APBC 2016 – ENCODE Short Course

How To: Run the ENCODE RNA-seq analysis pipeline on DNAnexus

Overview: In this exercise, we will run the ENCODE Uniform Processing Long RNA-seq Pipeline on the primary data from an ENCODE RNA-seq experiment on a stomach tissue sample.

The ENCODE Portal page for the experiment is here: (https://www.encodeproject.org/experiments/ENCSR000AFI/)

The pipeline was specified by the ENCODE RNA Working Group and implemented at the ENCODE Data Coordinating Center (DCC). Today we will run the pipeline on the DNAnexus cloud platform.

The ENCODE pipeline code is open-source and lives on github at: https://github.com/ENCODE-DCC/long-rna-seq-pipeline

Summary of Steps: Here is a high-level summary of what you will learn to do in this exercise.

• Find the ENCODE Uniform Processing Pipeline project on DNAnexus.
• Copy the pipeline software and files from that project to a new project in your account.
• Complete the specification of inputs to the workflow.
• Run the pipeline workflow on the cloud.
• Monitor the run’s progress.
• Visualize the output.

Step-by-step:

1) You will need to create an account on the DNAnexus website www.dnanexus.com. Log in to your DNAnexus account.

2) Choose “ENCODE Uniform Processing Pipeline” from the “Featured Projects” list on the left hand side of the page. You can also click on “Featured” to the right of “Projects”
3) Select the box by “long-RNA-seq” and, if you have not already copied it during another exercise, the box by “Reference Files” to select those folders.

4) Select “Copy” to copy a version of these files to your account.

5) This will bring up a pop-up window that will allow you to save these files. For this exercise, select the “New Project” button, which is at the top left corner. Making a new project will help organize all the files used in the demonstration.

6) Enter a name for your new project when prompted. Select the green “plus” sign to add when name is finished.

7) Select “Copy to this folder” to add the ENCODE Uniform Processing Pipeline” files to the new project.

8) When finished, the following pop-up window will appear:
9) Select the arrow at the top left of the browser window to return to the home page.

10) Select your project from the project management list.

11) To open the “long-RNA-seq” folder, select the “long-RNA-seq” text.

12) Select “ENCODE RNA-seq (Long) Pipeline – 1 replicate (paired-end)”

13) This will bring up the workflow for the ENCODE RNA-seq (Long) Pipeline using 1 paired end replicate. You will now populate the left side of the page (the orange boxes are input files) with the appropriate files. All the blue and green (outputs) boxes are auto-filled by the pipeline. The black boxes represent the various processing steps.
14) If you hover over the input boxes, it will prompt you as to which file is required.

15) Selecting the input box will open a new window that will list possible options. There is a query window in the upper right corner of the pop-up window that can be used to filter the files by reference genome, sex, or name.

16) Select “ENCFF646CCF_1-chr21hemi.fastq.gz” – this will now complete the first paired end input.
17) Click on the Read2 input box and select “ENCFF646CCF_2-chr21hemi.fastq.gz” – that completes the reads input.

18) Add the appropriate genome reference files for each analysis step.
   For TopHat, use Reference Files/hg19/rna-seq/hg19_male_v19_ERCC_tophatIndex.tgz
   For bam to stranded signals, use Reference Files/hg19/male.hg19.chrom.sizes
   For Star, use Reference Files/hg19/rna-seq/hg19_male_v19_ERCC_starIndex.tgz
   For RSEM, use Reference Files/hg19/rna-seq/hg19_male_v19_ERCC_rsemIndex.tgz

19) Specify the location of where these output files will be stored by selecting the “Set output folder button.” This will open a pop-up window where the folder can be selected. You may wish to create a new folder for the output.

20) When all input files are correctly selected and all analysis steps have been configured, the text within the black box will appear green and the “Run as Analysis” box in the upper right corner will become available. Click “Run as Analysis” to start the analysis.

21) Starting the analysis will bring up the “Monitor” tab which will display the details of the pipeline steps as they run.
22) If necessary, the Terminate button can be used to cancel the pipeline. Otherwise, when completed the status will change.

23) The output files can be found in the output folder you specified above.

24) To visualize as a custom track at the UCSC Genome Browser, select the output file you are interested in utilizing. Right click and select “Download.”

25) A new window will pop up. Select “Get URL.”

26) This will cause the window to change and a URL to the file will appear. Right click on this URL and select “Copy Link Address.”
27) Go to [http://genome.ucsc.edu/](http://genome.ucsc.edu/) and select “My Data” from the top options bar.

28) Select “Custom Tracks” from the options menu.

29) Paste the URL link into the text window. Be sure the reference genome is correct for this file (GrCH37 for this demo). Hit “Submit” when finished.

30) This will bring up the “Manage Custom Tracks” page. Select “Go to Genome Browser” to visualize this track.

31) Scroll down and your custom track will appear on the browser page. Change visualization settings as wanted.
32) Your custom track is now visualized in the genome browser at UCSC.
Other DNAnexus Tools:

To load data once you are in your own project

1) Start a “New Project” or find your own project in the DNAnexus homepage.

2) If new, name project in the upper left corner.

3) Select “Add Data” to select the files you want to use for analysis to your project.

4) When the “Add Data to Project” window pops up, select “From another DNAnexus project.”

5) Scroll down and select “ENCODE Universal Processing Pipeline” project to access the data.

6) Choose “Add Data” to select these files.
7) When these files are uploaded, the following window will pop up.

8) These files and associated applets will now appear in the Manage tab of your browser.

To import a fastq file directly from the ENCODE portal to DNAnexus

1) Go to the ENCODE portal (encodeproject.org) and find the fastq file you are interested in using. Right click on this file and select “Copy Link Address.”

2) In the manage tab, under “Add Data” select the “From a Server” option and paste the URL into the box. Select “Add Data” and the file will upload.
To share project with another user

1) In order to share your project, select the blue “Share” button at the upper right corner of the browser page.

2) This will bring up a pop-up window where you can add user names and select permissions to allow collaborators access to view, edit, or contribute to your projects.