APBC 2016 – ENCODE Short Course

**How To: Run the ENCODE histone mark ChIP-seq analysis pipeline on DNAnexus**

**Overview:** In this exercise, we will run the ENCODE Uniform Processing ChIP-seq Pipeline on a small test dataset subsampled from a mouse histone H3K9ac ChIP-seq experiment.

The ENCODE Portal page for the experiment is here: (https://www.encodeproject.org/experiments/ENCSR087PLZ/)

The pipeline was specified by the ENCODE Analysis 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/chip-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 boxes by “histone-chip-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 this 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 histone-chip-seq folder, click the “histone-chip-seq” text.

12) Select the “ENCODE histone ChIP” workflow.

13) This will open the workflow for the ENCODE histone ChIP pipeline. You will now populate the left side of the page (the orange boxes are input files) with the appropriate input files. All the blue and green (outputs) boxes are auto-filled by the pipeline. The black boxes represent the various processing steps.
14) Select the “reads1” input box for the “Map Rep1” stage (the first step in the workflow) and a new window opens where you will navigate to the input files. Expand the “histone-chip-seq” and then the “test_data” folders to see the list of data files.

15) Select “R1.fq.gz”. You have now specified the input fastq for replicate 1.

16) Repeat the process to populate the reads1 inputs for Rep2, Control1 and Control2. The Replicate 2 input is R2.fq.gz. The control inputs are C1.fq.gz and C2.fq.gz. Since the data for
this experiment are produced by single-end sequencing, there are no inputs for “reads2”. After you have populated all the “reads1” inputs, your workflow should look like this:

17) Specify the appropriate genome reference file for the “Map Rep1” step. Because all the mapping steps map to the same reference, you only need to specify the reference once. Select the input box for the reference genome, and navigate to the “Reference Files” … “mm10” folder. In that folder, select the male.mm10.tar.gz reference file, which contains the mm10 reference and supporting index files for mapping.

18) The last input to be specified is the chrom.sizes file used to validate the coordinates of the called peaks. Specify that input for the “ENCODE Peaks” stage. Scroll the workflow page
down until you see the “ENCODE Peaks” stage, click the “…” button to expand the list of inputs, and select the “chrom.sizes” input box.

19) Expand the “Reference Files” ... “mm10” folders and select “male.mm10.chrom.sizes”.

20) 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 to “Done”.

23) The output of the mapping stages can be found in the “encode_bwa” folder and the output of the signal-generation and peak-calling stages can be found in the “encode_macs2” folder.
24) In a production environment, you will develop procedures or scripts to visualize and archive the results of multiple pipeline runs. But temporary URL’s can be generated for all outputs and used to quickly visualize some of the pipeline results. For this example, let’s look at the pooled signal track and the final, replicated peak set. In the “encode_macs2” folder, select the following output files:

**Pooled fold-over-control signal:**
R1.raw.srtfilt.nodup.srt.SE-R2.raw.srtfilt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw

**Replicated peak set:**
R1.raw.srtfilt.nodup.srt.SE-
R2.raw.srtfilt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb

After selecting these two files, click the “Download” button.

25) A new window will pop up. Select “Get bulk URLs” and copy the list of URLs. These URL’s will link to your output files and will remain active for 24 hours.

26) In this example you will use the UCSC Genome browser to visualize the results you just calculated as “custom tracks”. In a new web browser window or tab, go to [http://genome.ucsc.edu/](http://genome.ucsc.edu/) and select “My Data” from the top options bar.
27) Select “Custom Tracks” from the options menu.

28) Paste the URLs you copied above into the text window. Be sure the reference genome is correct for this file (mouse mm10 for this demo). **Tip: The UCSC Genome Browser is sensitive to white-space at the end of URL’s. If there are spaces after the URL’s you’ve pasted, delete them and make sure each URL is on its own line.** Hit “Submit” when finished.

29) This will bring up the “Manage Custom Tracks” page. Select “go” to visualize the tracks.

30) Because the raw data were subsampled to only chromosome 19, enter a position on that chromosome. For example, chr19:4,769,592-4,909,161

Set the signal track to display in “full” mode. Do you see strong signal for H3K9ac (the target for this experiment)? The black blocks in the replicated peaks track are the peaks that passed a stringent thresholding requiring the peaks to be observed in both replicates.
H3K9ac is a mark associated with transcriptional activation. Can you see that the peaks are called around the transcription start sites of several genes in this region?

Congratulations! You have replicated an ENCODE analysis starting with primary data. You can repeat this process on your own data, and be assured that your results will be directly comparable to all the experiments the ENCODE DCC has analyzed.
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.

![Projects](image)

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

![ENCOD_Demo](image)

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

![ENCOD_Demo](image)

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

![Add Data to Project](image)

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

<table>
<thead>
<tr>
<th>Broad Inst Viral NGS</th>
<th>Viewer</th>
<th>1</th>
<th>0.11 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCOD Uniform Processing Pipelines</td>
<td>Viewer</td>
<td>13</td>
<td>349.28 GB</td>
</tr>
</tbody>
</table>

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.