EUROPE

SARAH PARKINSON, JOE FRANCOMBE, CAGLA STEVENSON, HAMISH EVANS,
ADVAIT DESHPANDE, SUSAN GUTHRIE

Wellcome Sanger Institute and Wellcome Genome Campus Landscape Review

For more information on this publication, visit www.rand.org/t/RRA215-1

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

RAND Europe is a not-for-profit organisation whose mission is to help improve policy and decision making through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Table of contents

Abbreviations

ARGO	Accelerate Research in Genomic Oncology
BIC	BioData Innovation Centre
CAP	Complementary Analysis Programme
BRAIN	Brain Research Through Advancing Innovative Neurotechnologies
COSMIC	Catalogue of Somatic Mutations in Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CWTS	Leiden University's Centre for Science and Technology Studies
DDD	Deciphering Developmental Disorders
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
FISH	Fluorescence in situ Hybridisation
GEL	Genomics England Ltd.
GRL	Genome Research Limited
GA4GH	Global Alliance for Genomics and Health
GATK	Genome Analysis Toolkit
GSK	GlaxoSmithKline
GWAS	Genome-Wide Association Studies
HCA	Human Cell Atlas
HCP	Highly cited publications
HHMI	Howard Hughes Medical Institute
ICGC	International Cancer Genome Consortium
IP	Intellectual Property
iPSC	induced Pluripotent Stem Cells
LMIC	Low- to Middle-Income Country
MIT	Massachusetts Institute of Technology
MNCS	Mean Normalised Citation Score
MNJS	Mean Normalised Journal Score

MRC	Medical Research Council
NHGRI	National Human Genome Research Institute
NHS	National Health Service
OGT	Oxford Gene Technology
OICR	Ontario Institute for Cancer Research
PCAWG	Pan-Cancer Analysis of Whole Genomes
PDNA	Plasmodium Diversity Network Africa
PI	Principle investigator
QA	Quality Assurance
QTL	Quantitative Trait Locus
REA	Rapid Evidence Assessment
SNP	Single Nucleotide Polymorphism
SPACE	Stakeholder Perspectives on the Ethical Challenges in the use of Artificial Intelligence for Cognitive Evaluation
SWOT	Strength, Weakness, Opportunities, Threats
TCGA	The Cancer Genome Atlas
WHG	Wellcome Centre for Human Genetics
WoS	Web of Science

Figures

Tables

Boxes

Executive summary

RAND Europe was commissioned by Wellcome to analyse the role and contribution of the Wellcome Sanger Institute and the Wellcome Genome Campus within the field of genetics and genomics, and within the wider research landscape. The purpose of this study was to understand the contributions that the Sanger Institute and Wellcome Genome Campus make to the research community and society as a whole, as part of Wellcome's ongoing evaluation of their strategic funding priorities. The study also looked at four comparator organisations[1] to understand their contribution to the field of genetics and genomics and their wider impact, in order to contextualise Sanger and the Wellcome Genome Campus's contribution and set it within the wider research landscape. The review consisted of desk research, interviews (internal and external), bibliometric analysis, and case studies.

The study resulted in a report for Wellcome's internal use to inform their decision making. This document included a detailed consideration of the strengths, weaknesses, opportunities and threats that the Sanger Institute and Genome Campus face, which are relevant to Wellcome's internal reviews and decision-making processes.

The report presented here is derived from the report delivered to Wellcome, and is intended to be relevant to a wider audience interested in the genetics and genomics research landscape. Some detail that was relevant to Wellcome has been removed from the original report, and additional context has been added to make the findings relevant to a wider audience. The purpose of this report is to provide an overview of the contributions of the Sanger Institute and Wellcome Genome Campus alongside other comparable organisations in the field of genetics and genomics, and to present important themes that have influenced and will continue to influence organisations working in the field of genetics and genomics.

Key contributions of the Sanger Institute and comparator organisations

The Wellcome Sanger Institute has made important contributions within and beyond the field of genetics and genomics research, some key examples of which are summarised in Figure 1.

[1] Broad Institute, Wellcome Centre for Human Genetics (WHG), Janelia Research Campus and the National Human Genome Research Institute (NHGRI).

Figure 1: Summary of key achievements of the Sanger Institute

Areas of key research contributions	Outputs and impact across these research areas	Broader contributions
Key role in the Human Genome Project. Subsequent genome sequencing projects including 1,000/10,000/100,000/500,000 genomes projects. Human genetics research: sequencing to identify genetic variation that causes developmental disorders. Cancer research: understanding somatic mutation and its role in cancer and other diseases. Infectious disease research: developing pathogen reference genomes and exploring the genetic diversity of infectious agents around the world.	**Publications** (4,720 articles in the last 10 years and c.335,000 citations). Reference **datasets**, e.g. Ensemble, the Catalogue of Somatic Mutations in Cancer (COSMIC), DECIPHER. Novel **data analysis software**. **Spin-out companies**, e.g. Congenica, Microbiotica. **Leadership** of international research programmes. Wide-ranging **collaboration** with health service providers, e.g. the National Health Service (NHS), World Health Organization.	Wide-ranging contribution to training and development, including in low- to middle-income countries (LMICs). Public engagement and public attitudes towards uses of genomic data.

Source: RAND Europe analysis

To better understand the Sanger Institute's role in the field of genetics and genomics, we compared these contributions to that of four comparator organisations[2]: Wellcome Centre for Human Genetics (WHG), Janelia Research Campus and the National Human Genome Research Institute (NHGRI). An overview of each institution is provided in Table 1 below, which provides some of the key characteristics of each organisation, their main activities and operating environment. Table 2 provides a brief summary of the key outputs and contributions of each institution, including each organisations' bibliometric output. The main takeaways from these comparisons are:

- **Training and capacity building:** The Sanger Institute provides training and capacity building through its multiple formal and more informal partnership-based training offers. NHGRI also emphasises training, in part reflecting its dual role as government agency and funder as well as research institution, alongside WHG's role as a university department.
- **Leadership:** The Sanger Institute offers research at the scale needed to take on leadership roles on large-scale international consortia, as do NHGRI and the Broad Institute. The evidence suggests all three institutions have taken these leadership positions individually and collectively over time. Sanger has also undertaken a convening role for networks of smaller actors, too, using its scale to bring actors together in both the UK and LMICs.
- **Datasets and tools:** Another key contribution made by Sanger is the profusion of open access datasets and tools brought to the research field – though we also see strong contributions of this

[2] Comparator organisations were chosen in consultation with Wellcome, from a long list of comparators provided by the project team after initial desk research.

nature from comparator organisations, such as the imaging resources made available by Janelia Research Campus.

- **Commercialisation:** Sanger's focus on commercialisation has not been as strong as in other organisations – notably the Broad Institute – but other routes to translation have been emphasised, particular the strong ethos of openness and data sharing.
- **Disciplinary strengths:** Sanger has strong disciplinary strengths in the core area of genetics and heredity, with comparable citation levels to key comparators such as the Broad Institute and NHGRI. Two particular areas where the Sanger Institute outperforms these two closest comparators (in terms of subject matter) on the level of citation are biochemistry and molecular biology, and infectious diseases.

Table 1: Summary of the key characteristics of each institution

Institute	Year founded	Key partners	Staff	Annual funding and core funding	Scientific focus and key contributions	Research tools and technologies
Wellcome Sanger Institute	1992	University of Cambridge, EMBL-EBI	Over 1100 employees and 63 post-graduate students	£152.4m – 64% from Wellcome (2019), although it has been a higher amount from Wellcome in past	Large-scale genomic data production and analysis, cancer, human genetics and disease, parasites and microbes, genetic basis of diversity	Tools for annotation, gene finding, processing sequence data, visualisation, sequencing facilities
Broad Institute	2004	MIT, Harvard, hospitals	15 core Institute members, 51 non-core, >300 associate members	US$547.4m (c. £440m) – mixed income sources (2019)	Large-scale genomic data production, genome editing, drug discovery, cancer	Data analysis software for genomics and clinical data, cloud-based data storage
Wellcome Centre for Human Genetics	1994	Oxford University	Over 400 active researchers, 70 administrative staff	£20m annually of competitively won grants (2019)	Human disease research, gene sequencing using nanopore technology, genomic medicine	High throughput sequencing, computing and cellular imaging; graduate studies
Janelia Research Campus	2006	Established by HHMI	350 scientists	US$150m (c. £65m)	Mechanistic cognitive neuroscience, molecular tools and imaging, computation	Molecular tools, imaging technologies, tools for data science, software for analysis

National Human Genome Research Institute	1989	Part of NIH	50 NHGRI investigators within the division of intramural research	US$500–600m (c. £400–480m) from federal funding (2019)	Genomic technologies and data science, genetic disease, precision medicine, cancer, clinical research	Software and analysis tools for genomic data including study of complex traits, analysing sequencing data

Source: RAND Europe analysis

Table 2: Summary of key outputs and contributions of each institution

Institute	Publications (2008-2017)	Datasets and research tools	Training and capacity building	Translation and commercialisation	Public engagement and outreach	Leadership and advancement of field
Wellcome Sanger Institute	n=4,720, MNCS[3] 2.59, % papers in HCP[4] (10%): 29.4%; (1%): 5.6%	Tools for annotation, gene finding, processing, visualisation, sequencing facilities; datasets, e.g. Ensembl, COSMIC, DECIPHER	PhDs – 4-year, clinical, affiliated; masters programmes including MPhil genomic science targeted at LMICs; courses for scientists and health	Spin-out companies, e.g. Congenica, Microbiotica; BIC hosts 8 companies; licensing of COSMIC	Connecting Science public engagement team of 10 providing regular programme of monthly events, one-off events, online materials, training for	Leadership of international research programmes; key role in Human Genome Project and subsequent 1,000 etc. genome projects

[3] Mean normalised citation score: a measure of publication impact using number of citations, normalised to account for different citation patterns across fields of science and for differences in the age of publications. When MNCS is above 1, it indicates that an organisation performs better than the world average. When MNCS is below 1, it means that, on average, an organisation produces publications that are not cited as often as the world average.

[4] Highly cited publications: the percentage of publications produced by organisation that are in top 10% or 1% of highly cited publications.

			professionals; apprenticeships; range of courses and conferences through Connecting Science		scientists in engagement	
Broad Institute	n=5,591, MNCS 3.15, % papers in HCP (10%): 39.2%; (1%): 7.8%	Many datasets, software and tools to improve data analysis for large-scale genomic and clinical data, including a cloud-based analysis and data storage portal	Post-baccalaureate programmes for research experience, scientific writing and communications internships, associations for graduate and post-docs	Exclusive licensing of CRISPR-Cas9 in mammals to Broad Institute spin out; drug discovery and gene editing spin outs; 100s of patents in tech, therapeutics, engineering	Community activities for educators, educational tours, lecture series for the public, undergraduate research opportunities	Leading role in large international collaborations around Big Science, starting from the Human Genome Project
Wellcome Centre for Human Genetics	n=2,874, MNCS 1.94, % papers in HCP (10%): 24.6%; (1%): 3.3%	Research facilities for high throughput genomics and cellular imaging available to Oxford researchers on a fee-for-service basis	Graduate studies funded each year through Centre	At least 3 start ups focused on therapeutics and drug discovery	Public events, including with a dance company to promote science, technology, engineering, the arts and mathematics; undergraduate research internships	Leader in nanopore technology for genomic sequencing and in sequencing for genomic medicine

Janelia Research Campus	n=1,158, MNCS 3.34, % papers in HCP (10%): 41.5%; (1%): 9.2%	159 (including data, software, lab tools)	Undergraduate summer schools, graduate programme, high-school internships, computing programme; small, highly specialised conferences (~2/mth), workshops for junior scientists (~2/yr)	29 tools with patents issued or pending	Public lecture series 'Dialogues of Discovery'; provides ~US$1m annually to the local community school district to support science education	Main area of advancement for the field is in imaging techniques and tools, e.g. most detailed map of the fly brain to date
National Human Genome Research Institute	n=2,887, MNCS 1.71, % papers in HCP (10%): 19.2%; (1%): 2.4%	18	Graduate partnership programmes, medical residency, undergraduate programmes, summer internships, training in social and behavioural research	8 patented or pending techs; 9 research materials listed which are available for licensing	Community Engagement in Genomics Working Group; education and community involvement branch	Leader in the Human Genome Project, PCAWG, contributed to the HapMap, NHGRI-EMBL-EBI GWAS catalogue

Source: RAND Europe analysis

Reflections and future outlook

Reflecting overall on the Sanger Institute's role in the field of genetics and genomics, they act as a leader in the progress and development of research, alongside other key actors such as the Broad Institute, the only comparator organisation that is of a similar scale and scope as Sanger. The flavour and direction of this leadership is shaped by its specific values and ethos around openness, data sharing and collaboration. This provides a unique and novel contribution to the field of genetics and genomics, including indirectly through other organisations and actors using research conducted at Sanger and by partner institutions at the Wellcome Genome Campus.

Currently, there is a challenge for Sanger and other organisations working in genome sequencing in that high throughput sequencing does not offer the same competitive advantage as in the past, as this technology has become more widely available in recent years. The next challenge will be to find a direction for these organisations to explore that fits with their scale, operating environments, unique strengths and values. In the case of Sanger, this means finding an approach that capitalises on their ethos of openness, data sharing and collaboration to shape the development of the field over the coming years.

1. Introduction

1.1. Background and context

RAND Europe was commissioned by Wellcome to analyse the role and contribution of the Wellcome Sanger Institute and Wellcome Genome Campus within the field of genetics and genomics, and within the wider research landscape. This landscape review covered the academic contributions of Sanger, as well as the wider translational, communication, networking and commercialisation activities associated with the Sanger Institute and the Wellcome Genome Campus, setting both within the context of other comparator organisations.

This report is derived from a separate internal report delivered to Wellcome, and is intended to be relevant to a wider audience interested in the genetics and genomics research landscape. Some detail that was relevant to Wellcome has been removed from the original report, and additional context has been added to make the findings relevant to a wider audience. The purpose of this report is to provide an overview of the contributions of the Sanger Institute and Wellcome Genome Campus alongside other comparable organisations in the field of genetics and genomics, and to present important themes that have influenced and will continue to influence the field of genetics and genomics.

1.1.1. Wellcome Sanger Institute

Sanger was established in 1992 to improve human health by delivering insight into human and pathogen biology. This continues to be enabled by research conducted at Sanger, as well as through the numerous partnerships they have with academic institutions and industry (e.g. pharmaceutical companies). The Sanger Institute is supported by core funding from Wellcome, which allows them to work towards their aim of conducting studies that are of a design and scale such that other biomedical research institutes and academic institutions may struggle to compete, particularly in the area of high-throughput genomics. A major theme of their research is genomic variation, which underpins their scientific programmes: Cancer, Ageing and Somatic Mutation; Cellular Genetics; Human Genetics; Parasites and Microbes; and the Tree of Life looking at evolutionary insight from genome sequencing across species.

Sanger is a strong advocate of open science, demonstrated through the open sharing of software tools, datasets and techniques they develop with the wider research community, and has a significant role in the training of new and upcoming scientists in genetics and genomics. Since its establishment, Sanger has been involved in a number of high-profile, collaborative research projects including the Human Genome Project, which sought to map and understand all the genes within the human body, as well as the 1000 Genomes Project, the HapMap project and the Human Cell Atlas (HCA) initiative. Sanger also continues to be

involved in large-scale projects such as Deciphering Developmental Disorders (DDD) and the newly announced, Tree of Life project.[5] In addition to research projects, Sanger has been involved in creating a number of spin-out companies, including Microbiotica (which aims to use microbiome research to develop highly specialised treatments) and Congenica (a digital health company that supports clinicians to interpret complex genomic data).

1.1.2. The Wellcome Genome Campus

The Sanger Institute is situated within the Wellcome Genome Campus in Hinxton, UK, which is home to a number of research organisations focusing on genomics and computational biology. Alongside Sanger the campus includes the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Genomics England Ltd. (GEL; set up by the Department of Health and Social Care to deliver the 100,000 Genomes Project), the BioData Innovation Centre (BIC; which provides a flexible working space for genomics and biodata companies) and Connecting Science (which connects researchers, health professionals and the wider public).[6] The shared physical location of these organisations is intended to help facilitate sharing of knowledge and research capacities across organisations, staff mobility and networking among those on the campus. Sanger, Connecting Science and the fabric and infrastructure of the Wellcome Genome Campus operate under the name, Genome Research Limited (GRL), which is wholly owned by Wellcome, which has shaped Sanger's leadership role within the physical research campus. Wellcome has invested £3.5billion into the campus in the last 25 years, which supports Sanger, Connecting Science and the physical campus space (Wellcome 2019e).

1.2. Scope of the work

The aim of the landscape review was to understand the contributions that Sanger and the Wellcome Genome Campus make to the research community and society as a whole. The work focused specifically on the Wellcome-funded elements of the Wellcome Genome Campus[7] – not covering EMBL-EBI, BIC or GEL in detail, beyond their collaboration and interaction with the other parts of the campus. The study also looked at a number of comparator organisations[8] to understand their contribution to the field of genetics and genomics and their wider impact, in order to contextualise Sanger and the Wellcome Genome Campus's contribution and set it within the wider research landscape.

1.3. Research questions

The research questions that guided this study are:

[5] As of 28 July 2020: https://www.sanger.ac.uk/programme/tree-of-life/; as of 28 July 2020: https://www.ddduk.org/

[6] As of 28 July 2020: https://www.wellcomegenomecampus.org/aboutus/#whoishere

[7] Namely, the Sanger Institute, Connecting Science and activities at partner organisations on the Genome Campus that are done in partnership with Sanger

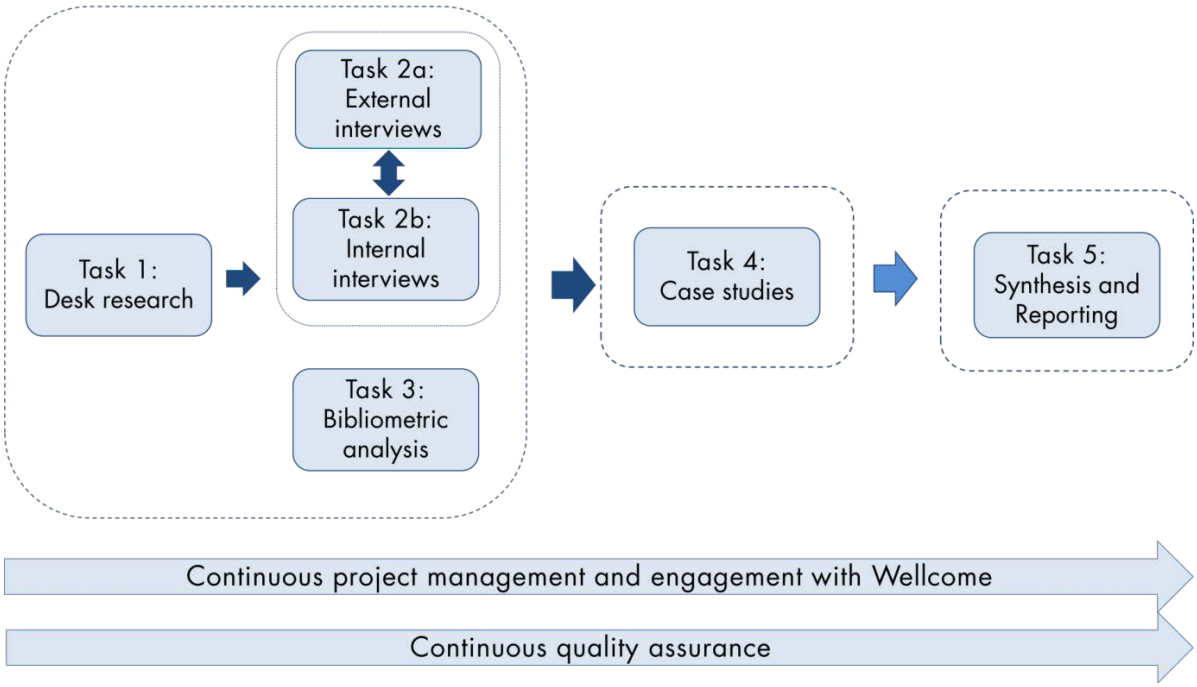
[8] The Broad Institute, WHG, NHGRI, and Janelia Research Campus. See below and Chapter 3 for more details.

1. What is the role of Sanger and the Wellcome Genome Campus in the genetics and genomics research landscape?
    a. What are the strengths of Sanger and the Genome Campus in the genetics and genomics landscape?
    b. What factors influence how Sanger and the Genome Campus contribute to the genetics and genomics landscape?
    c. What are the unique and distinctive contributions of Sanger and the Genome Campus?
    d. What is the contribution of Sanger and the Genome Campus relative to comparator organisations in the UK and internationally?
2. What are potential areas of opportunity for Sanger and the Genome Campus to further contribute to the genetics and genomics research landscape?

1.4. Approach

This study used a multi-method approach combining desk research, interviews, bibliometric analysis and case studies to understand a range of perspectives on the role and contributions of Sanger and Wellcome Genome Campus within the field of genetics and genomics and more broadly. Figure 2 provides an overview of the methodological approaches used in this study. A more detailed explanation of the methods used in this landscape review is presented in Annex A.

Figure 2: Methodological approach



Source: RAND Europe analysis

To benchmark and contextualise Sanger's contributions, this study explored the role and contributions of four comparator organisations using the same methods as those used to explore Sanger and the campus: the Broad Institute (Massachusetts, USA), the Wellcome Centre for Human Genetics (WHG; Oxford, UK), Janelia Research Campus (Virginia, USA) and the National Human Genome Research Institute (NHGRI; Maryland, USA). These comparator organisations were chosen in consultation with Wellcome, from a long list of comparators provided by the project team after initial desk research.

1.4.1. Desk research

Desk research consisted of a rapid evidence assessment of the literature available on the Sanger, Wellcome Genome Campus and each comparator organisation, as well as a review of each organisation's website and a review of any grey literature available. The focus of the desk research was to investigate the research contributions and wider activities of each organisation in the field of genetics and genomics, strengths and weaknesses within each organisation, and how the work of each organisation had been used, including by other researchers and the private sector. Desk research on Sanger and Genome Campus included a review of documentation provided by Wellcome for Sanger's quinquennial review, where information from the documents was extracted into an Excel template to aid analysis (see Table 12 in Annex A).

1.4.2. Interviews

We conducted 42 semi-structured interviews with 45 people for this study. Several groups of interviewees were included: internal Sanger and Genome Campus staff, including at least one interviewee from each Sanger Institute scientific programme; external experts in the field of genetics and genomics and external users of Sanger research; representatives of each comparator organisation; and those involved in case studies explored in this study.[9] A protocol for each group of interviewees was developed, which are provided in Annex B, although the interviews varied according to the knowledge and expertise of the interviewee. Before the interview, an information sheet was sent to all interviewees with information on the study and how the study team would use data from each interview. Interviews lasted approximately an hour and were conducted by telephone. Data from interviews are referenced throughout using identifiers; the characteristics of interviewees linked to these identifiers can be found in Table 13 in Annex A.

1.4.3. Bibliometric analysis

We conducted bibliometric analysis to provide metrics around publications and citations of Sanger Institute and each comparator organisation. The bibliometric analysis covered publications in the period 2008–2017,[10] and included metrics for each organisation based on: their number of publications, the number of citations received, the mean normalised citation score (MNCS) for all publications, the mean normalised journal score (MNJS) for all publications, and the percentage of publications within the top 1% and 10% of highly cited publications. The bibliometric analysis included metrics around the fields in which each

[9] Anonymous identifiers for interviews are as used as follows throughout this report: internal interviews (Int_01, Int_02,…), external interviews (Ext_01, Ext_02, …), comparator interviews (Comp_01, Comp_02, …) and case study interviews (CS_01, CS_02, …).

[10] Citations were counted up to and including 2018.

organisation published and the fields within which the work of each organisation was cited, as well as a collaboration network analysis map to visualise the collaborations of each organisation.

1.4.4. Case studies

Five case studies were chosen for further exploration into the contribution of the Sanger Institute, and other organisations, selected in collaboration with Wellcome at the mid-point of the study. The choice was based on initial desk research and interviews, as well as Wellcome's perspective on what information would be most useful to them. Open Targets, Deciphering Developmental Disorders (DDD), Tree of Life, the International Cancer Genome Consortium (ICGC) and Sanger's work in malaria research and the MalariaGEN network in particular, were selected. These case studies provided two perspectives on the contributions of Sanger: a forward-tracing view on how programmes and projects at Sanger contributed to the field and had impact; and a backward-tracing view on large, international collaborations and the role that the Sanger played within these projects.

1.4.5. Analysis and synthesis

We mapped data from interviews and document review into an Excel template, which captured information on structure and processes, and outputs by category. Team members synthesised evidence across parts of this framework and shared findings with the wider study team at an internal workshop. There, we revised these findings and discussed emerging themes. We triangulated and tested new issues and themes against the evidence collected.

1.5. Limitations of the study

We have identified a number of caveats and limitations to this analysis, the most important of which are set out below:

- **Reliance on interviews and self-report**: A significant proportion of the information used for this analysis is based on interviews and is therefore affected by the accuracy and completeness of the recall of participants, their willingness to disclose information and views, and their biases. We mitigate this by interviewing a range of individuals.
- **Bibliometric limitations**: All bibliometric analyses are subject to some common limitations, such as the use of citations as a proxy for research quality and the limits in coverage in bibliometric databases. In this project we identified relevant publications primarily through author addresses, which may not be used consistently.
- **Completeness of the range of contributions collected**: This report provides a picture of some of the important contributions to the field of genetics and genomics research, but is not a comprehensive record of all contributions made, because of the complex nature of research translation pathways.
- **Range of comparators**: We selected a sample of comparators to review and benchmark the performance of the Sanger Institute and present a context for their role in the wider landscape.

However, there are many other potential comparators that could have been chosen, which may have offered a different perspective.

1.6. Structure of this report

This report presents the findings of the landscape review and is intended to give Wellcome a resource to inform its decision making. Chapter 2 describes the role and contribution of Sanger and the Wellcome Genome Campus in the field of genetics and genomics. Chapter 3 provides similar information for each comparator organisation and contextualises Sanger in the research landscape, and compares each organisation's bibliometric output. Chapter 4 summarises each case study, and reflects on what Sanger and the campus can learn from each one. Lastly, in Chapter 5 the findings of this study are discussed, along with reflections on the study and on the research landscape for genetics and genomics. The annexes of this report contain detailed information about the methodology used in this study (Annex A), interview protocols (Annex B) and the structure of the case studies (Annex C).

2. Review of Wellcome Sanger Institute and Wellcome Genome Campus

2.1. Operation and key features

2.1.1. Sanger Institute

As of July 2020, around 2,500 staff work at the Wellcome Genome Campus and are employed by organisations on the campus, over 1,100 of whom are scientific employees at Sanger Institute. Sanger aims for a steady state of 35 core faculty, 20 associate faculty, and 15 honorary faculty at any one time (Wellcome 2019e). These faculty members create the strategy and the large-scale pieces of work that characterise the Sanger Institute's scientific strategy and portfolio (Wellcome 2019e, 5). They lead their own research groups at Sanger, focusing on a specific project within the five broad research programmes (Cancer, Ageing and Somatic Mutation; Cellular Genetics; Human Genetics; Parasites and Microbes; and the Tree of Life). There are also a number of scientific operation groups not uniquely associated with any one faculty member at Sanger working on various projects, for example on building genome databases and gene editing technologies, which are used across the institute. Similarly, there are also several information technology groups, working on informatics and software development within Sanger. They provide the infrastructure that is vital for large-scale, high-throughput science. Finally, within scientific groups at Sanger, there are several pipeline groups. They continually evolve and develop the efficiency and accuracy of sequencing, model organisms, cellular and analysis pipelines to enable Sanger researchers to study biological phenomena at scale. All of these scientific groupings contribute to the five broad research programmes at Sanger, which are described in detail below.

Cancer, Ageing and Somatic Mutation

The Cancer, Ageing and Somatic Mutation programme began in 2000. It aims to use high-throughput genomic technologies to obtain a richer understand of somatic mutations, alterations in deoxyribonucleic acid (DNA) that occur after conception, which can ultimately have clinical impact (Wellcome 2019a, 4). Researchers on the programme are founding members and leaders of the ICGC (explored in Section 4.5) (Wellcome 2019a, 4). Cancer genome sequencing studies completed in the programme have led to the establishment of the widely used Catalogue of Somatic Mutations in Cancer (COSMIC), now a comprehensive resource for investigating somatic mutations in cancer. Programme work on descriptive genomics has also led to an enhanced understanding of clonal dynamics in normal adult tissue maintenance and patterns of natural selection acting on somatic mutations (Wellcome 2019a, 4). In the next five years,

the programme will continue descriptive genome sequencing and functional genomics research using cellular models.

Cellular Genetics

The Cellular Genetics programme was established in 2012 with two overarching projects: defining tissue homeostasis and immune responses by 'cell atlasing', and high-throughput dissection of human development at cellular and molecular resolution (Wellcome 2019b, 3). To meet the goals of these projects, Sanger has had a leading role in the HCA project, an international consortium to map human cells and tissues using single cell approaches, which has been described as a 'Google Maps' of the human body (Wellcome 2019b, 4). The head of Cellular Genetics, Dr Sarah Teichmann, co-founded the project and continues as a co-leader of the HCA as a project network of 1,562 scientists (Wellcome 2019b, 11). The Cellular Genetics programme has also developed several new technologies in single cell genomics and high-throughput spatial genomics, including induced pluripotent stem cells (iPSC) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technologies (Wellcome 2019b, 20).

Human Genetics

Human Genetics is the longest standing programme at Sanger Institute, emerging from their inception work on the Human Genome Project (Wellcome 2019g, 4). The programme now has three specific areas: population genetics, complex diseases and traits, and rare genetic disorders. The Human Genetics programme focuses on specific diseases within these broad research areas, such as inflammatory bowel disease, where researchers work closely with the National Health Service (NHS) to harness clinical data at scale. The DDD study has been an important project within rare genetic disorders research at Sanger, providing the NHS with genetic diagnoses for thousands of previously undiagnosed children (CS_10). The Human Genetics programme is closely aligned with the Open Targets initiative (an Associate Research Programme at Sanger), a collaborative partnership including Sanger Institute, EMBL-EBI and several pharmaceutical companies (Wellcome 2019g). Open Targets and DDD are explored in some detail as case studies in Chapter 4.

There are several joint projects between the Human Genetics and the Cellular Genetics programmes, such as HipSci, which generates stem cell lines for the research community, and they frequently share technical expertise in single cell ribonucleic acid (RNA) sequencing, for example (Wellcome 2019g).

Parasites and Microbes

Although the Parasites and Microbes programme was only launched in 2018, Sanger has been a global leader in pathogen research for decades, as one of the earliest genome sequencing projects at the Centre studied the causal agent of malaria, *Plasmodium Falciparum* (Wellcome 2019h, 5). The programme focuses on sequencing the genomes of pathogens and their vectors, especially endemic diseases in low- to middle-income countries (LMICs) (Wellcome 2019h, 5). MalariaGEN, another project described in Section 4.4 as a case study, is an important part of the strategic aims of the programme, transferring technologies and upskilling LMICs in genomic sequencing technologies to combat malaria. One of the interviewees stated that Sanger Institute is at the forefront of making genomic technologies more accessible as standard tools of disease control programmes worldwide with this programme (C6_06).

Tree of Life

The Tree of Life is the most recent Sanger Institute programme, beginning in November 2019 (Wellcome 2019i, 10). The goal of the programme is to sequence the genomes of over 60,000 eukaryotic species in the UK, collaborating with organisations such as Kew Gardens, whose staff assist in the sample collection (C6_07). In completing this project, Sanger aims to generate a much-improved means of assessing ecosystems, and how they adapt to climate pressures. The Tree of Life project is seen by many as a departure for Sanger, typically focused on human health. It aims eventually to understand other species' toxins and susceptibilities to infection, and how the characteristics of agricultural pests are indirectly critical to human health (CS_09). The Tree of Life project is explored in Section 4.2.

In addition to the scientific groups, Sanger institute hosts a number of non-scientific groups to support the research. One such group is the translation office, created in 2011 to maximise the healthcare benefits of the Sanger Institute's research.[11] Non-scientific groups include a dedicated grants office, a library and scientific customer support, among several others.[12]

Scientific operations at Sanger Institute

Underpinning all the scientific programmes at Sanger Institute is the work of 23 scientific operations groups. These are specialised, small groups that provide research tools, scientific expertise and analysis, and data for projects at the institute. Many of the groups work to support the sequencing capacity of Sanger, which includes processing RNA and DNA sequencing data using software packages; creating whole-genome shotgun libraries from extracted material and single cells; and a data quality control team.[13,14] Sanger also has a specialised long-read sequencing scientific operations groups, which use Pacific Biosciences; Oxford Nanopore Technologies; Bionano Genomics; and 10x genomics sequencing platforms.[15] The latter group is particularly important for the Tree of Life programme of work, requiring de novo assembly for eukaryotic organisms with little pre-existing genomic data.

In addition to sequencing, there are several scientific operations groups specialised in informatics within Sanger. Broadly, these groups facilitate the harvesting, storage and analysis of DNA genotype and sequence information at the institute.[16] The stem cell informatics group offers dedicated biomimetics and software development expertise to the Cellular Genetics programme, and hosts a specific database for genome editing.[17] To enable biomedical research at Sanger , there is a scientific operations team working on genetically modified mice.[18]

---

[11] As of 28 July 2020: https://www.sanger.ac.uk/non_science_group/technology-translation/

[12] As of 28 July 2020: https://www.sanger.ac.uk/about/groups/#Administration

[13] As of 28 July 2020: https://www.sanger.ac.uk/group/bespoke-sequencing/

[14] As of 28 July 2020: https://www.sanger.ac.uk/group/data-quality-control-qc/

[15] As of 28 July 2020: https://www.sanger.ac.uk/group/long-read-sequencing/

[16] As of 28 July 2020: https://www.sanger.ac.uk/group/cellular-genetics-and-phenotyping-informatics/

[17] As of 28 July 2020: https://www.sanger.ac.uk/group/stem-cell-informatics/

[18] As of 28 July 2020: https://www.sanger.ac.uk/group/mouse-pipelines/

There is a series of scientific operations focused on cytology at Sanger. The cytometry core facility facilitates running cellular samples, data analysis and experimental design.[19] Staff have the instrumentation required for the physical and chemical measurement of cells. There is also a scientific operation dedicated to fluorescence in situ hybridisation (FISH) to enable chromosome analysis.[20]

Sanger has dedicated information and communications technology (ICT) groups, supporting the computing infrastructure and developing and maintaining databases.[21]

2.1.2. Partner organisations at the Wellcome Genome Campus

The EMBL-EBI

The EMBL-EBI– Europe's flagship laboratory for basic research in molecular biology – provides data services and training that help scientists realise the potential of 'big data' in the biological sciences.[22] Sanger has a long-standing partnership with EMBL-EBI. They work together on the Open Targets project, sharing a vision of the importance of large-scale biological data, and the openness of its provision, for the future of biomedical science (see Section 4.1 for more detail). The data distribution mission of EMBL-EBI complements Sanger's mission of data generation and broader ethos of open access to data (Wellcome 2019f, 6). As evidence of this collaboration, since October 2014, Sanger has contributed to the (EMBL-EBI) European Genome-Phenome Archive (EGA) by sharing 370 studies and 500 datasets (Wellcome 2019e, 10), and has also utilised data from this resource.

GEL

GEL was set up by the Department of Health and Social Care in 2012 to deliver the 100,000 Genomes Project, a flagship project that will sequence 100,000 whole genomes from NHS patients with a rare disease and their families, and common cancers.[23]Sanger Institute has close links with GEL, providing input on the rare disease strand of the 100,000 Genomes Project (Wellcome 2019g). Scientific leads on the DDD project also advise GEL on its future strategy (Wellcome 2019g). The presence of GEL on the Wellcome Genome Campus facilitates continued collaboration, and staff have recently embarked on a joint initiative with the Sanger Institute to develop new sequencing technologies. GEL also works closely with Sanger spin out and co-campus tenant Congenica, as it is the sole clinical decision support partner.

BIC

The BioData Innovation Centre (BIC) opened in 2016 and now accommodates nine small-to-medium companies that are fully embedded within the wider Genome Campus (Wellcome 2019e, 17). The BIC houses companies spun out of Sanger and EMBL-EBI and accommodates others which would benefit from, and contribute to, the scientific culture of the Genome Campus. In 2021 and 2026, the BIC will continue to provide facilities to small- and medium-sized commercial enterprises operating under the theme of

[19] As of 28 July 2020: https://www.sanger.ac.uk/group/cytometry-core-facility/

[20] As of 28 July 2020: https://www.sanger.ac.uk/group/molecular-cytogenetics/

[21] As of 28 July 2020: https://www.sanger.ac.uk/group/information-communications-technology/

[22] As of 28 July 2020: https://www.ebi.ac.uk/about

[23] As of 28 July 2020: https://www.genomicsengland.co.uk/

genomes and biodata (Wellcome 2019f). With the expansion of the campus as a whole the expectation is to increase this provision substantially over the next decade to include more, and potentially larger, companies.

Connecting Science

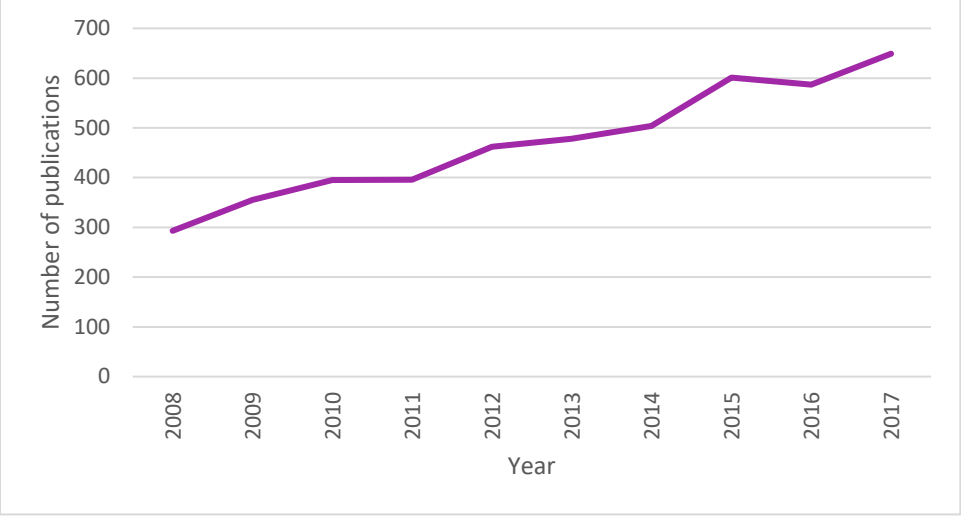
In 2016, Connecting Science was established as a new facility on Wellcome Genome Campus dedicated to public engagement with genomics and advanced training and courses. The Connecting Science building also operates as a conference centre. There is also a dedicated society and ethics research team within Connecting Science exploring the ethical questions posed by new genomic technologies. Between 2021 and 2026, Connecting Science will continue to deliver its programmes of courses and conferences, and aims to expand its global training reach to more participants in LMICs (Wellcome 2019f).

2.2. Outputs and contributions

2.2.1. Publications

Publications are a key output of Sanger Institute. Between 2008 and 2017, Sanger researchers contributed to **4,720** publications, spanning 116 Web of Science (WoS) subject categories.[24] Figure 3 shows how the annual publications output of Sanger grew consistently across this period, from 293 in 2008, to **649** in 2017.

Figure 3: Annual publications output of the Sanger Institute, 2008–2017



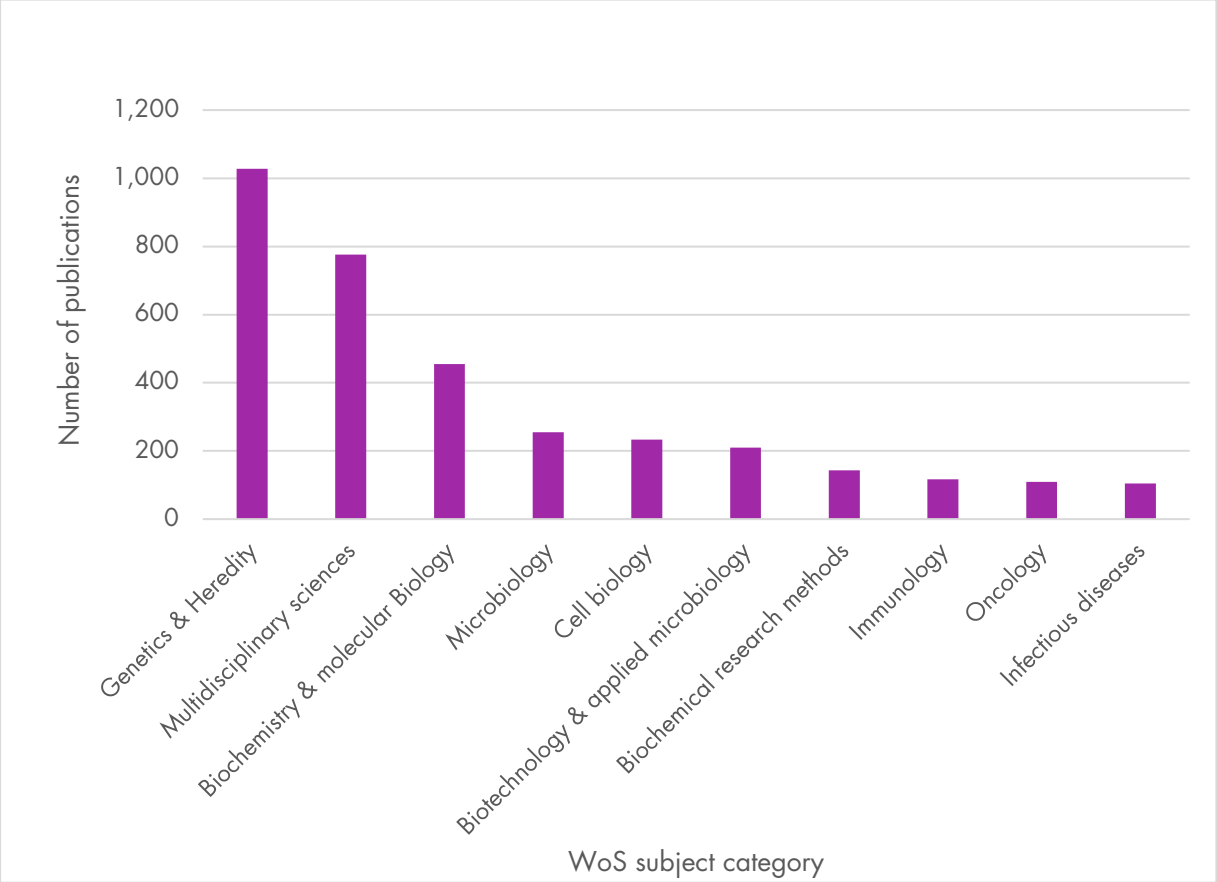
Source: RAND Europe analysis

The ten WoS subject categories in which Sanger Institute published the most papers between 2008 and 2017 were: genetics and heredity; multidisciplinary sciences; biochemistry and molecular biology; microbiology; cell biology; biotechnology and applied microbiology; biochemical research methods;

[24] Within WoS databases, each journal is assigned 1 (or more) of 252 subject categories. The subject categories enable measurement of the performance of papers (and journals) within the same or similar fields. Each published item inherits all subject categories assigned to the parent journal.

immunology; oncology and infectious diseases. Figure 4 shows the number of Sanger publications within each of these fields.

Figure 4: WoS subject categories in which Sanger Institute published the most papers in 2008–2017



Source: RAND Europe analysis

The MNCS of all Sanger Institute publications between 2008 and 2017 was **2.59**, showing that Sanger's publications were, on average, cited more than the world average in their field.[25] A MNJS of **2.16** indicates that Sanger publications are also published in journals with a higher impact, in citation count, than the world average.[26] Sanger also contributes a significant number of very highly cited publications. In total, **29.4%** of Sanger's publications fell within the top 10% most frequently cited publications in their field, while **5.6%** fell within the top 1% of most frequently cited publications.[27] Figure 5 presents the MNCS of

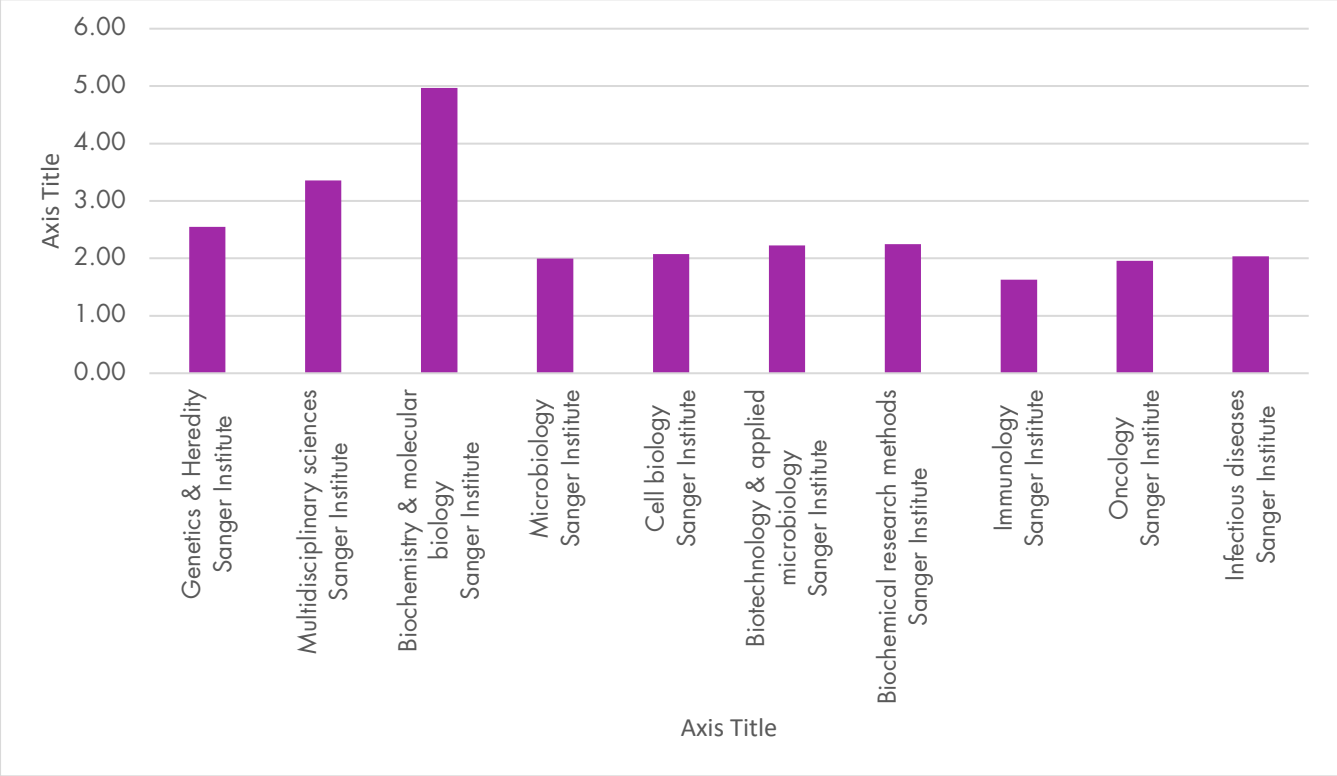
[25] MNCS is a measure of publication impact using number of citations, normalised to account for different citation patterns across fields of science and for differences in the age of publications. When MNCS is above 1, it indicates that an organisation performs better than the world average. When MNCS is below 1, it means that, on average, an organisation produces publications that are not cited as often as the world average.

[26] MNJS is a measure of the impact of the journals in which an organisation publishes. MNJS is calculated in a similar way to the MNCS, but uses the average number of citations of all publications published in the individual journals, normalised against the field and year of publication.

[27] Both these figures refer to the number and proportion of an organisation's publications that, compared with other publications in the same field and in the same year, belonged to the top 10% or 1% most frequently cited.

Sanger publications for each of the ten WoS subject categories in which it publishes most. The data suggests that Sanger publications within the field of biochemistry and molecular biology have been particularly highly cited.

Figure 5: The MNCS of all Sanger publications, 2008–2017, by Sanger's top ten WoS subject categories, by number of publications



Source: RAND Europe analysis

Most of Sanger's publication output is collaborative. The institutions with which Sanger co-authors most frequently are as follows: University of Cambridge; University of London; University of Oxford; Harvard University; and the United States Government.[28] Figure 6 presents a network map visualising the institute's co-authorship links with other organisations. The map shows the 200 organisations with which Sanger has the strongest connection in terms of co-authorship links. In Box 1, we summarise the key features of the network map.[29] The total link strength of Sanger within this network is **28,820.**

[28] These are the Sanger's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. These are the WHG's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. It should be noted that, in some cases, the listing of two institutions as authors may reflect a dual affiliation of a single authors (or authors). This may include dual affiliations between a home research institute and a parent organization.

[29] These key features are relevant to the co-authorship network maps presented in other sections of this report.

Box 1: Key features and concepts of the co-authorship network maps

Data sample of a network map: The data sample for each co-authorship network map is provided by the publication data for the institution that forms the focus of the network map, in this case the Sanger Institute. A co-authorship network map therefore shows patterns of co-authorship based only on publications on which the focus institution has also acted as a co-author.

Nodes: The basic feature of a co-authorship network map is the nodes. Nodes are the entities whose co-authorship connections are examined within the map, with each node represented by a circle. In the case of this analysis, each node represents a unique research institution. The size of each node is determined by its strength within the network, as measured by its total link strength (see definition below).

Links: Lines between nodes represent links (also referred to as edges), signifying the existence of co-authorship links between institutions.

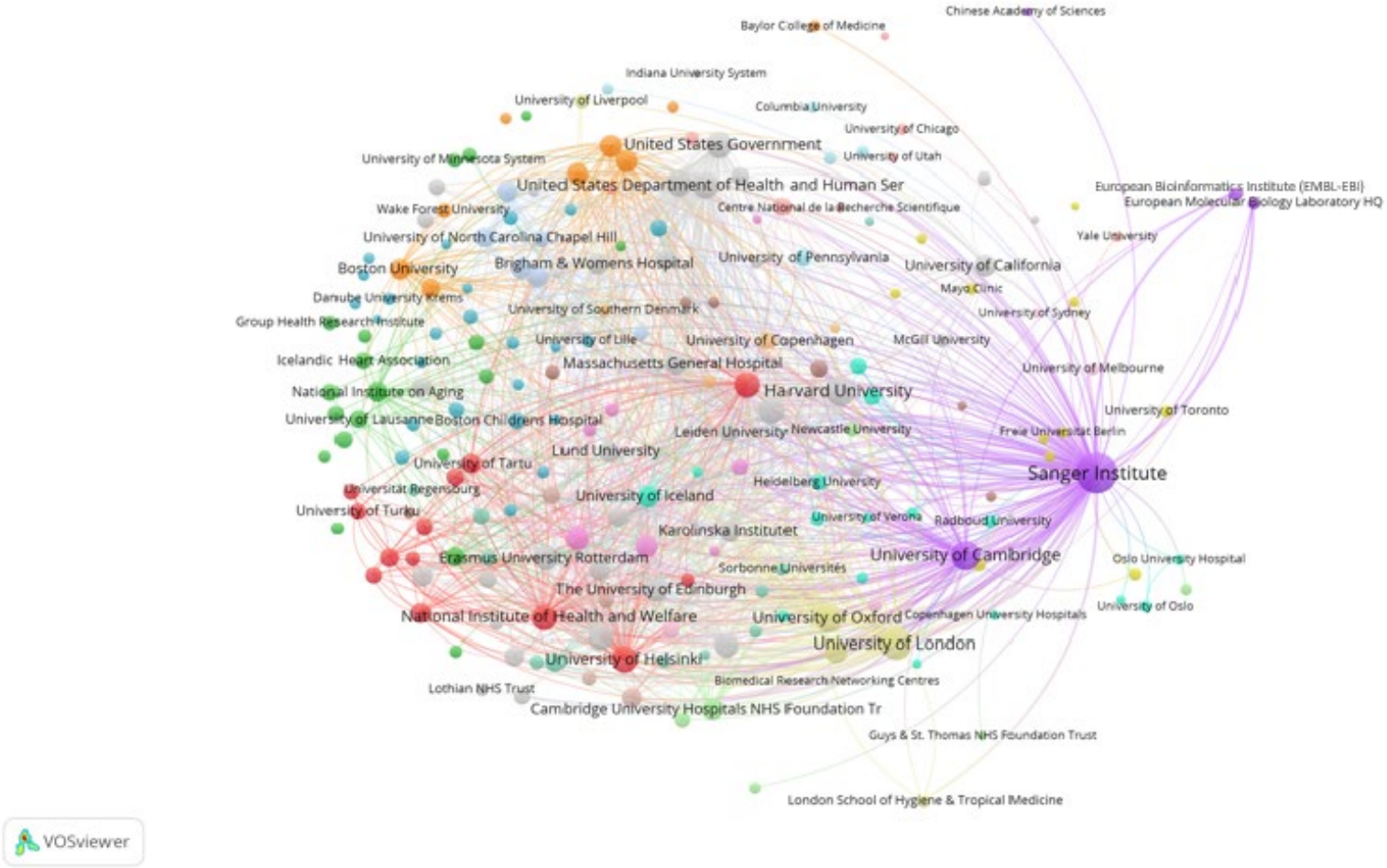
Link strength: Each link has a unique link strength (also referred to as weight), equivalent to the total number of publications on which the two institutions are listed as authors. The visual weight of the lines between nodes indicates the link strength.

Total link strength: Each node within the network map has a total link strength. The total link strength is derived from the sum of the weights of all edges that connect to a node.[30] The total link strength indicates the total strength of the co-authorship links of a research institution within the network. The size of a node is determined by its total link strength.

Clustering: With each network map, nodes are grouped into clusters of strongly connected nodes, as represented by colour variations. Clustering is performed using a modularity-based approach called 'smart local moving'. This algorithm takes into account data on link and link strength between nodes in order to identify locally well-connected communities within the network (Waltman et al. 2010). Clustering helps to visualise which nodes frequently co-author together when they co-author with the central node of the network map (e.g. the Sanger Institute).

[30] A single publication that involves multiple institutions will impart a weight of 1 on the edges between all involved institutions.

Figure 6: Sanger's co-authorship network map, 2008–2017



Source: Leiden University's Centre for Science and Technology Studies (CWTS) and RAND Europe analysis

2.2.2. Data and databases

As of June 2020, 23 open access reference databases freely available to the wider scientific community have been created at Sanger Institute. ENSEMBL, DECIPHER and COSMIC are among the most widely used of the databases and are described below. Beyond these specific examples, Sanger also contributes significant volumes of data to databases housed elsewhere, notably through its role in international collaboration.

Ensembl

Ensembl is a genome database project co-developed by Sanger Institute and EMBL-EBI in 1999 as a means of automatically annotating the genome and then integrating this information with other biological information.[31] Ensembl continues to function as a browser for vertebrate genomes and supports work in comparative genomics, evolution, sequence variation and transcriptional regulation.

DECIPHER

DECIPHER was developed in 2004 by then Sanger core faculty member Dr Nigel Carter and associate Sanger faculty and clinician, Dr Helen Firth.[32] DECIPHER has integrated genotypic and phenotypic data from 35,000 individuals and is available to researchers worldwide. The database was used and further developed in the DDD project, as it was the primary means by which clinicians and scientists shared data with each other. It is now an essential database for UK clinicians, and is also being used as part of the 100,000 Genomes Project.

COSMIC

Sanger Institute's cancer genome sequencing studies led to the establishment of COSMIC, which has become a comprehensive resource for investigating somatic mutations in cancer. COSMIC contains over 9 million coding mutations and a total of 41 million variants curated from 26,000 peer-reviewed publications and 84 large studies across all human cancer types (Wellcome 2019e).There are over 6,600 citations of COSMIC in peer-reviewed literature and the website has over 15,000 active registered scientists (Wellcome 2019e). In 2015, COSMIC introduced a licence charge for industry and for-profit organisations, while continuing to offer the full content to academic scientists for free. In 2019, the business aspect of COSMIC earned £1,300,000, enabling improved functionality and funding for the recruitment of twenty-six core staff (Wellcome 2019a).

Sanger Institute's role as a data producer and sharer

Since its inception, one of Sanger's primary roles within the genetics and genomics landscape has been as a data producer and sharer. Although Sanger holds some datasets in house (and shares them with partners), a significant volume of the data it produces contributes to wider (typically international) projects where the data is housed elsewhere. In either case, all partners share a goal of making the data accessible, shareable, interoperable and work together to ensure this regardless of where the datasets are housed. Sanger was one of the largest contributors to the Human Genome Project in the 1990s, and has continued to contribute

[31] As of 28 July: https://useast.ensembl.org/index.html

[32] As of 28 July: https://www.sanger.ac.uk/external_person/firth-helen/

large amounts of data to collaborative, cross-cutting projects such as the HCA, 1000 Genomes Project and the UK10K project (Int_02, Int_03, Int_05, Int_11, Int_13). One example of Sanger's role as a data producer is its contribution of genomic data to the 'data portal' of the ICGC. Through its leading role in four ICGC cancer genome projects (breast cancer, osteosarcoma, chronic myeloid leukaemia and prostate cancer), as well as its key role in ICGC's Pan-Cancer Analysis of Whole Genomes initiative, Sanger has been a key contributor of data to the ICGC portal, where data has been made freely available to researchers and clinicians around the world (for more information, see Section 4.5). Sanger has also made key data production contributions through the experimental arm of the Open Targets programme, where it has produced large datasets on the causal relationship between disease targets and diseases, which are made publicly and freely available through the Open Targets portal (CS_04, CS_05). EMBL-EBI has contributed expertise in bioinformatics and computing to help maintain and enhance this portal, which complements Sanger's role as a data producer for this portion of the project (Priego and Wareham 2018) (for more information, see Section 4.1). Another example of where Sanger will continue to build on its strength as a data producer will be through the Tree of Life programme, which will produce sequencing data related to biodiversity and feed into the wider Earth BioGenome Project (for more information, see Section 4.2).

When asked about the role of Sanger and Genome Campus in the genetics and genomics landscape, many internal staff identified data production as a key role of the institute (Int_02, Int_03, Int_05, Int_06, Int_11). However, some interviewees also expressed the view that Sanger has transformed beyond being exclusively a data generator to becoming a research institute in its own right, which a few described as a positive shift (Ext_01, Comp_03, Comp_04). If Sanger continues to pursue opportunities to build on its strength as a data producer, it should also continue to maintain its deep expertise in cellular biology, genetics and genomics, which allows it to be more than a 'sequencing factory' (Comp_04).

2.2.3. Training and capacity building

Another key area in which Sanger and the Genome Campus contribute is training and capacity building. Sanger's Graduate Programme aims 'to use the physical and academic resources of the institute to inspire and train the next generation of leaders in genomics' (Wellcome 2019d). It consists of three key programmes: a four-year PhD programme; a three-year clinical PhD programme run in partnership with the University of Cambridge, the University of East Anglia and Wellcome; and a one-year MPhil in genomic science programme. A key aim of the MPhil programme is to provide a stepping-stone to enable students from LMICs to move to PhD programmes (Int_06). To fulfil this aim, the programme is operated in partnership with nine institutions, offering Masters level training to three students from LMICs each year. (Wellcome 2019d).

Since the founding of Sanger Institute in 1993, a total of 279 individuals have undertaken an MPhil or a PhD at Sanger; 230 students have completed a PhD, and 24 students have completed the MPhil programme (Wellcome 2019d). A survey of 109 PhD students in the 2011–2015 PhD cohort revealed that there were at least 491 publications attributable to work undertaken by students in the cohort, including 409 unique publications (Wellcome 2019d).[33] Over 95% of this cohort had at least one publication and over three-

[33] This includes publications that had more than one outgoing student as an author.

quarters had one or more first author or joint first author publications (Wellcome 2019d). Aside from contributing to publications, graduate programme students have also participated in Sanger's various national and international collaborative research initiatives and presented their work at numerous national and international scientific meetings (Wellcome 2019d). According to data held by the graduate programme, approximately 69% of alumni are now working as researchers or PhD students in academia, industry, pharmaceutical companies or start-up companies, with 40% of these in senior roles. A further 6% of alumni are now either students in medical, veterinary, MD-PhD and MB-PhD programmes or working in clinical practice (Wellcome 2019d).

Alongside Sanger's graduate programme, the Wellcome Genome Campus makes a broader contribution to training through the work of Connecting Science, whose programme Advanced Courses and Scientific Conferences (ACSC) provides open post-graduate level training in genomics and biomedicine related topics across the following platforms: conferences and training courses hosted at the Genome Campus; overseas training courses delivered in LMICs; and free, open access online courses (Wellcome 2019c).[34] The focus of the training delivered through these platforms is to provide scientific researchers and healthcare professionals with an opportunity to learn more about genomics and to develop knowledge and skills that they can apply in their own work (Int_07). Many courses and scientific conferences are run in collaboration with other organisations on Wellcome Genome Campus, including Sanger Institute and EMBL-EBI. Between 2014 and 2019, ACSC delivered 272 conferences and courses across these four platforms, reaching 28,380 attendees (Wellcome 2019c). A review of feedback from 762 course participants showed that 63% of them stated that attending the course had contributed to the development of new research papers, while 48% said that it had helped them apply for grants (Wellcome 2019c). Partner organisations on the campus, such as EMBL-EBI, also have their own training offers delivered independently from the Sanger Institute.

2.2.4. Public engagement

Through Connecting Science, Genome Campus makes a broader contribution to public engagement in genomics research. This public engagement work takes of a number of forms, including campus visits, exhibitions, school visits, student placements, curriculum development and online learning resources (Int_07) (Wellcome 2019c). These initiatives create opportunities for public audiences to learn more about the science happening at organisations within the Wellcome Genome Campus, facilitating public awareness of, and engagement in, genomics (Int_07). Another important strand of Connecting Science's public engagement work is its contribution to research on ethical, legal and social issues connected to genomics (Int_07) (Wellcome 2019c). The Society and Ethics Research Programme conducts quantitative and qualitative social science research to explore the psychological, social and ethical impact of genetics and genomics. Recent work undertaken by this programme includes a global online survey focused on gathering public attitudes towards genomic data sharing, conducted in collaboration with the Global Alliance for Genomics and Health (GA4GH), and a study exploring stakeholder perspectives on the ethical challenges in the use of artificial intelligence for cognitive evaluation (SPACE).[35]

[34] As of 28 July: https://coursesandconferences.wellcomegenomecampus.org/

[35] As of 28 July: https://www.genomethics.org/

2.2.5. Translation and commercialisation

While academic publications are critical to Sanger Institute, it also has an interest in developing genomics research to be used in wider contexts. This work is facilitated by the translation office, created in 2011, which works to ensure that scientific projects lead to the development of new technologies, health and commercial benefits, and broader societal impacts.[36] The translation office works across all Sanger programmes and projects, with some stand out successes explored below.

DDD

As a result of the DDD project, a new biotech company, Congenica, was founded, which provides genomics expertise to the NHS. Sanger Institute staff work closely with Congenica. In addition, DDD staff, with help from Sanger's translation office, also worked with Oxford Gene Technology (OGT) to develop and commercialise new micro-array technology (CS_10). The new micro-array technology developed with OGT was, and continues to be, critical in detecting new pathogenic deletions and duplications. Clinical scientists have said that this higher resolution sequencing technology allows them to look for smaller changes and deletions in the exome than was previously possible.[37] As a result, the diagnostic yield (defined as a test's probability of returning information needed for a diagnosis) has increased from an average of 1–3% to 50% and is now used in many of the regional genetic services (Int_08). This improved array has been licensed, and has achieved more than £9m global sales, returning a 1% royalty to Sanger (Wellcome 2019g).

Open Targets

Open Targets is a public–private initiative, where Sanger acts as a key partner institute, aiming to develop new pharmaceuticals from genome-wide association studies (GWAS) and functional genomics data. Translation is clearly central to this programme, but as new drugs can take many years to be developed to market, it is not possible to determine its exact translational and commercial impacts (CS_05). However, Open Targets is already having a demonstrated financial impact: since its launch in 2014, £50m has been raised for research, with 40 projects awarded to Sanger at a value of £24.8m (Wellcome 2019e). Open Targets, a Sanger Associate Research Programme. is explored in more detail in Section 4.1.

Microbiotica (with Genentech)

Microbiotica was launched in December 2016 with the aim of becoming a global leader in new bacteriotherapies based on the human gut microbiome, founded at Sanger Institute (Wellcome 2019e). The company retains a close association with Sanger and its founders and is based at Wellcome Genome Campus. Microbiotica is developing a new class of therapeutics and has secured a US$500m co-development deal with Genentech (Wellcome 2019e).

[36] As of 28 July: https://www.sanger.ac.uk/non_science_group/technology-translation/

[37] https://www.ogt.com/resources/literature/1359_superior_detection_of_chromosomal_aberrations_using_the_latest_generation_of_exon-focused_constitutional_arrays

2.2.6. Leadership and collaboration

Leadership of various international consortia and research collaborations

Another key contribution of Sanger has been its role as a leader of cross-organisational research collaborations and consortia. Sanger can play this role through its infrastructure and strong research team. Beginning with its key role in the Human Genome Project, Sanger Institute has participated in and played a key leadership role within a large number of research collaborations, within the UK and internationally (Int_11). These include the 1000 Genomes Project, the UK10K project, the DDD project, the International Common Disease Consortium, the ICGC, the Cancer Dependency Map, the Human Cancer Cell Model Initiative, the HCA project, the Tree of Life project and the Global Alliance for Genomics and Health (Int_11). In addition to coordinating international research efforts and reducing duplication, these collaborations have had a number of other benefits, including sharing best research practices, developing solutions to common challenges and problems, and bringing together diverse genomic data in one place, and making it publicly available (Int_11, Int_13). By involving partners from the clinical field, some collaborations led by Sanger have also helped to translate genomics research into clinical practice. In Chapter 4 we consider a number of case studies that demonstrate the key role that Sanger Institute and has played in fostering and leading collaborations.

2.3. Summary of role in field and characteristics

Interviews with Sanger staff and associates, industry figures, comparator organisation staff, clinicians, collaborators and research users provided insight into the institute's role in the genomics landscape. In this section, we present common themes and summarise findings from these interviews regarding the characteristics and operation of the Sanger Institute.

2.3.1. Working at scale

Several internal and external interviewees emphasised that Sanger's capacity to undertake genetic and genomic science at scale is one of its key strengths (Ext_03, Int_13, Comp_04, Int_08, Int_05). One described Sanger as being one of few organisations, globally, able to complete high-quality 'science at scale' (Int_13). Interviewees at comparator organisations have said that strong analysis capabilities and large-scale sequencing programmes at Sanger allow the institute to complete projects that they would not be able to undertake at their organisation (Comp_04). This 'science at scale' is demonstrated in projects such as the DDD project, where Sanger was charged with sequencing and analysing genomic data from thousands of patients. Projects like this require large-scale sequencing capabilities, and Sanger is one of the few global genomic research centres positioned to lead in this domain (Int_08, Int_05).

2.3.2. Collaborative ethos

One internal figure at Sanger highlighted that conducting such large-scale scientific projects is only achieved through collaboration with co-tenants on the campus and external partners (Int_08). Another interviewee described Sanger's strong collaborative ethos as one of its defining characteristics, demonstrated by the typically long author list on its publications (Int_03). Sanger has formed a wide variety of working relationships and partnerships with many industry and academic institutions such as GEL, Health Data

Research UK and the Broad Institute. Sanger also works closely with the NHS, obtaining clinical samples while providing genomic data and analysis. Because of the concentration of genomic expertise at Sanger, its staff are well placed to collaborate in and lead international consortia, such as the ICGC and the HCA (Int_13). Another interviewee pointed to the presence of EMBL-EBI on the campus and general computational strength of Sanger as being central to leadership in such consortia (Ext_06).

2.3.3. Limited clinical access

Internal and external staff of Sanger Institute observed that one challenge faced by Sanger is that it lacks direct access to clinical samples and material (Ext_06, Comp_03), and needs to strategically partner with the Cambridge University Medical School and Cambridge University Hospitals to gain access to samples, including through its involvement with Health Data Research UK. This is compounded by the fact that comparator genomic research organisations, embedded within a medical school, have rich access to clinical samples and can subsequently perform more analysis. One interviewee also suggested that Sanger's relative isolation from a hospital setting means there is also difficulty in recruiting clinicians (Ext_01). While out of the 32-core faculty at Sanger, eight are clinically trained, one interviewee said that this was too few and limits the range of scientific questions and work that could be asked (Ext_03). As a result, Sanger is less able to translate research into actionable impact on human health or clinical practices. Finally, some interviewees said that, despite working closely with the NHS on DDD, the institute could be doing more work with national and international clinical centres (Ext_03, CS_12). However, we note that the Sanger has a wide range of collaborations with clinical partner which allows them to access clinical samples for a range of projects – but that this collaborative ethos is needed to help address what would otherwise be an important gap.

2.3.4. Openness and commitment to data sharing

This strong collaborative ethos goes hand in hand with Sanger institute's commitment to open access and data sharing (Ext_03). Interviewees external to Sanger have highlighted this as a key strength, benefiting the global genomics research community (Ext_01). While many interviewees identified the Broad Institute as being strongly collaborative and sharing data openly, some claimed that Sanger is stronger in this regard (Int_09, Int_12). In the Open Targets programme, all data is eventually released to the entire research community and pharmaceutical industry, increasing the likelihood of drug discovery (Int_09). Similarly, on the DDD study, data was shared globally, enabling many more diagnoses and publications than would have otherwise been possible (CS_11).

2.3.5. Translation and commercialisation approach

Linked to the ethos of openness, many interviewees identified Sanger Institute's approach to translation and commercialisation as a potential challenge (Ext_03, Ext_04, Ext_05, Int_08), although it is one that is also connected to benefits associated with open data policies and sharing research findings widely. Reasons cited for this challenge include scepticism of commercialisation among Sanger scientists (Int_08), insufficient support for spin-out companies (Ext_05) and an unwillingness to provide exclusive intellectual property (IP) licensing rights, which fails to attract investors (Int_08, Ext_05). While this culture has improved (one interviewee working in industry spoke highly of their commercial partnership and another

saw entrepreneurship as one of Sanger's key strengths), issues remain (Int_12, CS_09). It was suggested that Sanger has not been sufficiently entrepreneurial, missing opportunities around CRISPR technologies and gene editing, and being unwilling to take risks (Ext_04). Some suggestions that were provided to address this shortcoming include: an 'entrepreneur in residence programme' (Int_08); a more holistic approach that takes into account the full gamut of science and innovation activities on a structural level; and putting more funding into early technologies, which could help to improve commercialisation at Sanger Institute (Int_08).

As Sanger is strongly committed to open data, a potential conflict of interest emerges when spinning out new biotechnology companies: it cannot simultaneously release data to all and give a competitive advantage to spin outs through exclusive IP licensing, although this something that is considered in Sanger's open data policy. This challenge was recognised by both Sanger staff and an interviewee at a spin out, reflecting a broader question of priorities and general strategy (Ext_05, Int_08).

Overall Sanger's approach to translation could be considered to take a broader perspective, aiming to maximise the use of data and science through open sharing that enables widespread use and translation beyond and across commercial contexts.

2.3.6. Holistic approach to genomics

Several interviewees highlighted Sanger's joined-up approach to genetics and genomics research, including training, public engagement and bioethics research, as a strength (Ext_03). Central to this is the Connecting Science facility, which provides regular genomics training in the UK and LMICs. One interviewee claimed that this is a unique strength of Wellcome Genome Campus, as no other organisation is providing genomics training on this scale worldwide (Int_07). Similarly, the Open Targets programme has a specific outreach programme, regularly offering hands-on workshops on how to access and use its platform (Int_07). In doing so, Open Targets staff are upskilling the research community.

2.3.7. Core research grant

One interviewee claimed that core funding facilitated an open and collaborative research environment within Sanger (Int_06). Other interviewees have said that core funding enables Sanger to focus on longer-term projects (Int_07, Int_08). Perhaps because of this longer-term vision, external interviewees agreed that Sanger has been at the leading edge of developing new genomic technology: they were one of the first to buy into Illumina Next Generation Sequencing, and more recently has developed its own GWAS tools and methodologies (Ext_02, Comp_04). In contrast, at some comparator organisations, interviewees have said that funding is only provided for a short time, which according to one interviewee worries their scientists, as research programmes could be pulled at short notice (Comp_04). One interviewee even identified sustainable funding as *the* major challenge to their organisation (Comp_02)

2.3.8. Nimbleness

Some interviewees said that as Sanger Institute works on such a large scale, they are less able to respond to emerging issues and new technologies (Comp_04, Int_02). One interviewee even described Sanger as a 'production line' and not particularly strong at smaller, more detailed projects (Comp_04). An internal interviewee agreed with this, saying: 'we are not the right people to investigate the detail or nuance of

particular research questions', as the large-scale work they do requires 'rigidity' and fixed schedules (Int_02). It may be that due to its scale and long-term focus, Sanger is less able to respond quickly to new and upcoming topics in genetics and genomics.

2.3.9. Broader challenges for the field

Brexit

As in other scientific organisations across Europe, Brexit is a cause for concern. One Sanger Institute interviewee said that loss of access to European science funding would be a major challenge to the organisation; they are dependent on such funding for large international projects, and UK research councils do not provide similar opportunities (Int_12). Potential restrictions on the movement of scientific professionals and graduate students are another concern among Sanger staff (Int_06). Sanger has already been proactive in this regard, campaigning against Tier 2 Visa caps, a recommendation that has been adopted by the Migratory Advisory Committee. One interviewee expressed concern that Brexit could also affect Sanger's significant research and training initiatives in LMICs, such as MalariaGEN and the Connecting Science outreach programmes (Int_06).

Personnel

One interviewee said that the national shortage of bioinformaticians is a challenge to all genomic research, including at Sanger Institute (Int_13). However, Sanger staff have been proactive on this issue, working with Anglia Ruskin University to develop an undergraduate bioinformatics programme, and making a dedicated two-week bioinformatics courses available through the Connecting Science programme (Int_13). Moreover, one interviewee said that Sanger is unusual in the UK for having a wealth of expertise in bioinformatics, which is not always the case in other organisations (Ext_05).

2.3.10. Looking forward

While facing major organisational and scientific challenges, there are also opportunities for Sanger Institute to build on its strengths and existing projects. One Sanger interviewee identified the HCA, co-led by faculty in the Cellular Genetics programme, as a 'minefield of opportunity' for translation targets (Int_08). While the translation prospects from the Tree of Life project are less immediate, Sanger staff and external interviewees are confident that this project has the potential to make the institute, and the UK more broadly, a world leader in biodiversity research (CS_07, Ext_01). External interviewees see the Tree of Life project as being a pioneer in the use of genomics in ecosystems research and are already lending their expertise to international partners seeking to establish similar programmes (CS_08, CS_14).

While large-scale sequencing is increasingly competitive, Sanger will continue to provide large-scale DNA sequencing services for organisations like UK Biobank, where it will deliver 275,000 whole human genomes. It also has the opportunity to develop and implement new long-read sequencing technologies and nanopore technologies (Ext_01). Similarly, Sanger is at the forefront of developing new genomic engineering technologies, which combined with single cell sequencing offer great opportunities (Int_01). They could also complete more work in synthetic genomics, a current area of focus of some associate faculty, but with good potential to be extended into a core research focus (Int_08).

Sanger Institute has the opportunity to engage with the pharmaceutical industry by developing large-scale functional genomic screens, funnelling profits from this into other research areas (Int_08). In embarking on this kind of project, Sanger Institute would have to consider the mission of GRL to approach technology translation as 'a deliverer of societal benefit ahead of financial return' (Wellcome 2019f).

Finally, as the Genome Campus expands, there is the potential to grow the MPhil programme, and provide work experience to graduates students in the companies on campus (Int_06). This would help to establish a new generation of entrepreneurial genomics researchers, and contribute to the further integration of the campus. The expansion of the campus also provides the opportunity to increase its commercial presence, allowing more and larger companies in the BIC, for example.

3. Comparator institutions

In this chapter, we describe and analyse each of the four comparator organisations: the Broad Institute, the WHG, Janelia Research Campus and the NHGRI. We based our choice of these comparator organisations on initial desk research conducted by the study team in consultation with Wellcome staff. Each comparator organisation is explored, and their key features, contributions to the field of genetics and genomics, and impact in academia and beyond are described. Each comparator organisation is also compared with Sanger Institute, highlighting the key similarities and differences between the organisations. These comparator organisations help to contextualise Sanger within the wider research landscape, and benchmark Sanger's performance against comparators and competitors working in similar areas. Information in this chapter is taken from the website of each comparator organisation, unless otherwise noted.

3.1. Broad Institute

3.1.1. Background

The Broad Institute was founded in 2004, and was based on the work of the Human Genome Project and history of informal collaboration between the Massachusetts Institute of Technology (MIT) and Harvard University. The Broad Institute is located in the United States in Cambridge, Massachusetts, close to the Boston metropolitan area. It has a strong partnership with MIT, Harvard University and Harvard-affiliated hospitals. The Broad Institute's mission is to 'propel the understanding and treatment of human disease by tackling the most critical challenges in biology and medicine'[38] and is based on a model of integrated working across disciplines of biology, chemistry, mathematics, computation, engineering, medical sciences and clinical research. The Broad Institute's focus is wider than just genetics or genomics, also encompassing biology and other disciplines (Comp_02), although the focus of desk research and interviews with Broad Institute staff have focused primarily on genetics and genomics for the purposes of this study.

The Broad Institute's scientific focus areas are chemical biology and therapeutic science; drug discovery; genome regulation, cellular circuitry and epigenomics; immunology; medical and population genetics; and metabolism. Each of these areas is a scientific programme within Broad Institute. The Broad Institute focuses on disease areas through programmes encompassing cancer, cardiovascular disease, diabetes, infectious diseases, kidney disease, obesity, psychiatric diseases and rare diseases. The Broad Institute is organised into technological areas through platforms led by staff scientists in data science, genetic perturbation, genomics, imaging, metabolomics and proteomics. Programmes at the Broad Institute are

[38] As of 28 July: https://www.broadinstitute.org/strategic-alliances-and-partnering

characterised as communities of academic and staff scientists who share research and knowledge, while platforms provide the expertise, technology and organisation to coordinate projects. The Broad Institute contributes to a number of community-focused activities, including by providing community activities for educators, public lecture series and commentaries, post-doctoral fellowships for early career scientists. and undergraduate and post-baccalaureate opportunities for students.

The sheer scale of the Broad Institute and its broad expertise across areas in genetics and genomics, similar to Sanger Institute, have been important factors in its success, as the Broad Institute is able to take on projects for which smaller organisations may not have capacity (Comp_01, Comp_02).

3.1.2. Operation and key features

Funding sources

The Broad Institute is funded through a number of sources, including competitively won federal grants, industrial revenue, an endowment of over US$500m from Eli and Edythe Broad, and other philanthropic sources.[39] In 2019, the Broad Institute received over US$547m in revenue. It supports a small amount of research through its endowment and core funding, although much of the work conducted at the Broad Institute is supported through competitively won grants.

Staff

All faculty at Broad Institute are full employees of their home institutions (MIT, Harvard or Harvard-associated hospitals), including 15 core institute members with primary laboratories within Broad Institute facilities, 51 non-core institute members who lead projects within the Broad Institute and more than 300 associate members and affiliate members who participate in the Broad community; they sometimes lead projects and are eligible for internal funding from the Broad Institute. The Broad Institute also employs nearly 500 staff scientists who have a PhD or equivalent but are not part of the Broad faculty and work in platforms that crosscut the Broad Institute's research. A rotating group of 21 institute scientists are selected from this wider group every four years to play key roles in engaging connections across the Broad community, determining key scientific priorities and recruiting and mentoring scientists within Broad Institute.

Academic collaboration

The Broad Institute's faculty members are all full staff at their home institutions of either MIT, Harvard or Harvard-affiliated hospitals, which links Broad Institute to its partner institutions. An interviewee noted that this close relationship with MIT and Harvard has allowed the Broad Institute to attract world-class researchers and gain access to clinical samples from Harvard-associated hospitals, which has been mentioned as an important factor that has contributed to the Broad Institute's success (Comp_02). A Nature Index article on the role of physical proximity in fostering collaboration noted that the Broad Institute's physical proximity to Harvard and MIT in Cambridge, Massachusetts, is an important factor in its success, as it fosters the possibility of chance encounters, close collaboration and knowledge exchange (Savage 2016). The article identified the collaboration between MIT and Harvard as one of the strongest between two

[39] As of 28 July: https://www.broadinstitute.org/history

institutions in the world that is based on co-authorship, which is partially driven by the organisations' joint steering of the Broad Institute (Savage 2016).

3.1.3. Outputs and achievements

Data and databases

The Broad Institute makes data available publicly through epigenomic datasets where possible,[40] and has published many open access software and tools for data analysis.[41] For example, the Broad Institute resources were highlighted in resource reviews of multi-assay genomic data resources (Kannan et al. 2015) and of databases and web tools for cancer genomic studies (Yang et al. 2015), demonstrating the Broad Institute's utility to researchers in the field of genetics, genomics and cancer. The Broad Institute has also been a leader in cloud-based data storage and processing, including through a collaboration with Google to launch its Genome Analysis Toolkit (GATK) on Google's cloud platform (Terry 2015), reducing the cost of genome sequencing to approximately $5 per genome (Sheffi 2018). These activities have increased research capacity in genetics and genomics and contributed to further research in the field.

Translation and commercialisation

Along with the academic institutions and hospitals that are associated with the Broad Institute, pharmaceutical, biotech and technology companies are important partners for the Broad Institute. Since early in the Broad Institute's development, the institution has been able to attract private-sector investment, particularly for research that bridges the gap between laboratory results and results that can be taken up by the private sector (von Schaper 2017). The technology platforms at Broad Institute, along with providing essential scientific capabilities to Broad Institute faculty, provide genomic sequencing services to private companies on a fee-for-service basis. This helps keep their high-throughput pipelines filled while also maintaining leadership from scientists within Broad Institute and ensuring that interests between the Broad Institute and the private sector are aligned (Comp_01, Comp_02).

The Broad Institute has a number of spinoff companies, including Celsius Therapeutics, focused on drug development for autoimmune diseases and cancer immunotherapies (Kincaid 2018) and Beam Therapeutics and Editas Medicine, both focused on gene editing (DeAngelis 2019). The Broad Institute's approach to engagement is such that wherever possible, datasets are made widely available, and are shared with academic institutions and non-profit research institutes at no cost. Where there is potential benefit to health, the Broad Institute collaborates with industry through non-exclusive licensing wherever possible, although it has used exclusive licensing in some cases, such as through its exclusive licensing of CRISPR-Cas9 for genome editing in mammals to Editas Medicine. However, its patent for CRISPR-Cas9 has been hotly contested in the United States and Europe, resulting in considerable controversy for the organisation (Sherkow 2016).

The Broad Institute holds the patent for applying CRISPR-Cas9 to edit genomes of mammals and multiple cell organisms, opening up new possibilities for the treatment of genetic illnesses, and also developed

[40] As of 28 July: https://www.broadinstitute.org/epigenomics/data

[41] As of 28 July: https://www.broadinstitute.org/data-software-and-tools

SHERLOCK, an application of CRISPR where it can be used for a low-cost diagnostic tool (Cooke and Crew 2019). The Broad Institute has also worked on agricultural applications of CRISPR gene editing technologies, which may help to reduce food waste, limit pesticides and improve drought resistance (Pairwise 2019).

Leadership and collaboration

The Broad Institute has a strength in large-scale high-throughput genomic sequencing – the Genomics Platform at Broad Institute produces more than 500 terabytes of genomic data per month, and has reportedly been the largest producer of genomic information in the world for the past ten years.[42] Eric Lander, the president and founding director of the Broad Institute, played a lead role in sequencing for the Human Genome Project, and the Broad Institute's prowess in this area continues today with key roles in international genomic projects such as the HapMap project, 1000 Genomes Project, The Cancer Genome Atlas (TCGA) and the HCA initiative. The Broad Institute has claimed a leading role in mobilising and organising some international collaborations around genome sequencing in its official communications and website, including the HCA and 1000 Genomes Project (alongside Sanger Institute) and an international effort to make Ebola genome sequences openly available, which helped catalyse additional research (Yozwiak, Schaffner and Sabeti 2015). The Broad Institute has also made notable contributions to cancer genomics through projects such as TCGA, the Cancer Therapeutics Response Portal and the Cancer Dependency Map (DepMap) (Comp_02), as well as in the area of psychiatric illness, including through the largest genomic study to date on psychiatric disorders and a landmark 2016 study showing that schizophrenia is linked to a specific gene.

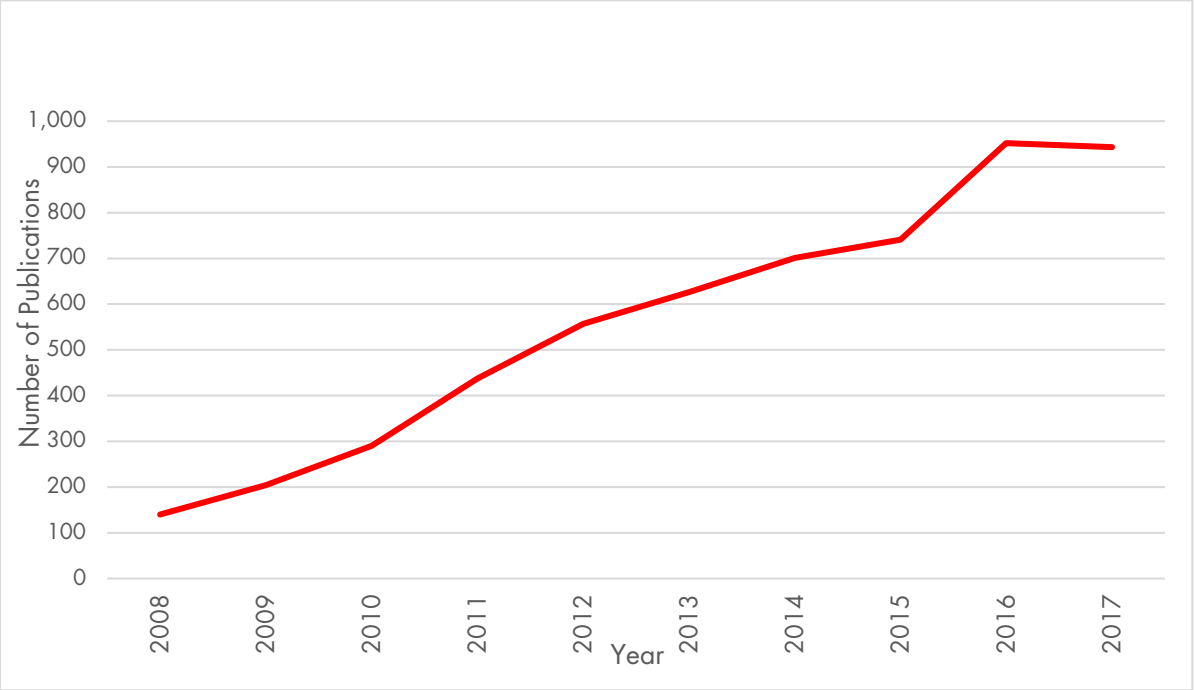
Publications

The Broad Institute produced **5,591** publications between 2008 and 2017, spanning 122 WoS subject categories. Figure 7 shows the annual publications output of the Broad Institute across this period, which rose from 140 in 2008, to **943** in 2017.

The ten subject categories in which the Broad Institute published the most papers between 2008 and 2017 were: **genetics and heredity, multidisciplinary sciences, biochemistry and molecular biology, microbiology, cell biology, oncology, biotechnology and applied microbiology, neurosciences, biochemical research methods, endocrinology and metabolism, and immunology**. Figure 8 shows the number of publications for each of these fields.

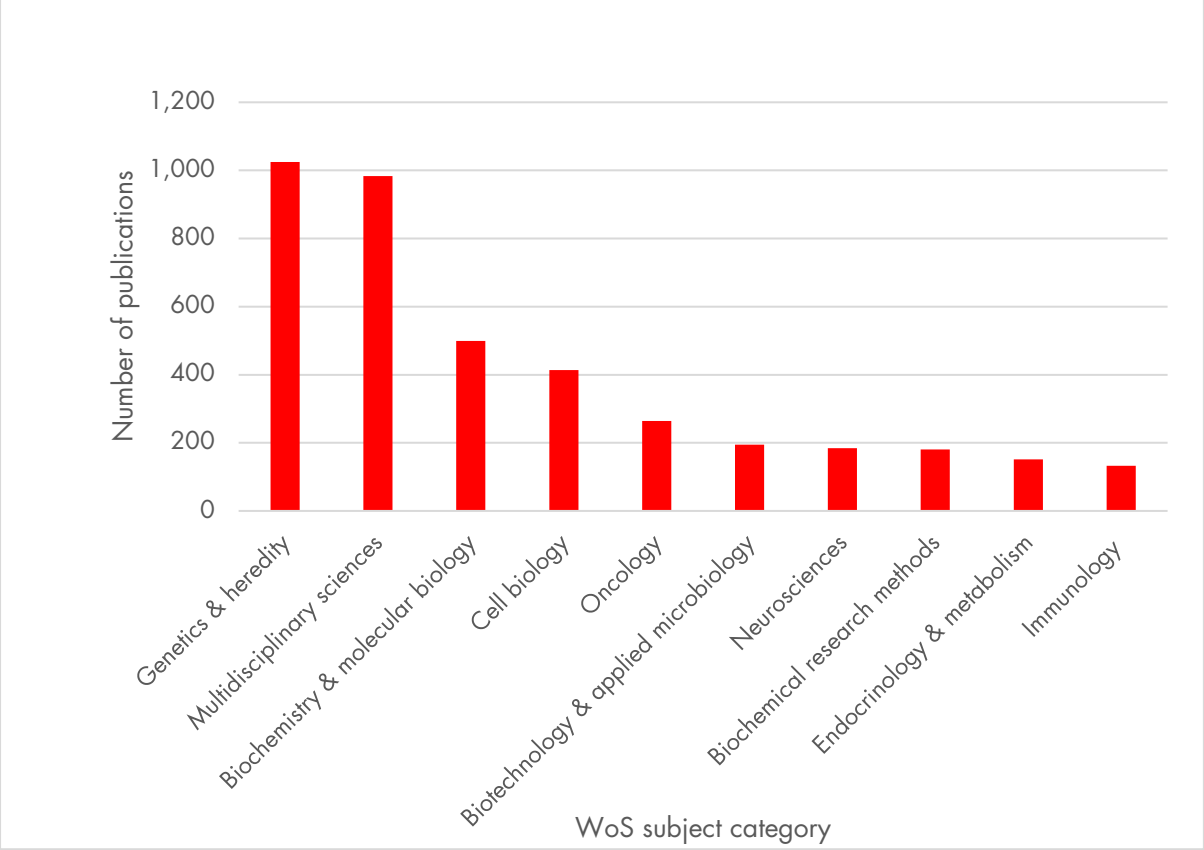
[42] As of 28 July: https://www.broadinstitute.org/genomics

Figure 7: Annual publications output of the Broad Institute, 2008–2017



Source: CWTS and RAND Europe analysis

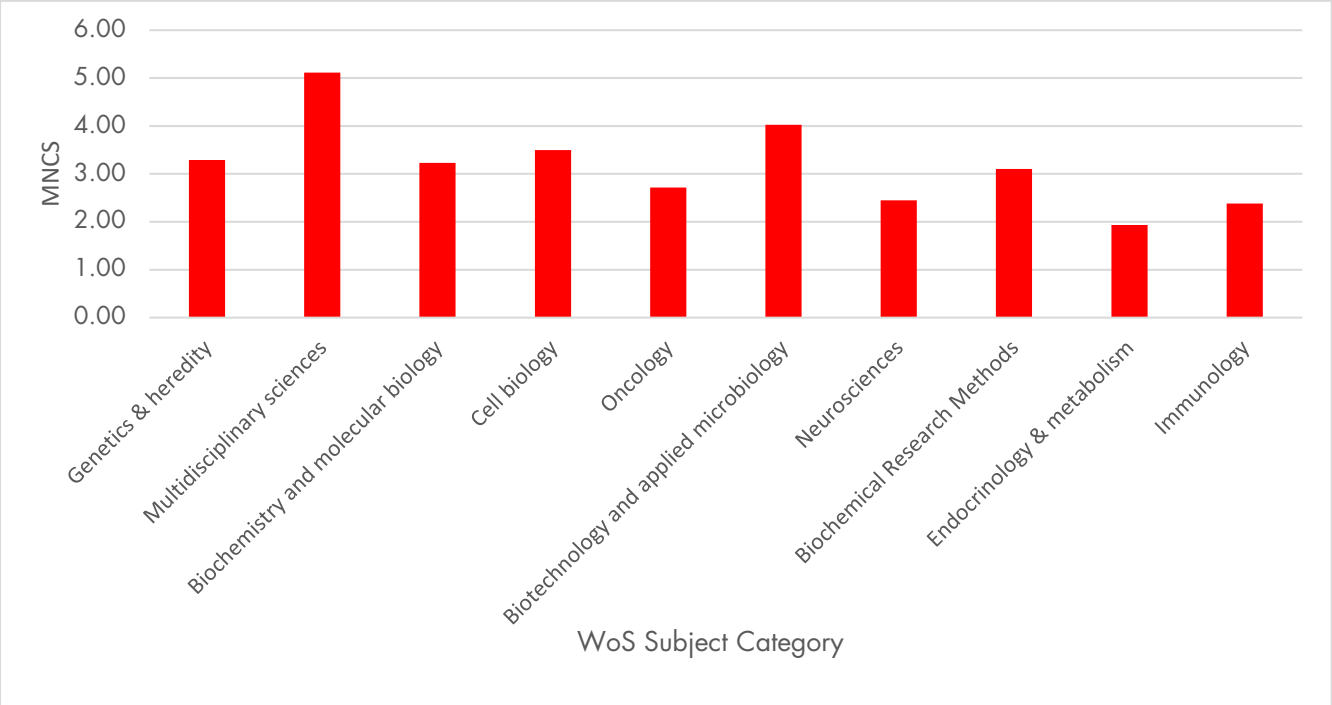
Figure 8: The WoS subject categories in which the Broad Institute published the most papers in 2008–2017



Source: CWTS and RAND Europe analysis

During this period the MNCS of Broad Institute publications was **3.15** and the MNJS was **2.79**. In total, **39%** of the Broad Institute's publications fell within the top 10% most frequently cited publications in their field. Meanwhile, **7.8%** were among the top 1% of most frequently cited publications. Figure 9 presents the MNCS of the Broad Institute's publications for the ten WoS subject categories in which it publishes most.

Figure 9: The MNCS of all Broad Institute publications, 2008–2017, by the Broad Institute's top ten WoS subject categories, by number of publications

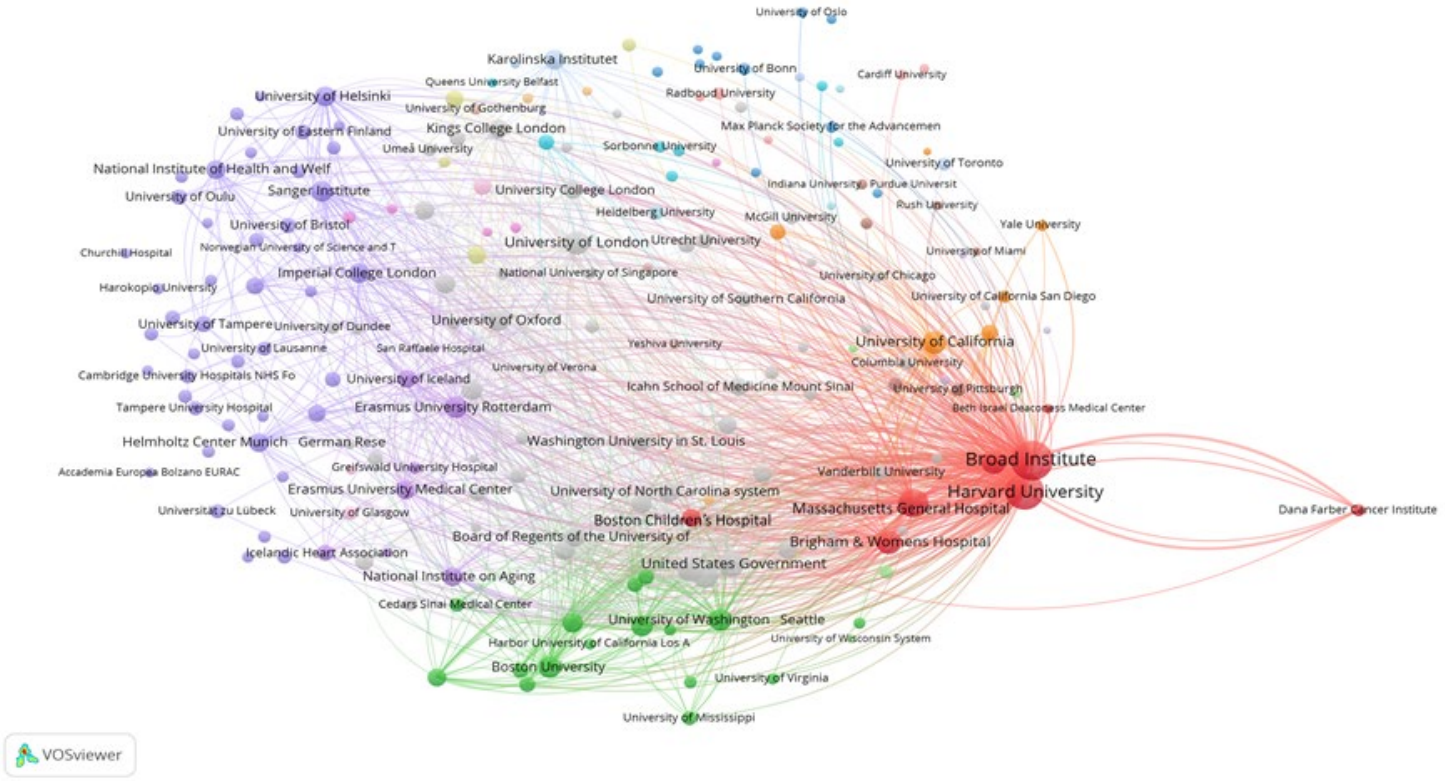


Source: CWTS and RAND Europe analysis

The institutions with which the Broad Institute co-authors most frequently are as follows: Harvard University; Massachusetts General Hospital; Massachusetts Institute of Technology; Brigham & Womens Hospital; and University of California.[43] Figure 10 presents a network map visualising the 200 organisations with which the Broad Institute has the strongest connections through its co-authorship links. Information on the key features of a co-authorship network map were provided in Section 2.2.1 above. The total link strength of the Broad Institute within this network is **39,338**.

[43] These are the Broad Institute's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. It should be noted that, in some cases, the listing of two institutions as authors may reflect a dual affiliation of a single authors (or authors). This may include dual affiliations between a home research institute and a parent organization.

Figure 10: The Broad Institute's co-authorship network map, 2008–2017



Source: CWTS and RAND Europe

3.1.4. Comparison to Sanger Institute

Table 3 summarises some of the key similarities and differences between the Broad Institute and Sanger Institute. Of the comparator organisations included in this study, Broad Institute is the most directly comparable to the Sanger Institute as they are similarly well funded, are large in scale and have an international reputation for excellence in genetics and genomics.

Table 3: Comparison between the Broad Institute and Sanger Institute

Key similarities	Key differences
• The Broad Institute and Sanger both excel in high-throughput genomics and the production of large genomic datasets. • The Broad Institute and Sanger have relatively broad areas of expertise, focusing across scientific areas, methods and disease areas. • The Broad Institute and Sanger and Wellcome Genome Campus both have physical campuses located in areas with high levels of research and innovation activity. • The Broad Institute and Sanger both work on large international collaborations, including HapMap, the 1000 Genomes Project and the HCA. Both institutions claim leading roles on large-scale projects.	• The Broad Institute is a leader in genome editing using CRISPR, while Sanger has not focused on this area. • The Broad Institute has used exclusive licensing agreements to help attract private funding, while Sanger has only pursued private sector involvement when this can be accomplished openly. • The Broad Institute and Sanger both have large amounts of core funding available, although Broad Institute researchers rely much more heavily on competitively won grants. • Sanger staff are fully associated with the institute , while all Broad Institute members are full staff members at member organisations. • The Broad Institute is closely aligned with academic institutions (MIT and Harvard) and Harvard-associated hospitals, giving them more access to clinical data.

The most notable differences between the Broad Institute and Sanger are the different approaches to commercialisation, with the Broad Institute entering into more willing to enter into closed agreements with individual companies than Sanger, which has entered into multilateral open agreements with the private sector. The Broad Institute's approach may spur private-sector innovation in areas that are less attractive when all organisations have the same access to information, although it also provides a competitive advantage to particular companies and limits open access to information. On the other hand, Sanger's strategy of open collaboration increases access to knowledge and information that can help spur innovation, although the private sector may find some areas unattractive for investment where they are unable to benefit from legal IP protections. The Broad Institute has closer ties to academic institutions and hospitals, which gives them access to clinical samples and research talent, but makes it more difficult to disentangle the contributions of these institutions from the unique contributions of the Broad Institute.

3.2. Wellcome Centre for Human Genetics (WHG)

3.2.1. Background

The WHG is a research institute of the Nuffield Department of Medicine at the University of Oxford, UK, which has its own purpose-built laboratories at the University of Oxford's Biomedical Research Campus in Headington, UK. It was founded in 1994 as the Human Genome Project was being conducted internationally, with financial support from Wellcome. The aim of WHG is to take our understanding of genetic inheritance and 'extend that understanding in order to gain a clearer insight into mechanisms and health and disease' and to 'pinpoint variant spelling and discover how they increase or decrease an individual's risk of falling ill'.[44]

WHG's areas of research are: genetics, genomics and structural biology; human disease; structural biology; statistical and population genetics; translational genetics; genomics; and transgenics. Within their human disease research, WHG focuses on five main disease areas: immunity, inflammation and infectious disease; cancer genetics; Type 2 diabetes; endometriosis; and cardiovascular disease. Alongside these areas of research, WHG also has a number of scientific cores that support research at WHG and within the wider community at the University of Oxford: The Oxford Genomics Centre, Bioinformatics and Statistical Genetics, Chromosome Dynamics, Cellular Imaging, Transgenics, Biomedical Research Computing and IT. These scientific cores provide various research services to the research groups within the WHG and more widely, including commercial technology to conduct high-throughput genomic sequencing, data management and analysis, high-performance computing and microscopy services.

3.2.2. Operation and key features

Funding sources

The WHG is funded by the University of Oxford, Wellcome and other sponsors (Cancer Research UK and the British Heart Foundation, among others), and researchers at the WHG are supported by approximately £20m in funding annually. Wellcome core funding supports around 30 members of faculty with the WHG and makes up approximately 12% of all funding (supporting laboratory work, administrative workers and IT), and the University of Oxford and other funders support 70 members of faculty. However, most funding for the WHG comes from competitively won grants, which at times has been a challenge as sustainable long-term funding is needed to conduct research in genetics and genomics (Comp_03). All services provided by the scientific cores at the WHG are on a fee-for-service basis, with principle investigators (PIs) within the WHG paying a lower price than outside researchers and commercial entities. The income from fee-for-service work helps cover expenses associated with the core facilities and wider work at the WHG, as the WHG is a charity and therefore not focused on making a profit from the services it provides (Comp_04).

[44] As of 28 July 2020: https://www.well.ox.ac.uk/about-us/about-us

Staff

There are about 400 research staff at the WHG, along with 70 people employed in administrative and support roles. Staff members of the WHG are employed by the University of Oxford (Comp_03), although there are also a number of external fellows of Wellcome, Cancer Research UK and the British Heart Foundation conducting research at the WHG. There are 40 PIs who lead research groups within the different areas of research within the WHG. PIs apply to join the WHG and are chosen for their academic excellence and how well their research fits into the priorities of the WHG (Comp_03).

3.2.3. Outputs and achievements

Training and capacity building

As a department within the University of Oxford, the WHG has a strong role in teaching and training new and potential scientists. For example, the WHG has a work experience programme for year 12 students, undergraduate internships to gain research experience, and graduate study programmes, fellowships and studentships, including five doctoral studentships per year in genomic medicine and statistics (funded by Wellcome) and studentships funded by the Nuffield Department of Medicine, National Institutes of Health (NIH) and the Medical Research Council (MRC) in areas such as structural biology, neuroscience, infection, immunology and translational medicine, cardiovascular medicine and basic sciences for clinicians.

Translation and commercialisation

The WHG's strengths lie primarily in human disease research, genetic discovery and statistical genetics, and the Centre particularly excels in areas where genetics overlaps with structural biology. While it does not have the same large-scale high-throughput sequencing capabilities as Sanger and the Broad Institute and operates at a much smaller scale, WHG can provide deep expertise within more limited areas of genetics and genomics (Comp_04). For example, WHG has been a pioneer in the use of nanopore technology for long-read sequencing, which helps sequence parts of the genome that are not sequenced well using short read sequencing technology.[45] WHG's expertise in nanopore technology which has led to a spin-out company (Genomics plc) focused on hand-held nanopore technology (Bowden et al. 2019), precision medicine and developing software applications to uncover the relationships between genetic variation and human diseases[46]. WHG has also led to spin outs in drug discovery and development for inflammatory diseases (Surface Therapeutics in 2004) and in biotherapeutics (Riotech in 2003).

The WHG's most notable contributions have been in the area of genomic (or precision) medicine, which has considerable potential to impact directly how patients are treated within genomic clinics in England. For example, the WHG's WGS500 project was a proof of concept that whole-genome sequencing can illuminate causative genes and help diagnose genetic disorders found that diagnoses can be found for up to 60% of patients with unknown genetic disorders (Taylor et al. 2015). This project was important in the lead up to the 100,000 Genomes project and the foundation of GEL, where the NHS sequenced 100,000 genomes of patients in genetic clinics throughout England (Ellis 2015). Research at the WHG was also

[45] As of 28 July 2020: https://www.well.ox.ac.uk/ogc/nanopore-genomes/

[46] As of 28 July 2020: https://www.genomicsplc.com/

instrumental in improving neonatal screening within a maternity hospital in Oxford, particularly in the area of Type 1 diabetes. These contributions to genomic medicine are facilitated by the WHG's close relationships with hospitals and clinicians, allowing staff to access clinical samples and understand how research can be used to impact healthcare practice (Comp_03).

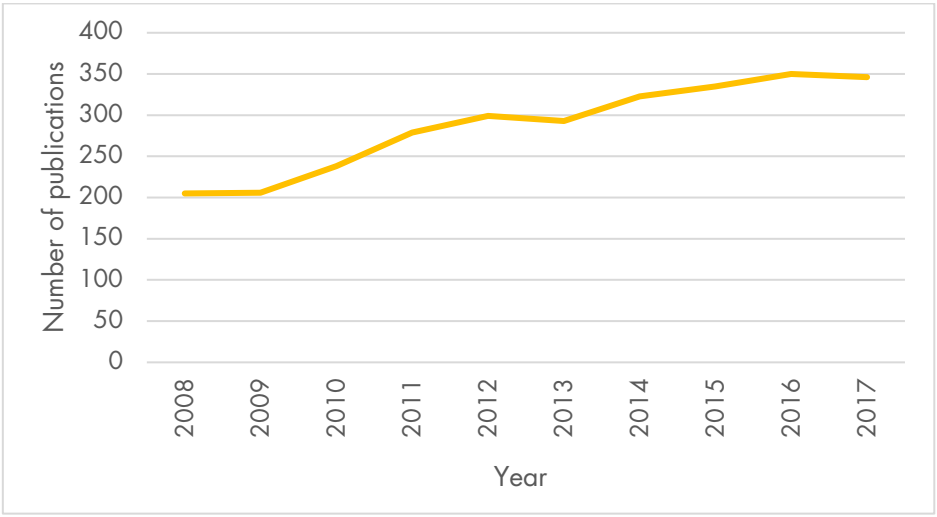
Leadership and collaboration

The WHG has contributed to several large international research collaborations, including the 1000 Genomes Project and the HapMap project, although it does not claim to have led these collaborations.

Publications

The WHG produced **2,874** publications between 2008 and 2017, spanning 103 WoS subject categories. Figure 11 shows the annual publications output of the Centre across this period, which rose from 205 in 2008, to **346** in 2017.

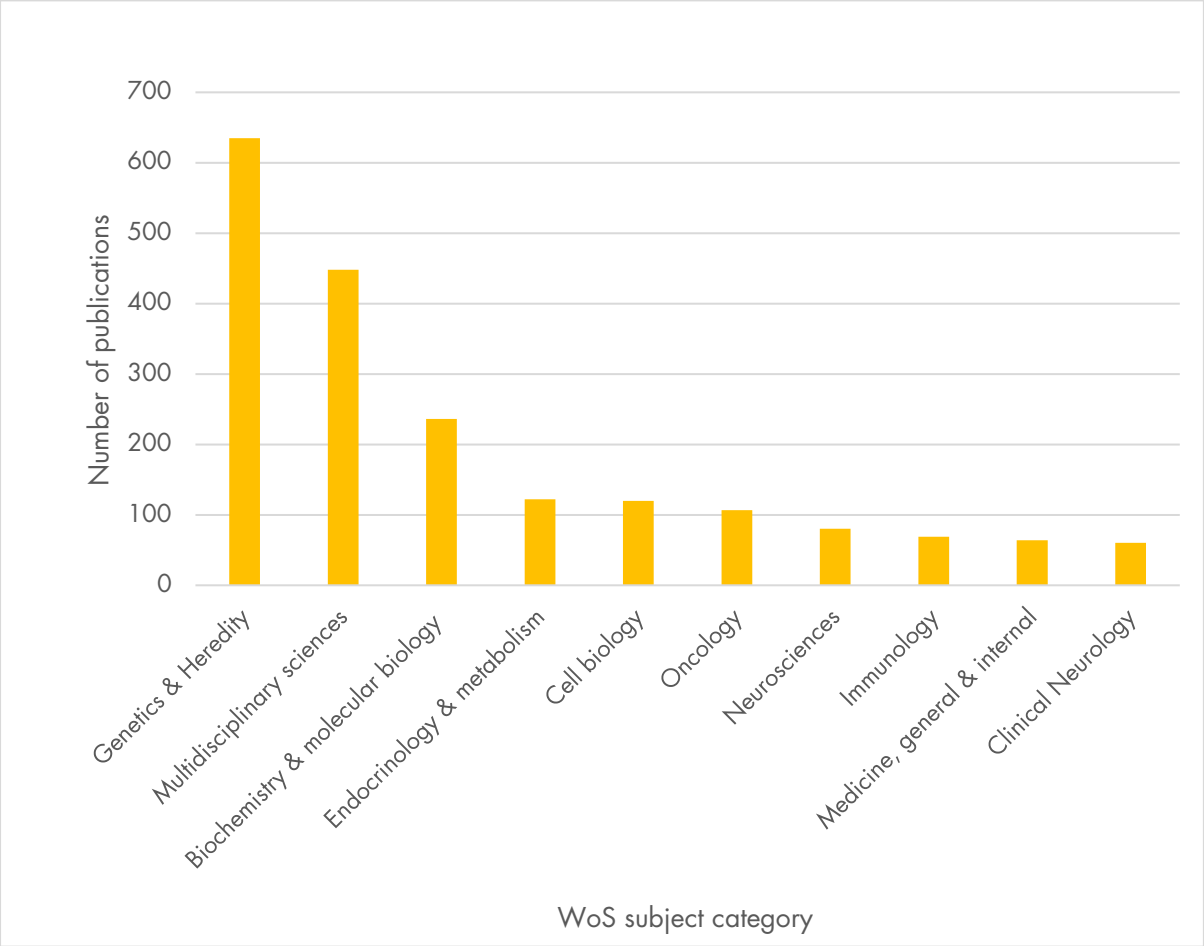
Figure 11: Annual publications output of the WHG, 2008–2017



Source: CWTS and RAND Europe

The subject categories in which the Wellcome Centre published the most papers between 2008 and 2017 were: **genetics and heredity; multidisciplinary sciences; biochemistry and molecular biology; endocrinology and metabolism; cell biology; oncology; neurosciences; immunology; medicine, general and internal; and clinical neurology**. Figure 12: The WoS subject categories in which the WHG published the most papers in 2008–2017.

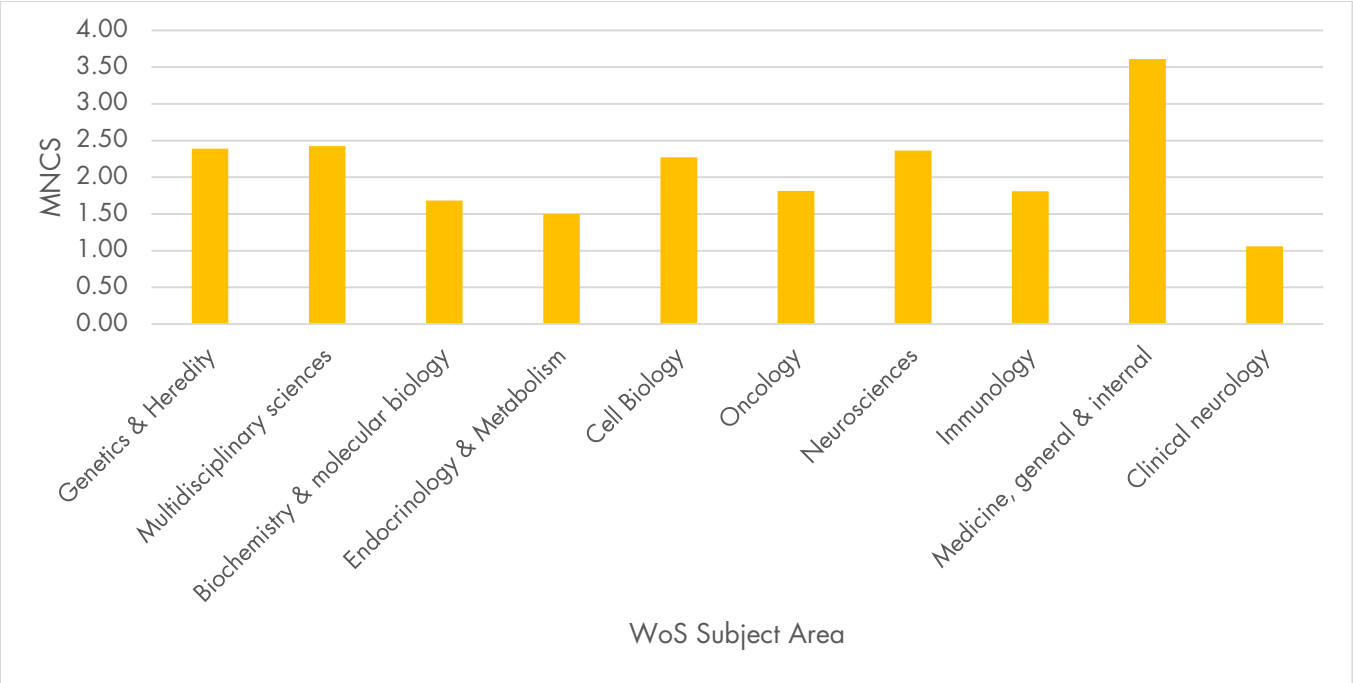
Figure 12: The WoS subject categories in which the WHG published the most papers in 2008–2017



Source: CWTS and RAND Europe

During this period the MNCS of WHG publications was **1.94** and the MNJS was **1.83**. **Nearly a quarter (24.6%)** of WHG publications fell within the top 10% most frequently cited publications in their field; **3.3%** of the WHG's publications fell within the top 1% most frequently cited publications in their field. Figure 13 presents the MNCS of WHG publications in its top ten subject categories by number of publications.

Figure 13: The MNCS of all WHG publications, 2008–2017, by the WHG's top ten WoS subject categories, by number of publications



Source: CWTS and RAND Europe

The institutions with which the WHG co-authors most frequently are as follows: University of Oxford; University of London; Sanger Institute; Harvard University and Oxford University Hospitals NHS Foundation Trust.[47] Figure 14 presents a network map visualising the 200 organisations with which the WHG has the strongest connections through its co-authorship links. Information on the key features of a co-authorship network map was provided in Section 2.2.1 above. The total link strength of the WHG within this network is **26,048**.

[47] These are the WHG's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. It should be noted that, in some cases, the listing of two institutions as authors may reflect a dual affiliation of a single authors (or authors). This may include dual affiliations between a home research institute and a parent organization.

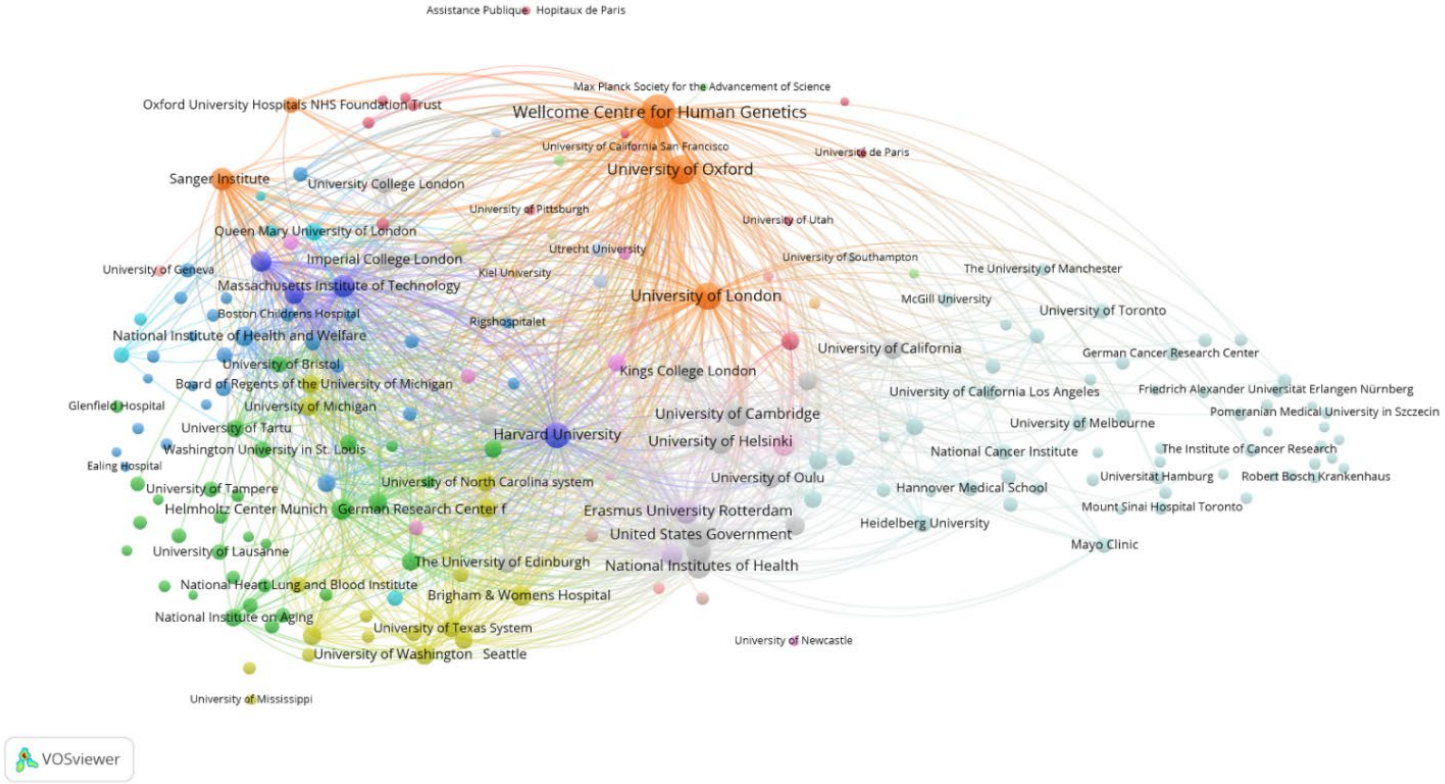


Figure 14: The WHG's co-authorship network map, 2008–2017

Source: CWTS and RAND Europe

3.2.4. Comparison with Sanger Institute

Table 4 provides some of the key similarities and differences between WHG and Sanger Institute. Overall, the WHG operates on a much smaller scale and with a much smaller budget than Sanger Institute, although it is a suitable comparison for some parts of Sanger's work, namely its work in human genetics.

Table 4: Comparison between the WHG and Sanger Institute

Key similarities	Key differences
• The WHG and Sanger Institute both work on large international collaborations, including HapMap and the 1000 Genomes Project, although Sanger tends to play a more leading role. • The WHG focuses on some of the same aspects of genetics and genomics as Sanger, including human diseases and genetic sequencing.	• The WHG is a much smaller scale than Sanger, and focuses more narrowly on human genetics. • The WHG is a part of the University of Oxford, while Sanger is generally not associated with academic institutions unless through formal strategic partnerships such as their partnership with the University of Cambridge. • Although both institutions receive Wellcome funding, WHG relies on competitively won grants and does not have the same level of core funding as Sanger • WHG offers services through their scientific cores on a fee-for-service basis. • WHG is more intensely focused on genomic medicine, and has access to clinical samples through a close relationship with hospitals in Oxford

3.3. Janelia Research Campus

3.3.1. Background

Janelia Research Campus was established by the Howard Hughes Medical Institute (HHMI) in 2006 in Virginia, US. Janelia was set up as the HHMI's first research campus, with the aim of bringing together researchers with different expertise and skillsets to work together in one space (HHMI 2003). The campus was designed by architect Rafael Vinoly to create a space that promoted collaboration and flexibility (HHMI 2003). This collaborative ethos is central to Janelia's work, which aims to enable researchers from different disciplines to work together to solve challenging research problems (Comp_06).[48] Janelia aims to address biomedical problems where progress requires technical innovation (Rubin and O'Shea 2019). An interviewee noted that Janelia focuses on high-risk projects, which require an interdisciplinary approach (Comp_06). This interviewee observed that the overall strategy of Janelia is to focus on advancing specific research areas, developing the tools and knowledge required to enable these areas to be advanced further by the wider scientific community whereby Janelia then targets another unexplored area (Comp_06). The

[48] As of 28 July 2020: https://www.janelia.org/our-research/overview

priority areas for research are decided through consultation, with researchers then given the freedom to pursue their research in these areas (Comp_06). An area of contribution from Janelia is its advancement of the field of imaging, such as microscope development, creating tools which are then shared with the wider research community (Comp_06).

3.3.2. Operation and key features

Funding sources

Janelia has an annual operating budget of US$130m (~£100m), funded by a core grant from HHMI,[49] and investigators do not pursue external funding.[50]

Staff

Within the campus there are approximately 600 employees,[49] with more than 350 scientists split between research labs, project teams and scientific support groups.[51] The scientific research conducted at Janelia is organised into research areas, and core research areas. Through the research areas Janelia aims to pursue a small number of scientific questions, with the potential for large impact. There are between one and three **research areas** at any one time, and they run for approximately 15 years. Currently there is one ongoing research area, mechanistic cognitive neuroscience, which explores how the brain enables flexible behaviour; a second research area will be launched later in 2020. The 15-year cycles enable researchers to approach more ambitious research questions, without the additional burden of having to re-apply for grant funding every four or five years. This is suggested as enabling researchers to focus on longer-term scientific goals rather than shorter term deliverables (Rubin and O'Shea 2019).

Alongside, the research areas, there are two ongoing **core research areas**: molecular tools and imaging, which focuses on novel reagents and tools to aid biological discovery; and computation and theory, which develops the quantitative skills to interpret data at scale. The core research areas do not follow the same 15-year cycles, but instead adapt over time, and consider the needs of researchers at Janelia, as well as those of the wider research community. In addition, there are multidisciplinary **project teams** who develop the tools required to tackle specific problems that would be difficult for individual researchers. Finally, there are **support teams** who partner with individual labs to provide made-to-order tools, software and equipment. There are approximately 50 individual research labs, which are purposely kept to a maximum of 6 researchers plus the group leader. Researchers at Janelia do not have tenure, with group leaders generally staying for a period of 10–12 years. This leads to a certain degree of turnover within the organisation (Comp_06), enabling the organisation as a whole to be flexible in the research areas it tackles.

[49] As of 28 July 2020: https://www.janelia.org/sites/default/files/About%20Us/janelia_flyer_FINAL.pdf

[50] As of 28 July 2020: https://hhmicdn.blob.core.windows.net/policies/Grants-Fellowships-and-Awards

[51] As of 28 July 2020: https://elifesciences.org/articles/44826#bib2

3.3.3. Outputs and achievements

Tools, data and resources

Janelia has been involved in the development of tools, data and resources which it makes freely accessible to the wider research community.[52] This includes the development of imaging instrumentation, laboratory tools, reagents, software and data. An area of contribution is the advancement of imaging tools and techniques (Ahrens et al. 2013; Chhetri et al. 2015; Liu et al. 2018; Tomer et al. 2012). For example, researchers at Janelia have made microscopes which can offer high resolution, even with very large samples (Perkel 2019). Other examples of developments include the use of genetically encoded imaging agents based on fluorescent proteins, which led to the GENIE project,[53] where imaging techniques are used to measure neuronal activity in the brain (Dana et al. 2016; Fosque et al. 2015; Sun et al. 2017). In addition to these technologies, Janelia offers use of its Advanced Imaging Centre, which gives scientists access to Janelia microscopes at no cost. Through the development of imaging techniques, Janelia has made several contributions to the field of neuroscience. One specific contribution has been through the mapping of all the neurons within the fly's brain, including the thousands of connections between them (known as the connectome) (Xu et al. 2020). This map can be used as a foundational tool on which to build greater understanding of the brain, and the data is now publicly available for other researchers to use.

Translation and commercialisation

Janelia has 29 tools which have patents associated or pending. One example is the patented technology Janelia Fluor® Dyes, which are fluorescent dyes that are bright, photostable, and cell permeable, making them useful for providing bright images for single molecule imaging experiments.[54]

Training and capacity building

Janelia is involved in training and capacity building, for example it hosts undergraduate summer schools, offers high school internships and has a joint PhD programme with the John Hopkins University where PhD students are trained at both institutions. In addition, Janelia has an extensive visitor programme, enabling visitors from a wide range of disciplines, including biology, chemistry, engineering and physics, to pursue their research interests alongside Janelia scientists.[55]

Public engagement

In addition to training and capacity building within the research community, Janelia hosts events for the wider community. It runs a public lecture series called Dialogues of Discovery, which invites thought leaders in science and related fields to give lectures to members of the public free of charge. Janelia also provides approximately U$1m a year to provide support for science education within the local community.

[52] As of 28 July 2020: https://www.janelia.org/open-science/tools-and-innovations

[53] As of 28 July 2020: https://www.janelia.org/project-team/genie

[54] As of 28 July 2020: https://www.janelia.org/open-science/janelia-fluor-dyes

[55] As of 28 July 2020: https://www.janelia.org/you-janelia/visiting-scientists

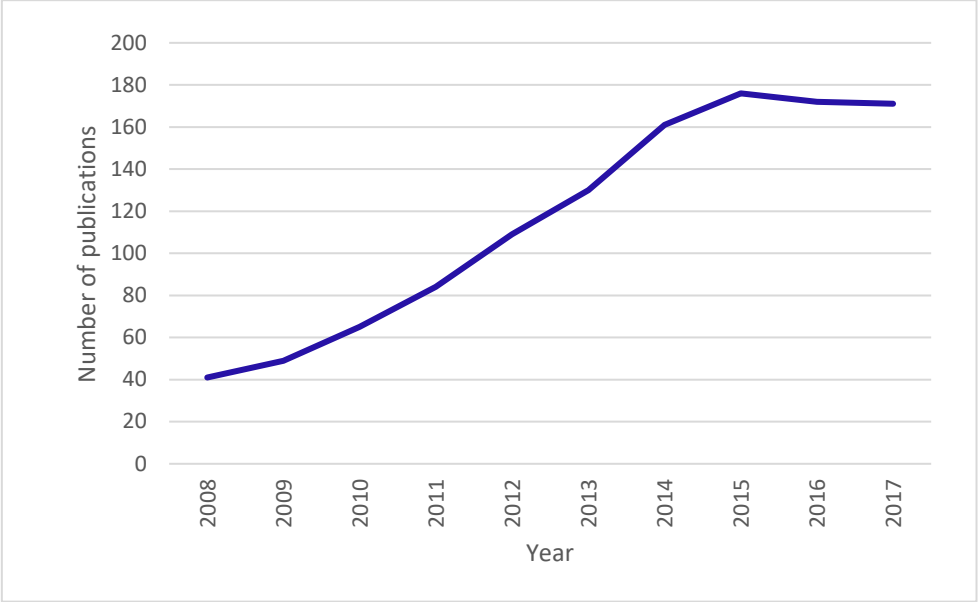
Leadership and collaboration

Janelia is part of the initiative Brain Research Through Advancing Innovative Neurotechnologies (BRAIN), which was launched in 2003. This is a collaborative programme to develop technologies to explore the human brain, looking into the connections and processes. This is a collaboration between research institutes and organisations, including federal and non-federal members.

Publications

Janelia Research Campus produced **1,158** publications between 2008 and 2017, the lowest figure of the organisations considered in this study. This may be due to the way affiliation and funding are acknowledged for this institution and thus represent a limited picture of its output. Within that limited set of publications, the diversity of fields covered by Janelia's publications is also narrower than for other organisations studied in this report, at 74 WoS subject categories. Figure 15 shows the annual publications output of Janelia across this period; it rose from 41 in 2008, to **171** in 2017 (with a high of 176 publications in 2015).

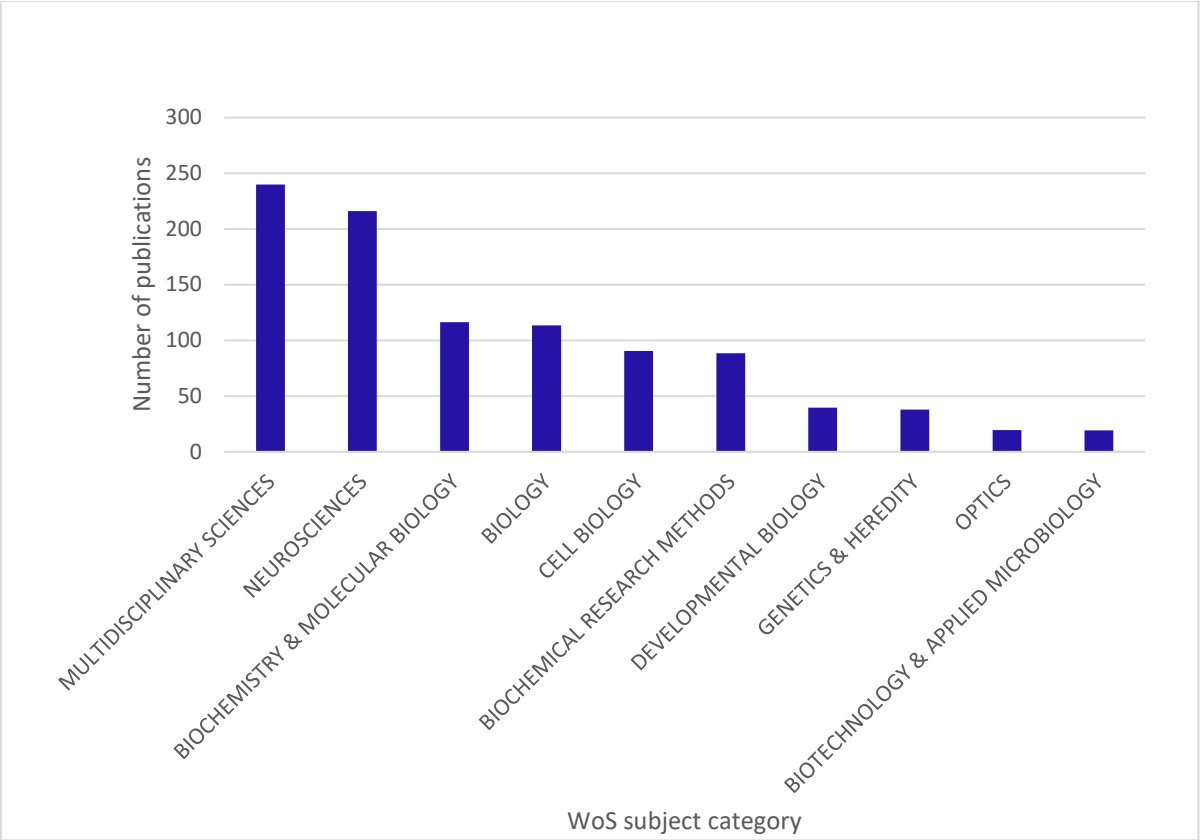
Figure 15: Annual publications output of Janelia Research Campus, 2008–2017



Source: CWTS and RAND Europe analysis

The ten subject categories in which Janelia Research Campus published most papers between 2008 and 2017 were: **multidisciplinary sciences, neurosciences, biochemistry and molecular biology, biology, cell biology, biochemical research methods, developmental biology, genetics and heredity, optics, and biotechnology and applied microbiology**. Figure 16 shows the number of Janelia publications across each of these fields.

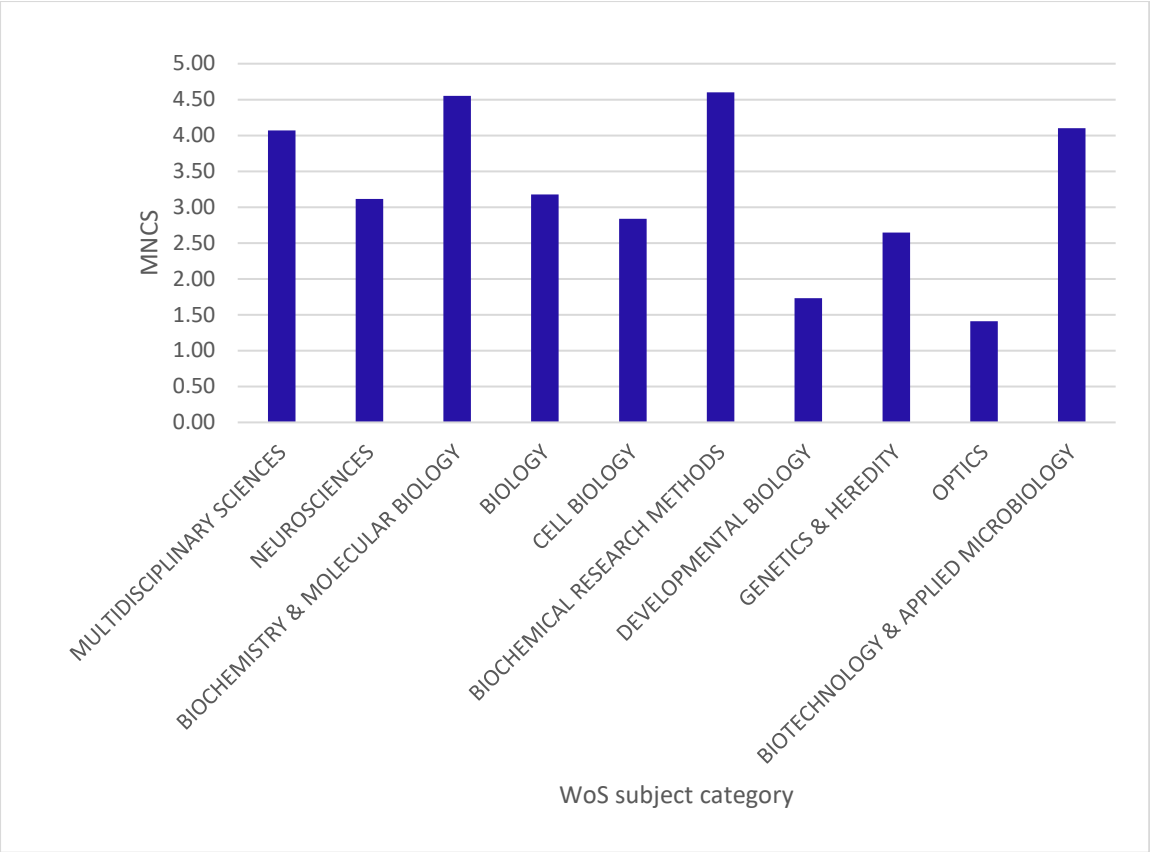
Figure 16: The WoS subject categories in which Janelia Research Campus published the most papers in 2008–2017



Source: CWTS and RAND Europe analysis

During this period the MNCS of Janelia publications was **3.34** and the MNJS was **2.62**. In total, **41.5%** of Janelia's publications fell within the top 10% most frequently cited publications in their field, while **9.2%** of its publications fell within the top 1% of most frequently cited publications. Figure 17 presents the MNCS of Janelia's publications across its top ten subject categories by number of publications.

Figure 17: The MNCS of all Janelia Research Campus publications, 2008–2017, by Janelia Research Campus' top ten WoS subject categories, by number of publications

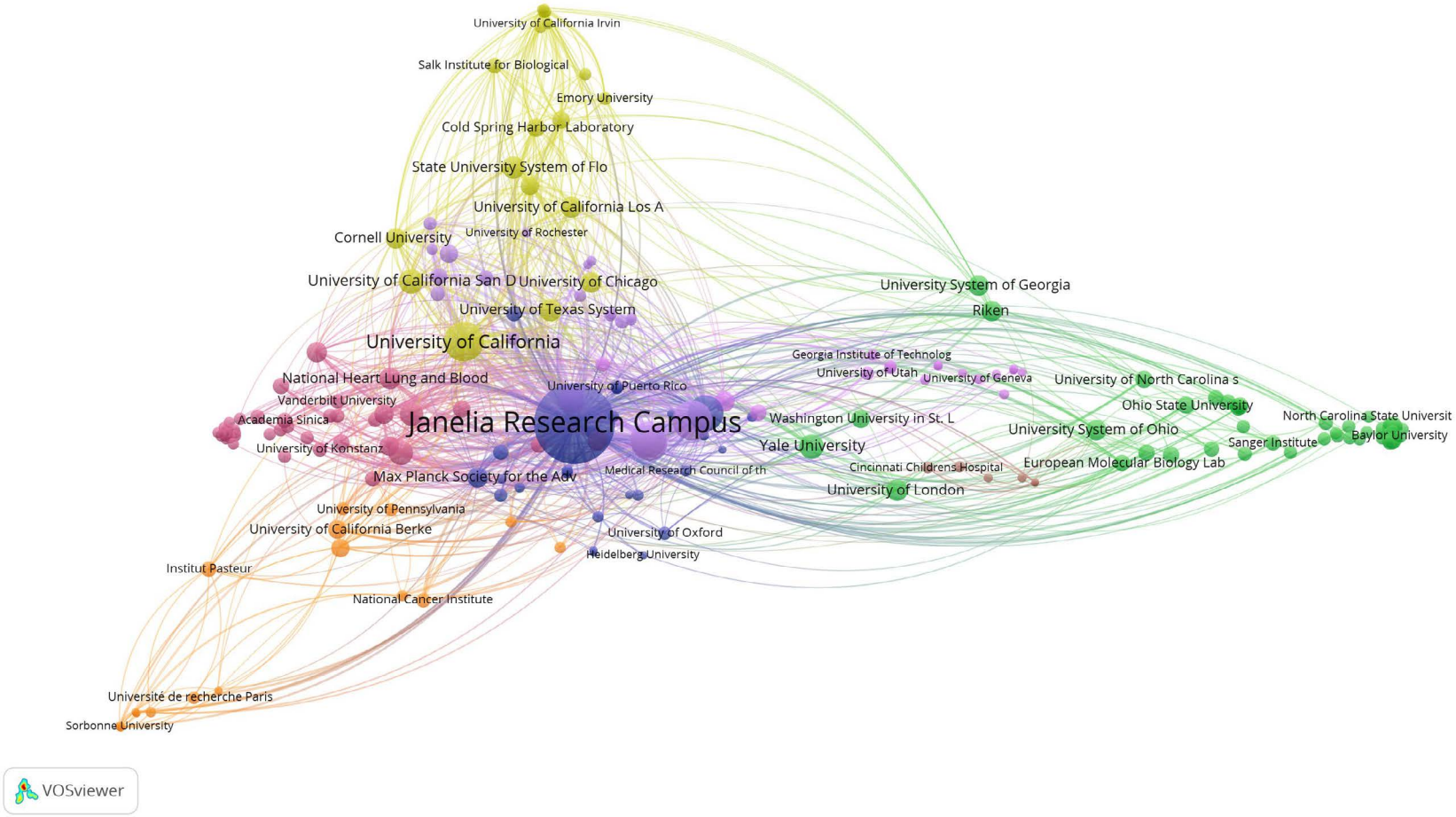


Source: CWTS and RAND Europe analysis

The institutions with which Janelia co-authors most frequently are as follows: University of California; United States Government; National Institutes of Health; United States Department of Health and Human Services; and Harvard University.[56] Figure 18 presents a network map visualising the 200 organisations with which Janelia has the strongest connections through its co-authorship links. Information on the key features of a co-authorship network map were provided in Section 2.2.1 above. The total link strength of Janelia Research Campus within this network is **2,437**.

[56] These are the Janelia's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. It should be noted that, in some cases, the listing of two institutions as co-authors may reflect a dual affiliation of a single authors (or authors). This may include dual affiliations between a home research institute and a parent organization.

Figure 18: Janelia Research Campus' co-authorship network map, 2008–2017



Source: CWTS and RAND Europe analysis

3.3.4. Comparison to Sanger Institute

Table 5 summarises some of the key similarities and differences between Janelia Research Campus and Sanger Institute and Genome Campus. Janelia was seen as a less appropriate comparator than other organisations included in this study because of its small scale and divergent research focus. However, the structure and ethos of Janelia was more closely aligned with that of Sanger than the other comparators.

Table 5: Comparison between Janelia Research Campus and Sanger Institute

Key similarities	Key differences
• There is a strong focus at both organisations on open access and creating tools for use in the wider academic community. • Both organisations have core funding, so researchers don't have to apply for grants continually. • There is a strong culture of fostering collaboration between labs at both organisations. • Both organisations have a strong strategic focus, and aim to foster a longer-term perspective. • Both Janelia and Sanger were set up by a charitable foundation (HHMI and Wellcome respectively). • Staff at Janelia Research Campus have no tenure.	• Janelia Research Campus operates on a smaller scale than Sanger • Janelia Research Campus is more oriented around the production of tools and software to aid research. • Janelia Research Campus has 15-year research cycles with few major research programmes at a time (currently one); it is less diverse than Sanger.

Source: RAND Europe analysis

Janelia and Sanger share several similarities in regard to their organisation and ethos but differ in the research questions they aim to answer (Comp_06). Like Sanger, Janelia has focused on bringing experts together to work on specific research challenges and offers core funding in order to achieve this. There is a top-down strategic focus at both organisations, and both encourage high levels of collaboration between researchers. In addition, both organisations heavily encourage data sharing. The differences between the organisations are that Janelia is far smaller than Sanger, and less diverse in the problems it attempts to tackle – Janelia has a singular vision and research focus. In addition, the approach taken by the organisations is different: Janelia primarily targets areas which are at a very early stage of development and aims to shift its focus after developing an area for a 15-year research cycle (Comp_06). Sanger focuses more on building up its knowledge base in key areas, and developing these areas further into projects which can offer translational benefits.

3.4. The NHGRI

3.4.1. Background

The NHGRI was established in 1989 to lead the US effort in the Human Genome Project. Since the completion of the Human Genome Project, NHGRI has continued to fund and conduct research with an aim of improving human health through advances in genomics research. NHGRI is part of the NIH, which is the primary medical research agency in the United States,[57] and is located on several NIH sites across Maryland, US. The main scientific focus of the NHGRI is to advance all scientific areas through the use of genomics (Comp_05). Unlike other NIH institutes, which focus on specific diseases, NHGRI aims to cut across these areas, providing cutting-edge genomics research which can be applied in various fields (Comp_05). In addition, it offers leadership in using genomics to advance particular areas of research. For example, in partnership with the National Cancer Institute, researchers were able to demonstrate how genomics approaches could be used to provide the tools and knowledge required to translate research projects (Comp_05). Like Sanger Institute, the NHGRI is recognised not only for conducting research on genetics and genomics, but also for its contributions to the ethical, legal and social issues around genetics and genomics (Comp_05).

3.4.2. Operation and key features

Funding

The NHGRI receives annual funding in the region of $500–600m from the US government, which is split between its responsibilities as a funder and a research institute. [58]

Staff

The NHGRI is organised into seven divisions, four of which are hosted within the Extramural Research Program, which supports the role of the NIH in its mission to advance genomics research.[59] These extramural divisions include the Division of Genomic Sciences, the Division of Genomic Medicine, the Division of Genomics and Society and the Division of Extramural Operations (which oversees grant applications and grants management). As well as these extramural divisions, there is also the Division of Intramural Research, which conducts the genomic research within the NHGRI's own laboratories (located primarily on the campus in Bethesda, Maryland).[60] Finally, the Division of Policy, Communications, and Education oversees activities such as policy development, outreach and media relations; and the Division of Management oversees the institute's financial and administrative services. Within the Division of Intramural Research, which is more akin to Sanger, there are 50 NHGRI investigators assigned to the 9 sub-divisions, which cover a range of areas relating to genomics research including cancer genetics,

[57] As of 28 July 2020: https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-human-genome-research-institute-nhgri

[58] As of 28 July 2020: https://www.genome.gov/about-nhgri/Budget-Financial-Information

[59] As of 28 July 2020: https://www.genome.gov/about-nhgri/Organizational-chart

[60] As of 28 July 2020: https://www.genome.gov/about-nhgri/Division-of-Intramural-Research

computational and statistical genomics, and genetic diseases, as well as other areas relating to genetics research.[60] The NHGRI is currently undergoing a strategic planning overview, and aims to 'identify, lead and support paradigm-shifting areas of genomics' (Comp_05). To do this staff will prioritise newly emerging areas of genomics which are not well defined, or lack significant investment from other organisations and institutions.[61] The full strategy will be launched later in 2020 (Comp_05).

3.4.3. Outputs and achievements

Data and databases

The NHGRI advocates data sharing and provides valuable resources to the wider research community (Comp_05). The NHGRI-EMBL-EBI GWAS catalogue provides a comprehensive resource of single nucleotide polymorphism (SNP)-trait associations, which are useful in genome-wide association studies (studies which aim to detect genetic variants associated with a trait) (Buniello et al. 2018). Previously, NHGRI was involved in the International HapMap Project, a publicly available map of the human genome, which enabled researchers to find genes associated with health and disease (R.A. Gibbs et al. 2003). More recently, NHGRI has been involved in the creation of a clinical genomic database, a manually curated database of conditions with known genetic causes for clinicians and researchers.[62]

Leadership and collaboration

A key feature of the NHGRI is thought to be its capacity to participate in team, consortium science (Comp_05). Beginning with its contribution in leading the Human Genome Project, it has continued to collaborate and lead on large-scale genomics projects and is thought to be highly regarded within the NIH ecosystem in this capacity (Comp_05). One example is TCGA, a multi-institution collaboration supported by the NHGRI and the National Cancer Institute. This was one of the first projects to characterise a host of different cancer types at the molecular level. Much of this work has now been published as the Pan-Cancer Atlas, a detailed genomic analysis of over 10,000 tumours representing 33 types of cancer.[63] One interviewee noted that a major role of the NHGRI is to share the knowledge in genomics, in order to help others advance the field (Comp-05).

Translation and commercialisation

From the beginning, NHGRI has played a large role in technology development, an example of which is the role it has played in the reduction of sequence costs (Comp_05). The NHGRI's Technology Transfer Office assists in the transfer of NHGRI-developed technologies to the private sector for further development.[64] This office is specifically evolved in the evaluation, and licensing of novel technologies developed by NHGRI investigators. There are several patented and patent-pending technologies including

[61] As of 28 July 2020: https://www.genome.gov/about-nhgri/strategic-plan/overview

[62] As of 28 July 2020: https://research.nhgri.nih.gov/CGD/

[63] As of 28 July 2020: https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html

[64] As of 28 July 2020: https://www.genome.gov/about-nhgri/Division-of-Intramural-Research/Scientific-Director-Office/Technology-Transfer-Office

technologies involved in targeted leukaemia therapy, micro-array technologies for rapid molecular profiling.[65]

Publications

The NHGRI produced **2,887** publications between 2008 and 2017, spanning 109 WoS subject categories. Figure 19 shows the annual publications output of the NHGRI across this period, which rose from 209 publications in 2008 to 336 in 2011. In 2017 the number was **317**.

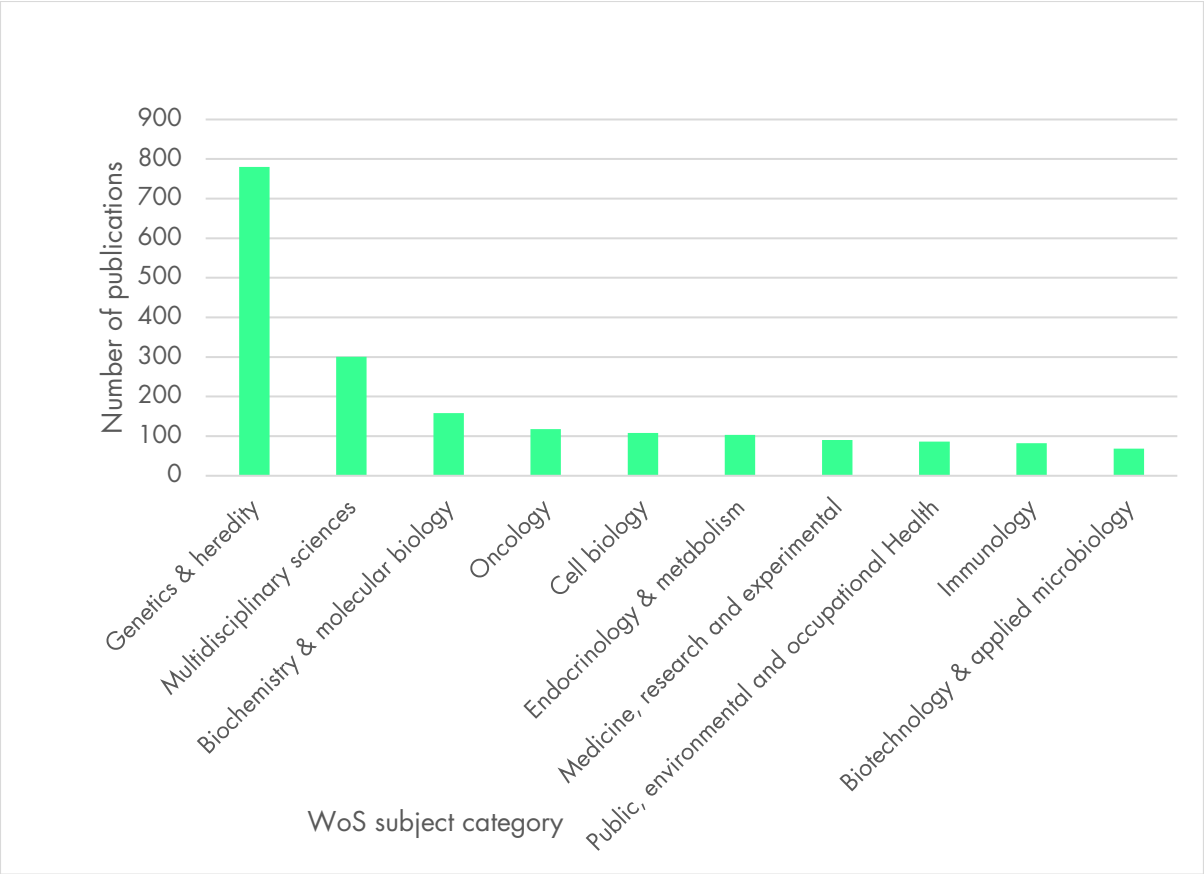
Figure 19: Annual publications output of the NHGRI, 2008–2017

Source: CWTS and RAND Europe analysis

The ten WoS subject categories in which the NHGRI published the most papers between 2008 and 2017 were: **genetics and heredity; multidisciplinary sciences; biochemistry and molecular biology; oncology; cell biology; endocrinology and metabolism; medicine, research & experimental; public, environmental and occupational health; immunology; biotechnology and applied microbiology**. Figure 20 shows the number of NHGRI publications for each of these fields.

[65] As of 28 July 2020: https://www.genome.gov/about-nhgri/Partner-with-NHGRI/technologies-available-for-licensing

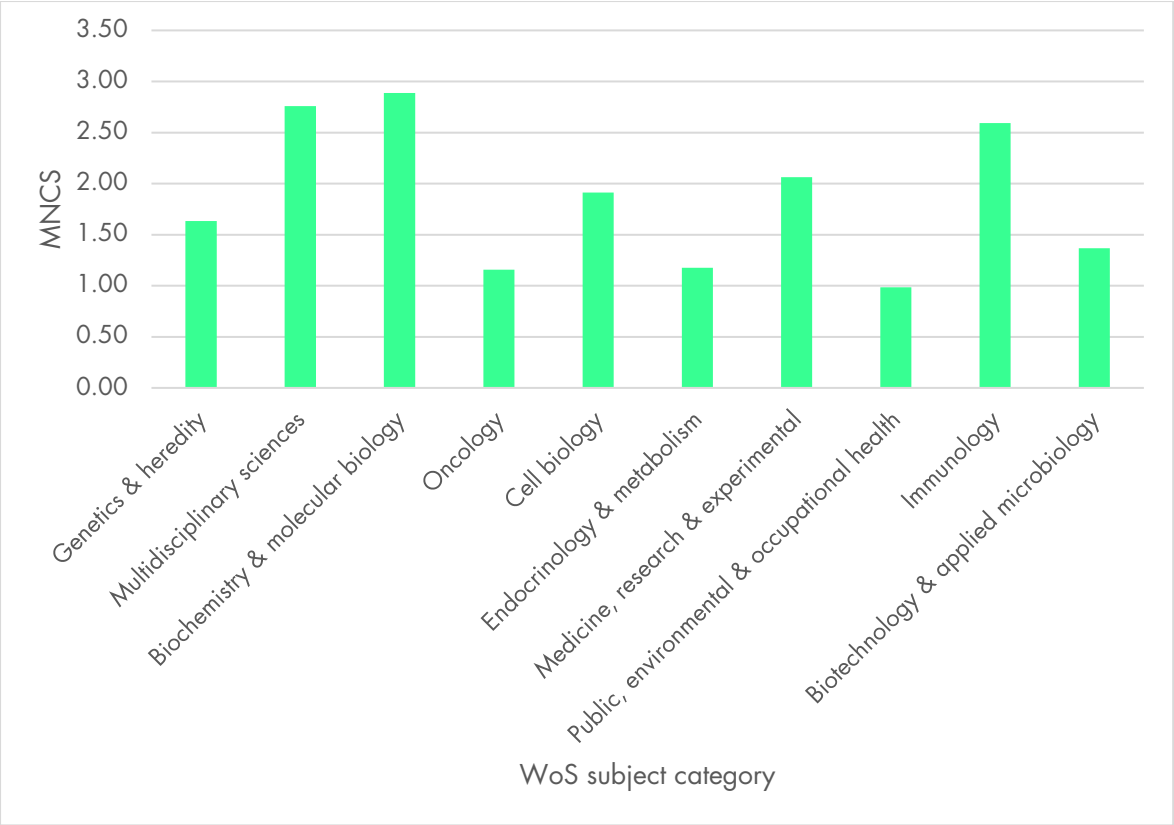
Figure 20: The WoS subject categories in which the NHGRI published the most papers in 2008–2017



Source: CWTS and RAND Europe analysis

During this period the MNCS of NHGRI publications was **1.71** and the MNJS was **1.60**. In total, **19.2%** of NHGRI's publications fell within the top 10% most frequently cited publications in their field, while **2.4%** of the NHGRI's publications fell within the top 1% of most frequently cited publications. Figure 21 presents the MNCS of NHGRI publications across its top ten subject categories by number of publications. The subject categories in which the MNCS is highest are biochemistry and molecular biology, multidisciplinary sciences and immunology.

Figure 21: The MNCS of all NHGRI publications, 2008–2017, by the NHGRI's top ten WoS subject categories, by number of publications

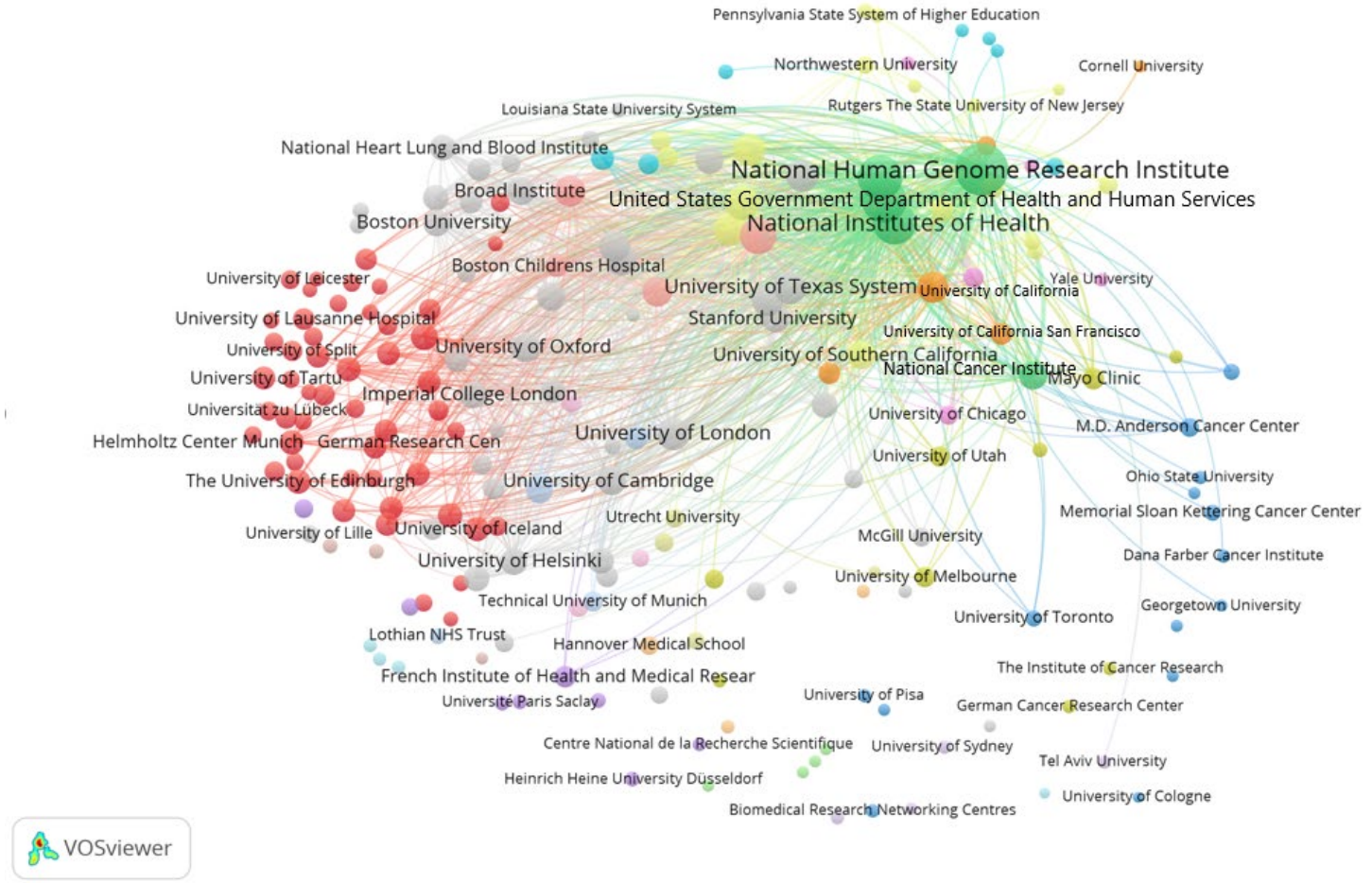


Source: CWTS and RAND Europe analysis

United States Government; United States Department of Health and Human Services; National Institutes of Health; National Cancer Institute; and Harvard University.[66] Figure 22 presents a network map visualising the 200 organisations with which the NHGRI has the strongest connections through its co-authorship links. Information on the key features of a co-authorship network map were provided in Section 2.2.1 above. The total link strength of the NHGRI within this network is **18,794**.

[66] These are NHGRI's top 5 collaborators as measured by the number of publications (P) on which both institutions are listed as authors. It should be noted that, in some cases, the listing of two institutions as co-authors may reflect a dual affiliation of a single authors (or authors). This may include dual affiliations between a home research institute and a parent organization.

Figure 22: NHGRI's co-authorship network map, 2008–2017



Source: CWTS and RAND Europe analysis

3.4.4. Comparison to Sanger Institute

Table 6 summarises some of the key similarities and differences between the NHGRI and Sanger Institute. The NHGRI was not seen as a good comparator organisation in this study as it is both a funder and a research institute. However, the common history between the NHGRI and Sanger Institute and the shared challenges they face as sequencing technology becomes more widely available, may provide useful insight to Sanger.

Table 6: Comparison between the NHGRI and Sanger Institute

Key similarities	Key differences
• Both organisations have a strong focus on genomics and genetics research to positively impact human health. • Both organisations have a strong focus on use of genomics technologies and genome sequencing. They are also interested in the biology relating to the immune system and infectious agents like viruses and pathogens. • Both have been involved in international collaborations (the Human Genome Project). • Both organisations focus on advancing data science, and the analytical methods needed to analyse sequencing data. • Like Sanger Institute, the NHGRI has the challenge that a greater number of institutes are able to sequence data, driving their 2020 vision and need for diversification.	• NHGRI is both a funder and a research institute. • NHGRI's Division of Intramural Research is of a much smaller scale than Sanger • NHGRI has a broader focus than Sanger

Source: RAND Europe analysis

The NHGRI and Sanger Institute share several similarities in their areas of research and capacity for international collaboration efforts but are very different as the NHGRI acts as a funder. Both NHGRI and Sanger have evolved from being primarily involved in the Human Genome Project, and both have had to diversify since this project was completed. NHGRI has strong links to researchers across the NIH, and therefore uses its expertise and funding to advance genomics across diverse research areas (Comp_05). Sanger appears to be more focused on building up its knowledge base in key areas, and developing these further into projects which can offer translational aspects.

3.5. Reflections

There are a range of similarities and differences between the different comparators identified here and Sanger Institute, but each offers a lens through which the operation, role and contributions of Sanger can be analysed. Table 7 provides a brief comparison of the structure, focus and organisation of each institution.

We note the shared historic origins of Sanger Institute and NHGRI, growing from the Human Genome Project, as well as the shared ongoing challenges in addressing a changing landscape where the historical advantages of high throughput sequencing no longer offer the same unique selling point they once did as sequencing technologies become much more widely available. NHGRI is still exploring its options for a new direction but is clearly looking to retain its status as being on the cutting edge of the field. As Sanger moves towards new basic research avenues, the Broad Institute, also facing this changing landscape, has shifted into a more commercially oriented model though still retaining its research strength underpinned by a close association with two internationally renowned academic institutions.

Table 7: Summary of the key characteristics of each institution

Institute	Year founded	Key partners	Staff	Annual funding and core funding	Scientific focus and key contributions	Research tools and technologies
Wellcome Sanger Institute	1992	University of Cambridge EMBL-EBI	Over 1100 employees and 63 post-graduate students	£152.4m – 64% from Wellcome (2019)	Large-scale genomic data production and analysis, cancer, human genetics and disease, parasites and microbes, genetic basis of diversity	Tools for annotation, gene finding, processing sequence data, visualisation, sequencing facilities
Broad Institute	2004	MIT, Harvard, hospitals	15 core Institute members, 51 non-core, >300 associate members	US$547.4m (c. £440m)– mixed income sources (2019)	Large-scale genomic data production, genome editing, drug discovery, cancer	Data analysis software for genomics and clinical data, cloud-based data storage
Wellcome Centre for Human Genetics	1994	Oxford University	Over 400 active researchers, 70 administrative staff	£20m annually of competitively won grants (2019)	Human disease research, gene sequencing using nanopore technology, genomic medicine	High throughput sequencing, computing and cellular imaging; graduate studies
Janelia Research Campus	2006	Established by HHMI	350 scientists	US$150m (c. £65m)	Mechanistic cognitive neuroscience, molecular tools and imaging, computation	Molecular tools, imaging technologies, tools for data science, software for analysis

National Human Genome Research Institute	1989	Part of NIH	50 NHGRI investigators within the division of intramural research	US$500–600m (c. £400–480m) from federal funding (2019)	Genomic technologies and data science, genetic disease, precision medicine, cancer, clinical research	Software and analysis tools for genomic data including study of complex traits, analysing sequencing data

Source: RAND Europe analysis

Comparison with the WHG provides a useful reflection on other models of funding in operation with Wellcome support, demonstrating a model much more reliant on leveraging wider funding, with Wellcome contributing only around 13% of total funding to the WHG, but providing a base against which wider funding, including institutional support and grant applications, can be built. This has not been without its challenges – with access to long-term support being noted as a potential issue – but has led to successful contributions to key international projects, though on a much smaller scale overall than for Sanger. Janelia has an interesting ethos and strategy, which is quite closely aligned to that of Sanger Institute – a particular focus on openness, collaboration (though more internally driven than in the case of Sanger), a campus setting and a long-term basic research but challenge-driven strategy. This is reflected in the publication outputs from Janelia, which are much lower in numbers, even accounting for institutional scale, but extremely highly cited, even compared with the high comparative baseline set by Sanger, suggesting a focus on quality over quantity.

Table 8 provides a brief summary of the key outputs and contributions of each institution. A particular area of contribution is in training and capacity building, with Sanger providing significant training and growth to the capacity of the field through its multiple formal and more informal partnership-based training offers. NHGRI emphasises training, in part reflecting its dual role as government agency and funder as well as research institution. Sanger offers research at the scale needed to take on leadership roles on large-scale international consortia, as do NHGRI and Broad Institute. The evidence suggests all three institutions have taken these leadership positions individually and collectively over time. A distinctive feature of Sanger is its propensity to take on convening roles for networks of smaller actors, too, using its scale to bring actors together in both the UK and LMICs. Examples of this are illustrated in Chapter 4 through the case studies. Another key contribution made by Sanger is the profusion of datasets and tools brought to the research field – though we also see strong contributions of this nature from the comparator organisations. Sanger's focus on commercialisation has not been as strong as in other organisations – notably the Broad Institute – but other routes to translation have been emphasised, particular the strong ethos of openness and data sharing. Sanger has strong disciplinary strengths in the core area of genetics and heredity, with comparable citation levels to key comparators such as the Broad Institute and NHGRI. Two particular areas where Sanger outperforms these two closest comparators (in terms of subject matter) on the level of citation are biochemistry and molecular biology, and infectious diseases.

Table 8: Summary of key outputs and contributions of each institution

Institute	Publications	Datasets and research tools	Training and capacity building	Translation and commercialisation	Public engagement and outreach	Leadership and advancement of field
Wellcome Sanger Institute	n=4,720, MNCS 2.59, % papers in HCP (10%): 29.4%; (1%): 5.6%	Tools for annotation, gene finding, processing, visualisation, sequencing facilities; datasets, e.g. Ensembl, COSMIC, DECIPHER	PhDs – 4-year, clinical, affiliated; masters programmes including MPhil genomic science targeted at LMICs; courses for scientists and health professionals; apprenticeships; range of courses and conferences through Connecting Science	Spin-out companies, e.g. Congenica, Microbiotica; BIC hosts 8 companies; licensing of COSMIC	Connecting Science public engagement team of 10 providing regular programme of monthly events, one-off events, online materials, training for scientists in engagement	Leadership of international research programmes; key role in Human Genome Project and subsequent 1,000 etc. genome projects
Broad Institute	n=5,591, MNCS 3.15, % papers in HCP (10%): 39.2%; (1%): 7.8%	Many datasets, software and tools to improve data analysis for large-scale genomic and clinical data, including a cloud-based analysis and data storage portal	Post-baccalaureate programmes for research experience, scientific writing and communications internships, associations for graduate and post-docs	Exclusive licensing of CRISPR-Cas9 in mammals to Broad Institute spin out; drug discovery and gene editing spin outs; 100s of patents in tech, therapeutics, engineering	Community activities for educators, educational tours, lecture series for the public, undergraduate research opportunities	Leading role in large international collaborations around Big Science, starting from the Human Genome Project
Wellcome Centre for Human Genetics	n=2,874, MNCS 1.94, % papers in HCP (10%): 24.6%; (1%): 3.3%	Research facilities for high throughput genomics and cellular imaging available to	Graduate studies funded each year through Centre	At least 3 start ups focused on therapeutics and drug discovery	Public events, including with a dance company to promote science, technology, engineering,	Leader in nanopore technology for genomic sequencing and in sequencing

		Oxford researchers on a fee-for-service basis			the arts and mathematics; undergraduate research internships	for genomic medicine
Janelia Research Campus	n=1,158, MNCS 3.34, % papers in HCP (10%): 41.5%; (1%): 9.2%	159 (including data, software, lab tools)	Undergraduate summer schools, graduate programme, high-school internships, computing programme; small, highly specialised conferences (~2/mth), workshops for junior scientists (~2/yr)	29 tools with patents issued or pending	Public lecture series 'Dialogues of Discovery'; provides ~US$1m annually to the local community school district to support science education	Main area of advancement for the field is in imaging techniques and tools, e.g. most detailed map of the fly brain to date
National Human Genome Research Institute	n=2,887, MNCS 1.71, % papers in HCP (10%): 19.2%; (1%): 2.4%	18	Graduate partnership programmes, medical residency, undergraduate programmes, summer internships, training in social and behavioural research	8 patented or pending techs; 9 research materials listed which are available for licensing	Community Engagement in Genomics Working Group; education and community involvement branch	Leader in the Human Genome Project, PCAWG, contributed to the HapMap, NHGRI-EMBL-EBI GWAS catalogue

Source: RAND Europe analysis

Figure 23 summarises the bibliometric performance of the five institutions using two citation-based metrics: the percentage of the institution's publications falling within the top 10% most highly cited publications in their field (PP top 10%) and MNCS. The citation performance is strong for all institutions, with all five having at least 19% of their publications in the top 10% by field and an MNCS above the world average of 1. The institution with the highest number of citation metrics is Janelia Research Campus, followed closely by the Broad Institute. Sanger is the next highest performing institution, followed by WHG and NHGRI, respectively. Figure 24 presents a combined co-authorship network map. The map visualises networks of co-authorship in the same way as the individual network maps featured earlier in this report. Here, however, the network map is based on co-authorship data for all organisations considered in the study, rather than focusing on one particular organisation. The map provides an opportunity to examine the strength and position of Sanger Institute within the broader co-authorship landscape within genetics and genomics research, and to compare this to the comparator organisations. Like the network maps shown above, the overarching network map has been capped to show connections between only the 200 most

strongly connected nodes within the network. Owing to its comparatively low number of co-authorship links, Janelia Research Campus does not feature within the network.

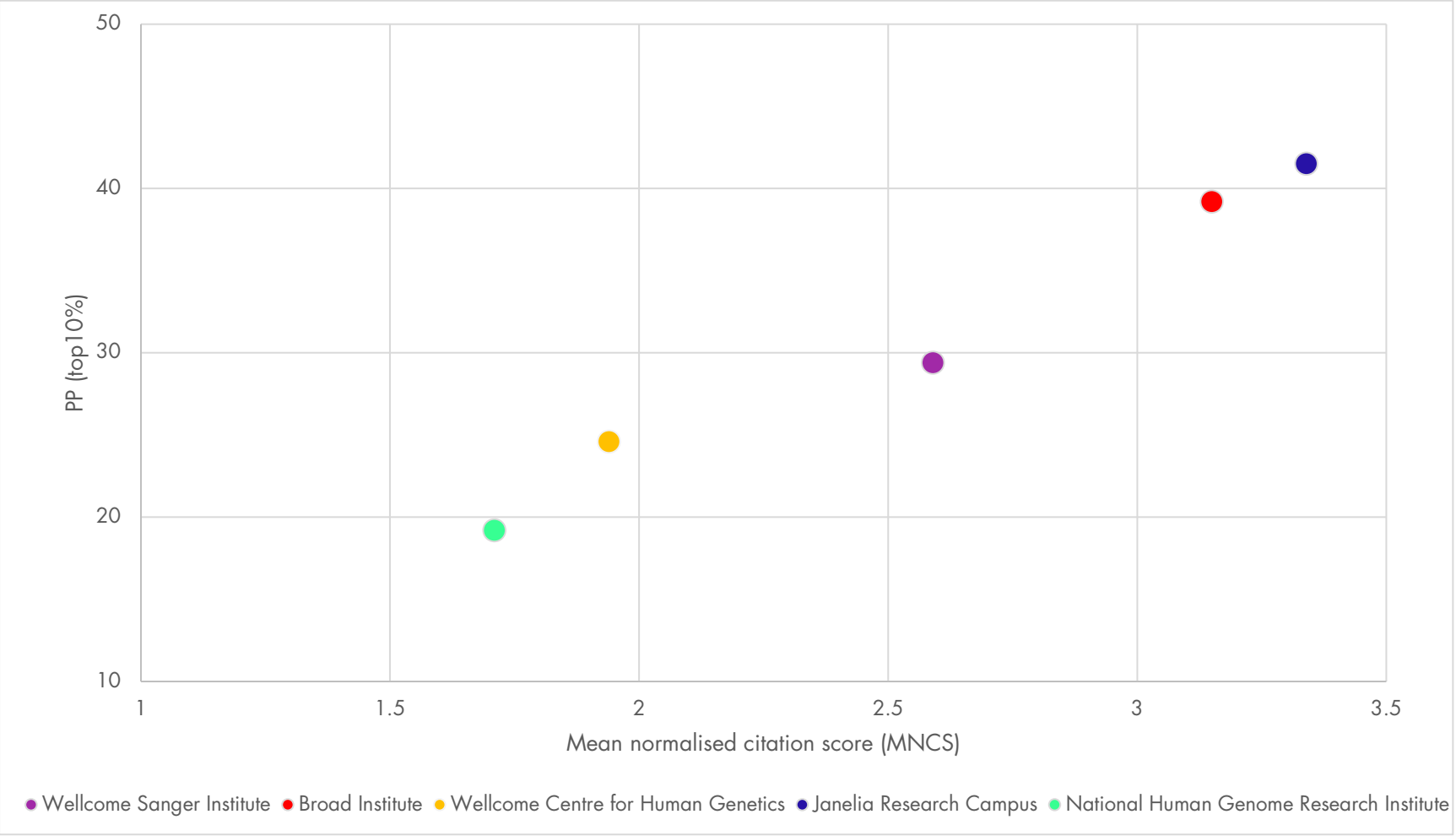
For the remaining four organisations– the Sanger Institute, the Broad Institute, the WHG and the NHGRI, a colour highlight has been applied reflecting the strength of each of the node nodes as total link strength within the network.[67] The closer to yellow in the colour spectrum, the greater the strength of the node. The organisation with the highest total link strength within the combined network is the Broad Institute (total link strength: 39,702). Sanger has the next highest (total link strength: 29,433), the WHG has the third highest (total link strength: 26,036) and the NHGRI the fourth (total link strength: 18,800).

Figure 25 shows the performance of the different institutions for the top ten research fields identified. We see here that Sanger makes significant contributions in the volume and quality of its publications across fields, with particular strengths in genetics and heredity, and biochemistry and molecular biology, as well as in microbiology, where Sanger produces a larger volume of publications than the other four institutions combined.

Reflecting overall on Sanger's role in the field it is evident that it acts as a leader in the progress and development of research, alongside other key actors such as the Broad Institute. But the flavour and direction of this leadership is shaped by its specific values and ethos, around openness, data sharing and collaboration. This provides a unique and novel offer, which has proved useful in the development of the field and which translates to effective delivery in slightly different contexts as shown by the work of Janelia Research Campus. The next challenge will be to find the best direction in which to take that ethos to shape the development of the field over the coming years.

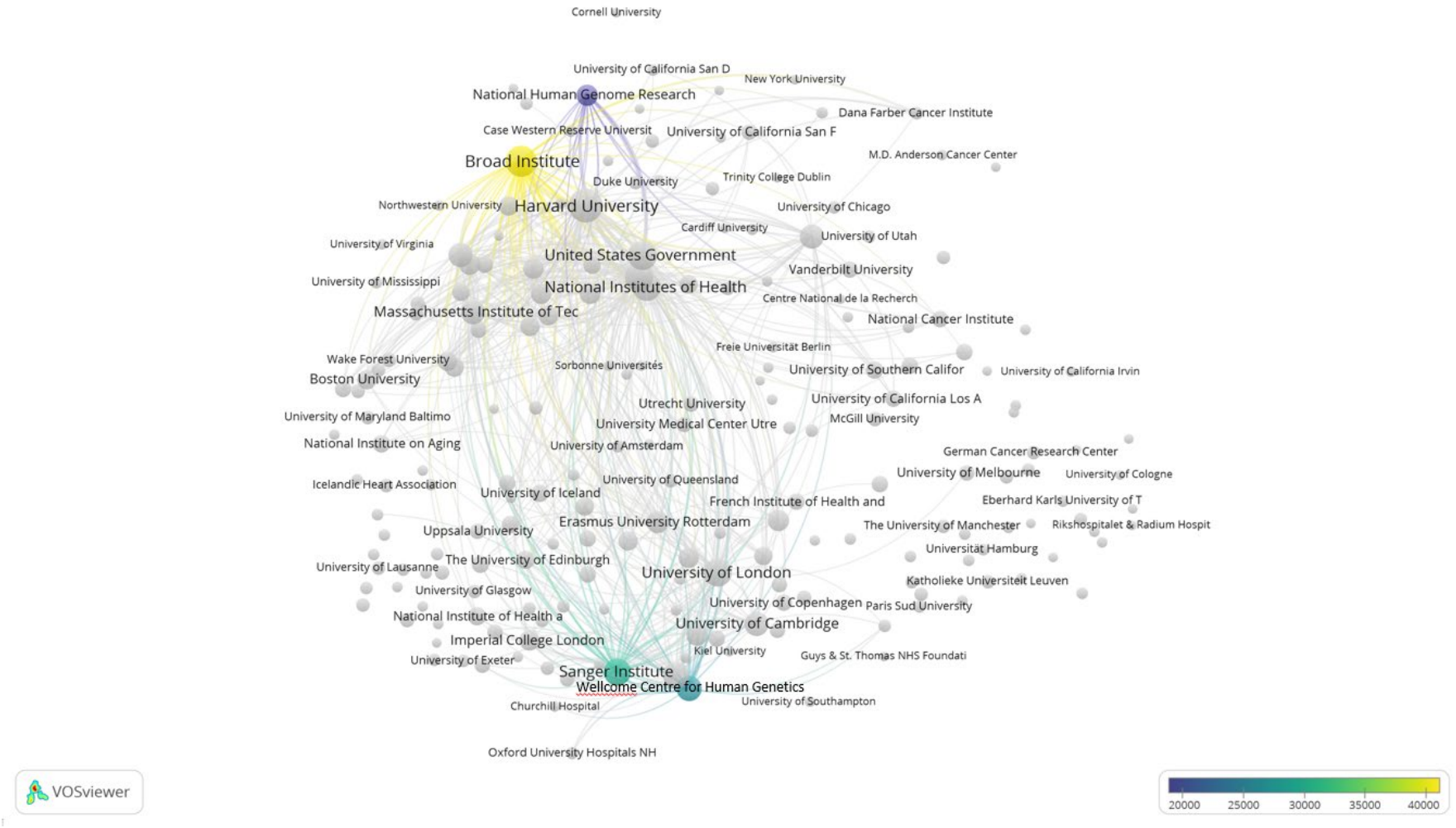
[67] The total link strength – a measure of the number of co-authorship links of a node within the network – is derived from the sum of the 'weights' of all 'edges' that connect to a node. The weights of an edge between two nodes represent the total number of publications on which the two organisations are both listed. A single publication that involves multiple partners will impart a weight of 1 on the edges between all involved partners.

Figure 23: The performance of institutions by top 10% highly cited publications and MNCS, against world averages of 10% and 1, respectively



Source: CWTS and RAND Europe analysis

Figure 24: Combined co-authorship network map of Sanger, the Broad Institute, the WHG and the NHGRI with other organisations



Source: CWTS and RAND Europe analysis

Note: As it has few co-authorship links, Janelia Research Campus does not feature within the network

Figure 25 The performance of comparator institutions, by discipline

	WSI		Broad		WCHG		Janelia		NHGRI	
	P	MNCS	P	MNCS	P	MNCS	P	MNCS	P	MNCS
Genetics & heredity	1028	2.55	1025	3.29	635	2.39	38	2.65	780	1.63
Multidisciplinary sciences	776	3.36	984	5.11	448	2.42	240	4.07	301	2.76
Biochemistry & molecular biology	455	4.97	500	3.23	236	1.68	116	4.55	158	2.89
Cell biology	234	2.07	414	3.49	120	2.27	6	1.56	108	1.91
Oncology	110	1.96	265	2.72	107	1.82	91	2.84	117	1.16
Biotehcnology & applied microbiology	210	2.22	194	4.02	58	1.46	19	4.10	69	1.37
Endocrinology & metabolism	82	1.48	151	1.93	122	1.50	89	4.60	103	1.18
Neurosciences	91	1.49	184	2.45	81	2.36	4	2.97	46	1.86
Microbiology	255	2.00	110	2.02	42	2.81	0	0.00	16	2.81
Biochemical research methods	143	2.25	181	3.10	49	1.50	0	0.00	28	0.94

	WSI
	P
GENETICS & HEREDITY	1028
MULTIDISCIPLINARY SCIENCES	776
BIOCHEMISTRY & MOLECULAR BIOLOGY	455
CELL BIOLOGY	234
ONCOLOGY	110
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	210
ENDOCRINOLOGY & METABOLISM	82
NEUROSCIENCES	91
MICROBIOLOGY	255
BIOCHEMICAL RESEARCH METHODS	143

Source: CWTS and RAND Europe analysis

4. Case studies

This chapter presents a set of case studies that illustrate the range and nature of the contributions made by the Sanger Institute. Each is selected to highlight different aspects of the work undertaken and the way in which it operates to support learning, and they are summarised in Table 9

Table 9: Summary of case studies

Case	Summary	Rationale for case study
Open Targets	Ongoing public–private partnership (since 2014) for conducting pre-competitive research in drug discovery by identifying and prioritising drug targets. This is an innovative way to encourage private-sector investment and collaboration, while still maintaining Sanger's ideals around openness.	To offer insights into Sangers commercialisation strategy
Tree of Life programme	Ongoing project (since 2019) to sequence 60,000 eukaryotic species (plants, animal, fungi) in the UK. Will compare genomic sequences to provide insights into evolution and conservation. There is collaboration across universities, funders, museums and horticultural organisations.	To explore Sanger's future direction and related decision-making processes
DDD	In collaboration with the NHS regional genetic services, Sanger produced and analysed novel genomic data, leading to diagnoses, spin-out companies and technologies, and improvements in clinical practice.	To explore an example of clinical application and impact
Malaria research and MalariaGEN	Founded in 2005, MalariaGEN is a scientific network that connects researchers and clinicians working to understand how genetic variation of humans, the malaria parasite and mosquitos affect the spread of malaria, and to use this knowledge to develop effective ways to control the disease. MalariaGEN has had a large impact on LMICs through large-scale collaborative and community projects, training, the development of a framework for equitable data sharing, and building local sequence capacity in order to establish surveillance mechanisms.	To provide insights into Sanger Institute's capacity building work and international collaboration
ICGC	Established in 2007, the ICGC comprises a global network of scientists, scientific groups and research funders and a centralised scientific organisation that serves to coordinate research projects. A primary goal of the ICGC has been to bring together all data on the genomic characteristics of cancers and to make this data freely available and accessible to the global research community.	To analyse the role Sanger Institute plays in large-scale, high-profile, internationally collaborative efforts

4.1. Open Targets

4.1.1. Introduction

Open Targets is an ongoing public–private partnership that brings together pharmaceutical companies, biotech companies and research institutions to conduct pre-competitive drug discovery research. The partnership focuses on the systematic identification of small molecules that are most suited to a disease's biological targets, and prioritises different drug targets that can potentially be used in drug development. This is done by harnessing the power of big data generated through genome sequencing, GWAS and other sources. The goal of Open Targets is to reduce the time and cost of drug discovery through collaborative research around target identification, after which pharmaceutical companies will engage in competitive research and development.

Open Targets was founded in 2014 by Sanger, EMBL-EBI and GlaxoSmithKline (GSK) (Koscielny et al. 2017). The partnership now includes Bristol Myers Squibb, Sanofi, Takeda and Celgene along with the founding partners, and Biogen was also a partner but has since left the partnership. Open Targets was originally called the Centre for Therapeutic Target Validation, but re-branded in 2016 to better reflect the focus of the partnership.

4.1.2. Background and context

Open Targets was founded to address a key issue in drug development: nearly 90% of all compounds that make it to the clinical trial stage of drug development fail to become licensed medicines owing to poor efficacy or patient safety concerns. This failure, which contributes to the high cost and long timelines associated with the drug development process, is often caused by a poor understanding of the biological target that the compound acts on (Bergauer et al. 2016). Traditionally, drug targets have been chosen for drug development on the basis of experimental evidence (Paananen and Fortino 2019). However, high-throughput genomic technologies have made it possible to generate large amounts of data on potential drug targets, including through GWAS, which presents opportunities to base choices of drug targets on genomic evidence and big data (Koscielny et al. 2017). Evidence has shown that using genetic data to select targets increases success rates in drug development, with drugs with genetic support for the drug mechanism having an estimated success rate twice of those without genetic support (Nelson et al. 2015).

Open Targets helps address a particular issue in genetics and genomics: some genes are much more studied than others, causing a 'streetlight effect' in which certain genes are studied at the expense of neglected genes because the evidence on them is more easily accessible. By bringing together data on all genes in a single platform, Open Targets helps make evidence on previously neglected genes more accessible (Dunham 2018).

Open Targets was founded in the context of there being an increasing number of public–private partnerships and pre-competitive research collaborations in the area of biomedicine and pharmaceuticals, with the average number of these horizontal, multi-party partnerships growing rapidly in comparison with bi-lateral agreements between single companies and research institutions (Vrueh and Crommelin 2017). These types of partnerships tend to focus on challenges that are too large for any one organisation to address effectively (Denee et al. 2012). They also reflect a shift in the pharmaceutical landscape from accumulating

proprietary data from experimental evidence, to taking advantage of the wealth of data available from a wide variety of sources and carefully using it to develop products for market (Bergauer et al. 2016).

4.1.3. Research process

Open Targets is located on the Wellcome Genome Campus, with offices in EMBL-EBI and laboratories in the Sanger Institute. Each private industry partner contributes to a common fund for Open Targets, and provides in-kind contributions by designating staff who work within the Open Targets partnership. Sanger and EMBL-EBI also provide in-kind contributions, such as dedicated staff, facilities, access to platforms and databases and research costs. Private partners gain value through Open Targets by getting early access to data before it is made public, exerting influence over the priorities of Open Targets, and gaining support in customising portals and incorporating proprietary data into the database (CS_03, CS_04). Sanger benefits from Open Targets as it gains support in translation and commercialisation, while retaining its commitment to open science values.

Open Targets has more than 90 staff members, who are funded by Open Targets within each member organisation. The Executive Leadership Team includes senior staff members from Sanger, EMBL-EBI and GSK who are responsible for providing strategic leadership and engaging new partners. There is also a Scientific Leadership Team with representatives from each partner organisation who are responsible for setting priorities within the partnership and approving research proposals, along with a Strategy and Operations Team and a Scientific Advisory Board. Although the intention of Open Targets is to focus on pre-competitive research, there is a protocol that will be followed if IP arises from Open Targets research that would allow for IP to be claimed by partners or members who are solely responsible for the contribution, but for the IP to be shared between partners for the purpose of Open Targets-related work (Priego and Wareham 2018).

There are two arms of Open Targets. The experimental programme generates new information about key therapy areas (oncology, immunity and inflammation, neurodegeneration) and data to support the causal links between targets and diseases. There is also a core informatics and data generation arm of Open Targets, which produces and maintains two important open resources in the area of target identification:

- The Open Targets Platform integrates public domain data from a wide variety of sources that associate targets and diseases to enable the identification and prioritisation of drug targets for further investigation and drug development. The platform allows users to search for targets or for diseases within the database, and produces a single score to help users understand the strength of the evidence around the association, while also allowing them to connect to the underlying evidence to investigate further.
- The Open Targets Genetics Portal integrates functional and biological data from a variety of sources, including GWAS, and uses statistical mapping to score the data and identify functionally important genes. This data can then be used to investigate pharmaceutical compounds and prioritise drug targets further.

Along with these resources, Open Targets also makes other open source bioinformatics tools available to aid in identifying and prioritising drug targets, including the Project Score portal that allows researchers to explore the results of CRISPR-Cas9 whole-genome drop out screens and the eQuantitative Trait Locus

Catalogue, which provides gene expressions and splicing quantitative trait loci from public studies on humans.

Some of the challenges that Open Targets have faced centre around bringing together academic research institutions and private companies, two sectors with separate cultures, norms and ways of working (CS_05). To overcome this challenge, the Executive Leadership Team encourages regular meetings between project teams and between members of the academic community and private industry, and organises an Open Targets integration day every four months to ensure that all partners come together to share learning, understand each partner's motivations and goals, and network. At times, there are challenges around priority setting, as the topics that academic researchers find most compelling do not always align with those that industry researchers understand as the topics with the most potential for translation and commercialisation. Where there have been disagreements, the Executive Leadership Team has sought a balance that suits both intellectual curiosity and opportunities for translation and commercialisation (CS_03, CS_04). In these ways, Open Targets helps provide an avenue for collaboration between academia and industry, two sectors that at times clash in their cultures, goals and ways of working.

An important facilitator of Open Targets has been the expertise that each organisation brings to the partnership, including in gene editing and CRISPR technology, induced pluripotent stem cells, single cell genomics, organoid and tissue culture, large-scale genomics and epigenomics, GWAS, next generation sequencing, bioinformatics and high-performance computing. Open Targets has benefited from areas where its focus overlaps with the deep subject-area expertise of partners, notably Sanger's expertise in cancer and EMBL-EBI's expertise in informatics (Priego and Wareham 2018), and the involvement of all partners in the management and organisational structure of Open Targets. Its physical location on the Genome Campus, which allows academic and commercial partners to collaborate and share ideas has been noted as a key facilitator of Open Targets (CS_03, CS_04).

4.1.4. Contributions to knowledge

As Open Targets is not an exclusive data generation programme, but rather a pre-competitive collaboration with a commitment to open data, all knowledge and data produced from the Open Targets programme have become a resource to those working in genetics and drug development. Much of the knowledge produced in the Open Targets experimental programme is within the focus therapeutic areas: oncology, immunity and inflammation and neurodegeneration. However, the tools and databases cover the whole genome, so present the opportunity to create knowledge in any disease or therapy area with a genetic basis. Notable outputs demonstrating the type of knowledge that has been produced through Open Targets include resources to identify and prioritise drug targets for cancer (Behan et al. 2019; Nalley 2019), schizophrenia (Gaspar and Breen 2017), Type 2 diabetes and Alzheimer's disease (Failli, Paananen, and Fortino 2019) and other diseases such as arthritis, bipolar disorder and multiple sclerosis (Picart-Armada et al. 2019), as well as studies that have contributed to the understanding of T-cell states and responses to immune diseases and inflammation (Cano-Gamez et al. 2020; Soskic et al. 2019).

4.1.5. Contributions to future research

Knowledge exchange between public and private partners is an important part of the work of Open Targets. For example, by working together through Open Targets, private industry partners learn from Sanger and

EMBL-EBI's deep expertise in genetics, genomics and bioinformatics, while academic researchers learn about translational research, commercialisation and private sector R&D from their industry colleagues. This knowledge exchange helps improve research efforts in both sectors, and gives both groups a view of how GWAS and other research moves towards target identification, drug development and other commercial activities (CS_03, CS_04). Along with collaborative working, knowledge exchange has also occurred through secondments of employees between public sector and private-sector partners (CS_03, CS_05).

Open Targets has produced important tools and databases for researchers in the field of genetics and genomics, and particularly those working in the area of target identification and drug discovery, and has led to additional research in the public and private sector around the targets identified through Open Targets tools. As it has been designed to be used by those without deep knowledge in bioinformatics, Open Targets allows more researchers to incorporate genomic evidence in their research (CS_04, CS_05).

Open Targets has worked in collaboration with researchers in genetics and genomics to harmonise data, creating a standard set of fields and format, as well as a proposed quality control process (Buniello et al. 2018). As Open Targets relies on public data with descriptive metadata to help with data linkages, the project has also supported work that promotes open science in Europe and the principle that data should be findable, accessible, interoperable and reusable (FAIR) (Halim 2018). This work not only contributes to the work of Open Targets, but also to wider research in the field of genetics, genomics and drug discovery.

4.1.6. Contributions to policy and product development

It is difficult to track and measure all of the potential private sector R&D activities and product development that has occurred as a result of Open Targets, although the Open Targets team attempts to engage private-sector users to track impact where possible (CS_03). As there are long timeframes in drug development, Open Targets will likely not yet have resulted in a new drug coming to market. However, companies are already using Open Targets data and resources to commit to drug targets during the drug development process (CS_04). It was reported that at least 6% of all disease-target pairs uncovered through Open Targets have resulted in a drug being developed, which is a conservative estimate based on what has been published (Priego and Wareham 2018). Open Targets data can also be used for drug repositioning, which given the high cost of novel drug development compared with drug repositioning is of increasing importance in the pharmaceutical industry. Open Targets has uncovered at least 2,540 potential new indications for 791 existing drug targets, and this number has likely increased since being reported (Khaladkar et al. 2017).

Open Targets has been used to inform decision making in private industry and to avoid costs. For example, one private-sector partner was presented an opportunity to invest in a potential line of research but was able to determine that the research was not replicable through Open Targets data, helping to avoid a wasted investment in this line of research (CS_03). Although all companies can use Open Targets data, partner organisations have early access to data, and are more likely to get a head start on the competition when this early access data is combined with proprietary data sources.

The European Commission conducted a case study of Open Targets to investigate the framework that it has created around open data sharing and user-driven platforms. It reported that Open Targets has promoted 'smart openness', in which openness helps to spur innovation, which is a counterbalance to either

complete openness or exclusive licensing of IP. The Commission also reported that Open Targets had encouraged public–private sector collaboration, which increased the impact of scientific research and data re-use (Priego and Wareham 2018).

4.1.7. Contributions to health and the health system

Although Open Targets has not yet resulted in a new drug coming to market impacting human health, there is potential for Open Targets to improve success rates in drug discovery by integrating genetics and genomics data into target selection and prioritisation (Koscielny et al. 2017; Nelson et al. 2015). All diseases with drug-genetic ties can potentially be studied using Open Targets, although the focus areas of oncology, immunology and neurodegeneration are particularly likely to benefit. Another area where Open Targets has potential is neglected and rare diseases, which have been previously been under-studied and considered less attractive by private-sector investors (Priego and Wareham 2018).

4.1.8. The role of Sanger Institute

The Sanger Institute was a founding partner of Open Targets, and according to an interviewee, the project would likely not have happened in the same way without Sanger's input at the initial stages of the project (CS_05). EMBL-EBI and GSK were also founding partners and contributed vital resources and expertise (in bioinformatics and drug development, respectively) that were needed to launch Open Targets.

Sanger's expertise in genetics, genomics, biology and disease areas such as cancer and neurodegeneration (through the Cancer, Ageing and Somatic Mutations programme and DDD) has been instrumental in Open Targets, which has contributed significantly to the impact that it has achieved (CS_05). Sanger contributes researchers to Open Targets, as all PIs within Open Targets are Sanger staff (CS_04). The Genome Campus acts as a physical premises for the Open Targets project through EMBL-EBI and Sanger, allowing researchers and staff from across all the partner organisations to work closely together and exchange knowledge, which has been cited as an important facilitator of the programme (CS_03, CS_04).

4.1.9. Lessons learned from the case study

Open Targets has led to knowledge exchange between public- and private-sector organisations, benefiting all parties involved in the collaboration by providing them with new knowledge and skills to carry out drug development research. The programme has led to new areas of research in drug discovery, and has potential to generate health impacts as research and development activities in both the public and private sector continue to mature. Open Targets is an example of where Sanger has been able to employ 'smart openness' through public–private partnership to encourage innovation in drug discovery while still promoting openness and ensuring that its research has benefits for human health. Sanger is considered less commercial than comparators such as the Broad Institute, which are more likely to collaborate with single companies individually rather than with a consortium of organisations, and which considers closed licensing of IP in some cases. Open Targets represents a potential way forward for Sanger and other organisations with a similar focus on openness to work with the private sector in a pre-competitive setting, which may help them achieve health impacts through the private sector while avoiding the potential controversy associated with closed collaboration with single companies. Additionally, Open Targets points to the 'softer' benefits of engaging with private industry, such as the technical and tacit knowledge that has been passed to Sanger

through translational research and commercialisation, arming staff with new pathways to have impact, which may be a helpful approach for other actors working in this space.

4.2. Tree of Life

4.2.1. Introduction

The Tree of Life programme has been operational since mid-2019, but was disrupted by Covid-19. There have therefore been no substantial research publications at this stage, and most of the impacts described here are based on future projections. Recognising the somewhat unique nature of the case study, we focus on exploring the thinking in creating the programme, the role Sanger Institute is playing in the programme, and understanding how it is likely to shape human health research. We also consider the extent to which the Tree of Life programme represents a departure for Sanger.

The Tree of Life programme is an umbrella term to identify Sanger's research encompassing the Darwin Tree of Life project.[68] Recognising that the terms may be easily conflated, for this case study:

- The Tree of Life programme is intended to investigate 'the diversity of complex organisms (eukaryotes) found in the UK through sequencing and cellular technologies'.[69]
- The Darwin Tree of Life project is a UK-wide initiative to read the genomes of all 60,000 complex species (eukaryotes) in Atlantic archipelago. Funded by Wellcome (£9.3m, awarded in 2019), its partner organisations include the Natural History Museum, Royal Botanic Gardens, Kew, Royal Botanic Gardens Edinburgh, European Bioinformatics Institute, University of Cambridge, Earlham Institute, the Marine Biological Association and the Wytham Woods long-term ecological observatory project at the University of Oxford. It is loosely affiliated with the Earth BioGenome Project, which 'aims to sequence, catalogue, and characterize the genomes of all of Earth's eukaryotic biodiversity'.[70]

Additionally, aligned with the Tree of Life programme, Sanger has also received £4.1m in funding from the Gordon and Betty Moore Foundation to sequence 2,000 aquatic organisms and their symbionts.

4.2.2. Background and context

The Tree of Life programme investigates 'the diversity of complex organisms (eukaryotes) found in the UK through sequencing and cellular technologies'. Its other main objective is to compare and contrast 'species' genome sequences to unlock insights into evolution and conservation'. Led by Professor Mark Blaxter, the Tree of Life programme covers a number of activities related to the Darwin Tree of Life project and faculty research.

[68] As of 28 July 2020: https://www.darwintreeoflife.org/

[69] As of 28 July 2020: https://www.sanger.ac.uk/programme/tree-of-life/

[70] As of 28 July 2020: https://www.earthbiogenome.org/

The original impetus for the Tree of Life programme is the evolving, emerging approach to human health identified as One Health.[71] The One Health approach examines the health of human societies in the context of the ecosystem services provided by the natural world (Atlas 2012; E. P. J. Gibbs 2014; Wellcome Sanger Institute 2020). From this perspective, the future of human health depends on conservation of biodiversity and the discovery of new biomaterials and pharmaceuticals. Genome sequencing of all species, whether animal, plant, fungal, protozoal or prokaryotic, is thus argued to be essential to harness the potential of a natural ecosystem for future medicine and human health applications (Wellcome Sanger Institute 2020).

This approach is behind the Earth BioGenome Project, which proposes the sequencing, in the next decade, of all 1.5 million described eukaryotic species, starting with reference genomes for the 9,500 taxonomic families.

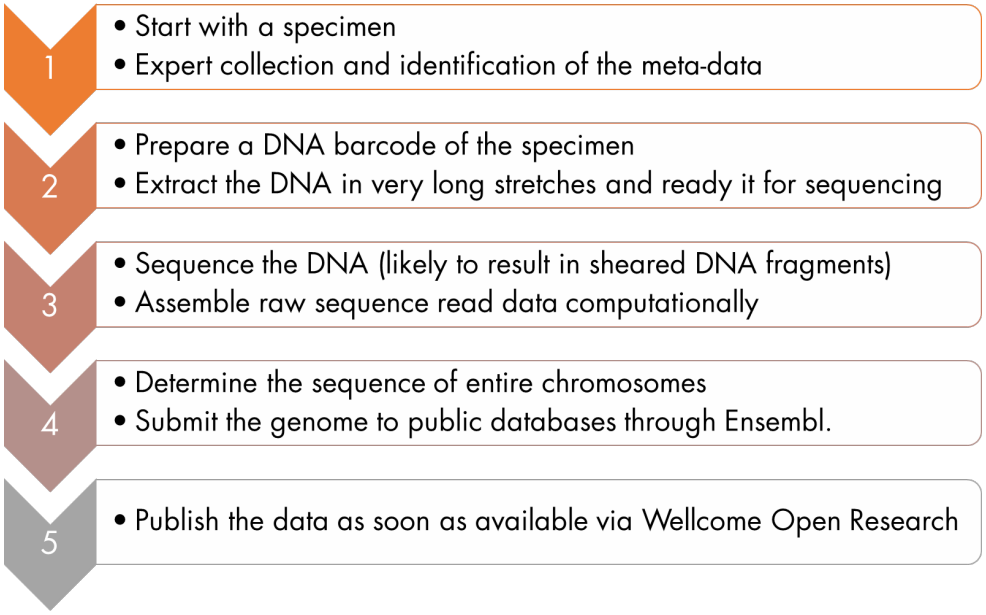
The Tree of Life programme is intended to build on Sanger Institute's track record and in-house expertise in genome sequencing to develop and deliver the goals of the Earth BioGenome Project. The Darwin Tree of Life project is intended to kick-start this work by pump priming the funding received through the Wellcome Discretionary award (Wellcome Sanger Institute 2020). The work will leverage developments in long-read sequencing technologies and algorithms for assembly, annotation and interrogation that Sanger has adopted (Mantere, Kersten, and Hoischen 2019), so the task of sequencing and assembling a 1 Gb genome at scale can now be performed at low cost (estimated at £1,250 reagent costs) and rapidly (days). It will also leverage the Sanger Genome Reference Informatics Team expertise, which specialises in genome assembly evaluation, working closely with the international Genome Reference Consortium, the Vertebrate Genomes Project and others (such as the 1000 Genomes Project) to deliver genome assemblies of the highest quality (Wellcome 2020).

4.2.3. Research process

Figure 26 depicts the current research process as part of the Darwin Tree of Life project at a high level.

[71] As of 28 July 2020: https://www.who.int/news-room/q-a-detail/one-health

Figure 26: The research process for the Darwin Tree of Life project



Source: RAND Europe analysis based on the description available on Sanger's website[72]

As Figure 26 shows, the research process starts with a specimen.[73] For this specimen, an expert collects and identifies metadata about where and when it was found. Sanger prepares a DNA barcode of the specimen. DNA is then extracted in very long stretches where possible and readied for sequencing. Since this process results in sheared DNA fragments, after sequencing, the raw sequence read data is 'assembled' computationally to determine the sequence of entire chromosomes. The genome is then submitted to public databases through Ensembl.[74] Sanger's stated aim is to publish all data as soon as available via Wellcome Open Research.

Starting with the core funding it has received for the Darwin Tree of Life project, Sanger intends to collect and complete reference genomes for every genus in the Atlantic archipelago via the additional external funding (Wellcome Sanger Institute 2020). Sanger has identified two projects to achieve this objective (Wellcome Sanger Institute 2020):

- Project 1 to create an infrastructure that innovates, builds and delivers reference genomes for thousands of diverse taxa and grows the Tree of Life
- Project 2 to capitalise on the infrastructure built in project 1 to investigate species diversity, species interactions and the evolution of genome structure.

[72] As of 28 July 2020: https://sangerinstitute.blog/2020/02/15/sequencing-and-the-tree-of-life/

[73] As described on the Sanger Institute website. See https://sangerinstitute.blog/2020/02/15/sequencing-and-the-tree-of-life/

[74] Ensembl is a joint scientific project between the EMBL-EBI and the Sanger Institute to provide a centralised resource for geneticists, molecular biologists and other researchers studying the genomes of our own species and other vertebrates and model organisms (Flicek et al. 2010). It is a genome browser – a graphical interface for display of information from a biological database for genomic data (Wang et al. 2013).

The Darwin Tree of Life project corresponds to the initial phase of these objectives. As part of the Darwin Tree of Life project Sanger aims to collect and supply specimens for 8,000 species from the Atlantic archipelago through the Genome Acquisition Labs (partners organisations such as Wytham Woods, Royal Kew Gardens, Royal Botanical Gardens Edinburgh) (Wellcome Sanger Institute 2020). Sanger's aim is to have sequenced, analysed and published 2,000 of these by working with EMBL-EBI and the Earlham Institute (the Analysis Hubs) (Wellcome Sanger Institute 2020).

Sanger proposes to base its prioritisation of genomes sequencing of species on the following criteria (not listed in a specific order) (Wellcome Sanger Institute 2020):

- Species of specific family or generic references
- Genome sequences that represent an ecosystem or group
- Genomes which have strong public recognition
- Genomes which are currently the subjects of investigation by Sanger or others because of striking phenotypes and landscape ecology
- Species that are keystone in ecosystems or are expected to be revealing of particular processes.

Undertaking the Tree of Life programme will be challenging, but will ultimately prove beneficial for Sanger Institute because it will strengthen its role as a leader in the field. Given the scope of the Tree of Life programme and the research process outlined above, sequencing 66,000 species is a massive logistical challenge, will require a lot of coordination on the part of Sanger to obtain reference material for every species, and will be a major effort when classifying the species targeted (CS_08). However, once the data is collected, the work on the Darwin Tree of Life project is likely to see Sanger lead on the collaboration over sequencing (CS_08). Sanger has a well-established repertoire of sequencing techniques to this end and, in the interim, data analysis techniques and software are expected to improve to handle the scale of the project (CS_08). Although assembly and annotation of genomes will be a challenge, Sanger's ambition and approach will be critically important in this context (CS_08).

Sanger's research approach is distinctive and yet important because previous technologies have only allowed sequences with lots of gaps and ignored differences between parental DNA strands (CS_09). Therefore, data was often highly fragmented and genomes were likely to be of low quality (CS_09). With the Darwin Tree of Life, Sanger's approach has been to carry out development and production work on the genomes in parallel (CS_09). As a result, Sanger is in a position to take a leadership role and demonstrate that high-quality genomes can be delivered at reasonable costs (CS_09).

When other ongoing efforts in this space are considered, although the Earth BioGenome Project is an international consortium, its mandate is to mainly encourage countries to contribute (CS_08). Sanger's Darwin Tree of Life is one of the early efforts in this space. The extent to which it is comparable with efforts in other countries can be understood only after the other similar efforts progress further.

4.2.4. Contributions to knowledge

The Darwin Tree of Life project is in its early stages and so far, Sanger has focused on publishing genomes of high quality. By sampling, extracting and sequencing at scale, Sanger expects that in five years, it will publish 20 papers a day.

The main contribution to scientific knowledge by the Darwin Tree of Life project since November 2019 has been the approximately 100 genomes generated, processed and made ready for release (Wellcome Sanger Institute2020). Working with Genome Reference Consortium partners, the Genome Reference Informatics Team have improved near-definitive references for human, mouse, zebrafish and chicken. Researchers are also working on other model references, such as rat (Wellcome Sanger Institute 2020).

According to one interviewee, Sanger aims to change the model of genomics research, with the intent of publishing genomes quickly and openly (CS_07). By publishing genomes as soon as they are available, the objective is to change the norms around scientific publication of genomes (CS_07). Although the norms are already changing to an extent, with the Darwin Tree of Life project, Sanger is well placed to push to change the model of publishing genomes completely (CS_07). Considering the interest generated by the Otter genome sequences generated by early 2020, the work being done by Sanger can be expected to play a key part in wildlife conservation in the UK (CS_07).

Beyond individual species, the work of Sanger is expected to enable investigations to take place related to fundamental scientific questions such as why humans have 23 chromosomes, and how and why chromosomes change (CS_07). The findings could be crucial in cancer research or biodiversity research studies (CS_07).

Sanger is committed to open science and has traditionally not been interested in patenting methods for genome sequencing (CS_09). However, Sanger often works with commercial organisations engaged in sequencing and DNA extraction and its commitment to publishing data and results openly was seen to be exemplary in terms of scientific contributions by one interviewee (CS_09).

The reference dataset of genomes is expected to broaden the existing field of genomics research to a range of species about which very little is known at this stage (CS_08). According to one interviewee the potential scientific advancements are likely to offer a range of benefits to human health research (CS_08).

4.2.5. Contributions to future research

Within the scientific field, the Tree of Life programme and the Darwin Tree of Life project are considered the first of their kind in attempting to produce high-quality non-human species genome datasets at scale as part of a highly coordinated effort (CS_08, CS_14 and CS_15). Sanger's work in the Tree of Life programme is likely to inspire similar efforts in the field and this augurs well for the future of genomics research and biodiversity research, according to one interviewee (CS_14).

The Tree of Life programme's Quinquennial Review 2021–26 highlights some of these spill over effects with similar initiatives emerging in Norway, Sweden and Spain (Wellcome 2020i). Sanger is discussing setting up Norwegian and Swedish Tree of Life programmes. In Spain, the Catalonian Society for Natural Sciences is planning a project in Catalonia with the aim of covering the Mediterranean basin (Wellcome Sanger Institute 2020). Sanger was invited to speak at the launch conference of the Catalonian Tree of Life project in early 2020 (CS_07, CS_14).

When discussing the potential impact of the Tree of Life programme on future research, four interviewees highlighted the current Covid-19 pandemic and how human health is inextricably linked with the broader natural world ecosystem (CS_08, CS_09, CS_14 and CS_15). According to one interviewee, the Darwin Tree of Life project is setting an example of how to do genome sequencing of non-human species and thus is likely to have a significant impact on future research in medical and pharmaceutical innovation (CS_09). The high-quality genome datasets published by the Darwin Tree of Life project are likely to become essential reference for genomics researchers (CS_09). According to another interviewee, Sanger's continued commitment to release the datasets in an open manner is a crucial catalyst to further developments in genomic knowledge (CS_08). Sanger's work on the Darwin Tree of Life project and the resulting high-quality datasets were likely to create increased scientific capacity in the UK related to genomics and genetics research (CS_08).

4.2.6. Contributions to policy and product development

In addition to the high-quality genome assemblies, Sanger Institute intends to develop new ways of generating sequencing libraries from difficult samples and plans to develop toolkits for data analysis (Wellcome Sanger Institute 2020). These products and tools are expected to have a cross-cutting impact across a number of genomics sub-disciplines, including biodiversity genomics, clinical genomics, human genomics, single cell genomics and pathogen genomics (Wellcome Sanger Institute 2020).

The work on the Darwin Tree of Life project is expected to encourage further investment in technologies for genome sequence generation and result in improved quantity, quality and length of data generated over the first five years of the Tree of Life programme (CS_07). Starting with the Darwin Tree of Life project, Sanger Institute is targeting computational efficiencies, the use of new hardware platforms, and data storage (Wellcome 2020). Towards this end, Sanger staff are engaged in working with bioscience companies and technology companies actively engaged in genomics research. Key examples are organisations such as Oxford Nanopore Technologies and Pacific BioSciences with whom Sanger is collaborating to deliver high-quality genome assemblies at reasonable costs (Wellcome Sanger Institute 2020).

One interviewee believed Sanger's policy of publishing all datasets openly is essential to technology and product development in the genomics space, which is highly iterative in nature (CS_09). Sanger's approach in this context was seen to distinguish it from other projects which have been imposing embargos on data sharing (CS_09). Such an approach was deemed to enable Sanger to maximise the potential of its datasets, and strengthen its reputation as a leader in genomics CS_09).

In order to have an impact on UK environmental policy, as part of the Darwin Tree of Life project, Sanger is engaging with the Department for Environment, Food and Rural Affairs; the Scottish Environment Protection Agency; Scottish Natural Heritage and Natural England as part of the sampling process (Wellcome Sanger Institute 2020). Additionally, Sanger Institute intends to engage with national monitoring agencies such as the National Biodiversity Network and the County Recorders scheme, non-governmental organisations such as Buglife, the Royal Society for the Protection of Birds, Butterfly Conservation and amateur naturalist societies to increase the outreach of the genomics datasets released through the Darwin Tree of Life project (Wellcome Sanger Institute 2020).

Connecting Science (which is part of Wellcome Genome Campus) is working with partners in the Darwin Tree of Life project to prepare a bid worth £450,000 to create engagement opportunities in the form of pop-up events and exhibitions and real-world DNA barcoding and metagenomics activities in schools and other public places (Wellcome Sanger Institute 2020).

4.2.7. Contributions to health and the health system

Although the Tree of Life programme potentially represents a different direction for Sanger Institute through its involvement in biodiversity research, four interviewees (CS_07, CS_08, CS_09 and CS_15) highlighted that this research is rooted in the One Health paradigm and thus expected to be increasingly important to human health, which is inextricably tied to agricultural produce, and the impact of pests on food production and the surrounding ecosystem (CS_09). Two interviewees highlighted that with the Tree of Life programme Sanger is well positioned to progress the knowledge of human genetics beyond the traditional anthropocentric view of the world (CS_09 and CS_15). One interviewee highlighted the increased risk of zoonotic transfer of animal pathogens to humans (CS_15). By deploying genetics and genomics at scale to biodiversity, Sanger is expanding the footprint of its earlier work on human genome sequencing (CS_15).

The contributions of the Tree of Life programme are thus intended to inform genome changes across life and further exploration of why humans can tolerate some changes but are pathogenic to others. The work in the Darwin Tree of Life project is expected to lead to identification of new models of ageing, disease and resistance (Wellcome Sanger Institute 2020).

4.2.8. The role of Sanger Institute

According to one interviewee, Sanger's Tree of Life programme is the first to have the ambition to do genome sequencing of non-human species at a large scale despite a number of other organisations or consortia (including the Vertebrate Genome project and Genome 10K project) being engaged with the Earth BioGenome Project at some level (CS_09). Another interviewee identified Sanger as the first organisation to have begun the work while others are in early stages of preparation (CS_07). Other projects such as the Swedish Tree of Life project and the Catalonian Tree of Life draw on the approach Sanger is pursuing and intend to emulate their strategy of engaging local partners to achieve extraction and sampling at scale (CS_07, CS_08 and CS_15). According to one interviewee, the Darwin Tree of Life project is thus unique because of the focused effort it will receive from an organisation of Sanger's expertise (CS_09).

According to one interviewee, Sanger has played a leading role in bringing together a consortia of different complementary skills and can thus be considered the central hub, effectively the engine and the control room, of the Darwin Tree of Life project (CS_15). Four interviewees believed that a consortium like this aiming to deliver genome sequences at scale is without precedent and would not have been possible without Sanger's involvement (CS_08, CS_09, CS_14 and CS_15).

In the broader context of developments in the scientific field, the Tree of Life programme can be an important strategic step for Sanger, since it has not traditionally engaged in biodiversity research (CS_08).Sanger taking a leading role in the Tree of Life programme can be expected to benefit the development of the One Health paradigm and the field of genomics and genetics (CS_08).

4.2.9. Lessons learned from the case study

This case study suggests that although the Tree of Life programme may be seen as a departure from the work Sanger has done in the past, it seems to align with the long-term transition within the field to the One Health paradigm to improving human health. Sanger's ability to complete genomics at scale and commitment to delivering open datasets continue to be seen as key differentiators within the field. Although the Tree of Life programme may not be perceived to be as ground-breaking in scientific and technological terms as the Human Genome Project, it has the potential to be ground-breaking in scale and the resources it aims to bring to sequencing non-human species (CS_08). Sanger is perceived to be leading the effort on high-quality genome sequencing of non-human species and poised to make important contributions to genomic science through the Tree of Life programme.

Despite the near consensus among interviewees on Sanger's leadership role and the potential of the Tree of Life programme to impact genomic science, however, some challenges to realising the ambitions can be discerned. Research initiatives such as the Earth BioGenome Project, the Genome 10K project, the Vertebrate Genome project or the International Barcode of Life project are loosely collaborative international consortia that rely on efforts at regional and national level pulling together. Sustaining funding and scientific effort for such projects over the long term is challenging. From this perspective, although the science behind the Tree of Life programme is fundamentally incremental, the programme provides a potential logistical model for other countries to learn from and replicate. However, the extent to which the Tree of Life programme will result in sustained global effort in sequencing all genomes of all life as demonstrated by the Earth BioGenome Project is not clear at this stage.

The Tree of Life programme has an ambitious long-term objective of genome sequencing of 18,000 species (Wellcome Sanger Institute 2020). At this stage the Darwin Tree of Life project is designed to run for 2.5 years plans to sample 8,000 species and deliver genome sequences for 2,000 species (Wellcome Sanger Institute 2020). Sanger has received funding from the Gordon and Berry Moore Foundation for the Moore Foundation Aquatic Symbiosis Genomics project (Wellcome Sanger Institute 2020). Although Sanger is in discussions with NERC and BBSRC, the funding and resourcing roadmap for fully realising the objectives of the Tree of Life programme is still emerging (Wellcome Sanger Institute 2020).

There are significant advantages to the centrally coordinated approach adopted for the Darwin Tree of Life project in order to deliver science at scale, make efficient use of available resources, facilitate economies of effort and create high-quality genome sequences. Such an approach overcomes the challenge of uncoordinated genome sequencing, which can result in fragmented, low quality data of no scientific utility.

4.3. DDD

4.3.1. Introduction

DDD was a Sanger-led project that aimed to diagnose children with disorders with an unknown genetic basis. So far, 35–40% of children participating in the study been diagnosed and over 200 scientific papers have been published (Caroline F. Wright et al. 2018). The project led to the establishment of Congenica and new micro-array technology (Int_08). DDD is now a model genome initiative, demonstrating the value of uniting fragmented testing practices and scaling it up to the national level (CS_13). DDD has influenced

GEL's study design and strategy, and inspired off-shoot international developmental disorder genomic programmes (CS_13).

4.3.2. Background and context

The DDD project initially began in 2009–2010 as a continuation of DECIPHER, a global clinical database developed at Sanger Institute in 2004 (CS_11) in an attempt to populate DECIPHER with consistent, high-quality data, ensuring it could be used universally (CS_10). Sanger faculty associated with the DECIPHER project sought to generate this kind of data themselves, collecting high-quality genetic and clinical information and providing more systematically genotyped and phenotyped patient data in DECIPHER. In doing this, they aimed to learn more about the genetic architecture of undiagnosed developmental disorders.

However, with the advent of exome sequencing technology just as the project was launched, the ambitions of the project grew beyond populating DECIPHER with higher quality data (CS_10). Exome sequencing allowed the DDD team to sequence all genes in the genome, so they could now complete comprehensive genetic testing on a massive scale. There was a clear scientific interest in conducting such a large-scale study: children with major developmental disorders were not included in most genomic databases, so, after sequencing their exomes, any differences observed from general population datasets were more likely to be significant in the onset of the disorder (CS_11). In short, they did not require tens of thousands of patients to identify genes associated with the disorder. The study also had potential clinical benefits. When DDD started, thousands of children were born each year in the UK with undiagnosed developmental disorders, many of which had a clear, but hitherto unknown, genetic basis (CS_11). By examining the genes of these children and their parents, the DDD team aimed to gain a diagnosis.

4.3.3. Research process

The programme was funded by the Health Innovation Challenge Fund, a joint venture between the Department of Health and Wellcome (Wellcome 2019g). Sanger received the funding and acted as the 'host institute' from the outset (CS_10). One interviewee has described Sanger Institute as the 'nexus of the project intellectually and organisationally', as Sanger and its associated staff were essential in structuring the organisation of the project (CS_10). For example, Sanger and associated faculty planned how DNA samples could be collected and shared across the UK. While Sanger played a central organisational and scientific role, DDD was also very much a collaborative effort. A crucial element of the collaborative effort was partnering with all 24 of the UK and the Republic of Ireland's regional genetic services. They placed a nurse in each centre 'to undertake systematic phenotyping and detailed genomic analysis for 13,000 children' (Firth and Wright 2011). The genetic centres provided the clinical interface for the study, recruiting families and collecting DNA samples and giving feedback once a diagnosis had been returned (CS_12). The inclusion criteria for the study were the presence of a neurodevelopmental disorder, congenital abnormalities, dysmorphic features, unusual behavioural phenotypes or a genetic disorder for which the molecular basis is unknown (Firth and Wright 2011). These criteria were deliberately designed so as to maximise the chance of finding a diagnosis for the children and families.

Before the DDD project, the genetic centres were disparate and used different testing procedures so within the NHS there was what one interviewee described as a 'postcode lottery' in the quality and resolution of

genetic data generated at each centre (CS_10). DDD worked to change this, as testing procedures were standardised. To ensure standardisation, all clinical geneticists were trained in the use of the Human Phenotype Ontology (HPO)(Wellcome 2019g). This helped the regional genetic services to conduct systematic clinical phenotyping and improve clinical annotations, which in turn helped Sanger staff with their analysis. Similarly, before the DDD study, the genetic centres did not openly share data between themselves (CS_10). Again, this changed over the course of the project and clinical geneticists could compare findings from across the country, rather than being limited to data from their local region.

Sanger received all the DNA samples and completed the sequencing and genomic analysis of all 40,000 of the samples collected by the genetic services (CS_10). These samples came from children and parents, employing what is known as a 'trio approach'. Such an approach enabled the DDD team to better determine the genetics of the disorders and the role of *de novo* mutations – those that are not present in either parent – in the onset of the disorders. Although they were not the first to use the trio approach, the DDD project geneticists were the first to demonstrate that it could work effectively at a large scale (CS_10).

A key part of DDD involved launching the empirical ethics project GenomEthics, led by Dr Anna Middleton, ethics researcher for DDD, which accompanied the scientific and clinical work.[75] GenomEthics explored the views of the public and professionals around genomics and the use of genomic data. This work has informed the project Your DNA, Your Say, run by Dr Anna Middleton, now Head of Society and Ethics Research at Connecting Science.

4.3.4. Contributions to knowledge

As genetic and phenotypic data was made available to the international research and clinical community via the DECIPHER database, publications using DDD data have come from around the world. As of May 2020, there have been 203 publications with 'DDD study' listed as an author, many appearing in *Science*, *Nature*, *The Lancet* and other prestigious journals. Many of these publications emerged from a complementary analysis programme (CAP): collaborative, smaller-scale research projects run by the regional genetic services (CS_12). Clinical geneticists, using information available from DECIPHER, could see colleagues with similar patient data and genetic mutations and establish their own research programme as a CAP. For example, one CAP involved geneticists in London and Edinburgh collaborating to complete a project around Gillespie syndrome, as they both had patients with very similar genetic data (CS_12) (McEntagart et al. 2016). Such CAPs provided more detailed publications than the large-scale analysis performed at Sanger Institute and boosted the research capacities of the genetic centres. One interviewee reported working within a hospital at a genetics service that, before DDD, did not have particularly strong research outputs, but now has the largest research output of the entire hospital. They credit this shift as being 'almost entirely due to the DDD study' (CS_12). Thus, the DDD study has helped to inform the genomics community and improve the standard of genomics research within the UK and globally.

The scientific leads of DDD have identified four particularly important publications coming directly from the DDD team that they felt changed the genomics landscape (CS_10):

[75] As of 28 July 2020: https://societyandethicsresearch.wellcomegenomecampus.org/project/genomethics

1. Their first publication in *Nature* in 2014 was crucial, as it demonstrated the success of the trio approach for large-scale genomics studies (C. F. Wright et al. 2019). In outlining 12 new genetic disorders, the team showed that the trio approach could, with statistical analysis, be an important diagnostic tool. Since then, collating data on over 31,000 families, they have identified 49 novel dominant disorders.[76]
2. Another publication described as key again featured in *Nature*, as the DDD team highlighted the global healthcare impact of *de novo* mutations (McRae et al. 2017). Here they estimated, for the first time, that 1 in 300 live births worldwide have a *de novo* mutation causing a developmental disorder, resulting in approximately 400,000 births annually of babies with developmental disorders. In this paper, the DDD team also identified 14 new developmental disorders.
3. In a third key publication, the DDD study demonstrated that non-coding *de novo* mutations in regulatory elements of the genome play a central causal role in the onset of many neurodevelopmental disorders (Short et al. 2018). This was the first time that this had been shown, as researchers had previously tended to only focus on genes. However, they were able to demonstrate, quantitatively, that non-coding regions played a role in the aetiology of developmental disorders. From this, they were the first to provide a 'robust estimate' of how much one could learn from exome sequencing compared to whole-genome sequencing (CS_10). According to one interviewee associated with the DDD study, this helped inform discussions at the National Institute of Health in the United States around what kinds of disease studies they should be discussing and funding (CS_10).
4. Finally, the scientific leadership of DDD identify a 2018 *Nature* publication as crucial: the DDD team demonstrated that there is a significant polygenic contribution to severe developmental disorder, and that polygenic mutations change the phenotypic expression of disorders typically seen as having a monogenic cause (Niemi et al. 2018). They identified that common variants can have a significant cumulative effect in increasing the risk of onset of developmental disorders, while most researchers had previously focused on rare genetic variants.

Separate from the genomic research undertaken, the GenomEthics strand of DDD published a study with the largest dataset – surveying 7,000 patients, clinicians and geneticists from 75 countries – to date of attitudes towards issues surrounding the return of incidental findings from sequencing research (Middleton et al. 2016). This final publication won the 2nd prize European Society of Human Genetics 2018 award for citations and impact.

4.3.5. Contributions to future research

Global partners have learned from DDD and have sought to replicate similar programmes in their own countries, borrowing Sanger's expertise. Even during the course of the project, Japanese and Singaporean scientists and officials came to the campus to learn how they could implement the project in their own

[76] As of 28 July 2020: https://www.sanger.ac.uk/news_item/milestone-reached-major-developmental-disorders-project/

countries and how Sanger could advise them (CS_10). One concrete example of this is DDD-Africa, where Sanger Institute acts as a partner. According to the DDD-Africa website, the 'project will be modelled on the (DDD)-UK study, which successfully facilitated the translation of genomic sequencing technologies for diagnosing DD [developmental disorders] in the UK'.[77] The NIH gave Sanger a grant to help in this endeavour (Wellcome 2019g). Additionally, following the model DDD-UK, Sanger Institute will conduct a similar large-scale exome sequencing and SNP genotyping study on patients with developmental disorders in India, called DDD-India (Wellcome 2019g).

4.3.6. Contributions to policy and product development

The biotech company Congenica Ltd emerged as a spin out from the DDD study. One interviewee involved in the study has described it as an effort 'to make sure that what we had learned was able to be taken up by the NHS' (CS_10). Congenica researchers work closely with the regional genetic services to develop new technology, such as rapid foetal exome sequencing, non-invasive exome sequencing, and using maternal blood samples to identify potential mutations in the child, and are currently working on a pharmaco-genetic project with Sanger researchers (CS_12). Additionally, Congenica is now the leading provider of clinical decision support software to GEL.

DDD staff, with help with Sanger's translation team, also worked with OGT to develop and commercialise new micro-array technology. They have co-designed an improved clinical micro-array for detecting pathogenic deletions and duplications. The new micro-array technology developed with OGT was, and continues to be, critical in diagnosing new disorders. This higher resolution sequencing technology enables clinicians and scientists to look for smaller changes and deletions in the exome than was previously possible (CS_11). This improved array has been licensed, and has achieved more than £9m global sales, returning a 1% royalty to Sanger Institute (Wellcome 2019g).

The DDD project has been heralded by Sanger and GEL interviewees alike as a 'trailblazer' for the 100,000 Genomes Project, informing its design and execution (CS_11, CS_13). DDD's influence on GEL can be seen in the latter's decision to obtain trio data for the rare disease strand of the 100,000 Genomes Project, and general clinical approach, using DECIPHER, and showing GEL how best to collect samples and return results to the NHS (CS_13). Moreover, the DDD project was most influential to GEL in demonstrating to policymakers that large-scale genomic sequencing projects could return results for the NHS (CS_13). In a 2012 UK Life Sciences strategy document, the DDD project was used as a case study of nationwide large-scale genomic research that could improve NHS diagnostic testing.[78] Without the precedent set by the DDD project, an interviewee at GEL said they would have been starting from 'ground zero' on the 100,000 Genomes Project (CS_13). Furthermore, scientific collaborations between the DDD study team and GEL are still ongoing, as DDD staff advise on the strategical planning of GEL, keeping them 'at the forefront' of scientific and technological developments (CS_10, CS_13).

[77] As of 28 July 2020: https://h3africa.org/index.php/consortium/deciphering-developmental-disorders-in-africa-ddd-africa-evaluating-clinical-exome-sequencing-in-an-african-setting/

[78] As of 28 July 2020: https://h3africa.org/index.php/consortium/deciphering-developmental-disorders-in-africa-ddd-africa-evaluating-clinical-exome-sequencing-in-an-african-setting/

However, others have questioned the comparison between DDD and the 100,000 Genomes Project. One clinical geneticist told us that DDD did not have 'as much as impact as it should have' on genomic medicine policy in the UK and GEL should have followed the DDD study design and approach more closely, observing that DDD was 'much more agile and attuned to clinical and patients' needs than the 100,000 Genomes Project, and also had a better diagnostic output (CS_12). A GEL interviewee conceded that diagnostic yield has been lower than for the DDD project, 25% and 40%, respectively, and GEL researchers are working with DDD and other Sanger staff to improve their yield (CS_13). In addition to the lower diagnosis rates of the 100,000 Genomes Project, one interviewee remarked more generally that working on the project has been more difficult than working on the DDD project, which made data collection and sharing easy (CS_12).

4.3.7. Contribution to health and the health system

Around 35–40% of the more than the 13,000 children taking part in the DDD study have gained a diagnosis,[79] which enables them to receive more tailored and focused treatment. A diagnosis can further personalise care as gene-specific growth charts become a possibility: clinicians can compare the development of the child with others with a similar disorder or mutation, so families are more aware of the development of the child (CS_11). These children may be progressing well given their disorder, but before being given an exact diagnosis, this was impossible to know as only comparisons with the general population were available.

According to a clinician closely involved in the DDD study, gaining a diagnosis also helps improve awareness of prognosis, knowledge of risks to other family members, and alleviates guilt from parents concerned that their child's disorder is their fault (CS_11). Furthermore, a diagnosis provides the opportunity to be part of a global network of families that have similar disorders and support groups. The DDD study has been proactive in collaborating with existing patient support groups, Unique and syndromes without name (SWAN), in co-writing informational leaflets,[80] which provide an overview of the disorders and what they mean for the family in an accessible format. The DDD website has a wealth of information and advice around genetic disorders, further helping those affected to understand their disorder.[81]

The DDD project improved ways of working within the NHS regional genetic services. Before the DDD study, these services did not openly share data with each other, but now operate as far more of a collaborative unit. Having seen the benefits of sharing data with each other in the DDD study, the regional services now share all of their patient data in an open consortium (CS_10). In DDD, various centres worked closely together on their own research programmes, CAPs, which further unified them as a national genetic service. Furthermore, the DDD project helped to familiarise them with exome sequencing technology that is now routinely available in NHS diagnostic labs. One clinician told us, 'DDD prepared us (clinicians) for this technology and enabled us to have conversations with scientists that we would not have had before'

[79] As of 28 July 2020: https://www.sanger.ac.uk/news/view/first-non-gene-mutations-behind-neurodevelopmental-disorders-discovered

[80] As of 28 July 2020: https://www.rarechromo.org/disorder-guides/

[81] As of 28 July 2020: https://www.ddduk.org/genome.html

(CS_12). The DDD project also cemented the DECIPHER database as a 'go-to database' for clinical genetics in the UK, whereas the database was initially seen as being little more than a repository of chromosomal data, following feedback from clinicians over the course of DDD, its functionality markedly improved and it became an invaluable clinical resource (CS_12). At least for this clinician interviewed, DECIPHER is now the primary source of information on genetic variation (CS_12).

4.3.8. The role of Sanger Institute

Sanger, along with the regional genetic services, played a central role in the DDD project. With the new exome sequencing technology, staff scaled up the research to a level that would not have otherwise been possible and provided scientific and bioinformatics expertise. One interviewee even remarked that DDD 'would not have happened without the Sanger' (CS_11). As a result of the open data policy and collaborative approach that is typical of Sanger projects, the DDD study was able to produce more publications than if data was restricted to researchers within the Sanger Institute (CS_10).

Sanger played an active role in unifying the previously disparate 23 regional genetic services. Staff facilitated meetings with representatives from the regional NHS centres at the Sanger Institute two years before the project formally began (CS_10). Such meetings were crucial to building trust, especially as some clinicians were wary of being handed masses of meaningless data. Sanger staff on the DDD project study further assuaged these concerns by returning only selected genetic information to the clinicians.

Sanger's translation office was heavily involved in assessing commercial partners for developing the array, and they provided a steer for partnering with OGT. Scientific leads on the DDD project have said that the collaboration with OGT worked well, in no small part thanks to the efforts of Sanger's translation team, facilitating meetings and collaboration. Sanger also played an active role in supporting Congenica, allowing its staff to spend up to a day a week establishing the company. Furthermore, in its first three years, Sanger staff and particularly those associated with the DDD study worked very closely with Congenica, lending their expertise and transferring their know-how. They still cooperate closely, and some DDD study staff continue to be associated with Congenica.

4.3.9. Lessons learned from the case study

On a technical level, the DDD study demonstrated that large-scale exome sequencing coupled with the trio approach could generate scientific and clinical benefits. Before the DDD study, this had not been demonstrated, but international partners are now embarking on similar kinds of projects using this methodology.

Second, the DDD study highlighted the benefits of collaboration. The technical achievements of the project were only possible because of the close collaboration between Sanger and regional genetic services. Before recruitment of patients had begun, they were discussing with clinicians how to gather and analyse samples and work most effectively together. Involving clinical scientists in the managing committee was also integral to the success of the project, as they could provide perspective on how to ensure buy-in from clinical geneticists and the NHS. Indeed, many of the one interviewee's criticisms of the 100,000 Genomes Project relate to the latter's relative lack of clinical engagement in comparison with the DDD study (CS_12). Finally, like many other Sanger projects, the open data policy reaped rewards over the course of the DDD

project. Over 200 papers have been published, most of which came from external scientists using DDD data.

4.4. Malaria research and MalariaGEN

4.4.1. Introduction

This case study explores MalariaGEN, and the contribution this made in LMICs. MalariaGEN is a scientific network that connects researchers and clinicians working to understand how genetic variation of humans, the malaria parasite (*Plasmodium falciparum*) and the vector (mosquito) affect the spread of malaria, and to use this knowledge to develop effective ways to control the disease. MalariaGEN's activities are widespread and include coordinating large-scale international projects, developing and promoting policies around data sharing, supporting research capacity in malaria-endemic countries, and providing an opportunity for future collaborations and networking opportunities.

4.4.2. Background and context

MalariaGEN was founded in 2005 to use genetic epidemiology data to provide insight into important aspects of malaria biology and spread (Achidi et al. 2008). The studies conducted through MalariaGEN use DNA sequencing technology to investigate the genomes of the human host, malaria parasite and mosquito vector to identify any genetic variants that may be important in determining the evolution and transmission of the disease. MalariaGEN was initially supported by joint funding from the Bill & Melinda Gates Foundation (through the Foundation for the NIH) and Wellcome, as part of the Grand Challenges in Global Health Initiative[82] awarded to Dominic Kwiatkowski. The first projects undertaken through MalariaGEN investigated the human genetic factors involved in resistance to malaria, and MalariaGEN's first Consortial Project produced a comprehensive dataset on the human genetic factors that determine resistance to malaria (Band et al. 2019).

At the time of MalariaGEN's inception, malaria research was often fragmented across different research groups, each pursuing their own small studies.[83] The small sample size of the datasets these approaches produced combined with the varying methodological approaches meant that the data were often difficult to interpret, with limited statistical power resulting in inconclusive results (CS_06). A key insight that researchers in MalariaGEN had was that to be able to gain a true picture of the disease, a much greater number of samples was required (CS_17). Combining the data enabled researchers to obtain more scientifically accurate results in order to gain a more comprehensive picture of the disease (CS_17). MalariaGEN enabled large-scale collaborative studies to be established, which could integrate genetic, clinical and epidemiological data to provide more robust insight into how to address the key scientific questions around malaria. The projects are supported by the MalariaGEN Resource Centre, based primarily between the WHG and the Big Data Institute at the University of Oxford, and Sanger Institute.[84] The

[82] As of 28 July 2020: https://gcgh.grandchallenges.org/sites/default/files/DominicKwiatkowski.pdf

[83] As of 28 July 2020: https://www.malariagen.net/about/our-approach

[84] As of 28 July 2020: https://www.malariagen.net/about/coordination/malariagen-resource-centre

work is supported by funding from several sources including The Sanger Institute Core Grant, the MRC, the National Institute for Health Research and the Bill & Melinda Gates Foundation (CS_17).

4.4.3. Impact in LMICs

Through large-scale collaborative and community projects, data sharing and capacity building, MalariaGEN has provided a platform from which to support impact in LMICs.

Large-scale collaborative and community projects

A key reason for the establishment of MalariaGEN was the need to bring multiple research groups together in order to provide more robust and comprehensive data into the causes and consequences of malaria (CS_06, CS_17). Through MalariaGEN, the Plasmodium Falciparum Community Project was established. This project connects multiple research groups across different geographical locations to build a catalogue of genetic variation associated with the malaria parasite. Researchers across the different locations can send the parasite samples they have collected to be centrally sequenced at Sanger Institute, and the resulting analysis provides a way of making accurate comparisons across the different samples possible. The project currently involves 49 partner studies across 28 countries and has produced a comprehensive open data resource that has been used by groups around the world in over 50 publications (CS_17). A second example is the Ag1000G Project, an international collaboration to use whole-genome sequencing to give a high-resolution view of the genetic variation within the natural populations of the *Anopheles gambiae* mosquito in Africa, the principal vector of the malaria parasite.[85] According to one interviewee, this project has been critical to bringing together research efforts to sequence the anopheles vector, enabling discoveries into the genetic variation within natural mosquito populations, and producing an open data resource for mosquito research and surveillance (CS_17).[86] This project combines the data for 1,142 wild-caught mosquito specimens collected from 13 countries spanning sub-Saharan Africa, providing a vast resource for researchers in the malaria research community (CS_17).

A framework for data sharing

An important aspect of large-scale collaborative projects is the ability to share data between the research groups in a way that allows for fair attribution for each researcher's contribution (CS_06, CS_17). MalariaGEN has been cited as being pioneering in the establishment of a data sharing framework, by which the malaria research community can share data in a transparent and fair way (Tessema et al. 2019). This includes: fair acknowledgement of the contributions made by researchers, fair attribution of discoveries made related to the data and fair sharing of any benefits associated with the data (CS_06). In establishing a framework, MalariaGEN has provided a way in which these issues can be clearly laid out, establishing the rules beforehand, and ensuring the process is transparent and clear. This has allowed researchers the flexibility to share their data, but also the security that the data they share belongs to them, and that they will be fairly acknowledged for the effort they have contributed to its collection. This framework has

[85] As of 28 July 2020: https://www.malariagen.net/projects/ag1000g

[86] As of 28 July 2020: https://www.malariagen.net/resource/27

provided a set of policies that enable the malaria LMIC research community to share data in an equitable way as the networks and collaborations develop (CS_06, CS_17) (Achidi et al. 2008).

Building sequence capacity

MalariaGEN has moved towards developing sequence capacity within LMICs (CS_06, CS_17). One project, funded by a grant through the National Institute for Health Research, is developing sequencing capacity in West Africa.[87] The aim of this project is to develop local surveillance mechanisms to monitor resistance within the parasite and mosquito populations, in order to inform the strategy of the national malaria control programmes (determining which drugs have been less effective). This project currently focuses on the establishment of laboratories in two locations (one in Ghana and one in The Gambia), and ultimately hopes to demonstrate how other surveillance systems can be deployed in other locations across the continent (CS_17). MalariaGEN has supported the establishment of the laboratories by procuring equipment, developing lab protocols, and giving advanced training workshops on the use of laboratory equipment and training on the bioinformatics and analytical processing of sequence data (CS_06, CS_17). A second project – GenRE-Mekong – is working to build local genetic surveillance in South East Asia,[88] where there is a large problem because malaria parasites are resistant to many of the available drugs, and therefore genomic information on the parasites is important for developing new drug treatments (CS_06). The genetic information gained through sequencing can then inform malaria control strategies, helping set policy on the therapies which can be used in future. Although both these projects are still at a relatively early stage, they demonstrate the shift from the work supported at the start of MalariaGEN, which primarily involved lab research, to the development of local sequencing capacity and operational surveillance, which is likely to drive national policy in LMICs (CS_06).

Developing networks

MalariaGEN has enabled researchers working internationally on malaria to develop networks and partnerships. From the initial funding, MalariaGEN established the Genomic Epidemiology of Malaria (GEM) conference, which runs every two years from Wellcome Genome Campus. This conference gives an opportunity for researchers to share knowledge and expertise on malaria, but also to develop partnerships (CS_06). Bursaries are available to improve participation from LMICs, and there are plans to host the conference internationally in the future (CS_06). From this conference, the Plasmodium Diversity Network Africa (PDNA) was established (CS_06).[89] This is an African-led research network which connects researchers interested in using genome sequence data to map the diversity of malaria parasites in Africa, and ultimately use this data to inform policy around malaria control. This initiative connects African researchers at all career stages, providing a support network, and enabling the sharing of knowledge around the analysis and interpretation of sequence data. PDNA researchers published a *Science* paper on sub-populations within the malaria parasite *Plasmodium falciparum* populations in sub-Saharan Africa (Amambua-Ngwa et al.

[87] As of 28 July 2020: https://fundingawards.nihr.ac.uk/award/17/63/91

[88] As of 28 July 2020: https://www.malariagen.net/projects/spotmalaria

[89] As of 28 July 2020: https://www.cggh.org/collaborations/plasmodium-diversity-network-africa

2019), and the network is considered to be important to develop these data skills among African researchers.[90]

Training

Between 2006 and 2010, MalariaGEN ran a data bursary scheme to support researchers in malaria-endemic countries to develop capacity in genetic data analysis. Participants were known as MalariaGEN data fellows, and they provided support and data management for specific studies within MalariaGEN. Many of these individuals have now gone on to lead their own research projects (CS_17). MalariaGEN has also been involved in several informal activities around training LMIC researchers and sharing facilities and expertise, which has enabled those researchers to gain access to the large datasets provided by Sanger, enabling them to gain experience in analysing and interpreting sequence data (CS_06).

4.4.4. The role of Sanger Institute

Sanger has played a key role in MalariaGEN from the start (CS_17). According to an interviewee, its sequencing capacity has been crucial to the processing and analysis of samples for the projects; samples sent from LMICs for sequencing have been examined in labs at Sanger, as well as at its core sequencing facilities (CS_17). The same interviewee noted that researchers at Sanger Institute have been critical to enable the training and expertise needed to guide these efforts and develop the protocols and assays required. Sanger's ethos for transparency and openness regarding data has meshed very well with the work done through MalariaGEN, and MalariaGEN has been able to build on this philosophy (CS_17).

4.4.5. Lessons learned from the case study

This case study illustrates the impact that MalariaGEN has had in LMICs. Sanger's contribution to building this network, employing sequencing technologies, training, developing local sequencing capacity and producing an equitable framework for data sharing have all had an impact on malaria research within LMICs, as well as policies which can be used in wider research areas. MalariaGEN has enabled a community of researchers to come together, and by sharing data, resources and knowledge they have been able to gain insight into malaria. Although still at an early stage, the work to develop local sequencing capacity is already having an impact on policy within LMICs, and it is likely that national surveillance systems will demonstrate how similar systems in other countries can be developed in the future. In addition, the framework for data sharing provides a set of principles, which can be adopted by researchers across disciplines to enable these large-scale collaborative projects, and give them insights into the genetic basis of many diseases.

4.5. ICGC

4.5.1. Introduction

The ICGC is an international collaboration established to coordinate research on cancer genomics around the world. Established in 2007, ICGC comprises a global network of scientists, scientific groups and research funders, and a centralised scientific organisation that serves to coordinate research projects,

[90] As of 28 July 2020: https://www.malariagen.net/blog/60-secs-with%E2%80%A6-dr-alfred-amambua-ngwa

maximise efficiency and promote the exchange of knowledge to address common challenges (ICGC 2008). A primary goal of the ICGC has been to bring together all data on the genomic characteristics of cancers and to make this data freely available and accessible to the global research community (ICGC 2008).

This case study examines the role of Sanger Institute in the formation, development and outputs of the ICGC. It highlights the key role that Sanger has played, alongside other organisations, in fostering and leading international collaborations, promoting open data access and knowledge sharing, as well as its significant contribution to cancer genomics research.

4.5.2. Background and context

Following the development of the first reference human genome under the Human Genome Project, the prospect of using genomic sequencing to better understand cancer biology had led to the initiation of a number of projects focused on the genomic characterisation of tumours (Hudson et al. 2010). The most notable of these were the Cancer Genome Project, established by researchers at Sanger in 2000, and TCGA, a US-based collaboration launched in 2005, backed by the National Cancer Institute and the NHGRI (Hudson et al. 2010; Jennings and Hudson 2013). Established in 2007, the ICGC was the product of a shared conviction among leading researchers and funders that an international forum would help to ensure the greater harmonisation and coordination of work within this expanding field (Hudson et al. 2010; Jennings and Hudson 2013).

The founding of the ICGC took place at a meeting held on 1–2 October 2007 in Toronto, Canada, convened by six organisations – the European Commission, Genome Canada, the National Cancer Institute, the NHGRI, the Ontario Institute for Cancer Research and Wellcome.[91] At the Toronto meeting, participants agreed that the principal aim of the ICGC would be to assist the development of a comprehensive reference database, freely available to the global research community, containing the genomic abnormalities in tumours of 50 major cancer types and subtypes (ICGC 2008).[92] At the same time, the ICGC would also serve to streamline research, reduce duplication, improve the standardisation of research practices and facilitate comparable studies across different types of cancer (Hudson et al. 2010; ICGC 2008).

Participants at the Toronto meeting agreed on an overarching governance structure for the ICGC. In its broadest sense, this structure consisted of two component parts: 'cancer genome projects', each comprising a funder (or funders) and a scientific research group, with each project focusing on the characterisation of one specific cancer type or subtype; and a series of 'central' ICGC bodies, including committees and working groups (ICGC 2008). The role of the central ICGC bodies was to coordinate and support the research undertaken by individual cancer genome projects, including addressing issues such as sample collection, consent, ethics, technologies, operating procedures and quality standards (ICGC 2008).

[91] The meeting attracted 122 participants from 22 countries, including world leaders in cancer genomics and related fields (Jennings and Hudson 2013).

[92] The initial target of the ICGC was to collect data to define the genomes of 25,000 primary untreated cancers (CS_01).

Sanger Institute played a key role in the formation of the ICGC. Mike Stratton, then Director of Sanger's own Cancer Genome Project, was a 'founding member' of the Consortium (CS_01). At the Toronto meeting, Stratton served on the interim ICGC Executive Committee and the Scientific Planning Committee and led an early working group on genome analysis (ICGC 2008).[93] According to one interviewee, a key early contribution of Sanger was to define quality metrics, analysis standards and data sharing principles to which all ICGC cancer genome projects should adhere (CS_16).

Another key contributor to the establishment of the ICGC was the Ontario Institute for Cancer Research (OICR) (CS_01). Led by its President and Scientific Director Tom Hudson, the OICR agreed to provide Secretariat services to the Consortium, with Hudson serving as the ICGC's Executive Director (ICGC 2008). Reflecting the work that both institutes had already initiated in the field of cancer genome analysis, the NHGRI, led by Francis Collins, and the Broad Institute, led by Eric Lander, were actively involved in the founding of the ICGC (CS_01). The NHGRI-backed TCGA, which had promoted collaboration in cancer genomics research within a US context, provided an important blueprint for the ICGC (Comp_02).

4.5.3. Research process

In the years after 2007, the research process under the ICGC followed broadly the blueprint set out by the Toronto meeting (CS_01). Individual cancer genome projects, most of which were organised by country, undertook research into the genomic characterisation of specific cancer types (Hudson et al. 2010; ICGC 2008). Meanwhile, the central ICGC bodies, including ICGC working groups, provided scientific coordination and exchange of learning, and facilitated the development of standards and guidelines for cancer genome projects (Hudson et al. 2010; ICGC 2008). From April 2010, the ICGC launched a centralised data portal providing a single location to which projects could send synthesised data in a universal format, thereby making it accessible to the broader research community (Behrman and Mazerik 2019; Hudson et al. 2010).

Within this research process, Sanger Institute has played various roles. One has been as a scientific research group responsible for the design and implementation of cancer genome projects. Since 2007, Sanger Institute has led or contributed to ICGC cancer genome projects on breast cancer, osteosarcoma and chronic myeloid leukaemia (CML) and prostate cancer (CS_16). According to one interviewee, Sanger researchers have contributed 'to a high degree' to the scientific coordination work of the central ICGC, including leading ICGC working groups on technologies, common samples and genome aberrations (CS_01). Through participation in ICGC working groups, Sanger also played an important role in establishing best practices for quality assurance of computational analysis (CS_16). Sanger has also led new strategic projects launched by the ICGC, for example a new ICGC strategic project called the Pan-Cancer Analysis of Whole Genomes (PCAWG) in 2013. Moving beyond the initial focus on sequencing cancer genomes, the aim of PCAWG was to apply whole-genome sequencing to identify common patterns of mutation in over 2,600 cancer genomes. The impetus for the PCAWG project came from a proposal prepared by Sanger Institute researchers (CS_16). Since its inception, Sanger has provided strategic leadership for this large-scale international collaboration, with Peter Campbell, Director of the Cancer and

[93] From 2009, he also led the ICGC working group on technologies (Hudson et al. 2010; ICGC 2009).

Somatic Mutations Programme, serving as chair for the PCAWG steering committee (Campbell et al. 2020).

There are some ways in which Sanger Institute's contribution to the research process faced shortcomings. Sanger was in a position to share its technological capacity with others who did not have sequencing capacity. However, some researchers avoided collaborating because of concerns about Sanger's level of control over such a project and the complexity of contractual arrangements that would be required (CS_02).

Sanger's contributions to the ICGC research process ran alongside contributions from other organisations. The Broad Institute and the NHGRI also contributed to the ICGC through cancer genome projects and participation in the broader scientific coordination work of ICGC committees and working groups (CS_01). The contribution of the Broad Institute and the NHGRI organisations to the cancer genome sequencing carried out by the ICGC was distinct in that it focused on exome sequencing of a larger number of tumour types, rather than more detailed genome sequencing (CS_16). Another difference of the American organisations was that sequencing was distributed across multiple different centres, with less clearly defined leadership roles for specific tumour types (CS_16). Chinese researchers actively contributed to the ICGC research process. Following the formation of the ICGC, the Chinese Cancer Genome Consortium (CCGC) was established, and CCGC member organisations led 15 cancer genome projects and participated in a large number of ICGC committee and working groups. In 2014, CCGC hosted the ninth ICGC scientific workshop in Beijing (Hu et al. 2015).

4.5.4. Contributions to knowledge

Through 86 cancer genome projects spanning 22 jurisdictions, the ICGC has collected data on 24,000 cancer genomes (just short of its initial target of 25,000), covering 50 tumour types (CS_01). A second major contribution of the ICGC has been in the area of whole-genome sequencing (CS_16). Through the PCAWG project, the ICGC has overseen the integrative analysis for 2,658 cancer whole genomes across 38 tumour types (Campbell et al. 2020). In January 2020, the latest findings of the PCAWG were published in a series of 23 papers published in *Nature* and its affiliated journals (Campbell et al. 2020). According to one interviewee, the Sanger Institute has made an important contribution to the sequencing and publication outputs of the ICGC (CS_16).

4.5.5. Contributions to future research

Under the ICGC, genomic data collected has been made available to the public through the Consortium's centralised data portal. The data portal is the 'first project to successfully federate large amounts of cancer genomics data and rich annotation data in a single access point' (Zhang et al. 2011).[94] Two interviewees highlighted that the portal now provides an 'invaluable resource' for researchers working in the field of cancer genomics, with many scientific papers citing its data (CS_01, CS_02).

Another significant contribution of the ICGC has been its role in developing and advancing methods for conducting research. Through ICGC working groups covering areas such as technologies, verification and

[94] While certain categories of data are restricted in order to protect patient privacy, the central ICGC Data Compliance Office can grant access to this data, where justified for research purposes. At present, over 300 research groups have such access (CS_01).

validation, ethics, epigenomic profiling, exome transcriptomics and bioinformatics analysis, scientists have shared experiences and learnings across different cancer genome projects and identified best practices for the conduct of research (CS_01). In many cases, working groups have also developed policies and benchmarks to ensure the standardisation of research practices (and data outputs) across ICGC projects (CS_02). The impact of this broader method development is reflected in the fact that some countries, such as India, undertook their first genome research projects under the auspices of the ICGC (CS_01). As noted above, Sanger has played a critical role in this method development work. While helping to develop best practices for quality assurance of computational analysis, Sanger has also shared its expertise in variance detection methodologies and mutational signature extraction with other researchers through the ICGC (Hu et al. 2015).

4.5.6. Contributions to health and healthcare

According to one interviewee, it is possible to link research performed by the ICGC to a number of clinical impacts (CS_16). Two key examples of this come from the field of mutational signatures – a field directly connected to research carried out by Sanger under the ICGC. Mutational signatures analyses patterns of mutation within cancer genomes and uses this to identify particular mutational processes that may increase the risk of certain types of cancer and identify what is the cause of those mutations. The global application of this approach to cancers has already demonstrated distinctive patterns of mutation for certain tumour types, linking this to unique patterns of exposure within those areas (CS_16). The most important example of this has been the use of mutational signatures to discover that certain herbal medicines in East Asia possess carcinogenic qualities (CS_16). More specifically, mutations that were identified as part of the ICGC studies in chronic blood cancers are now used as part of the routine screening strategies for suspected blood cancer patients within the NHS (CS_16).

Notwithstanding the above contributions, the focus of ICGC's first two projects (the 25k Initiative and PCAWG) was not explicitly on clinical translation (CS_01). However, the third strategic project, Accelerate Research in Genomic Oncology (ARGO) reflects a more concerted shift towards the clinical application of cancer genomics research (CS_01, CS_02, and CS_16). The aim of ARGO is to analyse samples from cancer patients with high-quality clinical annotations and data in order to better address questions relating to the treatment of specific cancer types. The project aims to collect data on 200,000 cancer patients around the world.[95]

The ARGO project is an ambitious next step for the ICGC. At the same time, the extent to which the project will work effectively is not necessarily clear (CS_02). Compared to the early work of ICGC projects, for example, coordinating research using clinically annotated data presents a series of new challenges, including the question of who will pay for that work and whether it will be possible to conduct it in a unified way across different countries, healthcare systems and cancer types (CS_02). There are also more fundamental questions about the extent to which an international consortium like the ICGC is the most effective way to go about translating cancer genome research into clinical practice, or whether this is better done at the national level (CS_02). One argument for a consortium-based approach is that if collaboration

[95] As of 28 July 2020: https://www.icgc-argo.org/page/64/about-icgc-argo

can be made to work effectively, it will produce more comparable studies, the results of which can be contrasted and linked in ways that would not be possible if done in isolation.

4.5.7. The role of Sanger Institute

While establishment of ICGC reflected broader developments within the field of cancer genomics research, and drew on the energies of a wide range of funding and research organisations, Sanger Institute played a major role. Already a world leader within the field cancer genome sequencing, Sanger was at the forefront of the international meetings that led to the formation of the ICGC, and provided leadership, in implementing its own cancer genome projects and through stewardship on ICGC committees and working groups. Sanger helped to ensure that adherence to quality metrics, analysis standards and data sharing principles sat at the core of the ICGC project.

As the ICGC has evolved, Sanger has, like several other organisations, acted both as a scientific research group and a key contributor to the Consortium's scientific and methodological development work. More uniquely, Sanger has also driven and provided leadership for new strategic initiatives launched under the ICGC, most notably the PCAWG project. Sanger has made an important contribution to the various outputs of the ICGC, from publications, to data, to methodological development and capacity building. Sanger has also been at the forefront of attempts to apply findings from ICGC research to clinical practice.

4.5.8. Lessons learned from the case study

As a case study, ICGC illustrates Sanger's expertise within the field of cancer genomics, but also its various strengths in building international research collaborations, providing strategic leadership, promoting open access to genomic data, and promoting quality and capacity building. Working in partnership with others, Sanger Institute used its own position at the forefront of cancer genome sequencing to foster an international research collaboration that helped to build broader research capacity, ensure the quality and comparability of research, and bring together the resulting data in a single access point, thereby making it freely accessible for the global research community. The result of this effort has been to not only increase the amount of cancer genome sequencing done, but also lay the foundations for a stronger global research infrastructure in this field. The case study has provided other insights, including the perceived challenges of collaborating with Sanger held by some smaller partners. It has also highlighted questions surrounding the extent to which international collaborations such as the ICGC, notwithstanding their varied contributions to scientific research, provide the best model for translating research into clinical practice.

5. Reflections

Bringing together evidence across sources, we can identify a range of strengths, weaknesses, opportunities and threats in relation to the Sanger Institute, the Wellcome Genome Campus and other comparator organisations working in the field of genetics and genomics research.

5.1. Reflections on the role of the Sanger Institute and the Genome Campus in the field and key characteristics

5.1.1. Working at scale

A key strength of Sanger and similar organisations such as the Broad Institute is the capacity to undertake genetic and genomic science at scale. However, advances in technology have brought significant widening in access to sequencing technologies so the ability to deliver high-speed, high-throughput sequencing is no longer a distinctive contribution to the field. The Sanger Institute, Wellcome Genome Campus and the Broad Institute have all played a key role in this broadening of access through the scientific and technological advances achieved, and the training, capacity building and strengthening and collaboration activities conducted. However, as the landscape changes, this presents a challenge for these institutions in finding a new role in the research landscape that fits with their expertise, strengths and ethos.

5.1.2. Collaboration and convening

Collaboration has been key to the evolution of the field and all the comparator organisations, along with the Sanger, have been involved in different internationally collaborative efforts, which have shaped the field of genomics research. The Sanger Institute exemplifies this collaborative ethos, which is a key strength of the organisation. This collaborative ethos extends within and across the physical research campus, and more widely to include external partners nationally and internationally. As illustrated in the case studies, much of Sanger's work is collaborative, and the organisation has significant convening power, using their reputation, scale and attitude of openness to facilitate collaborations at different levels. This convening power is also evident for other key players in the field such as Broad and NHGRI, which have played a key role in large-scale international collaborations and also partner in other ways. For example, the Broad is closely connected to local institutions such as MIT and Harvard, and NHGRI as a funder as well as a research institute has broad links across the field within the US and more widely. As well as contributing to large-scale international collaborations with peer organisations, such as the ICGC, Sanger also acts to bring together and support wider networks of smaller actors, taking on a leadership role in the field, such as through the DDD and MalariaGEN programmes. This collaborative approach can serve not just as a vehicle

for large-scale research advancement, but also as a vehicle for the broadening of participation and capacity development across the research landscape.

5.1.3. Role of the physical campus

Another strength of Sanger is the wider Wellcome Genome Campus, enabling a 'joined-up approach' to genetics and genomics research, including training, capacity building and public engagement alongside high-quality research on a single campus. This has been a key feature of how Sanger and other comparable organisations with a similar physical campus setting such as Janelia contribute to the field of genetics and genomics, as it allows for training and capacity building alongside ongoing research activities. The potential to build on and expand the advantages presented by physical campus settings are also an opportunity, for example by further strengthening relationships across research campuses and building closer links with on-site organisations, such as GEL and EMBL-EBI in the case of the Wellcome Genome Campus and Sanger. There may also be scope to scale up and strengthen the public engagement and training offer provided on the Wellcome Genome Campus and other physical research campuses in the field of genetics and genomics. Training in particular may be an opportunity, taking into account noted personnel challenges in the field, which can not only build research capacity but also promote Sanger's ethos of openness and collaboration to up and coming researchers in the field.

5.1.4. Open research

Another key strength of Sanger and other organisations at the Wellcome Genome Campus is a commitment to data sharing and open access to genomic data, which is highlighted across all the case studies conducted. This ethos of openness has contributed to many of the impacts that Sanger has been able to achieve – which may often be indirect through others building on the unique resources and datasets made publicly available. This interacts with the collaborative nature of Sanger's work as noted above, with sharing of data facilitating open and effective collaboration. However, it is also worth noting that this ethos of openness can present challenges in collaboration, too, in cases where other parties may have concerns about sharing of their information and ensuring appropriate credit for their contributions. Most interviewees thought that Sanger manages this effectively in most cases, with 'smaller' collaborators receiving appropriate acknowledgement for their contributions. Openness is also a key feature of other institutes, notably Janelia, which has a key emphasis on openness and sharing of data and particularly research tools. There is some suggestion that this openness, combined with a clear focus on fundamental research has contributed to a lower level of commercialisation activity for the Sanger Institute compared with some other organisations such as the Broad Institute. However, the case of Open Targets illustrates how that approach can be used to facilitate collaboration with the private sector in a more open and cross-collaborative manner, striking a balance between openness and pre-competitive access to data. Typically, Sanger has oriented its translational work in other directions, by making data publicly available to enable use, update and translation by all rather than necessarily via a commercial route, in alignment with Sanger's values and ways of working.

5.1.5. Long-term perspective

Linked to the scale of Sanger and the Wellcome Genome Campus and the funding model in place, another strength of Sanger is the ability to be strategic and long term in research focus. This longer-term funding

model is to a large extent similar to the comparator organisations, which have varying arrangements but typically are offered some degree of longer-term financial security as part of the institute status. This in turn enables them to take on long-term, open-ended and ambitious projects, with core funding acting as a key facilitator. This longer-term perspective has enabled Sanger to take on leadership roles in complex international consortia and collaborations effectively and contribute as leaders, alongside peers, in key developments in the field.

5.2. Looking forward: the future for genetics and genomics research and the role of the Sanger Institute and Wellcome Genome Campus

The Sanger and the Genome campus, along with other comparator organisations, have made wide ranging contributions to the field of genetics and genomics research as illustrated in the examples throughout this report, from training and capacity building to advancing science and improving health.

A challenge for the Sanger Institute and other organisations with similar scale and reputation will be to continue to maintain its role as a world leader in a changing research landscape. When Sanger was founded, high-throughput sequencing was a ground-breaking technology available to few researchers. As technology changes and sequencing becomes more widely available, key players in genomics research fields need to consider where they can continue to be ground-breaking and make novel contributions to the field. In the case of the Sanger Institute and the Genome Campus, some of their unique features include the scale and deep knowledge of genetics and genomics, paired with a reputation for openness, collaboration and the ability to serve as a 'focal point' to shape the direction of travel of the field, alongside peer institutes internationally. We start to see emerging evidence of Sanger Institute forging this new path through the Tree of Life programme, for example – a research programme that emphasises and capitalises on Sanger's strengths such as openness, data sharing and long-term, large-scale ambitious work. Other organisations have taken different directions – for example the Broad Institute and NHGRI, which are focusing on other aspects such as gene editing, CRISPR, and other human genetic and translational oriented work, emphasising their strengths and the context in which they operate. This diversification and differentiation between institutes is likely to be beneficial to the field enabling new avenues to open up and building resilience.

Another key observation is that the contributions that these large institutes make go beyond what is typically measured in terms of performance in research. As funders and organisation consider their future direction of travel in the field of genetics and genomics research, it will be crucial not just to consider their publications and direct contributions to the research field, but also to take into account the wider role they play as conveners and focal points for smaller actors in the space, and in particular their role as a source of data, tools and capacity building.

This illustrates another area that merits further consideration, which is to better understand and hence better facilitate impact through research. In the case of the Sanger Institute, much of the benefit it creates to society and the economy will be secondary, through the use of information and data, as well as knowledge, capacity and resources, by others outside the campus. This situation will likely also apply to other large institutes in the field. This presents challenges in capturing and understanding the range and nature of those benefits and the use of the resources and achievements of research institutes by others. Though challenging,

a deeper ongoing understanding of these benefits could be crucial in understand how effective Sanger's (and equivalently other research institutes') strategic approach – and indeed underpinning structure and values – are in delivering their intended outcomes. This will be crucial as the field evolves and institutions attempt to make strategic decisions about the contributions and leadership they wish to offer in the new research landscape.

Bibliography

Achidi, E., T. Agbenyega, S. Allen et al. 2008. 'A Global Network for Investigating the Genomic Epidemiology of Malaria.' *Nature* 456 (7223): 732–37. https://doi.org/10.1038/nature07632.

Achidi, Eric, Tsiri Agbenyega, and Steve Allen. 2008. "A Global Network for Investigating the Genomic Epidemiology of Malaria." *Nature* 456 (7223): 732–37. https://doi.org/10.1038/nature07632.

Ahrens, Misha B, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. 2013. "Whole-Brain Functional Imaging at Cellular Resolution Using Light-Sheet Microscopy." *Nature Methods* 10 (5): 413–20. https://doi.org/10.1038/nmeth.2434.

Amambua-Ngwa, Alfred, Lucas Amenga-Etego, Edwin Kamau, Roberto Amato, Anita Ghansah, Lemu Golassa, Milijaona Randrianarivelojosia, et al. 2019. "Major Subpopulations of Plasmodium Falciparum in Sub-Saharan Africa." *Science* 365 (6455): 813–16. https://doi.org/10.1126/science.aav5427.

Atlas, Ronald M. 2012. "One Health: Its Origins and Future." In *One Health: The Human-Animal-Environment Interfaces in Emerging Infectious Diseases*, 1–13. Springer.

Band, Gavin, Quang Si Le, Geraldine M. Clarke, Katja Kivinen, Christina Hubbart, Anna E. Jeffreys, Kate Rowlands, et al. 2019. "Insights into Malaria Susceptibility Using Genome-Wide Data on 17,000 Individuals from Africa, Asia and Oceania." *Nature Communications* 10 (1): 5732. https://doi.org/10.1038/s41467-019-13480-z.

Behan, Fiona M., Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, et al. 2019. "Prioritization of Cancer Therapeutic Targets Using CRISPR–Cas9 Screens." *Nature* 568 (7753): 511–16. https://doi.org/10.1038/s41586-019-1103-9.

Behrman, Shannon, and Jessica Mazerik. 2019. "The International Cancer Genome Consortium: Exploring the Cancer Genome." National Institutes of Health - Office of Cancer Genomics.

Bergauer, Tobias, Thorsten Ruppert, Laurent Essioux, and Olivia Spleiss. 2016. "Drug Target Identification and Validation: Global Pharmaceutical Industry Experts on Challenges, Best Strategies, Innovative Precompetitive Collaboration Concepts, and Future Areas of Industry Precompetitive Research and Development." *Therapeutic Innovation & Regulatory Science* 50 (6): 769–76. https://doi.org/10.1177/2168479016651298.

Bowden, Rory, Robert W. Davies, Andreas Heger, Alistair T. Pagnamenta, Mariateresa de Cesare, Laura E. Oikkonen, Duncan Parkes, et al. 2019. "Sequencing of Human Genomes with Nanopore Technology." *Nature Communications* 10 (1): 1869. https://doi.org/10.1038/s41467-019-09637-5.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2018. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12. https://doi.org/10.1093/nar/gky1120.

Buxton, Martin, and Stephen Hanney. 1997. "Assessing Payback from Department of Health Research and Development: Second Report. Volume 1: The Main Report." *HERG Research Report*, no. 19.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Cano-Gamez, Eddie, Blagoje Soskic, Theodoros I. Roumeliotis, Ernest So, Deborah J. Smyth, Marta Baldrighi, David Willé, et al. 2020. "Single-Cell Transcriptomics Identifies an Effectorness Gradient Shaping the Response of CD4+ T Cells to Cytokines." *Nature Communications* 11 (1): 1801. https://doi.org/10.1038/s41467-020-15543-y.

Chhetri, Raghav K, Fernando Amat, Yinan Wan, Burkhard Höckendorf, William C Lemon, and Philipp J Keller. 2015. "Whole-Animal Functional and Developmental Imaging with Isotropic Spatial Resolution." *Nature Methods* 12 (12): 1171–78. https://doi.org/10.1038/nmeth.3632.

Cooke, Jennifer, and Bec Crew. 2019. "New Tools for New Treatments." *Nature Index*. https://www.natureindex.com/news-blog/new-tools-for-new-treatments.

Dana, Hod, Boaz Mohar, Yi Sun, Sujatha Narayan, Andrew Gordus, Jeremy P Hasseman, Getahun Tsegaye, et al. 2016. "Sensitive Red Protein Calcium Indicators for Imaging Neural Activity." *ELife* 5 (March): e12727. https://doi.org/10.7554/eLife.12727.

DeAngelis, Allison. 2019. "CRISPR Gene Editing Startup Beam Therapeutics Plans $100M IPO - Boston Business Journal." *Boston Business Journal*. https://www.bizjournals.com/boston/news/2019/09/30/crispr-gene-editing-startup-beam-therapeutics.html.

Denee, Tom. R., Arnold Sneekes, Pieter Stolk, Antoine Juliens, Jan A. M. Raaijmakers, Michel Goldman, Daan J. A. Crommelin, and Jorg W. Janssen. 2012. "Measuring the Value of Public–Private Partnerships in the Pharmaceutical Sciences." *Nature Reviews Drug Discovery* 11 (5): 419–419. https://doi.org/10.1038/nrd3078-c1.

Dunham, Ian. 2018. "Human Genes: Time to Follow the Roads Less Traveled?" *PLOS Biology* 16 (9): e3000034. https://doi.org/10.1371/journal.pbio.3000034.

Ellis, Rosalyn. 2015. "Recruitment Begins on Major Genetics Project." NIHR Oxford Biomedical Research Centre. August 26, 2015. https://oxfordbrc.nihr.ac.uk/recruitment-begins-on-major-genetics-project/.

Failli, Mario, Jussi Paananen, and Vittorio Fortino. 2019. "Prioritizing Target-Disease Associations with Novel Safety and Efficacy Scoring Methods." *Scientific Reports* 9 (1): 9852. https://doi.org/10.1038/s41598-019-46293-7.

Firth, Helen, and Caroline Wright. 2011. "The Deciphering Developmental Disorders (DDD) Study." *Developmental Medicine and Child Neurology* 53 (June): 702–3. https://doi.org/10.1111/j.1469-8749.2011.04032.x.

Flicek, Paul, Bronwen L. Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, et al. 2010. "Ensembl's 10th Year." *Nucleic Acids Research* 38 (suppl_1): D557–62. https://doi.org/10.1093/nar/gkp972.

Fosque, B. F., Y. Sun, H. Dana, C.-T. Yang, T. Ohyama, M. R. Tadross, R. Patel, et al. 2015. "Labeling of Active Neural Circuits in Vivo with Designed Calcium Integrators." *Science* 347 (6223): 755–60. https://doi.org/10.1126/science.1260922.

Gaspar, H.A., and G. Breen. 2017. "Pathways Analyses of Schizophrenia GWAS Focusing on Known and Novel Drug Targets." *BioRxiv*, January, 091264. https://doi.org/10.1101/091264.

Gibbs, E Paul J. 2014. "The Evolution of One Health: A Decade of Progress and Challenges for the Future." *Veterinary Record* 174 (4): 85–91.

Gibbs, Richard A., John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, et al. 2003. "The International HapMap Project." *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Halim, Shakera. 2018. "Open Targets -Transforming Drug Discovery." *Health Europa Quarterly*, 2018. http://edition.pagesuite-

professional.co.uk/html5/reader/production/default.aspx?pubname=&edid=73e202a8-1e25-4d2e-afc3-1cd95c26e5ae.

HHMI. 2003. "Janelia Farm Research Campus: REPORT ON PROGRAM DEVELOPMENT." https://www.janelia.org/sites/default/files/About%20Us/JFRC.pdf.

Hu, Xueda, Huanming Yang, Jie He, and Youyong Lu. 2015. "The Cancer Genomics and Global Cancer Genome Collaboration." *Science Bulletin* 60 (1): 65–70. https://doi.org/10.1007/s11434-014-0692-9.

Hudson (Chairperson), Thomas J., Warwick Anderson, Axel Aretz, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, M. K. Bhan, et al. 2010. "International Network of Cancer Genome Projects." *Nature* 464 (7291): 993–98. https://doi.org/10.1038/nature08987.

"International Cancer Genome Consoritum: 2nd Scientific Workshop, June 22-24, 2009." 2009. International Cancer Genome Consortium.

"International Cancer Genome Consortium: Goals, Structure, Policies & Guidelines." 2008. International Cancer Genome Consortium.

Jennings, Jennifer, and Thomas J Hudson. 2013. "Reflections on the Founding of the International Cancer Genome Consortium." *Clinical Chemistry* 59 (1): 18–21. https://doi.org/10.1373/clinchem.2012.184713.

Kannan, Lavanya, Marcel Ramos, Angela Re, Nehme El-Hachem, Zhaleh Safikhani, Deena M.A. Gendoo, Sean Davis, et al. 2015. "Public Data and Open Source Tools for Multi-Assay Genomic Investigation of Disease." *Briefings in Bioinformatics* 17 (4): 603–15. https://doi.org/10.1093/bib/bbv080.

Khaladkar, Mugdha, Gautier Koscielny, Samiul Hasan, Pankaj Agarwal, Ian Dunham, Deepak Rajpal, and Philippe Sanseau. 2017. "Uncovering Novel Repositioning Opportunities Using the Open Targets Platform." *Drug Discovery Today* 22 (12): 1800–1807. https://doi.org/10.1016/j.drudis.2017.09.007.

Kincaid, Ellie. 2018. "Broad Institute Spinout Aims To Bring Precision Medicine To Autoimmune Disease." *Forbes*. https://www.forbes.com/sites/elliekincaid/2018/05/15/broad-institute-spinout-aims-to-bring-precision-medicine-to-autoimmune-disease/#1e4e922b3d72.

Koscielny, Gautier, Peter An, Denise Carvalho-Silva, Jennifer A. Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, et al. 2017. "Open Targets: A Platform for Therapeutic Target Identification and Validation." *Nucleic Acids Research* 45 (D1): D985–94. https://doi.org/10.1093/nar/gkw1055.

Liu, Tsung-Li, Srigokul Upadhyayula, Daniel E. Milkie, Ved Singh, Kai Wang, Ian A. Swinburne, Kishore R. Mosaliganti, et al. 2018. "Observing the Cell in Its Native State: Imaging Subcellular Dynamics in Multicellular Organisms." *Science* 360 (6386): eaaq1392. https://doi.org/10.1126/science.aaq1392.

Mantere, Tuomo, Simone Kersten, and Alexander Hoischen. 2019. "Long-Read Sequencing Emerging in Medical Genetics." *Frontiers in Genetics* 10: 426.

McEntagart, Meriel, Kathleen A Williamson, Jacqueline K Rainger, Ann Wheeler, Anne Seawright, Elfride De Baere, Hannah Verdin, et al. 2016. "A Restricted Repertoire of De Novo Mutations in ITPR1 Cause Gillespie Syndrome with Evidence for Dominant-Negative Effect." *American Journal of Human Genetics* 98 (5): 981–92. https://doi.org/10.1016/j.ajhg.2016.03.018.

McRae, J., S. Clayton, and Fitzgerald et al. 2017. "Prevalence and Architecture of de Novo Mutations in Developmental Disorders." *Nature* 542 (7642): 433–38. https://doi.org/10.1038/nature21062.

Middleton, Anna, Katherine I Morley, Eugene Bragin, Helen V Firth, Matthew E Hurles, Caroline F Wright, Michael Parker, and on behalf of the DDD study. 2016. "Attitudes of Nearly 7000 Health Professionals, Genomic Researchers and Publics toward the Return of Incidental Results from Sequencing Research." *European Journal of Human Genetics* 24 (1): 21–29. https://doi.org/10.1038/ejhg.2015.58.

Nalley, Catlin. 2019. "Genome-Scale CRISPR Screening Finds New Cancer Drug Targets." *Oncology Times* 41 (10). https://journals.lww.com/oncology-times/Fulltext/2019/05200/Genome_Scale_CRISPR_Screening_Finds_New_Cancer.22.aspx.

Nelson, Matthew R, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, et al. 2015. "The Support of Human Genetic Evidence for Approved Drug Indications." *Nature Genetics* 47 (8): 856–60. https://doi.org/10.1038/ng.3314.

Niemi, Mari E. K., Hilary C. Martin, Daniel L. Rice, Giuseppe Gallone, Scott Gordon, Martin Kelemen, Kerrie McAloney, et al. 2018. "Common Genetic Variants Contribute to Risk of Rare Severe Neurodevelopmental Disorders." *Nature* 562 (7726): 268–71. https://doi.org/10.1038/s41586-018-0566-4.

Paananen, Jussi, and Vittorio Fortino. 2019. "An Omics Perspective on Drug Target Discovery Platforms." *Briefings in Bioinformatics*, no. bbz122 (November). https://doi.org/10.1093/bib/bbz122.

Pairwise. 2019. "Pairwise Licenses CRISPR Technologies from Massachusetts General Hospital (MGH) and Broad Institute." Business Wire. 2019. https://www.businesswire.com/news/home/20190318005588/en/Pairwise-Licenses-CRISPR-Technologies-Massachusetts-General-Hospital.

Perkel, Jeffrey M. 2019. "The Microscope Makers Putting Ever-Larger Biological Samples under the Spotlight." *Nature* 575: 715–17.

Picart-Armada, Sergio, Steven J. Barrett, David R. Willé, Alexandre Perera-Lluna, Alex Gutteridge, and Benoit H. Dessailly. 2019. "Benchmarking Network Propagation Methods for Disease Gene Identification." Edited by Bjoern Peters. *PLOS Computational Biology* 15 (9): e1007276. https://doi.org/10.1371/journal.pcbi.1007276.

Priego, Laia Pujol, and Jonathan Wareham. 2018. "Open Science Monitor Case Study." European Commission.

Rubin, Gerald M, and Erin K O'Shea. 2019. "Looking Back and Looking Forward at Janelia." *ELife* 8 (February): e44826. https://doi.org/10.7554/eLife.44826.

Savage, Neil. 2016. "Does Proximity Play a Role in Science Collaborations?" *Nature Index*. https://www.natureindex.com/news-blog/does-proximity-play-a-role-in-science-collaborations.

Sheffi, Jonathan. 2018. "In Our Genes: How Google Cloud Helps the Broad Institute Slash the Cost of Research." Google. February 12, 2018. https://blog.google/products/google-cloud/our-genes-how-google-cloud-helps-broad-institute-slash-cost-research/.

Sherkow, Jacob S. 2016. "CRISPR: Pursuit of Profit Poisons Collaboration." *Nature News* 532 (7598): 172. https://doi.org/10.1038/532172a.

Short, Patrick J., Jeremy F. McRae, Giuseppe Gallone, Alejandro Sifrim, Hyejung Won, Daniel H. Geschwind, Caroline F. Wright, et al. 2018. "De Novo Mutations in Regulatory Elements in Neurodevelopmental Disorders." *Nature* 555 (7698): 611–16. https://doi.org/10.1038/nature25983.

Soskic, Blagoje, Eddie Cano-Gamez, Deborah J. Smyth, Wendy C. Rowan, Nikolina Nakic, Jorge Esparza-Gordillo, Lara Bossini-Castillo, et al. 2019. "Chromatin Activity at GWAS Loci Identifies T Cell States Driving Complex Immune Diseases." *Nature Genetics* 51 (10): 1486–93. https://doi.org/10.1038/s41588-019-0493-9.

Sun, Yi, Aljoscha Nern, Romain Franconville, Hod Dana, Eric R Schreiter, Loren L Looger, Karel Svoboda, Douglas S Kim, Ann M Hermundstad, and Vivek Jayaraman. 2017. "Neural Signatures of Dynamic Stimulus Selection in Drosophila." *Nature Neuroscience* 20 (8): 1104–13. https://doi.org/10.1038/nn.4581.

Taylor, Jenny C, Hilary C Martin, Stefano Lise, John Broxholme, Jean-Baptiste Cazier, Andy Rimmer, Alexander Kanapin, Gerton Lunter, Simon Fiddy, and Chris Allan. 2015. "Factors Influencing Success of Clinical Genome Sequencing across a Broad Spectrum of Disorders." *Nature Genetics* 47 (7): 717–26.

Terry, Mark. 2015. "Google Partners with Broad Institute of MIT to Corner Genomic Data Analysis Market." BioSpace. 2015. https://www.biospace.com/article/google-partners-with-broad-institute-of-mit-to-corner-genomic-data-analysis-market-/.

Tessema, Sofonias K., Jaishree Raman, Craig W. Duffy, Deus S. Ishengoma, Alfred Amambua-Ngwa, and Bryan Greenhouse. 2019. "Applying Next-Generation Sequencing to Track Falciparum Malaria in Sub-Saharan Africa." *Malaria Journal* 18 (1): 268. https://doi.org/10.1186/s12936-019-2880-1.

Tomer, Raju, Khaled Khairy, Fernando Amat, and Philipp J Keller. 2012. "Quantitative High-Speed Imaging of Entire Developing Embryos with Simultaneous Multiview Light-Sheet Microscopy." *Nature Methods* 9 (7): 755–63. https://doi.org/10.1038/nmeth.2062.

Vrueh, Remco L. A. de, and Daan J. A. Crommelin. 2017. "Reflections on the Future of Pharmaceutical Public-Private Partnerships: From Input to Impact." *Pharmaceutical Research* 34 (10): 1985–99. https://doi.org/10.1007/s11095-017-2192-5.

Waltman, Ludo, Nees Jan van Eck, and Ed C.M. Noyons. 2010. "A Unified Approach to Mapping and Clustering of Bibliometric Networks." *Journal of Informetrics* 4 (4): 629–35. https://doi.org/10.1016/j.joi.2010.07.002.

Wang, Jun, Lei Kong, Ge Gao, and Jingchu Luo. 2013. "A Brief Introduction to Web-Based Genome Browsers." *Briefings in Bioinformatics* 14 (2): 131–43. https://doi.org/10.1093/bib/bbs029.

Wellcome. 2019a. "Cancer, Ageing, and Somatic Mutaion Quinquennial Review 2021-26." Wellcome Sanger Institute.

———. 2019b. "Cellular Genetics Quinquennial Review 2021-26." Wellcome Sanger Institute.

———. 2019c. "Connecting Science Quinquennial Review 2021-2026." Wellcome Sanger Institute.

———. 2019d. "Graduate Programme Quinquennial Review, 2021-2026." Wellcome Sanger Institute.

———. 2019e. "GRL Ecosystem Quinquennial Review 2021-26." Wellcome Sanger Institute.

———. 2019f. "GRL Strategic Overview Quinquinnial Review 2021-26." Wellcome Sanger Institute.

———. 2019g. "Human Genetics Quinquennial Review 2021-26." Wellcome Sanger Institute.

———. 2019h. "Parasites and Microbes Programme Quinquennial Review 2021-26." Wellcome Sanger Institute.

———. 2019i. "Tree of Life Programme Quinquennial Review 2021-26." Wellcome Sanger Institute.

Wellcome Sanger Institute. 2020. "Tree of Life Programme: Quinquennial Review 2021-26." Internal.

Wright, C. F., E. Prigmore, D. Rajan, J. Handsaker, J. McRae, J. Kaplanis, T. W. Fitzgerald, D. R. FitzPatrick, H. V. Firth, and M. E. Hurles. 2019. "Clinically-Relevant Postzygotic Mosaicism in Parents and Children with Developmental Disorders in Trio Exome Sequencing Data." *Nature Communications* 10 (1): 2985. https://doi.org/10.1038/s41467-019-11059-2.

Wright, Caroline F., Jeremy F. McRae, Stephen Clayton, Giuseppe Gallone, Stuart Aitken, Tomas W. FitzGerald, Philip Jones, et al. 2018. "Making New Genetic Diagnoses with Old Data: Iterative Reanalysis and Reporting from Genome-Wide Data in 1,133 Families with Developmental Disorders." *Genetics in Medicine* 20 (10): 1216–23. https://doi.org/10.1038/gim.2017.246.

Xu, C. Shan, Michal Januszewski, Zhiyuan Lu, Shin-ya Takemura, Kenneth J. Hayworth, Gary Huang, Kazunori Shinomiya, et al. 2020. "A Connectome of the Adult Drosophila Central Brain." *BioRxiv*, January, 2020.01.21.911859. https://doi.org/10.1101/2020.01.21.911859.

Yang, Yadong, Xunong Dong, Bingbing Xie, Nan Ding, Juan Chen, Yongjun Li, Qian Zhang, Hongzhu Qu, and Xiangdong Fang. 2015. "Databases and Web Tools for Cancer Genomics Study." *Genomics, Proteomics & Bioinformatics* 13 (1): 46–50. https://doi.org/10.1016/j.gpb.2015.01.005.

Yozwiak, Nathan L., Stephen F. Schaffner, and Pardis C. Sabeti. 2015. "Data Sharing: Make Outbreak Research Open Access." *Nature News* 518 (7540): 477. https://doi.org/10.1038/518477a.

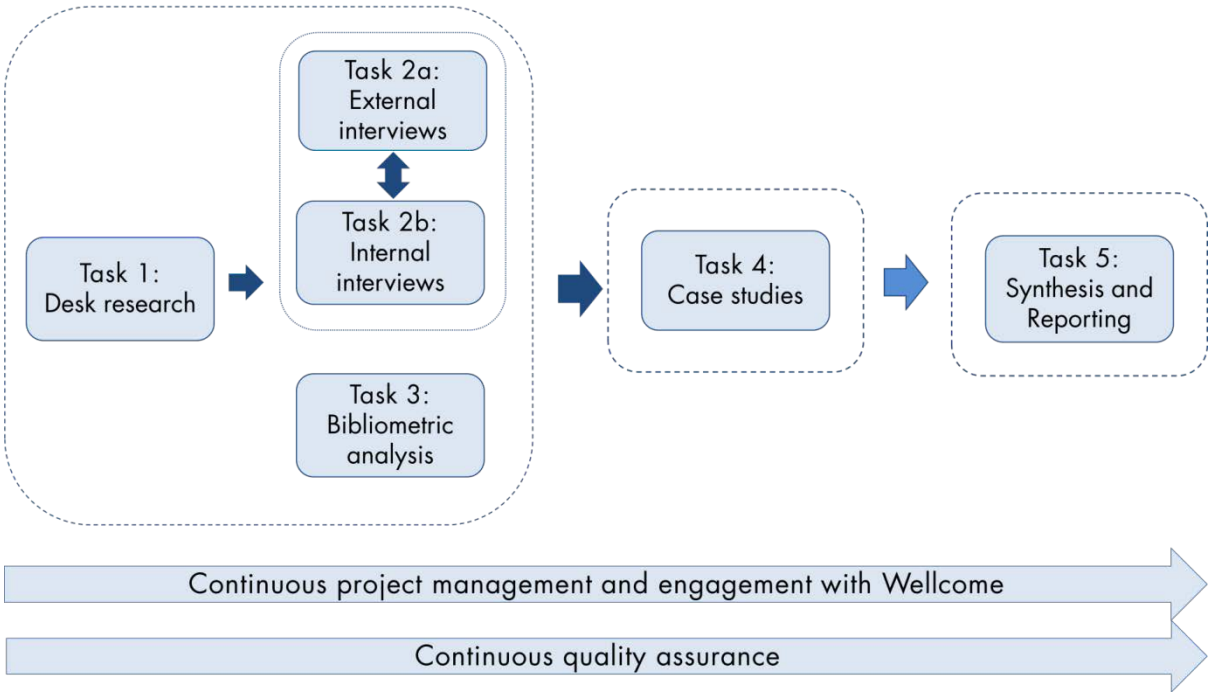
Zhang, Junjun, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong Liang, et al. 2011. "International Cancer Genome Consortium Data Portal—a One-Stop Shop for Cancer Genomics Data." *Database* 2011 (bar026). https://doi.org/10.1093/database/bar026.

Annex A. Methodology

We adopted a multi-method approach combining desk research, interviews, case studies and bibliometric analysis. This approach aimed to capture a range of perspectives on the role and contribution of the Sanger Institute and Wellcome Genome Campus within the field of genomics and the wider research landscape, setting this in the context of other comparators. In particular, we focused on four comparator organisations – the Broad Institute, the WHG, the NHGRI and Janelia Research Campus, chosen in consultation with Wellcome after initial desk research on similar organisations in the areas of genetics and genomics. The review covered the academic contributions of Sanger Institute and the wider translational, communication, networking and commercialisation activities associated with Sanger and Wellcome Genome Campus.

Figure 27 provides a visual overview of the primary tasks for this project, each of which is described in detail below. Two cross-cutting activities – project management and quality assurance – covered the entire duration of the project. Over the course of the project, we actively engaged with Wellcome through fortnightly project update calls during which we discussed progress, issues and the direction of the project.

Figure 27: Approach to undertaking the primary tasks for this project



Source: RAND Europe analysis

A.1. Task 1: Desk research

The purpose of this task was to use both the literature provided by Wellcome and other literature to develop a better understanding of the role that Sanger Institute and Wellcome Genome Campus play within the wider research landscape of genetics and genomics and biomedical research. This gave insight into their role in accomplishing Wellcome's overall goals more broadly.

Wellcome provided RAND Europe with documentation from Sanger's quinquennial review. In addition to this, we performed a rapid evidence assessment (REA), which allowed us to identify and compile existing evidence regarding the contribution of Sanger and the Genome Campus to the field of genetics and genomics, as well as the contribution of similar comparator organisations.

REAs are systematic reviews of the available literature on a topic. However, unlike a full systematic review, the scope and coverage of the review is restricted through search criteria, which are refined through the process to ensure that the review focuses on the most relevant literature to the scope of this work. The study team applied a structured approach to the REA based around three stages: literature searches, screening results and data extraction.

A.1.1. Stage 1: Literature searches

The study team conducted a structured review of academic literature, commentaries and grey literature. In our literature searches we limited our reviews to English-language articles published within the last ten years (2010–2019) in order to identify relevant up-to-date material. We focused on literature that evaluated or commented on Sanger Institute or the Genome Campus, and their role within the research landscape.

We identified relevant literature through two routes:

- **Literature search**: We conducted a search for publications from 2010 onwards. Table 10lists some of the search strings we used. This search was intended to result in a focused search, capturing any relevant information on measurements of the role of either Sanger or the Genome Campus or comparable research organisations in the field.
- **Snowballing** (the continuous, recursive process of gathering and searching for references within the bibliographies of shortlisted articles): We used snowballing from the reference lists of publications identified following screening.

Table 10: Indicative search strings used for the REA

Group	Search category	Search terms
1	Sanger or Genome Campus	'wellcome genome campus' OR 'wellcome sanger institute' OR 'sanger institute' OR 'genome campus'
2	Scientific field	'genom*' OR 'bioinformatic*' OR 'genet*'
3	Comparator organisations	'broad institute' OR 'wellcome centre for human genetics' OR 'janelia' OR 'national human genome research institute'

Source: RAND Europe analysis

A.1.2. Stage 2: Screening

This stage involved searches and selection, with titles and abstracts of identified studies screened for relevance against predefined inclusion and exclusion criteria. Table 11 lists some of the criteria used to determine if articles were relevant.

Table 11: Inclusion and exclusion criteria for the literature review on the wider contributions of Sanger Institute and Wellcome Genome Campus

Criterion	Include	Exclude
Topic relevance	Studies that comment on Sanger Institute or Wellcome Genome Campus, or comparators. Studies including strengths and weaknesses of Sanger ,the campus, and comparators. Studies that comment on how Sanger, and the campus, and comparators have contributed to the field of genetics and genomics.	Exclude research articles that focus on the pure science from either Sanger or the campus or comparator organisations alone.
Geographical location	No restriction	
Year of publication	2010 onwards	2009 or earlier, with exceptions if literature is highly relevant
Study characteristics	Academic publications; good quality grey literature; opinion and commentary pieces.	
Language	English	Other languages

Source: RAND Europe analysis

A.1.3. Stage 3: Data extraction and analysis

During this stage we extracted information from each included publication to facilitate cross-analysis against the key study questions and themes. Following a pilot and testing stage, researchers independently recorded data about each selected paper that met the inclusion criteria. Table 12 shows the extraction template in Excel we used to extract documents provided by Wellcome.

Table 12: Extraction template for document review

	Document 1	Document 2
Document Name.pdf		
Document Summary		
Future opportunities for Wellcome Sanger Institute and Wellcome Genome Campus		
How will Wellcome Sanger Institute and Wellcome Genome Campus make the most of these opportunities?		
Future challenges		
How could Wellcome Sanger Institute and Wellcome Genome Campus be better prepared to meet future opportunities and tackle challenges?		
Other points raised		
Role of Wellcome Sanger Institute and Wellcome Genome Campus within the genetics and genomics research landscape		
Relationship between Wellcome Sanger Institute, the broader Wellcome Genome Campus and collaborators		
Key strengths of Wellcome Sanger Institute and Wellcome Genome Campus		
Factors that enable Wellcome Sanger Institute and Wellcome Genome Campus to play this role		
Key contributions of Wellcome Sanger Institute and Wellcome Genome Campus; ways Sanger and the campus have impacted on others		
Key challenges facing Sanger and the campus		
Things that could be done to address these challenges		
Key contributions of specific Sanger or Genome Campus programme(s)		
Factors that enable Sanger or Genome Campus programme(s) to make these contributions		
Future opportunities for specific Sanger or Genome Campus programme(s)		
Key challenges facing specific Sanger or Genome Campus programme(s)		
Things that could be done to address these challenges		
Key observations on collaborating organisations: the Broad Institute		
Key observations on collaborating organisations: the WHG		
Key observations on collaborating organisations: Janelia Research Campus		
Key observations on collaborating organisations: the NHGRI		
Other collaborators		

A.2. Task 2: Interviews

We carried out a series of semi-structured telephone interviews of up to one hour with stakeholders external to Sanger and Wellcome Genome Campus and internal staff. This enabled us to develop a picture of the internal workings of these research organisations, what staff felt was working well or not, and also understand how experts regarded Sanger, the campus,and comparator organisations with regard to the wider research landscape. Semi-structured interviews allow all stakeholders to be asked a similar set of questions while enabling emergent issues to be explored. This gave us scope to explore key perceptions from the interviewee's point of view, while making it possible to compare responses between interviewees. Interview protocols are provided in Annex B, although interviewers were free to add questions not covered in the protocol, for example in response to an interviewee's previous responses.

When contacting potential interviewees, and at the start of the interview, the research team provided information regarding:

- Details of the study
- Consent to take part in the interview
- Attribution of information (information and quotes will not be attributed to individuals unless explicitly approved)
- Audio-recording of the interview (for accuracy and note-taking purposes, and only with the interviewee's consent).

We conducted 42 interviews with 45 individuals. Some interviewees were interviewed twice (focusing on different aspects of Sanger Institute, Genome Campus, other organisations and case studies in each interview), and some interviews had more than one interviewee. We held 13 interviews with internal employees of Sanger and Genome Campus, 6 interviews with external experts or research users, 16 interviews for the case studies, and 6 with representatives from comparator organisations. The characteristics of the interviews are given in Table 13.

Table 13: Interviewee characteristics

Anonymous identifier	Number of interviewees	Category	Description
Comp_01	1	External (comparator)	Broad Institute staff
Comp_02	1	External (comparator)	Broad Institute staff
Comp_03	1	External (comparator)	WHG staff
Comp_04	1	External (comparator)	WHG staff
Comp_05	1	External (comparator)	NHGRI staff
Comp_06	1	External (comparator)	Janelia staff
CS_01	1	Case study (ICGC)	ICGC – academic collaborator
CS_02	1	Case study (ICGC)	ICGC – academic collaborator
CS_03	1	Case study (Open Targets)	Open Targets – private-sector staff

CS_04	1	Case study (Open Targets)	Open Targets – Sanger staff
CS_05	1	Case study (Open Targets)	Open Targets – EMBL-EBI staff
CS_06	1	Case study (Malaria)	MalariaGEN – staff
CS_07	1	Case study (Tree of Life)	Tree of Life – Sanger staff
CS_08	1	Case study (Tree of Life)	Tree of Life – Academic researcher
CS_09	1	Case study (Tree of Life)	Tree of Life – Private sector
CS_10	1	Case study (DDD)	DDD – Sanger staff
CS_11	1	Case study (DDD)	DDD – Clinical collaborator
CS_12	1	Case study (DDD)	DDD – Clinical collaborator
CS_13	1	Case study (DDD)	DDD – GEL staff
CS_14	1	Case study (Tree of Life)	Tree of Life – Academic researcher
CS_15	1	Case study (Tree of Life)	Tree of Life – Public sector collaborator
CS_16	1	Case study (ICGC)	ICGC – Sanger staff
CS_17	2	Case study (Malaria)	Malaria – Sanger staff
Ext_01	1	External	Expert in genetics/genomics
Ext_02	1	External	Expert in genetics/genomics
Ext_03	1	External	Expert in genetics/genomics
Ext_04	1	External	Expert in genetics/genomics
Ext_05	1	External	Research user – private sector
Ext_06	1	External	Research user – private sector
Int_01	2	Internal	Sanger staff – Cancer, Ageing and Somatic Mutation
Int_02	2	Internal	Sanger staff – Cellular Genetics
Int_03	1	Internal	Sanger staff – Human Genetics
Int_04	1	Internal	Sanger staff – Parasites and Microbes
Int_05	2	Internal	Sanger staff – Tree of Life
Int_06	2	Internal	Sanger staff – PhD Programme
Int_07	2	Internal	Genome Campus – Connecting Science
Int_08	1	Internal	Genome Campus – Translation
Int_09	1	Internal	Sanger associated programme – Open Targets
Int_10	1	Internal	Sanger associated programme – Health Data Research UK
Int_11	2	Internal	Sanger staff – Overarching
Int_12	1	Internal	Sanger staff – Overarching
Int_13	1	Internal	Sanger staff – Overarching

A.2.1. Task 2a: External interviews

The objective of the external interviews was to collect information and views from staff at comparator organisations, as well as experts in the wider field of genomics and genetics, about their perception of the role of Sanger and the broader Genome Campus within the wider research landscape.

A.2.2. Task 2b: Internal interviews

The objective of the internal interviews was to collect information and views from staff at Sanger and the wider campus about their perception of the role of their work in the wider research landscape, any contributions they have observed from Sanger and Wellcome Genome Campus; ways in which the Sanger Institute and Genome Campus facilitated their work and its wider dissemination; and collaborations in which they have been involved. We aimed to gain a range of expertise, interviewing at least one programme or project leader from scientific programmes (a senior member of the team from Cancer, Ageing and Somatic Mutation; Cellular Genetics; Human Genetics; Parasites and Microbes; Tree of Life), as well as wider functions at the Genome Campus. Wellcome provided an initial list of suggested interviewees, and the study team added to this list and prioritised between the interviewees to get the range of expertise across the thematic areas necessary for this study.

A.3. Task 3: Bibliometric analysis

Task 3 consisted of bibliometric analysis of outputs associated with Sanger Institute and Wellcome Genome Campus, and of comparator organisations. For this task, we worked with CWTS, which was sub-contracted as a data provider for this portion of the work. We based our choice of publications for Sanger and comparators on the already defined institution classifications in CWTS's databases, which are based on author affiliations. However, the WHG is not included in that dataset, so we identified publications using a dual approach in the data acquisition process for the WHG. We searched for name variants in the address affiliations of the publications, and used specific funding numbers provided by Wellcome to search in the funding acknowledgements of the publications.

We developed the following indicators for all institutions:

- The number of publications (**P**) in international journals of the unit of analysis in the period
- The internal coverage (**Int_cov**) of a set of publications in the WoS measured by the percentage of references from that set, which are also covered by the WoS
- The number of citations received by P during the entire period, excluding self-citations (**TCS**)
- The average number of citations without self-citations per publication (**MCS**)
- The percentage of publications not cited by others (in the given time period) (**Pnc**)
- The **MNCS**: the actual number of citations (without self-citations) divided by the expected number of citations per publication; the expected number of citations was based on the worldwide average citation score without self-citations of all similar publications belonging to the same scientific field, and thus a field normalised score was calculated for each publication; next, the MNCS indicator was computed for each unit of analysis, by taking the average of these field normalised citation scores for individual publications; a value above 1 indicates that the mean impact for the unit is above world average whereas a value below 1 indicates the opposite
- The **MNJS**: the average citation impact of the journals in which the publications appeared that were published by the unit of analysis; the indicator was calculated using the same principles as for calculating the MNCS; it shows whether the publications originating from the unit of analysis were published in top or in sub-top journals for citation impact

- The number of highly cited publications (**P[top 10%] and P[top 1%]**) in international journals of the unit of analysis in the period
- The percentage of highly cited publications (**PP[top 10%] and PP[top 1%]**) The percentage of publications published by the unit that are among the upper top 10%/1% percentile of the citation distribution for similar publications belonging to the same fields.

A variable citation window and publications from 2008 to 2017 were used to calculate indicators. Publication outputs were analysed using full counting, while citations were analysed using fractional counting. As well as these core indicators, trend analyses were produced for: P, MNCS, MNJS, PP (top 10%) and PP (top 1%).

We also analysed the fields of research in which institutions publish, by WoS journal subject categories, and their performance (MNCS) for each of these categories. We looked at knowledge use, analysing the fields from which institutes received their citations, again based on WoS journal subject categories. Finally, collaboration networks maps were produced for each institution based on their co-authorship on publications with other institutions.

A.4. Task 4: Case studies

To illustrate Sanger and the Genome Campus' contribution to the field of genetics and genomics, we conducted in-depth case studies tracking their role in a number of specific developments. In particular, we used case studies to explore in depth the pathways through which wider impacts occur, the barriers and facilitators, and the role of Wellcome relative to other actors. We identified case studies by drawing on the REA, holding interviews with key stakeholders and following suggestions from Wellcome, focusing particularly on questions Wellcome was most interested in exploring at the mid-point of the study. Forward-tracing case studies started from specific activities conducted at Sanger Institute and the Wellcome Genome Campus and explored the range and nature of outcomes from those activities, while backward-tracing case studies started from recent major advances in the field, and traced backwards to identify the contributions to that development across the Sanger Institute, the Wellcome Genome Campus, the comparator organisations and others to assess the role and nature of those contributions, and how they came together to achieve progress. We explored these case studies in this study:

- Open Targets
- The Tree of Life programme
- Deciphering Developmental Disorders (DDD)
- Malaria research at the Sanger Institute, including MalariaGEN
- The International Cancer Genome Consortium (ICGC).

A.5. Task 5: Synthesis and reporting

We took a structured approach to collating and analysing the different evidence and views collected through this study. This involved synthesising the information collected in tasks 1–4 in order to triangulate findings

against the research questions listed in this report. First, we mapped evidence from each data collection tool against the research questions. Then, a member of the team reviewed the evidence against each question to identify key messages and issues emerging. We then conducted an internal workshop with the study team where findings from each task were considered with the research questions and sub-questions in mind. At this point, the team also considered the main messages emerging and discussed potential implications for Wellcome.

Along with this report, an inception report and interim presentation with slides was provided to Wellcome. In discussion with Wellcome we identified the elements of this report that will be useful and appropriate to share with a wider audience, which will result in an additional publicly available RAND report with an associated executive summary. All outputs have been and will be subject to RAND Europe's quality assurance processes.

A.6. Caveats and limitations of the analysis

We have identified a number of caveats and limitations to this analysis, the most important of which are set out below.

- **Reliance on interviews and self-report**: A significant proportion of the information used for this analysis is based on interviews. This content therefore depends on the accuracy and completeness of the recall of participants and their willingness to disclose information and views; it is subject to any biases or particular perceptions of those participants. We mitigated this by interviewing a range of individuals, both internal and external to Sanger and Genome Campus and triangulating against other evidence sources.
- **Bibliometric limitations**: Bibliometric analyses are subject to some common limitations, such as the use of citations as a proxy for research quality (where citations may be for a variety of reasons, including to highlight work considered poor or erroneous), the limits in coverage in bibliometric databases of some fields of research (although those of interest here are largely well covered) and language limitations in coverage (with English-language publications favoured). These apply to all bibliometric analyses, not specifically to this project. In this project we identified relevant publications primarily through author addresses (showing affiliation to the relevant institutions). However, authors may have dual affiliations and may not use them consistently – and this may apply more to some institutions than others – so there may be some limitations in the completeness and boundaries on the sample of publications included in the analysis.
- **Completeness of the range of contributions collected**: This report provides a picture of some of the important contributions to the field of genetics and genomics research made by Sanger and Wellcome Genome Campus and comparators. It by no means provides a comprehensive record of all contributions made, and indeed it is likely that many contributions are not even known to the researchers involved in the work because of the complex nature of research translation pathways. However, it aims to give a flavour of the range and nature of these contributions and showcase a number of illustrative examples.

- **The range of comparators**: A sample of comparators was selected to review and benchmark the performance of Sanger Institute and Wellcome Genome Campus and provide a context for their role in the wider landscape. However, there are many other potential comparators that could have been chosen, which may have provided a different perspective. This is not to negate the contributions of those other actors, rather a reflection of the pragmatic need to select a number of comparators, which offer different types of insights into the role and model at Sanger and the campus.

We note a specific concern around the publications identified in relation to Janelia. Janelia Research Campus has by far the lowest number of publications compared with the other institutions. We broadened our sample by including mentions in funding acknowledgements of the terms, 'Janelia Research Campus', 'Janelia Farm', 'Janelia Visitor Program' or 'HHMI Janelia' (and also included publications listing these as an affiliation). However, the number of publications included was still markedly lower than expected, suggesting a significant proportion of publications from researchers based at the institution may be published under another affiliation. This limits the completeness of the analysis for this particular comparator.

Annex B. Interview protocols

This annex sets out the protocol for external interviews conducted with experts in the fields of genetics and genomics, and individuals from the four comparator organisations (Section A.1), and for internal interviews conducted with individuals from Sanger and the Genome Campus (Section A.2).

B.1. Protocol for external interviews

Prior to the interview, all interviewees will receive a **privacy notice** introducing the study, the purpose of the interview and outlining how the data provided by the interviewee will be used. The privacy notice will also outline how interviewees can raise data privacy concerns with the study team, and their rights as participants in the study.

B.1.1. Introduction

Thank you for making the time to speak with us today. Before we begin, I will give you a brief overview of the study and the purpose of the interview.

RAND Europe has been commissioned by Wellcome to conduct a landscape review of the Wellcome Sanger Institute and the Wellcome Genome Campus. This landscape review is designed to understand the role of the Sanger Institute and Genome Campus in the field of genetics and genomics, including their strengths, limitations, contributions to the field and potential areas of opportunity. Along with desk research and interviews, the study will include bibliometric analysis of outputs associated with the Sanger Institute and Genome Campus, and a number of case studies looking at significant advances in the field of genetics and genomics to help us understand whether and how the Sanger Institute or Genome Campus may have contributed to these advances. We will also be looking at several comparator organisations, to see how their contributions to the field compare or contrast to the Sanger Institute and Genome Campus.

You are being interviewed as an expert in the field of genetics and/or genomics for this study. The purpose of this interview is to understand the role of the Wellcome Sanger Institute and Genome Campus and other comparable organisations in the field, and to understand what you see as their strengths and limitations. We are also seeking out what recent advances in the field of genetics and genomics may be most significant, and your views on how the Sanger Institute, Genome Campus and comparator organisations have contributed towards these advances. We understand you might not be able to answer every question, so if you do not know how to answer a question or if you do not want to provide an answer, please feel free to let me know and I can move onto the next question.

Do you have any questions?

Are you happy to proceed on this basis?

(Wait for confirmation.)

Before we begin the interview, would it be okay with you to record this interview? This recording will be only for internal note-taking purposes and will be destroyed after the study is complete.

(Wait for consent to record and begin recording.)

B.1.2. Background and role

1. Could you please briefly describe your current role and professional background?
    a. How long have you worked in the field of biomedical research?
    b. How long have you worked in the field of genetics and/or genomics?

2. Have you been involved with the Wellcome Sanger Institute or Wellcome Genome Campus, either currently or in previous projects or roles? If so, could you describe your involvement? *(probes: funding, working on individual projects, associated faculty of Sanger/Genome Campus)*
    a. Have you been involved with Wellcome more widely, outside of the Sanger Institute or Genome Campus? If so, could you briefly describe your involvement? *(Probes: funding, working on individual projects, associated faculty Wellcome)*

        (Note: Previous involvement with Sanger/Genome Campus and/or Wellcome does not disqualify the interviewee from responding. If asked, reassure the interviewee that we are just trying to understand how familiar they are with the work of Sanger/Genome Campus.)

B.1.3. Contribution of Sanger Institute and Wellcome Genome Campus

3. *(If not answered through question about association with Wellcome/Sanger/Genome Campus above)* How familiar would you say you are with the work of the Wellcome Sanger Institute and Wellcome Genome Campus?
    a. What are some of their scientific programmes or specific projects that you are most familiar with?

4. *(If interviewee describes familiarity with overall work or specific programme/project)* Could you describe how the Wellcome Sanger Institute and/or Genome Campus have contributed to the field of genetics and genomics? *{Probes: academic or intellectual contributions, societal impacts such as improved products or services, changes to medical practice or patient-level outcomes)*
    a. Are there particular programmes or projects that have been especially impactful? If so, could you describe how these have impacted on the field? *(Probes: breakthrough findings from individual projects, streams of work)*

   b. Are there any outputs other than publications, that have been especially impactful? For example, these may include tools, datasets or databases?

   c. Are there any ways of working at the Sanger Institute or Genome Campus that have been adopted more widely in the field or genetics, genomics or biomedical research? *(Probes: physical campus location, bringing in associated faculty, funding strategies, ways of working with private companies and spin outs)*

5. *(Ask in relation to a particular contribution)* If the Wellcome Sanger Institute and/or Genome Campus did not exist, what difference do you think that might have made to progress in the field, and to the particular contributions you have described? *(Probe: Would the work have been carried out by another organisation? Would it have had the same impact? Would it not have happened at all?)*

6. How has work from the Wellcome Sanger Institute or Genome Campus been used by others working in the field of biomedical research, genetics and genomics?

   a. How has work been used by academics and researchers?

   b. How has work been used by private companies?

   c. How has work been used by those working to improve health?

7. Are there any areas where the Sanger Institute or Genome Campus has particular strengths? If so, can you describe what makes them well suited to work in this particular area? *(Probes: Streams of work (such as genome sequencing and genome variation, cancer, ageing, cellular genetics, parasites/microbes, evolutionary insight, computational genomics) or wider ways of working (such as their approach to associated faculty, funding, spin outs, PhD programme))*

   a. Is this strength unique to the Sanger Institute and/or Genome Campus?

   b. How can the Sanger Institute and/or Genome Campus build on these strengths?

8. Are there any areas where the Sanger Institute or Genome Campus has particular challenges or weaknesses? If so, could you describe these challenges? *(Probes: particular areas within the field of genetics/genomics, particularly weak work streams or challenges associated with how they operate)*

   a. Do you feel this is an area that the Sanger Institute and Genome Campus can improve on? If so, how, and if not, why?

B.1.4. Contribution of comparator organisations

As mentioned, we are also including several comparator organisations in our analysis to help understand their contributions and how this compares and contrasts with the Wellcome Sanger Institute and Genome Campus. Could you please indicate your familiarity with the following organisations?

- (List of comparator organisations to be inserted)

(Note: Interviewer to focus on comparator organisations that the interviewee is familiar with.)

9. Could you describe how the (ORGANISATION) have contributed to the field of genetics and genomics? *{Probes: academic or intellectual contributions, societal impacts such as improved products or services, changes to medical practice or patient-level outcomes)*
    a. Are there particular programmes or projects that have been especially impactful? If so, could you describe how these have impacted on the field? *(Probes: breakthrough findings from individual projects, streams of work)*
    b. Are there any ways of working at (ORGANISATION) that have been adopted more widely in the field or genetics, genomics or biomedical research? *(Probes: physical campus location, bringing in associated faculty, funding strategies, ways of working with private companies and spin outs)*
    c. How have these contributions been taken up by academics and researchers, private companies and others working to improve health?

10. Can you comment on how the contributions of (ORGANISATION) and those of the Wellcome Sanger Institute and Genome Campus relate to one another?
    a. Do you see these organisations as complementary to one another, or competing with one another? Or alternatively do the two organisations have another type of relationship?
    b. How do the strengths and weaknesses of each organisation relate to one another?

B.1.5. Advances in the field of genetics and genomics

We'll also be working backwards, looking at major advances in the field of genetics and genomics and seeing if or how the Sanger Institute and Genome Campus contributed to these.

11. Looking back over the last 5–10 years, could you briefly summarise 1–2 major advances in the field of genetics and genomics that have been particularly significant?
    a. What impact have these advancements had? *(Probe: for academics/researchers, for private companies and the availability of new or improved products or services, for those working to improve health)*

12. Are you aware if the Wellcome Sanger Institute or Wellcome Genome Campus had any role in these advancements? If so, can you comment on their contribution?
    a. What was their unique contribution over and above other contributions?

13. Are you aware if (COMPARATOR ORGANISATIONS) had any role in these advancements? If so, can you comment on their contributions?

B.1.6. Conclusion

14. Are there any areas of opportunity that you feel are important for the Sanger Institute and Genome Campus to pursue? *(Probe: Particular areas within the field, particular research questions, ways of working from other organisations)*

15. Do you have any other areas that you think we should look at in this landscape review, or anything you did not have the chance to say that you feel is important?

Thank you for taking the time to participate in this interview.

B.2. Protocol for internal interviews

Prior to the interview, all interviewees will receive a privacy notice introducing the study, the purpose of the interview and outlining how the data provided by the interviewee will be used. The privacy notice will also outline how interviewees can raise data privacy concerns with the study team, and their rights as participants in the study.

B.2.1. Introduction

Thank you for making the time to speak with us today. Before we begin, I will give you a brief overview of the study and the purpose of the interview.

RAND Europe has been commissioned by Wellcome to conduct a landscape review of the Wellcome Sanger Institute and the Wellcome Genome Campus. This landscape review is designed to understand the role of Sanger and the Wellcome Genome Campus in the field of genetics and genomics, including their strengths, limitations, contributions to the field and potential areas of opportunity. Along with desk research and interviews, the study will include bibliometric analysis of outputs associated with Sanger and the Genome Campus, and a number of case studies looking at significant advances in the field of genetics and genomics to help us understand whether and how Sanger or the campus may have contributed to these advances. We

will also be looking at several comparator organisations, to see how their contributions to the field compare or contrast to Sanger and the campus.

You are being interviewed as an internal employee of Sanger and/or the Wellcome Genome Campus. The aim of the interview is to understand your perceptions regarding the role, contribution, strengths and weaknesses of both your own specific organisation and the campus as a whole. We will also be interested in your views on the contribution of the specific research programme(s) in which you have been involved. The interview will also touch upon the relationship between Sanger, Genome Campus, and comparator organisations, and discuss what you see as the major opportunities for Sanger and the campus in the future. We understand you might not be able to answer every question, so if you do not know how to answer a question or if you do not want to provide an answer, please feel free to let me know and I can move onto the next question.

Do you have any questions?

Are you happy to proceed on this basis?

(Wait for confirmation.)

Before we being the interview, would it be okay with you to record this interview? This recording will be only for internal note-taking purposes and will be destroyed after the study is complete.

(Wait for consent to record and begin recording.)

B.2.2. Background and role

1. Could you please describe your position within the Sanger/Genome Campus
    a. Which organisation do you work for, and what is your specific role within that organisation?
    b. Are you involved in any specific research programmes?
    c. How long have you worked at the Genome Campus? What programmes (if any) have you worked on in the past?
    d. What organisations (if any) have you worked for outside of the campus?

B.2.3. Role and contribution of Sanger Institute and Wellcome Genome Campus

The first part of the interview will focus on the role and contribution of Sanger and/or the Genome Campus to genetics and genomics research. Here, if possible, we would like you to focus your answers on the role and most important contributions of Sanger and the campus as organisations. There will be an opportunity to talk about the specific role and contributions of your own research programme later in the interview.

2. Please could you begin by describing the role of Sanger and the campus within the genetics and genomics research landscape?

a. How would you describe the working relationship between Sanger and the campus?

b. To what extent do you think these respective roles are actually fulfilled by Sanger and the campus?

3. Are there any areas in which you feel Sanger and/or the campus has particular strengths? *(Probes: particular streams of work (such as genome sequencing and genome variation, cancer, ageing, cellular genetics, parasites/microbes, evolutionary insight, computational genomics); particular programmes; particular ways of working (such as their approach to associated faculty); or particular forms of impact translation, such as spin outs and pipeline groups)*

    a. What makes Sanger/Genome Campus well suited to this particular area/form of work?

    b. Is this strength unique to them?

    c. In what ways could they build on these strengths?

4. What, in your opinion, are the major ways in which Sanger and the Wellcome Genome Campus have contributed to the field of genetics and genomics research?

    a. Are there specific programmes or projects that have been especially impactful?

    b. Are there any outputs other than publications, that have been especially impactful? For example, these may include tools, datasets or databases.

    c. Why was this contribution significant?

    d. What has this contribution led to? *(Probes: How has the work been used by academics and researchers, by private companies and/or by other actors and organisations working to improve health?)*

    e. What was it about Sanger and the Genome Campus that made these contributions possible? *(Probes: ways of working, structure of programmes, collaboration with industry etc.)*

    f. To what extent are these contributions unique to Sanger and the Genome Campus?

(It is possible that the interviewee will wish to talk about more than one contribution made by Sanger and Genome Campus. In this case, the interviewer should pose the same follow up questions (a-e) in each case.)

5. In your opinion, are there any areas in which Sanger Institute and/or Wellcome Genome Campus faces particular challenges or weaknesses? *(Probes: particular areas within the field of genetics/genomics, particularly weak work streams or challenges associated with how they operate)*

    a. Do you feel this is an area that Sanger and Genome Campus can improve on? If so, how, and if not, why?

B.2.4. Role and contribution of specific Sanger Institute/Genome Campus programmes

In this part of the interview, we will discuss the role and contribution of the specific Sanger and/or the campus' research programme (or programmes) in which you have been involved.

Discussion should be limited to a maximum of two programmes. If the interviewee wishes to discuss more than one programme, it is suggested that this is done sequentially, with questions 6–8 asked in each case)

6. Please could you begin by describing the role of Sanger/Campus programme(s) with which you are most familiar?
    a. What are/were the aims of this programme?

7. What, in your opinion, are the major ways in which this programme has contributed to the field of genetics and genomics research?
    a. Why was this contribution significant?
    b. What has this contribution led to? *(Probes: How has the work been used by academics and researchers, by private companies and/or by other actors and organisations working to improve health?)*
    c. What was it about the programme that made these contributions possible? *(Probes: ways of working, structure of programmes, collaboration with industry etc.)*

8. Are there any areas in which this programme has faced particular challenges or weaknesses?
    a. Are these weaknesses specific to the programme?
    b. Are there things that could be done to address these weaknesses?

B.2.5. Comparator organisations

In this section of the interview, we would like to discuss comparator organisations. The aim here is to understand your perceptions on the role and contribution of other organisations working in the field of genomics research, and to understand how this compares and contrasts with your perceptions regarding Sanger and Genome Campus.

To begin, could you please state your familiarity (if any) with the following organisations?

- List of comparator organisations to be inserted

(Note: Interviewer to focus on comparator organisations that the interviewee is familiar with.)

9. Please could you describe the role of the (comparator) within the genetics and genomics research landscape?

10. Are there any particular contributions to genetics and genomics research that you associate with the (comparator)?

11. Could you please comment on how the contributions of the (comparator) and those of Sanger/Genome Campus relate to one another?
    a. Do you see these organisations as complementary to one another, or in competition? Alternatively, do the two organisations have a different type of relationship?
    b. How do the strengths and weaknesses of each organisation relate to one another?

B.2.6. Future opportunities and challenges

The final part of the interview will explore your views on the future opportunities and challenges for Sanger/Genome Campus

12. What do you see as the main areas of opportunity for Sanger and/or Campus to contribute to genetics and genomics moving forward?
    a. Do you think Sanger/campus is well placed to take advantage of these opportunities? If so, why?
    b. What obstacles might Sanger/campus face that would prevent it from realising these opportunities?
    c. In what ways could it be better prepared?

B.2.7. Conclusion

That brings us to the concluding question of the interview:

13. Are there any other areas that you think we should look at in this landscape review, or anything which you feel is important, but have not had a chance to say?

Thank you for taking the time to participate in this interview.

Annex C. Case study structure

This landscape review includes five case studies (detailed in Chapter 4) analysing the contribution of Wellcome Sanger Institute and Wellcome Genome Campus towards key developments in the field of genetics and genomics. These case studies cover both individual findings that have been significant in the field, as well as wider ways of working staff at Sanger and the campus use, which have proved influential. Each case study was written up following a common template designed by the study team to aid analysis and cross-case comparisons. Box 2 shows the indicative structure for the case studies. It is based on an adapted version of the Payback Framework developed by researchers to analyse the impact of health services research (Buxton and Hanney 1997).[96] This format has been adapted for individual specifics of each case study.

Box 2: Indicative case study structure

Introduction

This section will include a short overarching description of the case study. Possible examples include a particular development within genomics in which Sanger/campus played a significant role, a particular way of working promoted by Sanger/campus, or any broader contribution of Sanger/campus. The section will also provide a brief introductory statement on the role of Sanger/Genome Campus in relation to the case study.

Background and context

This section will provide background and context to understand the case study and how it came about. This will include both the scientific background to understand the work conducted, but also the wider context and rationale for the activities described in the case study – such as the reasons why the research was conducted, who funded it, where the idea came from. The aim of this section is to situate the case study within the wider landscape both from the point of view of the science and in terms of the research system.

Research process

What was done, by whom and how? This section will explain the process of research – what was done, what did it involve, what were some of the challenges, anything novel or new about the process. In particular it will be important to capture the nature of collaboration in the process, and the role of Sanger within that. It would also be useful to explore the extent to which the set up within the Sanger and the Genome Campus – be that funding, culture, leadership or other factors – facilitated (or indeed impeded) the process. Reflections on the ways in which other organisations – our comparators or others – facilitated and contributed would also be important.

Contributions to knowledge

[96] The Payback Framework was originally developed by Martin Buxton and Stephen Hanney at the Health Economics Research Group (HERG), Brunel University. The framework assesses the impact of health-based research according to its contribution across five key categories: knowledge; benefits to future research and research use; benefits from informing policy and product development; health and health sector benefits; broader economic benefits.

This section will focus on the ways in which the case contributed to scientific knowledge. It will describe intellectual developments associated with the case, including publications within scientific journals and significant research reports.

Contributions to future research

This section will focus on the ways in which the case contributed to future research. It will consider the ways in which the case contributed to the development of research skills and research capacities, as well as collaborations between different organisation and stakeholders. It will also cover the ways in which the research contributed to focusing future research on productive avenues, or developed new methodologies or datasets that could be used elsewhere.

Contributions to policy and product development

This section will focus on the wider commercial benefits arising from the case. Key considerations here will include collaborations with industry and the development of new products and innovations linked to Sanger/Campus research. It will also cover any policy impacts of the work including changes in policy, to guidelines, impacts on training, effects on funding and support in research or health.

Contributions to health and the health system

This section will focus on the ways in which the case contributed to improvements in health and healthcare. It will consider new health innovations related to the case study and conduct a broad qualitative account of the impact of these interventions on health outcomes and the delivery of health services.

The role of Sanger Institute/Genome Campus

This section will analyse in more detail the specific role played by Sanger/Genome Campus in the developments explored in the case study. It will explore the particular role Sanger/Genome Campus programmes and/or research groups involved in the case, the specific ways in which these helped make the various contributions possible, and the role of these contributions relative to other organisations and stakeholders.

Lessons learned from the case study

This section will reflect on key lessons that can be drawn from the case study, with specific reference to Sanger/campus' s role and contribution. Building on the previous section, it will reflect on the distinctive features of Sanger/Genome Campus' contribution to the case, examine why these contributions were effective (as well as possible areas for development), and consider possible opportunities for Sanger/Genome Campus to build on these lessons moving forward.