

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
- Genome informatics
- Ensembl and vertebrate annotation
- Genome Reference Consortium
- Population genomics of molecular phenotypes



Wellcome Library, London

Next-generation sequencing. Work by Richard Durbin's and Ville Mustonen's teams will aid analysis of the large volumes of data being generated by the UK10K project and initiatives such as the International Cancer Genome Consortium (ICGC).

As well as being the tenth anniversary of the completion of the first draft of the human genome, one-third of which was sequenced at the Sanger Institute, 2010 was also a milestone year for the Ensembl project, founded to provide access to the human genome. Ensembl, a joint project with the European Bioinformatics Institute (EBI), provides the scientific community with an integrated vertebrate genome data resource. It is one of the three major human genome browsers, the others being at the National Center for Biotechnology Information (NCBI) and at the University of California, Santa Cruz. Since its launch in January 2000, it has grown to include 50 vertebrate genomes, integrating comparative data and information on variation. It has developed a

reputation for accurate, robust datasets, which are either automatically generated or, for human, mouse and zebrafish, also rely on manual curation by the HAVANA group.

One of Ensembl's strengths is its underlying software design, which allows data to be accessed through an Application Programming Interface (API). Over ten years this interface has remained unchanged while its functionality has been extended with each of the 50 or so database releases. As the sister project Ensembl Genomes at the EBI uses a similar approach, a wide range of genome data is available through the same API. Researchers are increasingly using the Ensembl API to carry out computational analyses of the huge datasets now available. A recent popular application is a pipeline that generates an initial interpretation for all the sequence variations present in an individual human genome sequence.

There is an increasing demand by large numbers of researchers for compute-intensive analysis. To avoid this becoming a bottleneck, Ensembl also hosts its datasets in the Amazon Web Services (AWS) 'cloud', so researchers can run their own analyses using the Ensembl API without having to download huge datasets. Ensembl is even beginning to use these services itself, with its useast.ensembl.org and asia.ensembl.org site website mirrors being hosted in the AWS cloud (complementing uswest.ensembl.org, which is hosted in a traditional datacenter). The mirrors significantly improve website performance for researchers based in the USA and Asia. The 1000 Genomes Project is similarly hosting some of its raw data in the AWS cloud.

Tim Hubbard, Head
 Alex Bateman
 Richard Durbin
 Ville Mustonen

Mix and match

Sequence comparisons have shed light on the evolution of proteins.

Most proteins contain two or more 'domains' – semi-independent modules with distinct structures and functions. Mixing and matching of different domains is one way in which novel protein functions might emerge. While some examples of such mixing have been identified, exactly how much it contributes to protein evolution is unclear. Using a bioinformatic approach, Institute researchers have provided a much broader picture of domain shuffling and the mechanisms that underlie it.

The analysis was made possible by the availability of genome sequences for a range of animal species. A comparison of related genes can reveal the likely genetic changes – including those associated with domain shuffling events – that occurred after species diverged along different evolutionary paths.

Well-established examples of domain shuffling events were examined first to identify DNA sequence signatures associated with particular mechanisms of rearrangement. There are several ways in which domains can be mixed, such as fusion of exons in neighbouring genes, retroposition (whereby an RNA is converted to DNA and inserted into the genome), and recombination (when related DNA sequences align and exchange regions of DNA).

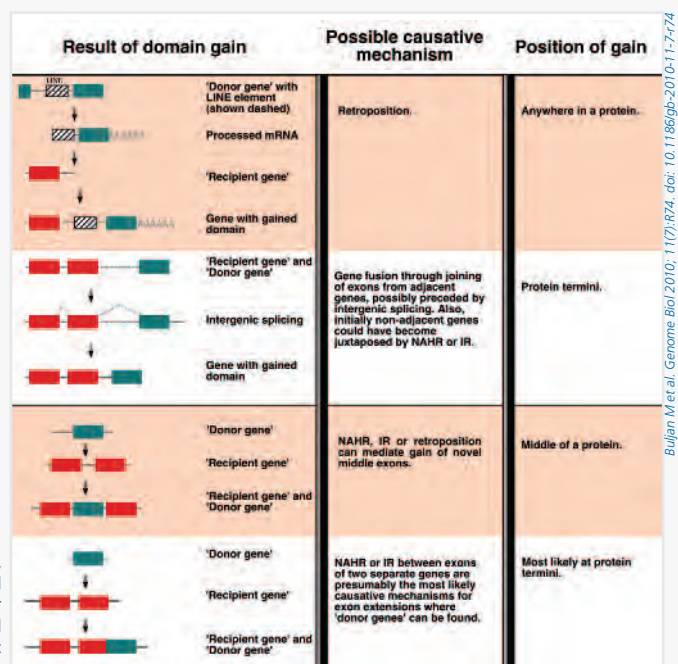
As we move into an era of sequence data from thousands of individuals, and transcriptomic and epigenetic data from multiple individuals, cell types and developmental stages, the effects of sequence variants on phenotype can be modelled more accurately. New approaches are being developed by Institute researchers in collaboration with external groups, while another Institute team is building models of somatic cancer mutations, calibrating an

The analysis suggests that the first of these processes, gene fusion, is by far the most common source of exon rearrangement, other processes accounting for perhaps 10–15 per cent of events. However, in some 80 per cent of cases, fusion was preceded by recombination, bringing novel exon combinations into closer proximity.

There is a notable evolutionary trend towards more complex domain architectures, suggesting that organismal complexity may derive at least in part from more sophisticated use of a protein domain toolkit. A combination of gene duplication by recombination followed by domain gain by fusion may be one mechanism by which this increase in complexity has come about.

Buljan M et al. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol* 2010; 11(7):R74.

Mechanisms for domain gains, including retroposition and gene fusion. IR stands for illegitimate recombination and NAHR for non-allelic homologous recombination.



evolutionary model using the Pfam database. Such models should help facilitate data analysis in the UK10K project and initiatives such as the International Cancer Genome Consortium (ICGC).

The transition from a single reference genome sequence to population studies also poses challenges to annotation. In some individuals a gene may appear to be functional, whereas sequence differences in other individuals imply that it is a non-functional pseudogene. The pilot phase of the 1000 Genomes Project found that, on average, each individual studied carried approximately 250 to 300 loss-of-function variants in annotated genes. This work has relied on the in-house development of new algorithms to identify variants and detect insertions and deletions.

In future, both reference genome sequences and gene annotation will have to incorporate these variations and link them to phenotype information. This will aid the interpretation of human genetics studies and personal genomes, allowing each sequence to be matched to the most appropriate reference sequence and annotation.

Similar issues apply across other areas of genome-based research. Sequences of multiple strains and entire populations are being collected for viruses, bacteria, pathogens and vertebrate model organisms. Large quantities of sequence variation and phenotype data will need to be organised and annotated.

Work began during the year on more complex data structures and annotation sets for the human genome. The Genome Reference Consortium (GRC), comprising the Sanger Institute, Washington University Genome Sequencing Center, St Louis, the EBI and the NCBI, is responsible for the reference genome sequences for human, mouse and zebrafish. It is beginning to represent regions of significant allelic diversity identified by the 1000 Genomes Project and other initiatives.

To represent alternative alleles, the GRC periodically releases 'patches' to the existing reference genome, which causes less disruption than the release of completely new assemblies. So far 70 patches have been released and can be displayed on Ensembl. Ensembl also began integrating disease and phenotypic annotations on sequence polymorphisms, adding nearly 60 000 phenotype annotations from genome-wide association studies. Germline mutations from external resources, including clinical locus-specific databases and the Sanger Institute's Catalogue of Somatic Mutations in Cancer (COSMIC), have also been incorporated.

With genome and transcript sequencing now commonplace, a major challenge for bioinformatics has been to draw on both the information and expertise of the entire community. Automatic annotation algorithms and tools to mine the published literature can extract some information, but need to be supplemented with expert curation. However, this is expensive and hard to scale.



Genome Research Limited

Large-scale population studies will require annotation to incorporate the uncovered variations and pseudogenes and link them to phenotype information. This will aid the interpretation of genomes, allowing each sequence to be matched to the most appropriate reference sequence and annotation.

Over the last 10 years Wikipedia has emerged as a hugely successful platform for cooperative organisation of textual information. Based on Wikipedia software, several sites have been set up as community annotation portals – including SNPedia (www.snpedia.com/index.php/SNPedia), which annotates human disease variants and is accessible from Ensembl.

An alternative is to use Wikipedia itself, which is something the Institute has pioneered for the Rfam database. Rfam continues to maintain the alignment of sequence data that makes up each RNA model, but textual descriptions are stored in Wikipedia. This hybrid has been so successful that a similar approach is planned for the much larger Pfam database. The group has also organised training across the Genome Campus to encourage contributions to wiki databases.

Bioinformatics continues to be dominated by the need to keep up with rapidly growing quantities of genome and transcriptomics sequence data. Large sequencing projects generate as much data as high-energy physics experiments such as the Large Hadron Collider at CERN. This is a challenge for algorithm development, database design and IT. It also underlines the importance of investment in data repositories, as the value of data is maximised through broad access and use. The Sanger Institute therefore supports the EBI-centred ELIXIR project to develop a long-term European infrastructure for biological information. ELIXIR is one of the priority European infrastructures of the ESFRI (European Life-science Infrastructure for Biological Information) process.



Building a genome

A new computational approach could greatly decrease the time needed to generate complete genome assemblies.

The primary outputs of genome sequencing are the individual 'reads' generated by sequencing machines. These range in length from up to 1000 nucleotides for conventional capillary sequencing down to less than 100 nucleotides for next-generation technologies. It is a considerable bioinformatic challenge to piece together millions of reads to assemble a contiguous sequence. A new technique has been developed that not only accelerates genome assembly from next-generation reads but is also flexible enough to handle the longer outputs produced by third-generation technologies.

Originally, genome assembly was based on identification of overlapping fragments, a process automated by specially designed algorithms. Unfortunately, this process is greatly complicated by DNA repeats – a major component of most genomes, particularly eukaryotic ones – which introduce ambiguities into assemblies. To get round the repeat problem, most software tools now chop up raw read sequence data into a series of identically sized fragments. In effect, these methods collate all copies of a short (30-50 nucleotides) repeat and treat them as a single entity.

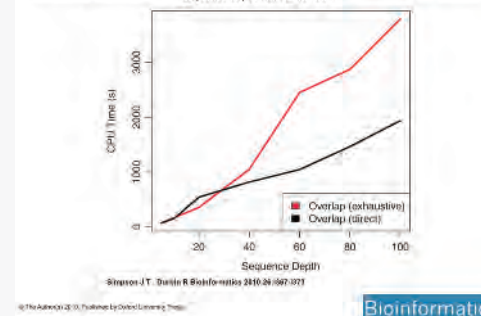
A variant of this method, known as the 'string graph', resurrected the overlap method while also collapsing repeats. Unfortunately, the computational effort of overlap-based methods is proportional to the square of the total amount of sequence being processed, leading to an unacceptable increase in computing load with additional sequence information. So despite having certain advantages, this method is rarely used at present.

To tackle this issue, the string graph approach is combined with a method developed to align sequence reads to a reference genome. Crucially, computing effort of the resulting algorithm is proportional to total sequence information rather than its square, enabling it to handle very large amounts of data.

The new method therefore enables genomes to be pieced together much more rapidly and with lower computing burden. Significantly, it is also flexible enough to deal with both short reads and the much longer reads of third-generation sequencers. Although not the finished article, the method is a significant step towards a software tool that could make automatic assembly of complete complex genomes, even human ones, a routine procedure.

Simpson JT, Durbin R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 2010; 26(12):i367–73.

The running time of the direct and exhaustive overlap algorithms for simulated E. coli data with sequence depth from 5x to 100x.



Simpson JT, Durbin R. *Bioinformatics* 2010; 26(12):i367–73
doi: 10.1093/bioinformatics/btq177

Using this new computational approach very large amounts of data can be handled. This is because the computing effort of the resulting algorithm is proportional to total sequence information rather than its square.

We classify protein and RNA sequences into families, to shed light on their function in health and disease.



Finn RD et al. Nucleic Acids Res. 2010; 38(Database issue):D211–22. doi: 10.1093/nar/gkp965



Wellcome Library, London

New alignment confidence display. The colour of the residues reflects the alignment uncertainty, and is based on the posterior probability that is calculated by HMMER3. A green residue indicates a strong likelihood that the alignment is correct. Where certainty decreases, the colour becomes closer to red.

Next-generation sequencing is creating a vast deluge of protein and RNA sequence data. Scientists are only ever able to characterise experimentally a small fraction of these sequences. By identifying similarities between the sequences of proteins and RNAs, we can predict the likely functions of these molecules. This greatly enhances the value of the sequence data generated at the Sanger Institute and around the world. We use sophisticated models to identify similarities, and group proteins and RNAs into families that we characterise and display in our databases: Pfam, Rfam and MEROPS.

During the past year we released Pfam version 24.0, which contains 11 912 families and is the first to use the new HMMER3 search software. Using HMMER3 has allowed us to increase the coverage of Pfam to 77.7 per cent of all known proteins. Due to its speed, it is now much easier for users to run Pfam analyses on their own data.

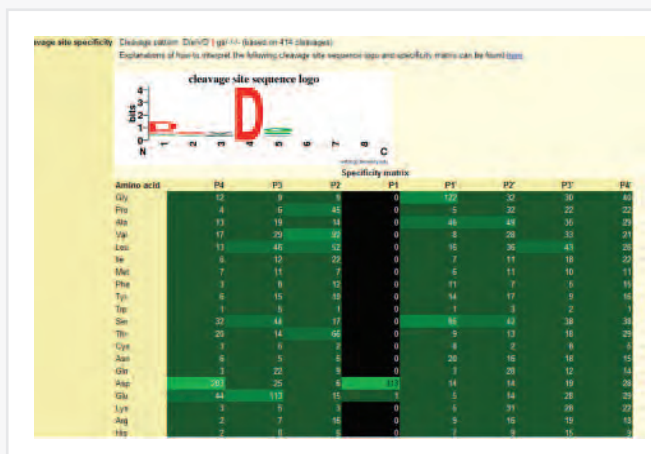
Version 10.0 of the Rfam database includes 1446 RNA families. The release incorporated 'clans', which cluster related RNA families together. This new level of classification has already stimulated the RNA community to

MEROPS has collected substrate cleavage data for more than 40 per cent of known peptidases. The cleavage pattern at the top is a textual representation of the logo, where the scissile bond is shown as a red cross, and the binding pockets separated by forward slashes. Below, the letters are coloured to indicate amino acid properties: blue (basic), red (acidic), black (hydrophobic) and any other (green). In the frequency matrix the more often an amino acid occurs a position, the brighter the shade of the cell (black for unknown).

create new software tools to detect these relationships. Rfam was the first major biological database to deposit all its annotation into Wikipedia. Over 1000 users have made more than 9000 edits to the Wikipedia articles that are displayed in Rfam.

MEROPS is a database of peptidases, their inhibitors and substrates. MEROPS has collected substrate cleavage data for more than 40 per cent of known peptidases. The size of the cleavage site collection has doubled to over 40 000 known sites. In the latest release of MEROPS, a completely new catalytic class of peptidases has been included – the asparagine peptidases class, which includes a variety of viral and microbial peptidases.

During the coming year we aim to increase coverage of known sequences and families as well as continue to engage the public and the scientific community through Wikipedia.



Rawlings ND. Database (Oxford) 2009; 2009:bap015 doi: 10.1093/database/bap015



Identifying the major mechanisms for gain of new protein domains during evolution.

Buljan M, Frankish A, Bateman A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol* 2010; 11(7):R74.



Latest Pfam developments, including the move to HMMER3 software for 100-fold speed improvements.

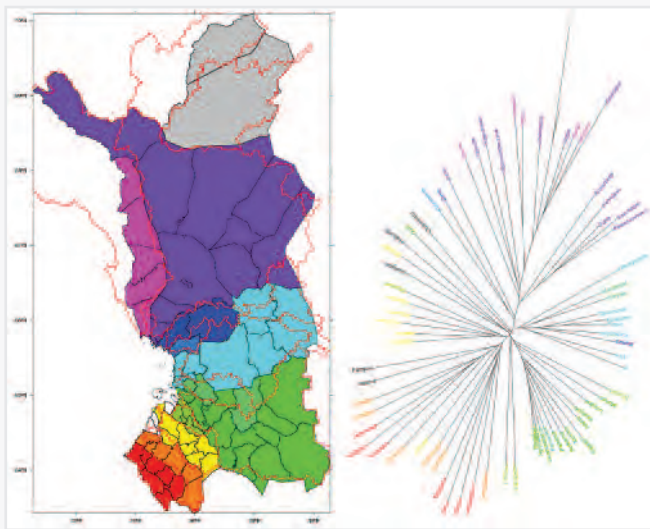
Finn RD et al. The Pfam protein families database. *Nucleic Acids Res.* 2010; 38(Database issue):D211–22.



Development of the peptidase cleavage collection in the MEROPS database.

Rawlings ND. A large and accurate collection of peptidase cleavages in the MEROPS database. *Database (Oxford)* 2009; 2009:bap015.

The Genome Informatics research group develops methods to use high-throughput DNA sequence data to discover genetic variation and understand its functional consequences.



Genome Research Limited

Population structure in northern Finland reflects geographic structure.

One major focus is using new technology sequence data to discover genetic variation. We have played a leading role in study design, methods development and primary data analysis for the 1000 Genomes Project (see page 21), and are taking on a similar role for a major new British medical sequencing project, UK10K. Alongside this, we are developing machine-learning methods to study the genetics of variation in high-dimensional data sets such as expression data.

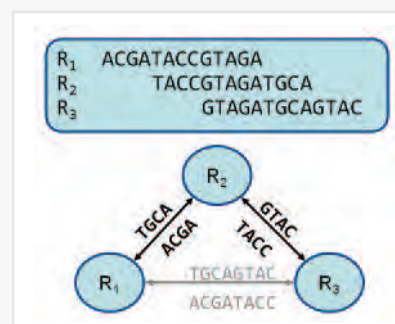
For the 1000 Genomes Project, we developed the QCALL and DINDEL methods, to identify variants (variant calling) and to detect insertions and deletions (indels), respectively. In addition, we extended our Burrows-Wheeler Alignment (BWA) read-alignment software to handle the long sequencing reads generated by the newest sequencing machines. We are also developing new sequence-assembly approaches based on the computer science methods underpinning BWA. As well as being relevant for de novo sequencing of new species, these methods will also facilitate reference-free analysis of genetic variation in humans.

Alongside the 1000 Genomes and UK10K Projects, we are also working on a strategy for efficient comprehensive analysis of rare genetic variation in isolated populations, and have developed a new approach to

haplotype analysis in such populations. We are currently applying these approaches to the Kuusamo isolate in north-eastern Finland, and will sequence a representative sample of its population in 2011, as a pilot for further work on isolated populations.

Finally, beyond genetic variation, it is now possible to obtain complex data sets from a population sample of cell lines, including data on gene expression, chromatin structure, metabolite concentrations and other cellular characteristics. These rich data sets allow more sophisticated modelling that can add considerable statistical power beyond independent association tests for each trait. We have been developing approaches to take advantage of this additional power in collaboration with Microsoft Research, Cambridge, TwinsUK at King's College London, and Manolis Dermitzakis in Geneva. As these cellular traits are intermediaries between genome variation and end phenotypes, we hope these methods will provide additional insight into the genetic components of disease.

An example of a string graph constructed from the overlaps between three reads. We aim to use such assembly methods to provide an unbiased view of human genetic variation, independent of the reference sequence.



Wellcome Library, London

A new fast and low-memory approach to DNA sequence assembly from short sequencing reads.

Simpson JT, Durbin R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 2010; 26:i367-73

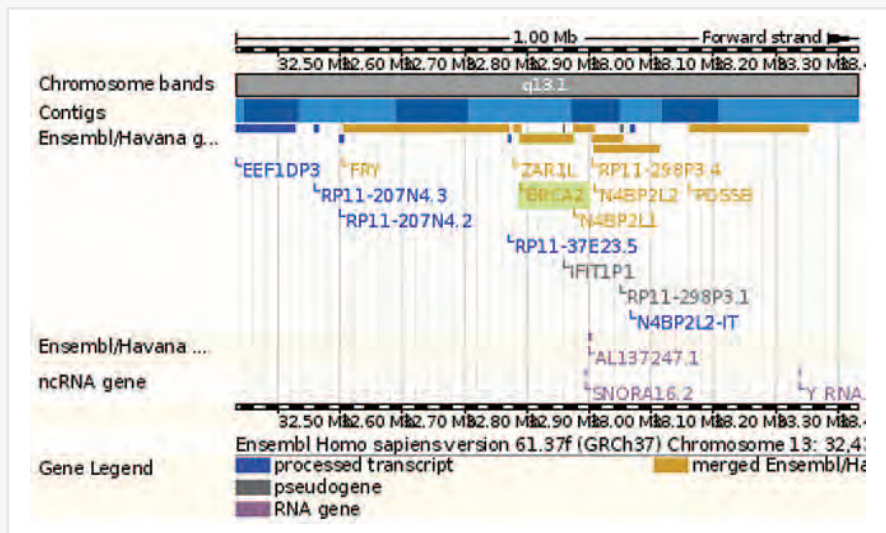
Use of factor analysis to remove a large fraction of the non-genetic variance in high-dimensional data, increasing power for genetic tests.

Stegle O et al. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 2010; 6:e1000770

A systematic RNAi screen for cell-division defects in human cells, requiring significant informatics input.

Neumann B et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* 2010; 464:721-7

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



Ensembl view of human gene *BRCA2*, incorporating GENCODE version 4 geneset. The GENCODE dataset has been adopted by major international consortia, including the 1000 Genomes Project.

Our programme consists of Ensembl, a joint project with the European Bioinformatics Institute (EBI), and the HAVANA annotation group. Ensembl provides integrated genomic information for 50 vertebrate genomes including genesets and regulatory annotation generated by the project. The human, mouse and zebrafish genesets are created jointly by HAVANA and Ensembl.

We have a strong reputation for genesets and have been involved in geneset-related analysis of nearly all the vertebrate genomes completed so far. In 2009/10 this has included horse, songbird and turkey. We are also working with more than 30 groups involved in a community pig genome annotation project. However the largest effort remains focused around improving human, mouse and zebrafish genesets.

For human we lead the GENCODE consortium of the ENCODE project, funded by the US National Institutes of Health. GENCODE aims to develop a comprehensive annotation of all coding regions within the human genome, by manual curation, computational analysis and targeted experimental approaches. This year has seen annotation of unitary, ribosomal and duplicated pseudogenes, and initial analyses of non-coding RNAs.

GENCODE version 4, launched as part of Ensembl release 58 in May 2010, is now the default human geneset and has been adopted by several international consortia, including the 1000 Genomes Project and the International Cancer Genome Consortium.

GENCODE incorporates the Consensus Coding Sequence (CCDS) project, an international collaboration that focuses on the protein-coding subset of the human and mouse geneset. HAVANA is also involved in the KOMP/EUCOMM knockout mice projects, and contributed to an analysis of mass spectrometry data that identified 40 novel mouse genes.

The availability of large amounts of transcriptome data collected using next-generation sequencing (RNAseq) is allowing us to extend our genesets to capture expression differences between cell types and developmental stages and the detection of non coding RNA genes (ncRNA). New Ensembl pipelines have been developed and applied to zebrafish transcriptome data as a pilot. In 2009 we setup the RNAseq genome annotation assessment project (RGASP) to evaluate progress of automatic gene building using this source of data. This will continue in 2011 with a third workshop as part of the Sequence Mapping and Assembly Assessment Project (SMAAP).



Wellcome Library, London

► **Ten years of the Ensembl genome browser.**

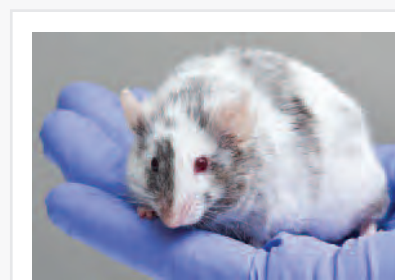
Flicek P et al. Ensembl's 10th Year. *Nucleic Acids Res* 2010; 38:D557-62.

► **High-throughput sequencing of 14 000 exons identifies mutations induced on mouse chromosome 11, confirming large-scale sequencing's viability for mutation detection.**

Boles MK et al. Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin. *PLoS Genet* 2009; 5(12):e1000759.

► **An analysis of gene loss by Havana and other members of the GENCODE project**

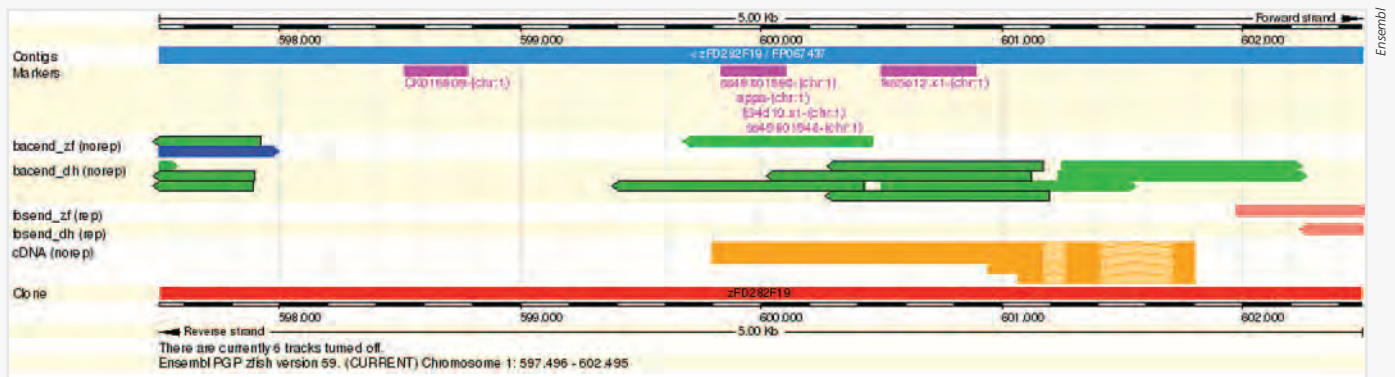
Zhang ZD et al. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* 2010; 11(3):R26.



Dave Sayer, Wellcome Trust

The Ensembl genome browser provides access to data from more than 50 organisms, from human to turkey. However the largest effort remains focused around improving human, mouse and zebrafish genesets.

The Genome Reference Consortium continuously improves reference genome assemblies of human, mouse and zebrafish to provide robust substrates for genome analysis.



Zebrafish genome in PGP viewer in Ensembl. A full PGP viewer can now be generated for a subset of a genome, allowing rapid trials and assessments of alternative paths.

The Genome Reference Consortium (GRC) comprises the Sanger Institute, Washington University Genome Sequencing Center, St Louis, the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). Its aims are to refine and complete reference genome sequences and, when necessary, to produce alternative assemblies of structurally variant loci.

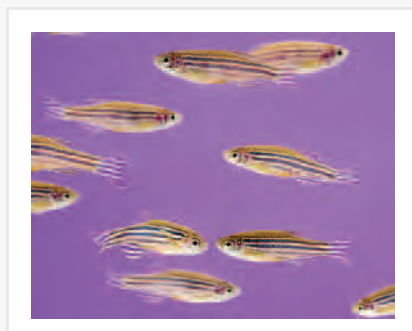
The original model for representing the genome assemblies was to use a single, preferred tiling path as a consensus representation of the genome. For most mammalian genomes, however, regions with complex allelic diversity cannot be represented by a single tiling path. The GRC is working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

One specific tool developed at Sanger to aid in this process is PGP viewer, an Ensembl-based tool for viewing genome assembly data. This has recently incorporated new datasets, including additional libraries of clone ends and, most notably, optical map data for mouse and human. A full PGP viewer can now be generated for a subset of a genome, allowing rapid trials and assessments of alternative paths.

For the human reference sequence, the GRC has introduced a new system to provide

updates outside the full assembly release. Users interested in a specific locus can now obtain an improved representation without affecting users who need chromosome coordinate stability. These 'patches' have been adopted by the major browsers (Ensembl, NCBI MapViewer) and are displayed as an option alongside the official reference assembly. So far, two updates have been released, containing 70 patches.

After the Sanger Institute released the zebrafish genome assembly Zv9 in July 2010, the GRC took over management of this reference. Work has started to replace whole-genome shotgun sequence with high-quality finished clone sequence, to close remaining gaps and to position as yet unlocalised sequence. The GRC is also refining an updated mouse reference assembly. So far whole-genome shotgun sequence has been substituted with clone sequence, closing 155 gaps.

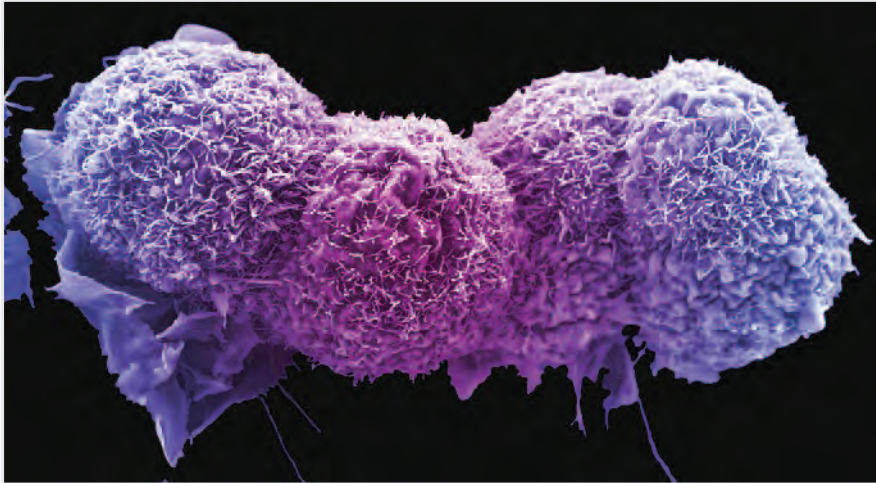


Dave Sayer, Wellcome Trust

The Genome Reference Consortium now has responsibility for the two major vertebrate model organisms, mouse and zebrafish.



We combine evolutionary theory and population genetic analysis to elucidate functional consequences of genomic variation.



Anne Weston, JRI, CRUK, Wellcome Images

Lung cancer cell dividing. Cancer development can be likened to Darwinian evolution inside the body, where the 'fittest' cells survive, multiply and eventually outcompete normal cells.

By analysing their evolutionary dynamics, we aim to tease out the impact that genomic alterations have on biological function. Evolution is a complex stochastic process, the outcome of which is decided at the level of populations of individual organisms or cells. How this process plays out depends on the interplay between three key evolutionary forces:

- mutation/recombination, which generates variation;
- genetic drift, the 'noise' in reproduction;
- selection, the differential reproductive success of individuals.

The individual roles of these factors are difficult to disentangle. However, this untangling will be critical if we are to understand the functional consequences of mutations. An important example is the need to distinguish cancer-causing 'driver' mutations from a background of 'passenger' mutations. As well as experimental studies, bioinformatics approaches can also address this question.

One systematic way to study cancer mutations is to use evolutionary theory to inform scoring systems for genomic alterations. During 2010 we have analysed somatic mutations in cancer using a germline fitness-based scoring system. The scoring

system is calibrated using Pfam domain alignments. Applying this approach to common germline variation suggests that calculated scores for genetic variants are indeed proportional to their germline fitness effects; as is evident by the scorings' ability to predict the main pattern of the observed germline variation.

There is no *a priori* reason why such scores should also reflect the somatic fitness of cancer cells. However, there is an interesting relation for mutations that fall onto predicted tumour suppressor genes, which show anomalously severe alterations.

We have applied the scoring system to somatic cancer mutations by integrating the observed variation at the level of loci and genes. The strongest signal, as assessed by the deviation between the null model and the observed mutations, is recorded at the gene level, suggesting that many distinct mutations ultimately give rise to a similar phenotypic effect.

As a next step, we plan to integrate our scoring system into a predictive model, which could be used to judge the likelihood that somatically mutated genes were oncogenes or tumour suppressor genes.



Genome Research Limited

Identification of key observables for adaptive dynamics and predicting their behaviour under various evolutionary scenarios.

Mustonen V, Lässig M. Fitness flux and ubiquity of adaptive evolution. *Proc Natl Acad Sci USA* 2010;107:4248–53