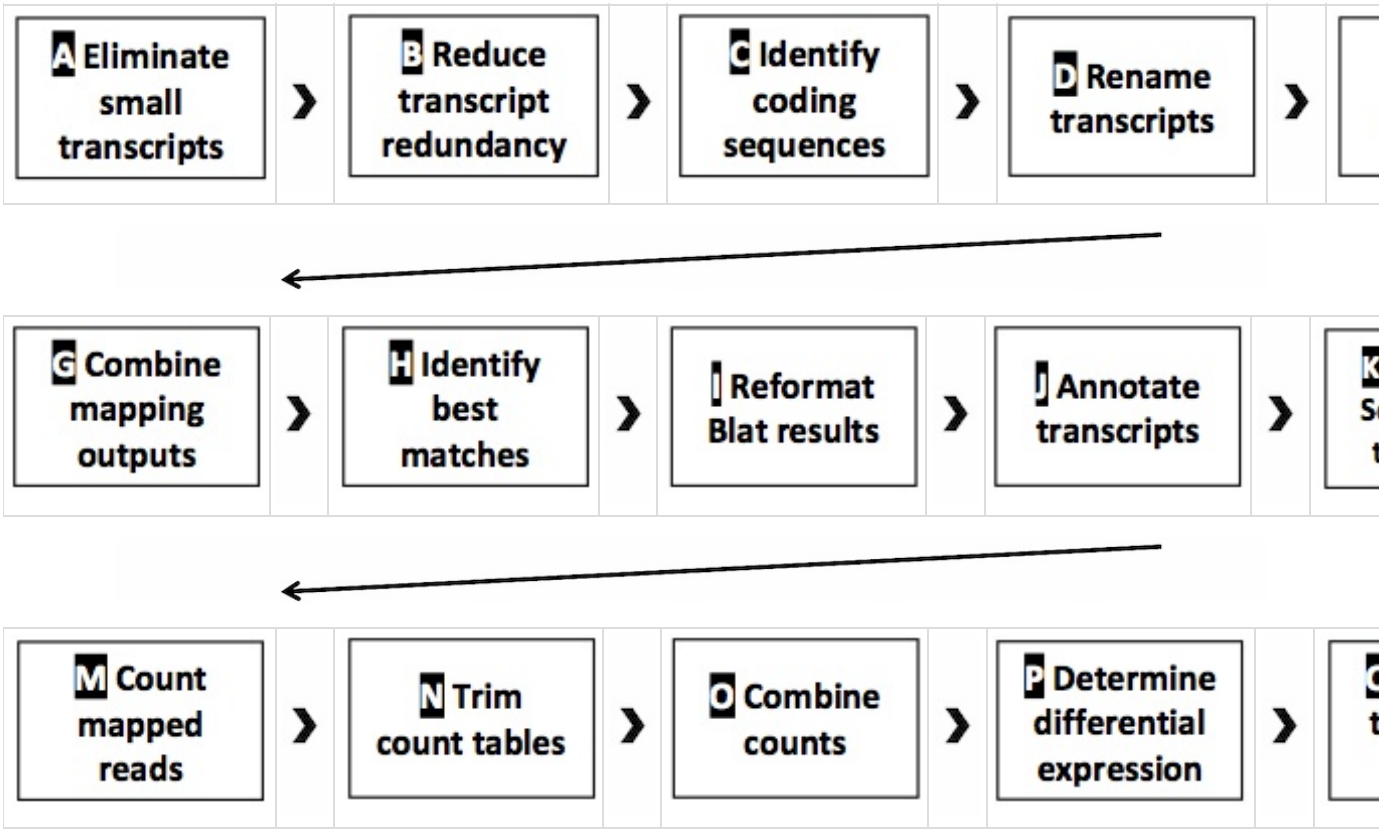
Tutorial: Characterizing Differential Expression With RNA-Seq (Without Reference Genome)

Alert:

The iPlant App Store is currently being restructured, and apps are being moved to an HPC environment. During this transition, users may occasionally be unable to locate or use apps that are listed in our tutorials. In many cases, these apps can be located by searching them using the search bar at the top of the Apps window in the DE. To increase the chance for search success, try not searching the entire app name and version number but only the portion that refers to the app's function or origin (e.g. 'SOAPdenovo' instead of 'SOAPdenovo-Trans 1.01'). In critical cases, please report your concern to the iPlant Ask forum or to support@iplantcollaborative.org. Thank you for your patience.

Tutorial under review

Please work through the tutorial and add your comments on the bottom of this page. Or send comments per email to bmlau@email.arizona.edu. Thank you.



Introduction and Overview

Author: Dr. Roger Barthelson, iPlant Collaborative/University of Arizona

Goal

Identify changes in gene expression levels between at least two sequenced transcriptome samples.

Approximate tutorial completion time: 3 hours (Using the pre-computed iPlant sample data from a study in Belgica antarctica (Teets et al., 2012).)

Rationale and Background

RNA-Seq refers to whole transcriptome sequencing of cDNA, generally using a high-throughput ("next-generation") sequencing technology. RNA-Seq generates deep-coverage information about samples' mRNA. This can be used for a variety of purposes, such as: transcriptome assembly, gene discovery and annotation. RNA-Seq is also used to detect differential transcript abundance between tissues, developmental stages, genetic backgrounds, and environmental conditions.

This RNA-Seq analysis tutorial differs from other RNA-Seq tutorials in that it does not require an assembled reference genome. It still requires an assembled transcriptome however; assembly of transcriptomes is described in other tutorials such as Transcriptome Assembly (de novo) and BLAST a Transcriptome.

The protocol addresses seven primary objectives:

- Prepare data sets (Section A. and B.).
- Identify coding sequences (Section C.).
- Identify putative functions for coding sequences (Section F.).
- Map coding sequences against reference transcriptome (Section K.).
- Count the reads that map to each coding sequence (Section M.).
- Identify differentially expressed coding sequences (Section P.).
- Differentiate between genes that are up-regulated and genes that are down-regulated (Section R.).

Additional sections consist of reformatting, splitting and combining result files (outputs) from a prior step into the inputs for a subsequent analysis.

Pre-Requisites

1. An iPlant account. (Register for an iPlant account at user.iplantcollaborative.org.)
2. An up-to-date web browser, java-enabled. (Firefox recommended. If you wish to work with large data sets of your own and upload them using the iDrop Lite web interface, Chrome is not suitable due to its issues in utilizing 64-bit Java.)
3. Assembled transcript set/transcriptome. (Use your own data or a sample data set provided.)
4. The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.

Test Data

The sample/test data is derived from a set of studies, performed by Nicholas Teets and others, with an antarctic flightless midge. The midge, living in an environment where liquid water is unavailable much of the year, must be able to tolerate dehydration in order to survive. The published RNA-Seq studies tested a number of conditions, including dehydration and compared them to control conditions. The RNA-Seq reads are Illumina Genome Analyzer II reads retrieved from the NCBI Sequence Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/sra/?term=Belgica%20antarctica. Because the reads, when tested with FastQC appeared to be already trimmed, no further trimming was done. The reference for the study is Gene expression changes governing extreme dehydration tolerance in an Antarctic insect, Nicholas M. Teets, Justin T. Peyton, Herve Colinet, David Renault, Joanna L. Kelley, Yuta Kawarasaki, Richard E. Lee, Jr, David L. Denlinger, Proc Natl Acad Sci U S A. 2012 December 11; 109(50): 20744--20749.

Workflow

A. Eliminate small transcripts (app: Select contigs)
B. Reduce transcript redundancy (app: CD-HIT-est 4.6.1)
C. Identify coding sequences (app: Transcript decoder 1.0)
D. Rename transcripts (app: Linux stream editor)
E. Split RefSeq file (app: Split FASTA file)
F. Map transcripts (app: Blat (with options))
G. Combine mapping outputs (app: Concatenate Multiple Files)
H. Identify best matches (app: Best Hit for Blat Output)
I. Reformat Blat results (app: Cut Columns)
J. Annotate transcripts (app: Rename contigs 2.0)
K. Map RNA-Seq reads to transcripts (app: Bowtie-2.2.1--Build-and-Map)
L. Reformat mapping output (app: SAM to sorted BAM)
M. Count mapped reads (app: Index BAM and get stats)
N. Trim count tables (app: Cut Columns)
O. Combine counts (app: Join multiple tab-delimited files)
P. Determine differential expression (app: DESeq)
Q. Separate transcripts by type (app: Numeric Evaluation of a Data Column)
R. Generate transcript lists (app: Cut Columns)

Approximate analysis durations for the iPlant sample data are provided in each step. With other data sets, depending on size, it could take less or more time. Using the sample data, users can skip through the workflow (a la 'cooking show'), returning later to examine the results of their own analysis.

Summary

1. This protocol takes the user through many steps that lead from RNA-Seq reads for an organism without a reference genome, and eventually ends with lists of differentially regulated transcripts. The transcripts are newly assembled and not firmly established. The annotation that is developed for them is also provisional, and so the results are a starting place for researchers to pursue their own interests with respect to determining which genes mediate the phenotypic differences between the conditions studied. Many different tools are presented to the user, and should be seen as part of a flexible set of procedures that can be followed in a workflow order that works best for the researcher. For example, many may prefer to leave out the minimal annotation of the transcripts in favor of annotating only the up-regulated or down-regulated transcripts after they are determined. Tools that the user is given in this tutorial include ones for manipulation of data in tab-delimited text files (Section O), or for renaming contigs (Section J). These tools become important to the researcher using any set of RNA-Seq tools, because they provide the ability to sort through results on a large scale.
2. For this tutorial, DESeq was used for the statistical evaluation of differential expression, but EdgeR would have been an equally good choice for this type of analysis. DESeq was presented here because EdgeR is presented elsewhere.
3. Overall, the method presented here is appropriate for studying differential expression for an organism with a previously unsequenced genome. The researcher may eventually have genome sequence data, and at that point, more annotation could be added to the results produced by this method.

REFERENCE:

Sample data: **Nicholas M. Teets, Justin T. Peyton, Herve Colinet, David Renault, Joanna L. Kelley, Yuta Kawarasaki, Richard E. Lee, Jr, David L. Denlinger, Proc Natl Acad Sci U S A. 2012 December 11; 109(50): 20744--20749.**