



Using Galaxy to facilitate RNA-Seq analysis

Luís Augusto Eijy Nagai, Felipe Rodrigues Silva, Leandro Carrijo Cintra, Poliana

Fernanda Giachetto, Adhemar Zerlotini Neto

Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária, Embrapa Informática Agropecuária

Galaxy is an open-source, web-based and friendly platform to creating and executing, among others, a complete RNA-Seq analysis workflow. The amount of RNA-Seq data available increased due to a simplified data analysis process, allied with improvements in sequencing techniques and cost, thus making RNA-Seq an essential tool for transcriptome research. Nevertheless, RNA-Seq experiments must be analyzed with robust, efficient and statistically grounded algorithms. Fortunately, researchers and developers are further advancing such software tools in order to make it possible to focus on data analyze rather than computational tasks such as software installation and file formats. The Galaxy project integrate all kinds of bioinformatics tools including the Tuxedo Suite for RNA-Seq analysis (Bowtie, Tophat, and Cufflinks). TopHat aligns reads to a reference genome and Cufflinks uses these mapping files in order to assemble the reads into transcripts and estimate their expression index (FPKM - fragments per kilobase of transcript per million fragments mapped). Cufflinks also have a differential expression tool that uses the FPKM index of different conditions and reports significant changes in expression using a rigorous statistical analysis. These tools are gaining wide acceptance and have been used in a number of recent high-resolution transcriptome studies. CummeRbund renders Cufflinks output in publication-ready figures and plots. A complete analysis of 200Gb of transcript data is feasible with no previous advanced computational knowledge and can be executed, from the raw data to the expression index calculation, in a period of less then two weeks. In this experiment, two RNA-Seq datasets were analyzed in order to evaluate both, the Galaxy platform and the RNA-Seq analysis suite. We were able to observe that such combination of softwares is really effective on allowing two biologists researchers to perform the whole transcriptome analysis through a intuitive web interface. The Galaxy project has a helpful growing community, tutorials and installation was simple in UNIX/Linux. Even though whole transcriptome analysis usually produces very large files, the Galaxy platform allows easy visualization through an embedded genome browser and filtering tools to selecting and sorting tabular data. Hence, we obtained satisfactory results in a short period of time executed by non computer experts without extensive bioinformatics training. Both the simplified workflow course and reduced learning curve are relevant points to motivate new users to get started using Galaxy for RNA-Seq analysis. Supported by: Embrapa.

Keywords: Galaxy, RNA-Seq, Tophat, Cufflinks

Concentration area: Transcrip Proteomic

