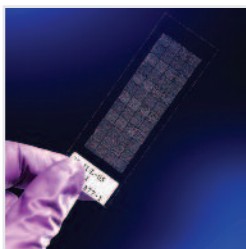


The Sanger Institute's core facilities provide expertise in a range of high-throughput technologies, from DNA sequencing to animal breeding, and are thus central to the Institute's high scientific productivity.

- IT infrastructure and data management
- Sequencing
- Sequencing informatics
- Genome analysis pipelines
- Research Support Facility
- Proteomic mass spectrometry



The microarray pipeline produces both transcript and comparative genomic hybridisation data.

The science conducted by the Sanger Institute is characterised by its large scale. This productivity is founded on its high-throughput core facilities, which include DNA and RNA processing pipelines, microarray analysis, genotyping and sequencing by conventional and next-generation technologies, and the animal handling facility, together with a mass spectrometry unit and cytogenetics facility. High-capacity informatics and IT systems analyse, store and share the data emerging from these activities.

Over the last year, the sample logistics and genotyping pipelines have handled the massive sample load and data output from the major genotyping projects undertaken by the Sanger Institute as part of the Wellcome Trust Case Control Consortium and other initiatives. They have also coped with data handling challenges created by the substantial increase in the data produced by the next-generation sequencers. A new laboratory information management system (LIMS) has been developed that tracks samples, projects and sequences through the entire pipeline.

A comparison of next-generation sequencing technologies led us to specialise on the Illumina platform. We purchased a further 11 Illumina sequencers, bringing our total to 37. We also run two Roche 454 FLX next-generation sequencers and ABI 3730 capillary machines, which are used primarily for whole-genome sequencing projects.



Loading samples in the genotyping facility.

Raw output has increased almost fivefold, to nearly 300 gigabases (Gb) per week, and is poised to increase further. The Sanger Institute is now capable of producing more than 15 terabases of sequence per year.

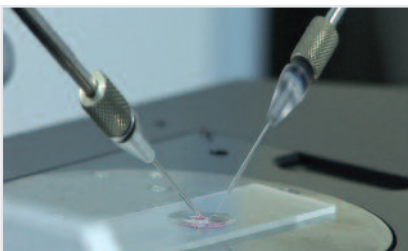
As well as adopting the new LIMS, bioinformatics analysis has moved to a rapid response software development model and web-based applications. As well as improving speed and responsiveness, this approach also eliminates any platform compatibility issues.

Our IT support services have shown a similar expansion, with disk capacity doubled to 3 petabytes (3000 terabytes) during the year and further expansion in the pipeline. An additional 2000 blade computers were added, bringing the total at the Institute to more than 5000. A major effort has been put into virtualisation, with physical servers being replaced by blade computers, generating space and energy savings.

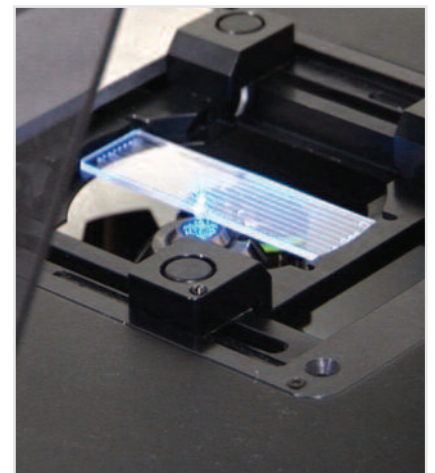
The focus of genome analysis teams has been the extra genotyping throughput required in genome-wide analyses. In addition, direct RNA sequencing techniques have begun to supersede microarray-based approaches in analyses of gene expression, a trend likely to continue into the future.

Such studies feed into genome annotation, as does work carried out by the proteomic mass spectrometry facility. The facility has also developed a new affinity tagging system so protein assemblies containing specific proteins can be isolated and characterised. This approach has been used in a variety of contexts, including an analysis of protein networks controlling stem cell activity.

The Research Support Facility increased its numbers of licensed animal technicians from 18 to over 40. It has supervised a considerable increase in the number of mutant ES cells that are converted into mice and has begun evaluating the phenotypes associated with the genetic changes in these mice. It has also been trialling a new caging system that would allow for a significant increase in capacity.



Close-up of one of four microinjection 'rigs' in the RSF. Under magnification, mouse blastocysts (3.5 day-old embryos) are injected with ~10 genetically modified ES cells. Genetic material from the ES cells may be incorporated into the embryo. Around 350 blastocysts are microinjected every week.



Raw output from next-generation sequencing technologies has reached almost 300Gb per week.

We provide high-performance IT infrastructures and data management to support large-scale science.



The Data Centre at the Sanger Institute.

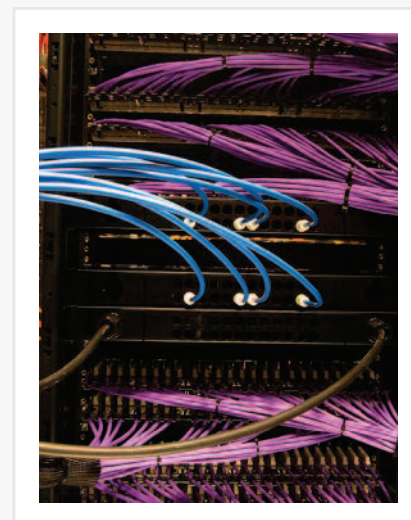
The IT infrastructure at the Sanger Institute is one of the most extensive in the life sciences. With new generation sequencers, it has continued to grow dramatically. Over the last year, installed disk capacity has doubled to 3 petabytes (3000 terabytes or TB) and we plan to add a further 800 TB in the coming year. We also added a further 2000 cores of blade computers, bringing the total number installed to more than 5000.

We have worked on an aggressive virtualisation project to reduce the number of physical racked servers. To date we have over 80 virtual servers installed on just three blade servers. Our aim is to virtualise most of our 300 racked servers into less than 20 blade computers, generating significant space and energy use savings. To support continued expansion of the IT operation, we are investigating ways to bring additional electrical power onto the Genome Campus so that we can fit-out the fourth quadrant of the data centre.

We have held discussions with all the other large-scale genome sequencing centres to discuss current challenges, and have agreed on a programme of work to investigate international models for future data sharing.

As our operation continues to grow, we will switch to a combination of on-site and off-site data facilities. This is necessary not only for disaster recovery purposes but also to establish resilient mirrored data operations (replication) as an alternative to large tape backups, which will not scale to multi-petabyte levels. We are also exploring emerging technologies such as 'cloud computing'.

Surprisingly, web traffic has not increased dramatically this year but has remained at a steady state of approximately 15–16 million hits (7–8 million page impressions) each week.



Data cabling for high-speed networks.

We test, develop and apply sequencing technologies to derive high quality genomic sequence data.

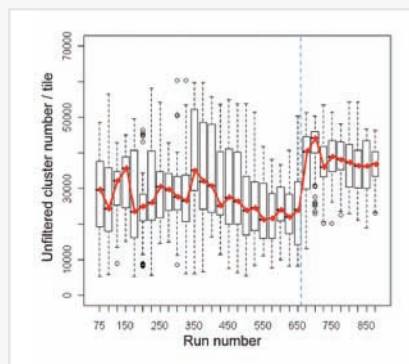


Installation of Illumina GALLx sequencing machines.

The Sanger Institute has one of the largest sequencing facilities in the world. Our raw output now approaches 300 gigabases (Gb) per week (up from 60 Gb per week last year), with further increases likely over the coming months. The Institute is capable of producing more than 15 terabases of sequence per year, the equivalent of nearly 250 human genome sequences at 20x coverage. For comparison, the total amount of sequence submitted to public databases before the advent of next-generation sequencing technology was 250 Gb.

As well as sequence-based projects, the huge numbers of short sequences provided by new technologies can be used in studies of transcriptomes (RNA-seq), protein-DNA interactions (ChIP-seq) and high-throughput transposon mutagenesis.

To assess two new sequencing platforms – the Illumina (Solexa) GAll, and the ABI SOLiD – we compared their throughput, error profile and per-base operating costs. Although they were broadly comparable, for practical reasons we have chosen to focus our efforts on a single platform, Illumina. We purchased a further 11 Illumina sequencers, bringing our total to 37.



Graph showing improvement in cluster density after introduction of qPCR for library quantification (vertical dotted line). Quail MA et al. *Nature Methods* 2008 5(12) 1005-10 doi:10.1038/nmeth.1270

Much of our effort this year has been focussed on refining protocols for the Illumina platform. Many have been taken up by other centres, and some have been incorporated into Illumina's standard commercial protocols. We also act as an advance testing site for Illumina, giving us early access to new technical developments. Several members of the group were involved in the analysis of the first full human genome sequence to be generated solely on the Illumina platform.

A sample preparation method for the Illumina platform tailored to (A+T)-rich genomes.

Kozarewa I et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 6: 291 (2009)

A summary of methods used to enhance the Illumina platform for high-throughput studies.

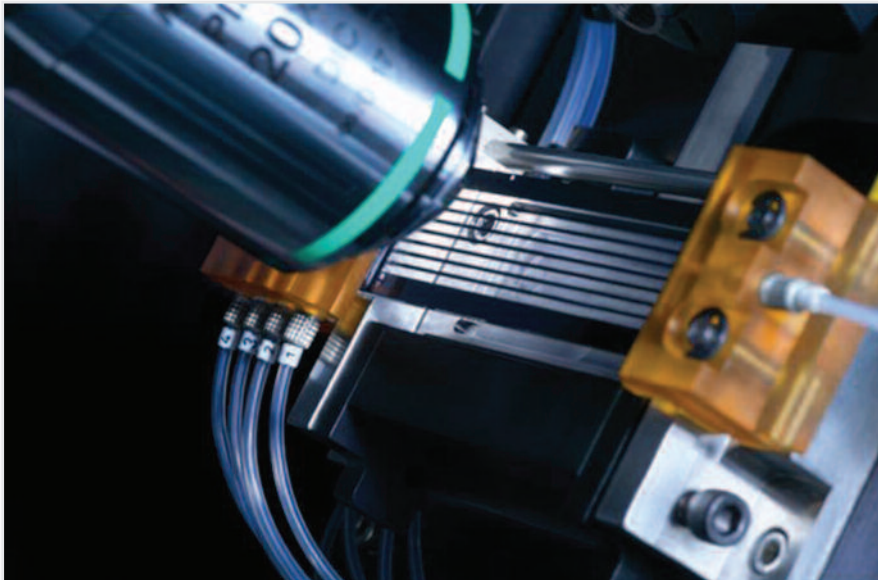
Quail MA et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5: 1005 (2008)

We also run a second next-generation platform, the Roche 454 FLX, which supports projects that require longer read-lengths. We currently run two of these instruments, which are used for projects such as bacterial *de novo* sequencing and viral sequencing.

We currently have 35 ABI 3730 capillary machines in production, down from 45 last year. These are mainly being used for the zebrafish and pig whole-genome sequencing projects, which will be completed at the end of 2009.

To ensure that the molecular biology pipelines can keep pace with this increase in sequencing capacity, we are investigating options such as commercial robotics systems to allow us to automate steps in the sample handling process.

We develop and apply informatic technologies to support the tracking of samples, projects and sequences through the sequencing pipeline.



High throughput DNA sequencing. Three billion bases of DNA sequence can be generated from DNA immobilised in this eight lane flow cell during one four-day run.

Changes in sequence production are driving new approaches to informatics support. We have developed a laboratory information management system (LIMS) that tracks samples, projects and sequences through the entire pipeline. Because of the speed of change, we have moved towards extremely flexible and modular software. We switched away from perl/TK-based applications to systems delivered entirely over the web, ensuring widespread take-up without major operating system or hardware compatibility problems.

We have also adopted rapid software development methods, in which high value features are delivered over short (typically two week) development cycles. Software feature requests are agreed with the users in advance of the development 'sprint', thereby ensuring that we are always working on the most important aspects of software.

We separated the two principal software systems into two modules: 'Sequencescape' to manage the laboratory activities and 'NPG' to monitor sequencing instruments and process sequence information as it is generated. These software modules work together but can be decoupled and developed independently.

We have also transferred our genotype tracking to Sequencescape. Combining lab tracking applications offers big advantages in code sharing as well as scientific benefits in making data across both sequencing and genotyping platforms visible in a single system. We also aim to replace our legacy capillary tracking system with Sequencescape over the next year.

The speed of change has left little time for research and development, which poses a challenge to our ability to manage expansion in future years. We are, however, attempting to evaluate the significant opportunities offered by new directions such as cloud computing and storage systems. These could offer major benefits for scientific projects that depend critically on high levels of collaborative activity and data sharing.



A section of an image from an Illumina machine. Each dot represents a DNA fragment and the colour represents the base at a specific position in that fragment.

Sanger pipelines provide high-throughput delivery in sample logistics, genome-wide data generation, quality control, data transfer and analysis.

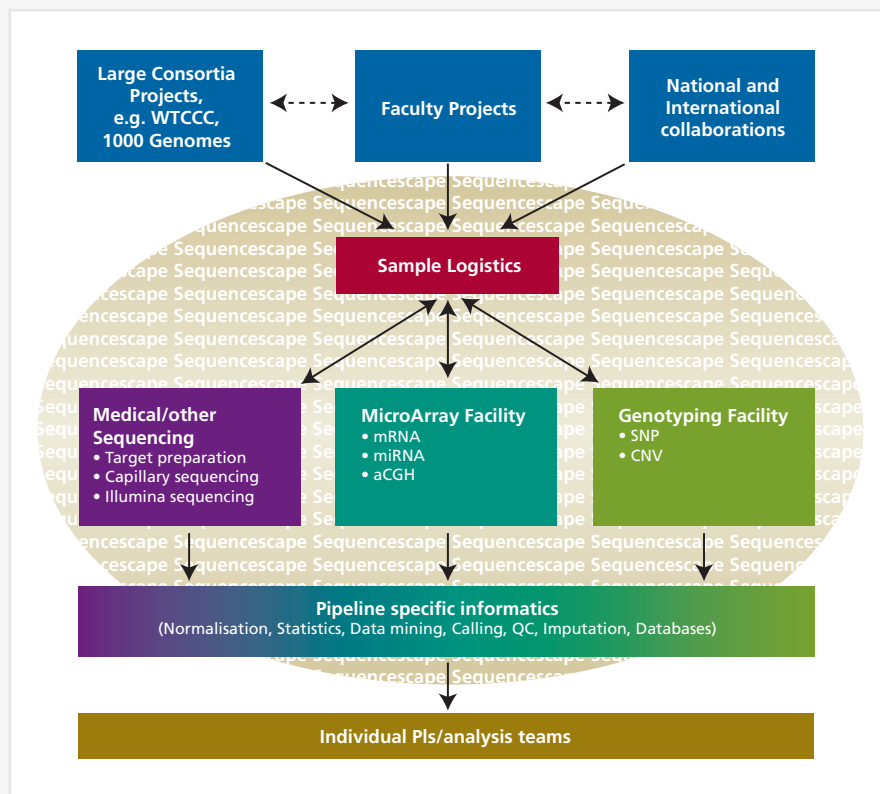
The sample logistics pipeline performs DNA or RNA extractions from a range of substrates and quality assures all samples sent to the Institute for genotyping or sequencing. Among several innovations, we have introduced a new laboratory information system and further increased the level of automation and developed tools to minimise manual data transfer and manipulation.

The genotyping pipeline utilises Illumina, Affymetrix, Sequenom and Taqman platforms. The main platform for genome-wide association studies uses the Illumina 660W quad BeadChip at a throughput of 3000 samples per week. Some 51 700 samples were processed for phase 2 of the Wellcome Trust Case Control Consortium, and a similar number were generated for faculty projects. Several other significant projects, both UK-based and international, have been completed and numerous cohort studies are complete or in progress. In collaboration with Gordon Dougan, bacterial samples from the *Salmonella* Typhi study have been typed.

The variation informatics group provides quality control and analysis for genotyping projects. An analysis pipeline exists for genome-wide association studies data. Further analysis, most commonly imputation, is provided according to the needs of the project. A secure website has been developed to allow external collaborators access to data.

The microarray facility is designed to produce data for small to medium scale Faculty projects reaching from tens to a few hundred samples. The pipeline produces both transcript and comparative genomic hybridisation data using commercial Illumina, Affymetrix and Agilent platforms. In 2009 the pipeline is projected to produce data for more than 3000 samples.

The microarray informatics team performs large-scale microarray data analysis. An Illumina sequence analysis pipeline for estimating RNA abundance levels, splice variants and sequence variants in the transcriptome has been constructed.



The Genome Analysis Pipeline. Sample Logistics co-ordinate sample import and delivery to the Genotyping, Microarray or Sequencing laboratories. Projects are tracked via Sequencescape LIMS, and Informatics support is provided for each application.

We maintain and care for the animals used in the Sanger Institute's research programmes.

As well as basic tasks such as husbandry, we also work closely with research teams on more sophisticated interventions such as breeding and colony management, procedures such as mouse ear clipping (and fish fin clipping) for genotypic analysis, tumour watch, irradiation, infectious challenge, tissue harvesting, blastocyst microinjection, mouse embryo cryopreservation, and re-derivation of new lines. We also coordinate animal exports, facilitating the Institute's strategic goal of advancing the scientific enterprise across the world.

We are committed to the principles of the 3Rs (replacement, refinement and reduction), which underpin all our current work and development plans. Our commitment to animal welfare has been recognised by the Home Office Inspector, who has used us as an example of a well-run facility.

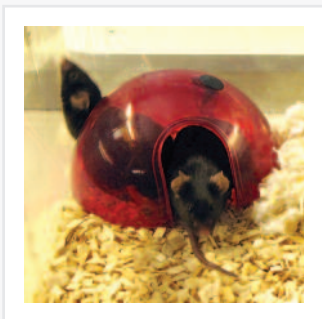
We provide training to staff from the research teams in specialised techniques, such as intraperitoneal injection and scrotal vasectomies (refinement). In the last year, we have helped to introduce new techniques, for example testing sperm-freezing methods for zebrafish (reduction) and the introduction of germ-free isolators for mice. We have also opened a new mouse holding area supporting the Malaria Programme.

We have been trialling a new caging system that, if fully implemented, would increase our overall capacity from 20 800 to 29 000 cages, while still meeting UK guidelines for stocking density.

Animals receive the best treatment possible – essential not only for their welfare (refinement) but also for the integrity and strength of our scientific results. All mice are housed in individually ventilated cages. They are tracked on a database, enabling us to implement a colony management regime that optimises the number of animals used (reduction). We can also track health and welfare concerns in a mouse-specific manner, enabling us to detect trends that may need further investigation (refinement).

The health and well-being of staff are of paramount concern. We promote a healthy work environment and strengthen staff's technical capabilities. We have implemented Laboratory Animal Allergen monitoring that permits us to identify areas where engineering controls are necessary and those where personal protective equipment is more suitable.

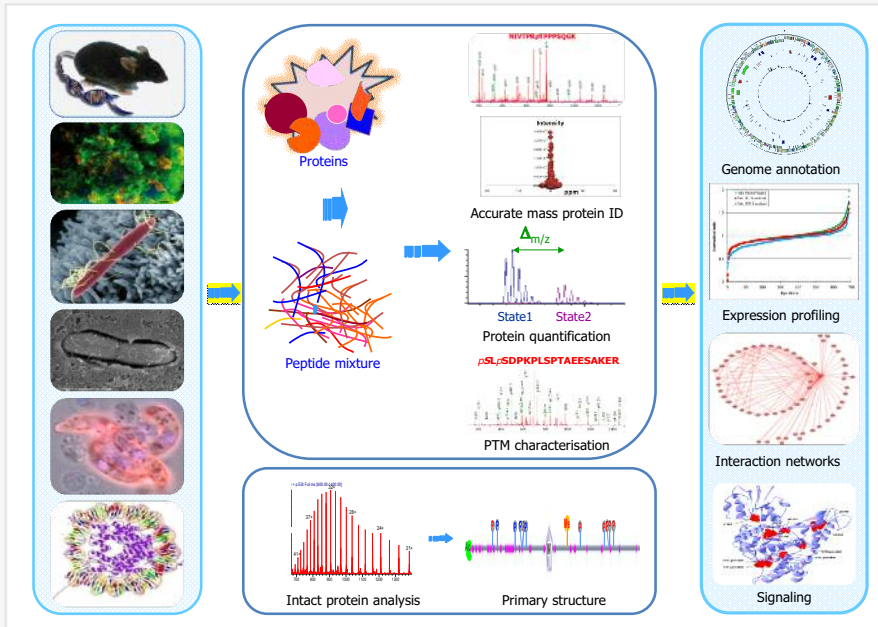
Staff members are given time and support to develop their skills, and we have implemented a training programme for new starters. During the year, the number of licensed animal technicians increased from 18 to over 40.



Mice exploring environmentally enriched cages in the Research Support Facility.



We develop and apply sensitive mass spectrometry-based approaches for proteome characterisation to complement genomic studies and genome annotation, as well as for mapping protein interactions and signalling pathways.



Overview of the Mass Spectrometry facility.

Mass spectrometry is a highly versatile analytical technique for rapid quantitative identification of proteins and post-translational modifications. It has become established as a key tool for protein expression profiling, characterising protein interactions and elucidating protein structure.

Our programme combines a range of disciplines (mass spectrometry, biochemistry, molecular biology and informatics) to facilitate proteome description and measurement. Our research on mass spectrometry and informatics methods aims to improve data capture and analysis. A substantial part of our technology development has also focused on affinity enrichment methods, for targeted proteome analysis.

In collaboration with Allan Bradley and Bill Skarnes, we have developed eTAP (endogenous tandem affinity purification) tagging technology, in which the endogenous gene is modified to include two small affinity tags, so protein assemblies associated with the targeted gene can be specifically recovered. We have validated this approach for mapping protein interactions, both for individual genes as well as for systematic large-scale applications.

We are applying eTAP-MS to several chromatin remodelling and transcription factors, to unravel the regulatory protein networks that control stem cell processes. These data have contributed to international programmes (EuTRACC and the International Regulome Consortium) on stem cell regulation and maintenance.

Proteomic analysis has been especially powerful in the Genes to Cognition programme's studies. Our work has focused on describing the molecular architecture, organisation and signalling networks at the synapse. Follow-up studies using our targeted proteomics approaches are underway to dissect quantitative changes in mutant mice as well as in human synapse proteomes.