

How To: Run the ENCODE Transcription Factor (TF) 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 containing reads from only chromosome 21 from a human ZBED1 ChIP-seq experiment. The biosample was the K562 CML cell line.

The ENCODE Portal page for the experiment is here:

<https://www.encodeproject.org/experiments/ENCSR286PCG/>

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>

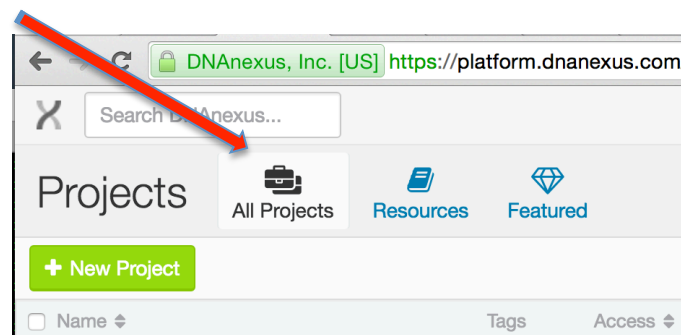
Summary of Steps: Here is a high-level summary of what you will learn to do in this exercise.

- **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.

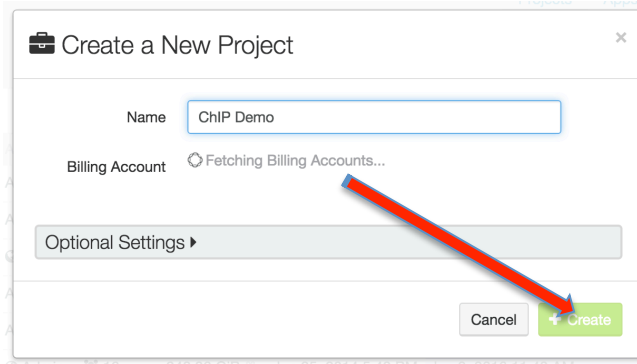
Skip ahead to step 9 if you have already copied the ChIP-seq pipeline files from the ENCODE Universal Pipelines project.

Step-by-step:

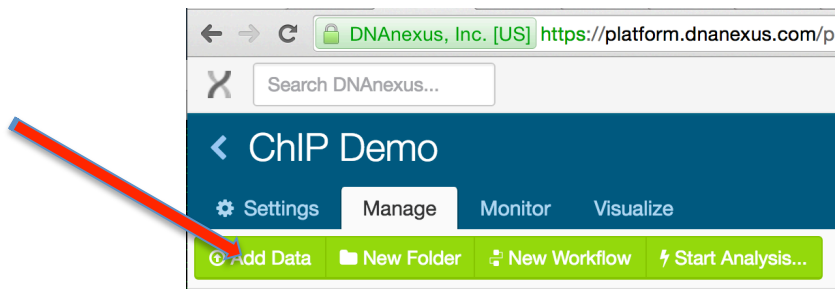
- 1) You will need to create an account on the DNAnexus website www.dnanexus.com. Log in to your DNAnexus account.
- 2) Once logged into your DNAnexus account, create a new project. Select "All Projects" and then click "New Project".



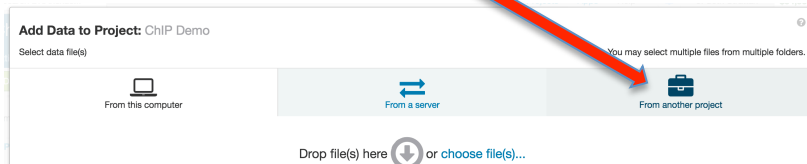
- 3) Give your project a new name and click "Create".



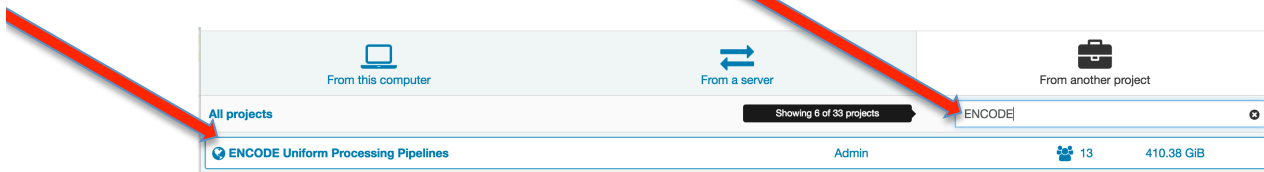
4) Select "Add Data" ...



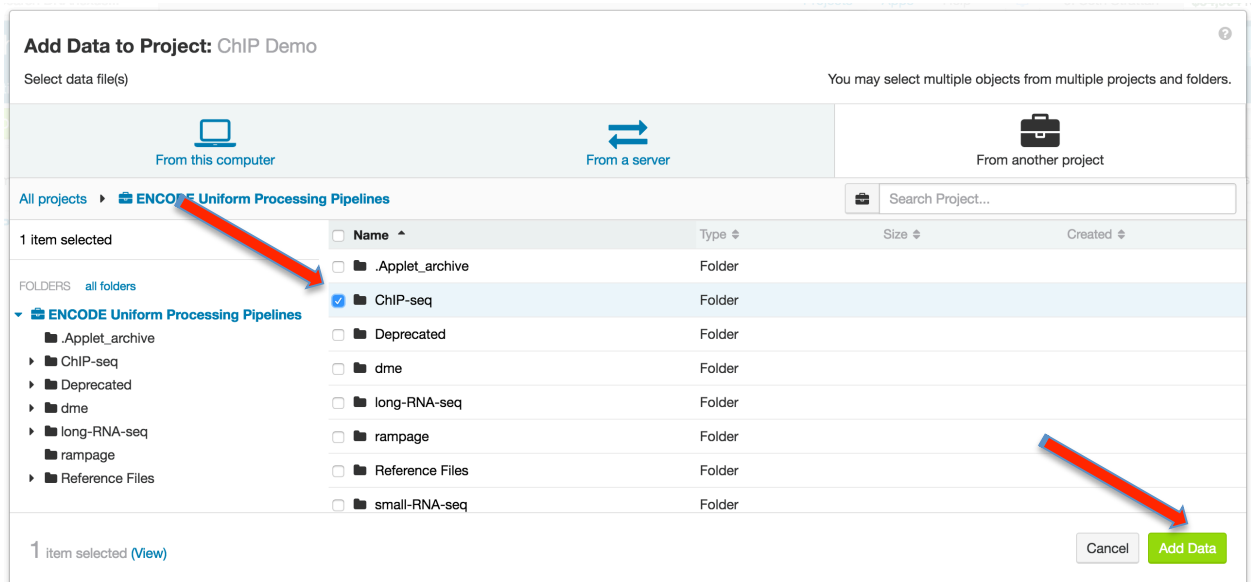
5) ... select "From another project" ...



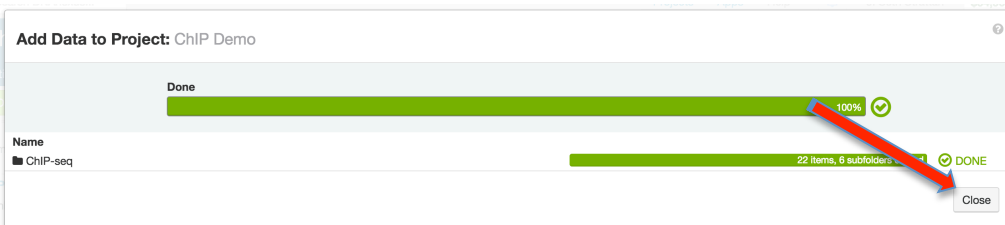
6) Type "ENCODE" in the search box and select "ENCODE Uniform Processing Pipelines"



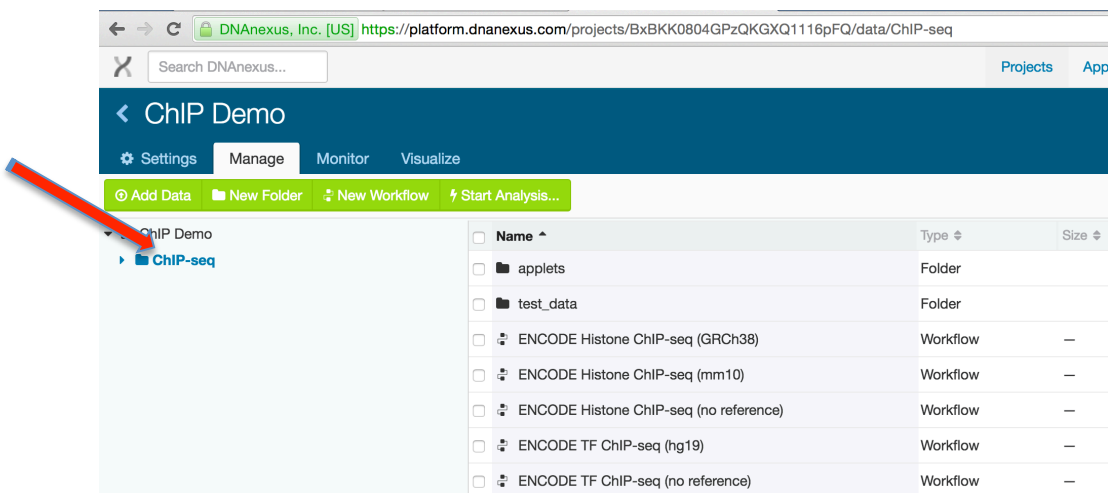
7) Click the box next to “ChIP-seq” and select “Add Data”.



8) When finished, the following pop-up window should appear. Click “Close”.

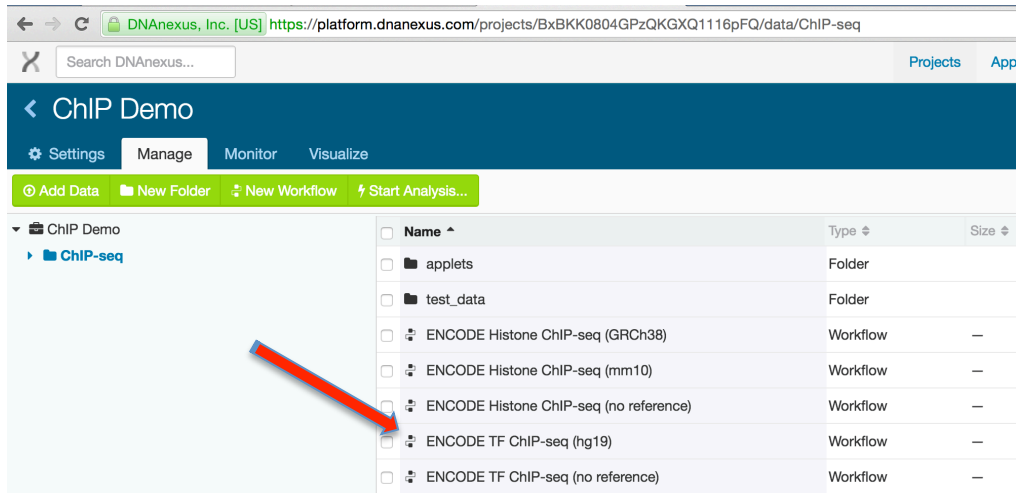


9) To open the ChIP-seq folder, click the “ChIP-seq” text. You should see the files copied to your project.

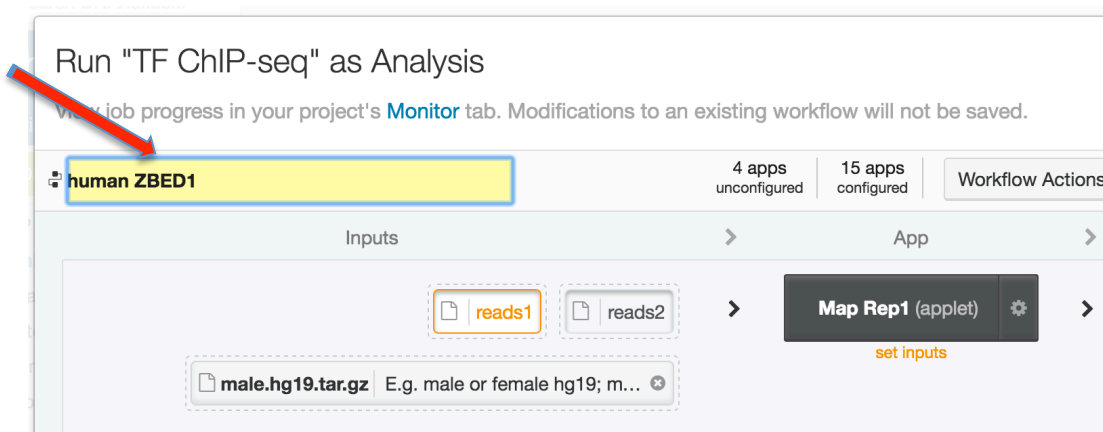


10) The example data in this exercise is from a human histone ChIP experiment, which we will map to the human hg19 assembly. Click on the “ENCODE TF ChIP-seq (hg19)” workflow to

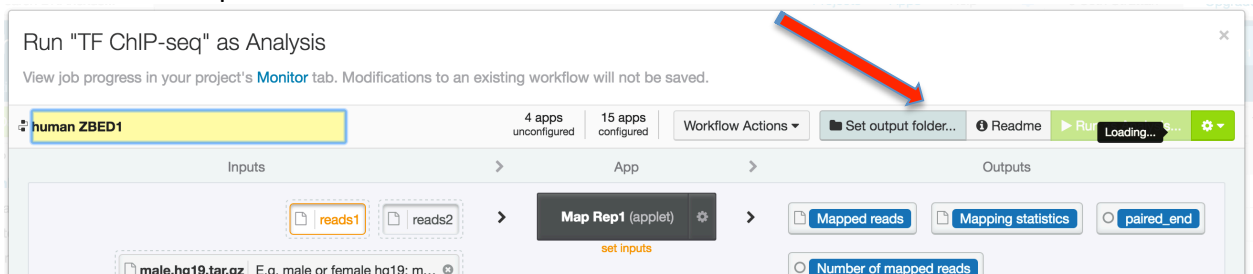
open it.



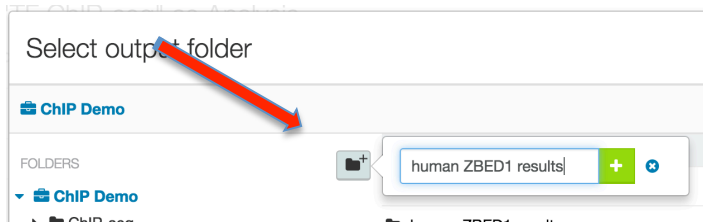
11) This window represents an “Analysis”, which is an instantiation of the TF ChIP-seq workflow. Give the analysis a name, like “human ZBED1”



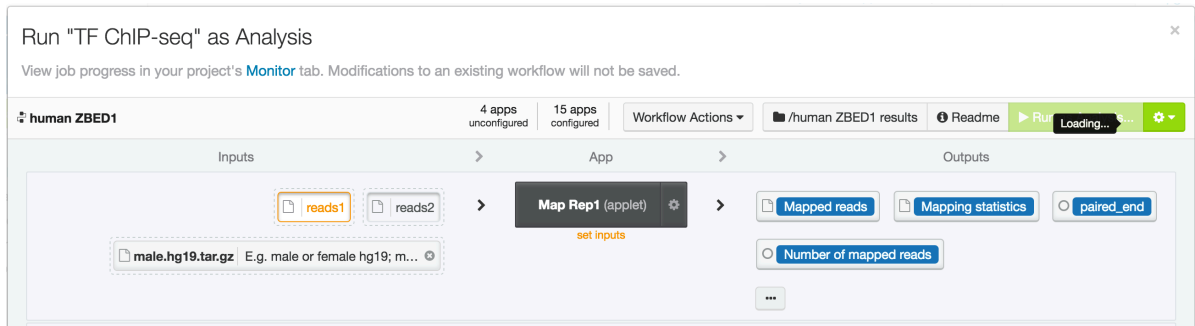
12) Click on “Set output folder ...”



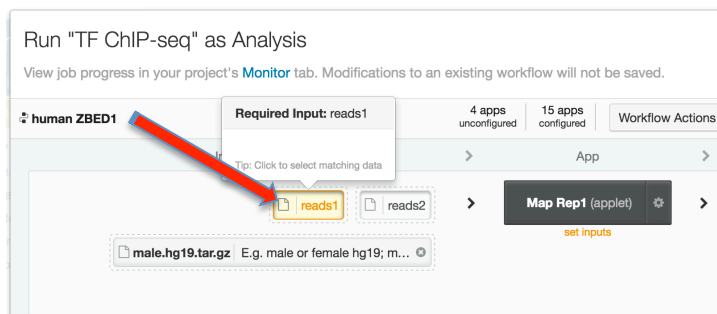
13) Click on the new folder button to create a new folder and name it something like “human ZBED1 results”.



- 14) Now you should have named your analysis and specified an output folder for the results. Your workflow window should look like this:

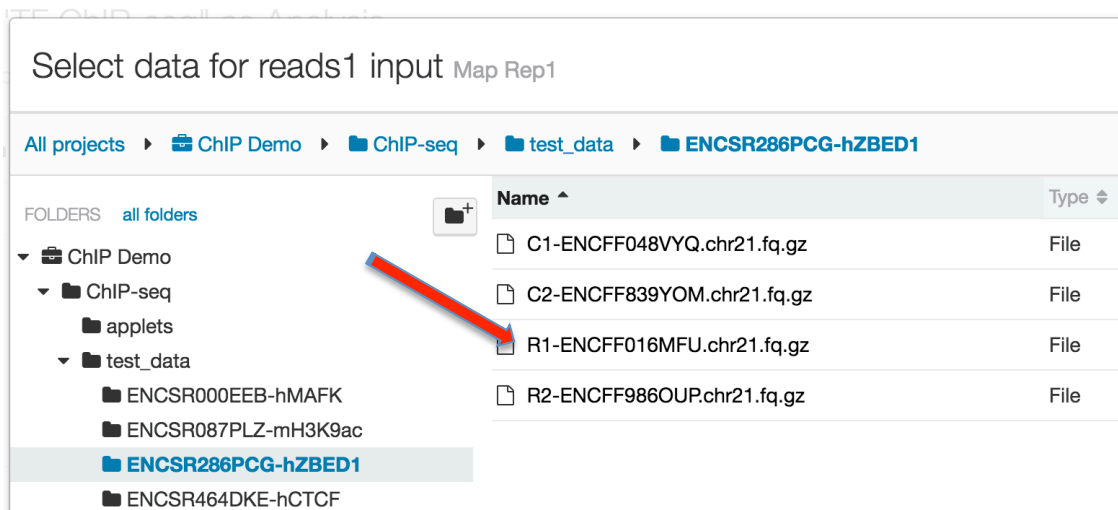


- 15) Select the “reads1” input box for the “Map Rep1” stage (the first step in the workflow). Note that the data in this example are from single-end sequencing, so all the “reads2” inputs will be left blank. In a paired-end experiment the second fastq of the paired reads for each replicate would go in “reads2”.

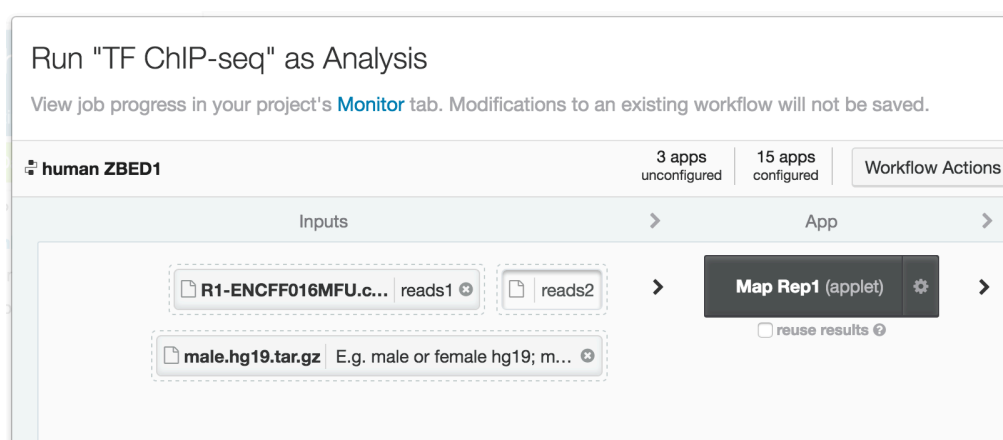


- 16) A new window opens where you will navigate to the input files. Expand the “Chip-seq” and then the “test_data” and then the “ENCSR286PCG-hZBED1” folders to see the list of data files. **Make sure to select the subfolder ENCSR286PCG-hZBED1 to limit the display to just the data for this experiment.**

The ENCSR286PCG-hZBED1 folder contains only reads for chromosome 21 from this experiment, for faster processing.



17) Select “R1-ENCFF016MFU.chr21.fq.gz”. You have now specified the input fastq for replicate 1 of this experiment.

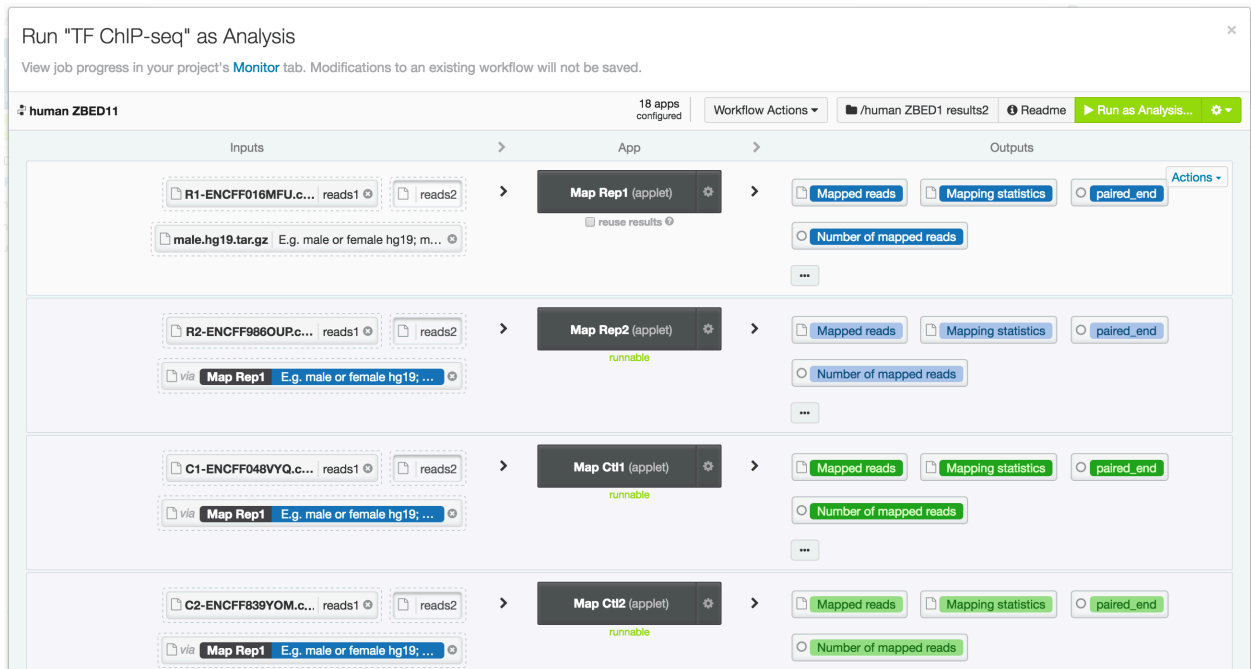


18) Repeat the process to populate the reads1 inputs for the “Map Rep2” step, the “Map Ctl1” step, and the “Map Ctl2” step. The Rep2 input starts with “R2”. The control inputs start with “C1” and “C2”, respectively. Since the data for this experiment are produced by single-end sequencing, there are no inputs for “reads2”. **Note: Make sure you choose the inputs that go with this experiment. They are all in the subfolder ENCSR286PCG-hZBED1.**

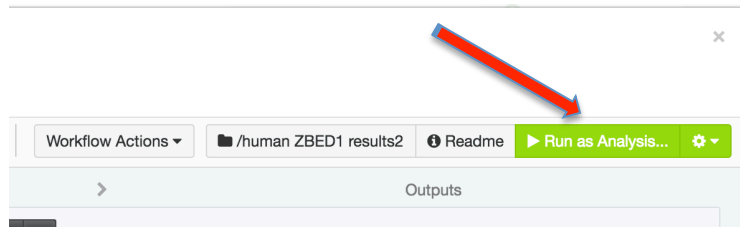
Here is a summary of the input files for this experiment:

Map Rep1: *R1-ENCFF016MFU.chr21.fq.gz*
 Map Rep2: *R2-ENCFF986OUP.chr21.fq.gz*
 Map Ctl1: *C1-ENCFF048VYQ.chr21.fq.gz*
 Map Ctl2: *C2-ENCFF839YOM.chr21.fq.gz*

After you have populated all the “reads1” inputs, your workflow should look like this:



19) All of the other inputs, including the indexed hg19 genome reference, have been pre-filled in this workflow. All input requirements are satisfied, so click “Run as Analysis” to start the analysis.



20) Starting the analysis will bring up the “Monitor” tab which will display the details of the pipeline steps as they run. Click on the “+” box to see all the analysis subjobs. If necessary, the “Terminate” button can be used to cancel the analysis.

DNAnexus, Inc. [US] https://platform.dnanexus.com/projects/BxBK0804GPzQKXQ1116pFQ/monitor

Search DNAnexus... Projects Apps Help- J Seth Strattan - Upgrade Now

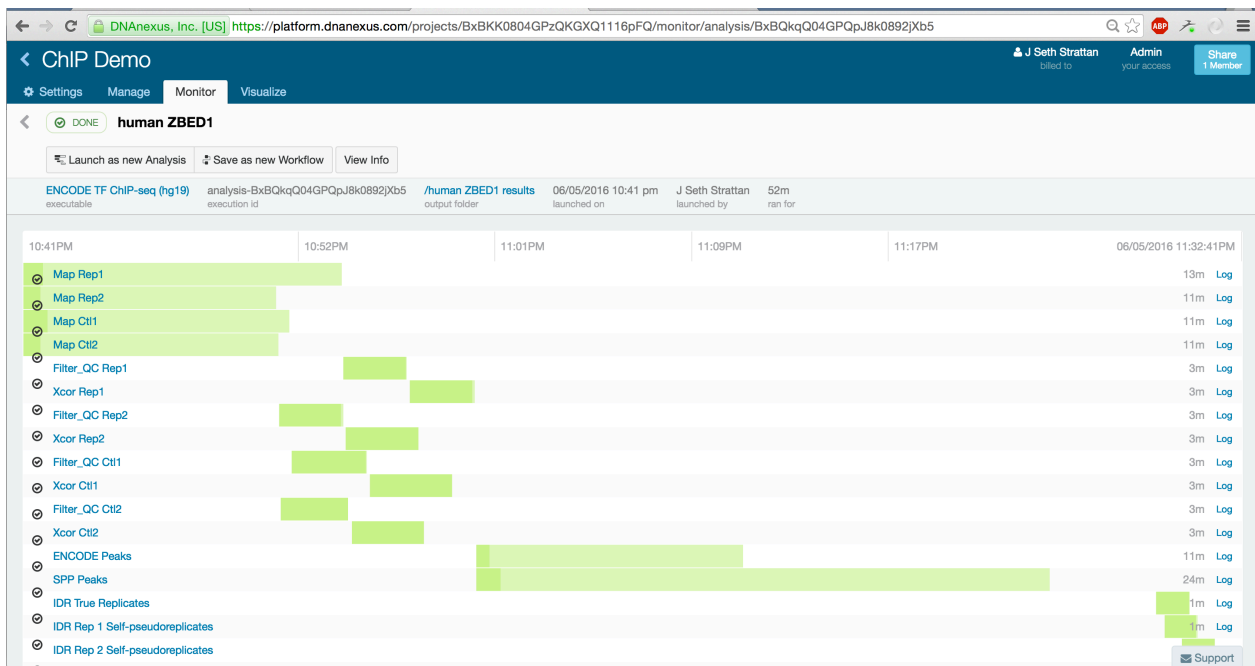
ChIP Demo J Seth Strattan Admin Share

Settings Manage **Monitor** Visualize

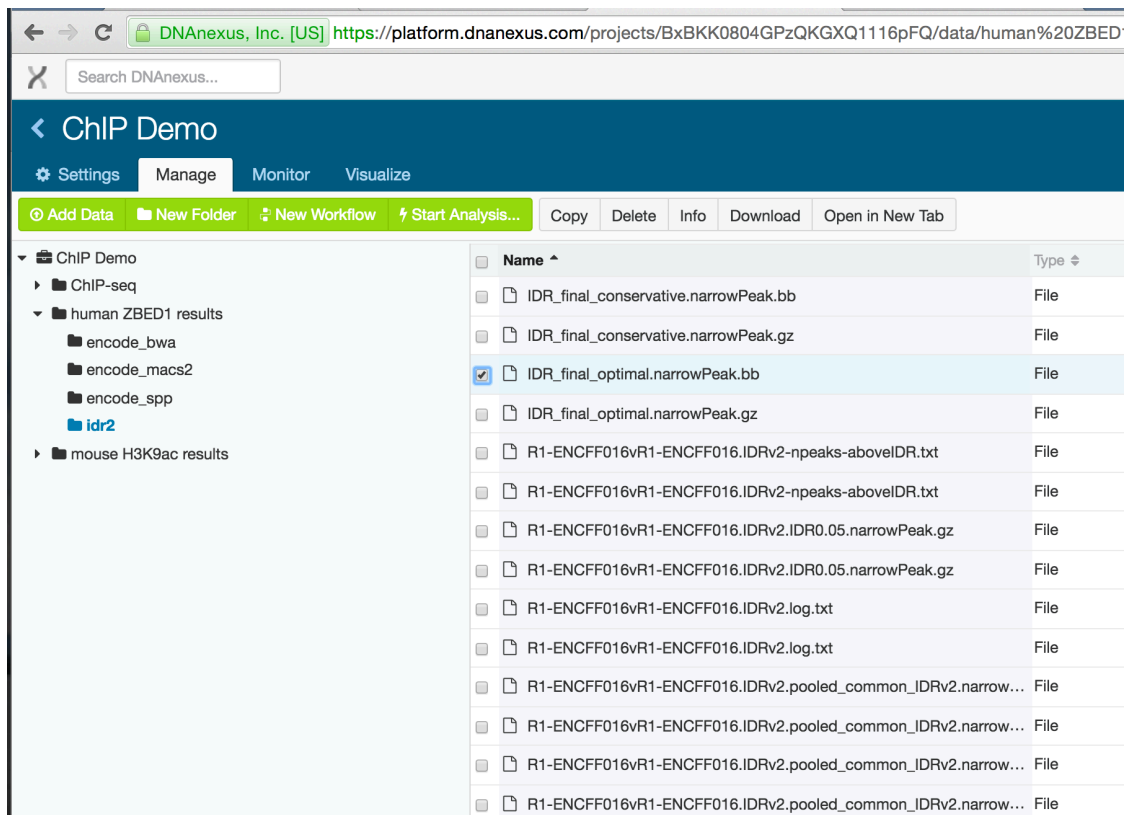
Bookmarks SEARCH SCOPE Root executions only STATE Any 2 recent jobs NAME Any ID Any CREATED Any Filters

Status	Name	Executable	Launched by	Started	Duration
Done	human ZBED1	ENCODE TF ChIP-seq (hg19)	J Seth Strattan	06/05/2016 10:41 pm	52m
Done	Map Rep1	ENCODE BWA	J Seth Strattan	06/05/2016 10:41 pm	13m
Done	Map Rep2	ENCODE BWA	J Seth Strattan	06/05/2016 10:41 pm	11m
Done	Map Ctl1	ENCODE BWA	J Seth Strattan	06/05/2016 10:41 pm	11m
Done	Map Ctl2	ENCODE BWA	J Seth Strattan	06/05/2016 10:41 pm	11m
Done	Filter_QC Rep1	ENCODE TF ChIP-seq Mapped Read Filter and QC	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Xcor Rep1	ENCODE TF ChIP-seq Cross-Correlation Analysis	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Filter_QC Rep2	ENCODE TF ChIP-seq Mapped Read Filter and QC	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Xcor Rep2	ENCODE TF ChIP-seq Cross-Correlation Analysis	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Filter_QC Ctl1	ENCODE TF ChIP-seq Mapped Read Filter and QC	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Xcor Ctl1	ENCODE TF ChIP-seq Cross-Correlation Analysis	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Filter_QC Ctl2	ENCODE TF ChIP-seq Mapped Read Filter and QC	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	Xcor Ctl2	ENCODE TF ChIP-seq Cross-Correlation Analysis	J Seth Strattan	06/05/2016 10:41 pm	3m
Done	ENCODE Peaks	ENCODE Peaks with MACS2	J Seth Strattan	06/05/2016 10:41 pm	11m
Done	SPP Peaks	ENCODE Peaks with SPP	J Seth Strattan	06/05/2016 10:41 pm	24m
Done	IDR True Replicates	IDR	J Seth Strattan	06/05/2016 10:41 pm	1m
Done	IDR Rep 1 Self-pseudoreplicates	IDR2	J Seth Strattan	06/05/2016 10:41 pm	1m
Done	IDR Rep 2 Self-pseudoreplicates	IDR2	J Seth Strattan	06/05/2016 10:41 pm	1m
Done	IDR Pooled Pseudoreplicates	IDR2	J Seth Strattan	06/05/2016 10:41 pm	1m
Done	Final IDR peak calls	IDR for ChIP-seq	J Seth Strattan	06/05/2016 10:41 pm	< 1m

21) Click on the analysis name (here we’ve named it “human ZBED1”) to watch the progress of each stage.



22) Within the output folder you specified above, the results of the mapping stages can be found in the “encode_bwa” subfolder, the output of the signal-generation stage can be found in the “encode_macs2” subfolder, and the final peak calls are in the “idr2” subfolder.



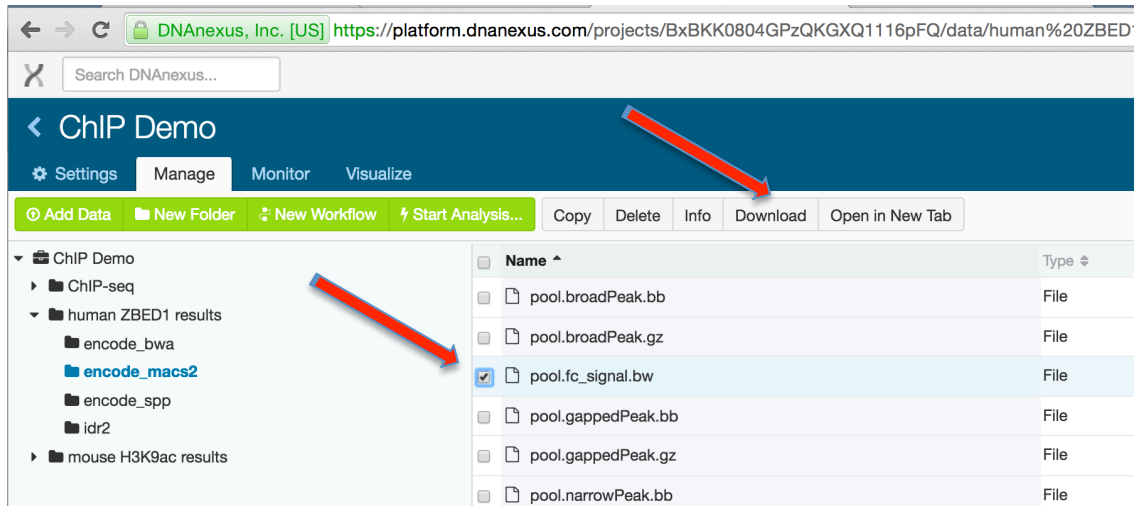
23) In a production environment, you will develop procedures or scripts to visualize and archive the results of multiple pipeline runs. But temporary URL’s can be generated for all outputs and used to quickly visualize some of the pipeline results. For this example, let’s look at the

pooled signal track and the final, replicated peak set. In the “encode_macs2” folder, select the following output file:

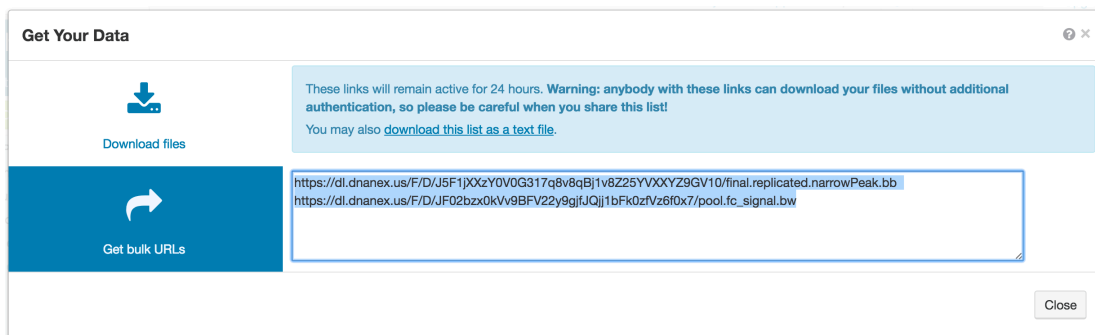
Pooled fold-over-control signal:

pool.fc_signal.bw

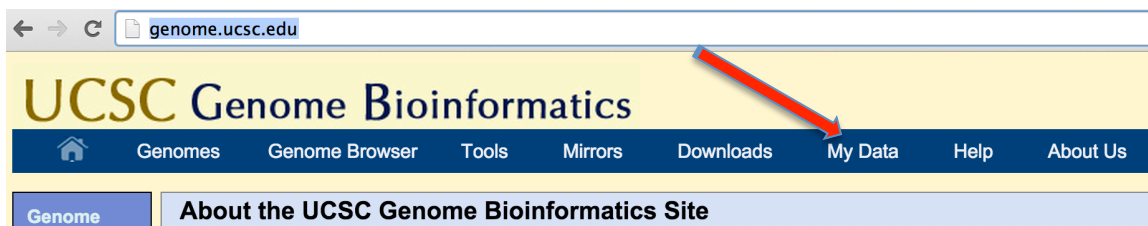
After selecting the file, click the “Download” button.



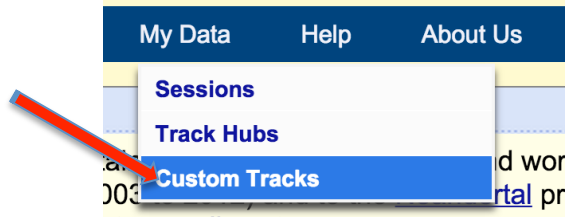
24) A new window will pop up. Select “Get bulk URLs” and copy the list of URLs. These URL’s will link to your output files and will remain active for 24 hours.



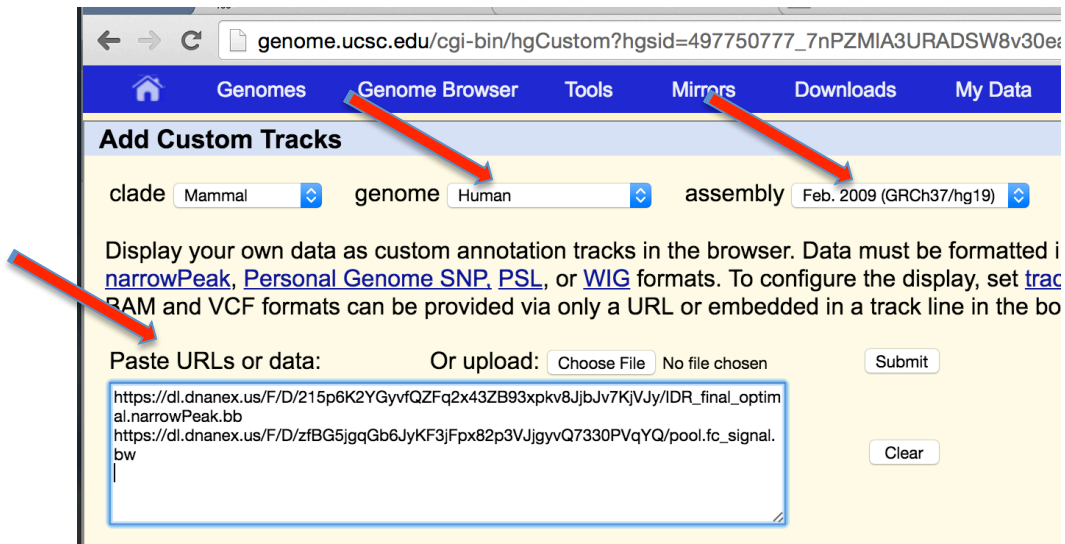
25) In this example you will use the UCSC Genome browser to visualize the results you just calculated as “custom tracks”. In a new web browser window or tab, go to <http://genome.ucsc.edu/> and select “My Data” from the top options bar.



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

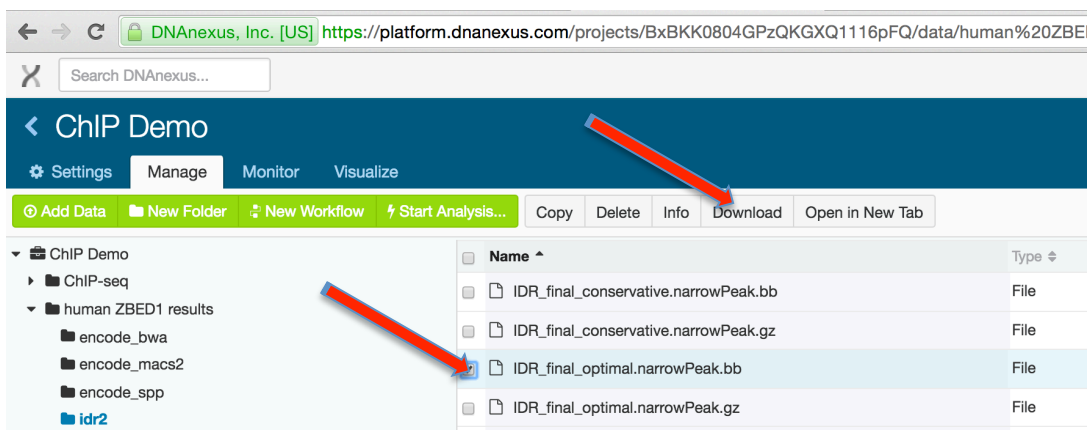


- 27) Paste the URLs you copied above into the text window. Be sure the reference genome is correct for this file (human hg19 for this demo). **Tip: The UCSC Genome Browser is sensitive to white-space at the end of URL's. If there are spaces after the URL's you've pasted, delete them and make sure each URL is on its own line.** Don't click "Submit", yet. We need one more track.

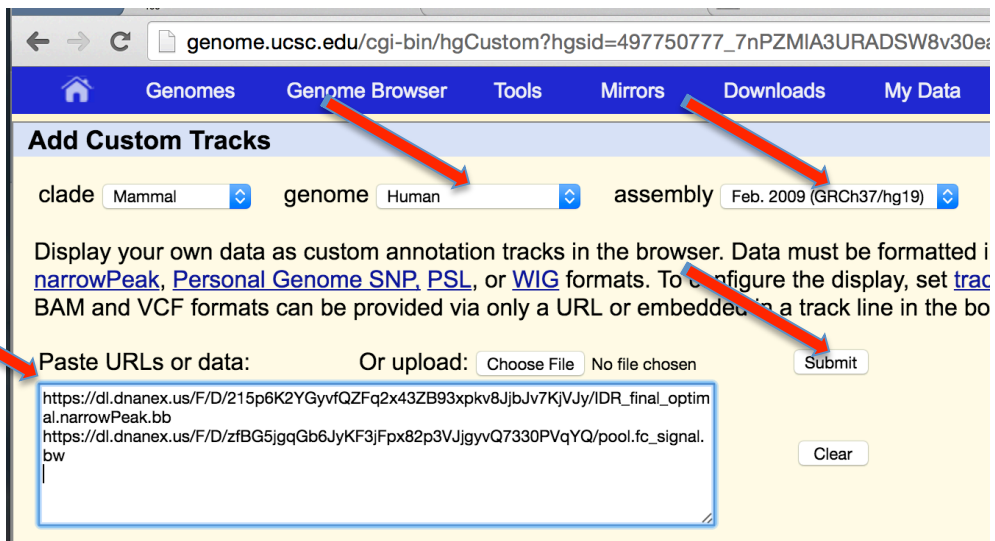


- 28) In the same way that you generated a URL for the signal track, go back and generate a URL for the final replicated peak set. It is in the subfolder "idr2" and it is called:

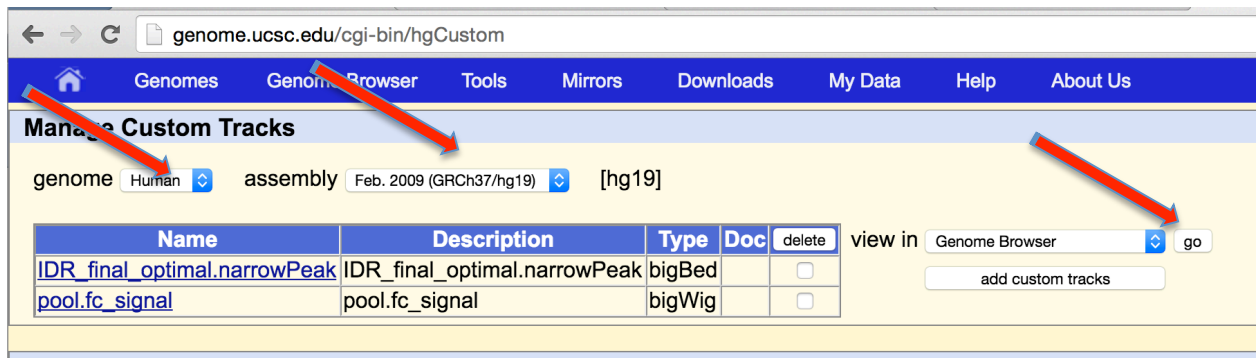
IDR_final_optimal_narrowPeak.bb



Generate and copy the URL in the same way you did above and paste it into your "Add Custom Tracks" UCSC Genome Browser window. Click submit when you have both URL's pasted onto their own lines.

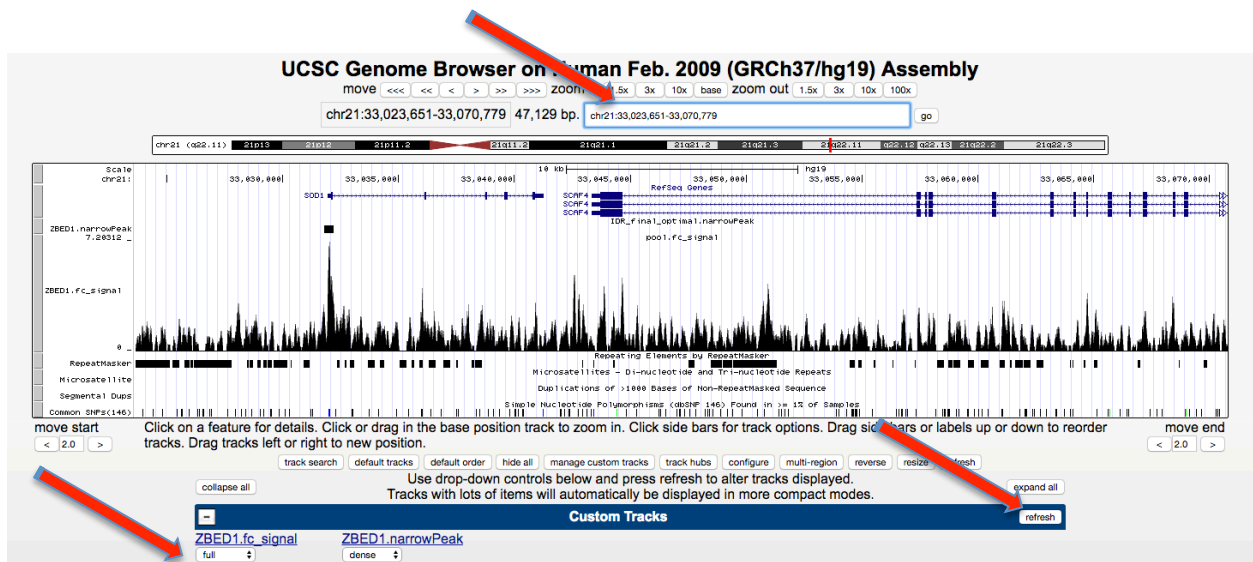


29) This will bring up the “Manage Custom Tracks” page. Double-check the assembly, and select “go” to visualize the tracks.



30) Because the raw data were subsampled to only chromosome 21, enter a position on that chromosome. For example, chr21:33,023,651-33,070,779

Set the signal track to display in “full” mode. Do you see the strong signal for ZBED1 (the target for this experiment) at the SOD1 promoter? 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.



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.

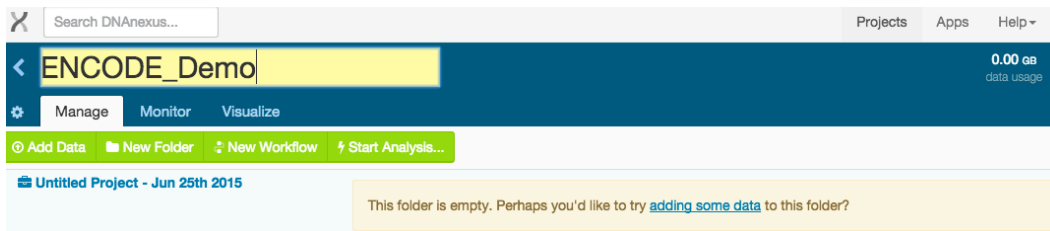
Other DNAnexus Tools:

To load data once you are in your own project

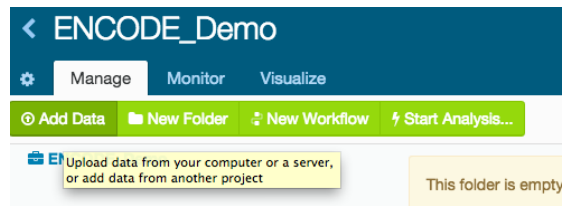
- 1) Start a “New Project” or find your own project in the DNAnexus homepage.



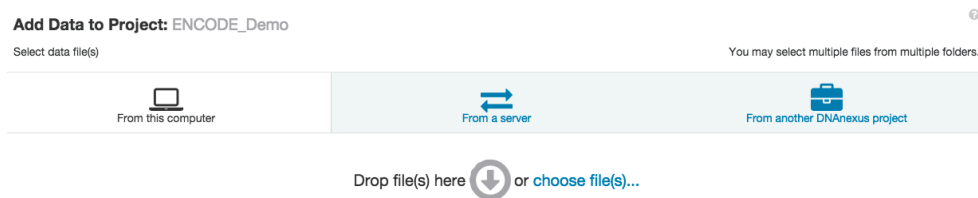
- 2) If new, name project in the upper left corner.



- 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.

Broad Inst Viral NGS	Viewer	1	0.11 GB
ENCODE Uniform Processing Pipelines	Viewer	13	349.28 GB

- 6) Choose “Add Data” to select these files.

7) When these files are uploaded, the following window will pop up.

Add Data to Project: ENCODE_Demo

Done 100%

Name	Progress	Status
long-RNA-seq	25 items, 4 subfolders copied	DONE
Reference Files	54 items, 6 subfolders copied	DONE

[Close](#)

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

ENCODE_Demo

Manage Monitor Visualize

[Add Data](#) [New Folder](#) [New Workflow](#) [Start Analysis...](#)

Name	Type	Size
long-RNA-seq	Folder	
Reference Files	Folder	
align-star-se (Fri Dec 12 01:41:16 2014)	Applet	1.16 MB
align-tophat-pe (Fri Jan 9 01:28:56 2015)	Applet	28.86 MB
align-tophat-se (Fri Dec 12 01:41:04 2014)	Applet	27.45 MB

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."

Files linked to ENCSR000AFI

Raw data

Accession	File type	Biological replicate	Technical replicate	Read length	Run type	Paired end	Mapping assembly	Lab	Date added	Validation status
ENCF001RNE Download 4.78 GB	fastq	2	1	101 nt	paired-ended	2		Thomas Gingeras, CSHL	2013-07-17	pending
ENCF001F Download 4.8 GB				101 nt	paired-ended	1		Thomas Gingeras, CSHL	2013-07-17	pending
ENCF001F Download 5.15 GB				101 nt	paired-ended	2		Thomas Gingeras, CSHL	2013-07-18	pending
ENCF001F Download				101 nt	paired-ended	1		Thomas Gingeras, CSHL	2013-07-18	pending

Right-click context menu options:

- Open Link in New Tab
- Open Link in New Window
- Open Link in Incognito Window
- Save Link As...
- Copy Link Address**
- Copy
- Search Google for 'Download'
- Print...

- 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.

Add Data to Project: ENCODE_Demo

Select data file(s)

You may add multiple URLs.

From this computer From a server From another DNAnexus project

<https://www.encodeproject.org/files/ENCF001RNE/@download/ENCF001RNE.fastq.gz>

Enter a URL...

Add Data to Project: ENCODE DEMO_June24

Done

Name

<https://www.encodeproject.org/files/ENCF001RNE/@download/ENCF001RNE.fastq.gz>

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.

Admin Private Share

your access access policy 2 Members

- 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.

Share project

Name	Access	Charges Allowed
Benjamin Hitz (hitz)	Viewer	Remove
Eurie Hong (euriehong)	Admin	\$

Add member...

Examples:
jsmith
user-jsmith

Close