ScienceDaily
Your source for the latest research news

Web address:
http://www.sciencedaily.com/releases/2012/10/121028142215.htm

Pushing Genome Data Analysis One Step Forward

ScienceDaily (Oct. 28, 2012) — Due to the exponential increase in sequencing capacity, efficient tools for data analysis are becoming essential to process the vast amount of biological data. The GEM project, led by Paolo Ribeca from the Centro Nacional de Análisis Genómico (CNAG) and including scientists from this center and the Center for Genomic Regulation (CRG), allowed the development of a tool for the interpretation of genomic data that is several times faster and much more accurate than other tools currently being used.

The study has been published in the journal *Nature Methods*.

If we use the well-known comparison of the genome with a book, then we can say without fear of being wrong that it is a very complicated book. It is thousands of times bigger than a regular book, with more than 3 billion letters in total, each one being an A, C, G or T, as per the four possible bases of the DNA code. One can see the genome as a sequence of millions of words without breaks between them nor capitalization nor punctuation. Most words occur only once in the genome, but some can be found thousands of times with small variations. And reading this book gets even more complicated when you can only see short sentences with few words, each one randomly extracted from the book.

Last generation sequencing techniques used at the CNAG and the CRG, involve breaking the genome into small pieces (alike to short sentences from the book), sequencing such pieces and trying to find them back in the genome. The next step, mandatory in most biological experiments, would be assigning the sentences to their correct original location. However, this can be an extremely difficult task: sentences might be misspelled (sequencing is not a perfect process, and introduces errors) or slightly different (the genome of the individual being sequenced usually contains small variations if compared to the reference one). In addition, each sequencing experiment produces billions of short sentences.

This is the starting point that led some researchers at the CRG and the CNAG to design a computer program that helps to find sequences in the reference genome, quickly and accurately: such tools, called 'mappers', are essential to interpret data in genomic studies, as they represent the first analysis step for many biological experiments. After 5 years of development the result is the GEM (Genomic Multitool) mapper.

The GEM mapper is several times faster than other reference programs in the field and delivers breathtaking performance, matching into the huge human genome of reference about 40 million sequences per hour on a single CPU core. As it uses algorithms that guarantee that it doesn't miss matches, GEM is also much more accurate than other comparable programs. In addition, GEM allows the parameters of the search to be tuned to the specific requirements of the biological

experiment being performed, offering a versatility that cannot be achieved with most existing tools.

The good performance profile of GEM will help to face a practical problem: the dramatic increase in the amount of sequencing data. As an example, the CNAG started operations in 2010 with a park of 12 second generation sequencers that generated roughly 50 Gbases per day. Thanks to the recent spectacular advances in sequencing technology, today, only 2 and a half years after, the CNAG generates almost 20 times more data with the same number of sequencing machines. However, it would have been impossible to increase the computing resources of the CNAG accordingly (and this is a problem common to biomedical research everywhere in the world). Hence, the development of more efficient analysis tools like GEM is essential to keep up with the increasing rate of production.

The GEM tools are a neat example of excellence research, and a world-class tool, entirely developed in Spain; although the project is lead by an Italian team member, the whole work has been carried out in Barcelona. This accomplishment was made possible by the very early adoption of next-generation sequencing machines at the CRG (in 2008), and the subsequent sustained investment in sequencing technologies by the Catalan and Spanish governments that culminated in the creation of the CNAG.

The research was funded by the Spanish Ministerio de Educación y Ciencia (Consolider program), by the US National Institutes of Health/National Human Genome Research Institute, and by the European Union (READNA and ESGI programs).

Story Source:

The above story is reprinted from materials provided by **Centre for Genomic Regulation**, via AlphaGalileo.

Note: Materials may be edited for content and length. For further information, please contact the source cited above.

Journal Reference:

1. Marco-Sola S, Sammeth M, Guigó R and Ribeca P. **The GEM mapper: fast, accurate and**

versatile alignment by filtration. *Nat. Methods*, 2012 DOI: 10.1038/NMETH.2221

Need to cite this story in your essay, paper, or report? Use one of the following formats:

(•) APA

( ) MLA

Centre for Genomic Regulation (2012, October 28). Pushing genome data analysis one step forward. *ScienceDaily*. Retrieved October 31, 2012, from http://www.sciencedaily.com/releases/2012/10/121028142215.htm

Note: If no author is given, the source is cited instead.