

<i>Pombe</i> postgenomics <i>Jürg Bähler</i>	37
Pathogen sequencing <i>Bart Barrell</i>	38
Metabolic diseases <i>Inês Barroso</i>	39
Protein and RNA families <i>Alex Bateman</i>	40
Immunogenomics <i>Stephan Beck</i>	41
Discovering genes involved in complex disease <i>David Bentley</i>	42
Mouse genetics <i>Allan Bradley</i>	43
Molecular cytogenetics <i>Nigel Carter</i>	44
Human genome sequence analysis <i>Panos Deloukas</i>	45
Analysis of human chromosome 22 <i>Ian Dunham</i>	46
Genome informatics <i>Richard Durbin</i>	47
Human disease analysis in the mouse <i>Lorraine A Everett</i>	48
Cell signalling in <i>C. elegans</i> <i>Andy Fraser</i>	49
Prediction research <i>Tim Hubbard</i>	50
Parasite sequencing and pathogen microarrays <i>Al Ivens</i>	51
Cell signalling, cell fate and proteolytic pathways in <i>C. elegans</i> <i>Patricia Kuwabara</i>	52
The atlas of gene expression <i>John McCafferty</i>	53
Bacterial sequencing and analysis <i>Julian Parkhill</i>	54
Bioinformatics <i>Kate Rice</i>	55
X chromosome: biology, evolution and disease <i>Mark Ross</i>	56
Gene trapping the mouse <i>Bill Skarnes</i>	57
Zebrafish genetics and embryology <i>Derek Stemple</i>	58
Cancer genetics <i>Mike Stratton</i>	59
Embryo gene expression patterns <i>David Tannahill</i>	60
Microarrays, transcriptional networks and human disease <i>Dave Vetrie</i>	61

Research interests





Research at the Sanger Institute centres on 'team leaders', who run independent groups focused on particular issues in genomics or the application of genome sequence data to understand animal or human biology.

The following pages provide brief summaries of each team's research interests and key publications. More detailed information can be found on their associated web pages.

Research interests

Pombe postgenomics

Jürg Bähler



In every cell, hundreds of genes and their protein products function together in complex and orchestrated networks that regulate various biological processes. Traditionally, molecular biologists and geneticists study one or a few genes at a time. With the increasing availability of entire genome sequences, DNA microarrays and other postgenomic technologies are now used to monitor the expression levels and interactions of all genes on a global scale. These are exciting times to work in this field, and genome-wide approaches will be crucial to fully understand life and disease.

The importance of differential gene expression is evident from the various cell types in our bodies, which all contain the same genome but differ strikingly from each other in function and appearance, reflecting differences in the use of available genes. We are using the increasingly popular fission yeast *Schizosaccharomyces pombe* as a model organism to understand genetic networks. We are keen to gain an overview of global patterns of gene expression and of regulatory strategies during normal cell growth and in response to unfavourable conditions.

Even the relatively simple yeast cells orchestrate and fine-tune their transcriptional programs in sophisticated ways, both during regular stages of cellular life (cell cycle, sexual differentiation), and in defence against environmental stress. This provides a valuable framework for understanding gene expression programs in more complex organisms. Our comprehensive analyses help to dissect regulatory circuits, and to discover transcriptional cascades and signalling pathways. Moreover, our data give insights into gene function, promoter motifs, and global mechanisms involved in the control of gene expression. We are also hosting collaborators from all over the world who work on a variety of complementing projects. Our website carries more details.

www.sanger.ac.uk/PostGenomics/S_pombe

Wood V et al. (2002)
The genome sequence of *S. pombe*. *Nature* 415: 871–80

Wood V and Bähler J (2002) How to get the best from fission yeast genome data. *Comp. Funct. Genomics* 3: 282–8

Mata J et al. (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nature Genet.* 32: 143–7

Chen D et al. (2002) Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* 14: 214–29

FIGURE 1

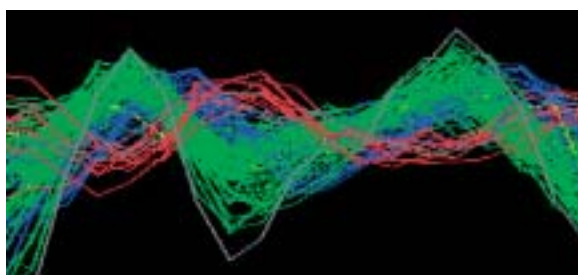
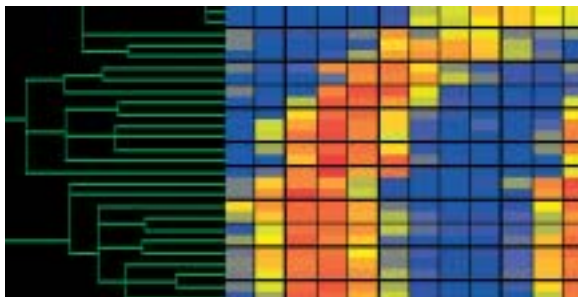


FIGURE 2

FIGURE 1
Gene cluster (top),
periodic cell cycle
transcription (bottom).

FIGURE 2
Fission yeast.

Pathogen sequencing

Bart Barrell



In the Pathogen Sequencing Unit (PSU) we specialize in producing high-quality finished sequence, which is annotated to a high standard. We sequence a diverse range of organisms and work on many genomes at once with projects at many different stages. This is accomplished by a team of project managers who are specialists in their field; they are supported by programmers and specialist annotators who systematically analyse these genomes. A summary of the PSU activities can be found on the pathogen sequencing pages. [Pages 30–31](#).

My own broad interest is in sequencing genomes and understanding how different kinds of information are encoded in DNA. Much of this focuses on predicting genes, their structure, function and organization and making this information easily accessible to the research communities. The programs we develop to analyse genomes are also designed to operate on PCs and Macs as well as Unix and Linux systems. This means that all the information that is accessible to us is also available to researchers at the bench.

We are still at the discovery stage in genome analysis and as we explore new organisms for the first time we can find tremendous variation in genome composition and structure and in the organization of genes. This can sometimes make sequencing difficult but also can

pose considerable challenges to the gene prediction tools that we use, particularly where very little sequence information existed previously for that organism and especially where the organism may only be distantly related to previously sequenced organisms. Here I can use my own experience in gene structure prediction for producing *de novo* training sets of spliced genes for gene-finding programs.

In different genomes we find considerable variation in the proportions of genes that are spliced and in the pattern of splicing, for example in the number of introns, or sizes of introns and exons. Also, there can be considerable differences in the composition of genes that may be affecting their expression. Particularly interesting are mosaics of different kinds of genes and gene structures in a genome. At a simple level these can be highly spliced small genes and very long unspliced genes. Although we can make some simple assumptions it is not known how these are related to, for example, core functions, parasite-specific genes, antigenic variation etc. By exploring different kinds of gene sets by comparative analysis of many different types of genomes we hope to understand the reasons for this variation.

www.sanger.ac.uk/Projects/Microbes
www.sanger.ac.uk/Projects/Protozoa

Cole S T *et al.* (1998)
Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.
Nature 393: 537–44

Gardner M *et al.* (2002)
Genome sequence of the human malaria parasite *Plasmodium falciparum*.
Nature 419: 498–511

Wood V *et al.* (2002)
The genome sequence of *S. pombe*.
Nature 415: 871–80



ACT comparison of sequences in three malarial parasite species – *P. falciparum*, *P. knowlesi* and *P. vivax*. These comparisons highlight differences between the genomes that may be important in disease progression.

RESEARCH INTERESTS

Metabolic diseases

Inês Barroso



The aim of the Metabolic Disease Group is to understand the genetic aetiology of complex diseases such as type 2 diabetes and obesity. The knowledge of genetic predisposition is important to help those found to be at increased risk for these disorders to make appropriate lifestyle choices, for example, increasing exercise, and to avoid risky behaviours such as high-fat diets. It can also lead to the development of new and better drugs that work in each affected individual. We are taking advantage of high-throughput systems to explore genetic predisposition to type 2 diabetes and obesity. This work involves close collaborations with the groups of Professor Steve O'Rahilly (Departments of Clinical Biochemistry and Medicine) and Dr Nick Wareham (Institute of Public Health) from the University of Cambridge.

Type 2 diabetes and obesity are complex diseases that result from gene-gene and gene-environment interactions. These diseases result from the challenge between environmental risk factors and genetically predisposed individuals (Figure 1). Because these diseases are complex, traditional genetic linkage studies have met with limited success.

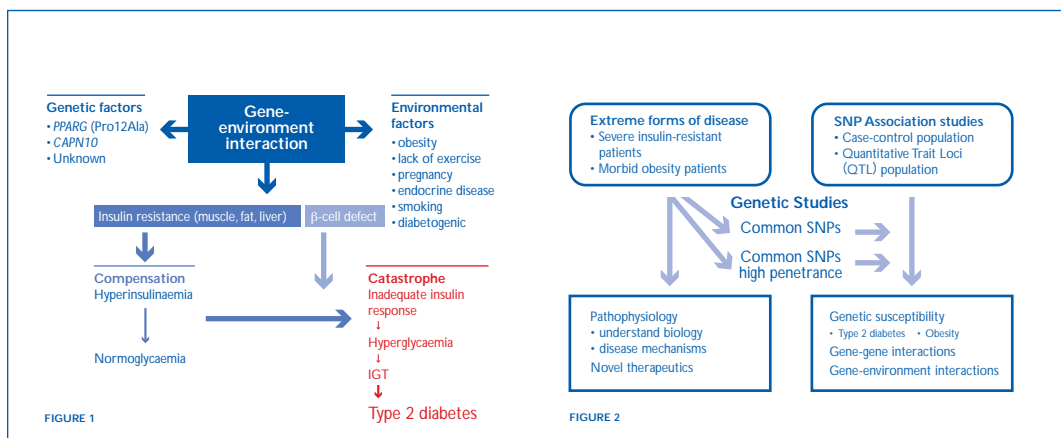
We use candidate gene studies to discover the genetic bases of type 2 diabetes and obesity. Using knowledge of these diseases' pathology and of the molecular pathways involved in glucose metabolism and feeding behaviour, we can make educated choices regarding which genes to study. Our strategy focuses on looking for mutations in these genes in extreme phenotype populations – extreme insulin resistance and extreme obesity cohorts – coupled with testing (genotyping) candidate gene DNA polymorphisms in population-based disease association studies (Figure 2). By studying extreme phenotypes, which tend to be genetically simpler than the more common complex diseases, obesity and type 2 diabetes, we have been able to implicate several genes as important in human severe insulin resistance diabetes. The population-based studies are also yielding insights into gene-gene and gene-environment interactions in metabolic disease.

www.sanger.ac.uk/Teams/Team35

Berger D *et al.* (2002) Genetic variants of Insulin Receptor Substrate-1 (IRS-1) in syndromes of severe insulin resistance. Functional analysis of Ala513Pro and Gly1158Glu IRS-1. *Diabetes Med.* 19: 804–9

Savage D B *et al.* (2002) Digenic inheritance in a family with extreme insulin resistance. *Nature Genet.* 31: 379–84

Barroso I *et al.* (1999) Dominant negative mutations in human PPAR γ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* 402: 880–3



Protein and RNA families

Alex Bateman



Our goal is to infer biological knowledge and accurately transfer this information to large data sets such as complete genomes. The research of the group is currently focused in two areas: protein and RNA family databases – Pfam, MEROPS and Rfam – and discovery of novel protein families.

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 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. The group includes the MEROPS database

that focuses on the classification of peptidases and provides the worldwide standard nomenclature for these proteins.

During 2002, we created Rfam, the first collection of non-coding RNA (ncRNA) families. We use covariance models to find new ncRNA genes in the nucleotide databases.

The classification of novel protein families continues to be a key way of transferring experimental results to new genomic data. This group has published many novel domains such as the SIS and PLAT domains. Our discovery of the PAZ domain predicted that the Dicer protein would be the dsRNA nuclease involved in RNAi some months before this was experimentally demonstrated. We also recently published a novel beta-lactam binding module called the PASTA domain found in bacterial cell surface receptors and the penicillin-binding proteins.

www.sanger.ac.uk/Software/Pfam
www.sanger.ac.uk/Software/Rfam
<http://merops.sanger.ac.uk>

Bateman A et al. (2002) The Pfam protein families database. *Nucleic Acids Res.* 30: 276–80

Rawlings N D, O'Brien E, Barrett A J (2002) MEROPS: the protease database. *Nucleic Acids Res.* 30: 343–6.

Bateman A and Birney E (2000) Searching databases to find protein domain organization. *Advn. Protein Chem.* 54: 137–57

Cerutti L, Mian N and Bateman A (2000) The PAZ domain: A new domain in gene silencing. *Trends Biochem. Sci.* 25: 481–2

Yeats C, Finn R D and Bateman A (2002) The PASTA domain: A beta-lactam binding domain. *Trends Biochem. Sci.* 27: 438–40

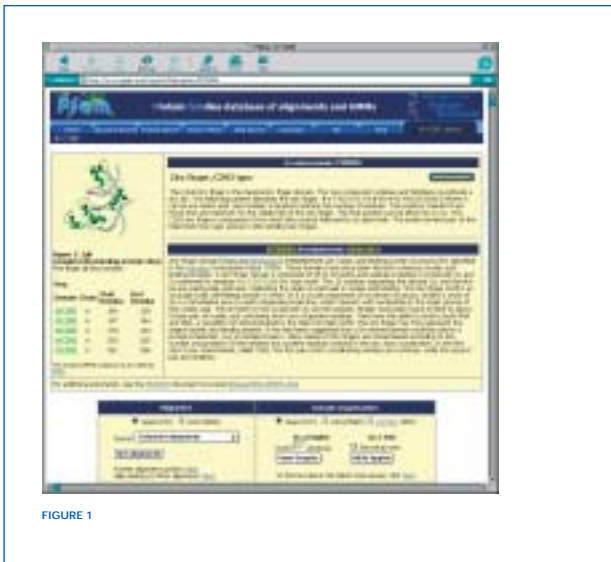


FIGURE 1

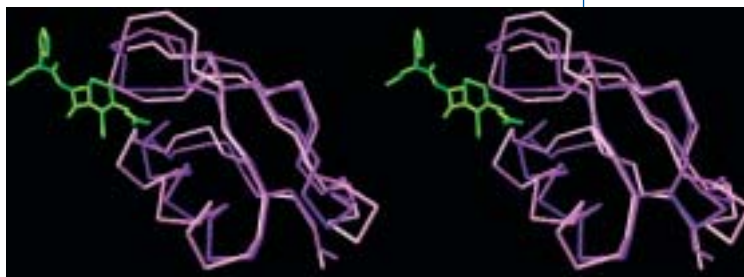


FIGURE 2

RESEARCH INTERESTS

FIGURE 1
The C2H2 zinc finger family at the Pfam website.

FIGURE 2
Stereoview of the PASTA domain binding to a beta-lactam antibiotic (green).

Immunogenomics

Stephan Beck



Our research is focused around the study of the organization, function and evolution of vertebrate defence genes, particularly those encoded by the major histocompatibility complex (MHC) and the leukocyte receptor complex (LRC). Both complexes form integral parts of the immune system and the MHC is the most important genetic region in relation to infection and common diseases such as autoimmunity. Driven by pathogen variability, immune genes have become the most polymorphic loci known, with some genes having over 500 alleles.

The main function of these genes is to provide protection against pathogens and they achieve this through complex pathways for antigen processing and presentation. However, even subtle changes in these pathways can lead to genetic miscommunication and result in disease, particularly autoimmune disease. This genetic balancing act also presents a major challenge to transplant medicine,

where the aim is to minimize the rejection of transplants while not having to compromise the patient's immune system. We employ both experimental and computational approaches in our investigations.

We use genomic sequencing and comparative analyses to characterize regions of immunological interest such as the MHC. So far, about 40 per cent of the identified genes are thought to be involved in immunity, but the function of many genes still remains unknown. By sequencing multiple MHC haplotypes, we aim to identify all functional polymorphisms and to provide a central resource for association studies of all MHC-linked diseases.

www.sanger.ac.uk/HGP/Chr6/MHC
www.sanger.ac.uk/Teams/Team50

Novik K L et al. (2002) Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* 4: 111–28

Younger R M et al. (2001) Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res.* 11: 519–30

Wilson I (2000) Plasticity in the organisation and sequences of human KIR/ILT gene families. *Proc. Natl. Acad. Sci. USA* 97: 4778–83

The MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401: 921–3

FIGURE 1
Some immune genes evolved over 400 million years ago (Mya) in early vertebrates whereas others emerged only fairly recently (80 Mya), highlighting the enormous plasticity of the immune system.

FIGURE 2
For functional analysis we employ microarray (left) and (right) methylation (CH₃) analyses to understand the expression and regulation of immune genes.

FIGURE 3
A gene map of the human major histocompatibility complex.

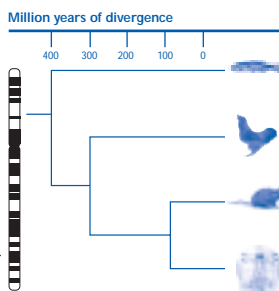


FIGURE 1

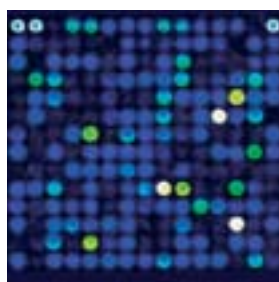


FIGURE 2a

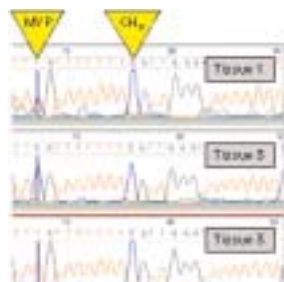


FIGURE 2b

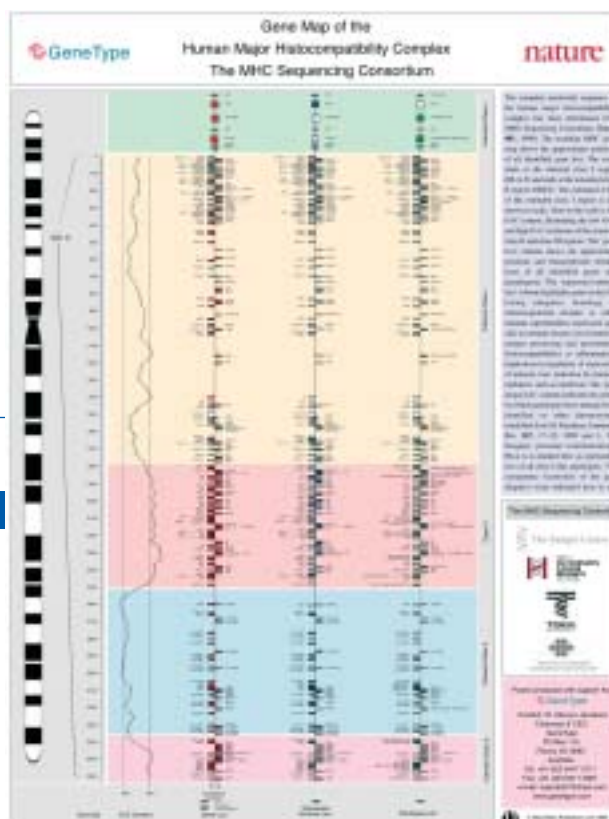


FIGURE 3

Discovering genes involved in complex disease

David Bentley



We are establishing projects to use large-scale genotyping approaches to search for novel genes that contribute to complex disease. Projects will involve the development of large-scale association studies which will test for the association of a specific allelic variant with a specific phenotype. Variants with positive associations can then be followed up to determine their functional significance. Initially this work will use variants in candidate genes, and whole genome scans will be developed in the longer term when information on common haplotype patterns in human populations is generated (see also the section on sequence variation, [pages 23–25](#)). An early focus of our work is the study of sequence variants in genes suspected of involvement in epilepsy and cardiovascular disease, and includes a focus on monogenic examples of related phenotypes.

The future of medicine will benefit enormously from improved knowledge of the genetic background of every individual. An area of particular interest to the group is to investigate the influence of specific sequence variants on the action, metabolism and clearance of administered drugs. Variations in drug response can act via either pharmacodynamic or pharmacokinetic mechanisms. An example of the former is provided by the mechanism of action of benzodiazepines (antiepileptic drugs) where variants in GABA_A receptor subunit gene sequences result in altered pharmacodynamic response to benzodiazepines.

Initially, specific collaborations will focus on discovering variants in genes that are involved in variable response to antiepileptic drugs, and also commonly used drugs such as paracetamol and codeine. Pharmacokinetic mechanisms of action are illustrated by the cytochrome P450 enzymes, which oxidize a variety of drugs and other toxins. For example, cytochrome P450 2D6 alone metabolizes over 80 therapeutic drugs, and variants in the 2D6 locus result in altered rates of metabolism of certain drugs (such as the antihypertensive drug debrisoquine). Population surveys carried out with selected well-characterized variants, as well as finding and testing new variants in candidate genes, will enable correlation of each variant with recorded or measured responses to drugs. The clinical information on drug response being gathered in prospective studies attached to collections such as the Avon Longitudinal Study of Parents and Children (www.alspac.bris.ac.uk), the study of the clinical outcomes and cost effectiveness of standard and new epileptic drugs or selected cancer or cardiovascular disease groups, will allow correlations to be made between specific drug responses and known or novel sequence variants.

www.sanger.ac.uk/Teams/faculty/bentley

Dawson E *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418: 544–8

Deloukas P *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 22. *Nature* 414: 865–71

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928–33

Coffey A J *et al.* (1998) Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* 20: 129–35

Vetrie D L P *et al.* (1993) The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. *Nature* 361: 226–33

Mouse genetics

Allan Bradley



The primary activities of our laboratory involve investigations of the function of genes using the mouse as a model genetic system. We generate and analyse mutations in mice using embryonic stem-cell technology, pioneered and developed by members of my laboratory over the last 20 years. Former graduate students have devised methods to make point mutations, engineer large deletions and chromosomal inversions.

My training philosophy is to provide a supportive laboratory environment that is rich technically as well as intellectually. There are many diverse projects in the group, which cover a wide range of biological questions ranging from embryonic development to cancer and DNA repair. This diversity reflects areas of biology into which investigations on different genes lead us. I view this diversity as enriching because all students and fellows are exposed to different projects and the experimental routes and hypotheses that are unique to different areas of biology. All students in the group are expected to generate and analyse mutations in mice, gaining experience in molecular techniques, ES cell culture, mouse genetics and some aspect of phenotypic analysis. Most projects in the laboratory have a genomic flavour, in keeping with our location at the Sanger Institute.

Current projects include: Genetic screens for genes in the mismatch repair (MMR) pathway using a cell line that lacks the Blm protein which allows us to pursue recessive genetic screens in culture. We are identifying new mismatch repair genes and genes in cell death pathways.

Hypothalamic genomics

The hypothalamus is a major control centre for many central aspects of mammalian physiology but has remained relatively unexplored using genetics. Knockouts of genes expressed in the hypothalamus have been generated and the phenotypes are being evaluated.

Analysis of Blm-deficient tumours

The Blm-deficient mouse offers opportunities to explore areas of the genome which are deleted in tumours. This mouse has an extraordinarily high level of sister chromatid exchange, and loss of heterozygosity. This allows us to map regions of the genome containing tumour-suppressor genes that are deleted in tumours using the technique of BAC array comparative genomic hybridization.

www.sanger.ac.uk/Teams/Team82

Luo G *et al.* (2000)
Elevated mitotic recombination and cancer predisposition in a mouse model of Bloom syndrome. *Nature Genet.* 26: 424–9

Zheng B *et al.* (1999)
Engineering a mouse balancer chromosome. *Nature Genet.* 22: 375–8

Mills A A *et al.* (1999)
p63, a p53 homologue required for limb and epidermal morphogenesis. *Nature* 398: 708–13

Sharan S K *et al.* (1997)
Embryonic lethality and radiation hypersensitivity mediated by *Rad51* in mice lacking *Brca2*. *Nature* 386: 804–10

Jones S N *et al.* (1995)
Rescue of embryonic lethality in *mdm-2* deficient mice by absence of p53. *Nature* 378: 206–8

Ramirez-Solis R, Liu P and Bradley A (1995)
Chromosome engineering in mice. *Nature* 378: 720–4



LEFT
Microinjection of ES cells into a mouse blastocyst.

BELOW
High frequency of sister chromatid exchange in Blm-deficient ES cells.



Molecular cytogenetics

Nigel Carter



My research interests have focused on applications of molecular cytogenetics particularly in relation to human disease, mammalian karyotype evolution and chromosome organization and structure. We have developed rapid methods for the mapping and sequencing of chromosome rearrangement breakpoints: over the next few years we will be sequencing many balanced, *de novo* translocations both from patients and from normal individuals. These methods use the clone resources developed during the sequencing of the human genome together with the power of flow sorting to separate the derivative chromosomes for sequence level analysis. These studies will not only identify genes responsible for the disease phenotypes but also help us understand the underlying mechanisms of chromosome rearrangement.

We also use the clone resources of the human sequence 'golden path' to develop microarrays for the detailed analysis of genomic copy number changes in tumours. By hybridizing tumour DNA labelled with a green fluorochrome together with normal DNA labelled in red onto arrayed clones (comparative genomic hybridization, CGH), analysis of the ratio of red to green on clones along the chromosome identifies regions amplified or deleted in the tumour.

Our initial arrays have a resolution of 150 kb every 1 mb across the genome, but higher resolution arrays and complete clone tiling path arrays for specific chromosomes have also been produced. We are using these arrays to make a detailed analysis of genomic changes in colorectal cancer.

Genomic arrays are also being used to study the biology of chromosomes. Initially we are using the genome-wide arrays and a chromosome 22q tilepath array to study chromosome replication timing.

We have also developed the genomic microarrays to speed up the process of breakpoint mapping mentioned above. By flow sorting a derivative, translocated chromosome and hybridizing onto a genomic array, the breakpoint is mapped to the resolution of the clones on the array in a single hybridization. With high-resolution arrays, breakpoint spanning clones can be identified directly.

www.sanger.ac.uk/Teams/Team70

Carter N P *et al.* (2002) Comparative analysis of comparative genomic hybridization microarray technologies: Report of a workshop sponsored by the Wellcome Trust. *Cytometry* 49: 43-8

McMullan T W *et al.* (2002) A candidate gene for congenital bilateral isolated ptosis identified by molecular analysis of a *de novo* balanced translocation. *Hum. Genet.* 110: 244-50

Fiegler H *et al.* (2003) DNA microarrays for Comparative Genomic Hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* (in press) 36:361-74.

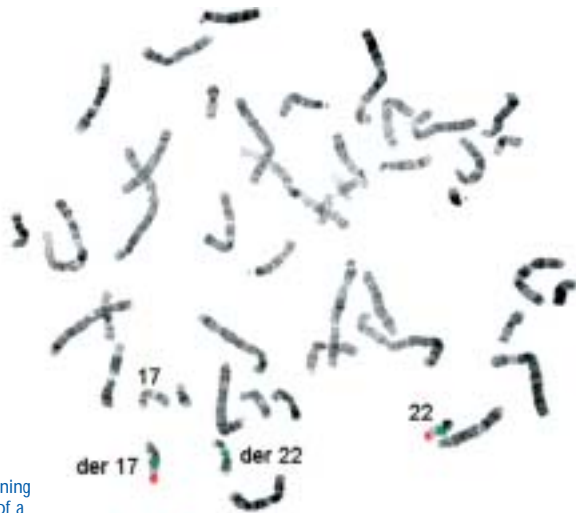
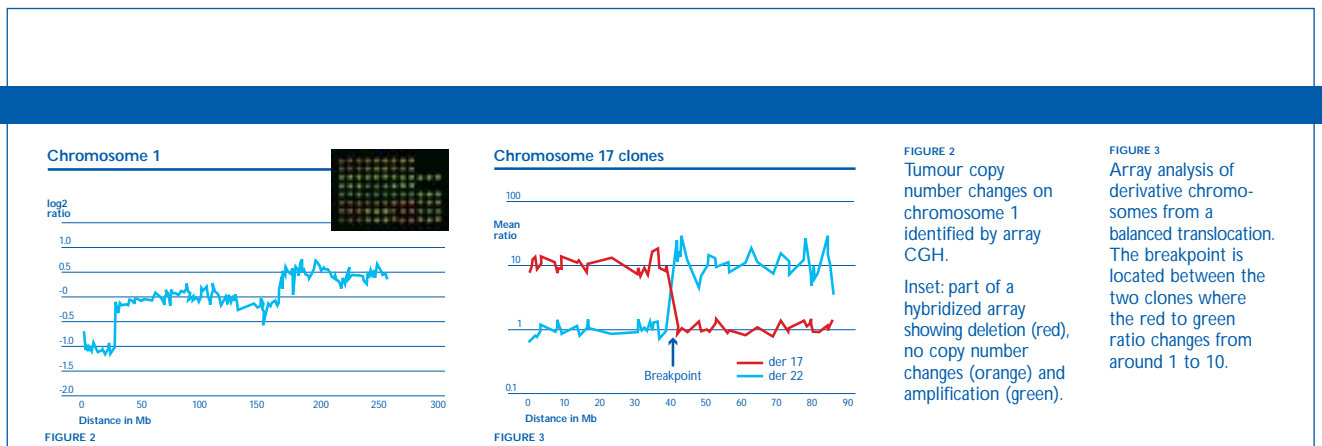
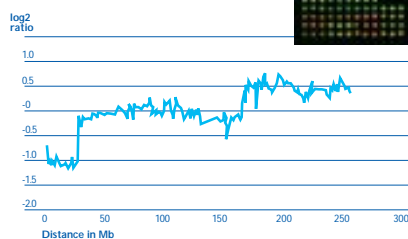


FIGURE 1 BAC clone spanning the breakpoint of a chromosome 17-22 translocation (green signals on chromosomes 22, derivative 22 and derivative 17).



Chromosome 1



Chromosome 17 clones

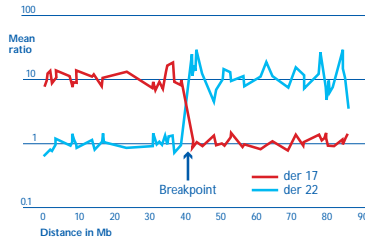


FIGURE 2 Tumour copy number changes on chromosome 1 identified by array CGH.

Inset: part of a hybridized array showing deletion (red), no copy number changes (orange) and amplification (green).

FIGURE 3 Array analysis of derivative chromosomes from a balanced translocation. The breakpoint is located between the two clones where the red to green ratio changes from around 1 to 10.

Human genome sequence analysis

Panos Deloukas



The genome sequence of an organism, annotated with all units of genetic information, is an essential tool in the study of its biology. The challenge to decode our own genome, identify all the genes and their regulatory elements and the use of genetic variation as a tool to dissect the molecular basis of human diseases are the broad research interests of the group. Our participation in the Human Genome Project started by constructing high-resolution physical maps using whole genome radiation hybrid mapping. A map of 30 000 human genes was reported in 1998 providing a valuable resource for both positional cloning of disease loci and anchoring the bacterial clone maps used to sequence the genome. We then went on to construct clone maps of human chromosomes 10 and 20 and coordinate the multidisciplinary effort to sequence and annotate them.

The chromosome 20 sequencing project was completed in 2001. We are expanding the initial annotation of selected regions, increasingly focusing on promoter and other regulatory regions as well as gene expression. We use a combination of experimental and

computational methods, with emphasis on comparative genome analysis. A 10 Mb region of mouse chromosome 2 which is syntenic to human 20q12–13.2 has been sequenced and analysis is in progress. Human 20q12–13.2 has been genetically linked to type 2 diabetes, obesity and Grave's disease. We are also working on the positional cloning of genes involved in eye disorders (corneal dystrophies) linked to chromosome 20.

The chromosome 10 sequencing project was completed in April 2003. We have an active interest in the 10q23–q26 region and are trying to clone various disease genes in collaboration with other groups.

The human genome sequence provides the basis to examine both the extent of sequence variation in human populations and the functional consequences of specific variants in their contribution to disease (page 23). The group is responsible for running a high-throughput genotyping facility (page 32) and research interests encompass studying the extent of linkage disequilibrium across the genome in multiple populations as well as using the haplotype map information to conduct association studies of diseases linked to chromosome 20.

www.sanger.ac.uk/Teams/Team67

Dawson E *et al.* (2002) A linkage disequilibrium map of human chromosome 22. *Nature* 418: 544–8

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

Morante-Redolat J M *et al.* (2002) Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy. *Hum. Mol. Genet.* 11: 1119–28

Deloukas P *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature* 414: 865–71

Bentley D R *et al.* (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* 409: 942–3

Deloukas P *et al.* (1998) A physical map of 30,000 human genes. *Science* 282: 744–6

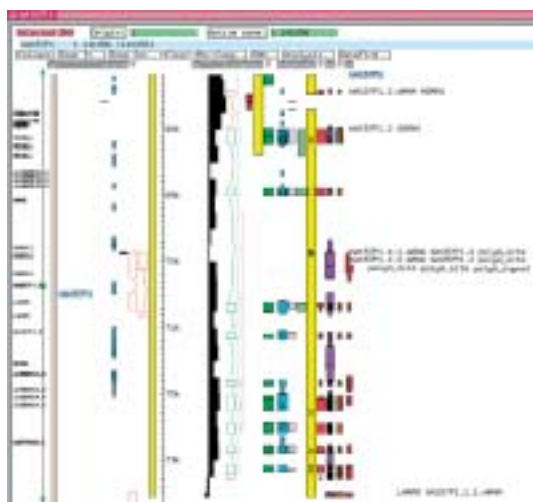


FIGURE 1

RESEARCH INTERESTS

FIGURE 1
Structural annotation of human chromosome 20. Database view (Acedb) depicting the various types of supportive evidence used to draw the exon-intron structure of the gene ADRM1 encoding adhesion regulating molecule 1.

FIGURE 2
Eye of a CHED (top) and a PPCD (below) patient. CHED and PPCD are diseases that affect the cornea.

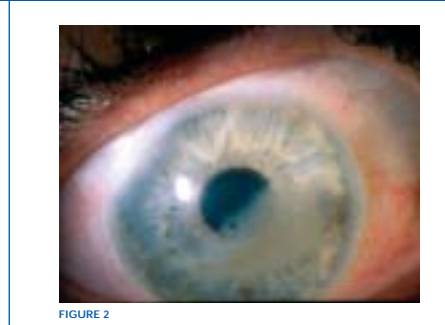
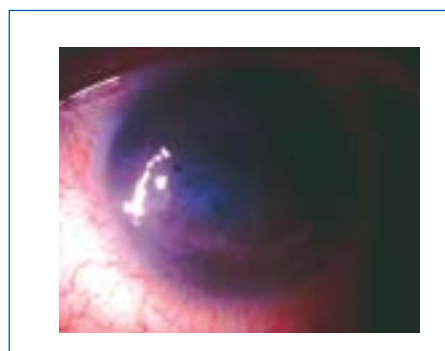


FIGURE 2

Analysis of human chromosome 22

Ian Dunham



Since 1991 the research of my lab has centred on the uses of human chromosome 22 as a model system for genome analysis. Chromosome 22 represents about 1 per cent of the whole genome, but is a relatively gene-rich chromosome and hence is a tractable model system for a number of genome-wide studies.

Initial research comprised physical mapping in yeast artificial chromosomes (YACs) to produce an extensive YAC map, which served as the backbone for future production of the DNA sequence. Following on from this, chromosome 22 was also mapped using radiation hybrids and bacterial artificial chromosomes (BACs) and other bacterial cloning systems. We also systematically dissected the structure of a series of low-copy repeats in 22q11.

During 1996–1999 our major concern was to bring the sequencing of human chromosome 22 to completion, as leader of the consortium of four sequencing groups and numerous collaborators. At the same time we established a benchmark level of gene annotation on the sequence. This annotation provided the basis for comparison by many external groups aimed at developing gene prediction and annotation procedures. We have continued to explore the limits of gene annotation based on cDNA and EST data combined with laboratory confirmation using cDNA library screening, RT-PCR and microarray expression analysis and have

recently released the latest version of this annotation. Currently chromosome 22 represents the best annotated region of the human genome, providing an excellent model system to develop functional genomic approaches.

Based on the established infrastructure of chromosome 22 a series of additional studies have been developed. First, we exploited the overlaps between clones from different haplotypes sequenced in the genomic tilepath to identify high-density SNPs in clusters. Second, the SNP Consortium approaches were piloted on chromosome 22 generating randomly distributed SNPs. Taken together these two sources provided an exceptional SNP resource and allowed us to explore the nature of linkage disequilibrium across the whole chromosome. Our current research interests build on the knowledge of human chromosome 22 as a defined subset of the human genome to develop approaches to studying gene expression and networks at the mRNA and protein level. These include microarray expression analysis, cloning of tagged genes and expression of their proteins, and study of protein intracellular localization. In collaboration with Nigel Carter, Kathryn Woodfine, a PhD student, is using the chromosome 22 tilepath and sequence to study replication timing across the chromosome.

www.sanger.ac.uk/Teams/Team62

Collins J E *et al.* (2003) Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* 13, 27–36

Dawson E *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544–8

Dawson E *et al.* (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res.* 11, 170–8

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

Dunham I (2000) The gene guessing game. *Yeast.* 17: 218–24

Mullikin J C *et al.* (2000) A SNP map of human chromosome 22. *Nature* 407: 516–20

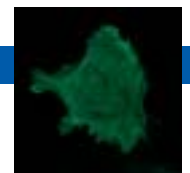
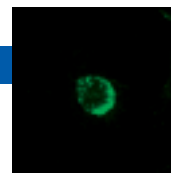
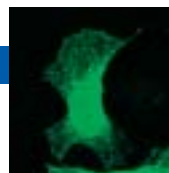
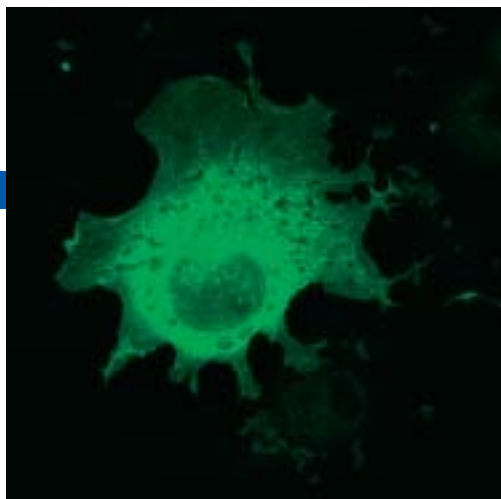
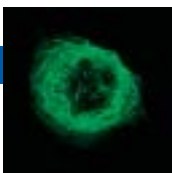
Dunham I *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* 402: 489–95

Collins J E *et al.* (1997) The organization of the gamma-glutamyl transferase genes and other low-copy repeats in human chromosome 22q11. *Genome Res.* 7: 522–31

Collins J E *et al.* (1995) A high-density YAC contig map of human chromosome 22. *Nature* 377: 367–79

Intracellular localizations of five representative genes from human chromosome 22. The localization of the expressed protein was identified by detection with fluorescently labelled antibody and visualized by confocal microscopy.

Images provided by Carol Edwards, Charmain Dunham and John Collins.



Genome informatics

Richard Durbin



I have broad interests in bioinformatics related to genomes. Although I have been involved in software and database development, and the primary systematic analysis of large genomes, I also have a research group working directly with me, which has during recent years focused primarily on the development of new sequence analysis methods. In general we have taken approaches based on probabilistic models, for example hidden Markov models.

A key question given a new genome sequence is how to identify the exon-intron structures of all the protein-coding genes, and hence all the proteins encoded in the genome. Although we know many of the genes from mRNA or EST sequencing, coverage of transcribed sequence from such sources is far from complete for any organism. We have been developing general tools for integrating data in gene prediction by dynamic programming and a method for using comparative data such as from human and mouse in combination to improve statistical prediction methods. We also investigated splice site and intron prediction for the rare variant U12 splicing system.

Another direction of interest has been to transfer methods from computational linguistics to biological sequence analysis; there are many similarities because both try to infer meaning from a sequence of more primitive symbols (letters or phonemes for linguistics/speech). We have seen a significant improvement in protein sequence classification using such approaches. We are also trying to optimize use of comparative information when there are multiple sequences involved from a variety of organisms. One project brings together molecular evolution approaches – such as those used for phylogenetic study – with gene-finding and other applied sequence analysis methods. Finally, I am also interested in the application of human genetic and population structure theory to very large-scale human genetic variation data sets, of the type being collected in many of the major programmes at the Sanger Institute.

Alongside these research groups, I also remain involved in WormBase, the international model organism database for *C. elegans*, and in the development of the genome database software package Acedb that is used by WormBase and many other genome databases. The Sanger WormBase group works on many of the sequence annotation aspects of the worm genome, as well as on data integration. This, together with other data resource projects in the division such as Ensembl, provides data sets and direct applications for the more theoretical developments described above.

www.sanger.ac.uk/Teams/faculty/durbin
www.sanger.ac.uk/Projects/C_elegans
www.sanger.ac.uk/Software/Acedb

Meyer I and Durbin R (2002) Comparative gene prediction using pair hidden Markov Models. *Bioinformatics* 18: 1309–18

Howe K, Chothia T and Durbin R (2002) GAZE: a generic framework for the integration of gene prediction data by dynamic programming. *Genome Res.* 12: 1418–27

Levine A and Durbin R (2001) A computational scan for U12 introns in the human genome sequence. *Nucleic Acids Res.* 29: 4006–13

Stein L et al. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 29: 82–6

Durbin R et al. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK

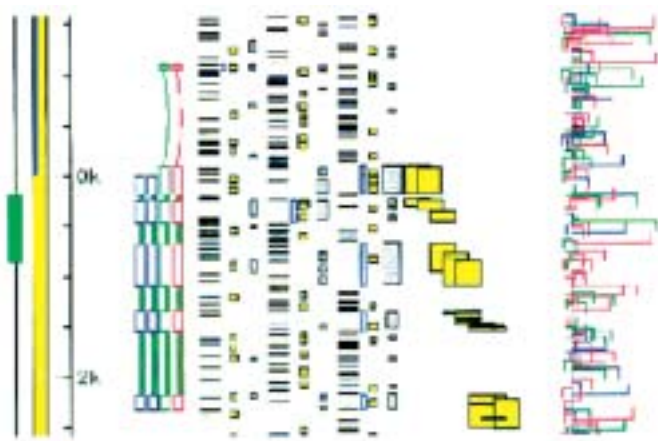


FIGURE 1

RESEARCH INTERESTS

FIGURE 1
Gene prediction using the GAZE algorithm can correctly determine the structure of 5' ends of *C. elegans* genes by incorporating special rules to handle trans-splicing.

FIGURE 2
The use of language modelling methods from speech recognition allows us to detect weak similarity to a TF_Otx domain in the CRX_HUMAN protein, because of its context adjacent to a homeobox domain. This allows functional

explanation of a mutation at position 242 which causes cone-rod dystrophy, which often results in blindness.

CRX_HUMAN cone-rod homeobox protein



Homeobox 40–96

Context: TF_Otx 164–250

FIGURE 2

Human disease analysis in the mouse

Lorraine A Everett



The research aims of our group all relate to using the power of the genomic sequence to yield a better understanding of human disease. We are currently adopting two complementary approaches to this end.

One of these uses a reverse genetics (from gene to phenotype) strategy to study the *SLC26* family of anion transporters and to investigate their involvement in disease. My interest in this gene family began in 1997 with the positional cloning of the Pendred syndrome gene (*SLC26A4*) and its subsequent characterization as well as the development of a mouse knockout. Together, these studies provided key insight into both thyroid physiology and the pathogenesis of deafness. It is now clear that there are 11 members of this gene family in mammals, each displaying a markedly different tissue expression pattern. It is interesting to note that the first three human members discovered are all implicated in human monogenic diseases (diastrophic dysplasia and congenital chloride diarrhoea as well as Pendred syndrome) and even more tempting to speculate on the pathogenic role of the other newer members. We are characterizing these novel genes and their encoded proteins further and central to this characterization is the

efficient generation of knockout mice lacking these genes and the subsequent thorough examination of their phenotype.

In our other research focus, we are using forward genetics approaches (from phenotype to gene) to identify the genes responsible for phenotypes of interest to us (primarily deafness as well as the insulin resistance syndrome and related pathologies such as type 2 diabetes). Various strategies will be used to achieve these aims, and include the use of mice acquired from large-scale mutagenesis screens, the use of sensitized pathway approaches coupled with more specialized phenotypic characterizations in our own ENU mutagenesis screens, and the identification of genes responsible for phenotypic differences between inbred strains of mice. While early positional cloning projects were lengthy tasks, the availability of the genomic sequence coupled with the powerful genomic techniques available when using the mouse as a model organism can allow such projects to be completed within a matter of months, and we are developing high-throughput approaches to such gene identification.

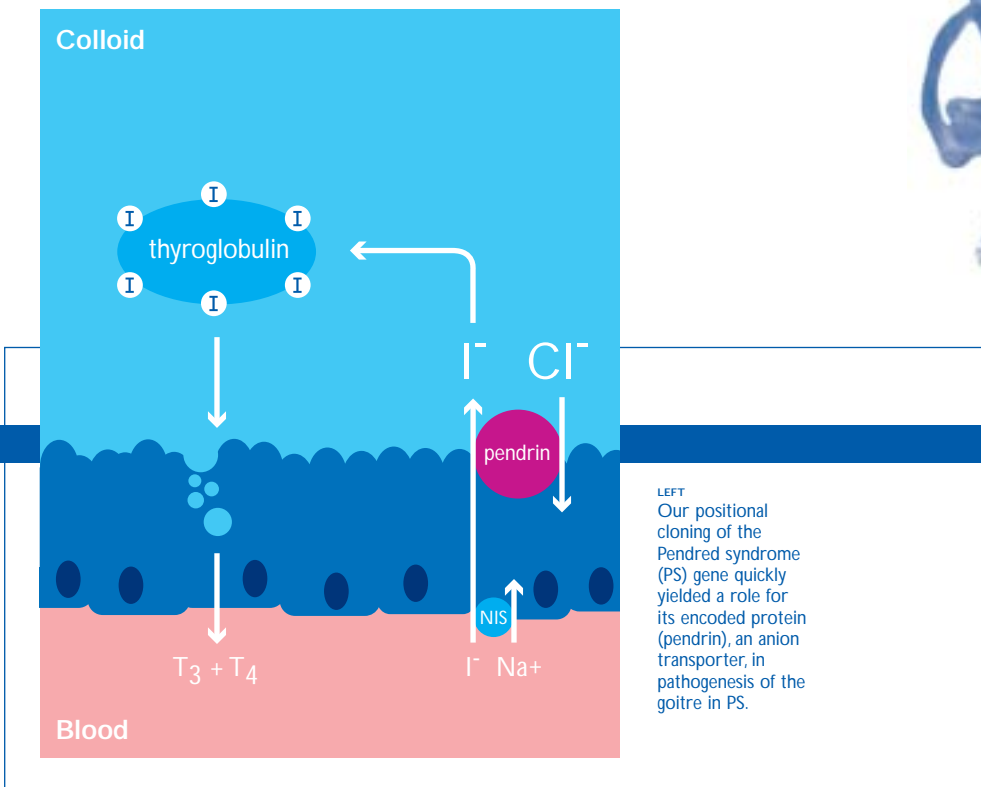
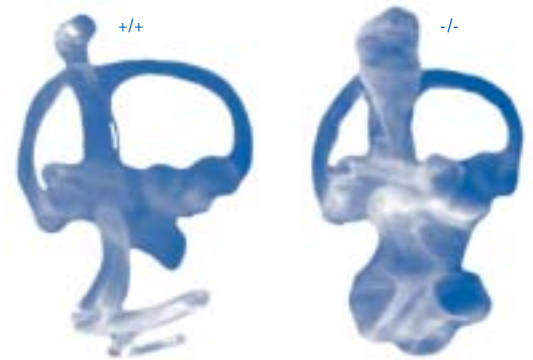
www.sanger.ac.uk/Teams/Team88

Everett L A et al. (1997) Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* 17: 411–22

Everett L A and Green E D (1999) A family of mammalian anion transporters and their involvement in human genetic diseases. *Hum. Mol. Genet.* 8: 1883–91

Everett L A et al. (1999) Expression pattern of the mouse ortholog of the Pendred's syndrome gene (*Pds*) suggests a key role for pendrin in the inner ear. *Proc. Natl. Acad. Sci. USA* 96: 9727–32

Everett L A et al. (2001) Targeted disruption of mouse *Pds* provides insight about the inner-ear defects encountered in Pendred syndrome. *Hum. Mol. Genet.* 10: 153–61



RESEARCH INTERESTS

ABOVE
Our development of a mouse knockout of the Pendred syndrome gene (*Slc26a4*^{-/-}) provided key insight into the aetiology of the deafness of PS. Shown here are paint-filled inner

ears (performed by Doris Wu, NIDCD) of wild-type and *Slc26a4*^{-/-} mice, showing the profound dilatation resulting from aberrant anion transport.

Cell signalling in *C. elegans*

Andy Fraser



Cell-cell signalling is essential for many biological processes ranging from developmental patterning to the regulation of cell proliferation. The main focus of my lab is to use complementary large-scale approaches in *C. elegans* both to identify the components of key signalling pathways and to understand how the information from multiple pathways is integrated in single cells. Many of the pathways that we study are very similar between the worm and humans, and previous work suggests that what we learn about the genes in the worm will help us to understand genes that may be involved in human cancer and other genetic diseases.

One of the many advantages of *C. elegans* as an animal model system is that it has a pattern of development that is essentially identical in all individuals. This invariant development means that we can use the worm for very sensitive screens for deviations from normal development. For example, the adult vulva normally consists of exactly 22 cells which derive from six precursor cells. Despite the fact that this tissue appears very simple,

screens have shown that vulval development requires highly complex signalling events involving a classical EGF-ras-raf-MAPK pathway along with Wnt and Notch pathways. Understanding how these pathways talk to each other in specifying this tissue is one of the major immediate goals for my lab.

We analyse signalling pathways by combining multiple large-scale analyses of gene function – these include using RNA-mediated interference (RNAi) to look individually at the loss-of-function phenotypes for almost every gene in the worm genome, and various biochemical approaches to map out the physical interactions between pathway components. Once we have done this for several pathways, we hope to be able to use this information to move beyond the view of pathways as a linear series of switches and instead look at whole signalling networks. Ultimately, we would like to extend our models of signalling to humans – if we can confirm that what we find in the worm is also true in humans, this would be the most important validation for our work.

Fraser A G *et al.* (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408: 325–30.

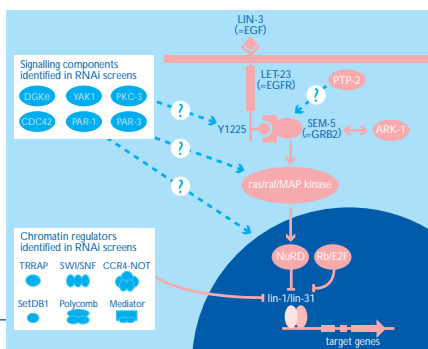
Kamath R S *et al.* (2001) Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *C. elegans*. *Genome Biol.* 2(1): research 0002.1–0002.10.

Kamath R S *et al.* (2003) Systematic functional analysis of the *C. elegans* genome using RNAi. *Nature* 421: 231–7.

Ashravi K *et al.* (2003) Genome-wide RNAi analysis of *C. elegans* fat regulatory genes. *Nature* 421: 268–72.

Dillin A *et al.* (2002) Rates of behavior and aging specified by mitochondrial function during development. *Science* 298: 2398–401.

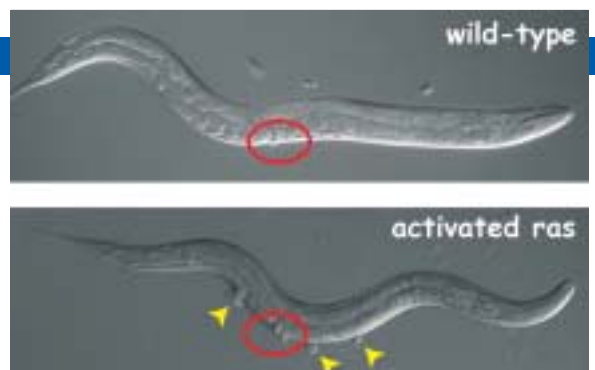
www.sanger.ac.uk/Teams/Team37



ABOVE
Forming a vulva requires activation of an EGF/ras/raf/MAP kinase signalling pathway in one specific cell; some of the components of the signalling pathway that have been identified by forward genetics are shown in pink. We have carried

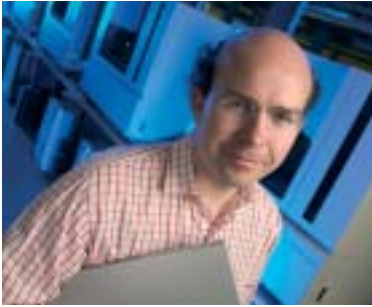
out RNAi screens which have identified new signalling components and chromatin regulators in the pathway. Determining how these new components interact with the known signalling network is a next key question for my lab.

RIGHT
Worms with an activating mutation in ras have both a normal vulva (circled in red) and multiple pseudovulvae (yellow arrowheads). We use this as the basis for some of our screens to identify genes that affect ras signalling.



Prediction research

Tim Hubbard



Bioinformatics adds value to biological data in two ways: firstly, through the construction of databases that organize data and thereby facilitate the discovery of new relationships within it; secondly, through the development of methods to predict properties of biological systems that would otherwise have to be determined experimentally.

For any type of prediction it is important that there are resources and test data sets to enable the efficiency of different methods to be evaluated. The group has carried out research into the evaluation of sequence similarity search methods based on the SCOP structural classification of proteins database, and provides a web-based evaluation service. The group is also involved in the Critical Assessment of Protein Structure Prediction (CASP) event held every two years. It has carried out similar evaluations of the accuracy of different gene prediction methods on genomic sequence.

Members of the group have been developing various methods to apply to genomic scale datasets such as to predict transcription start sites directly

and through the use of expression and comparative genomic sequence data to understand promoters and other structures in genome sequence. The group is very active in various 'open source' projects such as Ensembl and was responsible for the creation of biojava (www.biojava.org): client and server code to support the distributed annotation system and expression browser software.

www.sanger.ac.uk/Teams/faculty/hubbard
www.ensembl.org

Down T A and Hubbard T J (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* 12: 458-61

Hubbard T J et al. (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30: 38-41

Lepale R and Hubbard T J (2002) MaxBench: evaluation of sequence and structure comparison methods. *Bioinformatics* 18: 494-5

Lo Conte L et al. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30: 264-7

Moult J et al. (2001) Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* 45: 2-7

Pocock M R, Down T and Hubbard T J P (2000) BioJava: Open Source Components for Bioinformatics. *sigbio newsletter* 20: 10-12

Pocock M R and Hubbard T J (2000) A browser for expression data. *Bioinformatics* 16: 402-3

Dunham I et al. (1999) The DNA Sequence of human chromosome 22. *Nature* 402: 489-95

Hubbard T J (1999) RMS/Coverage graphs: A qualitative method for comparing three-dimensional protein structure predictions. *Proteins* 37: 15-21

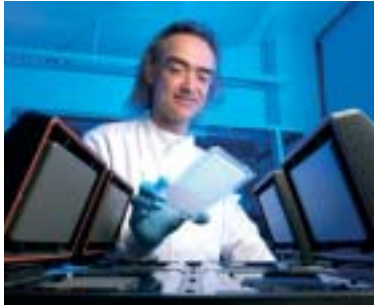
Brenner S E, Chothia C and Hubbard T J P (1998) Assessing sequence comparison methods. *Proc. Natl. Acad. Sci. USA* 95: 6073-8

Hubbard T J P and Park J (1995) Fold recognition and ab initio structure predictions using Hidden Markov models and beta-strand pair potentials. *Proteins: Struct. Funct. Genet.* 23: 398-402



Parasite sequencing and pathogen microarrays

Al Ivens



My interests are shared over two main areas: firstly, the management of several eukaryotic sequencing projects (*Leishmania*, *Eimeria*, *Schistosoma*), both at the sequencing and annotation levels (page 30), and secondly, pathogen microarrays.

As a functional extension of sequencing projects undertaken by the Pathogen Sequencing Unit, a pathogen microarray group has been established. The group is currently involved in three main projects, which have implemented many of the methodologies established by the Sanger Institute Microarray Facility (page 32), and *S. pombe* microarray group (page 37). To date, the gridded microarrays made by my group comprise PCR-generated segments of annotated genomes.

The projects are as follows:

(a) Comparative genomics of the enterics, concentrating on *Salmonella typhi* (which causes more than 17 million cases of typhoid fever a year, 600 000 of which are fatal) and related serovars. This aim is to identify which genes are present or absent when compared to the reference *S. typhi* CT18 genome. To date, 40 genomes, from both animal and human enteric organisms, have been analysed in this way: gene deletion events are clearly discernible. As these may well explain differences in attributes such as pathogenicity or host range, they provide excellent start points for

further laboratory-based studies of the organism(s). In related experiments, RNA expression profiling is being used to investigate the effect that gene-specific knockouts/mutations have on the expression of other genes.

(b) *Dictyostelium discoideum* is a model eukaryotic organism, with molecular genetics comparable to yeast, but a very different biology. It is often used for investigating chemotaxis, phagocytosis, intercellular signalling, development, cell differentiation and the evolution of multicellularity. It is also a convenient test system for investigating the mode of action of drugs such as lithium. We are producing a microarray resource to facilitate these investigations.

(c) *Plasmodium* arrays. An international consortium of eight laboratories, with interests that range from crystallography to transfection, was recently funded by the Wellcome Trust. This coincided with the publication of the human malaria parasite genome sequence, and the comparative sequencing of several non-human malaria parasites by the PSU (page 30). The role of my group is to assemble *Plasmodium* microarray resources based on these genomes, with a view to generating a pan-*Plasmodium* array. As malaria functional studies can be undertaken *in vivo* in rodents, the mouse arrays produced by the Sanger Institute Microarray Facility provide a valuable additional resource to study how the parasite interacts with its host to cause disease.

Almeida R *et al.* (2002) From genomes to vaccines: *Leishmania* as a model. *Philos. Trans. R. Soc. London. SerB. Biol. Sci.* 357: 5–11

Myler P J *et al.* (2002) The *Leishmania* genome project: new insights into gene organization and function. *Med. Microbiol. Immunol.* 190: 9–12

Tan T H *et al.* (2002). tRNAs in *Trypanosoma brucei*: genomic organization, expression, and mitochondrial import. *Mol. Cell. Biol.* 22: 3707–17

Siman-Tov M M, Ivens A C and Jaffe C L (2002) Molecular cloning and characterization of two new isoforms of the protein kinase A catalytic subunit from the human parasite *Leishmania*. *Gene* 288: 65–75

Denny P W *et al.* (2002) *Leishmania* major Rab7: characterisation of terminal endocytic stages in an intracellular parasite. *Mol. Biochem. Parasitol.* 123: 105–13

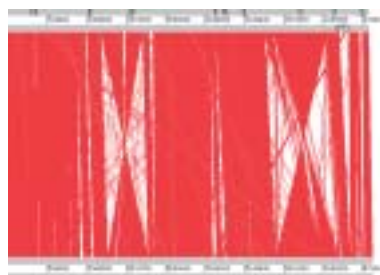
BELOW
The social amoeba *Dictyostelium discoideum* is a model eukaryotic organism for studies of differentiation and cell signalling. Shown above is the fruiting body, with elongating stalk (collaboration with Rob Kay).



www.sanger.ac.uk/Projects/Protozoa

www.sanger.ac.uk/PostGenomics/PathogenArrays

RESEARCH INTERESTS



LEFT
Sequence similarities between genomes of the closely related bacteria *Salmonella typhi* (top section of panel) and *S. typhimurium* (bottom section). Regions of similarity are indicated by the red regions: sequence-divergent regions are immediately obvious as clear areas.

Cell signalling, cell fate and proteolytic pathways in *C. elegans*

Patricia Kuwabara



The 100 Mb genome of the nematode *C. elegans* is the first genome of a multicellular organism to be sequenced in its entirety. Our group is taking advantage of the fully sequenced genome by asking how development and physiology are regulated globally, using functional genomic tools. We are applying techniques such as RNAi and DNA microarray expression profiling, to identify and to study the function of genes involved in cell signalling and cell fate determination, cell-cycle control and calpain-mediated proteolytic regulation.

C. elegans is an excellent model organism for pursuing these studies because it is genetically tractable, amenable to molecular and cell biological studies, and has a fully documented cell lineage and anatomy. Moreover, many of the nematode genes involved in these pathways and cellular processes have vertebrate homologues. Therefore, the information gained from studying such genes in the worm is likely also to shed light on the roles that these genes play in human development and disease.

www.sanger.ac.uk/Teams/Team20

Bergamaschi D *et al.* (2003) IASPP oncoprotein is a key inhibitor of p53 conserved from worm to human. *Nature Genet.* 33: 162–7

Kuwabara P E *et al.* (2000) A *C. elegans* patched gene, *ptc-1*, functions in germ-line cytokinesis. *Genes Dev.* 14: 1933–44

Kuwabara P E and Coulson A (2000) RNAi – Prospects for general technique for determining gene function. *Parasitol. Today* 16: 347–9

Sokol S B and Kuwabara P E (2000) Proteolysis in *C. elegans* sex determination: Cleavage of the membrane protein TRA-2A by the calpain TRA-3. *Genes Dev.* 14: 901–6

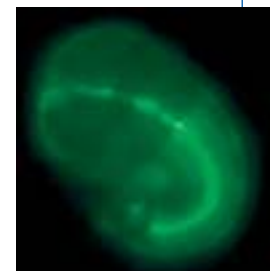
Kuwabara P E (1997) Worming your way through the genome. *Trends Genet.* 13: 455–60



Adult *C. elegans* hermaphrodite stained with DAPI to visualize DNA.

RESEARCH INTERESTS

Nomarski DIC micrograph of a live *C. elegans* embryo (left) expressing a green fluorescent protein reporter construct highlighting the adherens junctions of the pharynx and gut (right).



The atlas of gene expression

John McCafferty



The activity of genes in health and disease is manifested through the proteins which they encode. Proteins ultimately drive functional processes in cells and tissues and so knowledge of protein expression levels, modifications and sites of action can make an important contribution to our understanding of gene function. The 'Atlas of Gene Expression' project aims to systematically create a database describing expression level/localization of protein products in a wide range of human and murine tissues. To achieve this we will use antibodies as probes on tissue sections to identify sites of protein expression *in vivo*.

To meet these goals the Atlas group will express and purify recombinant proteins using a range of bacterial and eukaryotic expression systems. These recombinant proteins will be used as antigens to select recombinant antibodies from phage-antibody libraries. Phage display is an approach which potentially relieves the bottleneck in the production of antibody reagents against the multitude of genes

newly identified from sequencing the human genome. With phage display antibody molecules are physically linked to the genes which encode them, and so it is possible to select antibodies (and their associated gene) from large phage-antibody libraries by 'panning' on antigen. The resultant antibody clones are used to stain sites of protein expression in tissue sections.

To facilitate the use of tissue staining as a high-throughput 'read out' the group will introduce the use of tissue microarrays, automated immunocytochemical staining as well as automated data collection and screening. Ultimately this project aims to create a quality, information-rich database of protein expression profiles, which is easily accessible to the worldwide research community.

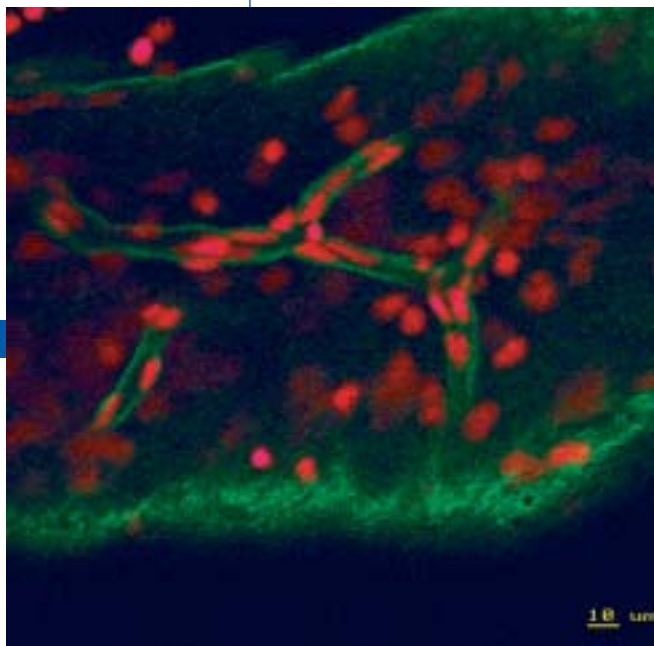
www.sanger.ac.uk/Teams/Team86

Patil M *et al.* (2001) Generation of baculovirus recombinants using PCR amplified fragments. *Biotechniques* 30: 1212-15

Osborn J K *et al.* (1998) Directed selection of MIP1a neutralising CCR5 antibodies from a phage display human antibody library. *Nature Biotechnol.* 16: 778-81

Vaughan T *et al.* (1996) Human antibodies with sub-nanomolar affinities isolated from a large non-immunised phage display library. *Nature Biotechnol.* 14: 309-14

McCafferty, J *et al.* (1990) Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 348: 552-4



LEFT
Fluorescent antibody detection of collagen IV in basement membranes.

BELOW
Electron micrograph of filamentous bacteriophage.
Courtesy Richard Perham



Bacterial sequencing and analysis

Julian Parkhill



My primary interest at the Wellcome Trust Sanger Institute is in bacterial sequencing. I am responsible for the management of each of the bacterial sequencing projects (page 31), both at the sequencing and annotation levels. All the bacterial genomes are sequenced using whole-genome shotgun procedures. Though initially straightforward, these often require a good deal of manual intervention during the finishing process, investigating and disentangling repeat structures and other ambiguities on a global scale. This is done in collaboration with the highly experienced finishers within the sequencing teams. Often we can gain useful and surprising insights into the biology of the organisms at an early stage through these methods.

When the genomes are finished we perform an unashamedly manual analysis and annotation of the sequence. This is

achieved by a group of lab-trained postdoctoral microbiologists with interests in different aspects of microbial biology. We attempt to go beyond a simple annotation of the genome to look at how we can interpret the data in terms of the biology of the organism. We are particularly interested in comparative genomics, and in trying to uncover the recent evolutionary history of the bacteria under investigation.

On the laboratory side, I run a group that is extending the comparison of genomes beyond the fully sequenced organisms. We are using comparative hybridization techniques to look for genes and regions that are present in close relatives of strains that have previously been sequenced. In this way we hope to identify genes involved in specific differential phenotypes such as virulence and host range. At present we are investigating members of the *Burkholderia cepacia* complex and of the enterohaemorrhagic *Escherichia coli* group.

www.sanger.ac.uk/Projects/Microbes

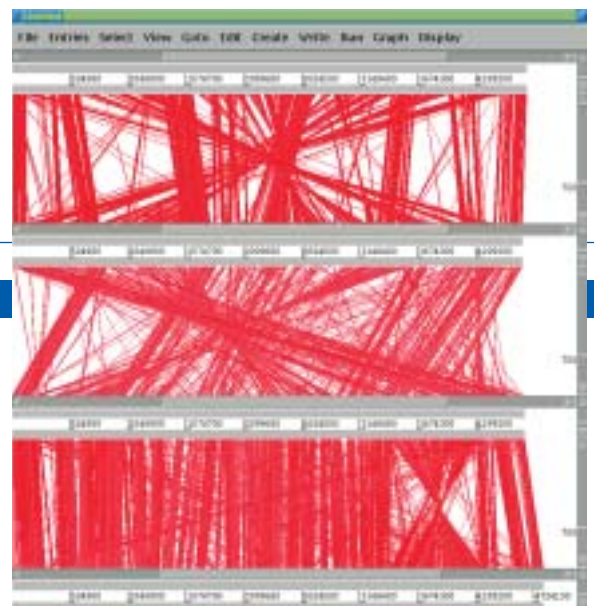
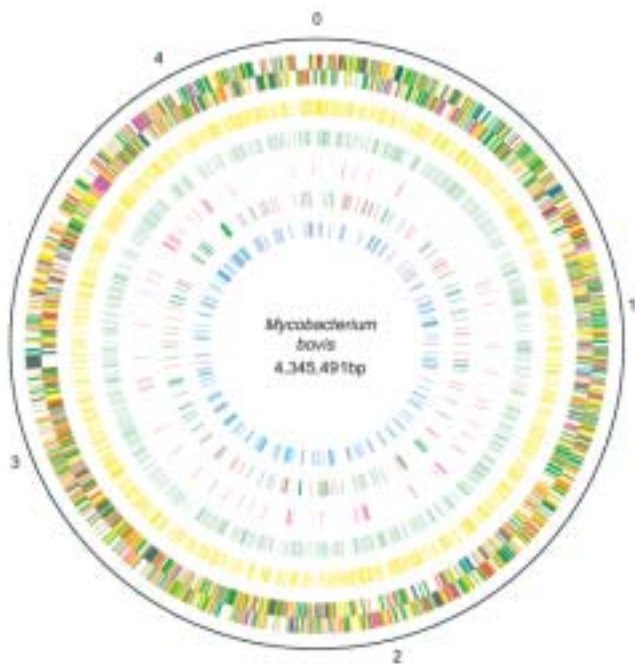
Parkhill J et al. (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 413: 848-52

Parkhill J et al. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 413: 523-7

Cole S T et al. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409: 1007-11

Parkhill J et al. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404: 502-6

Cole S T et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537-44



RESEARCH INTERESTS

ABOVE
Circular map of differences between the agents of human and bovine TB. Map of all the genes in *Mycobacterium bovis* (two outer rings) with every base-pair difference to *M. tuberculosis* indicated in the inner rings.

RIGHT
ACT comparison across diverse genomes: *Yersinia pestis* vs. *Yersinia enterocolitica* vs. *Escherichia coli* vs. *Salmonella typhi*.

Bioinformatics

Kate Rice



My interests lie in using informatics and large-scale analyses to remove uncertainties; to refine genomic information, and to use annotated genomes as a tool to understand human disease.

In a joint research project with Nigel Carter, we are using existing bioinformatics procedures, and developing novel approaches, to investigate chromosomal translocational breakpoints, which are important in a significant number of human diseases. We are using sequence analysis to search for patterns in breakpoint-spanning regions in order to understand better the events leading to translocation.

We are also actively involved in refining the use of microarray data. Microarrays are revolutionizing the way in which we look at gene expression and new methods are being developed at an extraordinary pace. It is crucial that existing and new data are analysed so that comparisons may usefully be made. Moreover, it is vital that we anticipate the questions posed and demands made by research groups in their use of such data.

The Sanger Institute's own microarray facility produces expression arrays for a wide range of organisms ([page 32](#)) and our current research areas include: estimation of the sequence redundancy in our array features for human, mouse and rat; determination of methods and rules for identifying cross-hybridization on chips; examination of array sequences for evidence of possible secondary structures, distribution of sequence length and composition.

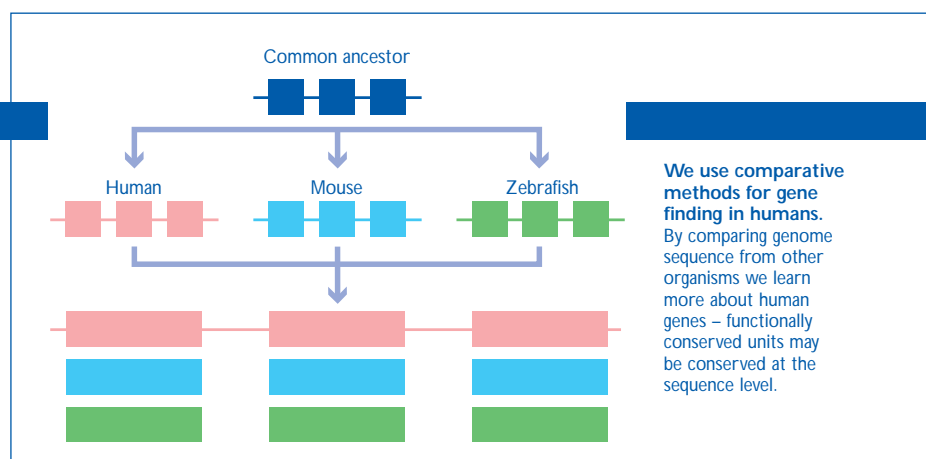
I also have a particular interest in genes of medical interest, such as those implicated in epilepsy, and have a research programme designed to improve annotation of these genes.

www.sanger.ac.uk/Teams/Team65

Bentley D R *et al.* (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* 409: 942–3

Dawson E *et al.* (2001) A SNP resource for human chromosome 22: Extracting dense clusters of SNPs from the genomic sequence. *Genome Res.* 11: 170–8

Mullikin J C *et al.* (2000) An SNP map of human chromosome 22. *Nature* 407: 516–20



X chromosome: biology, evolution and disease

Mark Ross



We are using sequence information as the basis to study the unique biological properties of the human X chromosome and to understand how these properties have been moulded by mammalian sex chromosome evolution. The dramatic divergence of the X and Y chromosomes is thought to be a consequence of their involvement in sex determination and has involved the substantial mutation and loss of genes on the Y chromosome. This process has had many important biological consequences, such as the need for a dosage compensation mechanism in females and the large number of X-linked diseases in males.

With our collaborators in the public Human Genome Project, we have essentially completed the X chromosome reference sequence. Sequence analysis and subsequent experimental work have allowed us to identify and describe many X-linked genes. We can predict that the chromosome is gene poor, containing only 3 per cent of the protein-coding genes in the genome. This contrasts with the observation that 11 per cent of described Mendelian conditions are X-linked. The genes implicated in several X-linked conditions, such as

lymphoproliferative syndrome and cleft palate, have been identified using the sequence data. Many others remain unidentified; these include cancer susceptibility and deafness genes, which we are attempting to identify in collaboration with other groups.

We are considering many questions relating to the biological consequences of sex chromosome evolution. Why do some genes evade silencing on the inactive X chromosome in females, and is this influenced by sequence context? How does the unusual pattern of recombination between X and Y in male meiosis affect mutation rate, gene duplication and the haplotype structure of the X chromosome? Is there a functional explanation for the high concentrations of certain repetitive element types on the chromosome? In order to address these questions, we are using the human reference sequence as the basis for gene expression studies, SNP development, haplotype mapping and comparison to the genomes of other mammals.

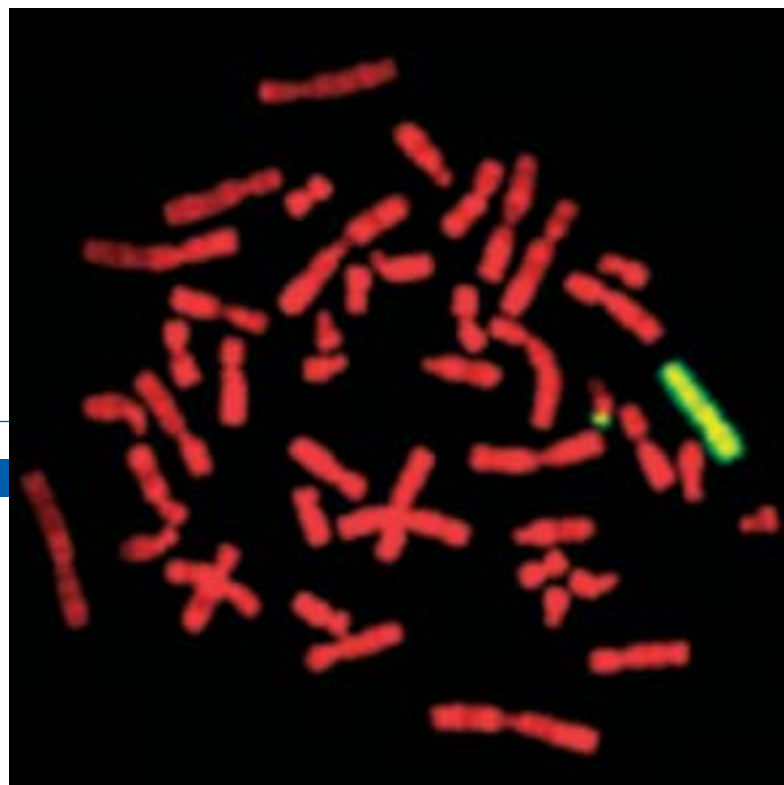
www.sanger.ac.uk/Teams/Team61

Stephan D A *et al.* (2002) Physical and transcript map of the hereditary prostate cancer region at Xq27. *Genomics* 79: 41–50

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

Bentley D *et al.* (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* 409: 942–3

Coffey A J *et al.* (1998) Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* 20: 129–35



RESEARCH INTERESTS

In this human metaphase chromosome spread, the entire X chromosome is made fluorescent using a chromosome 'paint' probe. The Y chromosome shows only a small region of fluorescence at the tip of its short arm.

Gene trapping in the mouse

Bill Skarnes



The primary research interest of my laboratory is centred on the use of gene-trapping technology to identify new genes critical for normal embryonic development in the mouse. Over the past 15 years, we have developed efficient and versatile gene trap vector designs for the generation of insertional mutations in mouse embryonic stem (ES) cells. This technology has been used successfully in recent years to recover mutations in 10 per cent of all protein-coding genes in mouse. The Sanger Institute Gene Trap Mutagenesis team will continue to expand this resource with the aim of isolating insertional mutations in every gene in the mouse. We are exploiting this library of gene trap ES cell lines as a source of mutant mice in genetic screens to identify recessive lethal mutations that perturb early embryonic development.

To identify new cell-signalling molecules required for normal development, we developed the 'secretory trap' approach to recover insertional mutations specifically in proteins targeted for secretion. From a screen of more than 80 mutations in secreted and membrane-spanning proteins, of which

one-third cause embryonic lethal phenotypes, several genes that play a critical role at gastrulation have been identified and selected for further detailed analysis. These include: (i) the Wnt co-receptors *Lrp5* and *Lrp6* which are essential at gastrulation for the formation and patterning of nascent mesoderm; (ii) a novel secreted protein expressed in the node that is required for the allocation of cells to the floorplate; and (iii) *Spint2*, a serine protease inhibitor important for proper formation of the epiblast.

In cells of the early embryo, lineage decisions are accompanied by dramatic changes in chromatin composition and conformation. We have initiated a genetic screen to study the role of chromatin proteins in embryonic development and their effect on global gene expression. Both loss-of-function and over-expression approaches in mouse ES cells will be used to identify novel chromatin proteins that influence cell fate decisions in the early embryo.

www.sanger.ac.uk/Teams/Team87

www.sanger.ac.uk/genetrap

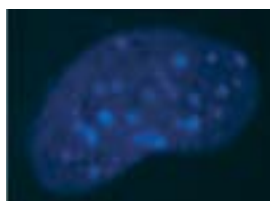


FIGURE 1



FIGURE 3

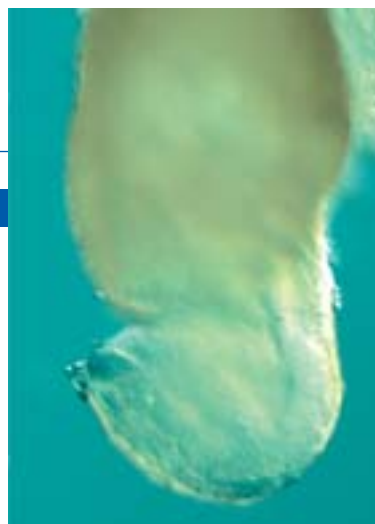


FIGURE 2

FIGURE 1
A gene trap insertion showing localization of the β -gal fusion protein to pericentric heterochromatin.

FIGURE 2
Abnormal morphogenesis of the epiblast in *Spint2* mutant embryos (right).

FIGURE 3
Ensembl genome browser showing position of gene trap insertions in the mouse *Lrp6* gene.

Zebrafish genetics and embryology

Derek Stemple



Work in our laboratory has primarily been directed at characterizing genes that control notochord development of zebrafish. The notochord is an essential organ for vertebrate development. It is both a skeletal element and a source of signals that pattern surrounding tissues. Genetic screens have led to the identification of many genes affecting notochord differentiation. We have successfully determined the identity of six of the affected genes and are working to identify a seventh.

Positional cloning revealed that three loci – *bashful*, *grumpy* and *sleepy* – encode individual $\alpha 1$, $\beta 1$, and $\gamma 1$ chains of the extracellular matrix protein laminin 1. There are surprisingly few details known of the *in vivo* requirements for the major isoforms of laminin. We found that the failure to produce either the $\beta 1$ or $\gamma 1$ chain results in complete loss of laminin 1 immunoreactivity and disruption of basement membrane, particularly surrounding the notochord.

In other work, we employ embryological manipulations to investigate development of axial tissues. In zebrafish, the dorsal

organizer gives rise to the entire axial mesendoderm, including the notochord. We developed an efficient micropipette-based method to transplant dorsal organizer tissue. Using this method we found that complete removal of the organizer region leads to the loss of axial mesendoderm. By contrast, if only the morphological shield is removed, the resulting embryos regenerate the missing tissue and develop normally. Such morphological shields are sufficient to induce the formation of a complete secondary axis upon transplantation to host embryos. Finally, by transplanting deep versus superficial fragments of the shield, we found that the organizer activity can be divided into separable head-inducing and trunk/tail-inducing regions.

Dorsal organizer specification depends on nodal signalling and one aspect of our research has been to investigate how the activity of these potent secreted proteins is modulated during early development. Two zebrafish nodals – Squint and Cyclops – have fundamentally different ranges of activity, with Squint acting at a distance to induce mesoderm, and Cyclops only able to act locally. We find that when the two nodal antagonists, Lefty-1 and Lefty-2, are simultaneously disrupted, unchecked nodal signalling results in a severe lethal gastrulation phenotype. By sharp contrast, when the Lefties are disrupted in *squint* homozygous mutants, the resulting embryos are able to gastrulate and form a complete embryonic axis. Thus during gastrulation Lefties play a more crucial role in restricting the activity of the long-range Squint than the short-range Cyclops molecule.

Feldman B *et al.* (2002) Lefty antagonism of squint is essential for normal gastrulation. *Curr. Biol.* 12: 2129–35

Parsons M J *et al.* (2002) Zebrafish mutants identify an essential role for laminins in notochord formation. *Development* 129: 3137–46

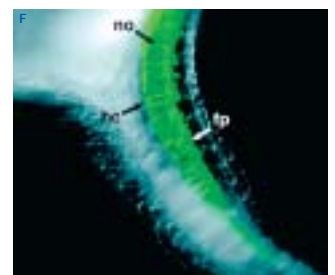
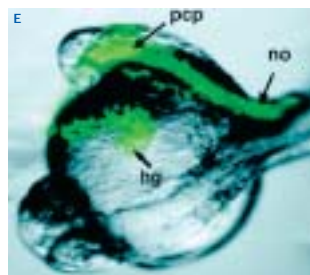
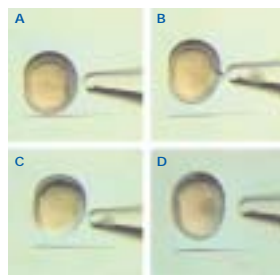
Saude L *et al.* (2000) Axis-inducing activities and cell fates of the zebrafish organizer. *Development* 127: 3407–17

BELOW
Laminin 1 expression is disrupted in *sleepy* mutant embryos.



www.sanger.ac.uk/Teams/Team31

RIGHT
Embryonic shield transplantation (A-D) leads to induction of a complete second axis (E,F) (pcp: pre-chordal plate, hg: hatching gland, fp: floorplate, no: notochord).



Cancer genetics

Mike Stratton



My research interests are in the study of cancer genetics, particularly the identification of the underlying mutated genes. The Cancer Genome Project, a group of about 50 people, is performing systematic searches for somatic mutations of various types in human cancers based upon the finished and annotated human genome sequence.

The most ambitious project in this portfolio is a screen for point mutations of every coding exon of the human genome in a series of 48 cancer cell lines. However, we are also characterizing cancer genomes in a variety of other ways, ultimately aiming to obtain complete descriptions that reveal all the underlying causative genes and reflect the abnormal influences that have shaped the cancer genome.

To achieve this end we have developed a high-throughput platform for mutation detection and have accumulated approximately 2000 cancer cell lines, the largest collection in the world, and an extensive series of primary tumours. Efforts are made to translate the discovery of new genes, where possible, into anticancer drugs and diagnostics.

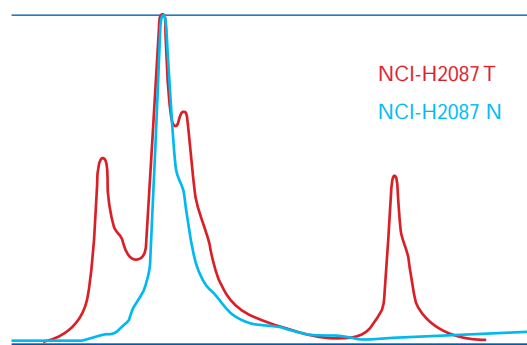
I also study predisposition to cancer using samples from families with multiple cases of cancer to map and identify new susceptibility genes for breast, testicular and other cancers. These discoveries are subsequently used in counselling at-risk individuals and in targeting preventive action.

www.sanger.ac.uk/CGP

Davies H *et al.* (2002) Mutations of the *BRAF* gene in human cancer. *Nature* 417: 949–54

Meijers-Heijboer H *et al.* (2002) Low penetrance susceptibility to breast cancer due to *CHEK2**1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nature Genet.* 31: 55–9

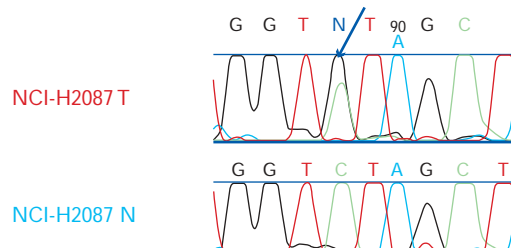
Capillary-based heteroduplex analysis



The top panel shows a comparison between a non-small lung cancer cell line (red) and normal DNA from the same person (blue) using our capillary-based heteroduplex analysis for mutation detection. The difference between these traces highlighted by our software resulted in the samples being sequenced (bottom) and a mutation (arrowed) in the *BRAF* serine threonine kinase being detected.

RESEARCH INTERESTS

Sequence traces



Embryo gene expression patterns

David Tannahill



My primary research interests are concerned with understanding the processes that govern how the adult body is constructed during development of the vertebrate embryo. Embryonic development is an amazingly complex process that begins with a single cell and ends up with an embryo containing all the different organs and tissues of the body, all organized in the correct anatomical relationship to each other. In order to understand development, we need to understand how cells make decisions to follow one developmental pathway and not another. We need to determine, therefore, how cells in the developing embryo interact with each other, and what signals they use to control their ultimate cell fate. These developmental decisions must also be tightly coordinated with the mechanisms that direct how cells grow, divide, die and move within the embryo to build the primordia of tissues and organs of the body.

As development proceeds, we also need to unravel the process of organogenesis so that an individual organ is constructed with the correct cell types, organized into the correct functional units. To understand development it is imperative, therefore, that we know where and when every gene in the genome is expressed within the embryo. This will provide us with

important information as to where genes normally operate and provides valuable insights into potential gene function, as well as the regulatory networks that control cell fate. Knowing the expression pattern of a gene is thus a prerequisite for determining its biological function.

This knowledge is also crucial in the understanding of the disease process. For example, in order to know what goes wrong in cancer, it is vital to know the factors that control normal cell division and differentiation. Furthermore, with a deep appreciation of how normal tissues are built, we are better able to design strategies for using stem cells and tissue engineering techniques to repair the damaged adult body. To achieve the goal of determining the expression pattern of all genes during development of the mouse embryo, I joined the Sanger Institute in the spring of 2003 to establish a new team that will develop high-throughput approaches for *in situ* hybridization and imaging of gene expression patterns in embryos.

www.sanger.ac.uk/Teams/Team39

Anderson C N G (2003) Molecular analysis of axon repulsion by the notochord. *Development* 130: 1123–33

Campbell D S *et al.* (2001) Semaphorin 3A elicits a stage-dependent collapse, turning and branching of *Xenopus* retinal growth cones. *J. Neurosci.* 21: 8538–47

Vermeren M M *et al.* (2000) Spinal nerve segmentation in the chick embryo: analysis of distinct axon repulsive systems. *Dev. Biol.* 225: 214–52

Britto J M, Tannahill D and Keynes R J (2002) A critical role for sonic hedgehog signalling in the vesicular expansion of the developing avian brain. *Nature Neurosci.* 5: 103–10

Keynes R J *et al.* (1997) Surround-repulsion of sensory axons in higher vertebrate embryos. *Neuron* 18: 889–97

FIGURE 1
Sonic hedgehog expression (blue) in the ventral midline of the developing brain.

FIGURE 2
Growing retinal axons (brown) avoid regions of semaphorin expression (blue) in the developing visual system.

FIGURE 3
Repulsion of growing axons by molecules secreted by the notochord (left).

FIGURE 4
Segmental expression of Neuropilin-2 (blue) in the developing trunk. Expression is localized to the dorsal root ganglia.

FIGURE 4

RESEARCH INTERESTS



FIGURE 1

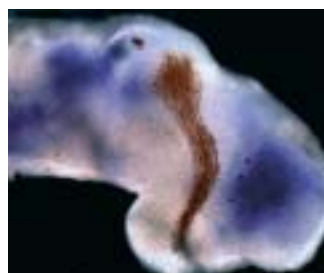


FIGURE 2



FIGURE 3

Microarrays, transcriptional networks and human disease

Dave Vetrie



The study of molecular genetics allows us to understand how inherited DNA alterations and variations in our genome sequence can give rise, or predispose an individual, to disease. With the availability of near-complete genome sequences of the human and model organisms such as the mouse, we are now in a position to study human biology and inherited disease pathologies in a manner that is unprecedented, accelerating improvements in human health. The main aims of my research interests over the last decade have been: (i) to study the molecular genetics of inherited human disease pathologies; and (ii) to develop and apply genomic tools and resources to aid in understanding disease pathologies.

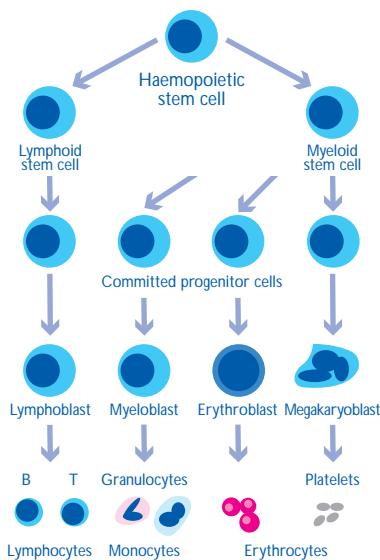
I have a long-standing interest in the molecular causes of several Mendelian disorders that were linked to genes found in the Xq22 region on the human X chromosome. This work culminated in the positional cloning of genes involved in two single-gene disorders: X-linked agammaglobulinaemia (*btk*) and X-linked syndromic deafness (*DDP*). More recently, I have studied complex common diseases which involve polygenic modes of inheritance as well as the determination of how our genetic makeup influences our response to medical treatments. Future projects will involve: (i) the development of large-scale association studies, to test for the association of a specific allelic variant with a specific disease phenotype present in cases but not controls; and (ii) the influence of specific sequence variants on the action, metabolism and clearance of administered drugs. Sequence variants identified in both types of studies can then be followed up to determine their functional significance and involvement in disease, pharmacokinetic and

pharmacodynamic mechanisms. An early focus of these studies will be the study of a large cohort of patients with epilepsy and their variable responses to antiepileptic drugs.

With respect to the second aim, my research group has been involved in the development of microarray resources which will allow the activity of genes in normal, experimentally perturbed and disease states to be studied in complex biological systems. My group has been instrumental in establishing the Sanger Institute Microarray Facility which presently produces high-quality genomic and expression arrays for our own use and for a variety of in-house and external collaborative projects. We are applying these microarray resources in the study of mammalian haemopoiesis. Haemopoiesis has served as a model process for studying stem-cell biology and is implicated in a wide range of human diseases. However, the transcriptional networks that determine these early cell fate decisions are still poorly understood. In collaboration with Professor T Green (Cambridge), we are pursuing a genome-wide microarray-based approach in combination with chromatin immunoprecipitation and RNAi to deduce transcriptional regulatory networks during haemopoiesis. We are interested in both identifying the direct gene targets of transcription factors as well as the regulatory DNA sequences to which these factors bind. An early focus of these studies has been the elucidation of the targets of the transcription factor SCL whose expression is critical for the development of all haemopoietic lineages (Figure 1) and has been shown to be involved in T-cell acute lymphoblastic leukaemia (T-ALL).

www.sanger.ac.uk/Teams/Team66

FIGURE 1
Mammalian
Haemopoietic
Lineages.



Jin H *et al.* (1999)
The human family of
Deafness/Dystonia
peptides (DDP) related
mitochondrial import
proteins. *Genomics* 61:
259–67

Woodward K *et al.*
(1998) Pelizaeus-
Merzbacher disease:
identification of Xq22
proteolipid-protein
duplications and
characterisation of
breakpoints by
interphase FISH. *Am.
J. Hum. Genet.* 63: 207–17

Jin H *et al.* (1996)
A novel X-linked gene,
DDP, shows mutations in
families with deafness
(*DFN-1*), dystonia, mental
deficiency and blindness.
Nature Genet. 14: 177–80

Vihinen M *et al.*
(1994) Structural basis
for X-linked
agammaglobulinaemia:
A tyrosine kinase
disease. *Proc. Natl. Acad.
Sci. USA* 91: 12803–7

Vetrie D *et al.* (1993) The
gene involved in X-linked
agammaglobulinaemia is a
member of the src family
of protein-tyrosine
kinases. *Nature* 361:
226–33

Careers at the Sanger Institute

Staff at the Sanger Institute fill a diverse range of roles demanding a broad band of skills and experience. In addition to scientific research and development there are opportunities in technical and administrative support, IT, engineering and robotics.

Recruitment targets all sectors of the market from entry-level graduate to individuals at postdoctoral and junior fellowship level and independent, established researchers at all levels. These people may be recruited locally, nationally and internationally resulting in an exciting cultural mix that provides a dynamic and stimulating environment in which to work.

The organization provides a structure that facilitates career advancement, demonstrated by the opportunities for promotion, training and further development that exist at all levels. The Sanger Institute is keen to retain high-calibre individuals and can provide an environment where new and exciting opportunities are always available.

Investment in staff training has increased year on year and opportunities range from specific skills training to personal development activities. Training events enable different groups of people at the Sanger Institute to work on common problems, learn from each others' experience and, we hope, have some fun!



Graduate students

We are very active in graduate student training, with more than 30 students currently enrolled. Most students are registered for a PhD with the University of Cambridge, although a few register with the Open University. Fully funded scholarships with registration at Cambridge are available through an annual competition, advertised internationally, and through our website www.sanger.ac.uk/careers/phd. Students are recruited from all over the world, providing a rich multicultural dimension to the Sanger Institute. Beginning in October 2003, these studentships will support a four-year programme, designed to expose students to different disciplines in genomics and to provide a rich training environment for PhD studies.

Open University registered students may be part-time or full-time. Students are mainly salaried staff, whose research programme falls within the scope of their role at the Institute.

Postdoctoral programme

Numerous postdoctoral training opportunities are available at the Sanger Institute to work with any of our independent faculty. Generous support offers an exciting opportunity to begin an independent research career in state-of-the-art genomics facilities. For more information see www.sanger.ac.uk/careers/postdoc

Faculty

The Sanger Institute provides a unique and supportive environment for young and established researchers who have developed hypothesis-driven or hypothesis-generating research programmes. Our special skills in high-throughput science, the breadth and depth of research support and the stimulating environment mean that unique projects can be undertaken. The breadth of our scientific enterprise is displayed in the research interests on pages 34.

The Sanger Institute provides a unique and supportive environment for young and established researchers



Other facilities



Library

The Sanger Institute houses a first-class library with a comprehensive reference book collection, and subscriptions to all relevant current journals. There is provision for site-wide access to most journals electronically.

Administrative and technical services

The Sanger Institute is served by a number of teams providing administrative and laboratory support. Administrative departments cover human resources, health and safety, finance, purchasing and stores. There is also a fully equipped reprographics facility.

Laboratory services include oligonucleotide synthesis, media production, glassware wash-up, building and electrical maintenance and a specialist waste disposal collection.

An in-house Robotics and Automation team designs and builds robots in response to specific project needs.

Social facilities

The DiNA is a cafeteria in a communal area where staff gather at all times of the day to chat over coffee or lunch, and at Hinxton Hall there is a staff restaurant.

There is a well-equipped gym on site, an active social and sports club and a crèche.



Contacts

Management of the Sanger Institute

Executive decisions are taken by the Board of Management, a team of eight senior staff chaired by the Director, which meets every two weeks. The Sanger Institute is built around teams dedicated to the research programmes, resources and support services: each is managed by a team leader, and groups of team leaders meet regularly within and across disciplines.

Contacts

General enquiries:

Tel: +44 (0)1223 834 244
 Fax: +44 (0)1223 494 919
 E-mail: recep@sanger.ac.uk

Job enquiries, HR Officer

Sarah Golland
 Tel: +44 (0)1223 494 943

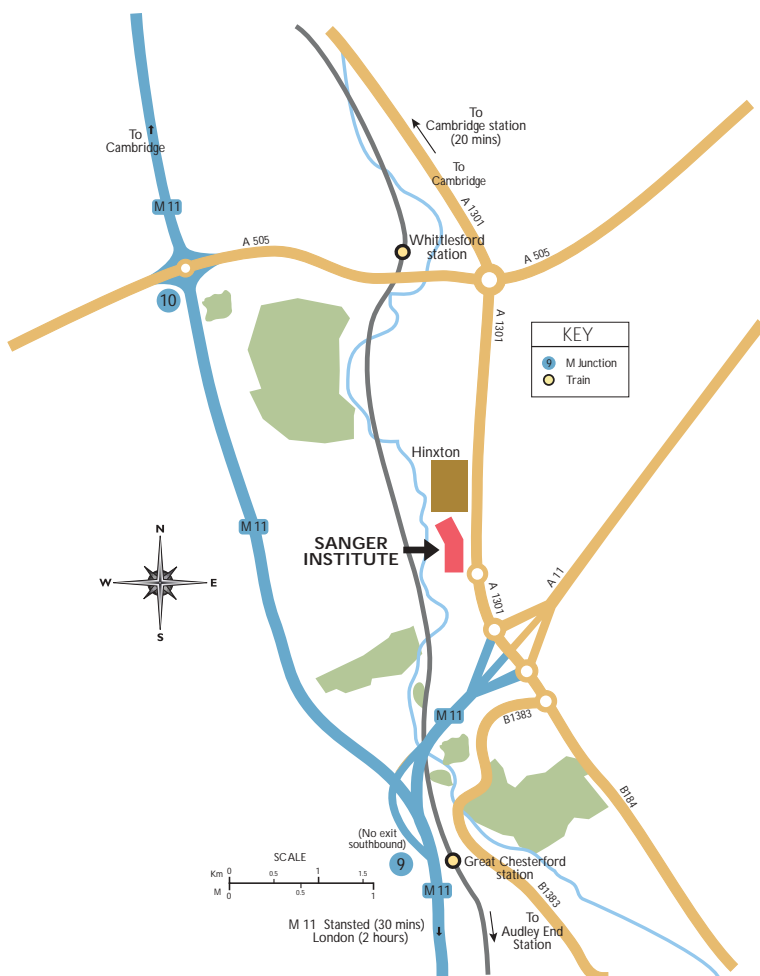
Graduate programme officer:

Christina Hedberg-Delouka
 Tel: +44 (0)1223 494 997

Media and PR officer:

Don Powell
 Tel: +44 (0)1223 494 956

www.sanger.ac.uk



Design: Sally Watts
 Photography: David Sayer

The Wellcome Trust is a registered charity, no. 210183. Its sole Trustee is The Wellcome Trust Limited, a company registered in England, no. 2711000, whose registered office is 183 Euston Road, London NW1 2BE.

The Wellcome Trust Sanger Institute is a research institute of Genome Research Limited (GRL), a wholly-owned subsidiary of the Wellcome Trust. GRL is a registered charity (no. 1021457) and a company registered in England (no. 2742969), whose registered office is 183 Euston Road, London NW1 2BE.

