Notes on RNA-seq Analysis Workflows

Amy Olex

## General RNA-seq Pipeline Steps

Below is a list of information I compiled during my research on how various people handle NGS data. Some of these preparation steps may already be performed automatically by the sequencing facility’s software. You may not need to follow all these steps as they can be project-dependent.

The Wikipedia page from here <http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools> lists tools for every step of processing RNA-seq data.

1. **FASTQ Preparation:** This is a list of steps other people recommend you do before alignment. You may not need to do all of these and many are already done by the sequencing facility.
   1. Demultiplex (usually done by facility)
   2. Trim to remove low-quality 3’ bases
   3. Remove adapter sequences (usually done by facility)
   4. Remove sequences that map to undesired RNA, like rRNA.
   5. Filter reads by k-mer coverage
   6. Normalize k-mer coverage
   7. Calculate mean and stdev of insert sizes for paired-end data.
2. **Alignment**
   1. DNA-seq
      1. Many aligners out there. I’ve used BWA and Bowtie2 so far.
   2. RNA-seq
      1. There are tons of aligners out there, but due to the splicing of transcripts you have to make sure you pick one that can adjust for this. So far I’ve used TopHat. You may also consider PRADA[1] as it is based on BWA instead of TopHat.
3. **Quantitative Analysis and Differential Expression**
   1. **Get read counts for DE analysis.** Right now I’m using HTSeq-count to get the read counts from the BAM files. This is one of the most popular, but there are others out there. You need to make sure you understand the setting for HTseq-count before using it as it can count reads differently depending. See this link for more details: <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>
   2. **Count normalization.** The paper by Dillies et al (2012)[2] gives a good review of the various normalization methods for RNA-seq data. Their conclusions were that DESeq and edgeR are the most robust, and they assume that most of the genes are not differentially expressed, which I think matches the biology.
   3. **Identify differentially expressed genes.** If using DESeq or edgeR, they both do this for you. There are also tons of other tools out there, but these are the most popular ones. The web interface provided by DEB [3] allows you to compare DE genes between DESeq, edgeR and bayseq. Also, the R package RNASeqDEPipeline [4] allows you to compare DE genes between the three just mentioned as well as Cuffdiff and one other.
4. **Enrichment analysis and visualizations**
   1. Once you have your list of differentially expressed genes you may want to perform enrichment analyses on them, or other analyses like clustering etc. These can be performed in a similar manner to microarray data. I am still compiling this information for this document.

**My Workflow for RNA-Seq Data**

## FASTQ Pre-processing

1. FASTQC: use to determine the quality of each FASTQ file.
2. Trimmomatic: use to trim reads if needed.

## TopHat Alignment, BAM Processing and HTSeq-count

I have written a script to run TopHat Alignment, BAM pre-processing and read counting with HTseq-count on my FASTQ files. To run this I need my reference genome and the gene annotations. The locations for my script and data are below:

Location of automated script:

/home/alolex/bin/rnaseq-pipeline

Location of Arabidopsis reference for TopHat2:

/data/iGenomes/Arabidopsis\_thaliana/NCBI/TAIR10/Sequence/Bowtie2Index/genome

Location of Arabidopsis annotations for TopHat2:

/data/iGenomes/Arabidopsis\_thaliana/NCBI/TAIR10/Annotation/Genes/genes.gtf

**Steps to use script (assuming it is in /home/usr/bin):**

1. Place pre-processed FASTQ files in working directory.
2. CD to directory
3. Create a text file with the names of the fastq files. For Single reads, list one file per line. For paired-end reads, list two files per line separated by a space. The first file of each line will be used to name the alignment output folders from tophat. Listing the full path to each file is recommended, but the script will work with just the file names, or relative paths as long as you execute the script in the appropriate parent directory.
4. Run the script: >> rnaseq-pipeline -f infile.txt -a /path/to/genes.gtf -i /path/to/Bowtie2Index/genome -o optionalOutDirectory
5. Your HTseq-count files should now be in the output directory.
6. Open R for DE analysis. Then you can follow the EdgeR or DEseq2 vignettes.

**Further Processing:** For my project I found out that they ran the same sample in multiple lanes. This means I have 4 FASTQ files for each sample. These were run through the alignment individually to get 2 BAM files for each sample. TO account for this I had to go outside of my script:

1. I merged the BAM files before running HTSeq Count. To do this I used PICARD MergeSamFiles tool.
2. I put all the merged bam files into one directory and then re-ran HTSeq Count.

## 

## 

## References

[1] PRADA: pipeline for RNA sequencing data analysis. Bioinformatics (Oxford, England), Vol. 30, No. 15. (1 August 2014), pp. 2224-2226, [doi:10.1093/bioinformatics/btu169](http://dx.doi.org/10.1093/bioinformatics/btu169) by Wandaliz Torres-García, Siyuan Zheng, Andrey Sivachenko, et al.

[2] A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics, Vol. 14, No. 6. (01 November 2013), pp. 671-683, [doi:10.1093/bib/bbs046](http://dx.doi.org/10.1093/bib/bbs046) by Marie-Agnès A. Dillies, Andrea Rau, Julie Aubert, et al.

[3] DEB: A web interface for RNA-seq digital gene expression analysis. Bioinformation, Vol. 7, No. 1. (2011), pp. 44-45 by Ji Qiang Q. Yao, Fahong Yu

[4] https://github.com/riverlee/RNASeqDEPipeline