Research impact:

Leading the world into the genomic era

Key contributions of the Wellcome Sanger Institute 1993-2023

Introduction

In the thirty years since our establishment, the Sanger Institute has been a world leader in pioneering genomics for fundamental discovery research, and in developing cutting edge genomic and bioinformatic resources that are relied upon by scientists globally. Evaluating the impact of this work is challenging given its breadth and scope. As such, we sought to develop an assessment framework that integrates quantitative indicators of impact wherever possible. However, metrics alone cannot convey the wide-ranging impact of our activities, which have dramatically altered how scientists view their field and how they tackle their research questions. This section, therefore, provides a narrative overview and qualitative assessment of our scientific contributions over thirty years. Intentionally brief, this is not an exhaustive list of all of the research we have conducted, but serves as a summary of many of our most influential initiatives and discoveries.

Contents

1. Providing fundamental resources for understanding biology: reference genomes

1.1 The reference human genome

Introduction

As the technologies underpinning DNA sequencing, genome mapping and data visualization improved through the 1980s, it became apparent to leaders in the field that with a coordinated effort, further technological improvement, and appropriate financial investment, it would be possible to sequence the complete human genome. John Sulston proposed to the Wellcome Trust that the UK should be a major contributor to this aspirational endeavour. They agreed, and the Sanger Centre at Hinxton was established in 1992 to deliver the UK and Wellcome's contribution to sequencing of the human genome.

Sanger Institute contributions

In February 2001, the Human Genome Sequencing Consortium, which consisted of sixteen institutions from six countries, published the first draft of the complete human genome.

Since then, the human genome has been iteratively improved to the very high quality complete sequence that exists today.

The Sanger Institute was the only UK partner in the consortium and contributed one third of the total sequence of three billion DNA base pairs.

In collaboration with the European Bioinformatics Institute (EMBL-EBI), the Ensembl database was established to host, organise, annotate, and display the human genome as a freely available resource.

Impact

The reference human genome sequence is widely considered to be among the most important milestones in the history of biology, and in the advancement of knowledge generally.

Sequencing of the human genome by a publicly funded consortium prevented its monopolisation and commercial exploitation by a single private company.

Working as an international consortium composed of large, coordinated teams of scientists and informaticians heralded a new scientific research culture and established a model for many subsequent research collaborations.

The principles articulated by the Human Genome Project and embodied in the Bermuda Principles for the rapid and public release of DNA sequence data, permanently transformed global practice in open data sharing.

The large volumes of data embodied in the reference human genome transformed computational, mathematical, and statistical biology, making irreversible and profound changes to the way that biology is studied.

The reference human genome transformed our understanding of human biology in an inestimable number of ways and is now underpins a large proportion of modern discovery and applied research relating to human beings.

It has enabled a huge shift in understanding of the biological basis of nearly all human diseases and has transformed the development of therapeutics and diagnostics.

The economic benefits resulting from biotechnological and pharmaceutical developments informed by the reference genome represent a substantial return on investment in the Human Genome Project.

1.2 Reference genomes of model organisms

Introduction

Our understanding of human biology in both health and disease has largely relied on the insights gained by studying non-human species as model organisms. The ability to determine which genes participate in which biological processes has enabled biological research to move beyond simple observation, to develop detailed insights into the underlying mechanisms at work. Having reference genome sequences for model organisms is vital to deciphering the molecular basis behind biological processes, and to relating these insights to human biology.

Sanger Institute contributions

We provided the reference genome sequences of most of the widely used model organisms:

- Yeast: *Saccharomyces cerevisiae* (1996), an experimental model organism that is widely used to study cellular biology and the functions and interactions of genes. The Sanger Institute made the single largest contribution to sequencing of what was the first complete genome from a eukaryotic organism.
- Bacterium: *Escherichia coli* (1997) is infectious gut bacteria and model organism that is used in molecular biology, biotechnology, and genetic engineering, critical for our understanding of bacterial biology and cellular processes such as DNA replication, transcription, and cellular metabolism.
- Nematode worm: *Caenorhabditis elegans* (1998) was the first reference genome of an animal and the first multicellular organism to be sequenced. Completed before the human genome, it helped scientists to interpret human genome sequence data and provided fundamental insights into the development of the nervous system and organs, and the process of programmed cell death.
- Mouse: *Mus musculus* (2002) has been the most widely used model organism for studying common diseases in humans owing to similarities in early development, reproduction, and basic organ systems.
- Zebrafish: *Danio rerio* (2013) is a model organism commonly used for studying embryonic development, neurology, and immunology.

Impact

These model organisms have been crucial experimental tools for understanding biology in general, and for providing insights relevant to human biology and gene function. They have been extensively used in the successful development of various types of therapeutic drugs and vaccines. Their use has had an inestimable impact on human health and provided economic benefits across many different fields. Their reference genome sequences have been central to these contributions.

1.3 Reference genomes of infectious microbes

Introduction

Infectious disease remains a major global cause of death, particularly in the young and in low- and middle-income countries. The identity of the microorganisms that cause most common human infectious diseases have been known for decades. To transform our understanding of their biology and identify new approaches to prevent and treat infections, researchers at the Sanger Institute (and other organisations) sought to generate reference genome sequences for the microbes behind the major human infectious diseases.

Sanger Institute contributions

1998	*Mycobacterium tuberculosis*, responsible for tuberculosis.
2000	*Neisseria meningitidis*, a cause of meningitis and other infections.
2000	*Mycobacterium leprae*, the bacterium causing leprosy in ancient civilisations, and still affecting hundreds of thousands of people today.
2000	*Campylobacter jejuni*, the leading cause of bacterial food-borne diarrhoeal disease worldwide, and the most frequent precursor to Guillain-Barré syndrome, a form of neuromuscular paralysis.
2001	*Salmonella* Typhi, the bacterium causing typhoid fever which kills hundreds of thousands of people each year.
2001	*Yersinia pestis,* the plague-causing bacterium, which killed 200 million people in the 14th century and became known as the Black Death. Drug resistant strains mean that plague is still a threat to human health today, with hundreds of cases reported every year.
2002	*Plasmodium falciparum*, the deadliest of the malaria-causing parasites that causes one million deaths each year, predominantly of young children in Africa.
2003	*Bordetella pertussis*, a bacterium causing whooping cough.
2003	*Corynebacterium diphtheria*, the toxin-producing bacterium which causes the potentially fatal infectious disease Diphtheria.
2004	Methicillin-resistant *Staphylococcus aureus* (MRSA), the bacterium causing a range of infections, including skin and wound infections, pneumonia, and bloodstream infections.
2005	*Trypanosoma brucei,* the protozoan causing sleeping sickness, a common and deadly disease in sub-Saharan Africa.
2005	*Entamoeba histolytica,* the protozoan causing amoebic dysentery, a significant cause of illness and death in low- and middle-income tropical countries.
2006	*Leishmania major,* the protozoan causing leishmaniasis, a spectrum of disfiguring and sometimes fatal diseases, affecting two million people in 100 countries each year.
2008	*Chlamydia trachomatis,* a bacterium commonly causing blindness and sexually transmitted infections.
2009	*Streptococcus pneumoniae*, a bacterium that can cause pneumonia, meningitis, and other infections.
2009	*Schistosoma mansoni,* the worm causing schistosomiasis, which affects 200 million people worldwide.
2013	Four *Cestoda* species of tapeworm, which cause diseases that can be fatal and are difficult to treat due to inefficient medication.

Impact

Reference genome sequences of these infectious disease causing microbes, and others, have transformed our understanding of their evolution and biology, and how they infect and cause disease in humans. Insights from the reference genomes have led to new therapeutics and vaccines, provided the basis for diagnostics tests, allowed us to determine how they spread and to track the evolution of antimicrobial and vaccine resistance. Collectively these reference genomes have had major impacts on the management of infectious diseases globally with many millions of lives affected.

1.4 Reference genomes of the Tree of Life

Introduction

Since the discovery of the structure of DNA and the invention of DNA sequencing technology, scientists have aspired to the notion that there should be a catalogue containing a reference genome sequence for all species on Earth. Advances in sequencing technologies and reductions in costs have recently enabled work towards this ambition to embark under the auspices of the Earth Biogenome Project.

Sanger Institute contributions

Together with a collaborative group of scientists, naturalists, and sample collectors, in 2019 the Sanger Institute established the Darwin Tree of Life project. This project will deliver high quality reference genomes of the ~70,000 known eukaryotic species of the UK and Ireland. This represents the largest such initiative in the world at present and, through the deep experience gained and problems solved, is leading the way for the Earth Biogenome Project to subsequently deliver at a global scale.

To date, more than 1,000 fully assembled reference genomes of animals, plants and fungi have been sequenced by the Darwin Tree of Life Project and deposited in public databases with production continuously accelerating.

Impact

Sequencing the genomes of all species will provide a comprehensive resource with which to study the evolution and biology of life on Earth. These genomes will provide the basis for deeper understanding of ecosystem and organismal diversity and how they change over time, notably in response to climate change and other influences. As such these genomes will provide information vital for humankind's efforts to manage life on Earth better in future than in the past. Furthermore, by providing access to the genes and proteins for every form of life, this project will help to accelerate the field of synthetic biology in previously unimaginable ways. Finally, there are likely to be major economic implications as investment in managing the environment, improving crops and many other yet unforeseen applications are developed using this information.

2. Powering cancer diagnosis and treatment through genomics

Introduction

All cancers are caused by somatic mutations, changes that occur over the course of an individual's lifetime in the copies of the human genome present in essentially every cell of the body. The advent of the reference human genome sequence provided the foundations for large-scale sequencing to detect all somatic mutations in cancer genomes. The Sanger Institute established its Cancer Genome Project in 2000 to systematically sequence cancer genomes and identify the "cancer genes", which, when mutated, cause normal cells to become cancerous and to understand the mutational processes behind these changes. This project was among the first cancer genomics initiatives in the world to be established.

Sanger Institute contributions

Our researchers sequenced thousands of cancer genomes of various tissue types, including breast, white blood cells, bone, soft tissue, oesophagus, kidney, head, and neck, colorectum, pancreas, skin, gallbladder, and other cancers, making a substantial contribution to the world's total.

Analysis of these cancer genome sequences led to the discovery of many previously unknown mutated cancer genes.

The proteins encoded by several of the mutated cancer genes identified have subsequently been used as targets for anti-cancer drug discovery programmes. For example, in 2002 Sanger scientists discovered mutations in the *BRAF* gene that were present in 70 per cent of malignant melanoma, which was untreatable at the time. Small molecule drug inhibitors of mutated BRAF were developed by pharmaceutical and biotechnology companies that caused BRAF mutant malignant melanomas to regress, and are now standard clinical treatments.

In 2008, we were a leading convenor and initiator of the International Cancer Genome Consortium, which provided coordination, oversight, strategy, and development of standards to deliver comprehensive cancer genome sequencing across the range of cancer types.

In 2010 our scientists published the first two complete whole cancer genome sequences, from a malignant melanoma and a small cell lung cancer, revealing their full repertoire of somatic mutation.

Sanger researchers developed the concept of 'mutational signatures', the patterns of somatic mutations caused by individual carcinogenic agents such as tobacco smoke chemicals and ultraviolet light. They then identified the repertoire of distinct mutational signatures operative in human cancers.

Several open data resources that are widely used to organise and present data on cancer genomes were developed at Sanger. One of these, COSMIC (Catalogue Of Somatic Mutations In Cancer), established in 2004, has become the world's largest database of somatic mutations in human cancers and is a routine resource for many researchers and clinical entities.

In the Pan-Cancer Analysis of Whole Genomes project, we led an analysis of more than 2,600 cancer genome sequences of 38 different tumour types in a collaboration involving over 1,300 scientists and clinicians from 37 countries, setting global conceptual and analytic standards for this type of analysis.

Established in 2009, the Genomics of Drug Sensitivity in Cancer initiative and database allowed researchers to identify the genetic abnormalities in a large panel of cancer cell lines that are predictive of response to the repertoire of anti-cancer drugs.

The Cancer Dependency Map integrates the work of multiple experimental and computational studies, to identify all essential genes in human cancers and helps establish dependencies in cancer cells for development of new therapies.

Impact

Discoveries from cancer genome sequencing at the Sanger Institute, and elsewhere, have transformed understanding of cancer biology and led to the development of successful new drugs for cancer treatment, new diagnostics, new approaches for early detection, new approaches for monitoring cancer burden and recurrence, and new insights for cancer prevention. All of these are, to varying extents, operating in current clinical practice. Cancer genome sequencing has also had substantial economic impact, with many biotechnology companies created, and pharmaceutical companies basing their drug strategies on the discoveries made. This includes the Sanger spinout Mosaic TX, which identifies selective vulnerabilities in cancer cells that can be exploited therapeutically to find new approaches to targeting difficult to treat cancers.

3. Exploring the full natural variation in human genomes

Introduction

The reference human genome sequence, coupled with advances in sequencing technologies and reductions in costs provided a foundation for the systematic large-scale study of inherited genetic variation in healthy individuals.

Sanger Institute contributions

Sanger scientists participated in the HapMap Project that discovered large numbers of common human sequence variations, using them to build linkage disequilibrium maps, which describe non-random associations between genetic loci. These maps allowed the effects of common genetic variation on any measurable phenotype to be studied across the genome.

The 1000 Genomes Project was an international collaboration led by Sanger scientists, to create a catalogue of common and rare inherited human sequence variations present in multiple populations.

As the costs of sequencing have continued to decrease, we have led or participated in ever larger-scale studies finding sequence variation human populations. These include the UK10K project analysing the genomes of 10,000 people in 2010 to sequencing of whole human genomes from 250,000 people in UK BioBank in 2021.

Impact

These studies have transformed our understanding of the sequence variation found in healthy human populations and hugely increased the number of inherited sequence variants known. This knowledge has provided the foundation for research into human evolution, the spread of *homo sapiens* from Africa, the existence of other humanoid species and many other topics. It has also provided information on genome and protein function and has been the basis for subsequent exploration of human disease through genome variation.

4. Identifying the roots of inherited genetic diseases for clinical use

Introduction

In addition to enabling the systematic study of inherited genetic variation in healthy individuals, the complete reference human genome sequence and subsequent technological advances also provided a framework to systematically investigate human sequence variation that contribute to inherited disease susceptibility.

Sanger Institute contributions

Mutations in genes on the X chromosome cause a wide range of inherited human diseases, including learning disability, particularly in males. Systematic sequencing of all X chromosome genes in individuals with learning disability led to the discovery of many mutated genes that cause learning disability. These genes have been incorporated into routine clinical genetics practice.

The landmark Deciphering Developmental Disorders (DDD) Study aimed to identify the mutated genes causing inherited severe developmental disorders. In a UK wide collaboration that sequenced and analysed the protein-coding regions (exome) of the genomes of individuals from more than 13,000

families, more than 70 new mutated genes contributing to developmental disorders were discovered. Most of the mutations identified occurred in parental germ cells, rather than being inherited through multiple generations, which has significance for genetic counselling. The study has already provided genetic diagnoses to more than 5,000 families, and will continue to search for further mutated genes contributing to rare childhood diseases.

The Prenatal Assessment of Genomes and Exomes (PAGE) study identified the inherited mutated genes underlying developmental anomalies identified during prenatal ultrasound screening, added to the gene list provided by the DDD Project.

Data arising from the DDD and PAGE initiatives, together with discoveries from other researchers in this field, are available through DECIPHER, an interactive database established at Sanger. DECIPHER facilitates the interpretation of disease-causing genome variants in rare human disease, supporting rare disease research and informing clinicians managing patients.

Our scientists also contributed to the UK collaborative Wellcome Trust Case Control Consortium (WTCCC), which used genome-wide association studies to explore the links between genome variation in the population and common diseases. The WTCCC substantially increased the number of genes known to have a role in the development of some of our most common diseases.

Impact

These studies have made profound overarching contributions to understanding of human disease genetics and to knowledge of gene function. Many of the inherited mutated genes in rare diseases that this work identified have been incorporated into routine clinical genetics practice globally. Furthermore, this research led to the spinout of Congenica, a company which uses genomic technologies to improve the diagnosis of rare paediatric disease.

5. Mapping all cell types and their functions in the human body

Introduction

Cells are the basic units of life. Historically, the identification of cell types and their organisation into tissues has primarily been based on microscopy. The Human Cell Atlas (HCA) is an international consortium aiming to produce a comprehensive catalogue of cell types in the human body by sequencing the array of mRNA transcripts (transcriptome) to understand the patterns of gene expression at a single cell level. The initiative will further map where each cell type is in the body and how each relates to other cell types in the formation of functioning tissues and organs.

Sanger Institute contributions

Founded in 2016 and co-led by Sanger scientists, the HCA has discovered many new cell types in adult and fetal tissues using single-cell transcriptome sequencing, with hundreds of projects contributing freely available findings and resources.

The HCA is providing foundational information on the full set of cell types in healthy humans and on their organisation and functional dependencies within tissues. This project will produce a map of the 37 trillion cells that form the human body that will transform medical research and healthcare worldwide.

Impact

Building the HCA data resource and analytical tools necessitated collaboration across multiple scientific disciplines including biology, medicine, computation, genomics, and technology development.

By 2023, cell atlases have been created for a broad range of organs and tissues, together with corresponding atlases of disease in the same organ from people with common complex diseases, cancers, rare diseases, infectious diseases, and other conditions.

6. Tracking infectious diseases with genomic surveillance

Introduction

Genomic technologies are revolutionising the way in which we can track the transmission of infectious disease pathogens and providing insight into their underlying biology, the genes and cellular processes that contribute to disease, and their ecological and evolutionary dynamics. Genomic surveillance is becoming an increasingly vital tool to monitor for spread of resistance to antimicrobial drugs and vaccines, and for newly emerging strains and potential pandemic threats. Our scientists work with investigators in countries around the world to understand the patterns of evolution and spread of range of infectious diseases, and link this to health outcomes.

Sanger Institute contributions

In 2005, the MalariaGEN network was established to enable genomic data-sharing to monitor and prevent the spread of anti-malarial drug resistance and support development of a vaccine. With partners in over 40 countries, the network undertakes surveillance of both the *Plasmodium* parasite and its *Anopheles* mosquito vector.

The Global Pneumococcal Sequencing (GPS) project uses large-scale whole genome sequencing to generate a high-resolution understanding of how the circulating population of *Streptococcus pneumoniae* bacteria adapts when susceptible strains are targeted by vaccination. Insights from GPS have been used to inform which serotypes to include in the latest pneumococcal conjugate vaccines under development.

Sanger scientists have used whole genome sequencing to study the evolution of the cholera-causing bacterium *Vibrio cholerae,* identifying historical transmission routes and redefining pandemic cholera as being caused by lineages of *V. cholerae* that have adapted for human-to-human transmission.

Whole genome sequencing was used to tracking the spread of methicillin resistant staphylococcus (MRSA) within in a UK hospital, demonstrating the potential for detection and management of disease outbreaks as part of routine hospital infection prevention and control measures.

Together with partners in the COVID-19 Genomics UK (COG-UK), the Sanger Institute established a pipeline to sequence and analyse tens of millions of samples obtained from across the UK, providing near real-time information on the spread and evolution of SARS-CoV-2. Through this work, the emergence of the Alpha, Delta and Omicron variants were detected early, providing vital data used to inform public health policy. Ultimately, the Institute has sequenced around 20% of the 16 million SARS-CoV-2 sequences generated globally.

Following the Sanger Institute's activity and experience gained during the early phases of the COVID-19 pandemic, "the Genomic Surveillance Unit" was established as a separate translational entity to provide services and products for genomic surveillance of a wide range of pathogens to the UK and internationally, with a focus on low- and middle-income countries.

Impact

Surveillance work undertaken at Sanger has enabled the sequencing of pathogen genomes at a previously unimaginable scale, which has provided fundamental scientific insights into pathogen evolution and

spread with an unprecedented degree of resolution.

As a consequence, genomic surveillance of pathogens, and their drug and vaccine resistance profiles, is becoming increasingly accepted as vital to informing national and international public health infectious disease control strategies.

The Institute's activities during the COVID-19 pandemic provided a globally unparalleled view on the evolution of SARS-CoV-2 to the UK government, providing information crucial to informing policy decisions on life-saving non-pharmaceutical interventions.

7. Understanding the genomic basis of ageing and disease

Introduction

While the detection of somatic mutations in cancer genomes has become standard over the past twenty years, detection of somatic mutations in healthy cells and tissues is a nascent area of research, as a result, our knowledge of normal cell genomes has remained relatively rudimentary.

Sanger Institute contributions

Sanger scientists developed multiple DNA sequencing approaches allowing detection of somatic mutations in normal tissues.

These techniques have allowed us to determine the mutation rates and mutational processes operative in many human cell types.

"Driver" mutations in cancer genes, previously found only in the genomes of cancer and premalignant tumours, were shown to be common in some normal tissues, increasing in prevalence during life.

The mutation rates, mutational signatures, and repertoire/prevalence of driver mutations changes owing to non-cancer disease states.

Impact

The Sanger Institute has led in establishing the foundations of this new domain of research concomitant with increasing interest and activity from other scientists and funding bodies.

An unexpected and rich landscape of differences between different cell and tissue types has been discovered, posing fundamental new biological questions pertaining to human biology.

This work is allowing us to identify individuals at elevated risk of cancer and revealed opportunities to modify cancer risk and non-cancer disease processes.

8. Understanding the effect of every possible variation in the human genome

Introduction

Understanding the consequences of changing (mutating) any one of the three billion 'letters' in the human genome has been a long-standing aspiration in human genetic research. The ability to do so will provide important insights ranging from a detailed understanding of the structure and function of the ~20,000 proteins encoded by the genome to improved use of genome sequencing in the diagnosis of human

genetic disease.

We have learned an enormous amount by examining the outcomes of natural random mutagenesis experiments, for example by sequencing cancer genomes and finding the mutated genes that convert normal cells to cancer cells, and by sequencing the genomes of individuals with genetic diseases and identifying the inherited mutated genes that cause these diseases.

A complementary approach is to artificially introduce mutations into the human genome in a directed and comprehensive manner in experimental systems. Combined with modern computational methods this approach holds great potential for understanding human biology.

Sanger Institute contributions

Sanger scientists have taken a leading role in an international collaborative initiative to generate all possible mutations in protein coding genes and regulatory sequences of the genome and to create an Atlas of Variant Effects. During its early stages ~11 million total variants been studied, which are already informing interpretation of clinical genetic tests, albeit thus far covering less than 1% of the human genome.

The Sanger Institute is moving beyond just generating comprehensive mutation atlases, by combining genomics, biophysics, mechanistic modelling, and artificial intelligence at scale to lay the foundations for ‘programmable biology’. This will enable scientists to design and produce new proteins and small molecules for disease treatment and bioengineering.

Impact

Atlases that chart the biological consequences of artificially induced mutations are providing fundamental insights into human biology, empowering the development therapeutics, and maximizing the utility of genomics for diagnosing genetic disease.