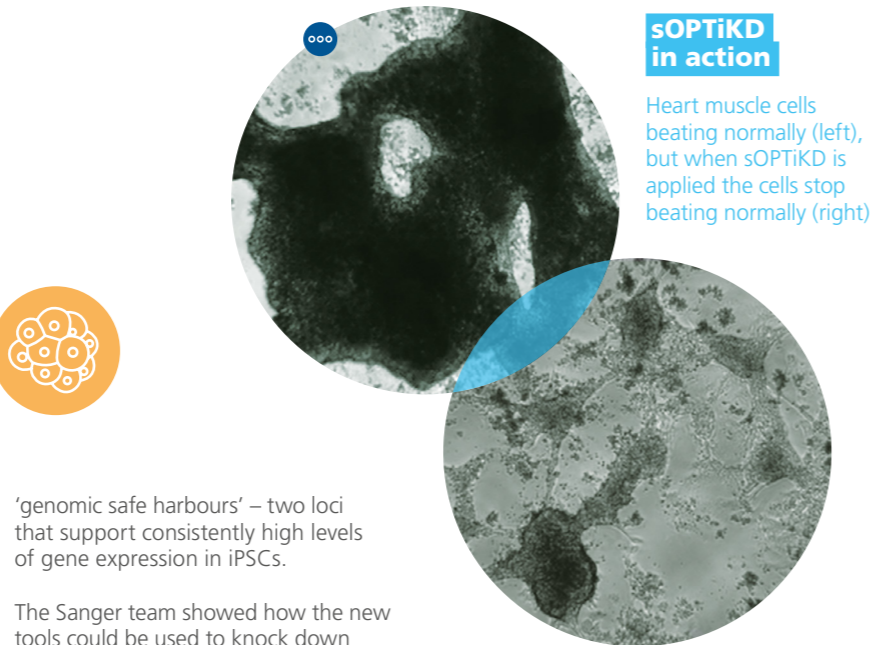
Sanger scientists have developed new methods to control differentiation of induced pluripotent stem cells (iPSCs) that, when combined with methods of targeted genetic manipulation, offer valuable new ways to explore cell function in health and disease.

Although methods exist to modify genes in human iPSCs, they have drawbacks. In particular, it is challenging to eliminate gene function at particular stages of development. Two new tools developed by Sanger researchers now provide this key capability.

The tools, known as sOPTiKD and sOPTiKO, use either short hairpin RNAs (shRNAs) to 'knock down' levels of gene expression (sOPTiKD) or the CRISPR/Cas9 system to eliminate specific genes (sOPTiKO).<sup>3</sup> Production of shRNAs or CRISPR/Cas9's DNA-targeting RNA is controlled by an inducible promoter active only when the antibiotic tetracycline is supplied. To ensure high levels of expression, the RNA-producing constructs are introduced into

**"As a cell develops the genes within it take on different roles. By allowing the gene to operate during the cell's development and knocking it out with sOPTiKO at a later developmental step, we can investigate exactly what it is doing at that stage."**

**Professor Ludovic Vallier**  
Group leader at the Sanger Institute



### sOPTiKD in action

Heart muscle cells beating normally (left), but when sOPTiKD is applied the cells stop beating normally (right)

'genomic safe harbours' – two loci that support consistently high levels of gene expression in iPSCs.

The Sanger team showed how the new tools could be used to knock down expression of specific genes at particular stages of iPSC differentiation. As a relatively quick, easy-to-use and robust pair of technologies, sOPTiKD and sOPTiKO open up new opportunities to explore the roles of genes in development, physiology and disease.

One of the most exciting manipulations of stem cells would be to correct genetic errors in the cells of people with inherited diseases – 'personalised cell therapy'. In 2011, Sanger researchers showed that this approach was technically feasible, correcting an abnormal  $\alpha$ 1-antitrypsin gene in human iPSCs.

One concern about the therapeutic use of iPSCs is the possibility they might proliferate excessively and cause cancer. Sanger-led work has shed important light on genomic changes in iPSCs – and found that the risk of cancer is low.

### How damaging is cell reprogramming?

Each cell in the body is ultimately derived from the fertilised egg, following multiple rounds of cell division. During this developmental journey, cells can acquire mutations – somatic mutations – so that the adult body is, in effect, a mosaic of cells with slightly different genomes. Further genetic changes may be introduced when cells are taken from the body and converted into iPSCs.

To explore mutational processes during development and reprogramming, Sanger scientists isolated and expanded individual adult cells, then generated multiple iPSC lines from each cell.<sup>5</sup> Genomic sequencing

of the original cell and its iPSC derivatives was then used to distinguish mutations that had arisen during development from those introduced during reprogramming.

Mutation rates were an order of magnitude lower in iPSCs – probably because genome-protection mechanisms are particularly active in stem cells. Notably, differences were seen in the types of mutation occurring during development and reprogramming, with those in iPSC often showing the hallmarks of oxidative stress-induced damage.

Most positively, there was no evidence of mutations likely to cause cancer. This is encouraging news for those aiming to use stem cells therapeutically, although cells would still need to be characterised genomically before therapeutic use.

### References

1. Yu Y *et al.* *Nature*. 2016; 539: 102–6.
2. Stubbington MJ *et al.* *Nat Methods*. 2016; 13: 329–32.
3. Bertero A *et al.* *Development*. 2016; 143: 4405–18.
4. Yusa K *et al.* *Nature* 2011; 478; 7369; 391–4.
5. Rouhani FJ *et al.* *PLoS Genet*. 2016; 12: e1005932.
6. Angermueller C *et al.* *Nat Methods*. 2016; 13: 229–32.
7. Macaulay IC *et al.* *Cell Rep*. 2016; 14: 966–77.
8. Ilicic T *et al.* *Genome Biol*. 2016; 17: 29.
9. Delmans M, Hemberg M. *BMC Bioinformatics*. 2016; 17: 110.

## Bench-based innovations revealing epigenetics



Sanger scientists developed new advances in laboratory techniques to add further value to scRNA-seq, for example to allow researchers to explore the relationship between epigenetic modifications and levels of gene expression. In 2016 Sanger researchers developed scM&T-seq, a method for simultaneously evaluating transcription and sites of DNA methylation in single cells.<sup>6</sup> When scientists applied the technique to 61 mouse embryonic stem cells, scM&T-seq not only confirmed known links between methylation sites and transcription of particular genes, but also revealed novel sites of DNA modification affecting the expression of genes that regulate stem cell pluripotency.

**"Parallel profiling of the methylome and transcriptome from the same single cell is feasible and may allow us to identify factors that regulate the relationship between transcription and DNA methylation."**

**Professor Wolf Reik**  
Associate Faculty at the Sanger Institute



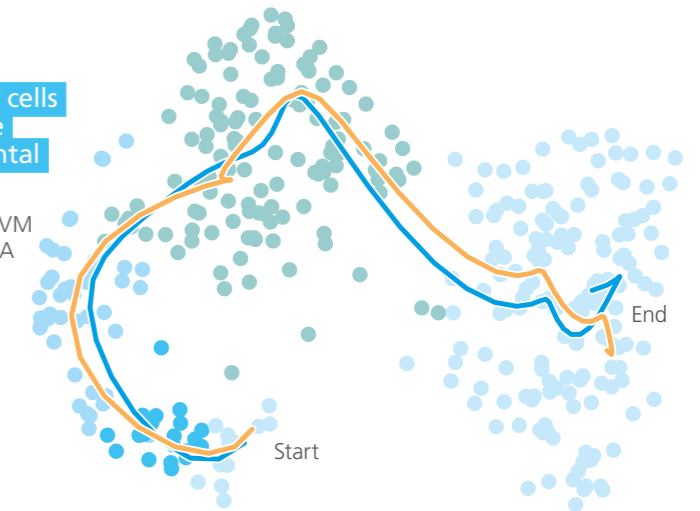
## Algorithms drive single-cell analysis

The Sanger Institute's increasing use of scRNA-seq is founded upon innovations in both experimental procedures and computational methodology that its scientists continue to develop and refine.

Single-cell research requires robust quality assurance: cells are often damaged by the procedures used to isolate them and data from such cells can contaminate the results. Manually inspecting cells can detect some abnormalities but is neither sufficiently rigorous nor readily scalable for supporting high-throughput analyses. Computational approaches may enable greater discernment at high volume,

### Ordering of cells through the developmental trajectory

- Bayesian GPLVM
- Clusters in ICA



## Gene expression reveals steps in blood cell development



Single-cell RNA sequencing is unlocking the gene expression programmes that underlie the development of blood cells. One Sanger-led study applied scRNA-seq in zebrafish to explore the differentiation of haematopoietic stem cells into thrombocytes (the fish equivalent of platelets).<sup>7</sup> Blood cell differentiation in zebrafish and mammals is similar,

but fish cells are easier to collect, separate and analyse.

The study revealed that gene expression shifts continually as cells pass along a pathway from stem cell to differentiated cell. Five milestones or developmental stages could be discerned along this pathway, characterised by the expression of distinct clusters of genes.

but existing data-cleaning methods often set arbitrary thresholds that could exclude viable cells. To overcome these challenges, Sanger researchers have developed computational methods to identify low-quality cells from raw scRNA-seq data.<sup>8</sup> Using data from thousands of visually assessed cells, the team developed a machine-learning algorithm to identify data features associated with low-quality cells. Around 20 data features were highly predictive of low-quality cells, allowing data to be automatically filtered out. Notably, the technique also identified cells that looked normal but were, in fact, compromised. The new approach was more accurate than existing methods and could readily be adopted by single-cell researchers worldwide.

The success of TraCeR – see 'Discovering how T cells respond to an infection' – illustrates how computational methods can extract useful biological information

from scRNA-seq data. A further computational innovation developed by Sanger researchers provides additional insight into the dynamics of gene expression: the 'D3E' tool.<sup>9</sup>

RNA-seq is often used to identify genes whose expression differs under two experimental conditions. This has traditionally been carried out on groups of cells, so comparisons of gene expression are averaged values across each cellular community. Unfortunately, this averaging effect could mask important biological patterns – in particular, variability in gene expression across cell populations. D3E analyses scRNA-seq data to map this variability and expose the previously hidden patterns, providing biological insights into the dynamics of gene expression.





# Boiling oceans of data for genomic gold

## In this section

20 Inflammatory bowel disease subtypes revealed by genomics

21 Applying UK10K and 1000 Genomes data yields new discoveries

22 Finding the BLUEPRINT for blood cell development

22 Rare mutations raise schizophrenia risk

23 Genetic roots of baby heart disease found

23 Man's explosive history hidden in the Y chromosome

As human genome sequence data accumulate at an ever-faster rate, Sanger scientists are at the forefront of efforts to identify genetic variants that affect our health – a key step towards understanding the mechanisms of disease and paving the way to new approaches to treatment. Although genetic factors make an important contribution to most common diseases, they do so in complex ways. Disentangling the knotty relationship between genes and disease calls for large-scale sequencing of both healthy and affected populations, international collaboration to pool data and resources, and new analytic tools – all areas in which the Sanger Institute is a world leader.

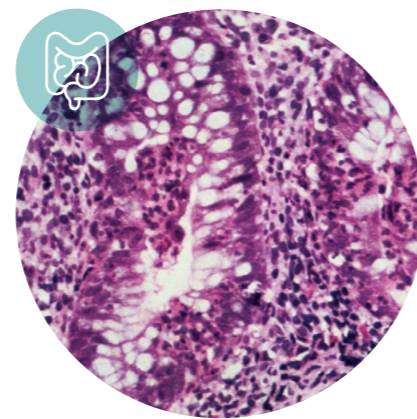
## Genomic scalpel dissects bowel disease

Over the past decade, the Human Genetics Programme has employed genome-wide association studies to identify dozens of genetic factors linked to common diseases whose roots are difficult to discover through other forms of enquiry – from neurological conditions to metabolic diseases. Sanger-led collaborations proved the continuing value of this approach by using more refined sweeps of larger numbers of patients to generate new insights.

For example, genetic studies of inflammatory bowel disease (IBD) and its

two main constituent conditions, Crohn's disease and ulcerative colitis, led by Sanger scientists and other research groups have identified 163 susceptibility loci for IBD, offering insights into the mechanisms of disease. Many of these loci increase the risk of both Crohn's disease and ulcerative colitis. Yet these two conditions display some differences in symptoms and progression, suggesting that there must be differences in their biology.

To explore this possibility, a Sanger-led consortium analysed data relating to more than 150,000 genetic variants in nearly 35,000 comprehensively assessed patients



### Inflammatory bowel disease

Crohn's disease can be genomically dissected into two subtypes: colon and ileum. They are as genetically different from each other as they are from ulcerative colitis

"We have been able to identify several biological pathways that likely play a role in primary sclerosing cholangitis... this gives pharmaceutical companies insight into biological systems to target."

Dr Carl Anderson  
Group leader at the Sanger Institute

from 49 centres in 16 countries.<sup>1</sup> The analysis suggested that IBD represents a range of disease, made up of three genomically characterisable types, not two. The researchers discovered that Crohn's disease can be genetically dissected into two subtypes, affecting the colon and the ileum, that are as genetically distinct from each other as each is from ulcerative colitis. These findings may help to explain why some treatments are more effective in some patients than others.

A separate study also used genome-wide association to clarify the relationship between IBD and a rare autoimmune condition that affects the liver – primary sclerosing cholangitis – a chronic progressive disease of the bile ducts, which can lead to liver cirrhosis and liver failure. Almost three-quarters of people with primary sclerosing cholangitis also develop IBD, suggesting that the two conditions may be intimately linked.

The international collaboration led by Sanger researchers carried out the largest-ever genome-wide association study of primary sclerosing cholangitis, identifying nine new susceptibility loci and more than doubling the total number of known risk loci to 16.<sup>2</sup> Only half of the newly identified regions were shared with IBD, indicating that primary sclerosing cholangitis is a distinct disease. The new loci also hint at biological processes that may be abnormal in the condition, suggesting possible new avenues for therapy development.

The study identified a further 33 loci increasing the risk of a range of inflammatory conditions that tend to co-occur with primary sclerosing cholangitis. These may reflect common underlying mechanisms of disease, and raise the possibility that drugs used to treat these other conditions could be beneficial in the disease.



## Statistical recycling powers new discoveries

While genome-wide association studies have been highly successful in identifying risk loci, several major challenges remain. Importantly, while such studies identify regions of the genome influencing risk of disease, further research is often required to pinpoint the specific genetic feature – the 'causal variant' – underlying this increased risk. In addition, they are particularly helpful in identifying risk loci that feature relatively commonly in the population but only make a minor contribution to disease development. Sanger researchers have developed computational tools that can predict (or 'impute') missing sequence

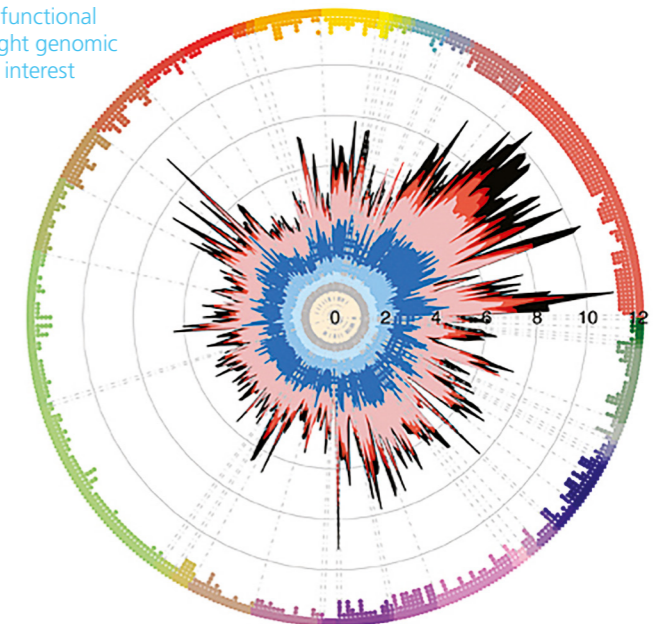
data, based on data from large-scale genome-sequencing projects such as the UK10K Project and 1000 Genomes Project. The result is that a much larger number of variants can be tested for association with disease, greatly increasing the research study's power.

As a result, a Sanger-led collaboration was able to analyse associations between 17 million genetic variations and 20 traits relevant to cardiometabolic disease – identifying 17 new associations, including six rare or low-frequency alleles.<sup>3</sup>

Importantly, the new tools also enabled regions of association to be narrowed down significantly, aiding the hunt for causal variants. In several cases, this narrowing down revealed strong candidates for causal variants with a potential role in disease processes.

### GARFIELD in action

GARFIELD (GWAS analysis of regulatory or functional information enrichment with LD correction) is a method that combines genome-wide association data with regulatory functional annotations to highlight genomic features of particular interest



**Finding needles in genetic haystacks**

30 million genetic variants were analysed for association with 36 blood-cell properties in more than 170,000 participants to identify 2,700 potentially causative variants

**Revealing blood cells' BLUEPRINTS**

The importance of rare variants is further emphasised by a major study, funded through the EU's €30m BLUEPRINT initiative, led by Sanger scientists as part of the International Human Epigenome Consortium (IHEC). A huge genome-wide association study published in *Cell* in 2016 tested 30 million genetic variants for association with 36 properties of red blood cells, white blood cells and platelets in more than 170,000 participants – identifying 2,700 variants affecting blood cell phenotypes, including hundreds of low-frequency and rare variants.<sup>4</sup> Crucially, these less-common variants had greater impact than the more-common variants identified previously.

The work has provided a wealth of data on potential biological mechanisms underlying increased risk of a host of

cardiometabolic diseases, as well other conditions that blood cells may be contributing to such as autoimmune diseases and even schizophrenia.<sup>3</sup>

In a related IHEC study, published at the same time in *Cell*, Sanger scientists led efforts to identify how genetic variation affecting gene control regions might influence risk of disease.<sup>5</sup> Focusing on three types of immune cell, the study mapped positions in the genome where genetic sequence variation or epigenetic modification (DNA methylation or histone modification) affects levels of gene expression. The study was also able to map 345 loci that have been implicated in immune-related disease onto this atlas of variation, suggesting routes by which they might be influencing immune cell biology and increasing the risk of diseases such as type 1 diabetes and multiple sclerosis.<sup>6</sup>

**Genomics reveals biology of mental health**

The underlying biology of mental health is difficult to discern. Sanger scientists sequenced the coding regions (exomes) of nearly 2,000 people with schizophrenia, and combined the information with data from more than 2,500 previously sequenced patients, to compare with the genes of unaffected people.<sup>6</sup> The approach revealed rare mutations in a gene, *SETD1A*, that increased the risk of developing schizophrenia 35-fold. These variations were also associated with increased risk of neurodevelopmental disorders.

Mutations in *SETD1A* are only found in 1 in 1,000 patients with schizophrenia, but the discovery provides new insight into the possible origins of schizophrenia. *SETD1A* codes for an enzyme that methylates histones, adding to growing evidence that abnormalities in chromatin modification play a significant role in schizophrenia.<sup>7</sup>

“The results were surprising, not only that we found such a high level of certainty that the *SETD1A* gene was involved, but also that the effects of the gene were so large. This is a really exciting finding for research into schizophrenia.”

**Dr Jeff Barrett**  
Group leader at the Sanger Institute

**Genomics gets to the heart of the matter**

Extensive sequencing has also provided new insight into congenital heart disease, which affects 1 in 100 newborns. In approximately 90 per cent of affected babies, only the heart is abnormal (non-syndromic); the remaining 10 per cent have additional abnormalities (syndromic).

Published in *Nature Genetics*, a Sanger-led study sequenced nearly 2,000 congenital heart disease patients and their parents, identifying three genes that had not previously been associated with the disease and uncovering important differences between syndromic and non-syndromic cases.<sup>7</sup> The researchers found that heart abnormalities in children with syndromic disease were typically caused by new mutations not seen in parents, but children with non-syndromic disease often inherited faulty genes from their apparently unaffected parents. These findings will help doctors advise parents of a child with congenital heart disease on their risks of having further affected children.<sup>8</sup>

“This is the first study to quantify the role that rare inherited variants play in non-syndromic congenital heart disease, and is extremely valuable as these patients make up 90 per cent of congenital heart disease patients worldwide.”

**Dr Mathew Hurles**  
Head of the Human Genetics Programme and Group leader at the Sanger Institute

**References**

1. Cleyne J *et al. Lancet*. 2016; 387: 156–67.
2. Ji SG *et al. Nat Genet*. 2017; 49: 269–73.
3. Iotchkova V *et al. Nat Genet*. 2016; 48: 1303–12.
4. Astle WJ *et al. Cell*. 2016; 167: 1415–29.e19.
5. Chen L *et al. Cell*. 2016; 167: 1398–414.e24.
6. Singh T *et al. Nat Neurosci*. 2016; 19: 571–7.
7. Sifrim A *et al. Nat Genet*. 2016; 48: 1060–5.
8. Malaspina AS *et al. Nature*. 2016; 538: 207–14.
9. Bergström A *et al. Curr Biol*. 2016; 26: 809–13.
10. Poznik GD *et al. Nat Genet*. 2016; 48: 593–9.
11. Schiffels S *et al. Nat Commun*. 2016; 7: 10408.

**Ancient Britons**

Comparing ancient genomes with hundreds of modern European genomes reveals that approximately 38 per cent of the ancestors of the English were Anglo-Saxons

**Tales of human evolution**

As well as risk of disease, human genome sequence data can also illuminate human history. For example, Sanger researchers have been involved in landmark studies into the evolutionary origins of the aboriginal populations of Australia.

By sequencing 83 genomes from Aboriginal Australians and 25 genomes from Papuans from the highlands of New Guinea, the Sanger-led study reported in *Nature* that the populations had split off from other human groups approximately 50,000 years ago, following a single ‘Out of Africa’ migration around 60–70,000 years ago.<sup>8</sup> Hence the populations had become genetically independent remarkably early in human evolution. By contrast, Asian and European populations diverged only about 10,000 years ago.<sup>9</sup>

Sequencing of 13 Y chromosomes of Aboriginal Australians by Sanger scientists also suggested that the continent was settled some 50,000 years ago.<sup>9</sup> Unlike other studies, the work published in *Current Biology* found no evidence for an influx of Asian Y chromosomes into Australia 4–5,000 years ago.<sup>10</sup>

Wider studies of the Y chromosome, led by Sanger researchers, have revealed intriguing insights into human expansion at various points in history. An international collaborative study published in *Nature Genetics* examined 1,200 men from 26 populations, tracing all current Y chromosome lineages back to a single origin 190,000 years ago.<sup>10</sup> Strikingly, the data appeared to show the sudden explosive expansion of certain lineages

at particular times in history – 50–55,000 years ago in Europe and Asia, and 15,000 years ago in the Americas, as well as various expansions in sub-Saharan Africa, western Europe, south Asia and east Asia at various times between 4,000 and 8,000 years ago. It is possible that the later expansions were linked to advances in technology – such as wheeled transport, metal-working or weaponry – providing an advantage to particular groups of men.<sup>11</sup>

“Using whole-genome sequencing allowed us to assign DNA ancestry at extremely high resolution and accurately estimate the Anglo-Saxon mixture fraction for each individual.”

**Dr Richard Durbin**  
Group leader at the Sanger Institute

Finally, and closer to home, genome sequencing has provided insight into the peopling of the British Isles. Genome sequences from ancient skeletons from the late Iron Age (around 50 BC) and Anglo-Saxon era (500–700 AD), unearthed near Cambridge, revealed that the Anglo-Saxon immigrants were genetically very similar to the modern Dutch and Danish.<sup>11</sup> They contributed some 38 per cent of the DNA of modern populations from east England and slightly less, 30 per cent, of the genomes of the modern Welsh and Scottish.<sup>11</sup>





# Discovering the enemy's family secrets

The Infection Genomics Programme discovers how pathogens evolve, acquire drug resistance and nullify immune systems. It applies DNA sequencing and analysis on a global scale to track the spread of bacterial and viral strains to enable more effective health planning, and generates reference genomes for neglected tropical diseases to find potential treatments. Within the body itself, its researchers explore the diversity of, and relationships between, bacteria in the gut microbiome to develop novel methods to modify and improve gut health.

## In this section

24 Genomics changes UK clinical infection control procedures

25 Europe exported *Shigella dysenteriae* worldwide

25 Unexpected sources of disease discovered

26 Blood flukes and black fever genomes reveal pathogen evolution

27 Tracking infections in real time

27 Targeting river blindness worm through its genome

27 Shining a light on the gut's dark matter

## Genomes change NHS practices

In 2016 the value of genomic monitoring was powerfully demonstrated in the journal *Science*.<sup>1</sup> Sanger-led research revealed the global spread of a potentially deadly bacterium among people with cystic fibrosis (CF), leading UK NHS clinicians to change their infection control and prevention procedures, and alter the design of new CF facilities.

Increasing numbers of CF patients have been acquiring infections of *Mycobacterium abscessus*, an innately multidrug-resistant and difficult-to-treat bacterium that can trigger severe lung damage. It had been assumed that *M. abscessus* is acquired sporadically from environmental reservoirs, but sequencing of more than 1,000 isolates from 517 patients in the UK, US and Australia by the Sanger Institute revealed that most cases are highly related.

The results suggest that new strains of *M. abscessus* have evolved that can be transmitted from patient to patient, most likely via contaminated surfaces or through airborne particles.

Using a combination of cell-based and mouse models, the researchers showed

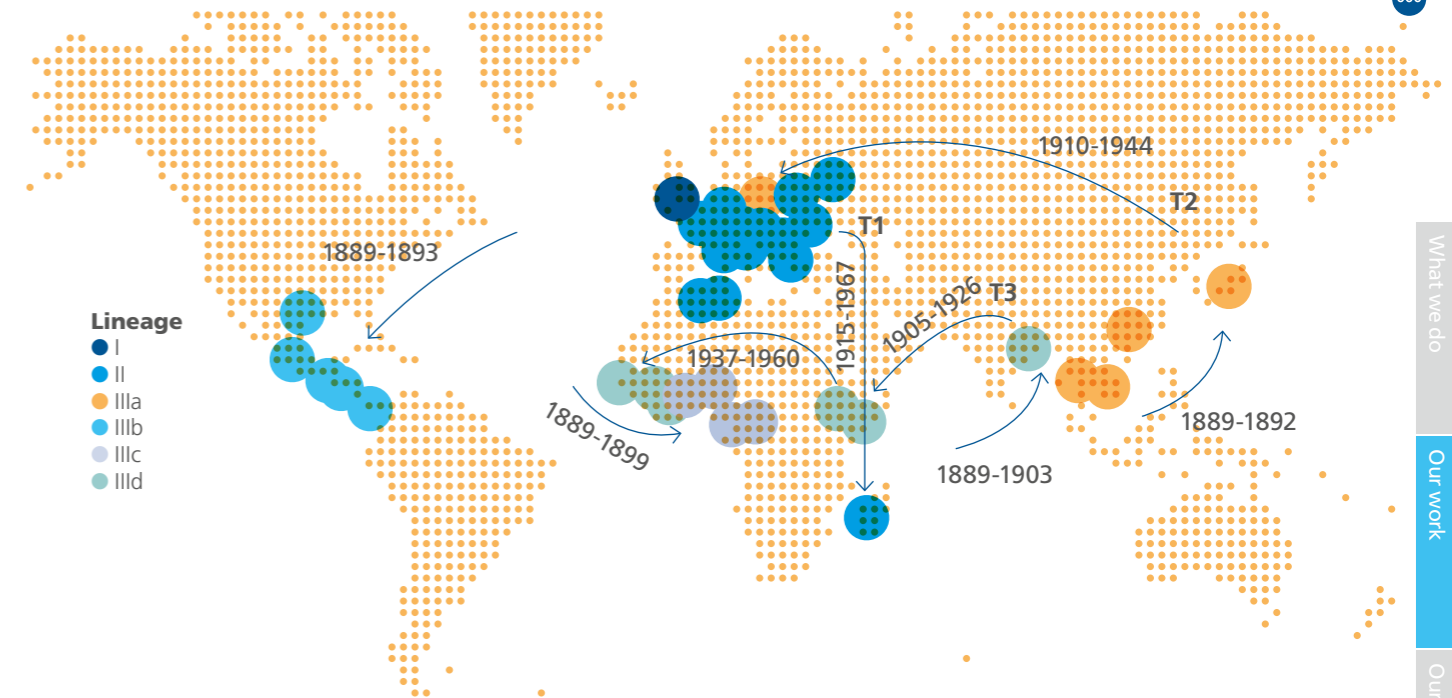
that the recently-evolved mycobacteria were more virulent, and likely to cause more serious disease.

Although it is a mystery how the new strains have travelled intercontinentally – transfer by healthy carriers is a possibility – the study highlighted important measures that could be used to minimise the risk of *M. abscessus* infection in CF clinics.

**“Now that we know the extent of the problem and are beginning to understand how the infection spreads, we can start to respond. Our work has already helped inform infection control policies and provides the means to monitor the effectiveness of these.”**

**Professor Julian Parkhill**  
Head of the Infection Genomics Programme at the Sanger Institute

## Geographic distribution and transmission patterns of *Shigella*



## Genetics reveals unexpected source of blindness

Analysis of genome sequence data can also reveal genetic changes affecting the biology of organisms – including their capacity to cause disease. For example, in work reported in *Nature Communications*, Sanger scientists' analysis of historical *Chlamydia trachomatis* samples in Australia revealed an unexpected cause of eye infections.<sup>3</sup>

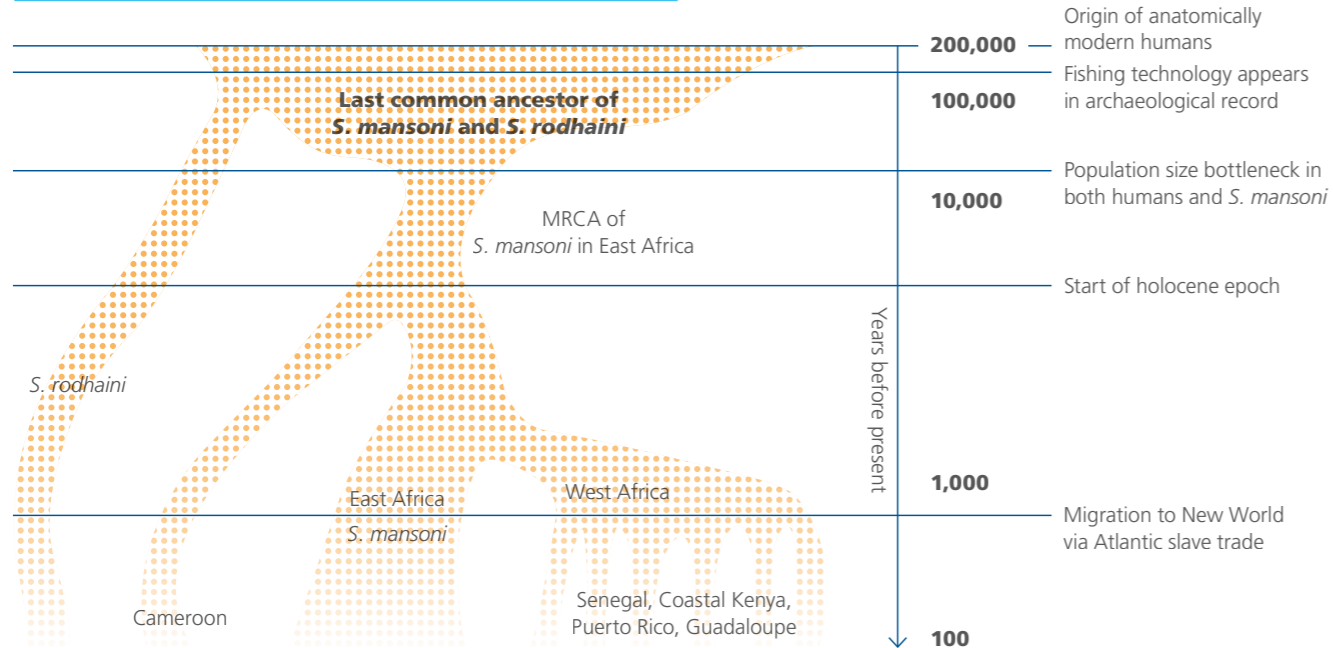
Different strains of *C. trachomatis* have specialised to live in different environments in the body. Some strains infect the eye, and repeated infection can lead to scarring and blindness – trachoma, a common and disabling condition in many low-resource countries.

In Australia, trachoma is mainly restricted to Aboriginal communities. Surprisingly, sequencing of *C. trachomatis* isolates collected in the 1980s and 1990s revealed that they were not 'classic' eye-infecting strains but urinary tract strains that had acquired genes coding for surface proteins required to infect the eye (*ompA* and *pmpEFGH*). The findings have significant implications for trachoma eradication programmes, as they suggest that urinary tract strains of *C. trachomatis* can evolve to infect the eye by capturing genes.

A further biological/geographical puzzle resolved in 2016 by Sanger researchers using genome sequencing was the anomalous behaviour of *Salmonella enterica* serovar Enteritidis (*S. Enteritidis*).<sup>4</sup> *S. Enteritidis* causes gastrointestinal upsets in most of the world, but it is a life-threatening blood infection in sub-Saharan Africa.

As reported in *Nature Genetics*, by sequencing and comparing 675 isolates from 45 countries, Sanger scientists found that *S. Enteritidis* has diverged significantly, with two African epidemic strains differing markedly from a global epidemic strain. African strains show distinct patterns of integrated prophage (bacterial viruses), loss of multiple genes, and gain of genes associated with antibiotic resistance.

The genetic changes are associated with adaptations to different environments. The global strain of *S. Enteritidis* has specialised for niches associated with intensive farming practices, and *S. Enteritidis* is generally acquired from contaminated foodstuffs. In Africa, it has acquired the capacity to overcome host defences in people in the region with impaired immune systems due to malnutrition, HIV infection or severe malaria.

Summary of *Schistosoma mansoni* population history

## Parasite genomes show effects of mankind's history

The blood fluke *Schistosoma mansoni* affects 250 million people. A Sanger-led consortium analysed 10 samples from around the world, and compared the *S. mansoni* genome with that of its rodent-infecting relative *Schistosoma rodhaini*. The results reveal fascinating details of its likely evolution alongside humans.<sup>5</sup>

Human and rat parasites diverged remarkably recently, less than 150,000 years ago, probably when early humans began fishing in East African lakes – the parasite lives in a pond snail for part of its life-cycle. Around 20–100,000 years ago, the parasite passed through a population bottleneck, possibly linked to climate change drying up lakes, or drops in human population sizes, before recovering. Notably, during the 16th to 19th centuries, new parasite populations were established in the New World – probably carried there through the slave trade.

Genome comparisons also revealed changes that may have enabled the parasite to adapt to life in humans. Modifications to the *VAL21* gene may protect *S. mansoni* from host immune

responses, while changes to an elastase gene may have helped the parasite to penetrate human skin and enter the body. It is hoped that this knowledge might offer new opportunities to develop preventive and therapeutic measures.

## DDT end linked to black fever resurgence

The protozoan parasite *Leishmania* is the second most deadly parasite after malaria, affecting nearly 300,000 people a year and killing up to 50,000. It affects the skin and internal organs, causing a condition known in the Indian subcontinent as *kala-azar* ('black fever'). In 2016 Sanger scientists published in *eLife* the results of their work with local researchers to sequence 200 isolates from across South Asia to explore how *Leishmania* has evolved over the past half century.<sup>6</sup>

The findings are consistent with a population bottleneck in the 1960s, at the time of effective vector control using DDT, and subsequent resurgence in the 1980s and 1990s when DDT use was stopped. DNA analysis also revealed a two-base-pair change associated with resistance to antimonial drugs, until recently the

standard treatment for *Leishmania* infections. This genetic change, in the *LdAQP1* gene, affects a transporter that imported antimonials into parasite cells. Looking forward, the new sequences will help researchers develop new genomic surveillance tools to track *Leishmania*, supporting regional attempts to eradicate the parasite.

## References

1. Bryant JM *et al. Science*. 2016; 354: 751–7.
2. Njamkepo E *et al. Nat Microbiol*. 2016; 1: 16027.
3. Andersson P *et al. Nat Commun*. 2016; 7: 10688.
4. Feasey NA *et al. Nat Genet*. 2016; 48: 1211–7.
5. Crellen T *et al. Sci Rep*. 2016; 6: 20954.
6. Imamura H *et al. eLife*. 2016; 5. pii: e12613.
7. Aanensen DM *et al. MBio*. 2016; 7. pii: e00444-16.
8. Argimón S *et al. Microbial Genomics*. 2016; 2. e000093.
9. Cotton JA *et al. Nat Microbiol*. 2016; 2: 16216.
10. Browne HP *et al. Nature*. 2016; 533: 543–6.

## Real-time tracking of epidemics

The ease with which the genomes of bacterial isolates can be sequenced is opening up new opportunities for 'genomic surveillance' – identifying and monitoring the spread of new strains of infectious disease.

For example, in 2016 Institute researchers worked with scientists and clinicians from 450 hospitals in 25 countries across Europe to show how sequencing of routinely collected MRSA isolates can reveal how antibiotic-resistant strains are evolving and spreading across the continent.<sup>7</sup> This study, published in *MBio*, relied on a free, easy-use web resource – Microreact.org. It was developed by the Centre for Genomic Pathogen Surveillance,

a partnership between the Sanger Institute and Imperial College London.

Microreact.org, allows researchers to upload evolutionary trees generated by sequencing projects, along with associated metadata such as date and place of collection of isolates.<sup>8</sup> Other researchers, clinicians and epidemiologists can then use the tool to see how strains are spreading and changing over time.

This information can then be shared via a permanent web link which can be published in a research paper. By putting information on Microreact, researchers can ensure that the data lives on – and in a format that other researchers can use as a basis for comparison or learning.

**New tools:** Microreact.org is a free web-based visualisation tool from the Centre for Genomic Pathogen Surveillance



The journal *Microbial Genomics* has been so impressed by the system that it is encouraging researchers to make data from prospective publications available through Microreact whenever possible.

## River blindness worm's genome offers clues to its own destruction

Research by Sanger scientists has revealed key differences in the biology of two related worms that can dwell side-by-side in the same person. In *Nature Microbiology*, Sanger researchers published the first complete genome of the parasitic worm *Onchocerca volvulus*.<sup>9</sup> Infecting some 17 million people, the parasite's larvae can migrate to, and damage, the eye resulting in river blindness.

Although an effective drug is available to treat *O. volvulus* – ivermectin – there is a risk that the parasite will develop resistance. Furthermore, the drug cannot be used in populations that are infected with a related parasite, *Loa loa*. This evolutionary near neighbour is also killed by ivermectin, but its death can trigger a potentially fatal inflammatory response in humans. Hence new drugs are urgently required.

Comparisons of the genomes of the two worms have shed light on key differences in biological pathways due to the presence of a commensal bacterium, *Wolbachia*, found in *O. volvulus* but not *L. loa*. *O. volvulus* has lost multiple genes in comparison with *L. loa*, relying instead on metabolic 'support services' provided by *Wolbachia*. These differences offer important opportunities to develop drugs that could target *O. volvulus* and leave *L. loa* untouched.



"Microreact takes disease tracking out of the hands of a privileged few and gives it to everyone who wants to understand disease evolution."

**Dr David Aanensen**  
Director of the Centre for Genomic Pathogen Surveillance and Group leader at the Sanger Institute

## Culturing the unculturable from the gut microbiome

Finally, returning to bacteria, a landmark study in *Nature* has overturned thinking about the populations of bacteria that live in the gut, the gut microbiota.<sup>10</sup> Conventional wisdom holds that most of these bacteria cannot be cultured, limiting detailed study of their biology. Sanger researchers have demonstrated that, in fact, many can be cultured, and have established a high-throughput pipeline for culturing, sequencing and analysing them.

The team successfully cultured 137 species of gut bacteria. Among several notable findings was the unexpected discovery that one-third of species can produce spores – suggesting that gut bacteria, which are killed by oxygen, may survive in the environment and spread between people in spore form. De-sporulation was triggered by certain bile acids when the spores re-entered the digestive tract.

This discovery and the new pipeline provide a valuable resource for researchers, opening up the study of the bacterial populations of our gut – which have been implicated in multiple aspects of health and disease, from obesity to mental health. This work has also helped to lay the foundations for a new spin-out company – Microbiotica.



# Hitting a moving target

## In this section

28 Malaria parasite evolution results in complete treatment failure

29 Evolutionary cousins shed light on human malaria

29 From gene to function

## Mapping antimalarial drug resistance

To identify the genetic changes underlying antimalarial drug resistance, the Sanger Institute draws upon its extensive partnerships across South-East Asia – a hotbed for drug resistance – to gather hundreds of samples of *Plasmodium falciparum*, the most common and deadliest human malaria parasite. In 2015, sequencing this international landscape of parasite genomes revealed mutations conferring resistance to artemisinin, a key component of artemisinin combination therapy (ACT), the front-line treatment for malaria.<sup>1</sup>

In 2016 the Malaria Programme identified mutations associated with resistance to another ACT drug, piperazine, which has led to complete treatment failure in Cambodia.<sup>2</sup> Sanger scientists analysed multiple samples from across Cambodia, finding that extra copies of genes coding for two proteins of the plasmepsin family and a mutation on chromosome 13 conferred resistance to piperazine, providing vital clues to the biological processes involved. In addition, identifying the genetic changes underlying resistance to artemisinin and piperazine will enable

The Sanger Institute investigates how millions of years of evolution and decades of antimalarial drug use have shaped the malaria parasite's genome – providing new leads for drug development and enhanced tools to track resistance at the local, national and international level.

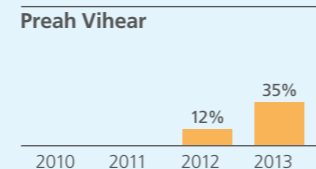
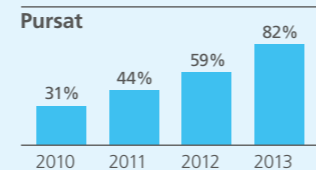
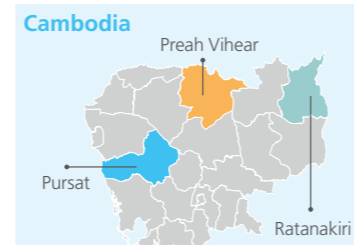
local health authorities to monitor the spread of resistance and recommend alternative treatments when necessary.

Genomic analysis is also providing insights into a human malarial parasite that is notoriously difficult to work with. *P. vivax* is extremely difficult to grow in laboratories and is found only in low numbers in patients. In one of the largest genomic studies of this species to date, Sanger researchers analysed *P. vivax* genomes taken from throughout South-East Asia.<sup>3</sup> They found that the parasite is evolving differently in Thailand, Cambodia and Indonesia in response to the specific mix of drugs used in these countries to treat *P. falciparum*. This knowledge has shed new light on the biology of *P. vivax* and will enable researchers to develop genomic tools to track the spread of resistance.

**“The emergence of piperazine resistance in Cambodian parasites has led to complete treatment failure there. This will threaten global attempts to eliminate malaria.”**

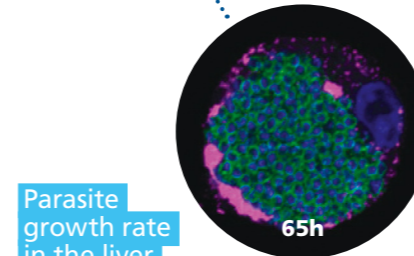
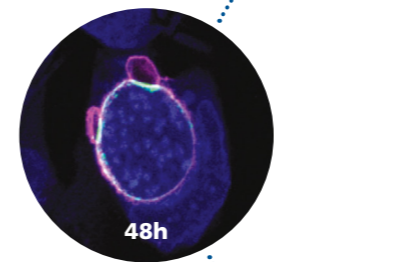
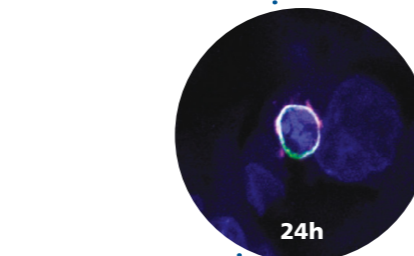
**Dr Roberto Amato**  
Senior staff scientist at the Sanger Institute

## Mutant allele frequencies in *exo-E415G*



Coloured bars indicate the mutant allele frequencies in each of the three provinces over time (no samples were available from Preah Vihear in 2010).

**212m**  
people worldwide infected with malaria in 2015 and there were roughly 429,000 deaths



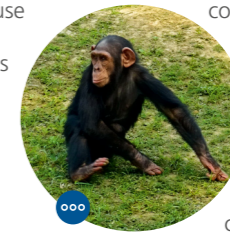
Parasite growth rate in the liver

**Revealed:** Gene knock-out techniques revealed that parasites lacking *DHHC3* struggled to grow *in vivo*, but were practically unaffected *in vitro* (shown above)



## Chimp and Gorilla malaria reveal human malaria's secrets

Studying the genomes of *P. falciparum* and *P. vivax* gives vital insights into their evolution in response to drug use over recent decades, but comparing malaria parasites from different host species reveals how parasites have adapted to their hosts over millennia – including the genomic changes that have enabled *Plasmodium* to infect humans.



In 2010, Beatrice Hahn at the University of Pennsylvania, US, and Sanger researchers compared the genomes of malarial species from wild living apes, finding that *P. falciparum* probably evolved from a gorilla-infecting species.<sup>4</sup> Building on this work, in 2016 they sequenced the genomes of two *Plasmodium* species from chimpanzees and compared them with other primate and non-primate malaria parasites.<sup>5</sup>

The work revealed that a gene family coding for proteins involved in remodelling of red blood cells has undergone a remarkable expansion in primate-infecting species. It also discovered that the forerunner of *P. falciparum* in the gorilla acquired a fragment of DNA containing two genes vital for invading human cells – potentially, a key step in the evolution of a human-infecting parasite.

## Using biological models to explore genomics

As well as large-scale genomic studies, Sanger researchers explore the function of individual genes to understand their biological role and their potential as therapeutic targets.

Highly variable regions in malaria parasite genomes point to genes that play a vital role in helping the organism evade host defences or gain entry into host cells. Sanger researchers found that two genes coding for parasite surface proteins *DBLMSP* and *DBLMSP2* show very high levels of genetic diversity, suggesting these proteins are targeted by the host's immune system.<sup>6</sup> Importantly, despite their great diversity, all the variants of these proteins bind to immunoglobulin M, which may provide a protective coat against the host's immune response.

The malaria parasite must adapt its biology to live in both its host species and its mosquito vector. One way it does this is by reversibly adding the fatty acid palmitic acid to many of its proteins. Sanger researchers have found that one of the enzymes responsible for this modification, *DHHC3*, plays a key role in parasite motility in both mosquitoes and humans.<sup>7</sup> Parasites lacking *DHHC3* were significantly less motile and less infectious, suggesting that the enzyme or its substrates could be possible new drug targets.

Finally, the Malaria Programme has developed a new approach for studying the molecular basis of red blood cell invasion. Sanger scientists generated erythrocyte-like cells from mouse embryonic stem cells and knocked out genes coding for surface proteins to see which the parasite uses to gain entry.<sup>8</sup> This revealed that glycophorin C, known to be important in invasion of human cells by *P. falciparum*, is also used by the rodent malaria parasite, *Plasmodium berghei*, to infect mouse cells. The team is now looking to adapt the method for use with human stem cells, to study the invasion of human cells.

## References

- Miotto O *et al.* *Nat Genet.* 2015; 47: 226–34.
- Amato R *et al.* *Lancet Infect Dis.* 2017; 17:164–73.
- Pearson RD *et al.* *Nat Genet.* 2016; 48: 959–64.
- Liu W *et al.* *Nature.* 2010; 467: 420–5.
- Sundaraman SA *et al.* *Nat Commun.* 2016; 7: 11078.
- Crosnier C *et al.* *J Biol Chem.* 2016; 291: 14285–99.
- Hopp CS *et al.* *Cell Microbiol.* 2016; 18: 1625–41.
- Yiangou L *et al.* *PLoS One.* 2016; 11: e0158238.

# Our approach

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



### 32 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.



### 34 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.



### 36 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.



### 38 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.



### 40 Connections

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

3,750bn  
bases of DNA  
sequenced a day

What we do  
Our work  
Our approach  
Other information





The Sanger Institute is able to conduct genomic research at a scale that few in the world can match. Its unique mix of computational, experimental, cell-line, animal and sequencing facilities enable Sanger researchers to explore and answer biomedical questions that are impossible for other organisations.

# Delivering science at scale

## Home of "cutting-edge science" says UK Prime Minister

In November 2016 the two newest buildings on the Wellcome Genome Campus were opened by UK Prime Minister Theresa May. The £42 million Bridget Ogilvie Building and Biodata Innovation Centre provide a cornerstone in the shared vision and commitment of the Institute and the Wellcome Genome Campus to drive pioneering research, innovation and discussion in the field of genomes and biodata.

The Bridget Ogilvie Building houses the Sanger Institute's sequencing centre on its first and second floors. Named in honour of the Director of the Wellcome Trust who was instrumental in establishing the Sanger Institute, the building has been designed to optimise the Institute's DNA sequencing and genotyping pipelines. Providing such an efficient, flexible and streamlined facility is vital to enable the Institute's delivery of science at scale: each day its teams read 3,750 billion bases of DNA, the equivalent of one 'gold standard' (30x) human genome every 35 minutes.

On the ground floor is a secure facility to supply the sequencing operations for the 100,000 Genomes Project run by Genomics England Limited. The Project will enable the NHS to become the first mainstream healthcare provider to offer genomic medicine as part of routine care.

Together, the two sequencing centres constitute one of the largest and most advanced sequencing facilities in the world.

The Biodata Innovation Centre focuses on driving innovation by nurturing small-scale computational ventures in the field of genomes and biodata. Its layout and services have been designed to support start-up companies whose technologies have enormous potential to deliver major advances in diagnostics, analysis and therapeutics.

Embedded into the fabric of the Wellcome Genome Campus both physically and intellectually, the Centre offers a well-resourced and scientifically stimulating environment for those wishing to translate and commercialise the fruits of basic research. So far companies from the US, Asia and Europe – including the Sanger Institute spin-outs Microbiotica and Congenica – have grasped this opportunity.

**"What I've seen on the Wellcome Genome Campus is an excellent example of research from across the UK and around the world coming together with commerce to deliver benefits for everybody including patients in the NHS."**

**Rt Hon Theresa May**  
UK Prime Minister



**World-class science:** Professor Sir Mike Stratton explains the power of the Institute's sequencing centre to UK Prime Minister Theresa May

**"We test hundreds of cancer drugs against cancer organoids growing in the laboratory. So, in the future, when we see a patient with the same genetic changes in their DNA we will know which drugs work and which ones don't."**

**Dr Mathew Garnett**  
Group leader at the Sanger Institute

## Organoids: adding a new dimension to cancer research

The Sanger Institute is at the forefront of creating and sharing cancer organoids – the next-generation of laboratory-based cellular models – to speed research and reduce duplication of effort. In July 2016, the Institute's Cancer, Ageing and Somatic Mutation Programme joined with Cancer Research UK, the National Cancer Institute (NCI) and the foundation Hubrecht Organoid Technology to launch the Human Cancer Models Initiative that will produce around 1,000 cancer organoids for researchers worldwide.

Until now, the majority of research into the drug sensitivity and biology of cancer has been conducted using cell lines grown in petri dishes. This approach has

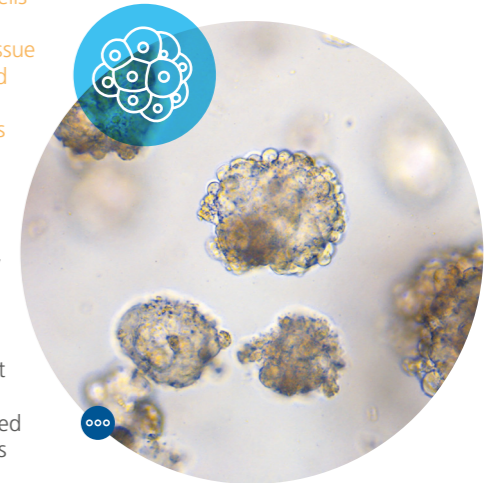
yielded many useful insights: Mathew Garnett's Genomics of Drug Sensitivity in Cancer project screened cell lines against hundreds of anti-cancer drugs. However, growing flat cultures of cell lines in a plastic-based environment means that key aspects of cancer development and biology may be missed and only the more 'aggressive' cell lines tend to thrive.

Mathew's team has joined with Cancer Research UK to build on its previous work to develop approximately 250 organoid cell models covering all stages of colon, oesophageal and pancreatic cancers. Individual tumour cells are grown in a 3D scaffold that allows them to self-organise into the different types of cell found in the original tumour. These 3D models more accurately reflect the genetic diversity, physical hallmarks, and cellular interactions of tumours in patients' bodies.

The team's organoids will be supplied with a complete history of the cells' origins, including the type of cancer

### Organoids

These balls of cells more closely resemble the tissue architecture and complexity of human tumours



and how advanced it was, the catalogue of genetic faults in the cells' DNA, and how the original tumour responded to treatment.

By enabling a broader range of cells, cancer stages and subtypes to grow and be stored, the team hope to produce valuable models of tumours that have been particularly hard to culture, such as oesophageal cancers. These models will allow study of many aspects of cancer's cellular biology, how tumours progress, drug resistance and allow the development of 'precision medicine' treatments.

## Delivering science at scale for European biobanks

New technologies and approaches open unexplored avenues of scientific enquiry. However, the fruits of this work cannot be shared across the research community due to a lack of established networks, coordination and access agreements. The result can be duplicated work, longer study timeframes, lack of reproducibility and erroneous data generation.

To prevent such issues hindering research using Human induced Pluripotent Stem Cells (HiPSCs), the Sanger Institute's Cellular Generation and Phenotyping (CGaP) core facility is playing a vital role in providing genomically and phenotypically characterised HiPSCs to secure, robust and sustainable supply chains. It is central to the Institute's efforts to deliver high-quality cell lines for the Human induced Pluripotent Stem Cell Initiative (HiPSCI) and to the European Bank for induced Pluripotent Stem Cells (EBiSC).

In March 2016 the first fruits of the Institute's involvement with EBiSC were shared with the worldwide research community. The consortium's online catalogue provides academic and commercial scientists with well-characterised induced Pluripotent Stem Cells (iPSCs) for use in disease modelling and other forms of pre-clinical research.

**"Two of the Sanger Institute's fundamental principles are to produce great science and to share it with the global scientific community."**

**Dr Chris Kirton**  
Head of Cellular Operations  
at the Sanger Institute

Founded in 2014 with €35 million from the European Union's Innovative Medicines Initiative, the 31-member consortium is gathering, curating, storing, and distributing 10,000 cell lines to internationally agreed protocols and standards. In particular, it seeks to provide the full range of relevant cell lines to enable full scientific exploration by distributing:

- original cell lines from patients
- 'isogenic control' cell lines; where the defective gene has been corrected in the patient's genome by gene editing to give a control for the same genetic profile
- healthy control cell lines.

The CGaP facility works with Institute Faculty to deliver science at scale by providing a central cell biology support service that can scale up and automate researchers' protocols. To supply EBiSC, CGaP draws on the cell lines from its work for HiPSCI.





The genomics revolution is producing discoveries and technologies that can be applied to real-world problems. The algorithms, techniques and bacterial libraries developed by the Sanger Institute, along with the therapeutic targets its researchers identify, have the power to transform pharmaceutical development and clinical practice.

# Translating innovations into scientific success stories


## Congenica delivers genome interpretation platform for the NHS

One Sanger Institute spin-out company, Congenica, is already transforming the care and treatment of patients in the UK National Health Service and internationally. It was one of four companies selected by Genomics England Limited to provide clinical interpretation services for the 100,000 Genomes Project and, in 2016, was the first UK company to deliver diagnostic reports.

The company's genome analysis platform; Sapiaientia™, was used to detect clinically relevant variants in more than 35 per cent of the patients tested, far greater than the 25 per cent that had been seen in similar large-scale national genome projects. Dr Matt Hurles and Dr Richard Durbin are among the original founders of the

company which was born from the pioneering work of the Deciphering Developmental Disorders (DDD) project. Congenica's collaboration with the NHS and UK Genomic Medicine Centres has enabled it to create a platform that supports the clinical workflow. Sapiaientia™ brings together in one secure, web-based user interface all the tools required to enable clinicians to collaborate remotely with other members of diagnostic multidisciplinary teams, speeding patient diagnosis.

The company is also helping to develop a novel assay for prenatal genetic diagnostic screening. Sapiaientia™ is being used in the PAGE (Prenatal Assessment of Genomes and Exomes) project, which is funded by a Health Innovation Challenge Fund Award, to help produce the evidence base needed for NHS adoption of this approach.

In addition Congenica's platform is also being used for research into drug development. An established partnership with Belgian company, UCB, is enabling researchers to bridge genomics and drug development and bring much-needed treatments to patients with rare diseases. 

1 in 17  
develop a rare  
disease at some  
point in  
their lives

80%  
of rare diseases  
are genetic  
in origin


95%  
of rare  
diseases have  
no treatment



## Microbiotica focuses on the body's 'forgotten organ'

The gut microbiome is sometimes called the body's 'forgotten organ' because of the key role its bacteria play in health. Imbalances in this complex community have been linked with obesity, allergies, Parkinson's disease and inflammatory bowel disease. The latest Sanger Institute spin-out company, Microbiotica, was launched in December 2016 to develop diagnostics and therapeutics from this under-developed resource.

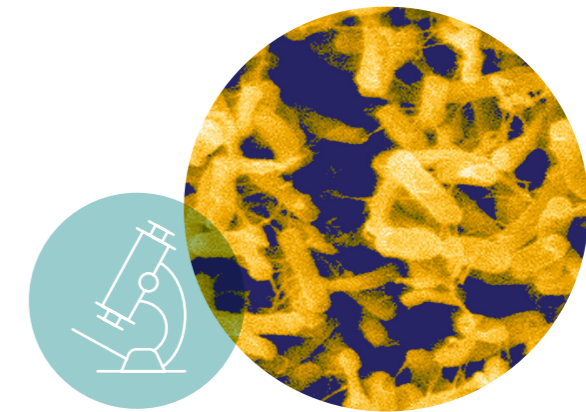
way Microbiotica will identify key differences in the constituency and prevalence of species to deliver robust clinical diagnoses.

In addition, the company is seeking to create live bacterial therapeutics; such as a pill containing a rationally selected and defined mix of bacteria that will successfully prevent reinfection with *Clostridium difficile* after antibiotic treatment. 



Dr Trevor Lawley and Professor Gordon Dougan have devised new methods to isolate gut bacteria and study their genomes, producing the world's largest culture collection and bioinformatics resource of the gut microbiome. This genomic database and sample library includes species from the gut's 'bacterial dark matter'; a large number of microbes that had previously defied culture and characterisation.

The company has unique access to these resources to identify disease-specific bacteria and stratify patients according to their intestinal communities. In this



"Our work at the Sanger Institute has shown that the human microbiome is important for health and disease, and is itself a therapeutic target."

**Dr Trevor Lawley**  
Group leader at the Sanger Institute and Chief Scientific Officer at Microbiotica


## Kymab set to start human clinical trials

Founded in 2010, Kymab is the Sanger Institute's longest-running and most successful spin-out venture so far, employing more than 130 people. In 2016 the company secured \$100 million series C financing, bringing its total funding to \$220 million. Its first human monoclonal antibody therapy will begin human clinical trials in 2017.

Kymab emerged from work of Professor Allan Bradley's laboratory that married embryonic stem cell technologies with genome engineering. Recognising the potential to develop a system to deliver novel human antibody-based therapeutics for the most challenging drug targets, the Institute worked with the Wellcome Trust to set up the spin-out company.

Building on the techniques developed at the Sanger Institute, Kymab researchers inserted almost 5.4 million DNA bases into the correct place in the mouse genome. The result was Kymouse™; a mouse that captures, in its engineered chromosomes, the entire diversity of the human immune system. This healthy mouse expresses

human, rather than mouse, antibodies and has the ability to generate highly selective and well-tolerated antibody therapeutics. So far Kymab has joined up with the Bill and Melinda Gates Foundation and Heptares Therapeutics to develop antibody-based medicines to fight infectious diseases, cancer and haematological disorders, as well as progress vaccine development for neglected diseases that affect populations in the developing world.

The company also works closely with a number of academic researchers worldwide, including a partnership with MD Anderson Cancer Center in Houston, Texas. 



### Genetic Engineering

Kymouse™'s chromosomes have been engineered to capture the entire diversity of the human immune system

"I am delighted that we have been able to take the first fruits of our basic research at the Sanger Institute and, with commercial support, develop it into fully-fledged technology."

**Professor Allan Bradley**  
Founder and Chief Technical Officer of Kymab, and Director Emeritus of the Sanger Institute



As the discoveries of genomic research impact all aspects of healthcare and society, our researchers are embracing new approaches to develop the necessary skills to span once-traditional divides of entrepreneurship, health service delivery and societal discourse.

# Building bridges across divides

## Cultivating an entrepreneurial spirit

For genomic research to reach its full potential and truly benefit healthcare, basic research findings must be carried by committed scientists from the bench to the bedside. The Sanger Institute is seeking to inspire in its researchers the same spirit that drove Sir Ernst Chain and Lord Howard Florey to translate Sir Alexander Fleming's curious finding in a culture plate into the first industrially produced antibiotic.

The Entrepreneurship and Innovation team was established in late 2016, and it will create an ecosystem of entrepreneurship across the Sanger Institute and Wellcome Genome Campus. Working on the principle that entrepreneurship is not a solo activity, it will seek to nurture networks of interested people for open dialogue and collaboration with external experts from industry and other research institutes and organisations.



To enable scientists to see the opportunities to translate their findings, the team will set up a range of activities to furnish researchers with the broad range of skills they need to carry a discovery through to commercial application. By providing engaging talks that highlight translational opportunities, coupled with challenge-driven events, scientists at all levels will be able to gain experience in identifying and validating ideas that could be developed for real-world application.

**“Entrepreneurship is a journey, not a destination. We help scientists develop skills that will help them to deliver real-world benefits, and enrich their academic research.”**

**Jo Mills**  
Entrepreneurship and Innovation Centre  
Manager, Wellcome Genome Campus



**“I was inspired by the stories of people who had started their own business, of the determination and hard work that is required to succeed.”**

**Alice Mann**  
PhD student at the Sanger Institute

## Students say YES to translation

In 2016 five Sanger Institute PhD students took part in the Biotechnology Young Entrepreneurs Scheme called Biotechnology YES. The competition, open to both PhD students and postdoctoral fellows, raises awareness of the processes involved in commercialising bioscience ideas. The competition aims to encourage an entrepreneurial culture among early-career researchers for the benefit of the UK economy.

The scientists took part in a three-day workshop with fellow biomedical researchers from other UK institutes to develop business plans for a hypothetical bioscience start-up company. During the first two mornings, the team was taken through many of the challenges and processes they would need to engage with to create a successful biomedical business plan.

These lectures were given by experts who had started their own businesses, patent attorneys and regulatory affairs officers. In the afternoons the researchers fleshed out the specifics of their plan while also benefiting from 1:1 mentoring sessions with financial directors, licensing specialists, GSK employees and university technology transfer officers.

On the third day, the team gave a ‘Dragons’ Den’ style presentation of their plan to ‘win’ equity investment.

The event has provided the students with many transferrable skills – from learning how to explain their science to a diverse range of audiences with differing levels of knowledge, to understanding the vital importance and benefits of teamwork.



The researchers found the experience invaluable and the group has cultivated a strong relationship with the Institute's Technology Transfer Office. They are now helping to develop ideas to encourage more scientists to engage with entrepreneurship.

## Listening to the Patients' Voice



setting, patient understanding of research and goodwill in agreeing to share their data is vital.

To enable meaningful dialogue and understanding between Sanger researchers and patients, the Public Engagement Team from Wellcome Genome Campus Connecting Science brought members of the charity Independent Cancer Patients' Voice together with Sanger cancer researchers. The result was a highly successful and thought-provoking day of collaboration and challenge as both groups shared their respective knowledge and experience of the disease.

To fully realise the power of genome-wide association studies and genomic comparison, Sanger scientists need access to vast numbers of biological samples, coupled with high-quality in-depth medical information. As genomics moves from a research environment into a clinical

After an introduction by the Public Engagement Team, Faculty member Mathew Garnett and senior scientist Haley Francies explained how patient samples provide the basis of the cell lines and organoids they use to screen cancer drugs against genome profiles. This is done to discover new potential treatments. These talks prompted lively discussion over lunch

where more members of the research team had the opportunity to learn about life with cancer and the concerns of patients.

Maggie Wilcox, President of the Independent Cancer Patients' Voice, summed up the day: “Our visit to the Sanger enhances our ability to provide educated and realistic patient involvement and we would be pleased to maintain our link in the future. We would be keen to assist in raising public awareness of cancer research and the need for informed lay involvement and donation of tissue and data.”

**“It was really good, everyone was helpful and friendly, and we learned a lot. We want to come again!”**

**Maggie Wilcox**  
President of the Independent Cancer  
Patients' Voice



From the Houses of Parliament to the Global Village, MPs, doctors and data scientists are wrestling with the legal, ethical, and technological challenges that the scale and speed of genomic research are posing. The Sanger Institute has been at the heart of these discussions, laying the foundations of a strong and vital future for genomic research, data sharing and clinical application.

# Building the future of genomic research



## Putting genomics on Parliament's agenda

For genomic research and its translation into health benefit to thrive requires a nurturing environment of strong funding, supportive regulatory frameworks and positive public opinion. To cultivate this consensus, the Sanger Institute informs health ministers and government bodies across the world.

The Sanger Institute has been working with UK parliamentarians to help them recognise the potential of new genomic technologies to benefit patients and to guide their adoption by the NHS.

In partnership with the PHG Foundation, the BMJ and Northern Health Science Alliance, the Sanger Institute set up, and guides, an All Party Parliamentary Group (APPG) on Personalised Medicine.

Although this cross-party group does not have an official status in Parliament, it provides a vital forum for political discussion and collaboration with experts on the future impact and potential of genomics.

The APPG on Personalised Medicine is made up of 24 members from all major political parties across both the House of Commons and the Lords. The politicians are keen to understand the opportunities that new genomic advances offer, and to identify and remove any barriers to their use due to complexity in the NHS or regulation.

In addition, the group will examine how data collection and data-sharing practices can be improved to allow genomic and clinical information to power future research. This is an area that the Institute is also exploring through the Global Alliance for Genomics and Health.

**“We are excited to be working with parliamentarians and our partners to explore the challenges of how to make personalised medicine a workable reality for the NHS, patients and the UK public.”**

**Dr Sarion Bowers**  
Research Policy Lead at the Sanger Institute



## Global Alliance seeks to share genomic riches

The wealth of genomic information being generated globally may never deliver on its promise unless access is opened up and simplified.

Large numbers of data sets are stored in incompatible formats and annotated using differing clinical and phenotypic terminologies. In 2013 the Sanger Institute partnered with the Broad Institute and the Ontario Institute for Cancer Research to co-host and co-fund a worldwide coalition dedicated to ensuring universal access to genomic information – the Global Alliance for Genomics and Health (GA4GH).

Since its inception, the Institute has been a leading contributor to the Alliance, identifying and disseminating best practices, and developing innovative technologies to accelerate genomic data sharing in areas including:

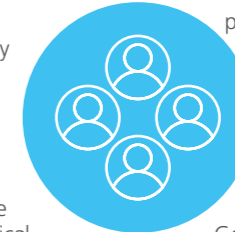
- Opening access to, and standardising analysis of, genomic data sets worldwide
- Linking phenotypic and clinical information to genomic data across healthcare systems throughout the globe
- Ensuring patient data privacy and security both nationally and internationally

- Harmonising international standards of consent, privacy and governance to enable data pooling.

Today the GA4GH comprises more than 400 organisations in more than 70 countries. The Alliance encompasses biomedical research institutions, healthcare providers, information technology and life science companies, funders of research, and disease and patient advocacy groups.

Many are participating in the Alliance's three pilots:

- **Beacon Project:** to enable discovery of genomic datasets across the internet. There are more than 70 beacons, including one at the Sanger Institute.
- **BRCA challenge:** an international collaboration to pool and curate information on *BRCA1* and *BRCA2* variants and clinical data for improved cancer risk assessment.
- **Matchmaker Exchange:** a federated network of databases across countries, including the Institute's DECIPHER, where clinicians, researchers and



patients can match rare genotypes and phenotypes through a single interface.

In 2016, the GA4GH's leaders announced their progress so far in the journal *Science*. The Alliance has developed the Genomics API to enable DNA data

providers and consumers to better share genomic information on a global scale, allowing disparate technology services to exchange genotypic and phenotypic data. It has also produced the Framework for Responsible Sharing of Genomic and Health-Related Data, outlining the basic principles and core elements for responsible data sharing.

But there is much to be done: from standardising computer readable consent forms to creating the flexibility and scalability needed to accommodate the unique data and terminology needs from around the world. The Sanger Institute continues to take a leading role in detangling these knotty problems.

## Capitalising on Cambridge's Biotech Cluster

Cambridge is at the leading edge of biomedical research and innovation and the Sanger Institute is ideally located to grasp the opportunities for collaboration that result. One such opportunity is the relocation of the Papworth Hospital, with its new Heart and Lung Research Institute (HLRI), to the Cambridge Biomedical Campus.

Papworth Hospital is the UK's largest specialist cardiothoracic hospital, and one of the largest hospitals of its type in Europe. As a result, its doctors treat some of the most challenging patients in the UK, meaning that the hospital's unique patient cohorts could prove especially fruitful to genomic enquiry.

To explore the possibilities and facilitate strategic collaboration, 12 senior directors, consultants and professors from Papworth Hospital and the HLRI met with 17 key

representatives from the Sanger Institute, including leading Faculty members, Associate Director Julia Wilson and Institute Director Mike Stratton.

The hospital's researchers outlined the specific clinical and translational questions they were seeking to answer to discover how the Sanger Institute's expertise and resources could help.

The joint workshop aimed to build on the success of the Sanger's and Papworth's existing collaborations and revealed a number of areas for beneficial collaboration that are being explored further. In addition, the HLRI's Interim Director will join Institute Faculty member Nicole Soranzo for a six-month research sabbatical to explore the potential of genomics to unlock the secrets of heart disease.



**“Our workshop with the Papworth Hospital and its new Heart and Lung Research Institute was a fantastic opportunity to explore how a strategic collaboration between two world-renowned institutes could work.”**

**Dr Julia Wilson**  
Associate Director, Sanger Institute



Almost all the research carried out by the Sanger Institute is facilitated by networks of knowledge and skills: within the Institute, across organisations on Campus, and globally. In the same fashion, Sanger researchers' findings and approaches drive further research endeavours worldwide through leveraging the power of the web to provide virtual networks and computational services.

# Using the power of networks



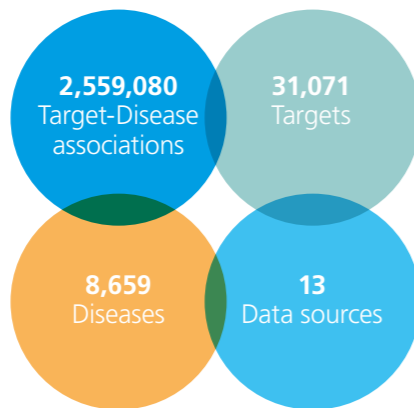
## Open sharing combines genetics and big data to speed drug development

Developing new drugs is a long and costly business, with a staggering attrition rate: it is estimated that up to 90 per cent of compounds entering clinical trials fail to become licensed medicines. To increase the success rate, in 2014 the Sanger Institute joined with EMBL-EBI and GSK to form an innovative partnership called Open Targets to speed new drug development.

The partnership works to integrate academic and industry knowledge, human and computer networks, and genomics and biodata in areas of mutual interest to determine how biologically valid newly identified therapeutic targets are. So far 30 experimental projects have been launched within the group and, in addition to the research papers that will be published over the next couple of years, data from the work are already being used with industry partners to prioritise new drugs.

However, the research is not a 'closed shop' that benefits only the partner organisations: all the work is precompetitive and is freely shared with the academic and industrial research community. Towards the end of 2015

### Online Target Validation Platform at the end of 2016



**"It is truly exciting to apply so many different areas of expertise, from cell biology to large-scale genome analysis, to the challenge of creating better medicines."**

**Dr Jeffrey Barrett**  
Director of Open Targets and Group leader at the Sanger Institute

the centre launched the online Target Validation Platform which weaves together many different types of data, including genetic associations, gene expression, literature mining and known drug targets.

The partnership is also open to new members and, in 2016, Biogen joined the group. The organisation is structured so that team working and knowledge sharing are built in to the working practices of the initiatives. All projects must have scientists from at least two partners and the centre regularly holds workshops on topics of mutual interest, along with scientific 'match making' to bring together researchers for maximal effect. Also, the partnership runs three 'integration days' each year to share all the teams' findings and discuss ideas for the future. To find out more, visit: [www.opentargets.org/](http://www.opentargets.org/)



**"The UMIC will enable storage and analysis of large datasets empowering African research programmes and institutions."**

**Professor Pantiano Kaleebu**  
Co-Director of the Uganda Medical Informatics Centre, Director of the MRC Uganda Unit, and Acting Director of the Uganda Virus Research Institute

## Building Sanger's African sibling

The flow of research is often inequitable, with much of the health-related data generated from population studies around the world being analysed and stored in American or European institutions. To redress this balance the Sanger Institute is committed to building genomic and biodata infrastructure and skills in low- and middle-income countries across the globe. In November 2016 this work took a major step forward with the official launch of the Uganda Medical Informatics Centre (UMIC) in Entebbe.

The high-throughput medical bioinformatics data centre significantly increases genomic and biomedical research capacity in sub-Saharan Africa by offering access to high-capacity servers that can store and analyse high-volume complex datasets. With genomic data on

3,000 individuals from across the continent it is one of the largest health research-orientated computational resources in Africa.

The centre has a computational capacity equivalent to 20 per cent of the Sanger Institute's data centre and can store up to 10,000 high-coverage full human genomes. By enabling the integration, curation and analyses of large-scale population health resources (including genomic, complex phenotypic and clinical data sets) it removes the need to send data to Europe for investigation.

The system is already supporting major research programmes including the International AIDS Vaccine Initiative (IAVI), TrypanoGEN (TPG), Makerere University/Uganda Virus Research Institute (UVRI) Infection and Immunity Research



Training Programme (MUII) and the MRC/UVRI Uganda Research Unit on AIDS (MRC/UVRI).

UMIC was built and managed by Sanger scientists in partnership with the Medical Research Council (MRC)-UVRI Research Unit, and the University of Cambridge. The project is a cornerstone of the African Partnership for Chronic Disease Research linking academic institutions and building academic-government and private partnerships across Africa.

## Travelling through time and space using the web

To truly understand the nature of infectious epidemics, researchers, healthcare workers and public health planners need to be able to visualise the transmission of a disease and the pathogen's evolution over time. Yet most of the information languishes in static publications or impenetrable databases. To break the data out of these confines and provide useful tools for scientists to share and 'move' through the information, the Sanger Institute partnered with Imperial College London to found the Centre for Genomic Pathogen Surveillance in 2014.

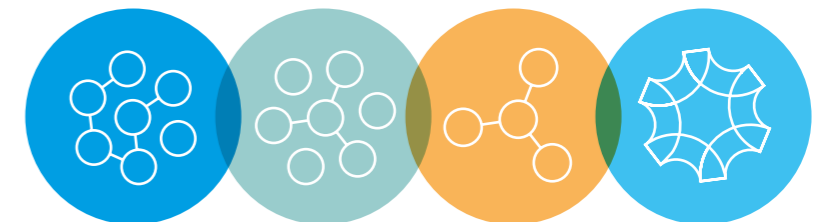
The Centre seeks to use the power of the internet to provide global access to online platforms that enable researchers to gather and visualise data related to antimicrobial resistance and genomic surveillance. Its key aims are to enable the identification of high-risk bacterial clones of public health importance, assess the level of risk they pose (in terms of resistance, virulence and transmissibility), and enable effective management.

Tools that the Centre offers include:

- **PhyloCanvas**: a completely online genetic relatedness tree viewer.
- **WGS** (Whole Genome Sequence Analysis): for processing bacterial genome sequences and visualising their evolutionary relationships and global distribution, with additional drug-resistance profile information.
- **Epicollect**: to enable swift data gathering using forms on mobile phones. The system was used for tracking Ebola in Liberia.
- **Microreact**: a real-time epidemic visualisation and tracking platform that allows researchers to move through the development of an outbreak over time.

**"Our study demonstrates the potential for combining whole-genome sequencing with internet-based visualisation tools to enable public health workers and doctors to see how an epidemic is spreading and make swift decisions to end it."**

**Dr David Aanensen**  
Director of the Centre for Genomic Pathogen Surveillance and Group leader at the Sanger Institute



## Other information

What we do

Our work

Our approach

Other information

## Image Credits

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

- Page 8** – Mosquito – Science Photo Library  
**Page 12** – Redhead woman – Getty Images  
**Page 13** – Acute Myeloid Leukaemia – Wellcome Images  
**Page 14** – Diagram adapted from Iorio F *et al. Cell* 2016; 166: 740–54.  
**Page 16** – PD-1 molecule – Protein Data Bank in Europe  
**Page 18** – sOPTiKD video images – Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute  
**Page 19** – Diagram adapted from Macaulay IC *et al. Cell Rep* 2016; 14: 966–77.  
**Page 20** – Crohn's disease – Wellcome Images  
**Page 21** – Crowd – KeithJJ, Pixabay  
**Page 22** – Red blood cells – Annie Cavanagh, Wellcome Images  
**Page 25** – Diagram adapted from Njamkepo E *et al. Nat Microbiol.* 2016; 1: 16027.  
**Page 26** – Diagram adapted from Crellen T *et al. Sci Rep.* 2016; 6: 20954.  
**Page 28** – Diagram adapted from Amato R *et al. Lancet Infect Dis.* 2017; 17: 164–72.  
**Page 29** – Malaria parasite growth – Hopp CS *et al. Cell Microbiol.* 2016; 18: 1625–41.  
**Page 29** – Chimpanzee – sarangib, Pixabay  
**Page 33** – Organoids – Marc van de Wetering, Hubrecht Institute for Developmental Biology  
**Page 35** – *Clostridium difficile* – Janice Carr, Public Health Image Library of the Centers for Disease Control and Prevention  
**Page 38** – Houses of Parliament – Nikki Gensert, Pixabay  
**Page 39** – Biomedical Campus – Cambridge University Hospitals NHS Foundation Trust  
**Page 40** – Pills – Sasha Dunaevski, Freeimages  
**Page 41** – UMIC launch event – UMIC

## Wellcome Trust Sanger Institute Highlights 2016/17

The Wellcome Trust 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 Trust Sanger Institute, 2017.

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

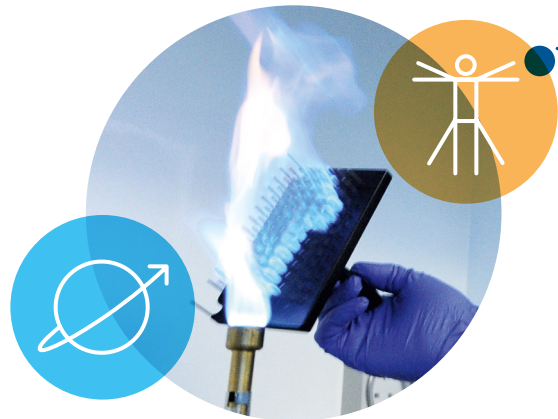
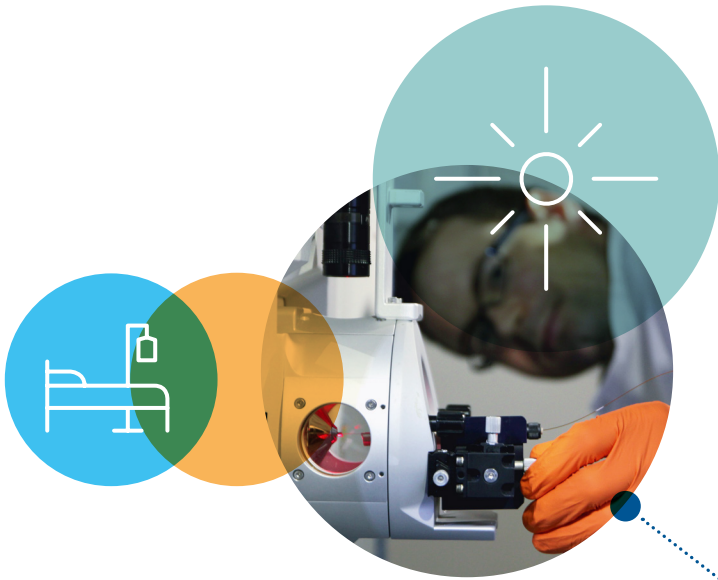


Printed by Park Communications on FSC® certified paper.

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

This document is printed on GenYouS, a paper containing 100% virgin fibre sourced from well-managed, responsible, FSC® certified forests.

Designed and produced by **CONRAN DESIGN GROUP**



Wellcome Trust Sanger Institute  
Tel: +44 (0)1223 834244  
[sanger.ac.uk](http://sanger.ac.uk)