>

Sanger Institute, CRG Launch RGASP 3 with Lens Trained on RNA-seq Mapping Tools

December 24, 2010

Sanger Institute, CRG Launch RGASP 3 with Lens Trained on RNA-seq Mapping Tools

By Uduak Grace Thomas

This article has been updated to correct the previously reported deadline for submissions.

The Wellcome Trust Sanger Institute and the Spanish Center for Genomic Regulation have launched the third round of the jointly organized RNA-seq Genome Annotation Assessment Project, or RGASP, which aims to enable "fair evaluation of different analysis methods within the community generating high-quality RNA-seq read alignments that can be used for efficient transcriptome characterization."

While the two previous RGASP challenges evaluated computational methods for producing quantitative estimates of gene and transcript activity or expression based on RNA-seq data, Roderic Guigo, who leads the bioinformatics and genomics program at CRG, told *BioInform* that RGASP 3 will focus specifically on evaluating tools used to map RNA-seq reads to reference genomes and annotated transcripts.

"One thing we realized from the previous RGASPs is that one of the key components of the process of quantification is mapping the RNA-seq reads onto the genome and a transcriptome" he said, adding that "this is more complicated than people expect."

Participants are expected to submit their predictions by Jan. 3, 2011.

Among other measures, assessors will evaluate the predictions using annotations provided by the GENCODE project, a subset of the Encyclopedia of DNA Elements, or ENCODE, initiative focused on annotating all evidence-based gene features in the human genome at high accuracy. RGASP 3 evaluators will also use standard gene prediction assessment metrics, such as sensitivity, specificity, and correlation coefficient at the nucleotide, exon, transcript, and gene level.

So far, participants have been provided with RNA-seq datasets for mouse whole brain and a human embryonic kidney cell line, as well as the relevant reference genomes. The organizers said that a third dataset comprised of simulated reads might be added at a later date.

David Adams at the Sanger Institute provided the mouse dataset and Thomas

Gingeras at Cold Spring Harbor Laboratory provided the human dataset.

The results of RGASP 3 will be presented along with the results of a related challenge that's focused on evaluating tools used for *de novo* genome assembly (*BI* 12/10/2010) at a workshop in Barcelona in April next year, organized and hosted by the International Center for Scientific Debate.

Creating a Gold Standard

On the project website, the organizers write that RGASP was launched last year to "assess the current progress of automatic gene building using RNA-seq as its primary dataset."

Guigo echoed similar sentiments in his comments to *BioInform* about the project. "RNA-seq is a very novel technique and it is still very unclear how to actually use it and how to use the RNA-seq results to get good ... estimates of transcript expression," he said. "RGASP [aims] to compare different methods [used] to produce ... estimates of gene and transcript expression based on RNA-seq data."

The challenge is organized within the framework of the ENCODE project and is modeled after the ENCODE Gene Prediction Workshop, or EGASP, which took place in May 2005. EGASP aimed to evaluate methods "to reproduce the manual and experimental gene annotation of the human genome" with an emphasis on protein-coding genes.

In RGASP 1, participants were provided with RNA sequence data and the corresponding reference annotations for human, fly, and worm; while in RGASP 2, groups received human RNA-seq data and the reference annotations. About 16 groups participated in the two rounds.

As part of the evaluative process, assessors compared predictions to the GENCODE annotation produced under the ENCODE project. Furthermore, predictions that weren't covered by GENCODE were validated experimentally, the organizers said.

While preliminary results of rounds one and two are available for RGASP participants, they haven’t been released more broadly yet because the assessors are still evaluating submissions, Guigo said.

He explained that one of the challenges they face is that there isn't a "good gold standard" in place that provides the "real expression of the transcripts and the genes" in the samples being analyzed. This makes it difficult, he said, to compare programs in terms of the accuracy of the predictions.

To help with the assessment, the RGASP team is working to create an independent standard by selecting several hundred loci from the data and performing "directed quantification" of the genes and transcripts with NanoString's digital gene expression platform.

In addition to the Sanger Institute and CRG, researchers at the California Institute of Technology, Yale University, University of Lausanne, and the University of California, Berkeley, have organized and contributed datasets to the three RGASP challenges.

Have topics you'd like to see covered in *BioInform*? Contact the editor at uthomas [at] genomeweb [.] com.

Related Stories

- Informatics Sector Sees Translational Research Opportunity, Scratches Cloud Computing Itch in 2010
  December 24, 2010 / BioInform
- 5AM Solutions Sets Up Shop out West to Tap into Bay Area Life Sciences Market
  December 24, 2010 / BioInform
- GenomeQuest, SGI Deploy Genome Analysis Platform for Ag Biotech Research at Biogemma
  December 24, 2010 / BioInform
- Downloads and Upgrades
  December 24, 2010 / BioInform
- People in the News
  December 24, 2010 / BioInform