



09

➤ Annual Review

The Wellcome Trust Sanger Institute led the world in decoding our human genetic makeup. As the largest single contributor to the Human Genome Project, the Institute results were released freely on the internet, to power the research of scientists worldwide.

Today, the Sanger Institute is using that foundation to understand the role of genes in health and disease through its own research programmes and through its leading role in international efforts such as the International Cancer Genome Consortium, the 100 Genomes Project, the MalariaGEN consortium of malaria researchers and the International Knock-Out Mouse Consortium. The Institute's research is founded on an ability to carry out biomedical research on a scale that is, in many cases, unmatched within Europe and its research findings are already being used in the development of new diagnostics and treatments for human disease.

The Wellcome Trust is the largest charity in the UK. It funds innovative biomedical research, in the UK and internationally, spending over £600 million each year to support the brightest scientists with the best ideas. The Wellcome Trust supports public debate about biomedical research and its impact on health and wellbeing.

04 ↘ Director's introduction

06 ↘ A year in review

14 ↘ Human genetics

32 ↘ Mouse and zebrafish genetics

48 ↘ Pathogen genomics

60 ↘ Bioinformatics

70 ↘ Core facilities

76 ↘ The Institute, Education and Communication

↘ Annual Review



Welcome to the Wellcome Trust Sanger Institute's Annual Review for 2008/09.

The Sanger Institute was founded in 1993 and rapidly established itself as Europe's leading centre of genome sequencing and analysis. It contributed one-third of the finished human genome sequence and has played a leading role in many international research consortia expanding our knowledge of the human, mouse and other genomes.

In recent years our emphasis has shifted to large-scale, high-throughput genetic analysis of humans, model organisms and pathogens. These studies are underpinned by a uniquely powerful technological and methodological infrastructure spanning sample preparation, bioinformatic analysis, computing and biological resource development.

Our aim has been to document genetic variation, to identify variation that influences human health, and to study the biological impact of this variation. In doing so, we can begin the process of translating the potential of genomic sequence data into practical applications that deliver medical benefits to patients.

As early adopters of next-generation sequencing technologies, we have significantly enhanced our productivity and maintained a leadership position across our key areas of interest. Our 800 scientists are driving forward their own research programmes and also developing resources widely used throughout the global research community.

This Annual Review highlights some of our major achievements during the year and provides brief summaries of our research programmes and projects and their likely future development.

Allan Bradley, Director
September 2009

As new sequencing technologies generate almost unimaginable amounts of data, there is a growing challenge to interpret genomic information and ultimately to apply it in a medical setting.



The Sanger Institute began life as a genome-sequencing centre. Although human health was the main driver of the Human Genome Project, the medical benefits were seen as accruing some time down the line. Initially, the Sanger Institute concentrated on sequencing. Over time, however, our emphasis has shifted significantly. While still generating and analysing very large amounts of sequence data – now the equivalent of several human genome sequences every day – we also have major programmes dedicated to answering biological questions of medical relevance. Our model organism and pathogen programmes carry out laboratory-based research activities that go beyond characterising genomes into experimental analysis of the biological function of genes and their variants, establishing their roles in health and disease.

Over the last few years the clinical relevance of our work and opportunities for medically relevant discoveries have been attracting a growing cadre of clinically qualified individuals to the Institute. As well as several faculty members with a clinical background, a significant number of clinicians are working at various levels in the Institute. We offer a range of training opportunities for clinicians through the University of Cambridge / Wellcome Trust Clinical PhD Programme, as well as at postdoctoral and more senior levels. We have established close links with Addenbrooke's Hospital in Cambridge and other clinical centres in the UK and elsewhere.

In addition to opening up novel areas of research, new sequencing technologies and the knowledge being generated by them will create potential new healthcare opportunities. Clinicians who have the vision, passion and first-hand experience of genomic advances will be key to the translation of this potential into tangible healthcare benefits.

Another way in which we enhance the clinical impact of our work is through collaborations. Nigel Carter, for example, has worked closely with Helen Firth at Addenbrooke's Hospital on the development of the DECIPHER database, which collates both genomic and clinical data on chromosomal abnormalities causing developmental disorders. This database, which holds data on cases from all over the world, is supporting both basic science discovery and clinical practice. In the microbiological area, Gordon Dougan and colleagues have established a mouse model of *Clostridium difficile* infection, and identified its spore as the main route of disease transmission. We believe that this will prove a valuable model in which to test infection control measures.

Clinically qualified researchers are a part of the Sanger Institute's large, young and diverse workforce. Our highly competitive international four-year PhD programme attracts many of the best graduate students worldwide. Over the last year recruitment into our faculty programme has focused on statistical genetics, which we achieved with the recruitment of Jeff Barrett and Eleftheria Zeggini, two researchers at early stages of their careers who are making significant contributions to the analysis of data from high-throughput genotyping and other studies. Manjinder (Manj) Sandhu has taken up a part-time position with us and is developing a genetic epidemiology programme. Manj holds a substantive appointment in the Department of Public Health and Primary Care, University of Cambridge, and will help to strengthen ties between the Institute and the university in this important area. We also recruited Paul Kellam to lead up research in virology, ensuring that our pathogen research spans all families of infectious organisms from viruses through bacteria to eukaryotic microbes and parasites. We are thrilled that Manolis Dermitzakis, a faculty member who started his independent career at the Institute just a few short years ago, is leaving to take up a significant leadership position in the University of Geneva, Switzerland.



George Vassiliou. Clinician scientist at the Wellcome Trust Sanger Institute.

We were delighted this year that Karen Steel was elected a Fellow of the Royal Society, following in the footsteps of Mike Stratton who received a similar accolade in 2008. Karen has made enormous contributions to studies of the genetic basis of deafness, including the recent discovery of a microRNA mutation causing deafness in mice and humans. Two of our microbiologists were honoured, with Gordon Dougan made a Fellow of the American Academy of Microbiology and Julian Parkhill elected to the fellowship of the UK Academy of Medical Sciences. New recruit Elizabeth Murchison was one of just four women awarded a 2009 L'Oréal–UNESCO UK and Ireland For Women in Science Fellowship. Such recognition is testament to the scientific strength of the Sanger Institute.

Constant change

Two years ago, we made the strategic decision to switch as rapidly as possible to next-generation sequencing technologies, which offered the prospect of vastly enhanced sequence output. This advance in technology demanded significant changes to the way we operate. The implementation of new sequencing and genotyping pipelines, and of new systems of data capture, storage and analysis, has been remarkably smooth.

The scale of this activity is worth reflecting on. At times the Wellcome Trust Genome Campus is responsible for a significant proportion of the data traffic across academic networks. As data from the 1000 Genomes Project becomes available, bandwidth demands will undoubtedly increase still further. Genome-based data generation is fast approaching the scale typically seen in physics mega-projects such as the Large Hadron Collider at CERN.

At the moment, a large chunk of sequencing capacity is dedicated to the 1000 Genomes Project. Other large-scale initiatives such as the Cancer Genome Project also have high demands, producing complete genome sequence information for a range of human cancers. However, as the degree of variation in the genome becomes fully documented, capacity will increasingly be focused on clinical studies, with sequence information being generated for patient samples. The major challenge will be to understand the significance of the variation identified and what it means to an individual. This has been a factor in our drive to build up our expertise in statistical analysis, to extract additional information from data, and in model organisms, to assess the functional significance of variation.

Significantly, high-throughput sequencing has been put to use across all our programmes. In pathogen studies, for example, entire populations of microbes can be sequenced, which is providing a much clearer view of how strains are distributed in the environment and how they are evolving. This can provide important insight into the nature of disease spread and also suggest ways in which transmission might be interrupted.

Resource generation

From its earliest days, the Sanger Institute has provided community resources, principally freely accessible genome sequence, which has greatly accelerated research worldwide. Resource development remains an important aspect of our work. As well as genome data, we generate a wide range of other resources that enable the research community to make best use of genome information. Many of our software tools have become *de facto* standards. Ensembl and its associated genesets are widely used by the community, as are our other biological databases such as Rfam and Pfam.



With the zebrafish genome sequence scheduled for publication early in 2010, much of our work is centred on the provision of mutants to the research community. These mutants have been used by numerous groups studying processes as varied as tumour development, muscle and bone formation, small RNA function, and even sleep.

We have also greatly increased our supply of mouse resources, primarily engineered mouse embryonic stem (ES) cells and phenotyped mice strains created by the Mouse Genetics Programme. Bill Skarnes' team have now generated more targeted mutant ES clones than have been produced by the entire academic community over the last 20 years, greatly accelerating mouse genetics research worldwide. We are currently in discussions to develop the use and value of this resource further through coordinated international collaborations.

The Sanger Institute maintains an international perspective, both in its recruitment and in its collaborations. The Human Genome Project established a new model for large-scale international collaboration in biology, and we have continued in its spirit, leading or contributing to many large international consortia. The 1000 Genomes Project is based on strong links with the USA and China. We are part of the ENCODE initiative, which is aiming to identify all functional elements in the human genome. The Cancer Genome Project has been a driving force in the establishment of the International Cancer Genome Consortium which is planning to sequence 500 examples of each of 50 different tumour types.

An important new collaboration is the global MalariaGEN network, led by Dominic Kwiatkowski, which is attempting to identify the genetic factors affecting human susceptibility to malaria – a task made more difficult by the high genetic diversity of African populations. This work has made major progress this year, and will benefit further from data from the 1000 Genomes Project.

Malaria is a relatively new area for us, but one where we have now established a coherent experimental base that in many ways is a microcosm of our more general strategy. We can now combine large-scale human genetic studies identifying genetic predispositions, investigate the impact of genetic variation in the parasite, and undertake experimental work in model systems to unpick important biological mechanisms. This is beginning to generate important findings that will shed light on the interaction between parasite and host and support the development of control measures.

A further important step forward has been our development of genetically unmodified induced pluripotent stem cells, in both mice and humans. This is particularly exciting as it greatly expands our ability to carry out work on a 'human model' system. This is one of the areas likely to grow further over the coming years, and again emphasises the increasing medical relevance of our work.





The Sanger Institute is characterised by a large-scale, high-throughput approach to research.



Sequencing

Since its inception, the central feature of the Sanger Institute has been the generation and analysis of DNA sequence information. Over the last two years the Institute has converted most of its sequencing capacity to next-generation technologies. It now has around 40 such machines, making it one of the largest facilities of its type in the world. The immense power of this platform enables the Institute to generate the equivalent of more than 100 haploid human genome sequences per week.

This sequencing capacity is directed at many projects, from genomes ranging in size from three billion base pairs, for example the human, to viruses of a few thousand base pairs. The projects aim to identify normal and disease-associated sequence variation in humans, somatic mutations in cancer, variation among strains of mice and to provide the reference gorilla, zebrafish and pig genome sequences. Among pathogens, reference sequences for parasites, such as the blood fluke *Schistosoma mansoni*, are being generated, and variation in the malaria parasite, bacterial and viral genomes, including influenza, is being explored in order to understand the ways in which these organisms cause disease and spread.

Informatics

High-throughput experimental methods have turned biology into a data-rich science where large-scale computational analysis is central to scientific discovery. The Sanger Institute's scientific success is underpinned by its depth of expertise in developing and applying computational methods and in the provision of large-scale IT systems.

The Institute's 1000 sq m data centre has a disk storage capacity in excess of three petabytes. More than 5000 computer 'cores' can be called upon by researchers for data analysis. These huge computer resources are being used to analyse and manage data from projects such as the Wellcome Trust Case Control Consortium and the 1000 Genomes Project. The Institute has a strong reputation for open release of data and in generating online resources that organise data in ways to maximise its utility to the scientific community. These include the Ensembl genome browser, developed in collaboration with the European Bioinformatics Institute (also located on the Wellcome Trust Genome Campus). These online resources are used by hundreds of thousands of researchers.

Research Support Facility

In 2001, the Sanger Institute began to build on its highly successful sequencing and informatics research portfolio, by introducing major programmes in model organism genetics. These programmes, focused on mouse and zebrafish, were designed to layer functional data onto vertebrate genomes to complement discoveries in human genetics. To support these new initiatives, a Research Support Facility was built.

One defining aspect of work at the Institute is its scale. Thus at inception the Research Support Facility was sized to support a large and ambitious model organism programme. Opened in 2006, with a capacity of 21 000 mouse cages and 4000 fish tanks, and supported by state-of-the-art automation and containment suites, this facility is now fully operational.

With the scale-up of our mouse and zebrafish genetics programmes, the Sanger Institute has now become the largest producer of new mouse and zebrafish knockouts globally, with more than 250 mouse and 100 zebrafish mutants produced each year. The genetic resources provide a foundation for many scientific programmes. We are committed to sharing these valuable resources with researchers worldwide.

