Bioinformatics

Project 2

RNA-seq Data Mapping to Genome and alternative splicing analysis

RNA-seq data:

1) You can search NCBI SRA database for RNA-seq data of the species of your interest. For example, SRR11006300 is a dataset from tomato. For data size <5 Gb, you can download the data directly from the user's interface. For data size > 5GB, you need to use a standalone sra tool kit to downland manually.

SRA database and SRA toolkit: https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/
The command is bellow, for example, you installed the toolkit in Windows:
\"program files\ncbi\sratoolkit\bin\fasterq-dump.exe" SRR11006300 –split-files

Genome data

2) Go to the NCBI Genome database for genome sequence downloading, for example, for tomato as: https://www.ncbi.nlm.nih.gov/genome/?term=tomato

After you have data available, you need to install the following tools (if you use your own computer. However, I have not tested if they work in Mac or Windows). Otherwise, **use the Linux computers in the lab.**

1) bowtie2 manual: http://computing.bio.cam.ac.uk/local/doc/bowtie2.html
2) bowtie2 download: https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.1.0/
3) tophat manual: https://ccb.jhu.edu/software/tophat/manual.shtml
4) tophat download: http://ccb.jhu.edu/software/tophat/downloads/
5) cufflinks manual: http://cole-trapnell-lab.github.io/cufflinks/manual/
6) cufflinks download: http://cole-trapnell-lab.github.io/cufflinks/install/

3) Easy access Linux, you can install the following software (for Window computer):
SSHSecureShellClient-3.2.9.exe
http://proteomics.ysu.edu/courses/BIOL4800_6900/LAB/

If you have a Mac computer, you can use your Mac terminal to access the Linux directly. You can learn basic Linux command using the following reference:
http://proteomics.ysu.edu/courses/BIOL4800_6900/ExtraReadings/ShellIntro.pdf

Procedure (to save time, you use the data I prepared for you, I suggest to use pineapple data as it is smaller, however, if you like to get more experience, you may choose a species you like to work on)

1) In Linux, after login, you first "mkdir your_name" to create a working directory.
2) Use one of the dataset (either pineapple or potato), download from
   http://proteomics.ysu.edu/courses/BIOL4800_6900/LAB/RNAseq/

   all the data files into your own directory. There are three or four files: genome (fasta); RNA files (1 in pineapple, 2 files in potato) (fastq format); Gff3 (annotation)
   Note: do not try to open the files -the file too big – will freeze your computer. Use right click of your mouse, then use "Save link as" to download the files.

The tools on Linux stations are installed at /home/tools/

How to run bowtie, TopHat and Cufflinks within your own RNAseq directory:

To run bowtie:
/home/tools/bowtie/bowtie2-build genome.fas genome.index

To run tophat:
1) if no gff file:
   /home/tools/tophat/tophat2 genome.index (from step1) RNA.fastq (fastq RNA-seq file)
2) for genome has a gff annotation file, use the following command:

/home/tools/tophat/tophat2 –G Genome_annotation.gff --transcriptome-index=gene_model/species_name genome.index RNA.fasta (RNA_1.fastq RNA_2.fastq)

3) After running tophat: A summary of the alignment counts can be found in tophat_out/align_summary.txt, and other files including "accepted_hits.bam".

To run cufflinks:
(1) using gff as a guide for mapping:
/home/tools/cufflinks/cufflinks tophat_out/accepted_hits.bam -g Genome_annotation.gff3

(2) if there is no annotaition file, just run as below
/home/tools/cufflinks/cufflinks /path/tophat_out/accepted_hits.bam

Note* I often use $mkdir to make a new directory for running cufflinks. You need to provide a right path for your input file

The output files from cufflinks contains gene and isoform expression information and the "transcripts.gtf" has the mapping information and the FPKM values.

Alternative Splicing Events Classification

After you run cufflinks, you will get "transcripts.gtf" file. Then use this file as input for next step.

Step 1: In the same director where your "transcripts.gtf" file is located, to run ($ is the terminal)
$/home/tools/astalavista-3.2/bin/astalavista -t asta -i transcripts.gtf -o AS.gtf.gz
Please note: the -o is to specify the output file, I suggest to use your data access number as part of the output file name, as you need to have ".gz" in the file – it is required by the software.

Step 2: Run the following command to unzip the output file
$gunzip SRR##_AS.gtf.gz

Step 3: Run the following:
$/home/tools/perl/gtf2events_standalone_parse.pl AS.gtf landscape.gtf
#note: AS.gtf as input; landscape.gtf as output -which contains readable AS classification. A file "summary.events" will be automatically created which contains the summary of AS events.

Step 4: The astalavista tool has a bug – the output file contains duplicated AS pairs – we need to correct the file output

Run the following command:
$ /home/tools/perl/uniqPair2.1.pl landscape.gtf landscape.gtf.uniq
Note: landscape.gtf – input which is generated in step 3. landscape.gtf.uniq – final output. A new summary file will also be generated automatically with a name events.summary".

Required submission –

Write a short lab report on the procedure and the output from tophat about the mapping information (data from the alignment summary file) and the AS events from **astalavista** server or stand-alone tool. Your project2 report – is like a short paper including Intro, Data and Methods, Result (only the mapping AS event summary table) and Discussion, Ref.

Extra-readings: read the following –

Clark S, Yu F, Gu L, Min XJ (2019) Expanding alternative splicing identification by integrating multiple sources of transcription data in tomato. Frontiers in Plant Sciences. 10:689. doi:10.3389/fpls.2019.00689.