

The Sanger Institute's research is underpinned by world-leading central facilities providing support for high-throughput sequencing, genome analysis, proteomics, data management and animal breeding.

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

➤ Core facilities



Genome Research Limited

Raw output from next-generation sequencing technologies has tripled to almost 1 terabase (1000 billion bases) a week.

The new era of genomic research is marked by an ever-increasing ability to generate, analyse and distribute huge volumes of data. It is a major challenge not only to keep up to date with technological developments but also to develop integrated systems to ensure that productivity gains are not lost due to bottlenecks in input pathways or data handling. Rigorous quality assurance must also be a high priority.

Next-generation sequencing platforms have enabled the Institute to dramatically increase its output. We are now fast approaching 1 terabase (Tb; 1000 billion bases) of raw sequence output every week. With the acquisition of additional capacity, output is anticipated to increase a further threefold in 2011, to 200 terabases a year (equivalent to 2500 complete human genomes sequenced at 30x coverage).

To ensure the Institute's flexibility to deliver its science, we operate a range of sequencing systems. While Illumina machines are our main platform, we maintain a range of solutions to provide us with the ability to deliver data in the most efficient way to meet the needs of Faculty. As a leading-edge genomics research institute, we are at the forefront of guiding and developing sequencing innovations. We are being given early access to third-generation sequencing instruments to formulate methods and protocols to exploit the power of these new processes before full deployment.

Next-generation technologies have also facilitated other high-throughput experimental approaches, including analysis of transcriptome and protein-DNA interactions and large-scale transposon mutagenesis.

Genotyping has continued to be a major area of work, with nearly half a million samples genotyped during the year for projects such as the Wellcome Trust Case Control Consortium and various international consortia. Some 2500 samples were processed on microarrays for gene expression or comparative genomic hybridisation analysis for small-scale faculty projects. Plans have been put in place to support high-density genotyping chips and fully automated platforms are being developed to maximise throughput and efficiency.

A particularly significant innovation has been a new 'pull-down' pipeline, enabling specific sequences, such as key sequences from gene-containing regions of a sample, to be selectively targeted. Typically used for human disease studies, this approach increases productivity tenfold. Throughput increased to around 200 exomes a week, and is set to rise further in 2011.

To keep pace with these step changes in output, software infrastructure has had to be significantly re-engineered. The main sequencing pipeline now has a 1 petabyte (1000 terabyte) storage array based on a farm of more than 1000 CPUs.

Similarly, storage capacity more than doubled over the year, to over 10 petabytes (10 000 terabytes). Computing capacity also grew markedly, to more than 8000 blade computer cores. Within the next few years, demands are likely to increase to 25–30 petabytes of storage capacity and 20 000 cores of computing capacity – uncharted territory for IT in life-science research.

Alongside these genomic pipelines, the Sanger Institute also has significant capability in mass spectrometry for proteomic studies. This expertise has been applied in several areas, including analysis of protein networks in the synapse and in stem cells and phosphopeptide detection.

High-throughput mass spectrometry is also being used for protein-level annotation of gene sequences. An analysis of 10 million spectra, from external and internal sources, identified many unsuspected protein products from known genes as well as 10 entirely novel coding genes. New data will be fed into the Ensembl genome browser.

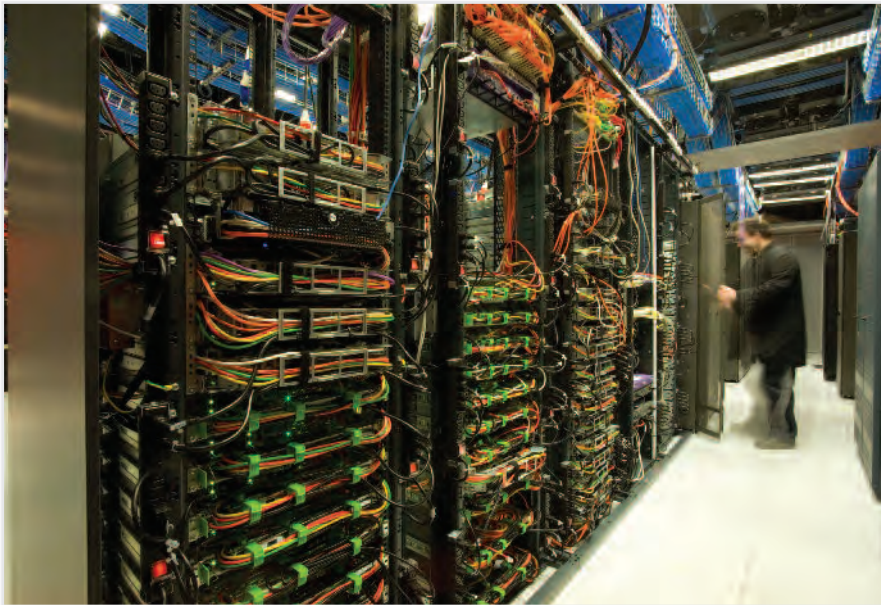
Biological studies rely heavily on the Research Support Facility. The Facility is one of the largest in the UK. With implementation of a new mouse database, the average number of cages per colony has been significantly reduced and the number of colonies increased. Such changes have enabled more efficient use of animals, and illustrate the Facility's commitment to the '3Rs' (replacement, refinement and reduction).

As well as hosting visits, the Facility also organised a Managing Mouse Colonies course which has now become an annual feature in the portfolio of Wellcome Trust Advanced Courses to which Core facilities teams contribute.



Our new mouse database had enabled us to reduce significantly the average number of cages per mouse colony, while increasing the number of colonies.

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



Genome Research Limited

Our virtualisation programme has been a great success, and we are aiming to accommodate more than 200 virtual servers in a resilient configuration.

The IT infrastructure within the Sanger Institute's data centre continues to expand exponentially in support of next-generation sequencing technologies. This year we have more than doubled our disk storage capacity, to over 10 petabytes (10 000 terabytes). Our computing capacity has also grown significantly, to over 8000 blade computer cores used for high-performance data analysis.

We are likely to require around 25-30 petabytes of storage capacity within the next five years, as well as at least 20 000 cores of computing capacity for data analysis.

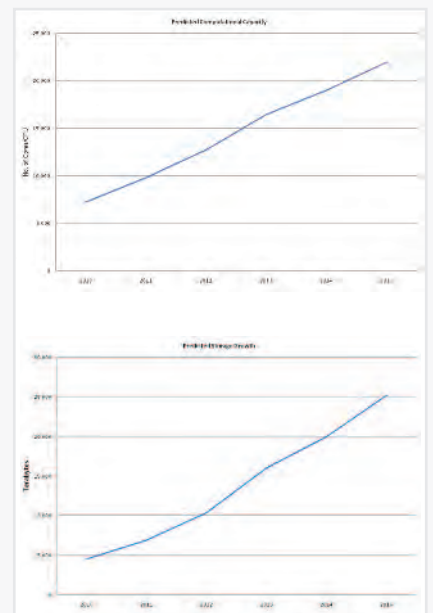
Our immediate priority is to plan for an anticipated fourfold increase in sequence production over the coming 12 months. One of our aims is to improve our onsite/offsite architecture and provide a higher degree of data replication to reduce risk for data held onsite.

Unsurprisingly, disk storage and data management are our highest priorities, and new management tools are in development to monitor and track the vast array of data held. We also liaise with international groups to provide input on data management tools being developed for the open source community.

We keep up to date with current technologies and new developments by maintaining contact and exchanging ideas and experiences with other large-scale genome sequencing centres. These are invaluable as we move into uncharted territory for this scale of IT in the life sciences.

Our virtualisation programme has been a great success over the past 12 months, and our current hardware system is being expanded to accommodate more than 200 virtual servers in a resilient configuration. Consolidation continues in this area to reduce the server hardware footprint, lower management overheads and achieve power savings.

The Sanger Institute website continues to be the first port of call for individuals and groups using our data. Web traffic has remained at a steady state of approximately 16 million hits (9.7 million page impressions) each week during the year.



Genome Research Limited

We predict that the Institute's increasing sequence production will require 25-30 petabytes of storage capacity within the next five years.

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



Dave Sayer, Wellcome Trust

Every week we generate some 1000 billion bases of sequence data – four times the world’s entire data output before 2008.

The Sanger Institute has one of the largest sequencing facilities in the world. Our raw output now approaches 1 terabase (Tb; 1000 billion bases) per week and, with the purchase of 20 Illumina HiSeq instruments, we can reasonably expect this output to at least triple during 2011. The Institute is capable of generating more than 200 terabases of sequence per year, the equivalent of 2500 whole human genomes sequenced at 30x coverage. For comparison, the total amount of sequence submitted to public databases before the advent of next-generation sequencing technology was 0.25 Tb.

As well as sequence-based projects, the huge numbers of relatively short sequences provided by the newest technologies can be used in studies of transcriptomes (RNA-seq and flowcell-based RNA Seq; FRT-seq), protein–DNA interactions (ChIP-seq) and high-throughput transposon mutagenesis. The latter application has enabled the Pathogen group to map all the essential genes for an organism, such as a medically relevant bacterium, within a single lane of an individual sequencing run.

A new ‘pull-down’ pipeline has been developed that allows us to pull out and selectively target specific sequences, such as key sequences from gene-containing regions of a sample. We typically use this for human disease studies, as it enables us to sequence at approximately 10 times lower cost or to analyse 10 times as many individuals for the same expenditure.

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

We currently have 12 ABI 3730 capillary machines in production, down from about 35 last year. These are mainly being used for the pig X/Y chromosome project, for the Genome Reference Consortium (page 69) and for verifying clones made for mouse embryonic stem cell knockout projects.

Into the future, we will take delivery of our first third-generation system, the PacBio RS sequencer. Although not yet competing with the Illumina platform in terms of cost and throughput, the RS should enable very long read lengths, up to many thousands of bases. Potential applications include highly rearranged tumour genomes, bacterial and viral sequencing, and assembly of novel genomes.

To ensure that the molecular biology pipelines can keep pace with this increase in sequencing capacity, we have purchased a number of commercial robotics systems to allow us to automate steps in the sample-handling and library-making process.

➤ **A novel technique for performing transcriptomics/ expression analysis directly on an Illumina flowcell.**

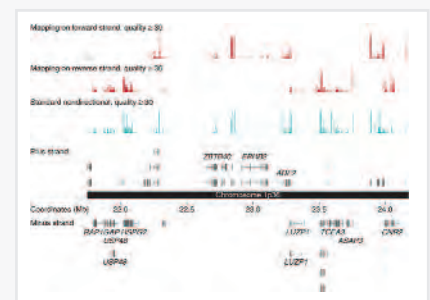
Mamanova L et al. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 2010; 7:130–2.

➤ **A review of methods for sequencing selective portions of genomes.**

Mamanova L et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010; 7:111–18.

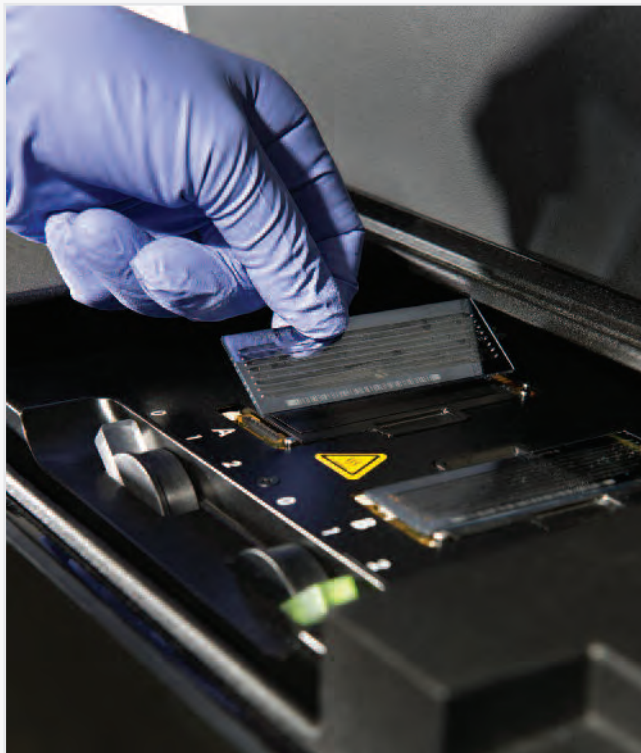
➤ **A simple high-throughput method to map all the essential genes in an organism using next-generation sequencing.**

Langridge GC et al. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res* 2009; 19:2308–16.



FRT-seq strands are created without amplification, avoiding standard process biases. FRT-seq sequences (red) compared with the standard preparation protocols and flowcell amplification (blue).

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



Dave Sayer, Wellcome Trust

Next-generation high-throughput sequencing using flow cells in an Illumina HiSeq machine. To cope with the data generated, the sequencing pipeline relies on a 1000-CPU storage array with a capacity of 1000 terabytes.

Large-scale genome resequencing has become a dominant area of genome research over the last year. The Sanger Institute is at the forefront of these efforts and is running projects on a scale almost unimaginable only a few years ago.

Next-generation sequencing technology has advanced rapidly over the last year and rates at which data are produced have soared. To be able to capitalise on the vast amounts of data these huge experiments generate, we have again had to re-engineer much of our software infrastructure to cope with increased demand for data transfer, safe storage, rapid analysis and delivery to the public domain.

The main sequencing pipeline now has a 1 petabyte (1000 terabyte) storage array that allows us to buffer and process sequence data using a farm of more than 1000 CPUs.

This rapid evolution of technology is not restricted to data-processing pipelines. Laboratory processes are constantly evolving to increase the efficiency and quality of our data output, and the software to support and track lab operations must keep pace with these changes.

This year we have united the IT systems for DNA sample reception, handling and quality assurance operations for sequencing and genotyping pipelines. This means we can now offer a unified system to manage projects across these two key operational areas.

The Sanger Institute is acquiring its first 'third-generation' sequencing instruments. Although they will initially be used as a research platform, our experience with these devices will give us insight into the direction that DNA sequencing technology will go and allow us to develop plans to support and exploit it when it becomes generally available as a large-scale platform.



Dave Sayer, Wellcome Trust

The Sanger Institute is about to acquire its first 'third-generation' sequencing machine, which uses single molecule sequences in real time on chips instead of using flow cells, avoiding the need for PCR amplification.

We are responsible for preparation of samples, data generation and quality control for high-throughput exome sequencing, genotyping and microarray analysis.



Genome Research Limited

Genotype BeadChip during imaging. Lasers scan the chip in swathes to detect genomic variation at the resolution of a single nucleotide.

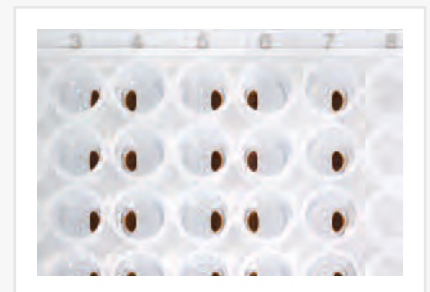
High-throughput genome analysis depends on reliable and efficient large-scale preparation and tracking of samples as well as management of data generated by sequencing, genotyping and other analyses. A sample logistics team manages import of quality-assured, bar-coded samples, which are formatted and delivered to the DNA pipelines for genotyping or sequencing.

During the past year, more than 470 000 samples have been genotyped using Illumina, Sequenom and Affymetrix platforms. We have delivered data for a range of initiatives, including Wellcome Trust Case Control Consortium studies, international consortia such as InterAct (diabetes), IMSCG (multiple sclerosis), MalariaGen and faculty-driven projects such as the Serious Adverse Event Consortium (SAEC), Morgam (cardiovascular risk), Cardiogenics and arcoGEN (arthritis). Equipment and processes have been updated to prepare for the next generation of high-density genotyping chips, including the Illumina 2.5 million SNP chip.

Some 2500 samples were processed on microarrays for gene expression or comparative genomic hybridisation analysis for small-scale faculty projects.

We have established a new production pipeline for pull-down and exome sequencing. New protocols and processes have been developed, with priority given to automation-based scalability, sequence yield and quality. New automation is being introduced to increase production capacity for bar-coded ('indexed') captured libraries feeding into the Illumina sequence production pipeline. Initially, our throughput will be 200 exomes per week in support of major programmes such as the UK10K project, the Zebrafish Mutation Resource, the Cancer Genome Project and the Deciphering Developmental Disorders project, as well as additional faculty projects, particularly those linked to analysis of human variation and medical resequencing. Capacity will be increased further in the second quarter of 2011.

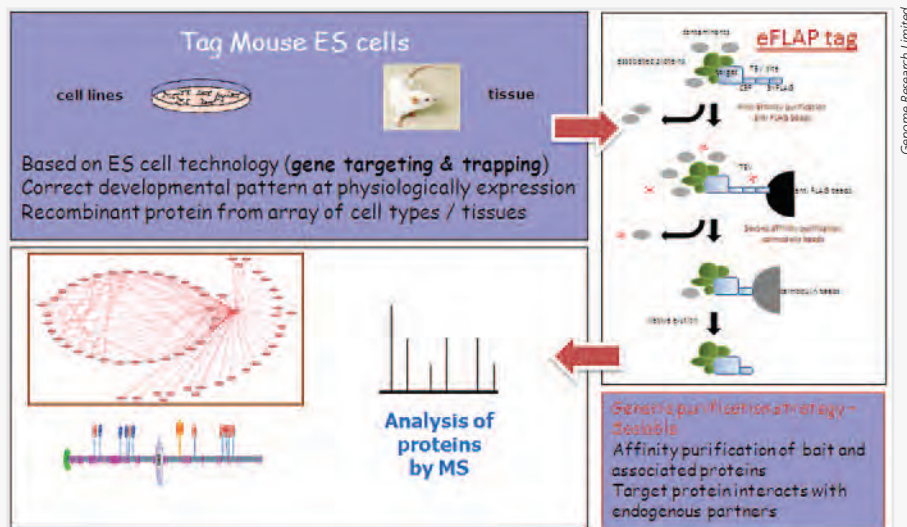
Informatics support is provided for sample tracking, data management, data formatting, quality control, preliminary analysis, data release and secure export. Robust pipelines exist for all of the main production activities, and will be reviewed and updated as our genome analysis production technologies evolve.



Dave Sayer, Wellcome Trust

Exome target enrichment using DNA probes labelled with magnetic beads. The probes bind with the genomic regions of interest and are pulled down using a magnet. The probes are washed and the beads removed, leaving the targeted DNA fragments for sequencing. A new production pipeline will bring sequencing capacity up to 200 exomes per week.

We develop and apply sensitive mass spectrometry-based approaches to proteome characterisation, to enhance understanding of gene function, protein interactions and cell signalling.



Genome Research Limited

The eTAP approach can be widely used and offers a systems biology description of protein interactions. This will be a driver in understanding at molecular level gene function and regulation mechanisms.

Many key aspects of gene function and regulation – such as post-translational modification and protein–protein interactions – can be directly explored only at the protein level. Mass spectrometry is a highly sensitive and versatile technique for characterising and quantifying proteins and their modification states.

The proteomics group supports diverse research activities within the Sanger Institute. Proteogenomics, high-throughput protein sequencing, is used to complement genome annotation and expression studies. Interaction proteomics, based on affinity purification mass spectrometry, maps native protein interactions to elucidate gene function and regulatory mechanisms. Pathway proteomics uses comparative phosphoproteomics to dissect cell signalling.

Mass spectrometry generates large amounts of data on the translation products of genes, but routine use of data for genome annotation has been limited. To address this issue, we are developing computational tools for large-scale data analysis. As well as more precisely defining gene products, these methods can also detect novel protein-coding sequences. Processing of more than 10 million spectra collected on mouse

proteomes, internally and from the public domain, revealed multiple alternative translation products as well as 10 entirely novel coding genes. This platform has also been used to characterise bacterial proteomes and phage proteins in *Salmonella*.

The eTAP (endogenous tandem affinity purification) tagging technology, developed in collaboration with Allan Bradley and Bill Skarnes, can be used to recover native protein assemblies from mouse embryonic stem cells or tissues. We have reported use of this technology to characterise protein complexes from synapses in the mouse brain, identifying schizophrenia susceptibility proteins, and to map interactions within the Oct4 network.

To support cell signalling studies, we have developed highly reproducible affinity-capture methods for phosphopeptides, which have enabled profiling of dynamic changes in phosphorylation. We are pursuing improved data acquisition and analysis procedures that will dramatically enhance our characterisation of phosphorylation. Initial studies, on the malaria parasite, suggest that sensitivities will match those achieved for whole proteome analysis, permitting us to explore signalling pathways at unprecedented depth.

Targeting of the Vi protein of *Salmonella* Typhi by a surprisingly diverse range of phage depends on a distinctive protein domain shared by all the phage.

Pickard D et al. A conserved acetyl esterase domain targets diverse bacteriophages to the Vi capsular receptor of *Salmonella enterica* serovar Typhi. *J Bacteriol* 2010; 192(21):5746–54.

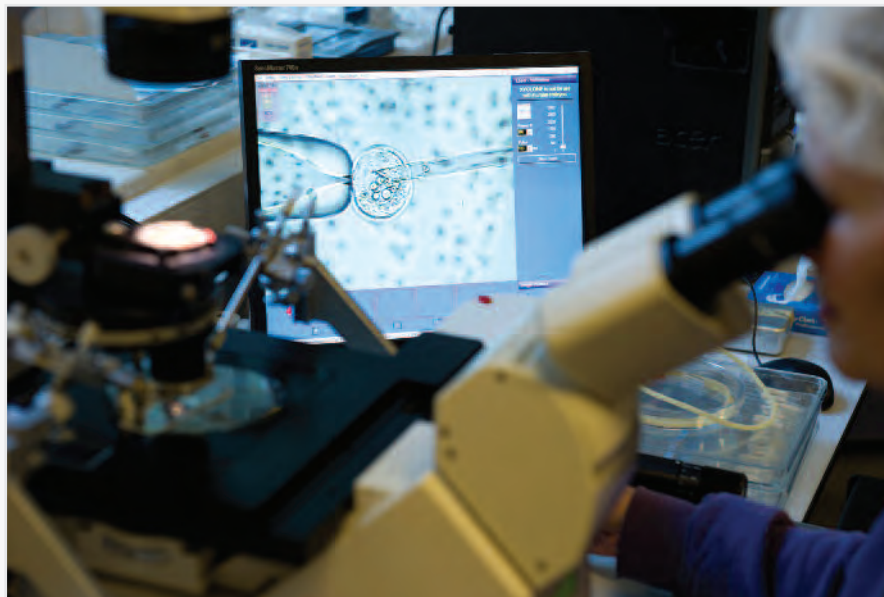
Use of the 'eTAP' technique to dissect a key synapse protein complex.

Fernández E et al. Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Mol Syst Biol* 2009; 5:269.

A network of almost 100 proteins interacting with the Oct4 stem cell factor will shed light on stem cell biology, embryonic development and human disease.

Pardo M et al. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell* 2010; 6(4):382–95.

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



Dave Sayer, Wellcome Trust

Microinjection in the Research Support Facility. We are involved in many collaborations with the external scientific community and have distributed 643 mouse strains in the past two years.

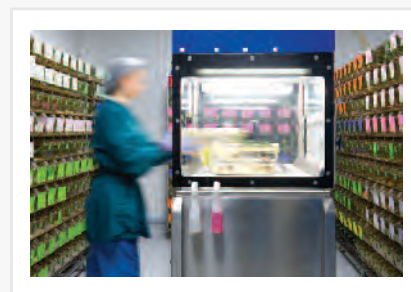
The last year has seen full implementation of our Mouse Database, which has allowed us to apply increasingly sophisticated breeding strategies for colony management. Using the system, we have been able to reduce the average number of cages per colony from 19 to 14, and increase the number of active colonies from 600 to nearly 1100. This not only achieves more efficient use of space but also meets the goals of the 3Rs, reducing the number of animals used. We hope to adapt the system for use with zebrafish.

The facility is approaching its working capacity and is now fully staffed, with 61 team members providing high-quality professional care and support. The Microinjection and Cryopreservation team has worked to refine and enhance its technological capabilities for production and archiving of genetically modified animals. Promising progress has been made with sperm freezing, even for the problematic C57BL/6N substrain. The team is also incorporating new techniques into daily routines, including non-surgical embryo transfers and inhaled anaesthesia.

With the increase in the facility's capabilities, we have a growing reputation as a world-leading facility. Visitors from all over the world come to view our facilities, databases and strategies for colony management, phenotyping and experimental manipulation. This commitment to training – of both technicians and research scientists – extends to a new Managing Mouse Colonies course, held for the first time last year and now incorporated into the Wellcome Trust Advanced Courses programme.

The facility has established numerous collaborations with the external scientific community, and the number of requests for modified organisms continues to grow. In the past two years, we have distributed 643 mouse strains across all programmes (including the Mouse Genetics Programme). We now account for around 4 per cent of UK procedures on animals (80 per cent of which simply involve breeding of animals). By removing the need for groups to develop their own strains, our distribution has potentially saved more than 300 000 animals.

Finally, we have also contributed to many EU programmes, including Infrafrontier, an EU-wide initiative linking mouse research facilities. With the Wellcome Trust, we are contributing to the new EU directive on the use of animals in research, which will need to be incorporated into UK law by January 2013. This will be a challenge, but we remain committed to achieving the highest possible standards of welfare for our animals.



Dave Sayer, Wellcome Trust

Technician cleaning mice cages. The Managing Mouse Colonies training for researchers and technicians has been incorporated into the Wellcome Trust Advanced Courses programme.