



 The Wellcome Trust
Sanger Institute
2003



The Wellcome Trust



Allan Bradley, Director.

A passion for discovery

The Wellcome Trust Sanger Institute is one of the world's pre-eminent research establishments. This level of recognition has been achieved in less than ten years. In part, this comes from our leading role in studying the human genome, but it also reflects the fact that we do much more than sequence genomes: we are committed to unravelling many of their secrets and sharing our results. Visitors who walk through our doors, or stay and work, often comment on the level of human energy, our excitement about our work, our willingness to share our time and resources, and our passion for discovery. They recognize that the Sanger Institute and its staff are unique.

While most researchers know our name and many have visited our state-of-the-art research facilities, few may appreciate the breadth of our scientific enterprise and the extent of our training programmes. The pages of this brochure provide some detail of our history and illustrate our programmes.

Introduction	02	Research interests	34	Cell signalling, cell fate and proteolytic pathways in <i>C. elegans</i>	52
From sequence to biology	06	<i>Pombe</i> postgenomics	37	The atlas of gene expression	53
Sequencing	12	Pathogen sequencing	38	Bacterial sequencing and analysis	54
Data mining	16	Metabolic disease	39	Bioinformatics	55
High-throughput analysis	20	Protein and RNA families	40	X chromosome: biology, evolution and disease	56
Genetic variation and disease	22	Immunogenomics	41	Gene trapping in the mouse	57
Model organisms	26	Discovering genes involved in complex disease	42	Zebrafish genetics and embryology	58
Pathogen genomes	30	Mouse genetics	43	Cancer genetics	59
Genomic Infrastructure	32	Molecular cytogenetics	44	Embryo gene expression patterns	60
		Human genome sequence analysis	45	Microarrays, transcriptional networks and human disease	61
		Analysis of human chromosome 22	46		
		Genome informatics	47	Facilities	62
		Human disease analysis in the mouse	48		
		Cell signalling in <i>C. elegans</i>	49	Careers	64
		Prediction research	50		
		Parasite sequencing and pathogen microarrays	51		



The Sanger Institute is the only genome centre in the UK, and is one of the largest in the world.

The Wellcome Trust Sanger Institute, now funded principally by the Wellcome Trust, is the only genome centre in the UK, and is one of the largest in the world. It comprises a unique blend of production and multidisciplinary research in genetics, sequencing, computational analysis and biology. It has a rich history of innovation and management of large projects and data sets, which provides a firm foundation for the next phase of genome analysis.

The Institute was jointly founded in 1992 by the Wellcome Trust and the UK Medical Research Council (MRC). Initially named the Sanger Centre, it was officially opened in October 1993 by Fred Sanger, who devised the prototype of modern day sequencing techniques. Until 1996, the Sanger Centre was located in existing buildings in the grounds of Hinxton Hall estate, near Cambridge, UK. With the success of a bid to bring the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to Hinxton, the Wellcome Trust decided to establish a permanent Genome Campus at Hinxton, with purpose-built buildings.

Hinxton Hall and its 55-acre grounds are owned by the Wellcome Trust. The Sanger Institute, EMBL-EBI, and the MRC Human Genome Mapping Project Resource Centre occupy the new buildings of the South Campus, while

the original hall and stables have been fully renovated and form part of the Conference Centre of the North Campus.

The Sanger Institute's current laboratories accommodate around 650 staff. Plans for further expansion have been approved, and construction began in November 2002. The new facilities will include additional laboratories, a rodent facility and a state-of-the-art data centre, and all will be available for occupation early in 2005.

Under John Sulston's founding directorship the Sanger Centre grew from a group of 15 personnel in April 1993 to 580 in 2000, at which time he retired as Director. During this period there was major investment in technological and IT infrastructure, and in attracting and retaining a team of dedicated and highly motivated staff.

Researchers have been responsible for a constant stream of scientific achievements and publications marking major achievements during the Institute's history. The Institute has developed world-renowned expertise in generating and analysing genome sequence data from many organisms ranging from humans to mice, fish, worms, yeasts and pathogens.



Introduction





MILESTONES IN THE SANGER INSTITUTE'S HISTORY

- 1992 Sanger Centre initiated by John Sulston

- 1993 Sanger Centre officially opened by Fred Sanger

- 1996 Sanger Centre moves into purpose-built accommodation

- 1997 Yeast genome completed: Sanger makes largest single contribution

- 1998 Wellcome Trust commits to fund an acceleration of the human genome sequencing at the Sanger Centre – from one-sixth to one-third of the genome
C. elegans genome completed: Sanger Centre and Washington University, St Louis, sequence the first multicellular organism
 Publication of the genome of *Mycobacterium tuberculosis*, which causes TB

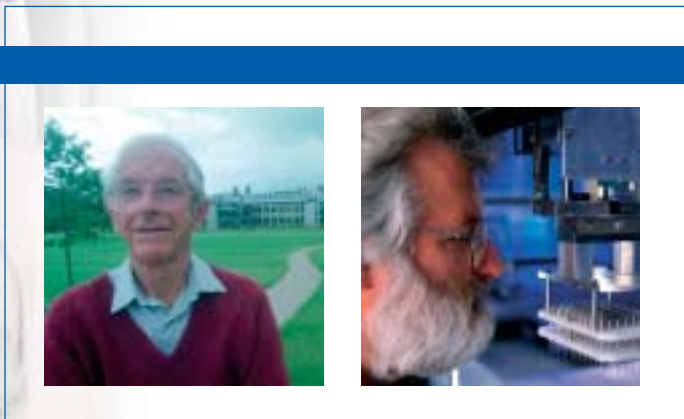
- 1999 Publication of the first finished sequence of a human chromosome, chromosome 22

- 2000 Allan Bradley takes over as Director of the Sanger Centre

- 2001 John Sulston knighted for services to genome research
 Publication of the draft human genome sequence in *Nature*
 New £300 million, five-year research programme announced. Sanger Centre becomes the Wellcome Trust Sanger Institute

- 2002 John Sulston awarded the Nobel Prize for work on the nematode worm *C. elegans*
 Publication of draft mouse genome sequence in *Nature*
 Publication of malaria parasite sequence in *Nature*

- 2003 Completion of the human genome sequence: the Sanger Institute makes the largest single contribution



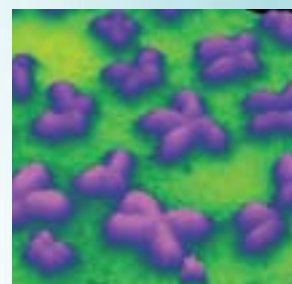
Far left Fred Sanger.
 Left John Sulston, founder Director.

Throughout its life the Sanger Institute has played a leading international role in promoting full and open cooperation in genomic research that is conducted in the public domain. The availability of DNA sequence in publicly accessible databases is transforming the strategies adopted to investigate biological and medical problems.



Left Main entrance to the Sanger Institute.

Below Human chromosomes.



INTRODUCTION





Overview	08	Genetic variation and disease	22
Sequencing	12	Model organisms	26
Data mining	16	Pathogen genomes	30
High-throughput analysis	20	Genomic infrastructure	32

From sequence to biology



Overview

The Sanger Institute's reputation has been based on its large-scale, high-quality genome sequencing projects. While these remain core to the Sanger Institute's work, it is equally committed to the development of tools to analyse and annotate genome sequence data, and to biological studies of gene action in living systems.

The interplay between these areas will provide unparalleled insight into the role of genomes in health and disease, and lay the foundations for new diagnostics and therapeutics based on an understanding of human health at a molecular level.

Sequencing

The Sanger Institute has made major contributions to the mapping and sequencing of the genomes of a host of organisms, including the human genome and those of yeast, more than 20 pathogens and the nematode worm *Caenorhabditis elegans*. [Pages 12,30.](#)

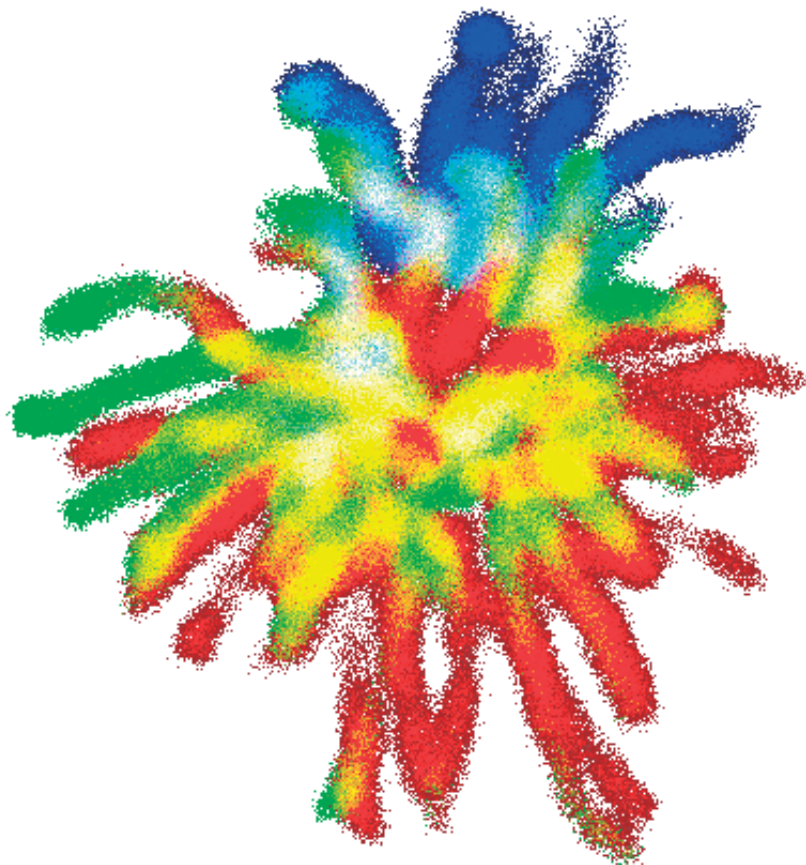
The Sanger Institute has been a major partner in the international programmes to map and sequence the mouse genome. It is also sequencing the entire zebrafish genome, which will be completed by 2005.

Data mining

Data pour continually from the Sanger Institute's sequencers: every day around 60 million bases of raw sequence data are generated. The rate of sequencing output has increased about fourfold every year ([see page 17](#)), and all the data are collected, organized and managed by the Sanger Institute's formidable computing systems.

Sophisticated software and databases keep the data tracked and organized, and underpin efforts to identify genes and other sequence features, and compare the human genome to those of other organisms. Results are analysed and presented to end users through websites such as Ensembl and Pfam. [Pages 18.](#)

As well as these computational approaches, Sanger scientists also collect experimental data to test computer predictions and further refine annotation of the genome. [Page 19.](#)





High-throughput analysis

Computational approaches can provide important clues to gene function, but need to be complemented by experimental studies if a full and accurate picture of a gene's activity is to be obtained.

Traditionally, researchers have worked with individual genes of interest, but the completion of genome projects and the development of new high-throughput tools have opened up new opportunities to explore gene function on a grand scale.

Two key techniques in use at the Sanger Institute are DNA microarrays and gene expression atlases – maps of mRNA and protein expression in living tissues. [Pages 20–21.](#)

Genetic variation and disease

The complete, annotated version of the human genome sequence provides the basis to examine both the extent of sequence variation in human populations, and how specific variants contribute to disease. [Page 22.](#)

The Sanger Institute is working to identify genetic variations such as single nucleotide polymorphisms (SNPs) and blocks of SNPs inherited together (haplotypes) ([see page 24](#)). These and other studies underpin attempts to identify particular genetic variants associated with common diseases.

Another area of interest is variation in the major histocompatibility complex, the area of the genome controlling the immune response and our defence against disease-causing organisms. [Page 24.](#)

Finally, a major initiative, the Cancer Genome Project, is systematically searching all human genes for genetic variations implicated in cancer – a quest that has already unearthed new cancer genes and potential new therapies.

[Page 25.](#)

Model organisms

The availability of the genome sequences of widely studied model organisms has enhanced their value as tools to help understand the function of genes in living systems.

Well-established projects at the Sanger Institute include studies in yeast and the nematode worm. [Page 29.](#)

A new project has been established to study the zebrafish as a model organism, which is ideal for genetic and developmental analyses. [Page 29.](#)

The most powerful model for understanding human biology and disease is the mouse, and a major focus of future studies will involve mouse model systems. [Page 26.](#)

Above Development of the zebrafish eye. The genome of zebrafish is being sequenced at the Sanger Institute and researchers are studying early development.

Left Chromosomes in a dividing cancer cell. The Cancer Genome Project is systematically searching for genes that cause cancer.

Our purpose is to further the knowledge of genomes, particularly through large-scale sequencing and analysis.



Pathogens

Since its inception in 1996, the Pathogen Sequencing Unit at the Sanger Institute has sequenced more than 20 pathogen genomes. It is currently working on some 40 other bacterial genomes and those of important eukaryotic pathogens. [Page 30.](#)

The sequencing and analysis teams in the Pathogen Sequencing Unit also work on important model organisms such as yeasts and microbes of medical importance. [Page 30.](#)



Genomic infrastructure

Sanger Institute research teams rely on world-class core facilities to support genome sequencing, analysis and research. [Page 32.](#)

A major factor in the Sanger Institute's success is a sophisticated IT set-up, with plans afoot for further major upgrades. Central support services also provide specialist input into the Sanger Institute's experimental work – including expert assistance with gene mapping and finishing, genotyping, exon sequencing, and microarraying.

Major new resources are currently being developed on site to support studies on the mouse.

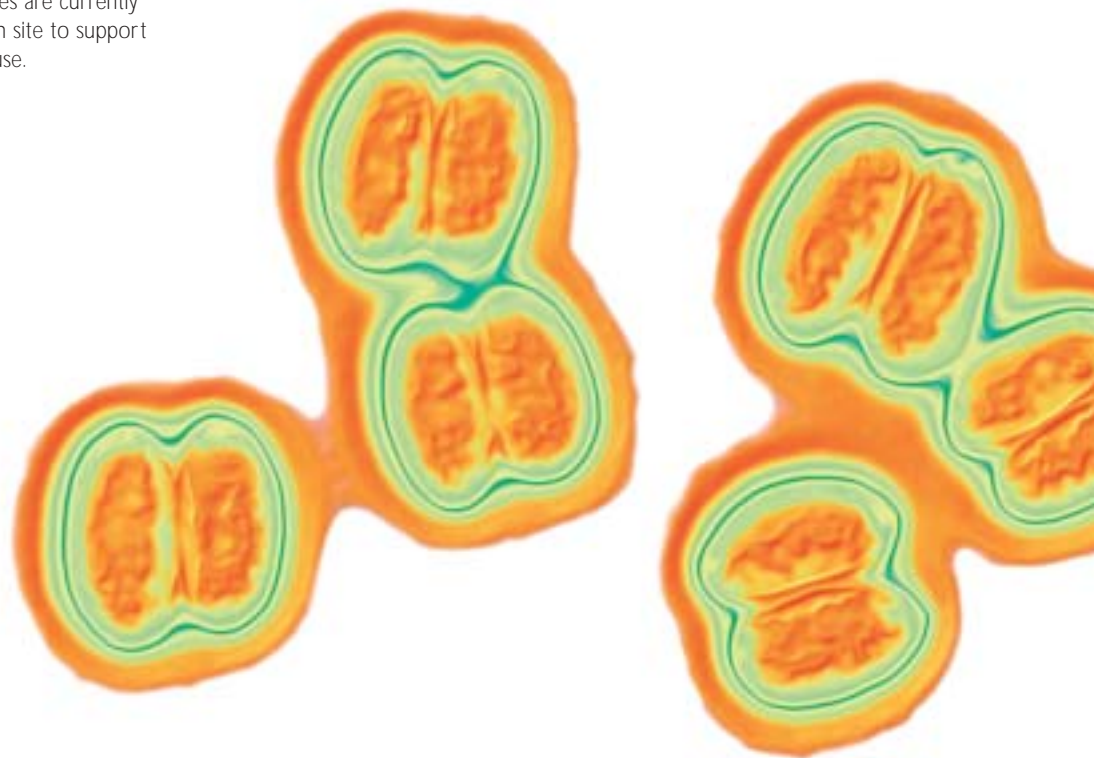
Facilities

The Sanger Institute benefits from a comprehensive library and information service, and first-class administrative and technical support. It even has its own in-house expertise in robotics and automation development, staff who work with research teams to maximize throughput and productivity. [Page 64.](#)

Its social facilities include a cafeteria, staff restaurant, and a sports and social club.

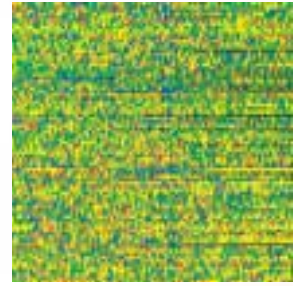
New investigators

To prime its new programmes, the Sanger Institute is recruiting talented researchers at the postdoctoral, fellow and senior fellow levels ([see page 62](#)). The teams created through this recruitment exercise will work with existing staff to develop multidisciplinary projects that utilize the ever-widening range of core skills and resources.



Right
Neisseria meningitidis,
a pathogen
sequenced by the
Sanger Institute.
SPL

Sequencing



From its founding in 1992, the role of the Sanger Institute has been to promote the understanding of living organisms through the determination of the complete DNA sequence of their genomes, and in particular to translate the findings with model organisms, such as yeast and the nematode worm *Caenorhabditis elegans* to the human genome. With this information, it would then be possible to undertake the systematic and painstaking work of understanding how the instruction manual is used to generate and control life itself. Over the past ten years, this concept has been translated into reality for a few of the best studied organisms.

One of our precepts is that the biological community benefits from as rapid release of sequence data as possible: it is a tremendous challenge to establish protocols to ensure this release is as error-free as possible and can be updated as improved sequence is generated.

Individual sequence traces and early assemblies of sequence are released to public databases where they can be accessed and searched. Draft whole genome assemblies have also been released for the human, mouse and zebrafish genomes and automated annotation of these assemblies is provided via Ensembl and similar genome browsers.

The Sanger Institute has established a high-throughput approach to sequencing genomes, and generates approximately 40 million reads of raw data per year. The process has been broken down into a pipeline of tasks that are undertaken by specialist teams and the flow of sequencing projects through the various stages is monitored by an Oracle-based tracking database.

The large-genome projects in which the Sanger Institute has participated are shown on page 15. Although draft sequences are enormously helpful to bench scientists, they contain many gaps and many problems with sequence assembly and resolution of duplicated

regions are not resolved until the sequences are finished. It has been a high priority to increase the efficiency of finishing and the Sanger Institute currently leads the world in generating over 600 million bases of finished sequence per year.

In addition to sequencing genomes, the Large Genome Sequencing Division also has projects underway to identify and map variations found in the genomes of individuals on to the reference sequences and to map clone libraries derived from different strains on to the reference genomes. Expressed sequence tag collections from frog and chicken are also being sequenced which will assist gene annotation of whole genomes.

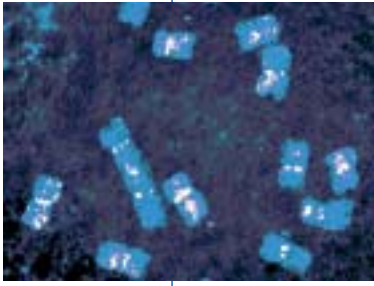
In all these projects, our goal is to provide the foundation of genome sequence and genomic variation that will help researchers interpret better the changes associated with human and animal disease.

The role of the Sanger Institute is to promote the understanding of living organisms through the determination of the complete DNA sequence of their genomes.

Far, left *C. elegans*,
the nematode
worm sequenced
and studied at the
Sanger Institute.
SPL

Left Output from
a DNA sequencing
machine.





FROM SEQUENCE TO BIOLOGY



Sanger sequencing

The methodology used for sequencing DNA is based on the chain termination reaction using dideoxynucleotide bases described by Fred Sanger and colleagues in 1977, for which Sanger was awarded his second Nobel Prize for Chemistry in 1980.

Fluorescent dyes are used to label the DNA fragments and automated DNA sequencers 'read' the sequence of 96 samples at a time. However, even the most efficient DNA sequencers can only read the sequence of between 500 and 1000 bases from each reaction. Thus, to determine the sequence of genomes that are between a few million bases (such as bacteria) and several thousand million bases in length (such as human at 2.9 gigabases or mouse at 2.5 gigabases) requires the sequencing of many thousands or millions of DNA fragments.

Sufficient sequence reads are generated so that they can be identified and assembled with a high degree of confidence. Each base is sequenced on average six to ten times. The process of generating random sequence fragments is called 'shotgun sequencing', derived from the random nature of the process of generating the fragments by breaking up the genome using a mechanical shear or sonication process.

The fragment collection is then cloned into M13 bacteriophage or plasmid vectors for subsequent replication and manipulation.

Assembly of the random shotgun sequence fragments results in extensive regions of high-quality sequence, but also gaps arise because of cloning bias, regions of repetitive sequence that cannot be assembled unambiguously, or sequence motifs that disrupt the sequencing enzymes. The term 'finishing' has been coined for the process of obtaining sequence across gaps and resolving ambiguities in the sequence with no more than one error per million bases, and is essential for production of the highest quality archival genome sequences.

Two basic strategies are used to sequence genomes: whole-genome shotgun sequencing or shotgun sequencing of pre-mapped clones, an approach often referred to as 'hierarchical shotgun sequencing'. These approaches can be used alone or in combination. A combined approach has the advantage of rapidly generating sequence over the entire genome and, at the same time, generating mapped resources that are used for finishing and biological experiments. A hierarchical approach was used to sequence the human genome and we are currently employing a combined strategy to generate finished genome sequences of the mouse and zebrafish.

Sequencing

WGS: Whole genome shotgun
 YAC: Yeast artificial chromosome
 BAC: Bacterial artificial chromosome
 PAC: Phage artificial chromosome

Table footnotes

¹ The *C. elegans* Sequencing Consortium (1998)
 Genome sequence of the nematode *C. elegans*:
 a platform for investigating biology.
Science 282: 2012–18

² www.sanger.ac.uk/Projects/C_briggsae

³ International Human Genome Sequencing Consortium
 (2001) Initial sequencing and analysis of the human
 genome. *Nature* 409: 860–921

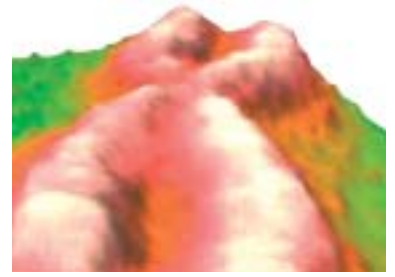
⁴ Mouse Genome Sequencing Consortium (2002)
 Initial sequencing and comparative analysis of the
 mouse genome. *Nature* 420: 520–562

⁵ www.sanger.ac.uk/Projects/D_rerio

GENOME	SIZE	APPROACH	SANGER CONTRIBUTION	COMPLETION DATE
<i>C. elegans</i> (worm)	100Mb	Hierarchical shotgun (mapped cosmids + YACs)	50%	December 1998 ¹
<i>C. briggsae</i> (worm)	90Mb	WGS	50%	September 2001 ²
<i>H. sapiens</i> (human)	2.9Gb	Hierarchical shotgun (mapped BACs/PACs, YACs)	30% (chromosomes 1,6,9,10,13,20,22,X)	Draft: June 2000 Finished: April 2003 ³
<i>M. musculus</i> (mouse)	2.5Gb	WGS + hierarchical shotgun (mapped BACs)	20% (chromosomes 2,4,11,X)	Draft: June 2002 Finished: 2003 ⁴
<i>D. rerio</i> (zebrafish)	1.5Gb	WGS + hierarchical shotgun	100%	Draft: April 2003 Finished: 2005 ⁵







Data mining

Mining large genomes – computational approaches

The problems:

difficult to find genes, large data sets

The human genome sequence is 30 times larger than that of the worm *C.elegans* and yet contains less than twice as many genes. The worm genome itself is ten times larger than yeast with less than three times as many genes. Decoding the function of DNA sequence is never easy, but becomes progressively more difficult as genomes get larger. More of the sequence is historical baggage – non-functional copies of genes (pseudogenes) and remains of virus sequences – making it harder to identify the real genes. We have not yet fully understood the language of the sequence that specifies where genes start and stop. As a result our computer-based interpretation is imperfect and largely depends on the knowledge of external experimental data. It can also frequently be improved by expert annotators.

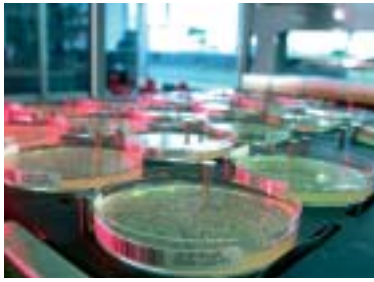
Size also brings its own problems. Genomes may be packaged into a few chromosomes of continuous DNA sequence, but these are too large to be processed efficiently by computers. It is necessary to break them into fragments, analyse each and then reassemble the results. For such large amounts of data the book-keeping problems alone – making sure every fragment has been analysed completely – are huge.

Processing large genomes has become a systems engineering problem, where thousands of computers need to be coordinated in order to analyse a genome rapidly. Rapid analysis and instant availability of the results is required to satisfy the hundreds of thousands of scientists worldwide who are basing their experiments on genome sequence data. Analysis needs to be regularly repeated to take advantage of the ever-increasing amount of external experimental data that can be linked to the genome sequence, each with the potential to lead to new insights.

In addition, all organisms are related to each other by evolution, and the comparison of genome sequences has the potential to unify biology at the molecular level. Great insight can be gained when it is realized that two genes in different organisms are equivalent and that the experiments carried out on each are investigating equivalent phenomena. Constructing lists of these gene 'orthologs' requires the comparison of whole genome sequences, which is another vast computing job.

Above Scanning probe image of human chromosome 1, one of the chromosomes sequenced by the Sanger Institute.

Decoding the function of DNA sequence is never easy, but becomes progressively more difficult as genomes get larger.



Top, left
Large-scale growth of bacterial colonies.

Top, right
The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is next door to the Sanger Institute and collaborates on projects such as Ensembl.

Right Structure of anthrax lethal factor, part of the anthrax toxin. Information on this protease can be found in the MEROPS database.

The solutions:

genome comparisons, Ensembl

Ensembl (www.ensembl.org) provides annotation of complete genomes through an automated analysis pipeline and gene building system. Genomes currently annotated through the Ensembl system are human, mouse, rat, *Anopheles* and *C. briggsae*. Annotation from other projects (such as *Drosophila*, *C. elegans*) is imported into the system and can be used in similar ways to the home-grown product. As well as serving data on multiple genomes, relationships between genomes are calculated and can be viewed from the chromosomal scale to the DNA level, along with associated lists of putative orthologous genes.

The Ensembl website integrates its gene predictions with information from external databases and provides interfaces for data mining and data export. Users can view their own data alongside Ensembl annotation, either by setting up their own DAS server or by simply uploading a file. A variety of text and sequence search facilities are also provided.

Ensembl is a joint Sanger/EMBL-EBI project. It is also a free, open-source software project, allowing anyone to become involved in code development. Academic and commercial groups are equally free to download the entire system and data either to set up a local mirror or apply the system to their own data. The openness of the Ensembl project, with its community discussion lists and collaborative nature, has contributed to its wide adoption.

Gene hunting in Havana

The Human and Vertebrate Annotation and Analysis (Havana) team is analysing the finished sequence produced at the Sanger Institute and trying to find new genes. In the area of informatics there are many ways to predict genes. *Ab initio* programs like Genscan or Fgenesh use rules obtained from training on a known data set of characterized genes to make ORF predictions. Another approach is to use homology data, that is cDNA, EST or protein sequences which match the genomic sequence, and this is probably the better method of making predictions. However, some full-length cDNAs are difficult to isolate and therefore we need to rely on non-perfect homologies to proteins and anonymous ESTs to try to build up a gene.

All these techniques can be used automatically to predict transcripts, as in the Ensembl project, and whole genomes can be analysed within a few days. However, automatic annotation has its drawbacks – currently pseudogenes are not predicted and differentiating alternative splicing remains a difficult area. Manual annotation provides the only reliable method to fully annotate the finished human genome to a high quality. The Havana team is responsible for annotating 40 per cent of the human genome. In collaboration with the experimental gene annotation group it is experimentally confirming predictions, to produce a 'gene set' for each chromosome. Although a slow process, in collaboration with the other sequencing centres, an accurate gene set for the whole human genome should be available by the end of 2003.

Pfam

In recent years it has been realized that the millions of proteins in nature can be organized into just a few thousand protein families. The Pfam database (www.sanger.ac.uk/Software/Pfam) is an attempt to organize proteins into a library of protein families, providing a 'periodic table' of biology. It consists of a large collection of alignments and profile-HMMs, currently amounting to over 6000 families that match to over 70 per cent of known proteins.

Pfam is used around the world to help annotate complete genomes such as the worm, fly and human genomes. It is also used by many biologists to help understand their proteins of interest in terms of their domain organization.

Recently the Pfam group has produced a new database of RNA families called Rfam (www.sanger.ac.uk/Software/Rfam). This database is the first large-scale database for RNA families and is already helping to discover novel RNA genes from our sequencing projects.

Proteases are necessary for the survival of all living creatures, and are encoded by about 2 per cent of genes in all kinds of organisms. They are an exceptionally important group of enzymes in biology, medical research and biotechnology. The MEROPS database (merops.sanger.ac.uk), which has recently moved to the Sanger Institute, is an integrated resource providing many kinds of information on proteases to scientists in academia and industry.

Data mining

Mining large genomes – experimental approaches

The availability of multiple complete genome sequences provides an extremely valuable reference for researchers trying to understand the role of genes in health and disease.

The availability of multiple complete genome sequences provides an extremely valuable reference for researchers trying to understand the role of genes in health and disease. Information on gene structure, evolution and family relationships can be extracted and predictions of biochemical function can be made through sequence comparisons. This detailed structural annotation in turn provides the basis to explore new approaches for large-scale functional annotation of genomes.

Experimental gene annotation

Initial computational work provides a set of high-confidence partial or putative gene structure elements (such as exons, introns and untranslated regions) which can then be confirmed or rejected on the basis of additional experimental data. The experimental annotation group works in close collaboration with the *in silico* team (Havana) to confirm and extend their results (page 18). Initially the goal is to determine a complete coding sequence for each gene in the genome, in concert with the establishment of a full-length cDNA clone. Knowledge of the full range of human gene structures will require annotation of every promoter, polyadenylation site, alternatively spliced exon, and regulatory element in each transcription unit. For both levels of annotation, the acquisition of high-quality experimental data confirming gene features will have the added benefit of providing important validated test datasets for further refinement of sequence analysis and prediction tools.



High-throughput analysis



Many different types of experimentally derived information are required to truly understand gene function. Examples include biochemical activities, the position of a gene product within pathways, its interacting partners or substrates, time and location of expression, response to cellular/environmental changes, the cellular or physiological consequence of interfering with expression and the effect it has on expression of other genes.

Historically this integrated information set has been created by researchers focusing on individual genes of interest. The availability of complete genome sequences and recent technological developments provide opportunities for more systematic and coordinated approaches to capturing and sharing this information. To capitalize on these opportunities, the Sanger Institute has developed new gene expression initiatives in the areas of DNA microarrays and gene expression atlases.

DNA microarrays

The analysis of global gene transcript expression patterns is a key new area of functional genomics because the physiological status of a cell, tissue or organism is, to an appreciable extent, directed by the entire complement of genes expressed therein. Furthermore, changes in almost all biological systems are likely to be accompanied by alterations in the abundance of transcripts for at

least some genes. The ability to use expression microarrays to survey the transcriptional profile for all genes expressed in a cell type or tissue, and to reveal differences in mRNA transcript levels between two different cell or tissue types, has now firmly been demonstrated to be a powerful and direct way of using sequence data for experimental functional studies.

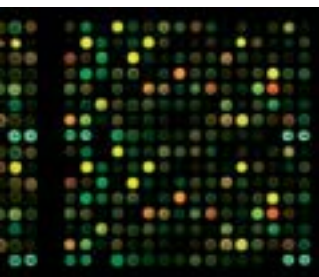
Expression microarrays consist of high-density arrays of cDNA sequences from genes attached to a solid support matrix that is typically the size of a standard microscope slide, or smaller. Fluorescently labelled cDNA derived from cells or tissues of interest are hybridized to the array, and the resulting hybridization events are measured and quantified to reflect the transcriptional levels of genes expressed in the sample.

Knowledge of these transcriptional changes can provide insights into the genetic regulation of biological systems and the functions of genes. For example, genes whose transcriptional profiles behave similarly in single or multiple microarray experiments are likely to be involved in similar biochemical pathways and be controlled by similar regulatory mechanisms in the cell. Furthermore, the levels of individual transcripts or groups of transcripts may also be indicative of abnormalities in cellular pathways associated with disease. Thus, expression microarrays are also having an increasingly significant impact on the prognosis and treatment of disease.

The Sanger Institute is both manufacturing and using large numbers of high-quality expression microarrays in a coordinated effort to provide large amounts of experimental evidence for the function and regulation of genes. This work covers a broad range of organisms including human, experimental model organisms such as mouse, rat, the nematode worm, and yeast as well as pathogenic microorganisms important to human health.

Gene expression atlases

While microarray approaches permit quantitative measurements on many thousands of genes simultaneously, using relatively small sample sizes, information on the pattern of expression within cells in a tissue is lost. Microscopy-based methods such as *in situ* hybridization examine lower numbers of genes but permit the detection of mRNA directly within tissue sections. Labelled DNA or RNA probes are hybridized with the tissue (as with microarrays) resulting in signals which can be detected microscopically at sites of expression. This gives a direct view of the state of gene activation within the tissue architecture. The pattern of *hox* gene expression during embryogenesis provides a beautiful example of differential gene expression preceding and guiding cellular differentiation during development.



Although measurement of mRNA expression may suggest the levels of protein expression, there is not a total correlation between mRNA expression and the levels of the encoded protein which accumulate in cells (due to differential translation, stabilization or degradation). In addition proteins can be modified, transported and targeted to different cellular locations. Since proteins ultimately drive function, knowledge of protein expression levels, modifications and sites of action may have a more direct correlation with phenotype than mRNA levels. The technique of immunohistochemistry, where antibodies are used as probes to detect the presence of proteins in tissue, allows this important spatial information on proteins to be captured.

Expression atlases and DNA microarray databases can thus provide high-quality complementary information on expression profiles to the worldwide research community. The Sanger Institute is active in developing these areas. The internal consistency and efficiency achieved by executing such initiatives in a single location, or across locations using optimized and standardized procedures, will add to their value. Establishing where and when a gene is expressed is, however, only part of the puzzle of determining function. This is a puzzle of many pieces and other types of data from other techniques are needed to truly understand gene function.

Above, left Microarray studies reveal changing patterns of gene expression.

Above *In situ* hybridization shows where particular genes are expressed (in this case the *Sema3A* gene in the developing spinal column and surrounding tissues of the normal chick embryo).



Genetic variation and disease

Variation and disease

DNA is a dynamic molecule which, over time, accumulates changes in the form of insertions, deletions or single base substitutions. These changes can be spread through the population via the cellular mechanisms of heredity. This process is influenced by natural selection: specific genetic variants may confer an advantage or disadvantage on the individual for survival in a particular environment. Survival of the individual in turn leads to contribution of those genetic variants to the gene pool in subsequent generations.

Additional variation arises as a result of recombination events, which re-shuffle the DNA segments and give rise to an ever-increasing number of new combinations of previously existing variants.

In evolutionary terms, *Homo sapiens* is a young species with little genetic variation. Most changes occur in DNA sequence with no apparent function (for example in regions between genes, or in introns). A small fraction of variants occur in functional units of the genome (such as protein-coding exons, or gene regulatory elements) and some of these have a biological effect. These are the variants which constitute the genetic component that governs our genetic individuality. They influence the way we look, our risk of developing disease,

and how each one of us responds to external stimuli. Understanding the molecular basis of genetic diseases and the response to treatment requires the identification of these underlying functional, or 'causative' variants.

The availability of a reference genome sequence has enabled us to start characterizing these sequence variants. If any two randomly chosen copies of the human genome are compared, they differ, on average, in one position every 1200 bp. The most abundant form of sequence variation is the single nucleotide polymorphism (SNP). SNPs are responsible for approximately 90 per cent of all sequence variation, the remainder being insertions or deletions.

Through large-scale sequencing efforts at the Sanger Institute and elsewhere, over 2.2 million unique SNPs have been discovered and deposited in the public domain, and this number is increasing all the time. The bi-allelic nature of SNPs, and their abundance, makes them ideal markers for undertaking genetic analyses on a large scale.

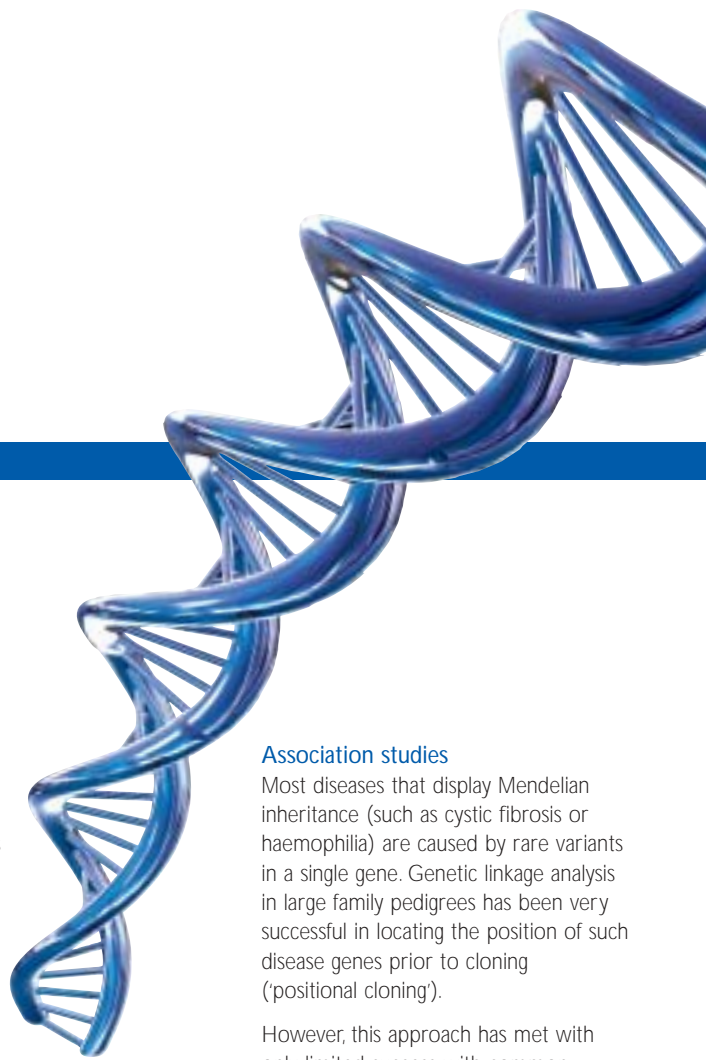
In parallel to this genome-wide effort, there is a systematic study at the Sanger Institute to compile a comprehensive list of all functional variants in the human genome by re-sequencing every exon in the genome in multiple individuals. This valuable resource is supporting disease-oriented and pharmacogenetic projects.

Association studies

Most diseases that display Mendelian inheritance (such as cystic fibrosis or haemophilia) are caused by rare variants in a single gene. Genetic linkage analysis in large family pedigrees has been very successful in locating the position of such disease genes prior to cloning ('positional cloning').

However, this approach has met with only limited success with common diseases such as diabetes, hypertension or asthma. For diseases such as this, the positional cloning approach has been restricted to situations where individual genes exert a strong effect on the disease phenotype and the disease shows a Mendelian pattern of inheritance.

By and large, however, genetic predisposition to common diseases is most likely to involve a combination of rare and common alleles in multiple genes. A powerful way to move forward in this field is to test sequence variants systematically in collections of healthy (control) and affected (case) individuals to establish the association of a particular variant allele with the disease ('association studies').



Genetic variation and disease

Linkage disequilibrium

Optimal use of SNPs as genetic markers requires prior knowledge of their distribution within and between populations. Recent studies suggest that linkage disequilibrium (LD) – the non-random association of nearby SNP alleles – fluctuates across the genome. In some regions there is substantial LD between SNPs that stretches over long distances (50–100 kb or more). These regions are derived from ancestral chromosomal segments that passed almost unmodified from generation to generation due to little historical recombination. As a result, the original haplotypes have been conserved in commonly occurring blocks. Other regions have little or no LD between SNPs, and ancestral recombination has disrupted the original haplotypes. It is important to establish the patterns of LD in the human genome, and to define the common haplotype patterns, so that SNPs can be selected to maximize the effectiveness of disease association studies in every region of the genome.

In August 2002, scientists at the Sanger Institute reported a linkage disequilibrium map of human chromosome 22 in Caucasians ([page 46](#)) and the construction of such a map in multiple populations for chromosome 20 ([page 45](#)) is under way. The Sanger Institute is now expanding this effort to other chromosomes as part of an international project launched in October 2002, to construct a map of LD and common haplotype patterns throughout the human genome (the 'HapMap' Project).

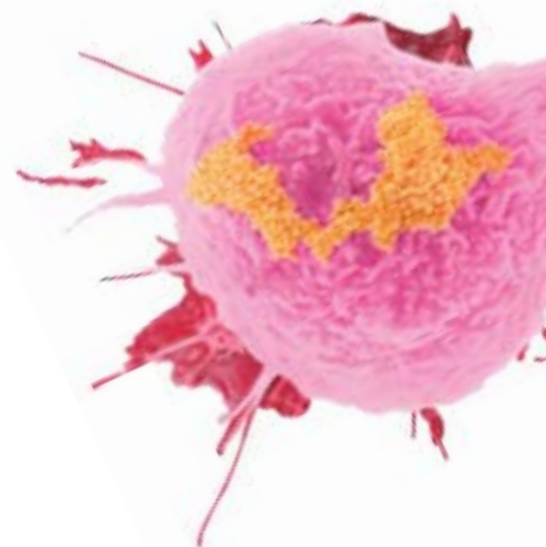
Candidate genes

For many diseases, it is possible to select genes which are particularly good candidates for involvement in a disease phenotype. The selection might be done on the basis of the known function of the encoded protein. For example, variants in the insulin receptor gene might be expected to contribute to diabetes. The candidate gene approach is being adopted for some studies within the Sanger Institute. Genes are selected, and their exons re-sequenced in DNA samples from multiple individuals in order to enrich the catalogue of sequence variants in each candidate gene. The SNPs are then tested for association in larger case-control studies. The advantage of this method is that both rare and common functional variants may be identified and studied. We are using this strategy in a number of our current disease association studies: for example obesity ([page 39](#)), type 1 diabetes ([page 46](#)), type 2 diabetes ([page 39](#)), cardiovascular disease and epilepsy ([page 42](#)).

Major histocompatibility complex

The major histocompatibility complex (MHC) is the most important genetic region in relation to infection and autoimmune disease, such as type 1 diabetes and multiple sclerosis. Driven by pathogen variability, MHC genes have become the most polymorphic loci in the human genome, with some genes (such as HLA-B) having over 500 alleles. By cataloguing all variations between the most common MHC haplotypes, the MHC Haplotype Consortium aims to provide a framework and resource for association studies of all MHC-linked diseases. [Page 41](#).

The Sanger Institute has a commitment to exploit genomic information to advance our understanding of disease and response to treatment with the aim to improve medical care. Understanding human sequence variation is key to this endeavour. Advancing this understanding and providing basic information and new research tools will underpin many new multidisciplinary projects developed within the Sanger Institute in collaboration with other centres.





Variation and cancer

All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, each of which ultimately confers growth advantage upon the clone of cells in which it has occurred. Most abnormalities are acquired through somatic mutation during the lifetime of the individual, but some are transmitted through the germ line and may manifest as inherited susceptibility to cancer.

Identification of the genes that are mutated and contribute to the genesis of neoplasia (referred to as cancer genes) is a central aim of cancer research. Investigation of these genes/proteins

Below Cancerous human cell dividing. The Cancer Genome Project is systematically searching for cancer-causing genes. SPL

forms the foundation of current understanding of biological abnormalities in neoplastic cells. Increasingly, strategies for the development of new therapeutic and preventive agents in cancer are dependent upon modulation of these critical molecular targets. There has been considerable progress in the identification of genes that are mutated in human cancer. However, several lines of evidence indicate that there are likely to be many remaining to be discovered.

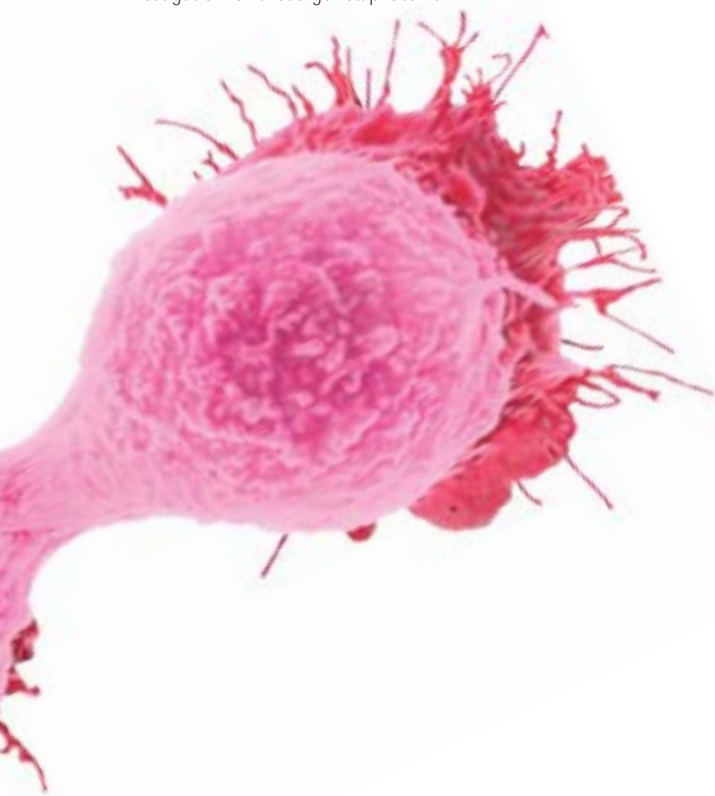
The aim of the Cancer Genome Project (CGP) is to identify somatically mutated genes in human cancer by systematic genome-wide searches using high-throughput mutation detection platforms applied to the template of the human genome sequence. There are several different classes of somatic mutation in human cancer that result in activation or inactivation of cancer genes. These include small intragenic mutations (single base substitutions and small deletions/insertions), large homozygous deletions, gene rearrangements, gene amplifications and epigenetic changes.

The CGP has currently implemented platforms to search systematically for most of these classes of mutation. The major platform we have developed is a capillary-based heteroduplex assay designed to detect small intragenic mutations. We aim to apply this platform

to mutational screening of every coding exon and splice junction in the genome in a set of 48 cancers. We have also implemented platforms that detect homozygous deletions and gene amplifications and will be evaluating novel procedures for the detection of gene rearrangements.

The nature of the output from the CGP is exemplified by the first novel cancer gene that has emerged from these searches. Using the capillary-based heteroduplex assay, we discovered that *BRAF*, a member of the *RAF* gene family of serine/threonine kinases, carries somatic mutations in approximately 70 per cent of malignant melanomas, 10 per cent colorectal cancers, and in 40 per cent of low malignant potential ovarian cancers. The mutations activate the *BRAF* kinase and confer transforming activity in *in vitro* assays. Further studies are now in progress to assess the use of *BRAF* kinase inhibitors as therapies for cancers in which *BRAF* mutations are found.

Meanwhile the search for novel cancer genes continues using the suite of mutation detection platforms. Ultimately, we aim to generate full descriptions of the set of DNA abnormalities present in individual cancers, hence determining the number of causative changes and waves of clonal expansion that are required for the emergence of symptomatic human neoplasms.





FROM SEQUENCE TO BIOLOGY

Model organisms

Top, left The mouse, a model for human biology.

Top, right An early mouse embryo.

Below In the 'gene trap' technique, expression of a marker gene can be detected if it inserts into an expressed gene.

Mouse

The human genome is being intensively scrutinized, but how much functional information can be deciphered from sequence analysis alone? Computational prediction and sequence alignment programmes can already identify many genes and predict some aspects of gene structure. Yet computers remain inferior to cellular machinery in recognizing genes, where they start and stop and when they should be turned off and on.

Computers can also classify genes into families based on conserved motifs. Although these comparisons are useful in predicting some aspect of a protein's biochemical activity or cellular distribution it is very difficult to predict the physiological role of any gene from its sequence. For instance, transcription factors and components of the membrane can be recognized from motifs in the encoded proteins, but it is not possible to predict in which cell type they would be expressed and who their partners might be and the consequence of a mutation.

One of the most informative experimental approaches to examine gene function is to analyse mutants.

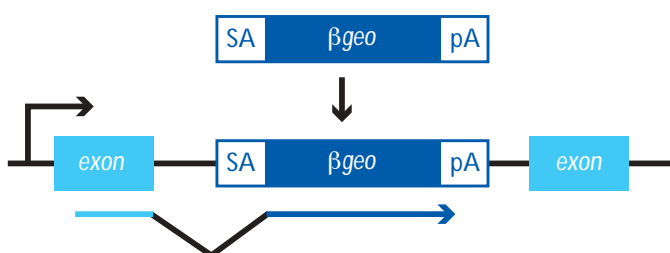
Spontaneous and induced mutations in the mouse have been studied for more than 100 years, but in the past two decades there has been an explosion in their use. The isolation of embryonic stem (ES) cells and demonstration that these cultured cells were capable of re-colonizing the mouse germ line were the two fundamental discoveries that led to generation of the first 'knockout' mouse in 1987, via a genetic modification that had been engineered *in vitro*. Today, it is possible to engineer mice with genetic changes as subtle as a single nucleotide substitution or with major alterations of the genome such as the deletion, duplication or inversion of millions of base-pairs. The genetic tractability of ES cells has made the mouse uniquely accessible for genetic studies compared with every other multicellular organism.

Because the mouse is a mammal it has many physiological, anatomical and metabolic parallels with humans. Although the anatomical differences between humans and mice may appear to be striking, these primarily reflect alterations in size and shape – detailed analysis of organs, tissues and cells reveals many similarities, extending to whole-organ systems, physiological homeostasis, reproduction, behaviour and disease. It is generally accepted that

the mouse is an excellent surrogate for exploring human biology – that disease processes in the mouse can accurately reflect those in humans. Thus the mouse is widely used to investigate diverse aspects of mammalian biology and pathology, ranging from embryonic development to metabolic disease, behaviour and cancer in adults.

Manipulation of genes within the living mouse is routine and can now be done with extraordinary precision. It is now possible to determine the function of each and every component gene in the genome by experimental manipulation and evaluation in the context of the whole organism. And because of the maturity of the technologies by which the mouse genome can be genetically manipulated, these studies can progress very rapidly.

Despite the success of these technologies, the combined output of the mouse genetics community over the past ten years has described mutations in just a few thousand genes, 10 to 15 per cent of the predicted gene content of the mouse. At the Sanger Institute we are investing in approaches to generate mutant mice more rapidly. This includes indexing of libraries of gene-targeting vectors (made possible by the genome





Model organisms

sequence) and gene trapping, in which genes are tagged for sequence retrieval by insertional mutagenesis. Gene trapping generates hundreds of different mutations from a single experiment, though these are in random sets of genes. Over the next few years the Sanger Institute will establish an extensive gene trap resource which will be used for internal programmes as well as being available to the community.

Over the past decade, knocking out genes by targeting and trapping has provided a rich source of information about gene function. But there is considerable uncertainty in predicting the phenotype(s) which will be displayed by the mutant mouse. Sadly, our knowledge of conserved domains, expression patterns, biochemical activity, protein-protein interactions and molecular structure is still inadequate to predict function. A knockout phenotype may shamelessly display our collective ignorance about gene function. In many cases, knocking out individual candidate genes may not efficiently identify genes specific to certain functions or disease – for instance, the genes involved in diabetes.

Genetic screens can identify the players in a specific process and this approach is widely used in other genetic organisms. Recently genetic screens have been conducted in mice, using chemical

mutagenesis. Chemically induced mutants provide a resource for future studies, but the underlying genetic lesions have to be identified so that the molecular mechanisms relating the lesion to the observed phenotype is understood. At the Sanger Institute we plan to conduct small scientifically focused screens in specific areas of biological expertise.

Whatever method is used to generate a mutation, understanding the mechanistic cause of the observed phenotype is central to determination of gene function. This understanding not only depends on knowing the mutated gene, but having a very detailed picture of the phenotype.

Progress in understanding the mouse genome will involve the input of diverse experimental approaches and the analysis of large numbers of mutant mice. Thus we are constructing a state-of-the-art mouse facility to house and evaluate this genetic resource.





Other models

Zebrafish: A model vertebrate for understanding gene function

Zebrafish have become a major focus of biomedical research over the past few years. There are many features that have made them so popular: they are easy and inexpensive to raise and maintain; a single pair of adults can produce hundreds of embryos weekly; embryos are virtually transparent and develop rapidly, allowing one to witness, with a simple dissecting microscope, the development of nearly every organ rudiment during the first 24 hours of development. These qualities combine to make zebrafish ideal for genetic and embryological studies and several systematic screens have identified many thousands of mutations affecting zebrafish development.

Whole-genome sequencing efforts have identified many genes whose function is not known. By microinjection of antisense oligonucleotides into early-stage embryos, zebrafish researchers can disrupt any such gene and study its function during development, offering insight into the biology of all vertebrates. This has already led to the discovery of a variety of gene functions relevant to understanding human disease. [Page 58.](#)

Schizosaccharomyces pombe

The fission yeast *S. pombe* is a unicellular eukaryote that is becoming increasingly popular as a valuable model organism to study a variety of basic biological problems. The genome of fission yeast has been completely sequenced by a European consortium led by the Sanger Institute. Fission yeast is evolutionarily distant from the budding yeast *Saccharomyces cerevisiae* and has dozens of genes (including several disease genes) present in multicellular eukaryotes but lacking in budding yeast.

The fission yeast genome is well annotated and contains some 5000 genes distributed over three chromosomes. In many aspects, such as centromere structure, promoter organization, cell cycle checkpoints, and mechanisms of cell division, fission yeast is closer to higher eukaryotes than budding yeast. Fission yeast is easy to handle, experimental conditions can be strictly controlled, and it is amenable to straightforward genetic manipulations. For these reasons, it provides an attractive complementary model system to the more commonly used budding yeast, and comparative studies in the two yeasts have proven to be fruitful in elucidating central biological concepts such as cell cycle regulation. The availability of the compact and low-complexity fission yeast genome is setting the stage for global expression profiling and other functional genomics approaches, which we expect to give much insight into eukaryotic gene function and regulation. [Page 37.](#)

Caenorhabditis elegans

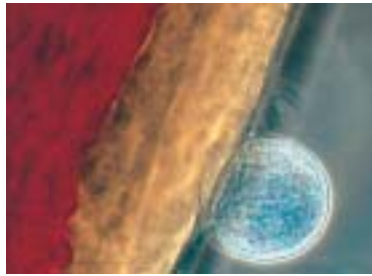
Many animal model systems, such as flies or mice, are very complex – not only do they have vast numbers of cells, but the precise organization of each animal is different. In the worm the picture is very different: adult hermaphrodite worms have only 959 cells which appear in exactly the same arrangement in every animal. This reproducible cell lineage allows us to understand the development and function of the worm at the resolution of individual cells and, furthermore, to carry out very sensitive screens to look for subtle changes in the animal. Such screens have defined roles for many key genes in processes ranging from the regulation of cell death to pathways that affect the lifespan of the animal.

Although the relatively simple development of the worm makes it seem at first sight very different to the bewildering complexity of a human, this is not the case at the molecular level. Over half of all worm genes have a human homologue, including many that are involved in genetic diseases such as cancer. Furthermore, many of the genes identified in screens in *C. elegans* not only look like human genes, but play very similar roles. Understanding how genes work in the simple worm can therefore suggest functions for their human counterparts; these insights can ultimately lead to new therapies and a better understanding of human biology as a whole. [Page 49.](#)

Left *C. elegans*.

Top, left Zebrafish, *Danio rerio*.

Top, right
Fission yeast, *Schizosaccharomyces pombe*.



FROM SEQUENCE TO BIOLOGY

Pathogen Sequencing Unit

Several teams at the Sanger Institute are dedicated to sequencing and analysing the genomes of pathogens and other microorganisms. These sequencing and bioinformatic teams form the Pathogen Sequencing Unit (PSU), which was set up by the Wellcome Trust in 1996 to sequence the genomes of organisms that affect human and animal health. The PSU brings together specialist sequencing techniques and analysis tools to deal with the different genomes of microbes. Much of the sequencing is carried out by whole-genome or whole-chromosome shotguns, which are then assembled into contigs and subsequently pieced together by 'finishers'. Eliminating every gap and/or ambiguous region is often difficult because there can be extremes of composition and repetitiveness which continually challenge current techniques for sequencing and assembly.

The genomes of pathogens are very different from those of their hosts, being much smaller, but with considerably more genes per kilobase. The genomes of bacteria are typically a few megabases in length, encoding a few thousand genes, whereas their hosts have genomes a thousandfold larger yet with only perhaps tenfold more genes. The PSU has developed specialist software to deal with the compact nature of these information-rich genomes and has a team of programmers supporting the specialist needs of pathogen sequencing and analysis. Using this software, microbiologists and parasitologists analyse and annotate every gene in a systematic fashion so that the genome data can be interpreted

with reference to the biology and pathogenesis of the organism. Sequence data are released onto the web as soon as they are produced and into the public databases (EMBL/GenBank/DDBJ) as soon as they are finished, annotated and published.

These fundamental data facilitate and accelerate the applied research of scientists worldwide, leading to greater understanding of disease-causing organisms, and allowing rational searches for, and design of, new antimicrobial drugs and vaccine targets.

Genome completion

Since its inception, the PSU has analysed and published the genomes of the microorganisms responsible for many of the world's most deadly infectious diseases. Our first published genome, in 1998, was *Mycobacterium tuberculosis*, the agent of TB, which infects one-third of the world's population, and is responsible for nearly 1.7 million deaths per year. The genome of *Plasmodium falciparum*, sequenced as part of an international consortium, was published in 2002. *P. falciparum* causes malaria, which kills over one million people a year, predominantly children under the age of five.

Other bacterial genomes that have been published include the causative agents of leprosy (*Mycobacterium leprae*), typhoid fever (*Salmonella typhi*), plague (*Yersinia pestis*), epidemic bacterial meningitis (*Neisseria meningitidis*) and food poisoning (*Campylobacter jejuni*). We have also published the genome of *Streptomyces coelicolor* which, at 8.7 Mb and nearly 8000 genes, is the largest bacterial genome yet sequenced. *S. coelicolor* is a harmless soil bacterium,

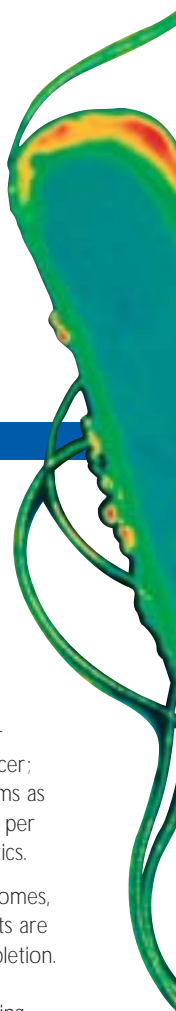
not a pathogen, but is of particular interest as it is an antibiotic producer; the *Streptomyces* group of organisms as a whole produce in more than 75 per cent of naturally occurring antibiotics.

In addition to these published genomes, about 40 bacterial genome projects are currently at various stages of completion. These cover a broad spectrum of human and animal pathogens causing diseases ranging from botulism to whooping cough. Bacterial diversity is immense and many of these projects involve comparisons between different strains of the same species, such as *Escherichia coli* or *Salmonella enterica*, both of which can cause diverse human and animal diseases. Using comparative microarrays it is possible to identify rapidly differences in gene content (for example deletions) between strains and relate these observations to the known pathology or otherwise of the organisms.

Eukaryotic genomes

The PSU also has a major interest in eukaryotic parasite genomes. We are undertaking comparative sequencing of model *Plasmodium* parasites that infect primates and mice in order to identify genes that interact with the host. Many of these genes will be useful vaccine targets. Other Apicomplexan parasites being sequenced include *Toxoplasma gondii* which causes toxoplasmosis and can infect almost all warm-blooded animals, *Eimeria tenella*, which infects chickens, and *Theileria annulata*, which infects cattle.

We are also sequencing the genome of the parasite that causes sleeping sickness in humans, *Trypanosoma brucei*, and a



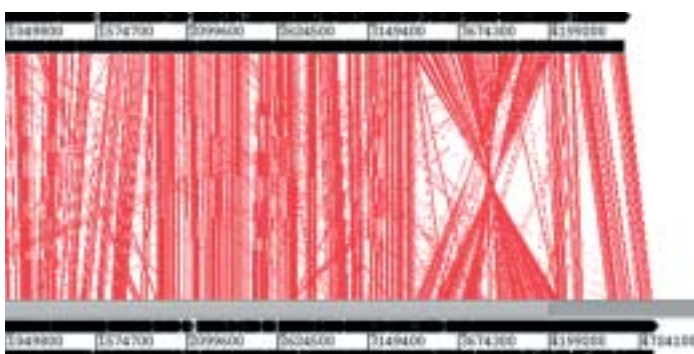
In addition to these published genomes, we currently have about 40 bacterial genome projects at various stages of completion.

Pathogen genomes

closely related parasite, *Leishmania major*, which causes leishmaniasis. These organisms are not only important because of the diseases they cause but are of evolutionary significance as deep-branching eukaryotes and are tractable models for molecular biology and cell biology.

As well as protozoan parasites we are also sequencing multicellular parasites – *Brugia malayi* and *Schistosoma mansoni*, which cause filariasis and schistosomiasis respectively in humans, and *Haemonchus contortus*, which infects sheep. We are also undertaking a large comparative nematode EST sequencing project which should reveal novel vaccine and drug targets and generate new data to provide insights into the evolutionary history of this important taxonomic group.

Alongside pathogens, the PSU is also interested in model organisms. The PSU developed from the team of individuals that made the largest single contribution to the genome sequence of the budding yeast *Saccharomyces cerevisiae* (baker's yeast), sequenced by an international consortium, and published in 1997–8. We have continued this interest by leading the consortium that in 2002 published the sequence of *Schizosaccharomyces pombe*, another yeast, whose cell cycle provides a powerful model for the analysis of human cell cycle control and hence cancer. We are also playing a dominant role in an international consortium generating the genome sequence of the social amoeba *Dictyostelium discoideum* which, aside from being a model eukaryotic organism in its own right, is ideal for the study of the signalling processes that regulate cell behaviour and differentiation during development.



Top, left Malaria parasites in an oocyst on the stomach wall of a mosquito.

Above *Salmonella typhi*, the cause of typhoid fever. SPL

Left A genomic comparison of *E. coli* (upper line) and *S. typhi* (lower line).

Genomic infrastructure

***In silico* resources, systems and informatics**

Analysis of the wealth of genomic data amassed over recent years is dependent on large-scale data storage, data mining and automated computational techniques. From this we will learn more about gene expression patterns, protein interactions, and produce comprehensive annotations and visualizations. In the future simulations will be run 'on chip', a type of computational biological analysis often referred to as *in silico*.

In support of this the Sanger Institute has a leading edge information technology (IT) infrastructure providing *in silico* resources for all of the projects at the Institute.

We have a complex IT installation consisting of large memory (256 gigabytes) multi-processor systems for analysing, comparing and assembling genome sequences. We also have large-scale compute farms with large numbers of commodity computer systems running very high-throughput sequence pattern matching algorithms.

The Institute has around 2000 computer processors and more than 85 terabytes of storage capacity running on gigabit LANs and high-speed Storage Area Networks. Plans are envisaged for petabyte-sized data stores containing data being analysed by tens of thousands of computer processors.

As a net exporter of data, our website takes over 3 million hits per week and demand is growing rapidly. In the future, we will be exploiting high-speed global networks using international compute and data GRIDS to improve global access to the web services and data held on-site.

Developing the science and the IT infrastructure is continuing and in 2005, new laboratory and research facilities and a new state-of-the-art data centre with 1000 m² of computer room space will be opened. This will be one of the largest bioinformatics IT centres in Europe and will house *in silico* resources – compute systems that will rank among the best in the world to maintain our position as one of the world's leading genome centres.

Experimental resources

Researchers at the Sanger Institute have access to a wide range of biological resources, core technologies and experimental expertise.

Genome mapping and clone resources

The Mapping Core Group maintains a wide collection of genomic clones and applies these to the mapping of large genomes. These collections include the RPCI BAC libraries, which are the predominant source of clones for sequencing the human, mouse and zebrafish genomes. The group uses a variety of techniques for map assembly, and specializes in high-throughput restriction fingerprinting to assemble clone maps. The Map Finishing Group coordinates the final stages of genome mapping, during which any remaining gaps are targeted using specialized libraries and cloning procedures. Gaps that cannot be cloned are sized using fibre FISH by the Molecular Cytogenetics Group. These approaches have been applied successfully to the human genome, both for the Sanger Institute mapping projects and on behalf of other genome sequencing centres.

DNA collections and genotyping for the study of human disease

The study of a complex human disease condition typically requires the collection of DNA from several hundreds of affected and control individuals. The recently established DNA Collections Group will be responsible for processing large numbers of anonymized blood samples to produce DNA for



genotyping. For candidate gene studies, the required single nucleotide polymorphisms (SNPs) are identified by the Exon Re-sequencing Group. Whole-genome SNP collections for association and haplotype studies come from public databases and are currently being expanded by further sequencing at the Sanger Institute and elsewhere. The Genotyping Facility provides the technology for high-throughput analysis of the sequence variants in DNA samples. Currently, the facility is based on MALDI-TOF mass spectrometry and can generate up to 70 000 genotypes per day. The facility is also installing another platform, based on fibre-optic bead arrays and with a capacity of 0.5–1 million genotypes per day.

Resources for the study of gene expression and genome rearrangements

High-density microarrays of cDNAs, oligonucleotides or genomic DNAs are powerful tools for the analysis of gene expression and genomic rearrangements. The Microarray Facility makes approximately 2000 such microarrays each month by robotically spotting PCR products onto the surface of glass microscope slides. Human, mouse and rat cDNA arrays representing 15 000, 11 500 and 12 000 genes, respectively, are produced on a routine basis for gene expression studies. For human and mouse, the full-length Mammalian Gene Collection cDNA clones are available. Whole-genome expression arrays are also fabricated for the *S. pombe* and *C. elegans* projects. Genomic arrays of human and mouse BAC clones are generated for comparative genomic hybridization studies to identify

duplications, translocations and deletions. An array comprising 3000 human clones spaced at 1 Mb intervals across the genome is available for comparative genomic hybridization. The Map Finishing Group is assembling a collection of tiling-path clones for the entire human genome and arrays for some chromosomes are already available.

Mouse collections

Mice which carry defined mutations in their endogenous genes or which exhibit specific disease traits are an important resource for discovering the function of genes and modelling human diseases. The Sanger Institute has recently initiated several new programmes which are dependent on state-of-the-art mouse genetic technologies. These include mice with mutations of just a single base-pair to mice which lack millions of base-pairs of genetic information. We envisage that over the next few years we will be able to examine hundreds of different genetic modifications in mice, providing important insights into human gene function. One aspect of this programme includes the construction of facilities to house these strains under optimal conditions. Another aspect of this programme includes generating many unique strains to facilitate genetic analysis using coat colours.

Embryonic stem cell gene-trap resource

A very convenient way of making and storing a collection of mice is as a frozen stock of embryonic stem cells, which can be used to reconstruct mice. We will be producing a large library of embryonic stem cells, each carrying a tagged sequenced mutation. Investigators at the Sanger Institute and elsewhere will be able to access this resource of 200 000 potential mouse strains to generate mice which carry a mutation of their gene of interest.

Archiving of biological resources

The scale of the projects underway and planned at the Sanger Institute – in areas such as protein localization, mouse gene targeting, gene expression analyses in model organisms, and study of human mutation and variation in disease – requires the assembly of vast numbers of biological resources. The Archives Group has been established to manage our collections of oligonucleotides, DNA samples, cell lines, clone collections and other resources, and to enable resource sharing between groups in an efficient manner. The group will also distribute some resources to our collaborators outside the Institute.