Icahn School of Medicine at Mount Sinai LINCS Center for Drug Toxicity Signatures

Standard Operating Procedure: Generation of Gene Read Counts

DToxS SOP Index: CO-3.1
Last Revision: March 24, 2016
This version is a revision of CO-3.0. It contains changes to steps 10b and 11
Written By: Yuguang Xiong and Eric Sobie

Approvals (Date): Joseph Goldfarb (3/25/16)
Marc Birtwistle (3/30/16)
Eric Sobie (3/30/16)
Ravi Iyengar (3/30/16)

Quality Assurance/Control (QA/QC) steps are indicated with green highlight.

Metadata recording is highlighted with yellow highlight and superscript indices.

---

1) Set up a Linux, Unix, or Mac OS X operating system platform which uses *Bash* as its default shell program.
2) Download, compile, and install the mRNA-sequencing alignment software Burrows-Wheeler Aligner (BWA)[1], and add the location of the executable file *bwa* to the system's `PATH` variable. We have been using BWA version 0.7.12-r1039.
3) Install the Python[2] programming development software to the system. We have been using Python version 2.7.6.
4) Create a top alignment directory `[top-align-dir]` to store all the datasets, e.g. `~/LINCS/Alignment`
5) Create 3 directories under `[top-align-dir]` to include the datasets for sequences, alignment and read counting:
   a) `"Seqs"` for mRNA-sequencing data.
   b) `"Aligns"` for alignment data.
   c) `"Counts"` for read counts data.
6) Download (upon request at DToxS website: see e-mails below under Correspondence) a set of gzip-compressed FASTQ-format RNA-sequencing data files of a released dataset to the "`Seqs`" directory created above. The data file names should have the same form as the following:
   RNAseq_[date]_Lane1_R1.fastq.gz, *RNAseq_[date]_Lane1_R2.fastq.gz*, *RNAseq_[date]_Lane2_R1.fastq.gz*, *RNAseq_[date]_Lane2_R2.fastq.gz*, etc
   where:
   `[date]` is the date tag of the dataset.

These separate files would correspond to the raw mRNA-seq reads from an experiment that used multiple lanes of a flowcell and ran sequencing across multiple flowcell runs.

7) Create a reference directory "`References`" under `[top-align-dir]`, e.g. `[top-align-dir]/References`, denoted as `[ref-dir]`
8) Download (upon request at DToxS website: see e-mails below under Correspondence) the gzip-compressed reference genome library file[3] to the reference directory `[ref-dir]`, and extract the compressed library file to create a directory, e.g. `[ref-dir]/Broad_UMI`, which contains all needed library files:
   a) The human genome reference library for mapping mRNA fragments to genes (the `Human_RefSeq` folder).
   b) The plate barcodes data files for assigning sequence read counts tagged with particular barcodes to corresponding plate wells:
      i) 96-well plate: barcodes`_trugrade_96_set2.dat` and `barcodes_trugrade_96_set4.dat`.
      ii) 384-well plate: `barcodes_trugrade_384_set1.dat`

      Note: the barcode configuration is preset in the shell script `run-alignment-analysis.sh` for each released dataset described below:
9) Create a program directory `Programs` under `[top-align-dir]`, e.g. `[top-align-dir]/Programs`, denoted as `[prog-dir]`.
10) Download (upon request at DToxS website: see e-mails below under Correspondence) the gzip-compressed program file and extract it to the program directory `[prog-dir]`, e.g. `[prog-dir]/Broad-DGE`, which contains all needed programs:
    a) The main program: run-alignment-analysis.sh
    b) The function programs: `split_and_align.py` and `merge_and_count.py`.
11) Open the shell script file `run-alignment-analysis.sh` in the `[prog-dir]` directory with a text editor and update the following program variables located at the beginning of the script file:
    a) In Section 1.1 `Global`:
       i) Set the top alignment directory `TOP_DIR` to `[top-align-dir]` created at *Step 4*.
    b) In Section 1.2 `Dataset`:
       i) Set the series number `SERIES` to that of the downloaded dataset.
       ii) Set the number of lanes `LANES` to the maximum lane number shown in the name of the data files at *Step 6*.
       iii) Set the data directory `DATA_DIR` to `[top-align-dir]` created at *Step 4*.
    c) In section 1.3 `Reference`:
       i) Set the reference directory `REF_DIR` to `[ref-dir]` created at *Step 7*.
    d) In section 1.4 `Program`:
       i) Set the program directory `PROG_DIR` to `[prog-dir]` created at *Step 9*.
       ii) Set `THREAD_NUMBER` to the number of processors or cores available on the computing platform.
12) Run the shell script `run-alignment-analysis.sh` using the following command line:
    `[prog-dir]/run-alignment-analysis.sh >& [top-align-dir]/run-alignment-analysis.log`
    where:
    a) `[prog-dir]` is the program directory created at *Step 9*.
    b) `[top-align-dir]` is the dataset directory created at *Step 4*.
    c) `run-alignment-analysis.log` is a text file that logs the screen outputs of the script.

    Note: this step may take several hours to finish depending on data volume and system configuration.

13) When *Step 12* finishes, check the data files generated by the shell script in corresponding directories:∑
    a) The `Aligns` directory contains:
        i) The data files with a name form of *RNAseq_[date].Lane[number].[number].fastq* are the multiple split parts of each gzip-compressed FASTQ format mRNA-sequencing data file *RNAseq_[date].Lane[number].fastq.gz* described at *Step 6*.
        ii) The data files with a name form of *RNAseq_[date].Lane[number].[number].fastq.sam* are the alignment outputs for mRNA sequence fragments.
    b) The `Counts` directory contains multiple files starting with *RNAseq_[date]*, among which:
        i) *RNAseq_[date].unq.refseq.umi.dat* contains the UMI read counts for all annotated genes in the reference library at all drug-treated conditions, which will be used as the inputs of subsequent analysis of differential gene expressions.
        ii) *RNAseq_[date].unq.well_summary.dat* contains a summary of uniquely aligned UMI and non-UMI sequence read counts for each sample.
        iii) *RNAseq_[date].unq.log.dat* contains various statistics of unique alignment of different types of sequences for entire sample set, e.g. the total read counts, the read counts assigned to samples, and the read counts aligned to reference genome library.
        iv) All other data files are about the uniquely or non-uniquely aligned UMI or non-UMI read counts for different types of sequences.

Metadata

1. **Burrows-Wheeler Aligner (BWA):**
   version 0.7.12-r1039,
   http://bio-bwa.sourceforge.net.
2. **Python:**
   version 2.7.6,
   https://www.python.org.
3. **Reference Genome:**
   A tailored version of UCSC Human Genome hg19 made by the Broad Institute (received in April 2015),

Correspondence

Yuguang Xiong
E-mail: yuguang.xiong@mssm.edu

Evren Azeloglu
E-mail: evren.azeloglu@mssm.edu