Using Galaxy to pre-process RNA-Seq data (FASTQ files) for importing to BRB-ArrayTools

Galaxy is a web-based tool through which users can process and analyze their next-generation sequencing (NGS) data. The basic procedure of processing the RNA-Seq data through Galaxy is described in the following steps,

1) Input data file
   At the Galaxy website (http://main.g2.bx.psu.edu/), the user can click on "Get Data" -> "Upload" files to input their FASTQ files to Galaxy. For files smaller than 2 Gb, the users can upload file directly to Galaxy by clicking on "choose file". For files larger than 2 Gb, it is recommended to use ftp protocol to upload files.
2) QC and data manipulation
   1. FASTQ groomer
      The format of FASTQ files obtained from different sequencing software can be different. For some FASTQ files, the FASTQ groomer needs to be run to convert the FASTQ file to standard format that can be used by Galaxy for downstream processing. For some FASTQ files already in standard format, this step might not be necessary. However, running groomer on a FASTQ file with standard format will produce the same file, which will not affect downstream processing/analyses.
   2. FASTQ joiner
      For paired end data in two separate files (end 1 and end 2), FASTQ joiner can be run to join them into one single-end file. This step is not necessary for single-end data file.
   3. FASTQ summary statistics
      By running "FASTQ summary statistics" tool, the summary statistics of a FASTQ file is created. The quality scores of each sequencing cycle are included in the file.
   4. Boxplot of the quality scores
      Boxplot of the quality scores for each sequencing cycle can be plotted by the "Boxplot" tool. This can be used to determine whether the reads with low median quality score needs to be removed by the "FASTQ trimmer" tool.
   5. FASTQ trimmer
      The "FASTQ trimmer" tool can be used to trim the end of reads whose median quality score did not pass the threshold. For instance, if a median quality score of 14 and lower is considered a bad score, the reads whose median quality scores are lower than 15 will be trimmed. A new FASTQ file will be generated for future use.
3) Mapping the reads by "Tophat for Illumina"
   For data from different sequencers there are different ways of mapping the reads against a reference genome. Here we only focus on the Illumina data. The "Tophat for Illumina" tool can be used to map the processed reads to a selected reference genome. This tool produces a BAM file than can be further used by "Cufflinks" tool to estimate the FPKM (fragments per kilobase of exon per million fragments mapped) for each assembled transcript or gene. *"Tophat is 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."(Excerpted from Galaxy website).

4) Transcript assembly and FPKM estimation by "Cufflinks"
   Using the BAM file produced by "Tophat for Illumina" as the input file, the "Cufflinks" tool assembles the reads into transcripts and estimates their abundance indicated by FPKM. It produces three files. One contains information for each assembled transcript isoform information. The other two contain the expression information at the transcript (isoform-based) and gene (gene-based) level, respectively. For each assembled transcript, the location of the transcript, but no gene identifier information is included. Therefore, "Cuffcompare" needs to be run to compare the assembled transcripts to a reference annotation.
5) "Cuffcompare"
   For all the above steps, each run is done on one sample. For multiple samples, the above sequence of steps must be run. The different assembled transcripts obtained need to be compared and aligned to a common reference annotation file so as to obtain the annotation information across multiple samples. This is done by the "Cuffcompare" tool. The user needs to input a reference annotation file in GTF format to be compared with the cufflink output files from multiple samples. The GTF file can be prepared by the user, or imported by clicking on "Get Data" at Galaxy to different websites of interest. For instance the user can click on "GetData"-> "UCSC Main table browser" to download the annotation file of interest, such as the "refGene" file in the correct genome build in GTF format."Cuffcompare" produces several files, one of which contains the annotation information with combined transcripts from different samples. The type of annotation information is dependent on the reference annotation file the user selects. For instance, if the refGene file is used as the reference annotation file, the annotation information will be RefSeq Ids, such as "NM_00001".
6) Obtaining output files that can be imported into BRB-ArrayTools using "Cuffdiff".
   After running "Cuffcompare", "Cuffdiff" can be run to obtain the estimated FPKM values of each gene or transcript (isoform) for each sample with annotations. The input files are the BAM output files generated from "Tophat" for all samples, as well as the "combined transcripts" file obtained from Cuffcompare. "Cuffdiff" also produces several files. The "gene FPKM tracking" file is the one to be imported into BRB-ArrayTools. It contains the tracking Id for each transcript associated with the annotation Id (such as the RefSeq Id), as well as the FPKM values for each sample. This file should be ready for importing into BRB-ArrayTools through the Data Import Wizard.