

The bioinformatics programme develops and applies methods to process, store and analyse data generated by high-throughput projects. Our principal aims are to infer genomic knowledge through computational analysis and integration of data and to generate resources of lasting value to biomedical research.

- Protein and RNA databases
- Population and comparative genomics
- Genome informatics and 1000 Genomes Project
- Ensembl and vertebrate annotation
- Genome Reference Consortium

The dominant theme in 2008/09 has been the impact of next-generation sequencing on both infrastructure and research projects. Projects such as the 1000 Genomes Project are collecting orders of magnitude more data than previous sequencing projects. Richard Durbin is leading (with David Altshuler of the Broad Institute) the 1000 Genomes Project, an international consortium that aims to identify essentially all polymorphism (genetic variation down to 1% allele frequency) in humans in multiple populations. The Sanger Institute has led the mapping and variant calling for the project. Data handling is coordinated with the European Bioinformatics Institute (EBI) which, with the National Center for Biotechnology Information (NCBI), is the Data Coordination Centre for the consortium.

The theoretical underpinning of the 1000 genomes strategy has come largely from the Durbin group, which during the year has published the results of pilot projects in yeast (70 strains from two species; see box: The history of yeast) and human (one individual sequenced to high depth in collaboration with Illumina).

Another rapidly growing application of next-generation sequencing is transcriptomics, where gene expression in individual cell lines, tissues, developmental stage and individuals from a population can be determined and compared. Such datasets will have a significant impact on genome annotation, and during the year informatics pipelines have been developed for processing transcriptome data as part of vertebrate genome annotation. These methods are initially being applied to a large set of zebrafish transcriptome samples generated by Derek Stemple's group.



Wellcome Library, London

More than 90 000 distinct computers access the Ensembl genome browser through its website every month.

Transcriptomics is being widely adopted within initiatives such as ENCODE – an international consortium set up by the US National Human Genome Research Institute to identify all functional elements in the human genome. The Sanger Institute participates in ENCODE as leader of GENCODE, which is concentrating on gene-related information. Under GENCODE a competition (RGASP or RNAseq Genome Annotation Assessment Project) is being organised to assess techniques for interpreting transcriptomics data. This is modelled on the successful EGASP (ENCODE Genome Annotation Assessment Project) competition organised at the Sanger Institute to assess gene annotation methodologies in 2005.

Tim Hubbard, Head
Alex Bateman
Manolis Dermitzakis
Richard Durbin



Wellcome Library, London

DNA bases represented by the letters. A, C, G, and T represent the four bases that encode the information in a DNA double helix.

Transcription is also being assayed to investigate the functional significance of genetic variation. Manolis Dermitzakis's group has been using array technology to profile gene expression in cell lines of most of the extended HapMap samples (850) from eight different populations. This has shed light on the impact of genetic variation on gene expression in diverse populations, and also on the patterns of population differentiation and recent natural selection. An analysis of three different cell types has also provided insight into tissue-specific effects of variation (see box: Ups and downs).

Ups and downs

Many genetic variants affecting gene expression act in a tissue-specific way.

As the extent of genetic variation becomes more apparent, it becomes increasingly important to understand the impact of that variation. Although this can be examined on a case-by-case basis, genome-wide studies are a valuable complement, generating information about many thousands of sites of variation.

Manolis Dermitzakis and colleagues have previously examined the extent to which genetic variation affects gene expression, a step towards understanding the biological impact of variation. Working with Stylianos Antonarakis in Geneva, his team has now extended this analysis, examining how genetic variation affects gene expression in different cell types.

The project looked at how genetic variants affected gene expression in three different cell types – fibroblasts and transformed B-cells and T-cell lines, all derived from cord and cord blood samples, respectively. Variants influencing gene expression were found next to around 1000 genes, and in the majority of cases, they had different effects in different cell types. Around a quarter of the variants affected expression only in fibroblasts, and similar proportions affected only B cells and T cells.

The results emphasise that studies examining the effects of variation on gene expression will need to be carried out on multiple cell types, otherwise some effects may be missed. It also highlights and explains the extent of tissue specificity of genetic disorders, largely due to the tissue-specific effects of their causal variants.

Dimas AS et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009 Jul 30; doi:10.1126/science.1174148. PMID: 19644074

The Dermitzakis group started the profiling of 1000 individuals from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort and multiple tissue samples from TwinsUK. It also used transcriptomics sequencing to initiate the profiling of 60 individuals from the 1000 Genomes Project. In May 2009 Manolis Dermitzakis moved to take up a Chair at the University of Geneva after a very productive five years at the Sanger Institute. Following Anton Enright's move to the EBI last year we have been conducting a faculty search and have recruited Ville Mustonen as a junior group leader. Dr Mustonen will start in October 2009.

The relentless growth in sequence data, which has been dramatically accelerated by next-generation sequencing, presents huge challenges for databases and algorithm development. Algorithms such as MAQ, developed by Richard Durbin's group to map sequence to reference genomes, have had significant impact and become *de facto* standards. The development of BWA represents a further step forward. The Durbin group is also making important contributions to the development of improved formats for storage and analysis of data.

Detecting distant relationships between sequences, which is a necessary step in classifying them, is an even harder algorithmic problem and is central to the Bateman group's activities. The newly introduced HMMER3 software, developed by Sean Eddy in the Pfam consortium, is 100 times faster and more sensitive, offering dramatically faster searches for users, as well as the ability to identify more distant biological relationships. Another highlight has been the pioneering use of Wikipedia by the Rfam team to facilitate community annotation. The success of this innovation is likely to encourage other databases to adopt similar models.

The Sanger Institute is now jointly responsible for the reference genomes of human and mouse as part of the Genome Reference Consortium, which was officially announced to the community in May 2008. Having established experimental and informatics infrastructures, in March 2009 the Consortium released a first update of the human genome reference assembly (GRCh37). Data from the 1000 Genomes Project, and other human genome sequence information, will be used to refine the human reference sequence further, with the identification of significant haplotypes in the human population that are not currently represented.

More than 90 000 distinct computers access the Ensembl genome browser through its website every month, generating millions of web hits each week. One of the 2008/09 highlights of the Ensembl and vertebrate annotation programme has been the release of a new web interface to Ensembl, which is significantly faster and easier to navigate. To further improve speed of access, an Ensembl mirror is now running at uswest.ensembl.org in California. Several new projects are reusing Ensembl software infrastructure, including the 1000 Genomes Project website (browser.1000genomes.org) and the EBI project Ensembl Genomes, which provides access to non-vertebrate genomes (www.ensemblgenomes.org).



Wellcome Library, London

A scientist examining a map of the complete *Campylobacter jejuni* genome. This bacterium is a causative organism of food poisoning and is one of the most common causes of diarrhoea in the world.

Updates to the next-generation sequencing machines and improved chemistry have meant that each machine produces around four times as much data as a year ago, requiring additional investment in IT infrastructure. The amount of raw sequence data being transferred between centres in the 1000 Genomes Project and other initiatives is becoming comparable to that generated by large high-energy physics experiments such as the Large Hadron Collider at CERN. Indeed, movement of biological data, especially from the Wellcome Trust Genome Campus, has at times been a dominant user of the UK academic network.

This growth underlines the need for investment in national, European and international bioinformatics infrastructure, as the value of the data is maximised through broad access and use. Up to now, biology projects have received much less informatics infrastructure investment than physics projects such as the Large Hadron Collider. The Sanger Institute is therefore supporting the EBI-centred ELIXIR project to develop a long term European infrastructure for biological information. ELIXIR is recognised as one of the priority European infrastructures of the ESFRI process (European Life-science Infrastructure for Biological Information).



Wellcome Library, London

The Data Centre at the Sanger Institute.



The history of yeast

A comparison of 70 strains of yeast has shed light on its domestication.

Yeasts have been used by humans for thousands of years. Of particular importance is baker's yeast, *Saccharomyces cerevisiae*, which has come to be the dominant species used in cooking and brewing worldwide.

The availability of a complete genome sequence opened up the possibility of *S. cerevisiae* population genomics – linking genetic variation to geographic distribution, and thereby deciphering its evolutionary history.

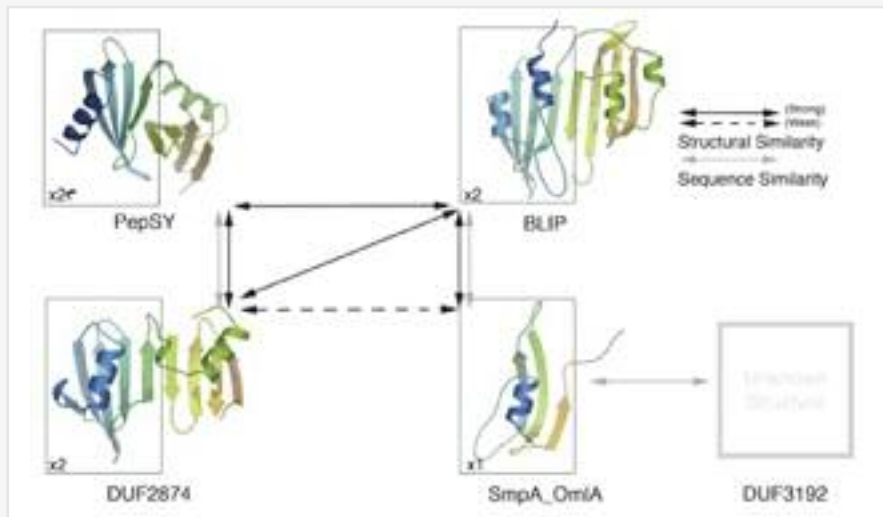
Because of its association with humans, *S. cerevisiae* has an unusual evolutionary history. Richard Durbin and colleagues therefore compared it with a closely related wild yeast species, *S. paradoxus*. *S. cerevisiae* strains were gathered from a wide range of continents and contexts, alongside *S. paradoxus* samples from tree bark.

The distribution of *S. paradoxus* conforms to the patterns expected for a wild species, with geographically discrete patterns of variation correlating with phenotypic differences between strains. *S. cerevisiae*, by contrast, shows some highly unusual patterns of genetic variation. The picture that emerges suggests that yeast was domesticated on a number of occasions, in different geographic locations, then extensively cross-bred to generate new strains, with some escaping back into the wild.

As well as providing fascinating insight into yeast's evolutionary history, the work also serves as a pilot study for the extensive genetic characterisation of different human populations in the 1000 Genomes Project.

Liti G et al. Population genomics of domestic and wild yeasts. *Nature*. 2009 Mar 19;458(7236):337-41. PMID: 19212322

We manage the Pfam (protein families), Rfam (RNA families) and MEROPS (peptidase) databases.



A novel superfamily of BLIP-like proteins identified using Pfam and structures from the JCSG consortium.

Pfam is the world-leading resource for protein families. Pfam release 23.0 contains 10 340 families (adding 1063 new families). We are moving Pfam over to the latest version of the HMMER software, which will enhance sensitivity and speed. We recently started a blog for both Rfam and Pfam to communicate changes to the community.

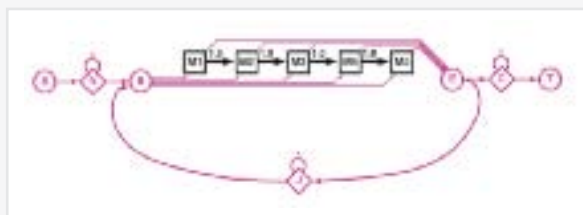
During 2008 we began using the ADDA database rather than PRODOM to generate Pfam-B. ADDA is released more frequently and helps us to increase coverage. We have also been collaborating with the Joint Centre for Structural Genomics consortium to improve annotation.

Our calculations suggest that a further 27 000 protein families are likely to be required to achieve high coverage (95%) of known proteins.

Rfam is a key resource for RNA informatics. Rfam release 9.1 contains 1372 families of non-coding RNA genes and RNA elements, a doubling in the number of families in the last year. For each RNA family, Rfam contains multiple sequence alignments that include RNA secondary structure annotation, functional annotation as well as covariance models for searching novel DNA sequence.

Rfam also became the first molecular biology database to make all its descriptive annotation available for editing via Wikipedia. This has led to increased quality annotation and allows closer collaboration with the RNA community. We also work with the journal *RNA Biology* to provide a track of papers on RNA families, where authors submit a Wikipedia article alongside a traditional paper.

The MEROPS peptidase database has added 566 new peptidases and over 20 000 substrate cleavages during 2008. Data can be displayed to show interactions between peptidases and their protein inhibitors and cross-references of pharmacological interest. New displays show significant gains, losses and expansions for organisms with completely sequenced genomes. Data for cleavages are now being collected from large-scale proteomic experiments.



The MSP profile from HMMER3 makes unaligned alignments that provides the 100-fold speed improvement compared to HMMER2.



Wellcome Library, London

➤ **Bioinformatic analysis suggests that phospholipid scramblases are part of a novel superfamily of membrane-tethered transcription factors that includes Tubby, a mouse protein that causes obesity when mutated.**

Bateman A et al. Phospholipid scramblases and Tubby-like proteins belong to a new superfamily of membrane tethered transcription factors. *Bioinformatics* 25: 159 (2009)

➤ **A guide to the latest version of the Rfam database.**

Gardner PP et al. Rfam: updates to the RNA families database. *Nucleic Acids Res* 37: D136 (2009)

➤ **An overview of the project linking Wikipedia to Rfam.**

Daub J et al. The RNA WikiProject: community annotation of RNA families. *RNA* 14: 2462 (2008)

Sean R Eddy, Janelia Farm Research Campus

We aim to identify the impact of genetic variation on phenotypic variation, principally patterns of gene expression.

Manolis Dermitzakis left the Sanger Institute in May 2009 to join the University of Geneva Medical School.



Slides in the genotyping facility.

Our main focus is genome-wide analysis of variation in gene expression and its association with nucleotide variation, particularly in disease susceptibility genes. We profile RNA in large samples and correlate genetic variation at the DNA level with phenotypic variation at the cellular level, as represented by gene expression. This approach helps us understand how genetic variation manifests its effects in the cell, and ultimately influences phenotypic variation at the tissue, organ and organismal level.

We have profiled gene expression in cell lines from most of the extended HapMap samples (850 in total) from eight populations. These populations are also genotyped for 1.6 million SNPs and copy number variation. This has allowed us for the first time to assess the effects of genetic variation on gene expression in diverse populations, providing insight into the patterns of population differentiation and recent natural selection.

We have also profiled B-cell, T-cell and fibroblast cell lines derived from umbilical cords of 85 newborns in Geneva, Switzerland, in collaboration with Stylianos Antonarakis. These individuals are also genotyped for 550K SNPs. The results provide insight into tissue specificity of the effect of genetic variation on gene regulation.

We are exploring how quantitative variation in gene expression could influence complex disease associations. We are also looking at links between sites of variable expression affected by distant genetic variation, to gain insight into the control of gene activity through genetic networks and nuclear architecture.

We are also profiling 1000 individuals from the ALSPAC (Avon Longitudinal Study of Parents and Children) cohort and multiple tissue samples from TwinsUK. Finally, using next-generation sequencing technology, we have profiled 60 Caucasians from the 1000 Genomes Project.



Cell lines in culture.



Wellcome Library, London

SNPs affecting gene expression are more likely to have been subject to positive selection than randomly chosen SNPs, suggesting that evolution is acting on gene regulatory sites.

Kudaravalli S et al. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol* 26: 649 (2009)

Analysis of genotype and gene expression data from four HapMap populations reveals that variants affecting protein sequence and those affecting regulatory sequences can act together to affect gene expression.

Dimas A et al. Modifier effects between regulatory and protein-coding variation. *PLoS Genet* e1000244 (2008)

Around three-quarters of regulatory variants have cell-type-specific effects on gene expression.

Dimas AS et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science advanced online publication* (30 July 2009)

Wellcome Library, London

We use whole-genome resequencing to study natural genetic variation in populations.



Potential populations to be sequenced in the 1000 Genomes Project. Four clusters totalling 500 samples each are planned to be sequenced, from Europe, Africa, East Asia, and the Americas. Eight sample sets come from the extended HapMap project.

A detailed catalogue of single nucleotide, insertion/deletion (indel) and structural variants will transform human genetics. To this end, following pilot projects in yeast (70 strains from two species) and human (one individual sequenced to high depth, in collaboration with Illumina), we have led the formation of an international consortium, the 1000 Genomes Project, to identify essentially all polymorphism (genetic variation down to 1% allele frequency) in humans in multiple populations.

In 2008, the 1000 Genomes Project collected sequence data for three pilot projects, with approximately 20% of the total collected at the Sanger Institute. We have led in the mapping and variant calling from these data, due for publication in 2009/10. From the pilots alone we expect to confirm approximately 7 million SNPs and short indels from dbSNP and discover at least a similar number of novel variants, providing near complete detection of SNPs with over 5% allele frequency. Collection of over five times as much data is due to be completed in 2009 or early 2010, with analysis during 2010.

To analyse resequencing data, we have developed software for read mapping and variant discovery. This includes MAQ, which has become a *de facto* standard, and more recently BWA, which is around ten times faster than MAQ. With colleagues in the 1000 Genomes Project we have also pioneered the SAM format for sequence alignments, and followed several approaches to identify structural variants, including short indels.

In other research, we released TreeFam 7.0, which includes 777 321 genes from 68 species in 16 141 families. We showed for the first time a quantitative effect of weak selection on protein-coding domains. WormBase continued to add related nematode genomes and collaborate with the Ensembl and ModEncode projects.



Wellcome Library, London

A whole-genome sequence of an African male produced by next-generating sequencing technologies.

Bentley DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53 (2008)

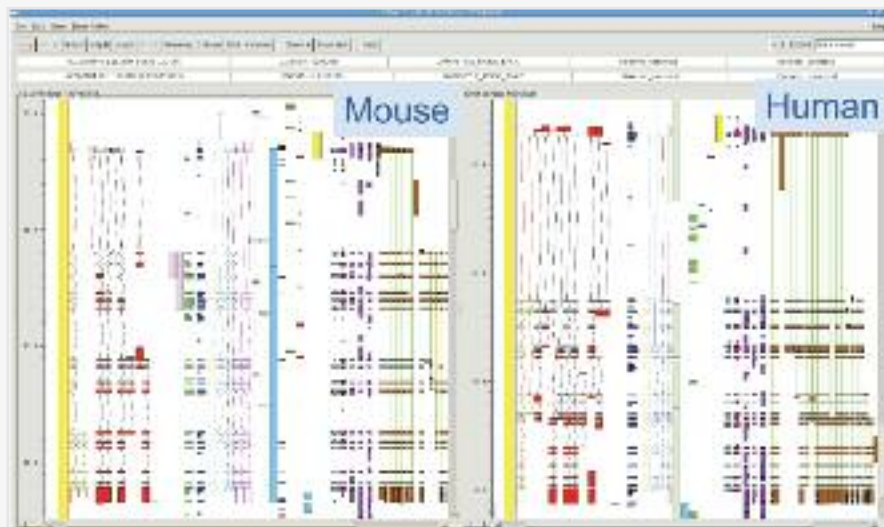
A comparison of 70 isolates of *Saccharomyces cerevisiae* and a wild relative sheds light on the likely evolution of 'domesticated' strains of baker's yeast.

Liti G et al. Population genomics of domestic and wild yeasts. *Nature* 458: 337 (2009)

A description of MAQ software, used to map short sequence reads to a reference genome.

Li H et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851 (2008)

We provide access to vertebrate genome information and reference genesets through the Ensembl genome browser.



Comparison of gene structures in human and mouse in the ZMAP genome annotation tool.

Our programme consists of Ensembl, a joint project with the European Bioinformatics Institute (EBI), and the Havana annotation group. Ensembl generates genesets for around 40 species using automatic pipelines. Havana works with Ensembl to create enhanced curated genesets for human, mouse and zebrafish. Ensembl also provides access to genome annotation and a wealth of other data.

Our genesets are the *de facto* world standard for vertebrate genome analysis. We have been involved in geneset-related analysis of nearly all the vertebrate genomes completed so far, including cow and platypus in 2008/09.

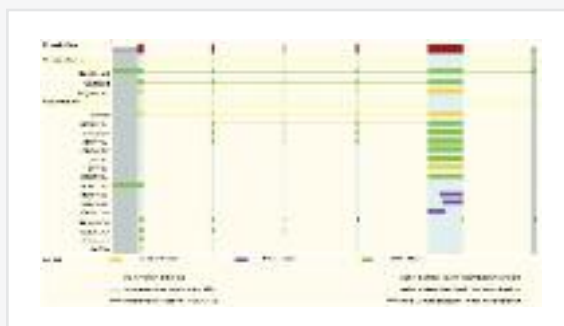
We play a leading role in the Consensus Coding Sequence and GENCODE projects, which are improving data consistency between databases and genome browsers. We are also working to standardise genome sequence alignments with the UCSC browser and gene family relationships with Treefam. We also play a major role in larger international genome annotation consortia such as the NIH-funded ENCODE and EU-funded Biosapiens and perform additional specialist annotation for the KOMP/EUCOMM mouse knockout projects.

We have contributed significantly to the development of technologies such as DAS and biomaRt, which are used globally for

dynamic integration of different genome datasets. Our core software infrastructure has also been widely adopted.

In 2008/09 we introduced a new web interface to Ensembl, based on extensive consultation with users, featuring improved navigation and speed. New projects reusing Ensembl software infrastructure include the 1000 Genomes Project website (browser.1000genomes.org) and the EBI project Ensembl Genomes, which provides access to non-vertebrate genomes (www.ensemblgenomes.org).

Transcriptome sequence datasets underpin genome annotation. The massive amounts of transcriptome data that can be generated by next-generation sequencing will allow transcription to be dissected according to cell type and its variation explored across individuals, something that was previously impractical. Under GENCODE, Havana is organising an international competition



Wellcome Library, London

➤ **The genome sequence of the cow, and comparisons with other mammalian genomes.**

Elsik CG et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324: 522 (2009)

➤ **A review of the main computational approaches used to identify protein-coding regions of the genome.**

Harrow J et al. Identifying protein-coding genes in genomic sequences. *Genome Biol* 10: 201 (2009)

➤ **An overview of the latest version of the Ensembl genome browser.**

Hubbard TJ et al. Ensembl 2009 *Nucleic Acids Res* 37: (Database issue), D690 (2009)

(RGASP) at the Sanger Institute in late 2009 to compare different approaches to processing this data, including new pipelines developed by Ensembl. This builds on the successful EGASP competition organised in 2005.

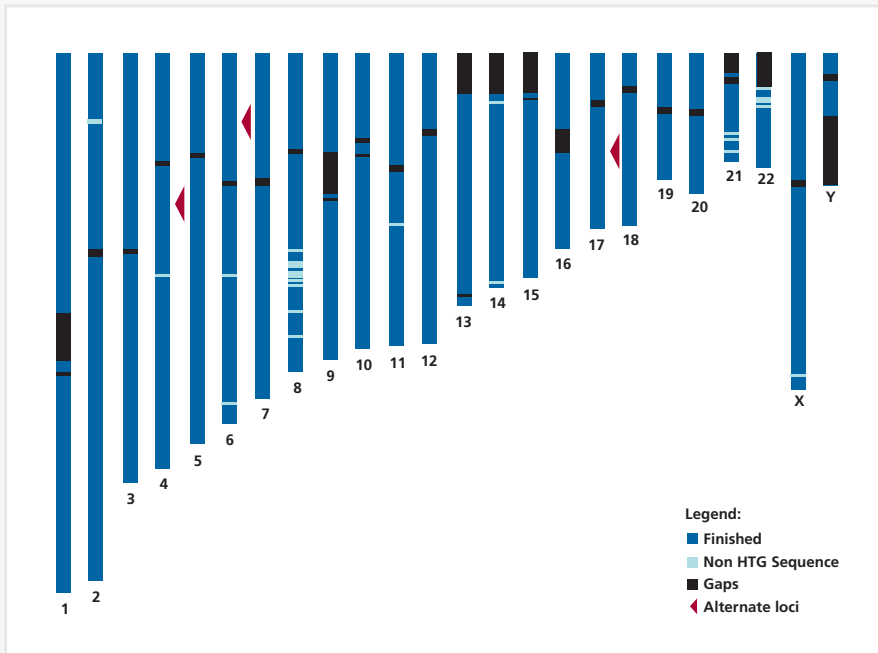
An example view from the new Ensembl website (www.ensembl.org) showing the experimental evidence that supports the annotation of one of the ~20 000 human protein coding genes.



The Genome Reference Consortium provides a long-term mechanism for updating the reference genomes of both human and mouse.



Wellcome Library, London



GRCh37: A graphical representation of the human genome assembly. The genome is coloured with respect to the genomic component used to build the genome assembly at that location. The red triangles mark regions where alternate loci have been provided.

The Genome Reference Consortium (GRC), comprising the Sanger Institute, Washington University Genome Sequencing Center, St Louis, the European Bioinformatics Institute and the National Center for Biotechnology Information, was formed in 2007 and officially launched in May 2008. The objectives of the consortium are to correct the small number of regions that are currently misrepresented, to close gaps and to produce alternative assemblies of structurally variant loci when necessary.

The first update of the human genome reference assembly (GRCh37) was released to the community in May 2009. GRCh37 is a significant update compared with the previous NCBI36 assembly. Improvements include the closure of 25 unspanned gaps; the resolution of over 150 issues reported as problems; the addition of alternate loci for three complex regions, including the MHC; and the standardisation of the entire genome assembly, including the addition of biological gap information.

GRC members have developed software systems and experimental standard operating procedures while updating the assembly and have made several internal test assemblies to assess systems and quality control processes. Plans are in place to update the mouse genome assembly, beginning with the construction of internal test assemblies. There is also continued experimental work in progress on the human sequence.

The GRC has formed a scientific advisory board, which met for the first time in May 2009 to review progress and advise on future strategy. A publication about the GRC, highlighting the GRCh37 release, is in preparation.

For more information please see www.sanger.ac.uk/sequencing/grc/ and www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/



Rows of different individuals' DNA sequences aligned at the same position in the genome.

Wellcome Library, London