ER-flow Application Description Template

Application Name: Tophat

Application domain: RNA Sequencing

Brief description of application
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park

input data:

- List of files containing paired-end reads in FASTQ or FASTA format, both left and right reads.

output data:

- A list of read alignments in SAM/BAM format

sample data: (lfn: /grid/vlemed/AMC-e-BioScience/RNASeq/data)
application (http://ccb.jhu.edu/software/tophat/index.shtml)
docs (http://ccb.jhu.edu/software/tophat/manual.shtml)
publication (http://www.nature.com/nprot/journal/v7/n3/full/nprot.2012.016.html)

Execution environment
DCI: (EGI, SRM/LFC, vlemed VO)
middleware: gLite workflow system: WS-PGRADE

Execution characteristics
input: 10MB-100GB output: 10MB-100GB
processing time (per unit): 1 - 36 hours
memory usage: n.a. disk usage: ~10MB-100GB

Target users
Bioinformatitions (LUMC, AMC); number of users: 10+; user type: end-user

Usage scenario for workflow in the ER-FLOW
A WS-PGRADE workflow has been implemented to port this application to EGI for the vlemed VO. The workflow is published on the SHIWA repository with appropriate documentation, metadata and sample data.

Contact information (workflow developers)
name: Vladimir Korkhov e-mail: vkorkhov@gmail.com