FAST BREAKING PAPERS - 2008

December 2008

PDF

Paul Flicek talks with *ScienceWatch.com* and answers a few questions about this month's Fast Breaking Paper in the field of Biology & Biochemistry.

Article Title: Ensembl 2008
Authors: Flicek, P, et al.
Journal: NUCL ACID RES
Volume: 36
Issue:
Page: :D707-D714
Year: Sp. Iss. SI JAN 2008
* European Bioinformat Inst, Wellcome Trust Genome Campus, Cambridge CB10 1SD, England.
* European Bioinformat Inst, Cambridge CB10 1SD, England.
* Wellcome Trust Sanger Inst, Cambridge CB10 1SD, England.
(addresses have been truncated)

SW: Why do you think your paper is highly cited?

Our paper describes one of the fundamental bioinformatics resources on which many other researchers build analysis and bioinformatics tools. One resource within the project that many people find integral to their work is the collection of Ensembl gene annotations. These genes sets are created across all of the genomes we support (approximately 45 chordates) and ensure a high standard for all downstream analysis.

Ensembl also provides an open infrastructure for genome information. We encourage all researchers to write custom applications that reuse Ensembl code and directly access the Ensembl data. By providing an open-source code base, we are integral to the bioinformatics analysis pipelines of many groups around the world and a number of custom applications based on our code and our results are listed on our web site.

SW: Does it describe a new discovery, methodology, or synthesis of knowledge?

We are integrating multiple and diverse data sources into consistent and sensible results. As more biological data is created, integration analysis leading to comprehensive genome annotation is required to make sense of the various sources and create knowledge that can be used in other ways.

The Ensembl human gene set is an example which now sensibly integrates information from protein sequences, cDNAs, manual annotation, CAGE, and ditag sequences and sequence-based expression data. Going forward, we have created a first, "best guess" computational annotation of genome regulatory regions by integrating data from multiple sources.

SW: Would you summarize the significance of your paper in layman's terms?

The role of computers, databases, and bioinformatics in the field of biology has risen dramatically over the last two decades. We are one of the reference databases for genomics and we create and maintain annotation on the collection of sequenced mammalian and chordate genomes. Our results are used by many other researchers in many ways to support their work in many fields of modern biology, including human and model organism genetics, comparative genomics, and bioinformatics.

SW: How did you become involved in this research, and were there any problems along the way?

The Ensembl project was started approximately nine years ago and has grown steadily since its beginnings. I joined the project in 2005 and I am now joint head of the effort. Our project is spread over two institutes, several dozen scientists, and hundreds of thousands of lines of computer code.

We are continually developing our methods and expanding the scope of our project to address the challenges of more and larger data sets supporting genome annotations. Our largest challenges and greatest problems are currently associated with using next generation sequencing data sets. To put things in perspective, the Ensembl project started before the human genome was finished and at a time when the total amount of DNA sequence in GenBank was less than the amount now produced daily by major genome sequencing centers.

SW: Where do you see your research leading in the future?

In the future, a major goal for a genome annotation resource such as Ensembl will be to understand and annotate human variation and connect the variation with phenotype. We are associated with the 1000 Genomes project and the data produced in this project will be a major step toward a more complete catalog of human variation. This type of variation data and some of the other resources that we and others are creating will likely be used in many future studies with clinical importance.

Do you foresee any social or political implications for your research?

Our project already includes the genome sequences of James Watson and Craig Venter (see also). In the very near future, we will include additional individual genome sequences from the 1000 Genomes project and other sources. These first cases are just the beginning of the era of personal and individual genome sequences, which has already captured the public imagination as evidence by the response to companies such as "23andme" and "decodeme."

As more people are able to access their own genetic information, some (possibly many) of them, will come to Ensembl to investigate the data resources and results that we have produced as they seek to understand their own genome. The social and political impacts of personal genomics will be incredibly large. How this impact will affect or guide our research remains to be seen, but it will surely play some role.

View screen shots from the new Ensembl website.

PDF GreeTree: Gene PRF1: Perforin-1 Precursor (P1) (Lymphocyte pore-forming protein) (PFP) (Cytolysin).

PDF Homepage: Ensembl homepage.

PDF Location: Location-based displays: Chromosome 17: 7,497,044-7,547,043.

PDF Location2: Location-based displays: Chromosome 17: 7,497,044-7,547,043.

Paul Flicek, D.Sc.
Team Leader, Vertebrate Genomics
EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton
Cambridge, UK
And
Honorary Faculty Member
Wellcome Trust Sanger Institute

Keywords: ensembl gene annotations, ensembl human gene set, ensembl code, ensembl data, open infrastructure for genome information, reference database for genomics, bioinformatics resource, genome annotations, sequenced mammalian and chordate genomes.

PDF