

# DNAneXus

## NGS DATA ANALYSIS FROM A BIOLOGIST'S PERSPECTIVE

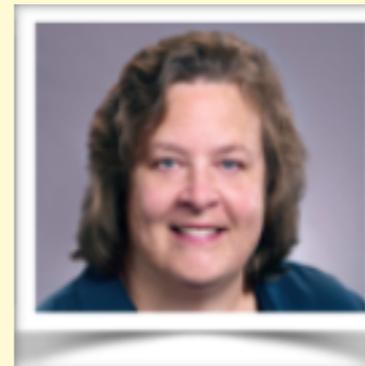
Peter C. Fitzgerald



Carl McIntosh



Amy Stonelake



Desiree Tillo



# TODAY AGENDA

- Introduction to DNAnexus
- Introduction CCR's Pilot program with DNAnexus
- DNAnexus Apps
- St. Jude Apps and data
- Highlight of CCR support resources
- Follow on classes:
  - DNAnexus Development Environment - Bioinformaticists  
Friday April 12th, 10:00-11:30 am. - NIH Bldg 37, Rm 2041/2107

# What is DNAnexus ?

**DNAnexus** is a bioinformatics company that provides a cloud-based data analysis and management platform for DNA sequence data. It was founded in early 2009 as a spin-off from Stanford University

"**DNAnexus** provides a **cloud-based platform** optimized to address the challenges of security, scalability, and collaboration, for organizations that are pursuing genomic-based approaches to health, in the clinic and in the research lab."

**DNAnexus** provides a simplified, structured and managed access two cloud-based service providers (AWS and Azure).

**AWS = Amazon Web Services**

**Azure = Microsoft Cloud Services**

Each environment virtually identical - BUT they are distinct spaces and difficult to move data and apps between the two

# DNAexus Projects

## Project-Centric World

Projects are the main unit of control and data management

Data and applications reside within a project

Sub Folders within a project are used to ease to task of data management

(A Structured project with sub-folders is essential for successful data management)

## Project Level Controls

- **Viewer** - can **view** and download data
- **Uploader** - can **upload** data, but cannot edit data or run apps
- **Contributer** - Can manage data and **run analyses** (can incur charges)
- **Administrator** - Can manage data, **membership** and run analyses (can incur charges)

# CCR DNAnexus Pilot

## **CCR-ORG**

- We have established an Organizational account
- 60% discount on standard rates
- Initial costs subsidized and managed by OSTR
- Support for use and customized development

## **Questions we hope the Pilot will answer**

- Will this resource be adopted by:  
biologists for data analysis and/or  
bioinformaticists for batch analysis and sharing results?
- Will it work for managing and sharing data on a large scale?
- Is the platform effective in disseminating software solutions?
- Is it a solution for patient data analysis (security, speed)?

# THE PROJECT

- The main Work Unit
- Can be Shared
- Often need to copy applications into the project folder

# FILE/FOLDER MANAGEMENT

- Files can have duplicate names  
*(but it can cause issues on occasion)*
- Use Folders Wisely
- File Filtering tools provide easy data navigation  
but not intuitive

# Analysis Tools

## ● Applets

Applets are lighter-weight executables that can be used as scripts for project-specific analyses or ad hoc data manipulations, proprietary analysis pipelines, or development/testing versions of apps. Unlike apps, **they reside inside your Project folder** alongside data

## ● Apps

Apps represent general-purpose tools, striving for compatibility, ease of use, and robustness. They're published in a dedicated section of the website, and typically include extensive metadata and documentation.

## ● Workflows

Workflows represent a series of executables (apps or applets) that are linked together by dependencies, e.g. one executable's outputs may be another's inputs. It is easiest to create a workflow in the web interface. **These also reside in the Project folder - pro tip for how to speed up APPs - create workflow out of APP**

# Why DNAnexus ?



St. Jude Cloud

ADVANCING CURES THROUGH DATA AND DISCOVERY

St. Jude Cloud is a data-sharing resource for the global research community. Explore unique next-generation sequencing data and analysis tools for pediatric cancer and other life-threatening diseases.

## Data



Mine one of the world's most comprehensive repositories of pediatric cancer genomics data.

[Access Data](#)

## Tools



Analyze genomics data using sophisticated computational pipelines built for speed and ease of use.

[Run Tools](#)

## Visualizations



Use our intuitive, field-tested visualization tools to explore data in a secure cloud environment.

[Visualize Results](#)

# CCR/GAU RESOURCES

- Help pages on the Web  
(<https://gau.ccr.cancer.gov/dna-nexus-pilot-program/>)
- Slack Channel for CCR\_DNANexus Pilot ([dnaxpilot.slack.com](https://dnaxpilot.slack.com))  
(help, general, development)
- Custom Built Work Flows (RNASEQ workflow, IGV\_session\_maker, ADAP, *Pausing Peak Aligner\**, *Tumor Mutation Burden\**)
- DNANexus Applications By Category Page  
(<https://dl.dnanex.us/F/D/jpyVIBVZKZjzf8IIQXfg7Xl3P8xIZ4lP7zKVygpX?inline>)
- Management of DNANexus Account, Funding and cost management

# RNA-SEQ

*Peter FitzGerald*

Head, Genome Analysis Unit

Custom Work Flow developed by  
Carl McIntosh and Peter FitzGerald (GAU)

# Salmon

—Don't count . . . quantify!

## Overview

Salmon is a tool for quantifying the expression of transcripts using RNA-seq data. Salmon uses new algorithms (specifically, coupling the concept of *quasi-mapping* with a two-phase inference procedure) to provide accurate expression estimates very quickly (i.e. *wicked-fast*) and while using little memory. Salmon performs its inference using an expressive and realistic model of RNA-seq data that takes into account experimental attributes and biases commonly observed in *real* RNA-seq data.

The mapping-based mode of Salmon runs in two phases; indexing and quantification. The indexing step is independent of the reads, and only need to be run one for a particular set of reference transcripts. The quantification step, obviously, is specific to the set of RNA-seq reads and is thus run more frequently.

Genes can have multiple transcripts (alternate splicing, alternate starts/stops).

Transcript expression is the expression of a specific transcript.

Gene expression means the overall expression of all transcripts of a gene.

(i.e. counts from a all transcripts of give gene are summed to yield a gene expression value)



HUMAN

GENCODE 29 (GRCh38.p12)

MOUSE

GENCODE M20 (GRCm38.p6)

The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation

## Statistics about the current GENCODE Release (version 30)

Total No of Genes	<b>58870</b>	Total No of Transcripts	<b>208621</b>
Protein-coding genes	19986	Protein-coding transcripts	83688
Long non-coding RNA genes	16193	- full length protein-coding	57687
Small non-coding RNA genes	7576	- partial length protein-coding	26001
Pseudogenes	14706	Nonsense mediated decay transcripts	15550
- processed pseudogenes	10663	Long non-coding RNA loci transcripts	30369
- unprocessed pseudogenes	3525		
- unitary pseudogenes	221	Total No of distinct translations	61870
- polymorphic pseudogenes	42	Genes that have more than one distinct translations	13709
- pseudogenes	18	Total No of Transcripts	208621
Immunoglobulin/T-cell receptor gene segments		Protein-coding transcripts	83688
- protein coding segments	408	- full length protein-coding	57687
- pseudogenes	237	- partial length protein-coding	26001

# THE SAMPLES

RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues

[Proteomics. Tissue-based map of the human proteome](#). Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigartyo CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. *Science* 347(6220) (2015), [PMID:5613900](#)

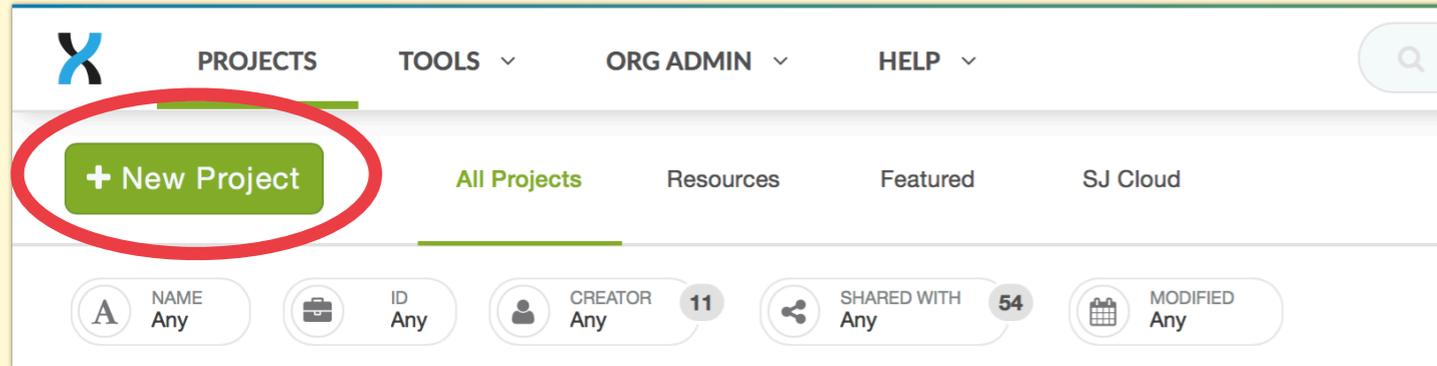
- Paired end sequences
- Two conditions - 3 replicates each  
Brain x3 vs Muscle x3

# RNASEQ

## The End Point

<file:///Users/fitzgepe/Downloads/Jupyter%20Notebook%20Viewer.webarchive>

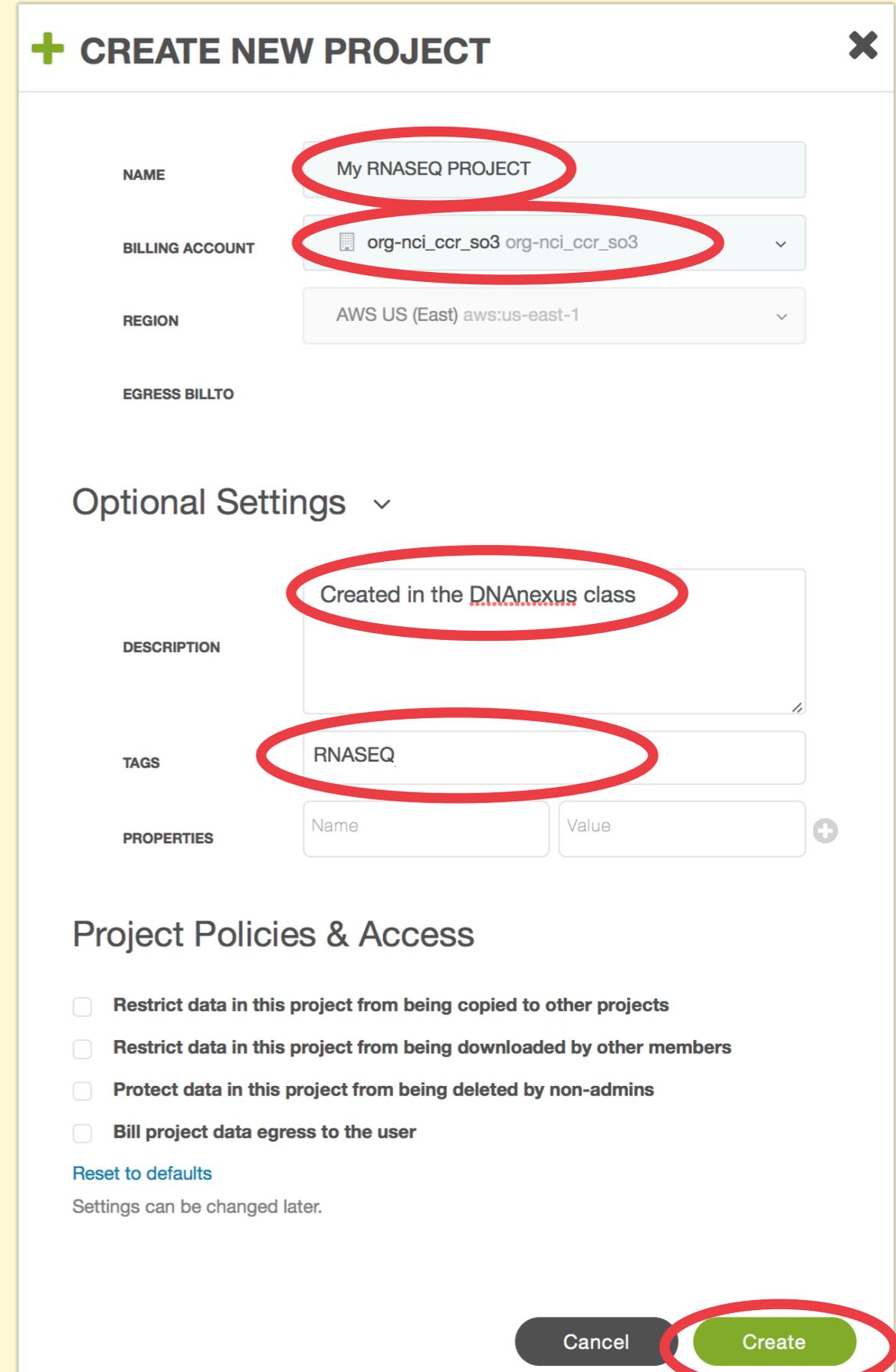
# Step I: Create an new project



Select the **New Project** button

In the pop-up dialogue box fill in the following info:

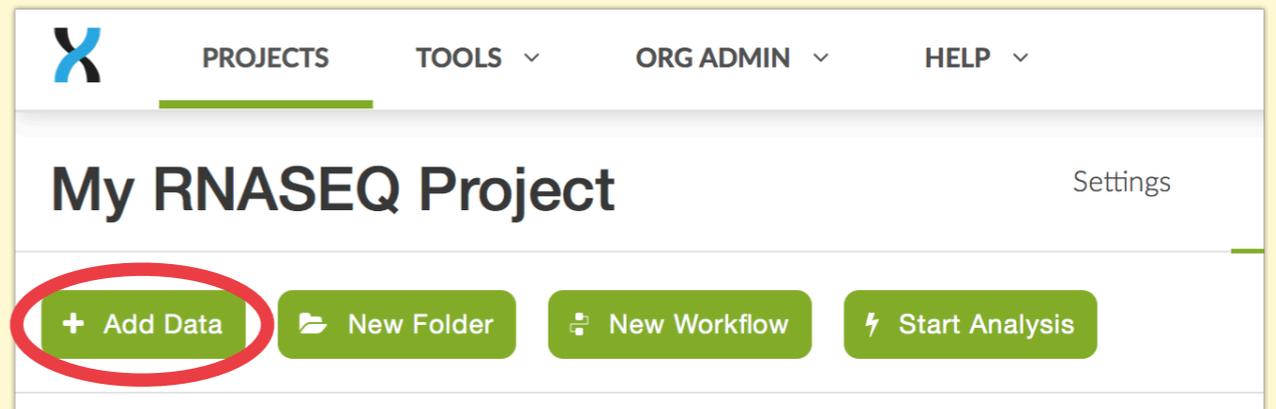
- Give the Project a meaningful **name**
- Select the **billing** - personal or org
- The **region** should be AWS-east
- Provide a **description**
- Optionally use **TAG** for later filtering
- Hit the **Create** button

A screenshot of the 'CREATE NEW PROJECT' dialog box. The fields are filled out as follows: 'NAME' is 'My RNASEQ PROJECT', 'BILLING ACCOUNT' is 'org-nci\_ccr\_so3 org-nci\_ccr\_so3', 'REGION' is 'AWS US (East) aws:us-east-1', 'DESCRIPTION' is 'Created in the DNAnexus class', and 'TAGS' is 'RNASEQ'. The 'CREATE' button at the bottom right is circled in red. Below the 'Optional Settings' section, there are four unchecked checkboxes under 'Project Policies & Access': 'Restrict data in this project from being copied to other projects', 'Restrict data in this project from being downloaded by other members', 'Protect data in this project from being deleted by non-admins', and 'Bill project data egress to the user'. There is also a 'Reset to defaults' link and a note 'Settings can be changed later.'

# Step 2: Copy some data and applets in to the project

We're using the common project *CCR\_Resources* to get the data and applets

Select the **Add Data** button



Select the following:

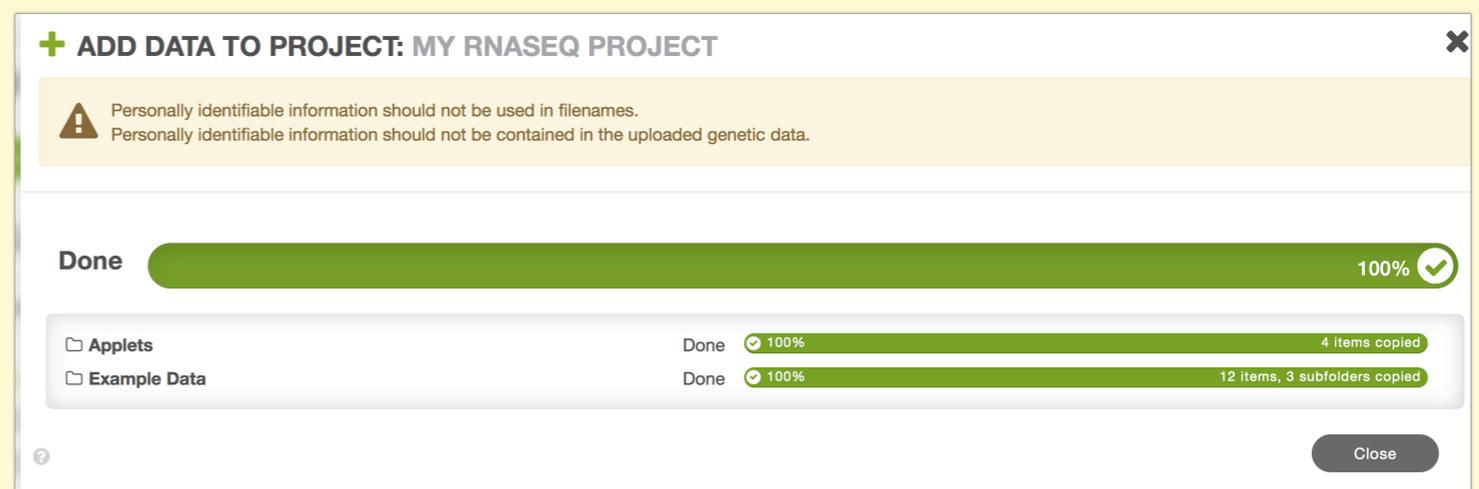
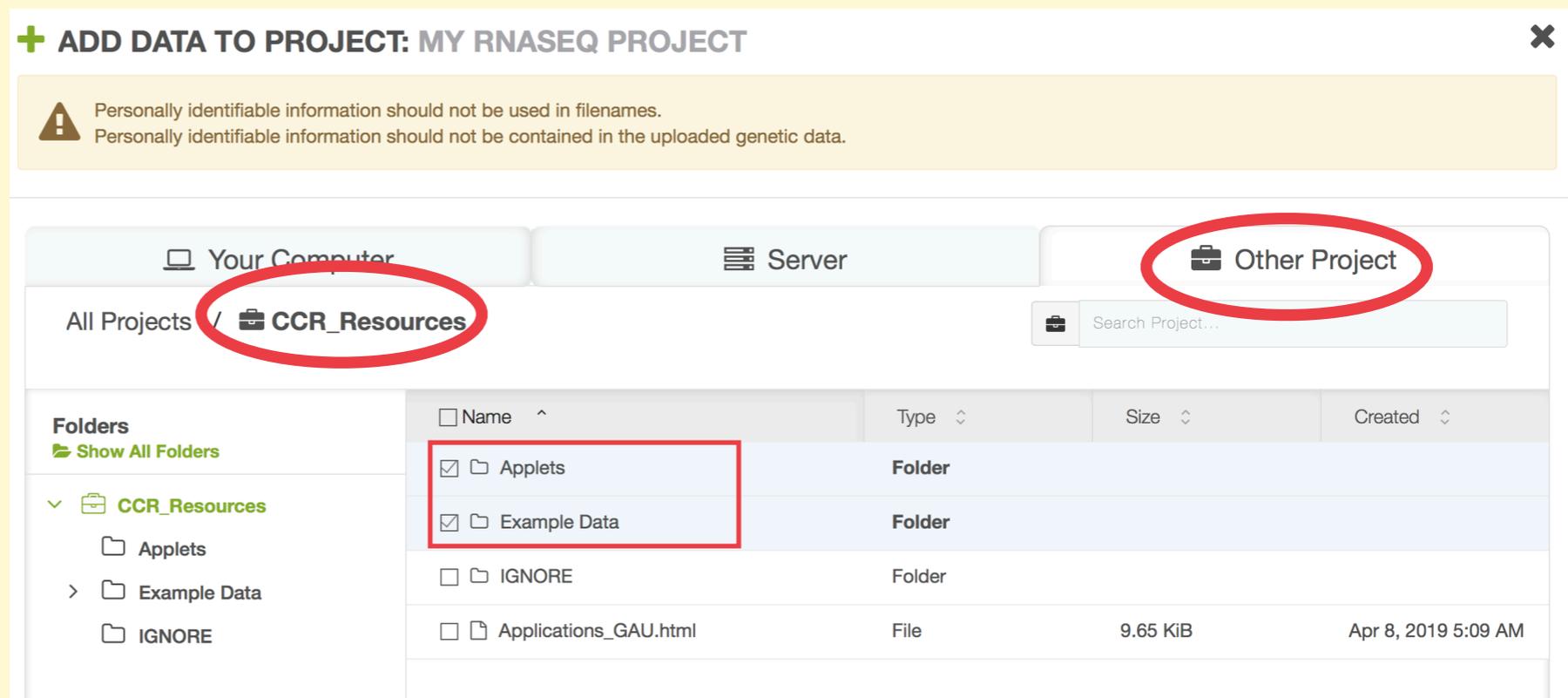
- Other Project
- Project CCR\_Resources
- Folder CCR\_Resources

Check the boxes for

- Applets
- Example Data

Select copy

Close the transfer dialogue



# Step 3: Select the workflow and choose files & parameters

The screenshot shows the 'My RNASEQ Project' interface. At the top, there are navigation tabs: PROJECTS, TOOLS, ORG ADMIN, and HELP. A search bar and user information (Peter Fitzgerald) are also visible. Below the navigation, there are buttons for 'Add Data', 'New Folder', 'New Workflow', and 'Start Analysis'. A filter bar is present with options for 'SEARCH SCOPE', 'NAME', 'ID', 'TYPES', 'MODIFIED', 'TAGS', and 'PROPERTIES'. The main content area shows a table of Applets:

Name	Type	Size	Created
igv_session_maker	Applet	37.15 KiB	Apr 9, 2019 11:39 ...
quant_sf2express_table	Applet	1.11 GiB	Apr 8, 2019 11:24 ...
Salmon-RNAseq	Workflow	—	Apr 10, 2019 1:00 ...
salmon_spg_wf	Applet	1.13 GiB	Apr 8, 2019 9:27 AM

In the left sidebar, the 'Applets' folder is circled in red. In the table, the 'Salmon-RNAseq' row is also circled in red.

The Workflow consists of two applets:

- 1) Generates the gene and transcript count data
- 2) Combines the sample count data into a single matrix file and make the final \*.html output

Select the following:

- Applet folder
- Salmon-RNAseq -workflow

Select the:

- Example Data
- Database

Set the parameters in both applets

The screenshot shows the 'RUN "SALMON-RNASEQ" AS ANALYSIS' configuration window. It displays a workflow diagram with two applets: 'salmon\_spg\_wf' and 'quant\_sf2express...'. The 'salmon\_spg\_wf' applet has two input fields circled in red: '\*fastq.gz | fastq\_gz\_list [array]' and '\*salmon\_idx.tar.gz | salmon\_idx\_file'. The 'quant\_sf2express...' applet has a 'configure params' button circled in red. The right side of the window shows the output files for each applet. At the top right, there is a 'Run as Analysis...' button.

# Step 4: Selecting the sample data

 **SELECT DATA FOR FASTQ\_GZ\_LIST INPUT** ✕  
salmon\_spg\_wf

All Projects /  **My RNASEQ Project**  Search Project...

**PATTERNS** clear

▼ **Files (\*.fastq.gz)**

**Folders**

- ▼  My RNASEQ Project
  -  Applets
  - >  Example Data

<input checked="" type="checkbox"/> Name ^	Type	Size	Created
<input checked="" type="checkbox"/>  brain_rep1_R1.fastq.gz /Example Data/RawDa	<b>File</b>	1.52 GiB	Apr 4, 2019 10:55 PM
<input checked="" type="checkbox"/>  brain_rep1_R2.fastq.gz /Example Data/RawDa	<b>File</b>	1.53 GiB	Apr 4, 2019 10:55 PM
<input checked="" type="checkbox"/>  brain_rep2_R1.fastq.gz /Example Data/RawDa	<b>File</b>	1.78 GiB	Apr 4, 2019 10:56 PM
<input checked="" type="checkbox"/>  brain_rep2_R2.fastq.gz /Example Data/RawDa	<b>File</b>	1.78 GiB	Apr 4, 2019 10:56 PM
<input checked="" type="checkbox"/>  brain_rep3_R1.fastq.gz /Example Data/RawDa	<b>File</b>	1.61 GiB	Apr 4, 2019 10:56 PM
<input checked="" type="checkbox"/>  brain_rep3_R2.fastq.gz /Example Data/RawDa	<b>File</b>	1.62 GiB	Apr 4, 2019 10:57 PM
<input checked="" type="checkbox"/>  muscle_rep1_R1.fastq.gz /Example Data/Rawl	<b>File</b>	1.30 GiB	Apr 4, 2019 11:03 PM
<input checked="" type="checkbox"/>  muscle_rep1_R2.fastq.gz /Example Data/Rawl	<b>File</b>	1.30 GiB	Apr 4, 2019 11:04 PM
<input checked="" type="checkbox"/>  muscle_rep2_R1.fastq.gz /Example Data/Rawl	<b>File</b>	1.60 GiB	Apr 4, 2019 11:04 PM
<input checked="" type="checkbox"/>  muscle_rep2_R2.fastq.gz /Example Data/Rawl	<b>File</b>	1.61 GiB	Apr 4, 2019 11:04 PM

**Suggestions**  My RNASEQ Project

≡ 12 Items Selected

Cancel Select

# Step 5: Select the transcriptome file - this “lives” in the Helper directory in the Applet folder

**SELECT DATA FOR SALMON\_IDX\_FILE INPUT** salmon\_spg\_wf ✕

All Projects / **My RNASEQ Project**

**PATTERNS** [clear](#)

**Files (\*salmon\_idx.tar.gz)**

Name ^	Type ^	Size ^	Created ^
 human_gcv29_salmon_idx.tar.gz /Applets/Helpers	File	2.88 GiB	Apr 5, 2019 3:48 PM
 mouse_gcvM20_salmon_idx.tar.gz /Applets/Help	File	2.44 GiB	Apr 5, 2019 3:49 PM
 yeast_S288C_salmon_idx.tar.gz /Applets/Helpers/	File	135.66 MiB	Apr 5, 2019 3:50 PM

**Folders**

- My RNASEQ Project
  - Applets
  - Example Data

**Suggestions** **My RNASEQ Project**

# Step 6: Provide an output directory name

## ⚙️ CONFIGURE: SALMON\_SPG\_WF (APPLET) ✕

✓ SSH is allowed for this app.

salmon\_spg\_wf

### About Applet ...

Salmon Scatter-Process\_Gather Workflow

This applet process a batch of pair-end \*.fastq.gz read files and runs [Salmon](#).

To use the developer's words:

Salmon is a tool for **wicked-fast** transcript quantification from RNA-seq data. It requires a set of target transcripts (either from a reference or de-novo assembly) to quantify. All you need to run Salmon is a FASTA file containing your reference transcripts and a (set of) FASTA/FASTQ file(s) containing your reads. Optionally, Salmon can make use of pre-computed alignments (in the form of a SAM/BAM file) to the transcripts rather than the raw reads.

Developed by: [Fitzgerald, Peter (NIH/NCI) [E]] ([fitzgepe@mail.nih.gov](mailto:fitzgepe@mail.nih.gov)) and [McIntosh, Carl (NIH/NCI) [E]] ([mcintoshc@mail.nih.gov](mailto:mcintoshc@mail.nih.gov))

Group: [Genome Analysis Unit](#)

### Required Input Files

**FASTQ Gzip Compressed Paired-end Files** - A batch sample PE read files with the form \*\_R1.fastq.gz and \*\_R2.fastq.gz.

**Salmon Index tar.gz File** - A Salmon Indexed genome files with the form \*\_salmon\_idx.tar.gz .

### Input Parameters

**Output Folder** - Provide an output directory name for result files

\* Fields are required

**Name**  \*

**Output Folder**  \*

**Instance type**  Select ▾

### COMMON

**Bootstrap Value**  \*

# Step 7: Provide an output directory name and a file prefix

## CONFIGURE: QUANT\_SF2EXPRESS\_TABLE (APPLET)

✔ SSH is allowed for this app.

Convert quant.sf files to expression tables.

### About Applet ...

This applet converts a batch of \*\_quant.sf input files generated by applet **salmon\_spg\_wf** and produces expression tables with sample names in columns and genes names in rows.

Developed by: [Fitzgerald, Peter (NIH/NCI) [E]] ([fitzgepe@mail.nih.gov](mailto:fitzgepe@mail.nih.gov)) and [McIntosh, Carl (NIH/NCI) [E]] ([mcintoshc@mail.nih.gov](mailto:mcintoshc@mail.nih.gov))

Group: [Genome Analysis Unit](#)

### Required Input Files

**Quant SF Files** - Selected files name ending in \*\_quant.sf produced by applet **salmon\_spg\_wf**.

### Input Parameters

**Output Folder** - Provide an output directory name for result files.

**Instance type** - For this applet, asking for more computer resources will not reduce run time, but will cost more.

### COMMON Input Parameters

**Prefix** - A prefix to pre-append resulting files.

### Output Files

**Expression HTML File** - An output file that provides useful links, *DNAexus* job information and instructions on submitting to [BioJupies](#) and [iDep](#) which provide downstream \_RNA\_seq analysis for the expression tables.

**Raw Counts Table File** - File containing table with unprocessed raw counts.

**TPM Counts Table File** - File containing table TPM (*transcripts per million reads*) counts.

\* Fields are required

**Name**  \*

**Output Folder**  \*

**Instance type**  Select ▾

### COMMON

**prefix**  \*

[Reset to applet defaults](#)

# Step 9: Monitor the Job (or not)

My RNASEQ Project Settings Manage Monitor 1 Visualize Access: Admin 1 Share

SEARCH SCOPE: Root executions only STATE: Any 1 recent job NAME: Any ID: Any CREATED: Any LAUNCHED BY: Any FILTERS SAVED FILTERS

Status	Name	Executable	Launched by	Started running	Duration	Price	Worker URL
In Progress	Salmon-RNAseq	Salmon-RNAseq	Peter Fitzgerald	-	< 1m	~ \$0 as of Apr 10, 1:12 PM	

My RNASEQ Project Settings Manage Monitor 1 Visualize Access: Admin 1 Share

SEARCH SCOPE: Root executions only STATE: Any 1 recent job NAME: Any ID: Any CREATED: Any LAUNCHED BY: Any FILTERS SAVED FILTERS

Status	Name	Executable	Launched by	Started running	Duration	Price	Worker URL
In Progress	Salmon-RNAseq	Salmon-RNAseq	Peter Fitzgerald	-	2m	~ \$0.0038 as of Apr 10, 1:14 PM	
Running	salmon_spg_wf	salmon_spg_wf	Peter Fitzgerald	04/10/2019 1:13 pm	1m	-	
Waiting on Input	quant_sf2express_table	quant_sf2express_table	Peter Fitzgerald	-	-	-	

These images show different stages of the process. The job can be terminated at any time by clicking on the Terminate button

States are:

Waiting 

Running 

Error 

Processing Sample Batch DONE Launch as new Job View Log View all Inputs/Outputs View Info

EXECUTION ID: job-FXg28yQ0ZxF24qkGP2xgX2B2 PARENT EXECUTION ID: job-FXg27zQ0ZxFG2kKv80xB8k81 LAUNCHED ON: 04/10/2019 1:14 pm RAN FOR: 15m EXECUTABLE: salmon\_spg\_wf

LAUNCHED BY: Peter Fitzgerald OUTPUT FOLDER: /

Time	Job Name	Duration	Log
01:17:07PM	Processing Sample Batch	15m	Log
01:21:17PM	brain_rep2	14m	Log
01:24:12PM	brain_rep3	10m	Log
01:27:08PM	brain_rep1	13m	Log
04/10/2019 01:32:32PM	muscle_rep2	11m	Log
	muscle_rep3	10m	Log
	muscle_rep1	9m	Log

INPUTS: process\_input, array\_of\_scattered\_input

OUTPUTS: Batch of abundance\_h5 files (abundance\_h5\_s), brain\_rep2\_abundance.h5, brain\_rep3\_abundance.h5, brain\_rep1\_abundance.h5, muscle\_rep2\_abundance.h5, muscle\_rep3\_abundance.h5, muscle\_rep1\_abundance.h5

# Step 10: The final output from the DNAnexus Workflow

Since DNAnexus does not currently provide truly interactive utilities we have chosen to provide the option of using two external utilities.  
BioJupies  
iDep

Thus the **final step** is to download the count matrix file to your local machine and then upload to one or both of these external resources

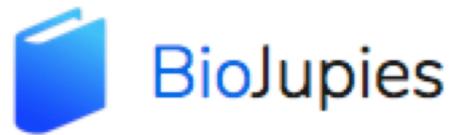
The screenshot shows a web browser window with the URL [dl.dnanex.us/F/D/7Y9bYjzyBy4vzJbVYbjbz85PqgFJ1j7Py5qxKXqZ?inline](https://dl.dnanex.us/F/D/7Y9bYjzyBy4vzJbVYbjbz85PqgFJ1j7Py5qxKXqZ?inline). The browser's address bar and tabs are visible at the top. The page content is organized into several sections:

- About this Applet**
  - [About GAU](#)
  - [About BTEP](#)
- Usefull Links**

See a complete summary at [DNAnexus Job Monitor](#).
- Input Transcript Quant File Summary**
  - [brain\\_rep2\\_quant.sf](#)
  - [brain\\_rep3\\_quant.sf](#)
  - [brain\\_rep1\\_quant.sf](#)
  - [muscle\\_rep2\\_quant.sf](#)
  - [muscle\\_rep3\\_quant.sf](#)
  - [muscle\\_rep1\\_quant.sf](#)
- Input Gene Quant File Summary**
  - [brain\\_rep2\\_quant\\_genes.sf](#)
  - [brain\\_rep3\\_quant\\_genes.sf](#)
  - [brain\\_rep1\\_quant\\_genes.sf](#)
  - [muscle\\_rep2\\_quant\\_genes.sf](#)
  - [muscle\\_rep3\\_quant\\_genes.sf](#)
  - [muscle\\_rep1\\_quant\\_genes.sf](#)
- Instructions**
  - Download Expression Tables from *DNAnexus*
    - Download [RAW Counts Table for Transcripts](#)
    - Download [TPM \(Transcripts Per Million\) Counts Table for Transcripts](#)
    - Download [RAW Counts Table for Genes](#)
    - Download [TPM \(Transcripts Per Million\) Counts Table for Genes](#)
  - Download and Edit Design Table in *Excel* or Text Editor for use in *iDEP*
    - Download [Design Table](#)
  - Select Analysis Site and Upload Expression Table
    - Upload an Expression Table File to [BioJupies](#)
    - Upload an Expression Table File to [iDEP](#)

At the bottom of the page, it says "Created by: [Genome Analysis Unit](#)".

# Two excellent Analysis Options - Shiny App Servers (R)



## Step 1. Upload or Fetch RNA-seq Data

- Upload your raw or processed RNA-seq data
- Fetch >8,000 public RNA-seq datasets published in the Gene Expression Omnibus

## Step 2. Select Data Analysis Tools

- Select from multiple state-of-the-art RNA-seq data analysis tools
- Contribute your computational tool as a plugin

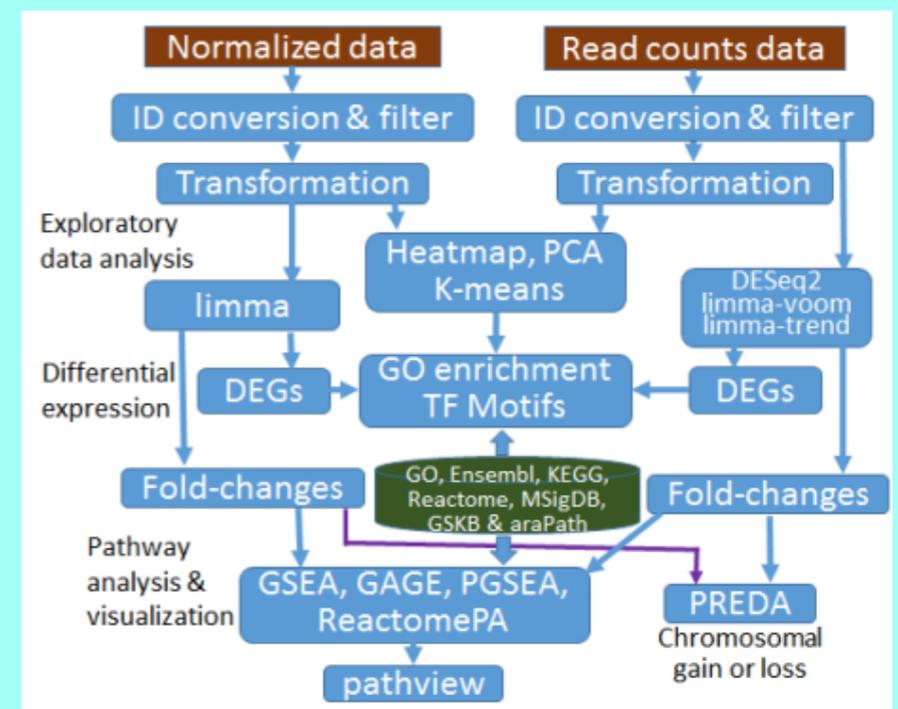
## Step 3. Generate Your Notebook

- Access and share your results through a permanent URL
- Download, rerun and customize your notebook using Docker

## BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface <https://amp.pharm.mssm.edu/biojupies/upload/table>

**iDEP** (*integrated Differential Expression and Pathway analysis*) is a web-based tool for analyzing RNA-seq data, available at <http://bioinformatics.sdstate.edu/idep/>. It reads in gene-level expression data (read counts or FPKM), performs exploratory data analysis (EDA), differential expression, pathway analysis, biclustering, and co-expression network analysis. iDEP also accepts DNA microarray data or other gene-level expression data, such as those from Chip-seq or proteomics studies.



# WHAT WE HAVE ACHIEVED

- ~**20** people have analyzed 6 RNASEQ samples (not subsetted) in ~10 steps, in less than an hour, at the cost of ~\$0.35/person or ~\$7 total
- Each sample(6) ran on its own 4 cores which comes to 24 cores/person for a total of 480 cores and 120 samples !!

# ChIP-SEQ

*Desiree Tillo*

Genome Analysis Unit

Custom Work Flows developed by St. Jude and ENCODE

## ChIP-seq processing workflows in DNAnexus

- St. Jude Cloud (<https://platform.stjude.cloud/tools/chip-seq>)  
*[Note this runs in the AZURE space]*
- ENCODE (<https://platform.dnanexus.com/projects/featured>)  
(select under “ENCODE Uniform processing pipelines”)

## Summary of the St. Jude ChIP-seq workflow

**Input:** fastq files from a ChIP seq run (case only or paired case+control)

### Mapping:

- Align reads (fastq.gz files) to reference genome (human, mouse, drosophila) using bwa
- Post processing of reads (removing multiple mapped reads, removing duplicated reads).

### Peak-calling:

- MACS2 (narrow peak analysis: transcription factors, certain chromatin marks)
- SICER (broad peak analysis, certain chromatin marks e.g. H3k27me3)

### Output:

- Peaks as BED (.bed) and big BED (.bb) files.
- Genome coverage files as bigWig (.bw) file for each input fastq.
- QC plots and files:
  - A cross correlation plot (for measuring fragment length and computation of data quality)
  - Sequencing quality metrics (output from fastqc)

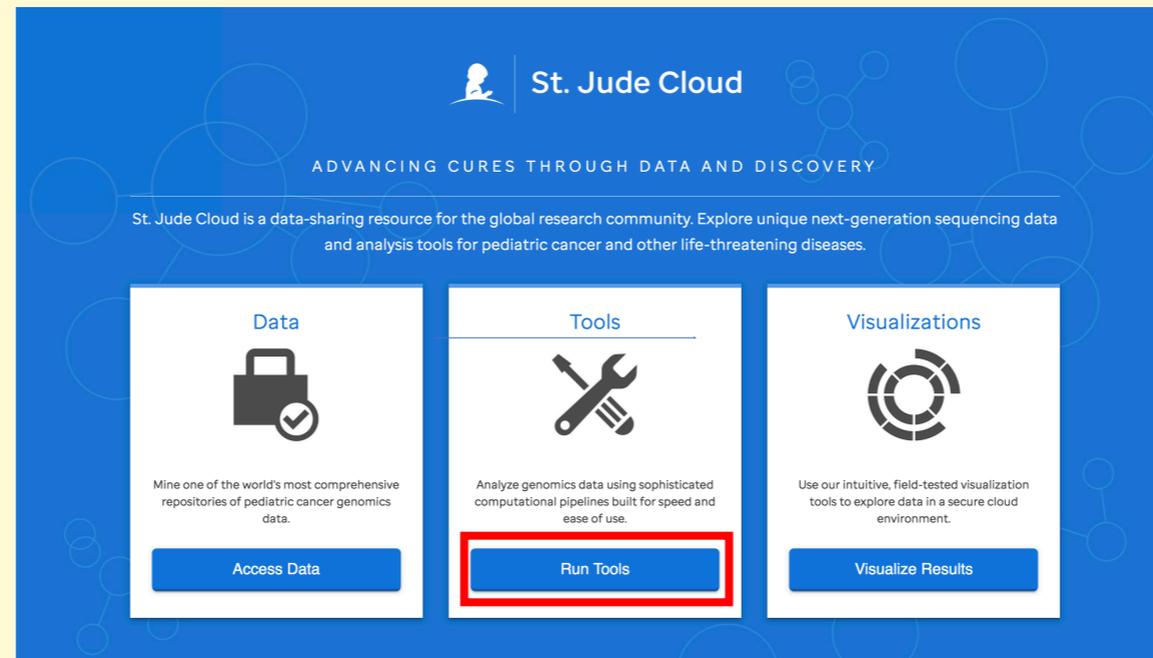
<https://stjude.github.io/sjcloud-docs/guides/tools/chipseq/>

# Running ChIP-seq workflows on the St. Jude cloud

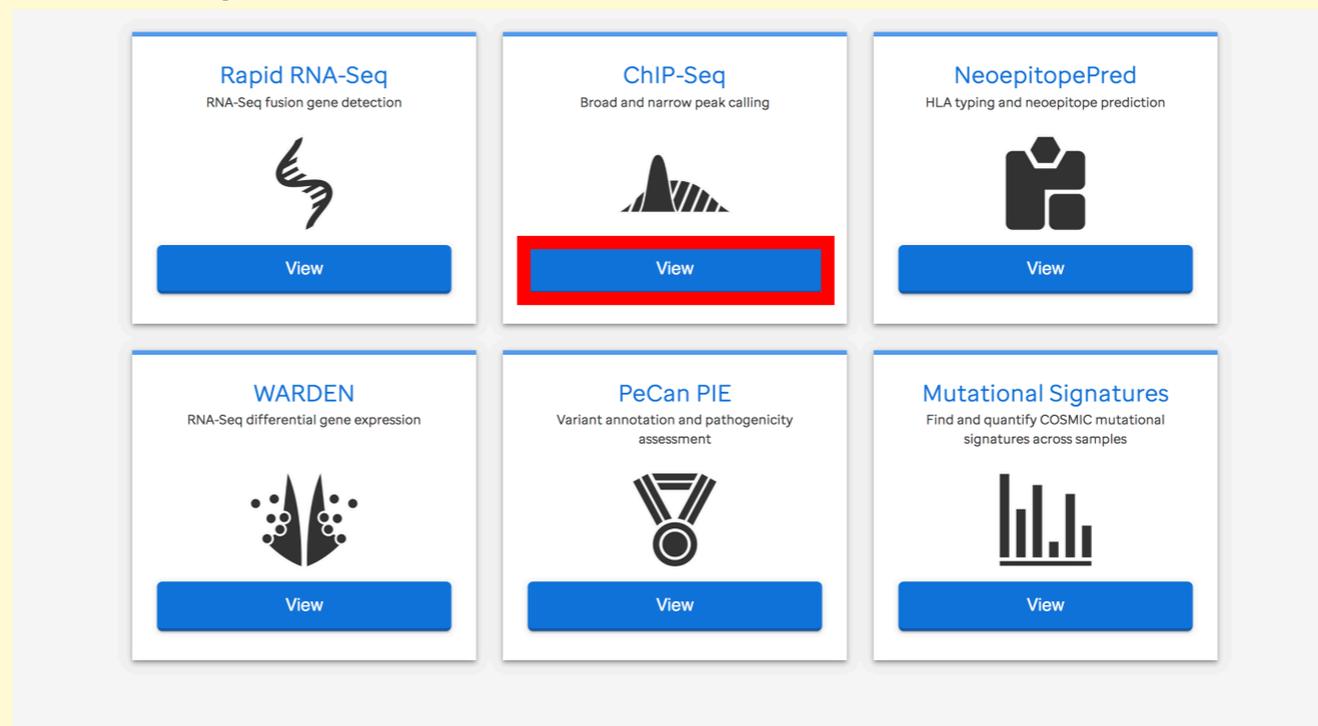
## Step 1. Set up

Go to St. Jude Cloud (<https://www.stjude.cloud/>)

Click on “Run tools”:



Click on the “View” button under “ChIP-seq”:



On the left panel select "Log in to launch this tool", and log in with your DNAnexus account:

**St. Jude Cloud Platform** DATA TOOLS VISUALIZATIONS User

## CHIP-Seq

*Broad and Narrow Peak Calling*

**Authors** Xing Tang, Yong Cheng  
**Publication** N/A (not published)  
**Input** Paired CHIP-Seq FASTQ files  
**Output** Peak coordinates BED file, coverage BigWigs, cross correlation plots to show enrichment quality.  
**Supported Genomes** HG19 (GRCh37), HG38 (GRCh38), MM9, MM10 (GRCm38), DM3 (BDGP5)  
**Technical Support** [Contact Us](#)  
**Open Source** Stars 0 Watch 6 Issues 0 open

[Log in](#) to launch this tool

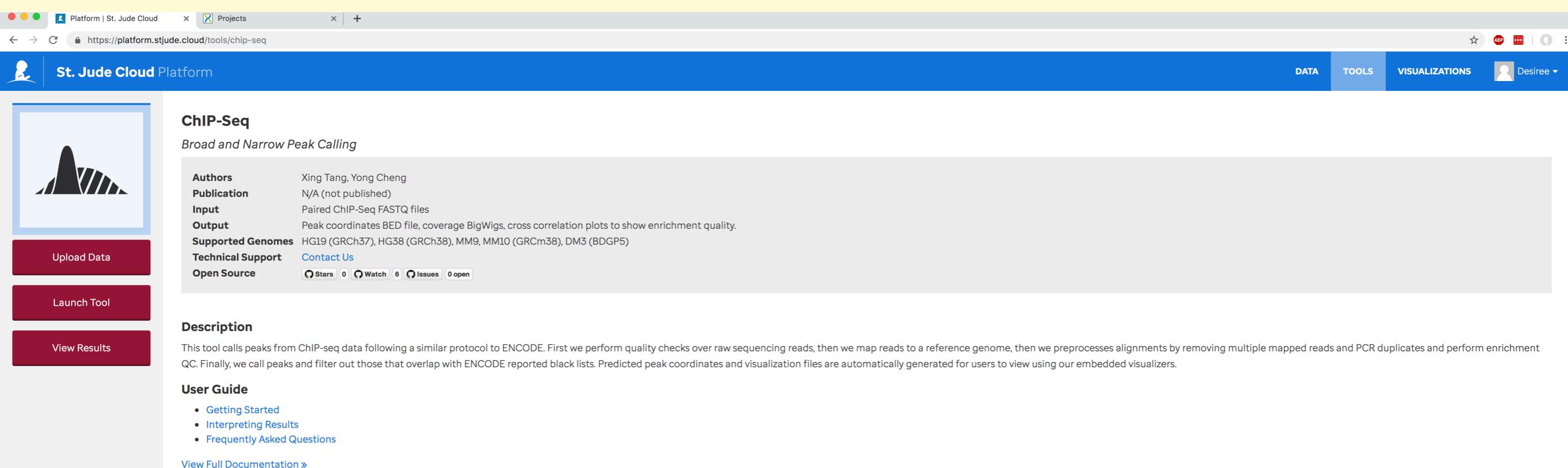
This generates a project called "ChIP-Seq" in your list of projects on your DNAnexus account. In it will be all of the available ChIP-Seq workflows (Broad Peak Caller, Narrow Peak Caller, etc):

<input type="checkbox"/>	Name ^	Type ^	Size ^	Created ^	<input type="checkbox"/>
<input type="checkbox"/>	Results	Folder			
<input type="checkbox"/>	uploads	Folder			
<input type="checkbox"/>	ChIP-seq Broad Peak Caller (Case + C...	Workflow	—	Sep 20, 2018 10:17 AM	
<input type="checkbox"/>	ChIP-seq Broad Peak Caller (Case)	Workflow	—	Sep 20, 2018 10:23 AM	
<input type="checkbox"/>	ChIP-seq Narrow Peak Caller (Case + ...	Workflow	—	Sep 20, 2018 10:07 AM	
<input type="checkbox"/>	ChIP-seq Narrow Peak Caller (Case)	Workflow	—	Sep 20, 2018 10:14 AM	

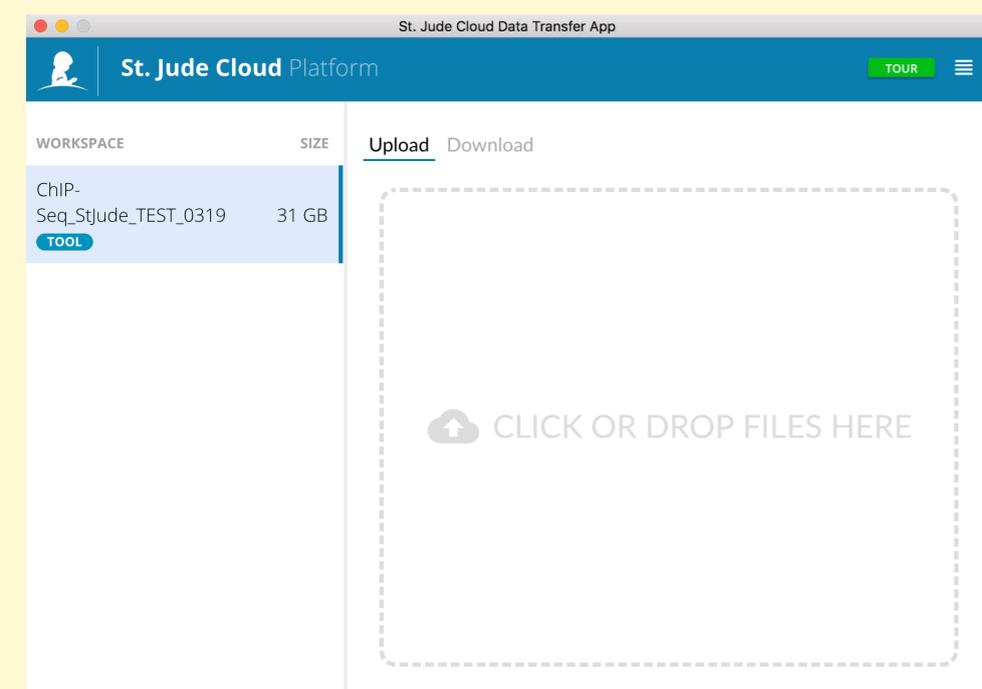
## Step 2. Upload your data (.fastq files from a ChIP seq run)

### Option 1: Use the St. Jude drag and drop data transfer app:

Click on “Upload data” on the St. Jude ChIP-seq tool page, which will take you to another page to either open or download the app.



The screenshot shows the St. Jude Cloud Platform interface for the ChIP-Seq tool. The page title is "ChIP-Seq" with the subtitle "Broad and Narrow Peak Calling". The authors are Xing Tang and Yong Cheng. The publication is listed as "N/A (not published)". The input is "Paired ChIP-Seq FASTQ files" and the output is "Peak coordinates BED file, coverage BigWigs, cross correlation plots to show enrichment quality." The supported genomes are HG19 (GRCh37), HG38 (GRCh38), MM9, MM10 (GRCm38), and DM3 (BDGP5). There is a "Contact Us" link for technical support and an "Open Source" section with 0 stars, 6 watches, 0 issues, and 0 open items. The description states that the tool calls peaks from ChIP-seq data following a similar protocol to ENCODE, performing quality checks, mapping reads, preprocessing alignments, and enrichment QC. The user guide includes links for "Getting Started", "Interpreting Results", and "Frequently Asked Questions". A "View Full Documentation" link is also present.



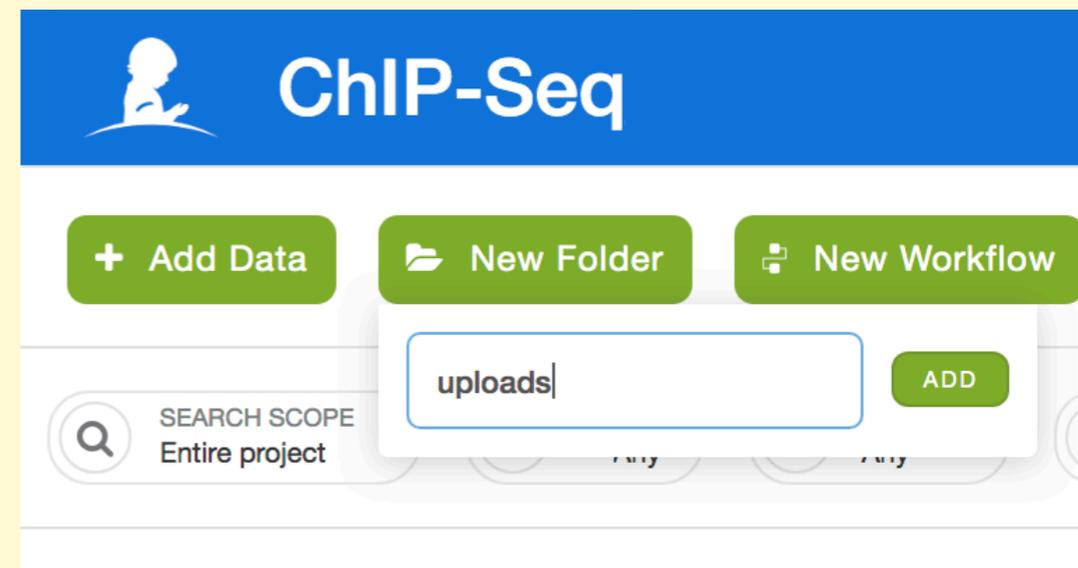
The screenshot shows the St. Jude Cloud Data Transfer App interface. The workspace is named "ChIP-Seq\_StJude\_TEST\_0319" and has a size of 31 GB. The app is currently in "Upload" mode. A large dashed box in the center of the workspace contains a cloud upload icon and the text "CLICK OR DROP FILES HERE".

Uploading data using the app will place the data in a folder called “uploads” in your DNAnexus workspace

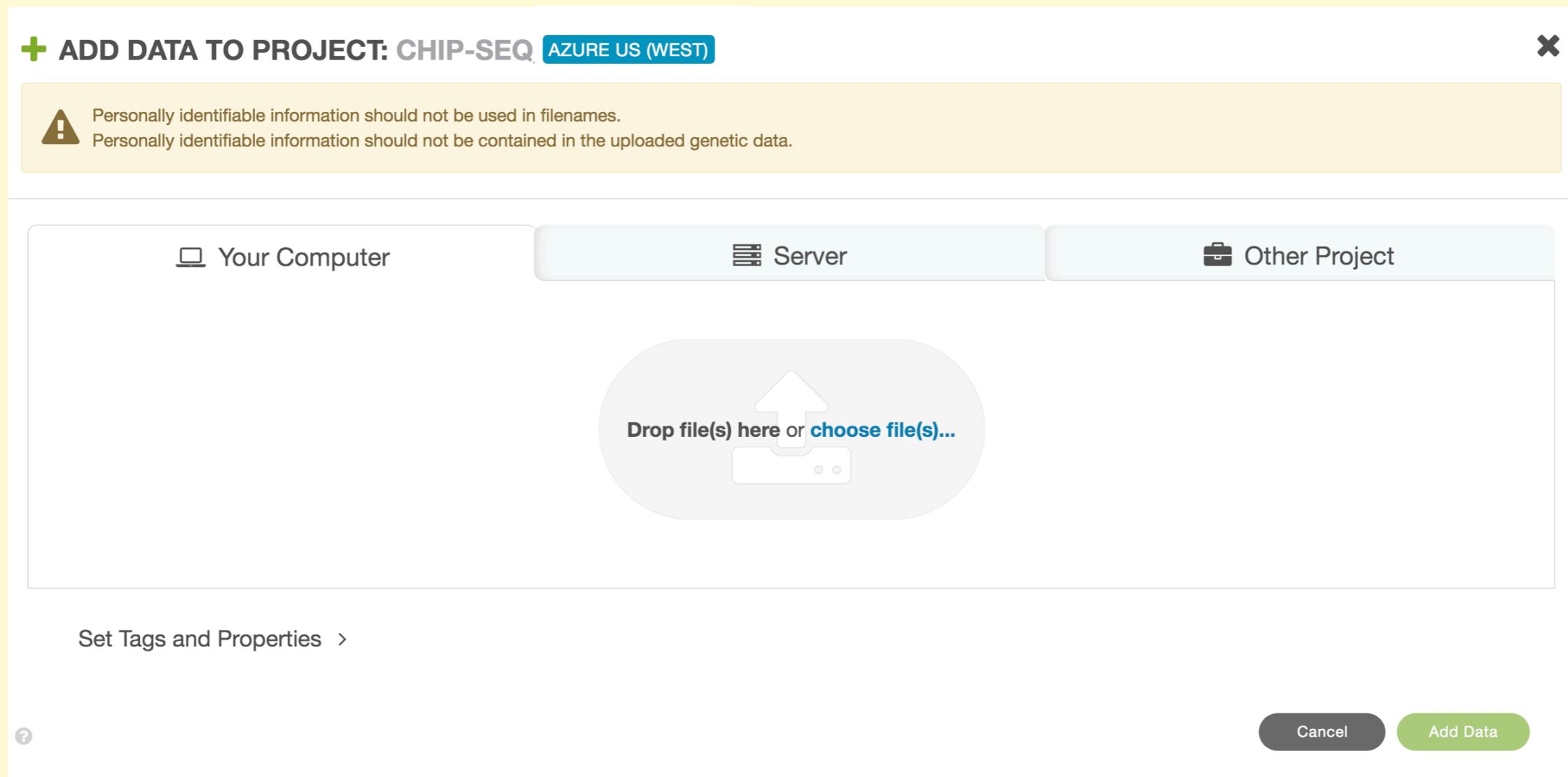
# Step 2. Upload your data (.fastq files from a ChIP-seq run)

## Option 2: On DNAnexus

Create a data/uploads folder in your St. Jude ChIP-Seq workspace (click “New Folder”)



Navigate to your data folder, and click “Add data”

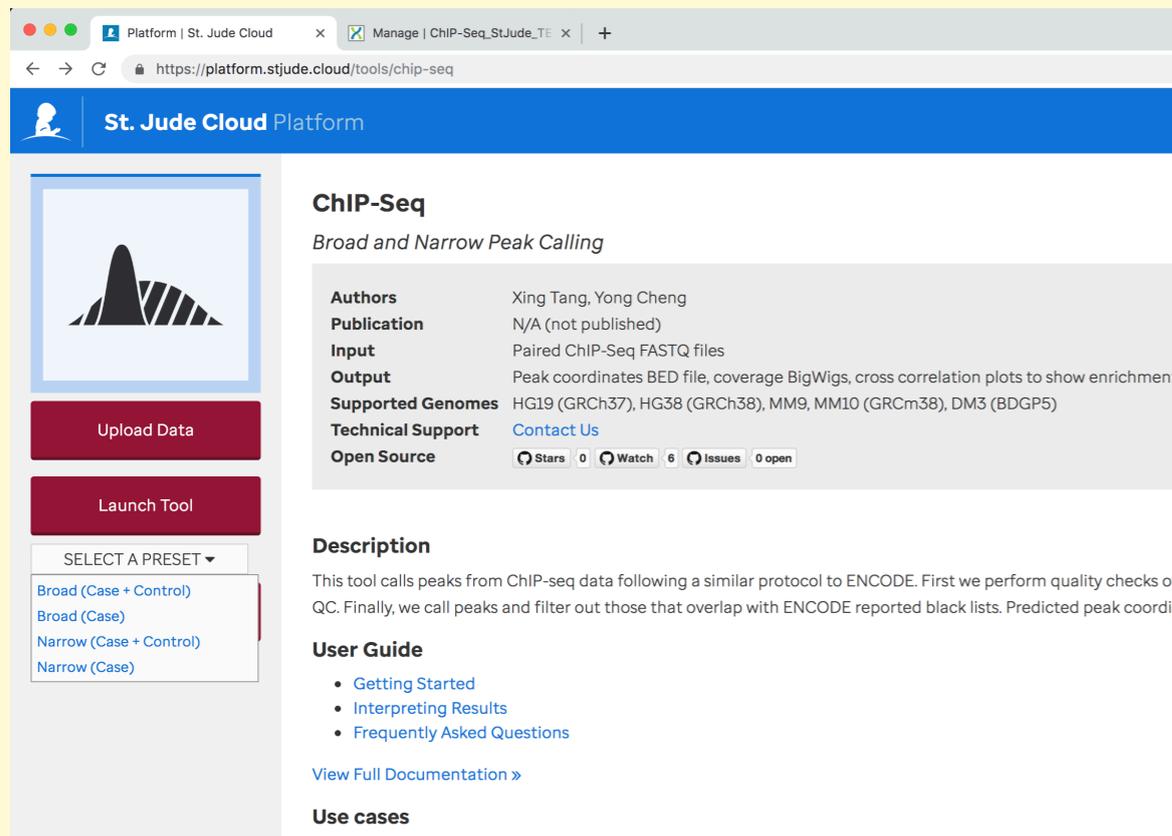


# Step 3. Launch the workflow

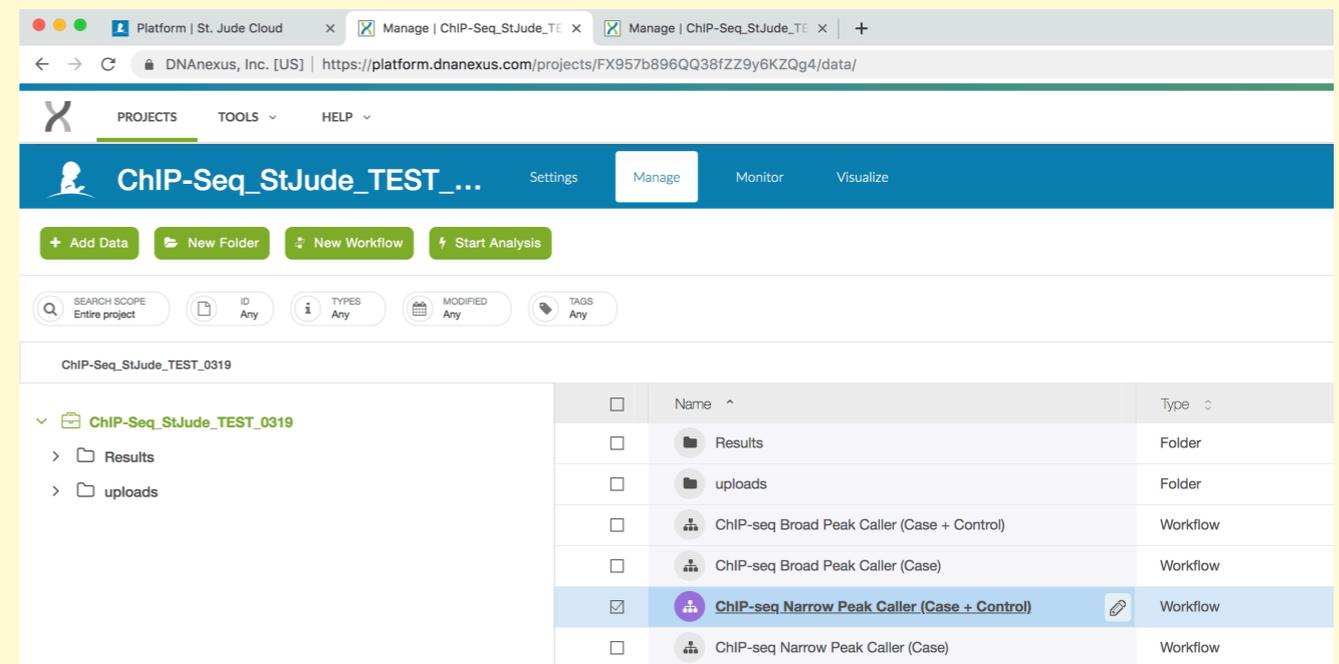
## Option 1: St. Jude portal

Click launch tool, select the workflow from the dropdown menu that works best for your needs

## Option 2: Directly on DNAnexus



The screenshot shows the St. Jude Cloud Platform interface for the CHIP-Seq tool. The page title is "CHIP-Seq" with the subtitle "Broad and Narrow Peak Calling". The authors listed are Xing Tang and Yong Cheng. The publication is noted as "N/A (not published)". The input is "Paired ChIP-Seq FASTQ files" and the output is "Peak coordinates BED file, coverage BigWigs, cross correlation plots to show enrichment". Supported genomes include HG19 (GRCh37), HG38 (GRCh38), MM9, MM10 (GRCm38), and DM3 (BDGP5). Technical support is available via "Contact Us". The open source status shows 0 stars, 6 watches, and 0 open issues. A "Description" section explains that the tool calls peaks from ChIP-seq data following a protocol similar to ENCODE, performing quality checks and filtering out peaks overlapping with ENCODE reported black lists. A "User Guide" section includes links for "Getting Started", "Interpreting Results", and "Frequently Asked Questions". A "View Full Documentation" link is also present. The "Use cases" section is partially visible. On the left side, there are buttons for "Upload Data" and "Launch Tool", and a "SELECT A PRESET" dropdown menu with options: "Broad (Case + Control)", "Broad (Case)", "Narrow (Case + Control)", and "Narrow (Case)".



The screenshot shows the DNAnexus interface for the project "CHIP-Seq\_StJude\_TEST\_0319". The interface includes a navigation bar with "PROJECTS", "TOOLS", and "HELP" menus. Below the navigation bar, there are buttons for "Add Data", "New Folder", "New Workflow", and "Start Analysis". A search bar is present with filters for "SEARCH SCOPE", "ID", "TYPES", "MODIFIED", and "TAGS". The main content area displays a list of workflows under the project "CHIP-Seq\_StJude\_TEST\_0319". The list includes folders for "Results" and "uploads", and several workflows. The workflow "ChIP-seq Narrow Peak Caller (Case + Control)" is selected, indicated by a checkmark in the first column and a blue highlight.

	Name	Type
<input type="checkbox"/>	Results	Folder
<input type="checkbox"/>	uploads	Folder
<input type="checkbox"/>	ChIP-seq Broad Peak Caller (Case + Control)	Workflow
<input type="checkbox"/>	ChIP-seq Broad Peak Caller (Case)	Workflow
<input checked="" type="checkbox"/>	ChIP-seq Narrow Peak Caller (Case + Control)	Workflow
<input type="checkbox"/>	ChIP-seq Narrow Peak Caller (Case)	Workflow

# Either method will open the workflow (Option 1 will open a new tab)

Input your uploaded fastq into the input boxes

St. Jude Cloud

CHIP-seq Narrow Peak Caller (Case + Control)

1 app unconfigured | 6 apps configured | Workflow Actions | Run as Analysis...

Inputs: \*.fq.gz \*.fastq.gz ChIP Reads | \*.fq.gz \*.fastq.gz Control Reads

App: Parameter Wrapper (set inputs and params), FastQC (IP), FastQC (Control), BWA (IP), BWA (Control), MACS2, Report

Outputs: the file with parameters, \*.stats-fastqc.html FastQC Report, \*.stats-fastqc.txt FastQC Stats, \*.bam Sorted mappings, \*.bai Sorted mappings index, \*.log log for bwa, \*.bed bed file of identified peaks, \*.bb big bed file of identified peaks, \*.metrics.txt file of metrics, \*.xls raw peak file directly output, \*.out\_report

Set parameters (reference genome, prefix for output) by clicking “Parameter Wrapper”:

CONFIGURE: PARAMETER WRAPPER VERSION 0.0.14

Wrapper application for ChIP-seq pipeline

St. Jude ChIP-seq Parameter Wrapper

This app is a setup step for the ChIP-seq pipeline. It configures the pipeline based on the chosen settings.

The output folder is not necessary, and can be left as '/'. The output prefix is required.

The final outputs will be found in the following path: OUTPUT\_FOLDER/Results/OUTPUT\_PREFIX. This path will be created by the pipeline.

If the output path exists before the pipeline run and is non-empty, then a part of the unique job id will be added to the folder name.

Fields are required

Version: 0.0.14

Name: Parameter Wrapper

Output Folder: [empty]

Instance type: mem2\_ssd1\_x4

COMMON

prefix for output: hg38\_ctcf\_chr21\_test

reference genome: mouse: mm9(MGSCv37), mm10(GRCm38); human: hg19(GRCh37), hg38(GRCh38); drosophila: dm3(BDGPv5) | GRCh38

output big wig file or not. The wiggle files can be uploaded to genome browser to view reads distribution along chromosomes. | True (default) False

remove peaks from black list or not | True (default) False

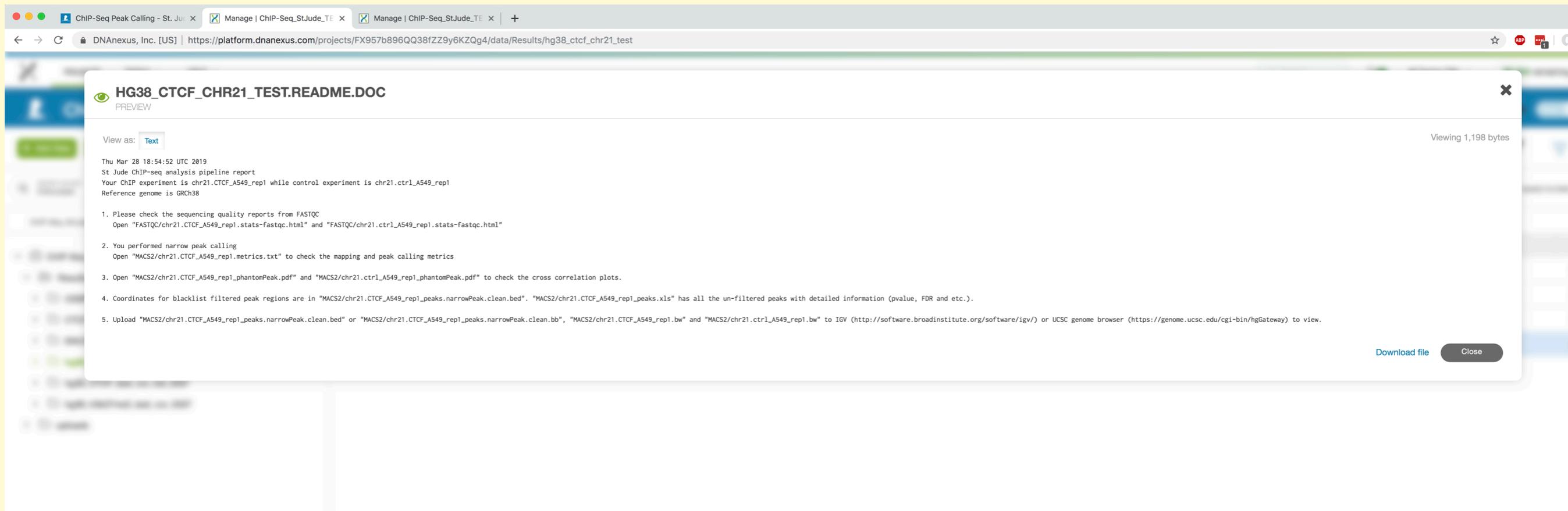
This provides you an option to run peak calling with a specified fragment length instead of estimating it based on cross correlation. Input the number of base pairs like: 200. | NA

Reset to app defaults | Save

Hit save, and click “Run as analysis”

**Results will be in results/<prefix>**

**A summary of the output will be in results/<prefix>.doc**



The screenshot shows a web browser window with a document preview overlay. The document title is "HG38\_CTCF\_CHR21\_TEST.README.DOC" and it is in "PREVIEW" mode. The document content includes a timestamp, a description of the St Jude ChIP-seq analysis pipeline report, and a list of five instructions for checking sequencing quality, peak calling metrics, cross-correlation plots, blacklist filtered peak regions, and uploading files to IGV or UCSC genome browser.

**HG38\_CTCF\_CHR21\_TEST.README.DOC**  
PREVIEW

View as:  Viewing 1,198 bytes

Thu Mar 28 18:54:52 UTC 2019  
St Jude ChIP-seq analysis pipeline report  
Your ChIP experiment is chr21.CTCF\_A549\_rep1 while control experiment is chr21.ctr1\_A549\_rep1  
Reference genome is GRCh38

1. Please check the sequencing quality reports from FASTQC  
Open "FASTQC/chr21.CTCF\_A549\_rep1.stats-fastqc.html" and "FASTQC/chr21.ctr1\_A549\_rep1.stats-fastqc.html"
2. You performed narrow peak calling  
Open "MACS2/chr21.CTCF\_A549\_rep1.metrics.txt" to check the mapping and peak calling metrics
3. Open "MACS2/chr21.CTCF\_A549\_rep1\_phantomPeak.pdf" and "MACS2/chr21.ctr1\_A549\_rep1\_phantomPeak.pdf" to check the cross correlation plots.
4. Coordinates for blacklist filtered peak regions are in "MACS2/chr21.CTCF\_A549\_rep1\_peaks.narrowPeak.clean.bed". "MACS2/chr21.CTCF\_A549\_rep1\_peaks.xls" has all the un-filtered peaks with detailed information (pvalue, FDR and etc.).
5. Upload "MACS2/chr21.CTCF\_A549\_rep1\_peaks.narrowPeak.clean.bed" or "MACS2/chr21.CTCF\_A549\_rep1\_peaks.narrowPeak.clean.bb", "MACS2/chr21.CTCF\_A549\_rep1.bw" and "MACS2/chr21.ctr1\_A549\_rep1.bw" to IGV (<http://software.broadinstitute.org/software/igv/>) or UCSC genome browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>) to view.

[Download file](#) [Close](#)

**Can view .bw, .bb, files on IGV or UCSC genome browser**

# Summary of the ENCODE uniform processing pipeline for ChIP-seq on DNAnexus

<https://github.com/ENCODE-DCC/chip-seq-pipeline>

**Inputs: fastq.gz files (case only, case+control) from a ChIP seq run.**

- Can be SE or PE, and replicates

**Mapping (mouse, human, custom)**

- Map reads with BWA, mark duplicates with Picard, and remove duplicates.
- Enrichment QC: Estimate library complexity and calculate NRF (non-redundant fraction), PBC1, PBC2 (PCR bottleneck coefficient).
- Calculate cross-correlation analysis with SPP/phantompeakqualtools.

**Signal tracks**

- Generate p-value and fold-over-control signal tracks for each replicate and replicates pooled with MACS2.

**Peak-calling (histone marks)**

- Call peaks with MACS2.
- Calculate and report overlapping peaks from both replicates.

**Peak-calling (transcription factors)**

- Call peaks with SPP.
- Threshold peaks with IDR.
- Report IDR-thresholded peak sets, self-consistency ratio, rescue ratio, reproducibility test.

**Output:**

- peak coordinates (.bed, .bb)
- signal tracks: coverage, fold enrichment/control, p-value (.bw)
- Various QC plots and files:
- Mapping stats (flagstat)
- A cross correlation plot (for measuring fragment length and data quality metrics)
- IDR output (measures consistency between replicates, uses reproducibility in score rankings between peaks in each replicate to determine an optimal cutoff for significance )

# Running the ENCODE ChIP-seq workflow on DNAnexus

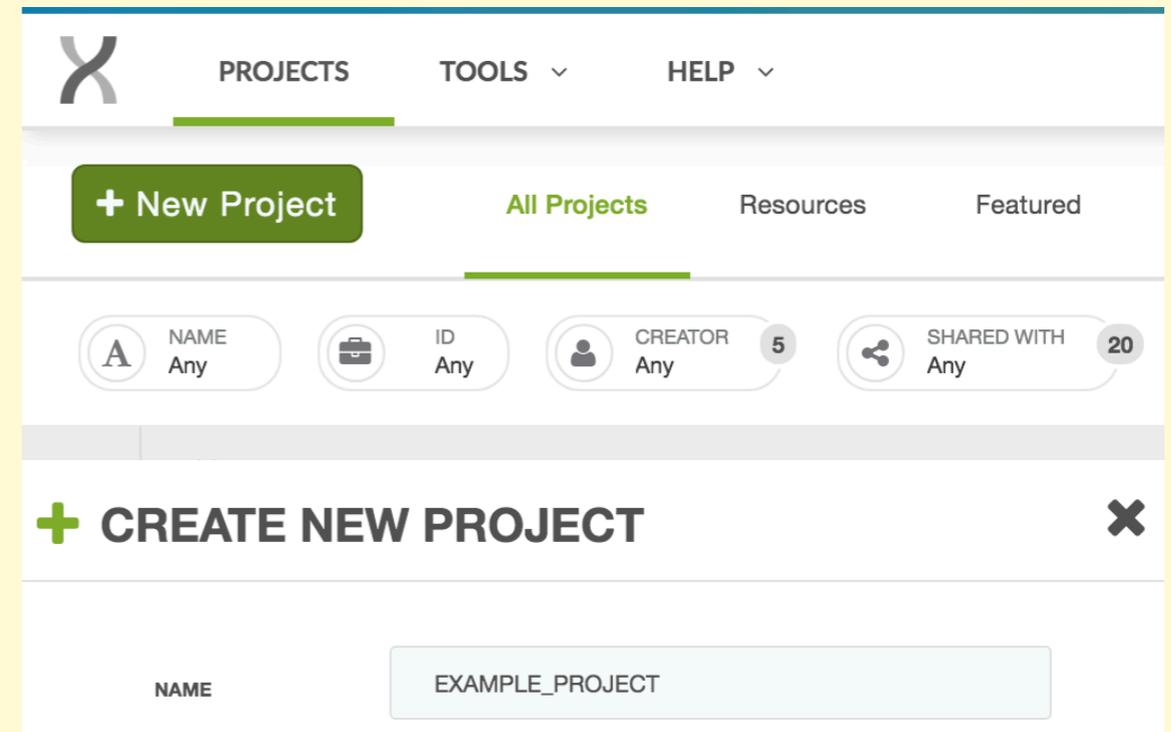
<https://www.encodeproject.org/tutorials/chip-pipeline-howto/>

## Step 1. Set up+Data upload

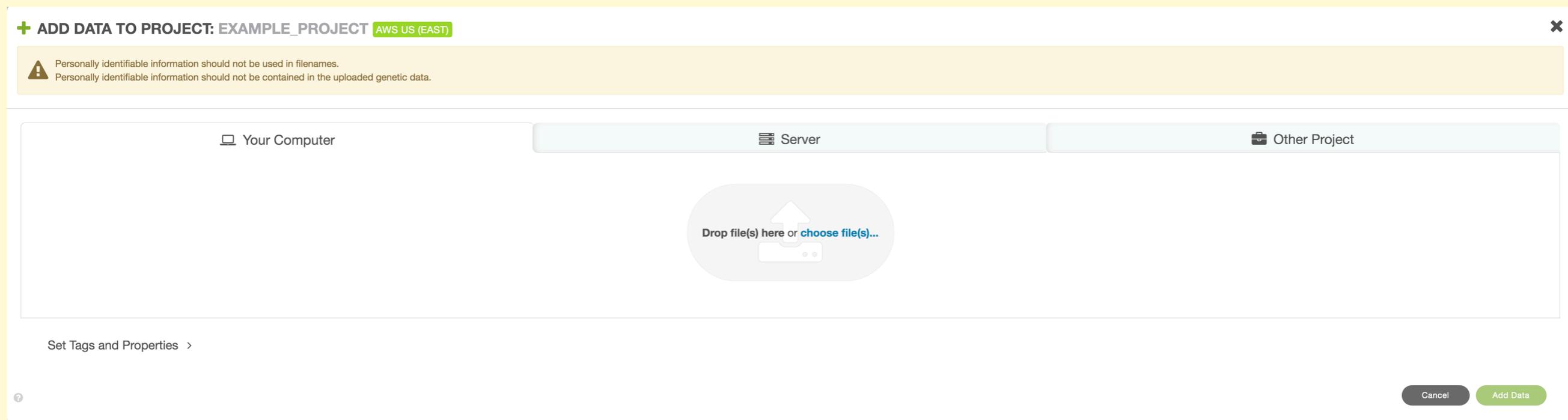
Sign into DNAnexus

Create a project (click on New Project)

Name your project



Add your fastq.gz files



# Add the ENCODE ChIP-seq workflows:

**+ ADD DATA TO PROJECT: EXAMPLE\_PROJECT** AWS US (EAST) ✕

Personally identifiable information should not be used in filenames.  
Personally identifiable information should not be contained in the uploaded genetic data.

Your Computer | Server | Other Project

All Projects

\* Only Non-PHI projects in the same region as the project are shown.

Project Name	Viewer	Count	Size
GTable Reference Data [deprecated]	<input type="button" value="VIEWER"/>	1	0.00 GiB
GTable Exome Pipeline Demo [deprecated]	<input type="button" value="VIEWER"/>	1	0.00 GiB
GTable Wiki Demo [deprecated]	<input type="button" value="VIEWER"/>	1	0.00 GiB
Exome Analysis Demo	<input type="button" value="VIEWER"/>	3	9.81 GiB
<b>ENCODE Uniform Processing Pipelines</b>	<input type="button" value="VIEWER"/>	5	757.10 GiB
Broad Inst Viral NGS	<input type="button" value="VIEWER"/>	1	0.23 GiB

**+ ADD DATA TO PROJECT: EXAMPLE\_PROJECT** AWS US (EAST) ✕

Personally identifiable information should not be used in filenames.  
Personally identifiable information should not be contained in the uploaded genetic data.

Your Computer | Server | Other Project

All Projects / **ENCODE Uniform Processing Pipelines**

**Folders**

- ENCODE Uniform Processing Pipelines
  - .Applet\_archive
  - ATAC-seq
  - ChIP-seq
  - ChIP-seq2
  - data
  - Demo
  - Deprecated
  - DNase-seq
  - long-RNA-seq
  - pipeline-genome-data
  - pipeline-test-samples
  - rampage
  - Reference Files
  - RNA-seq
  - small-RNA-seq
  - WG Bisulfite (Methylation)

Name	Type	Size	Created
.Applet_archive	Folder		
ATAC-seq	Folder		
<input checked="" type="checkbox"/> ChIP-seq	Folder		
ChIP-seq2	Folder		
data	Folder		
Demo	Folder		
Deprecated	Folder		
DNase-seq	Folder		
long-RNA-seq	Folder		
pipeline-genome-data	Folder		
pipeline-test-samples	Folder		
rampage	Folder		
Reference Files	Folder		
RNA-seq	Folder		
small-RNA-seq	Folder		
WG Bisulfite (Methylation)	Folder		

1 Item Selected

# Step 2. Launch the workflow

Go to ChIP-seq, workflows, select genome version (e.g. GRCh38)

The screenshot shows the DNAnexus web interface for a project named 'EXAMPLE\_PROJECT'. The breadcrumb navigation is 'EXAMPLE\_PROJECT > ChIP-seq > workflows'. A table lists the following items:

Name	Type	Size	Created
GRCh38	Folder		
hg19	Folder		
mm10	Folder		
ENCODE histone ChIP-seq (specify reference)	Workflow	-	Apr 4, 2019 5:03 PM

The 'GRCh38' folder is highlighted with a red box.

Select the workflow that best suits your experiment  
(in this example, we have a single replicate of 1 case, 1 control)

The screenshot shows the DNAnexus web interface for the 'GRCh38' folder. The breadcrumb navigation is 'EXAMPLE\_PROJECT > ChIP-seq > workflows > GRCh38'. A table lists the following workflows:

Name	Type	Size	Created
ENCODE histone ChIP-seq (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE histone ChIP-seq Unary Control (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE histone ChIP-seq Unary Control Unreplicated (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE histone ChIP-seq Unreplicated (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE TF ChIP-seq (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE TF ChIP-seq Unary Control (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE TF ChIP-seq Unary Control Unreplicated (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM
ENCODE TF ChIP-seq Unreplicated (GRCh38)	Workflow	-	Apr 4, 2019 5:03 PM

The workflow 'ENCODE TF ChIP-seq Unary Control Unreplicated (GRCh38)' is highlighted with a red box.

# Add data (.fastq.gz) by clicking the appropriate box

**RUN "TF CHIP-SEQ" AS ANALYSIS**  
View job progress in the [Monitor](#) tab. Modifications to an existing workflow won't be saved. Try the new batch tool runner beta!

ENCORE TF ChIP-seq Unary Control Unreplicated (GRCh38) 2 apps unconfigured | 8 apps configured | Workflow Actions | Readme | Run as Analysis... | Settings

Inputs	App	Outputs
<input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Forward reads (SE or PE) <input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Reverse reads (PE) <input type="text" value="GCA_000001405.15..."/> Reference files for bwa	Map Rep1 (applet) set inputs	Mapped reads   Mapping statistics   True if input reads were PE   Number of mapped reads
<input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Forward reads (SE or PE) <input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Reverse reads (PE) via Map Rep1 Reference files...	Map Ctl1 (applet) set inputs	Mapped reads   Mapping statistics   True if input reads were PE   Number of mapped reads

# Set the output folder

**RUN "TF CHIP-SEQ" AS ANALYSIS**  
View job progress in the [Monitor](#) tab. Modifications to an existing workflow won't be saved. Try the new batch tool runner beta!

ENCORE TF ChIP-seq Unary Control Unreplicated (GRCh38) 10 apps configured | Workflow Actions | Readme | Run as Analysis... | Settings

Inputs	App	Outputs
<input type="text" value="chr21.CTCF_A549_r..."/> Forward reads (SE or PE) <input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Reverse reads (PE) <input type="text" value="GCA_000001405.15..."/> Reference files for bwa	Map Rep1 (applet) runnable	Mapped reads   Mapping statistics   True if input reads were PE   Number of mapped reads
<input type="text" value="chr21.ctrl_A549_rep..."/> Forward reads (SE or PE) <input type="text" value="*.fastq *.fastq.gz *.fq *.fq.gz"/> Reverse reads (PE) via Map Rep1 Reference files...	Map Ctl1 (applet) runnable	Mapped reads   Mapping statistics   True if input reads were PE   Number of mapped reads
via Map Rep1 Mapped reads	Filter_QC Rep1 (applet) runnable	Mapped reads surviving th...   Indexed reads surviving th...   Post-filtering mapping stati...   Duplication metrics from M...

- Update workflow with changes
- Set output folder**
- Edit template
- Save as template
- Copy workflow to project...

Create new folder by clicking the icon, select the newly created folder then click "add"



**SELECT OUTPUT FOLDERS**

EXAMPLE\_PROJECT

Folders

- EXAMPLE\_PROJECT
  - ChIP-seq

OUTPUT | + | x

ChIP-seq

# Click "Run as analysis"

**RUN "TF CHIP-SEQ" AS ANALYSIS**  
View job progress in the Monitor tab. Modifications to an existing workflow won't be saved.

10 apps configured | /OUTPUT | Workflow Actions | Readme | **Run as Analysis...**

**ENCODE TF ChIP-seq Unary Control Unreplicated (GRCh38)**

Inputs: chr21.CTCF\_A549\_r... Forward reads (SE or PE), \*.fastq \*.fastq.gz \*.fq \*.fq.gz Reverse reads (PE), GCA\_000001405.15... Reference files for bwa

App: Map Rep1 (applet) - runnable

Outputs: Mapped reads, Mapping statistics, True if input reads were PE, Number of mapped reads

Inputs: chr21.ctrl\_A549\_rep... Forward reads (SE or PE), \*.fastq \*.fastq.gz \*.fq \*.fq.gz Reverse reads (PE), via Map Rep1 Reference files...

App: Map Ctrl1 (applet) - runnable

Outputs: Mapped reads, Mapping statistics, True if input reads were PE, Number of mapped reads

## Step 3. View results

Results will be in output directory set in Step 2.

**EXAMPLE\_PROJECT** Settings Manage Monitor Visualize

+ Add Data | New Folder | New Workflow | Start Analysis

SEARCH SCOPE: Entire project | ID: Any | TYPES: Any | MODIFIED: Any | TAGS: Any

EXAMPLE\_PROJECT > OUTPUT

Name	Type	Size
encode_mac2	Folder	
encode_map	Folder	
encode_spp	Folder	
idr2	Folder	

EXAMPLE\_PROJECT

- ChIP-seq
- OUTPUT**

← Signal files (.bigwig)  
← Mapped reads  
← ChIP-seq QC plots  
← Final peak coordinates (.bb, narrowPeak)

## Workflow comparison: Run time and cost

Typical ChIP-seq experiment

Transcription factor (TF) : 10-20M reads/replicate

Narrow chromatin marks: >20M reads/replicate

Broad chromatin marks: min 40-50M reads/replicate (ENCODE says >45M), more for primary cells/tissues

Pipelines suggest min 50bp reads, longer for broad peak detection

Guidelines from: Jung et al., NAR 2014 and the ENCODE DCC

Factor	Peak type	Sequencing run parameters	# reads	St. Jude		ENCODE	
				Time	Cost	Time	Cost
chr21_CTCF*	Narrow	SE,50bp	100K IP, 100K ctrl	20m	\$0.08	47m	\$0.49
CTCF	Narrow	SE,50bp	21M IP, 21M ctrl	1h 21m	\$0.84	6h 43m	\$7.01
H3K27me3	Broad	SE,75bp	67M IP, 60M ctrl	5h 23m	\$3.53	7h 5m	\$5.61

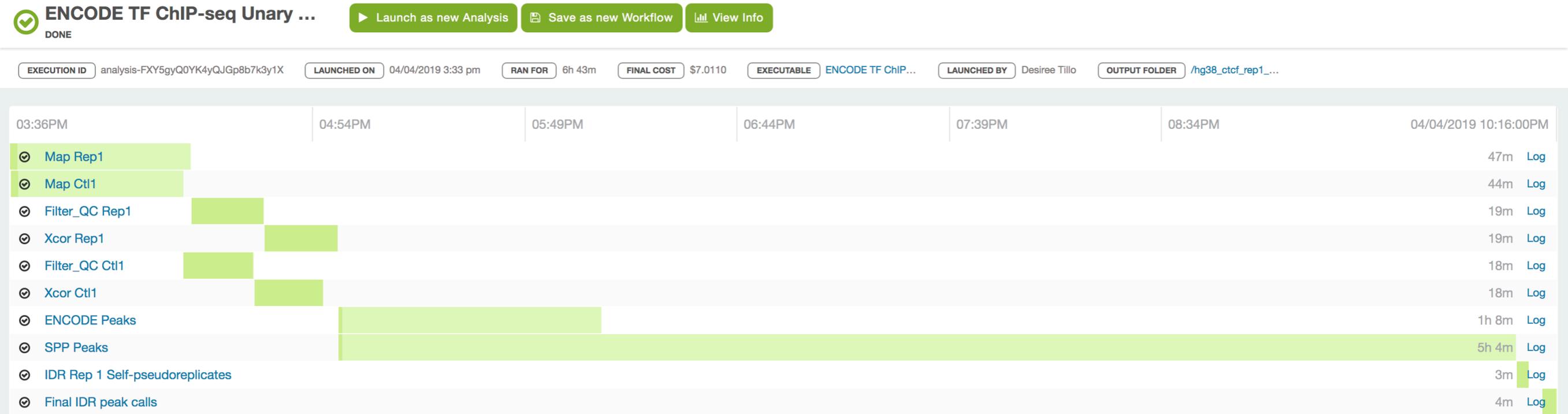
# Workflow Comparison

	St. Jude	ENCODE
Mapping	bwa	bwa
Signal output	Yes: coverage for each input fastq	Yes: Fold enrichment/control, p-values
Narrow peak calling (e.g. transcription factors, H3K4me2/3, H3K27Ac)	macs2	spp for peak calling, IDR for thresholding
Broad peak (certain chromatin modifications, H3K36me3, H3K27me3, H3K9me1/2)	SICER	macs2
QC	sequencing quality: fastQC ChIP-Seq quality: quality:spp/ phantompeakqualtools	ChIP-Seq quality: spp/ phantompeakqualtools, IDR (peak reproducibility)
Supported genomes	human, mouse, drosophila	human, mouse, custom
Paired-end?	no	yes
Combining replicates?	no	yes (concatenates fastqs)
Cost for typical experiment	\$1	\$7
Time for typical experiment	~1 hour	~6 hours

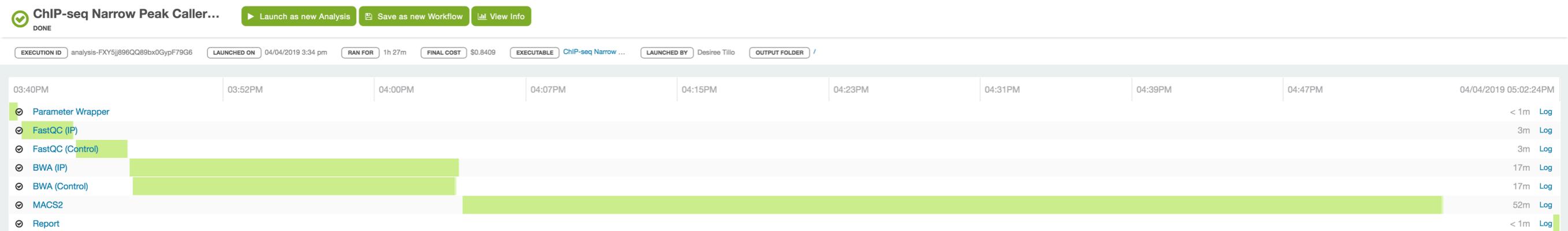
# Biggest bottleneck at the second peak calling stage (SPP)

Same input (CTCF\_rep1)

## ENCODE



## St. Jude



# DNA-SEQ

*Peter FitzGerald*

Head, Genome Analysis Unit

Overview of Available Work Flows

# Sentieon

## *Various Programs and Pricing*

Use this app to identify variants (SNPs and indels) using a pipeline that reimplements the GATK 3.x, MuTect, and MuTect2 best practices method.

Sentieon DNaseq FASTQ to VCF	<b>Sample Type</b>	<b>Price/Sample</b>
Sentieon DNaseq FASTQ to VCF (Panel)	Panel (fastq.gz $\leq$ 5 GiB)	\$3.30
Sentieon DNaseq FASTQ to VCF (WES) <small>Use this app to identify variants (SNPs and indels) using a pipeline that reimplements the GATK 3.x, MuTect, and MuTect2 best practices method.</small>	WES (5 GiB < fastq.gz $\leq$ 20 GiB)	\$9.90
Sentieon DNaseq FASTQ to VCF (WGS)	WGS (20 GiB < fastq.gz $\leq$ 90 GiB)	\$47.00
Sentieon DNaseq FASTQ to VCF (WGS2)	WGS2 (90 GiB < fastq.gz $\leq$ 200 GiB)	\$70.00 to ~\$100.00
Sentieon TNseq FASTQ to VCF	<b>Sample Type</b>	<b>Price/Sample</b>
Sentieon TNseq FASTQ to VCF (WES)	WGS (40 GiB < fastq.gz $\leq$ 300GiB)	\$80 to ~\$130
Sentieon TNseq FASTQ to VCF (WGS)	WES (fastq.gz $\leq$ 40GiB)	\$23 to ~\$30



PROJECTS

TOOLS ▾

ORG ADMIN ▾

HELP ▾

# My RNASEQ Project

Settings

+ Add Data

New Folder

New Workflow

Start Analysis

### SELECT A TOOL

The app or applet will be added to your workflow

Filter tools

- FASTQ Read Trimmer v1.4.0** - Read Manipulation  
\*.FQ.GZ... \*.FQ.GZ... > \*.FQ.GZ \*.FQ.GZ
- Gzip File Compressor v1.1.2** - File Manipulation  
\*.GZ
- Parliament2 v0.1.9**  
\*.BAM... \*.BAI \*.FA.GZ... \*.FAI > \*.VCF \*.TAR.GZ \*.LOG
- QIAGEN's Ingenuity Variant Analysis™ v1.0.6**  
\*.VCF... \*.TSV \*.BAM \*.BAI >
- Salmon (St. Jude) v1.0.0**  
\*.FA.GZ \*.FASTQ.GZ \*.FASTQ.GZ >
- SAMtools Mappings Indexer v1.1.2** - Mappings Manipulation  
\*.BAM > \*.BAI
- Sentieon TNseq FASTQ to VCF v0.6.0** - Variant Calling  
\*.FQ.GZ... \*.FQ.GZ... \*.TXT \*.VCF.GZ \*.VCF.GZ.TBI \*.VCF.GZ \*.VCF.GZ.TBI  
\*.FQ.GZ... \*.FQ.GZ... \*.VCF.GZ \*.VCF.GZ.TBI \*.BAM \*.BAI  
\*.VCF.GZ \*.VCF.GZ \*.CSV > \*.DUPLICATION\_METRICS... \*.BAM \*.BAI \*.TABLE  
\*.BWA-INDEX.TAR.GZ \*.FA\*.GZ \*.BAM \*.BAI \*.BAM \*.BAI \*.DUPLICATION\_METRICS...  
\*.RESOURCE.BUNDLE.TAR.GZ \*.BAM \*.BAI \*.BAM \*.BAI \*.TABLE \*.BAM  
\*.BED \*.BAI
- STAR Mapping v1.3.3** - Read Mapping  
\*.STAR-INDEX.TAR.GZ \*.FASTQ... \*.FASTQ... > \*.BAM \*.BAI \*.LOG.GZ \*.SJ.OUT.TSV.GZ...  
\*.VCF... \*.TSV.GZ... \*.GTF... \*.BAM \*.STAR-INDEX.TAR.GZ
- STAR RNA Alignment (St. Jude) v1.1.1**  
\*.FASTQ.GZ \*.FASTQ.GZ > \*.BAM \*.BAM.BAI

# My RNASEQ Project

Settings

Manage

Monitor

Visualize

Access: Admin

1 Share

Untitled Workflow - Apr 10th 2019 4:00pm

1 app  
unconfigured

Workflow Actions

Readme

Autosaved

Start Analysis...

Inputs

App

Outputs

\*.fq.gz \*.fastq.gz Tumor Reads [array]

\*.fq.gz \*.fastq.gz Tumor Reads (right mates) [a...]

\*.fq.gz \*.fastq.gz Normal Reads [array]

\*.fq.gz \*.fastq.gz Normal Reads (right mates) [...]

Sentieon TNseq FASTQ...

set inputs and params

\*.txt TNsnv: Call-stats

\*.vcf.gz TNsnv: Variants

\*.vcf.gz.tbi TNsnv: Variants VCF index

\*.vcf.gz TNhaplotyper: Variants

About this app

+ Add a Step

# RUN "SENTIEON" AS ANALYSIS

View job progress in the [Monitor](#) tab. Modifications to an existing workflow won't be sav...

[Try the new batch tool runner beta!](#)

Sentieon

1 app  
unconfigured

Workflow Actions ▾

▶ Run as Analysis...



Inputs

App

Outputs

\*.fq.gz \*.fastq.gz **Tumor Reads [array]**

\*.fq.gz \*.fastq.gz Tumor Reads (right mates) [a...

\*.fq.gz \*.fastq.gz Normal Reads [array]

\*.fq.gz \*.fastq.gz Normal Reads (right mates) [...

\*.vcf.gz Panel of Normal

\*.vcf.gz COSMIC -- Catalogue of So...

\*.csv Read group information

\*.bwa-index.tar.gz **BWA reference genome index**

\*.fa\*.gz **Reference genome FASTA file**

\*.resource.bundle.tar.gz GATK resource bundle

\*.bed Target coordinates [array]

**Sentieon TNseq FA...**

set inputs and params

\*.txt **TNsnv: Call-stats**

\*.vcf.gz **TNsnv: Variants**

\*.vcf.gz.tbi **TNsnv: Variants VCF index**

\*.vcf.gz **TNhaplotyper: Variants**



[About this app](#)

Close



# SELECT DATA FOR REFERENCE GENOME FASTA FILE INPUT



Sentieon TNseq FASTQ to VCF

All Projects / **Reference Genome Files: AWS US (East)**



PATTERNS [clear](#)

Files (\*.fa\*.gz)

## Folders

[Show All Folders](#)

Reference Genome Files: AWS US (...)

- C. Elegans - Ce10
- D. melanogaster - Dm3
- H. Sapiens - GRCh37 - b37 (1000 ...)
- H. Sapiens - GRCh37 - hs37d5 (1...
- H. Sapiens - GRCh38 with alt cont...
- H. Sapiens - GRCh38 without alt c...
- H. Sapiens - hg19 (Ion Torrent)
- H. Sapiens - hg19 (UCSC)
- M. musculus - mm9
- M. musculus - mm10

No data available here

## Suggestions

- [DNAnexus Reference Genomes: AWS US-east](#)
- [My RNASEQ Project](#)

Cancel

Select

# Development/Batch Support

DNAnexus Development Environment - Bioinformaticists

*Friday April 12th, 10:00-11:30 am. - NIH Bldg 37, Rm 2041/2107*

- dx-toolkit - command line access
- Development languages (python, bash, docker)
- Applet development
- Cloud workstation application
- Batch processing
- Resource selection and optimization

# CCR/GAU RESOURCES

- Help pages on the Web  
(<https://gau.ccr.cancer.gov/dna-nexus-pilot-program/>)
- Slack Channel for CCR\_DNANexus Pilot ([dnaxpilot.slack.com](https://dnaxpilot.slack.com))  
(help, general, development)
- Custom Built Work Flows (RNASEQ workflow, IGV\_session\_maker, ADAP, *Pausing Peak Aligner\**, *Tumor Mutation Burden\**)
- DNANexus Applications By Category Page  
(<https://dl.dnanex.us/F/D/jpyV|BVZKZ|zf8|IQXfg7X|3P8x|Z4|P7zKVygpX?inline>)
- Management of DNANexus Account, Funding and cost management