bcbio-nextgen: Automated, distributed next-gen sequencing pipeline

Roman Valls Guimera, Science for life genomics staff, Brad Chapman

Harvard School of Public Health, Bioinformatics Core; Cambridge, MA USA Science for Life Laboratory, Stockholm, Sweden

http://www.hsph.harvard.edu/research/bioinfocore/

http://www.scilifelab.se/

bcbio-nextgen is an Python framework for next generation sequencing analysis. The fully automated pipeline interacts with the sequencing machine, runs sequences through configurable processing pipelines, and uploads the data into Galaxy for visualization and additional processing. The variant calling analysis pipeline handles alignment to a reference genome, variant identification with GATK and preparation of summary PDF files for assessing run quality.

The pipeline is fully distributed and will run on single multicore machines or in compute clusters managed by LSF, SGE or SLURM. The CloudBioLinux and CloudMan projects utilize this pipeline for distributed analysis on Amazon cloud infrastructure.

The Galaxy web-based analysis tool can be optionally integrated with the analysis scripts. Tracking of samples occurs via a web based LIMS system, and processed results are uploading into Galaxy Data Libraries for researcher access and additional analysis.

Relevant Web sites

1. http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/
2. http://bcbio.wordpress.com/2011/09/10/parallel-approaches-in-next-generation-sequencing-analysis-pipelines/
3. https://github.com/brainstorm/bcbb/tree/master/nextgen
4. http://www.slideshare.net/chapmanb/developing-distributed-analysis-pipelines-with-shared-community-resources-using-cloudbiolinux-and-cloudman