MedicalXpress

Samtools CRAMS in support for improved compression formats

August 15 2014

Computer scientists at the Wellcome Trust Sanger Institute have released a major upgrade of Samtools, one of the most popular next-generation sequence analysis tools. The revised Samtools 1.0 enables researchers to easily compress, share and analyse genomic sequence data, reducing costs and supporting genomics research around the world.

The Global Alliance for Genomics and Health, in which the Sanger Institute is a partner, has been set up to enable researchers and clinicians to work together using standardised and efficient DNA sequence data formats to find the genetic variants responsible for disease. Samtools 1.0 supports this initiative by enabling researchers to read and write data in the new CRAM format, which was recently adopted by the Global Alliance, in addition to the existing SAM and BAM file formats for genomic sequence information.

The benefits of using CRAM are immediate: it gives a size reduction of 10-30 per cent. In addition, in a similar fashion to the JPEG format for images, CRAM supports much greater compression – up to a hundred fold – in "lossy" mode which preserves almost all of the important information.

"This major rebuild of Samtools reflects our commitment to supporting the global use of sequencing data," says Dr Richard Durbin, Head of Computational Genomics at the Sanger Institute. "Genome science worldwide relies on fast and efficient data analysis and storage, and Samtools 1.0 fulfils this need by supporting new sequencing and analysis

technologies."

Samtools software is embedded in many bioinformatics pipelines and is the foundation of many thousands of genomic research papers. Since its creation in 2009, the program has been downloaded more than 225,000 times. Samtools 1.0 is freely available at http://www.htslib.org/. This new version was substantially rewritten to support the highly efficient genomic data format CRAM, add new functionality, and integrate more cleanly with other tools.

"Samtools 1.0 embeds CRAM into genomic data analysis pipelines and removes the need for additional processing," says Dr John Marshall, from the Sanger Institute. "This development paves the way for widespread uptake of this highly efficient file format in genomic research and will lead to lower storage costs."

The significant savings in storage that can be achieved are due to incorporating data compression techniques developed jointly by the Sanger Institute and the EMBL-European Bioinformatics Institute.

"It has been exciting to work on implementing CRAM into Samtools," says James Bonfield, at the Sanger Institute. "The great flexibility of CRAM has allowed a number of new compression techniques to be incorporated, which when combined with Samtools 1.0 will help to future-proof genomic data storage and analysis."

Provided by Wellcome Trust Sanger Institute