Data resources:

Genomic data resources and analysis tools created at the Sanger Institute 1993-2023

Introduction

We have a long tradition of generating, annotating and analysing genomic data and making it accessible to the scientific community. This section provides an overview of the genomic data resources and analysis tools we have created over the past thirty years, which facilitate and inform further research.

Data sharing at the Sanger Institute is based on the principles of Open Science. Our core policies promote the rapid sharing of datasets and databases, as well as protocols, software and analysis tools in order to maximise access to knowledge, increase inclusivity and collaboration, and enhance the impact of our research.

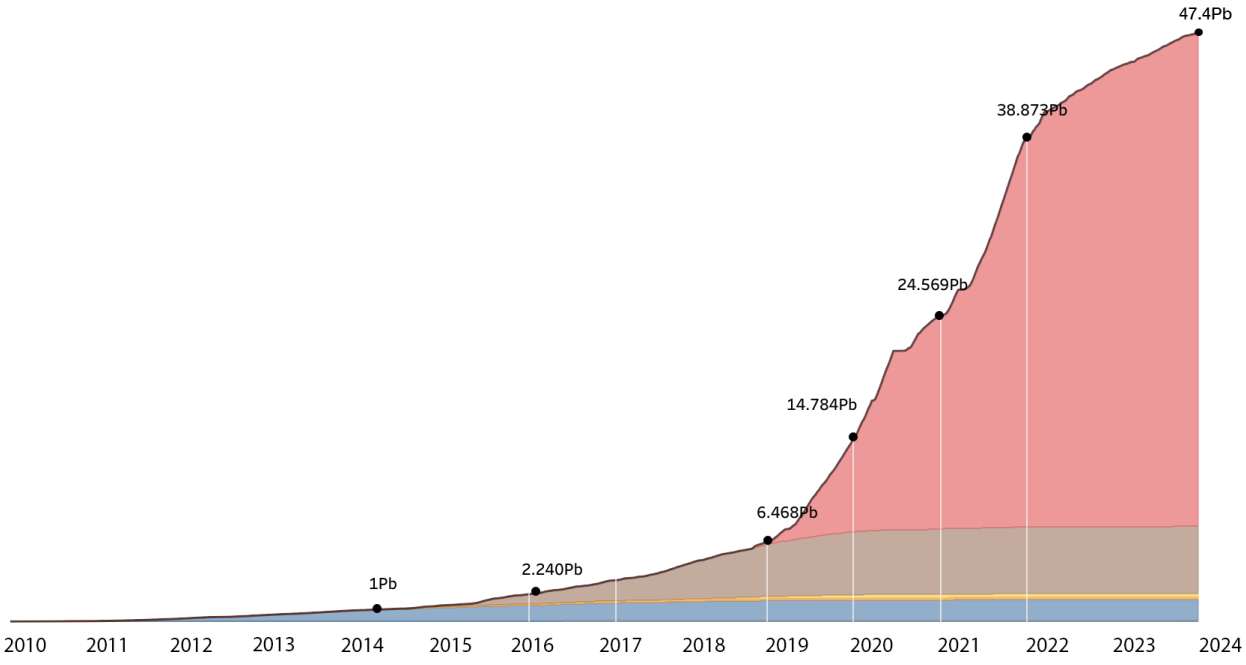
Sanger data resources are shared through portals and websites with the aim to enable easy and effective access to facilitate further research. Global access is often best facilitated by relying on well-established data repositories hosted by our long-term strategic partner EMBL-EBI. Many data resources initiated at the Sanger Institute are transferred to EMBL-EBI upon reaching maturity. In other cases, the data resources remain hosted at the Sanger Institute, are made available through websites managed by Sanger or our collaborators, or through open-access specialist portals such as protocols.io and Github.

Genome sequencing output

Total DNA sequencing contribution

The Sanger Institute is home to a large sequencing facility that delivers genome sequencing of a diverse range of species at scale. Genome sequencing at the Institute generated over 47.4 Petabases of sequencing data between January 2010 and June 2023 (Figure 1). The volume of the data produced increased rapidly from 2019, driven by the project to sequence 250,000 whole genomes from UK Biobank (completed in 2021). The sequencing output increased further in 2020, reflecting the Institute's rapid scale-up in capacity to process 64,000 SARS-CoV-2 virus samples per week as part of the pandemic response.

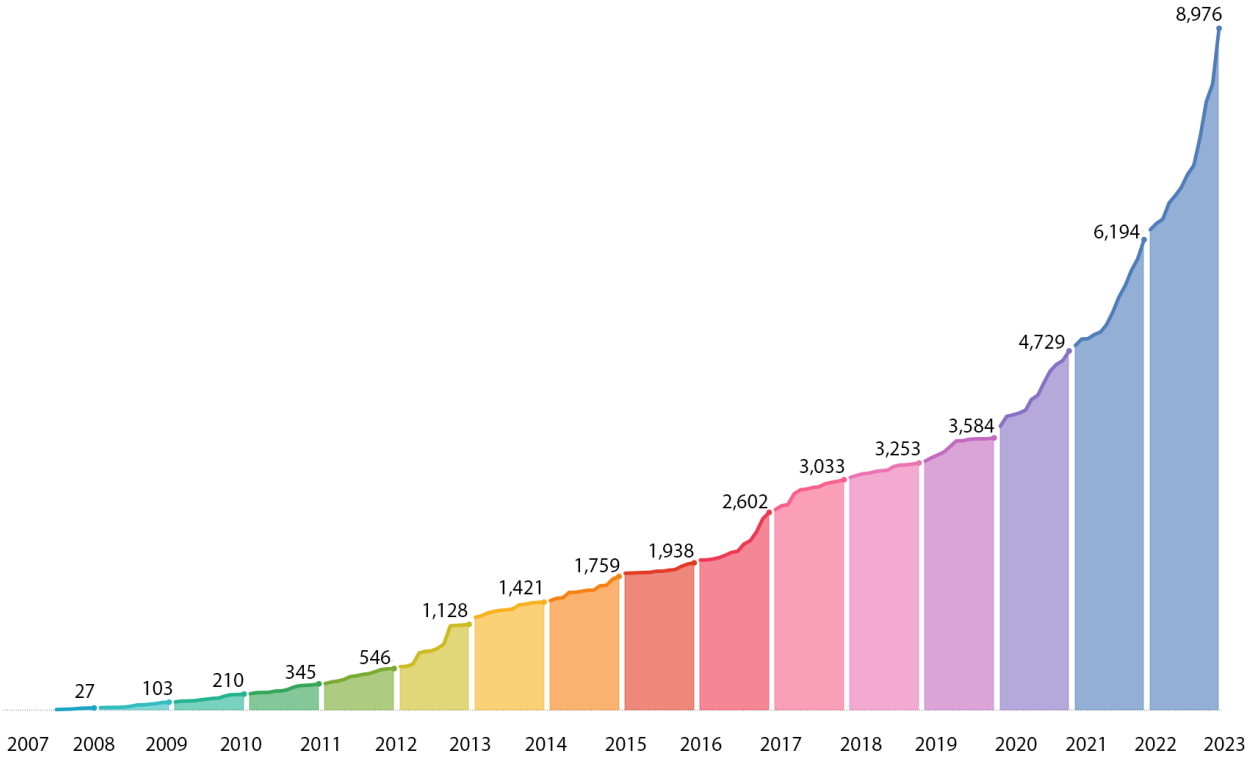
Figure 1. Cumulative sequencing output by the Sanger Institute, 2010-2023.



Number of species sequenced

The number of species diversity of the species sequenced at the Sanger Institute has increased over time with the cumulative total reaching 8,976 by October 2023 (Figure 2). These genomes come from a taxonomically diverse range of species, and will continue to grow in the coming years as the Tree of Life programme scales up efforts to sequence all eukaryotic species in the British Isles. The Tree of Life programme has produced 1000 new reference genomes by June 2023. These are being made accessible to researchers through the European Nucleotide Archive (ENA), hosted by EMBL-EBI and via a Genome Note.

Figure 2. Total number of species Sequenced, 2007-2023



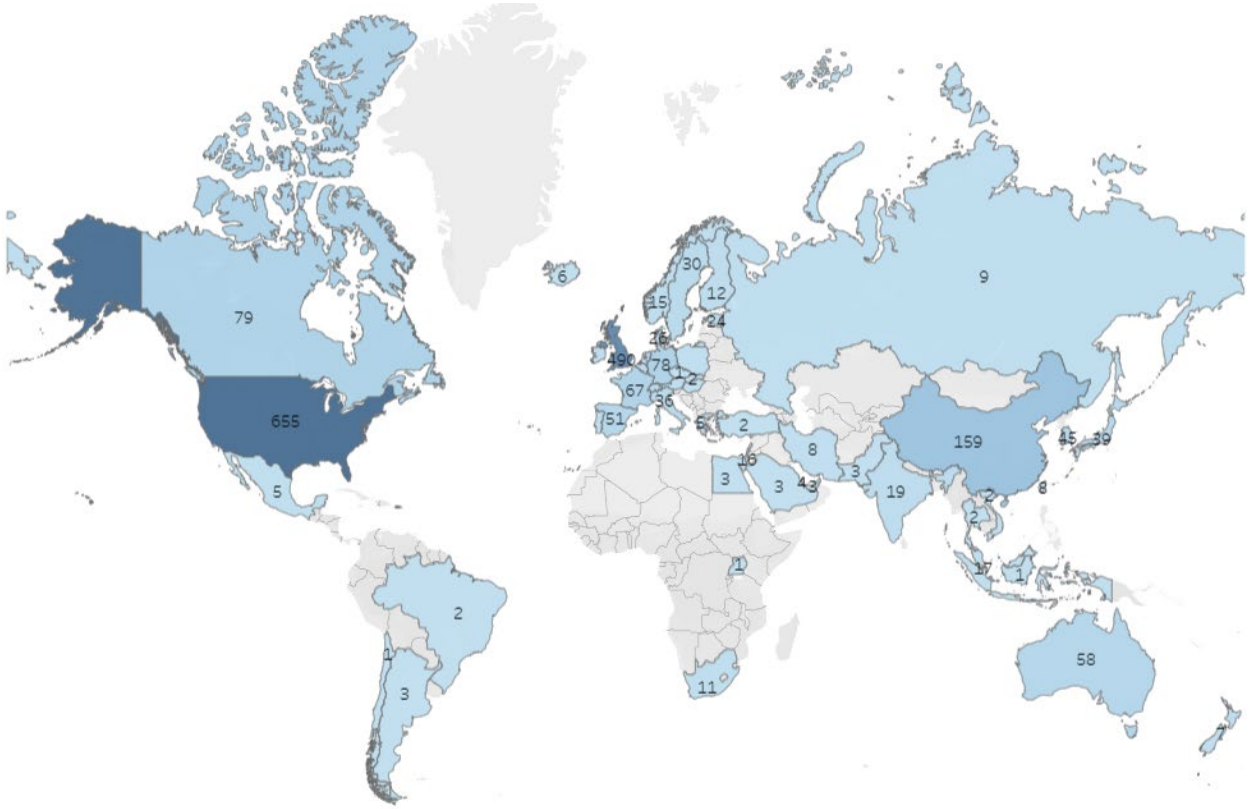
Contribution to the human genome data repository

Since October 2014, we have shared 538 studies and 710 datasets via the EMBL-EBI European Genome-Phenome Archive (EGA), one of the world's largest databases for storing and distributing human genetic and phenotypic data. These data have been reused by scientists globally, contributing to a further 354 publications which combined have received 19,115 citations by July 2023. The data that we have generated accounts for 8.5% of all studies and 6.9% of all datasets shared on EGA. On average, 107 million requests are made to EMBL-EBI websites every day.

Data Access Agreements

Where ethical considerations necessitate access to data to be regulated, we use Data Access Agreements (DAAs) to share data. Since 2018, Sanger has approved 2,174 different DAAs for human data with academic or commercial institutions and hospitals in 53 countries (Figure 3). The 83% of DAAs are requested by research institutes or universities, 11% by commercial entities and 6% by hospitals.

Figure 3. Sanger datasets shared through the Data Access Agreements, 2018-2023



Dedicated databases

As champions of open science, the overwhelming majority of the data from our projects are made available to the worldwide scientific community through freely accessible public databases, along with other scientific resources, computational tools and protocols.

While not exhaustive, the below list provides an insight into some of the largest and most impactful databases initiated at the Sanger Institute, listed in chronological order.

Pfam (1998)

This open access database was established to enable biologists to classify protein sequences based on hidden Markov model profiles. Pfam has transformed our ability to understand biological sequence data in its evolutionary context, allowing us to define “gene families”, and democratising knowledge.

Ensembl (2000)

Co-created by Sanger and EMBL-EBI scientists, Ensembl enables access to genome databases for vertebrate species and integrates these data with other biological information on genome structure and function. Used widely by the global research community, it now includes genome sequence data from thousands of vertebrate species, plants and fungi to provide researchers with timely access to the latest data.

WormBase ParaSite (2000)

This database was created to address the historical lack of investment in studying parasitic worms, which are responsible for some of the most neglected tropical diseases, including river blindness, schistosomiasis and hookworm disease, and which affect the lives of over a billion people globally. It provides a systematic approach to the integration, presentation and analysis of genome and

transcriptome data for 240 species of flatworm and nematode parasites. WormBase ParaSite has been recognised by the Global Biodata Coalition as a Global Core Biodata Resource whose long-term sustainability is critical to life science and biomedical research worldwide.

ENCODE - Encyclopaedia of DNA Elements (2003)

ENCODE was established to identify all functional elements of the human and mouse genomes, including information on gene locations. It is used as a reference for many subsequent large-scale genomics projects including the Human Cell Atlas, the Cancer Genome Atlas, and the NIH Roadmap Epigenomics Mapping Consortium.

DECIPHER - the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (2004)

DECIPHER includes data from two landmark longitudinal studies: Deciphering Developmental Disorders (DDD) and the Prenatal Assessment of Genomes and Exomes (PAGE). It facilitates the interpretation of disease-causing genome variants in rare human disease, supporting research and informing clinicians managing patients. Three hundred centres across the globe contribute genomic and patient data into DECIPHER, which now has more than 45,000 patient records linked to more than 184,000 physical observations, with more than 46,000 records of copy number variations and more than 12,000 sequence variants through open-access. DECIPHER is now hosted by EMBL-EBI.

COSMIC - the Catalogue of Somatic Mutations in Cancer (2004)

COSMIC is a comprehensive database of cancer mutation data and analytic tools. It contains records of more than 23 million mutations and 6,800 precise forms of human cancer from 29,000 peer reviewed publications and 42,000 whole genome screen samples. The COSMIC database has been cited more than 20,000 times, and has 50,000 academic and commercial registered users, with 30,000 users/month visiting the website.

HipSci - Human Induced Pluripotent Stem Cells Initiative (2012)

This initiative explores how genome variation influences cellular characteristics and has generated over 900 human induced-pluripotent stem cell (iPSC) lines from healthy volunteers and individuals with genetic disease. It has become one of the largest iPSC resources, and includes rich genomic information facilitating exploration of disease mechanisms. The HipSci portal provides a central catalogue of all of the cell lines generated by the project, providing direct links to all open access assay data.

The Human Cell Atlas – HCA (2016)

HCA was co-founded and co-led by Sanger scientists as a major global initiative to map and characterise every cell type in the human body, using advanced single-cell genomics. Single-cell atlases have been created for a broad range of organs and tissues through new cross-disciplinary collaborations. By June 2023 the HCA has over 3,000 members in 95 countries, providing foundational information on the human cell organisation and functional dependencies within tissues.

Cancer Dependency Map (Cancer DepMap) project (2017)

In strategic collaboration with the Broad Institute (Boston, USA), this project aims to identify all genetic dependencies in cancer cells and use this information to be exploited to develop new therapies. It now brings together several resources, including:

The **Cell Model Passports** (CMP), a resource of over two thousand cancer cell line and organoid models and access to various genomic and functional datasets including CRISPR and drug sensitivity screens, mutational data and proteomics. As part of the Organoid Derivation Project, Sanger researchers have created 276 organoids, of which the datasets for 145 have been made available as a resource for the

scientific community. These have been accessed by researchers from 160 countries, through 64,000 site visits to the website (up 172 per cent from previous year). In addition, 45 of these organoid models are commercially available as a biological resource to the scientific community.

The **Genomics of Drug Sensitivity in Cancer** (2009), a resource for therapeutic biomarker discovery in cancer cells, which can be used to identify patients most likely to respond to anticancer drug, has screened over 500 compounds for genomic markers predicting drug response.

Data analysis tools

More than 200 tools, software, protocols and services are downloadable from our website, or dedicated repositories, which enable the processing, analysis and management of research data or laboratory information. These include protocols and resources relating to statistical analysis, laboratory animals, vectors and flow cytometry. For example, the Burrows-Wheeler Alignment (BWA) tool aligns short reads of DNA sequence against a reference sequence, such as the human genome, more accurately and faster than existing tools. This freely available software was developed in 2009 and remains in use today, having been cited over 34,000 times.

Sanger researchers have invented numerous computational formats and tools to standardise and compress genome data for effective storage and management. Developed in 2009, the Sequence Alignment/Map (SAM) provided a common format for storing sequence read alignments against reference genome sequences, making downstream processing less complicated. SAM became the standard format for storage of genomic data before it, and BAM (the binary compressed version of a SAM file), were superseded by CRAM in 2011. Developed in collaboration with EMBL-EBI scientists, CRAM provided a compressed version of BAM which halved the data storage required for a human genome. All three have been adopted as the industry standard for genomic data compression, storage and transfer and are still used today.

The Sanger Institute is a co-founder and host of the Global Alliance for Genomics and Health (GA4GH, 2013), a collaborative project to create the protocols and frameworks needed to open up the world's genomic databases to the global scientific community to deliver truly seamless sharing of genomic and clinical data.

Conclusion

Open Science is central to the Sanger Institute's scientific strategy to enable rapid and responsive sharing of data, software and tools across the globe. Genome sequencing underpins our substantial contributions to the global research ecosystem. Through our numerous international collaborative projects, we deploy genomic science at scale and play a leading role in driving advances in the capabilities and applications of genomic technologies. We ensure that the foundational datasets that we generate are accessible to the global scientific community to enable others to address their own research questions, while providing insights that transform our understanding of biology, in health and disease, and facilitate the advancement of knowledge.