

Vertebrate genomics



Paul Flicek

DSc Washington University, 2004. Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008. At EMBL-EBI since 2005, Team Leader since 2007, Senior Scientist since 2011.

DESCRIPTION OF SERVICES

Our team creates and maintains the genomic resources of the Ensembl project, a joint project with the Wellcome Trust Sanger Institute, and is responsible for data management for a number of large-scale international projects, including the 1000 Genomes Project and, in collaboration with the Functional Genomics team, the International Mouse Phenotyping Consortium. We also maintain and develop EMBL-EBI's major variation databases, including the European Genome-phenome Archive (EGA) and the DGVa database of copy number and structural variation. All of these resources are made publicly available and are widely used by the scientific community and by the team itself as part of our research into evolution, epigenetics and transcriptional regulation.

Our specific research projects, largely done in collaboration with Duncan Odom's group at the University of Cambridge, focus on the evolution of transcriptional regulation. Recently we have expanded 'comparative regulatory genomics' techniques including mapping the same DNA-protein interactions in matched tissues in multiple species to understand how gene regulation has evolved while the tissue-level functions are largely conserved. We are also interested in the role of chromatin conformation in tissue-specific gene regulation and have investigated both the CTCF and cohesin complex in this context.

SUMMARY OF PROGRESS

- Issued five releases of Ensembl encompassing extensive expansions of human variation and regulatory data in addition to new species, new genome assemblies and other new features;
- Launched the International Mouse Phenotyping Consortium (IMPC);
- Published 27 scientific papers in peer-reviewed journal articles, representing all of the group's major activities.

MAJOR ACHIEVEMENTS

The five major releases of Ensembl in 2011 featured extensive updates to the core genomic resources provided for human, mouse, rat and zebrafish. They also included five new supported species, including turkey and the endangered gibbon and Tasmanian devil genomes. We continued to support updates to the human reference assembly, including the recent 'patch version' releases. Our variation data resources such as the Ensembl Variant Effect Predictor (VEP) support a large, integrated collection of sequence variation associated with disease as well as the reference data included in the most recent version of dbSNP, which comprises the comprehensive, world-wide variation data created by the 1000 Genomes Project. Our participation in the ENCODE project included both extensive analysis of whole-genome functional data as well as the incorporation of the most important ENCODE datasets and results into Ensembl. The Ensembl outreach team has conducted nearly 100 hands-on training courses across Europe and around the world.

The European Genome-phenome Archive (EGA) nearly doubled the number of available studies in 2011 to approximately 150 and introduced a new web interface that makes it much easier for users to find and submit data. The DGVa database of structural and copy number variation now contains nearly every available CNV/SV dataset. Streamlined data-submission and -exchange procedures were

introduced with its peer database, dbVar, at the NCBI and the project's main collaborator at the Database of Genomic Variants in Toronto.

In collaboration with Helen Parkinson in the Functional Genomics team (see page 38), the mouse informatics team will lead a five-year effort to provide the data management and coordination for the NIH-funded Knockout Mouse Phenotyping Program (KOMP2). The KOMP2 effort, our on-going work with the European Mouse Mutant Archive (EMMA) and the Infrafrontier European infrastructure are all key components of the International Mouse Phenotyping Consortium, which was formally launched in September 2011.

Data from the 1000 Genomes Project was accessed by a steadily growing number of users, with a considerable increase following the October 2010 initial project publication. We released the full phase 1 data set and developed several new tools to enable community access to the data. We also started the process of collecting data from phase 2 of the project.

Our research into the comparative regulatory genomics of the DNA-binding protein CTCF, led at EMBL-EBI by PhD student Petra Schwalie, demonstrated that over evolutionary time CTCF's binding profile in mammalian genomes has been dramatically affected by waves of retrotransposon insertions. Specific retrotransposon repeats containing the

CTCF binding site have spread through the mammalian genome multiple times, including within the past 20-30 million years in the mouse and rat genomes, leaving behind a conserved hierarchical signature. In addition, the same data demonstrated that CTCF has an evolutionary conserved 34 bp binding site, which is approximately twice as long as previously known.

FUTURE PLANS

The rapidly growing volume and diversity of data across the scope of genomics research is the major challenge for projects like Ensembl. We see an ever-increasing number of whole-genome sequences as well as comprehensive variation, regulatory, disease and phenotype data in human and other species. We have created a number of analysis and visualisation methods to summarise and present dense and complex regulation data (see Figure) and will continue to expand these to other species and disease states. At the same time, the EU Blueprint project and the NIH-funded KOMP2 project are both expected to produce their first major data sets in 2012. This means that we will continue to play an end-to-end role in major genomics projects from raw-data management for the project to summary-data presentation to the wider scientific community.

Selected publications

Flicek, P., Amode, M.R., et al. (2011) Ensembl 2011. *Nucleic Acids Res.* 39 (Database issue), D800-6.

Lindblad-Toh, K., Garber, M., et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

Locke, D.P., Hillier, L.W., et al. (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469 (7331), 529-33.

Marth, G.T., Yu, F., et al. (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.* 12 (9), R84.

Renfree, M.B., Papenfuss, A.T., et al. (2011) Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.* 12 (8), R81.

Our research projects are expanding in number of species, tissues and specific DNA-protein interactions. We will also focus on understanding the differentiation process and components of cell- and tissue-specific regulation. These questions will be addressed both in the context of our established collaborative projects with the Odom group and as part of other collaborations, including larger EU-funded projects.

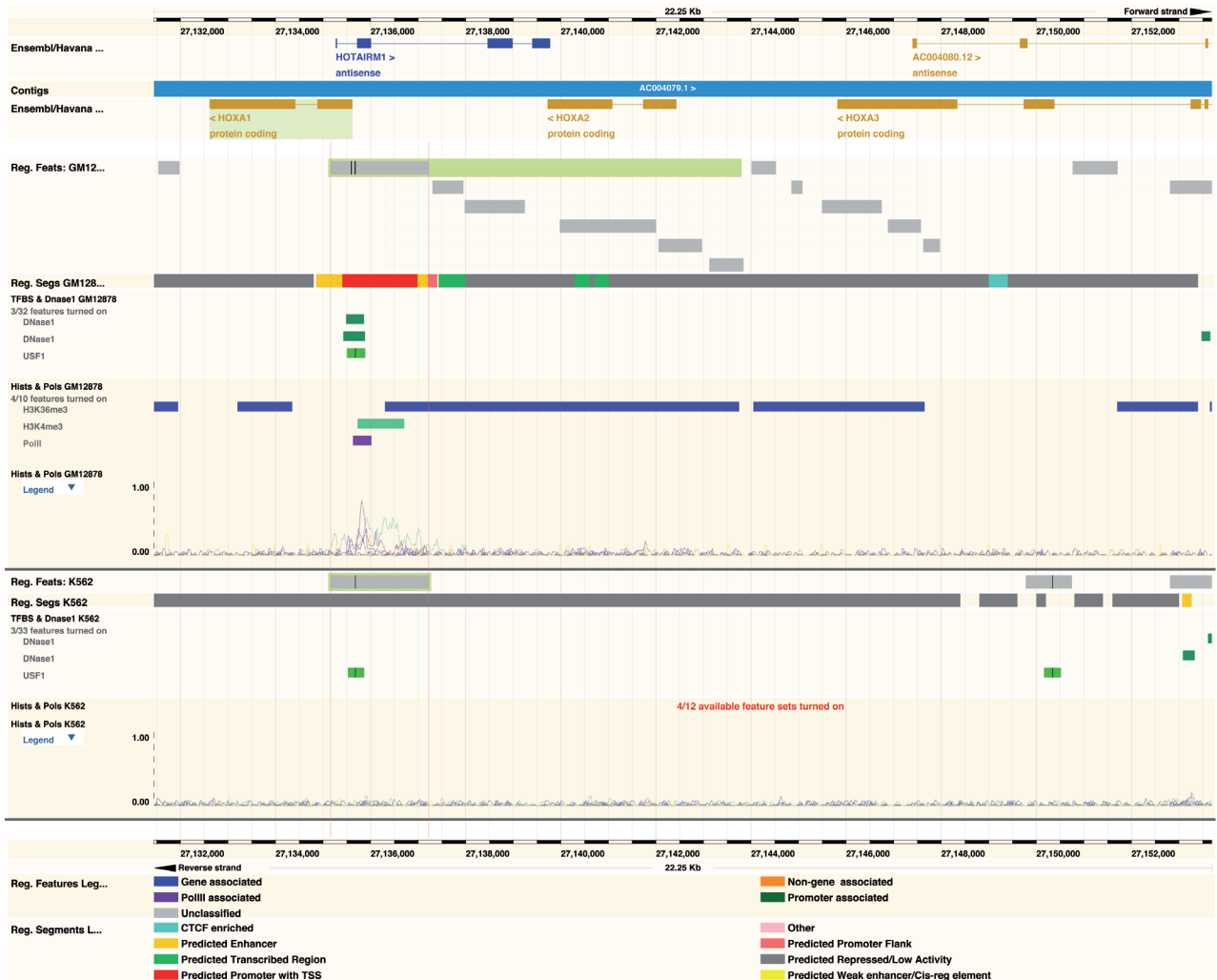


Figure. Raw, processed and summarised data from two human cell lines in region around the HOXA1 gene on chromosome 7 as presented in Ensembl. Lymphoblastoid (GM12878) and myelogenous leukemia (K562) cell lines from the ENCODE project show differences in raw data (wiggly tracks), processed data (coloured rectangular boxes in the tracks labelled DNase1) and summary level multi-coloured segmentation tracks demonstrating the difference at the chromatin and regulatory level between the active gene in GM12878 cells and the inactive gene in K562 cells.