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

DNAnexus 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*)

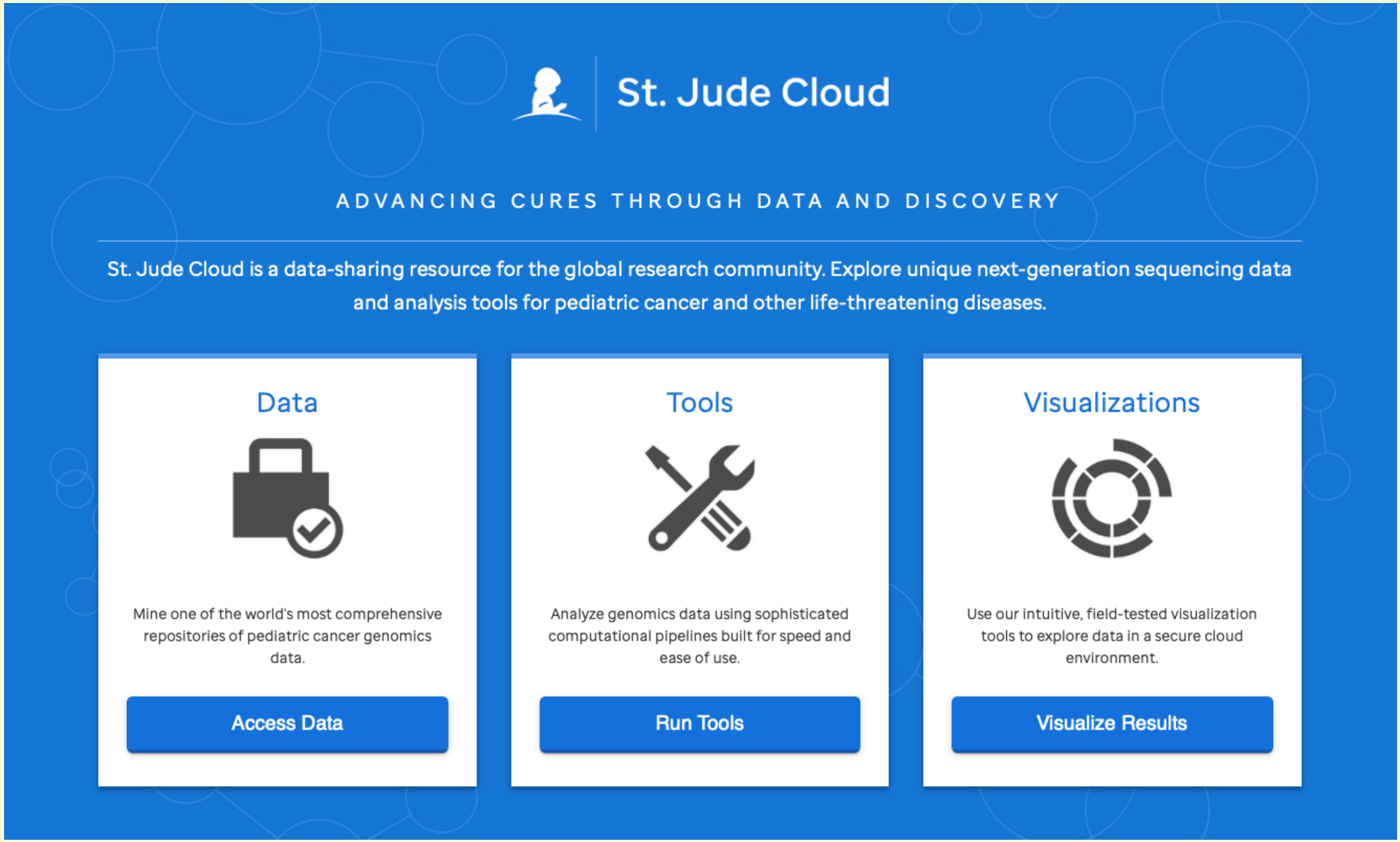
- Use Folders Wisely

- 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 ?



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*)
  (help, general, development)

- Custom Built Work Flows (RNASEQ workflow, IGV_sesssion_maker, ADAP, *Pausing Peak Aligner*, Tumor Mutation Burden**)

- DNAnexus Applications By Category Page
  (https://dl.dnanex.us/F/D/jpyV1BVZKZJzf811QXfg7X13P8x1Z41P7zKVygpX?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)

GENCODE

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	**58870**	Total No of Transcripts	**208621**
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, Szigyarto 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

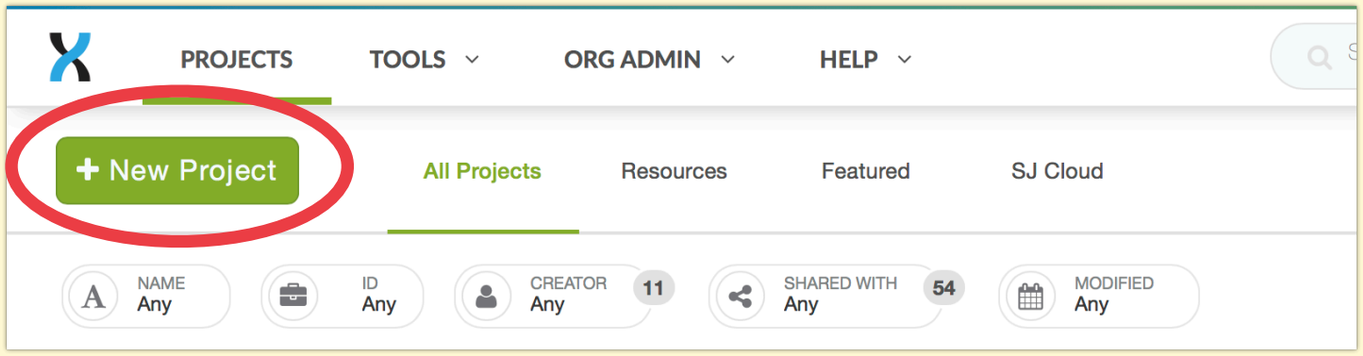
  Brain x3 vs Muscle x3

RNASEQ

The End Point

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

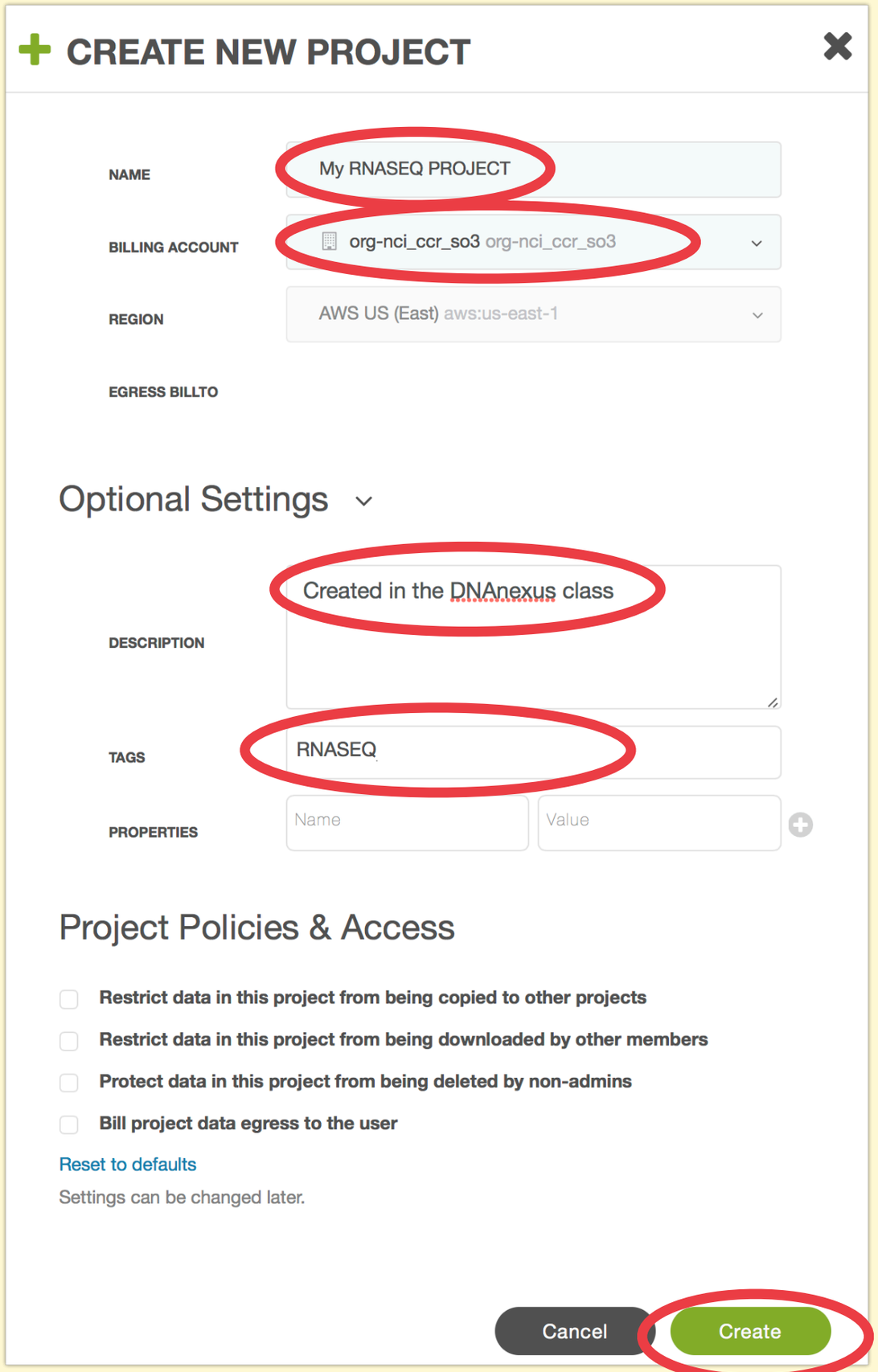
Step 1: Create an new project



Select the New Project button

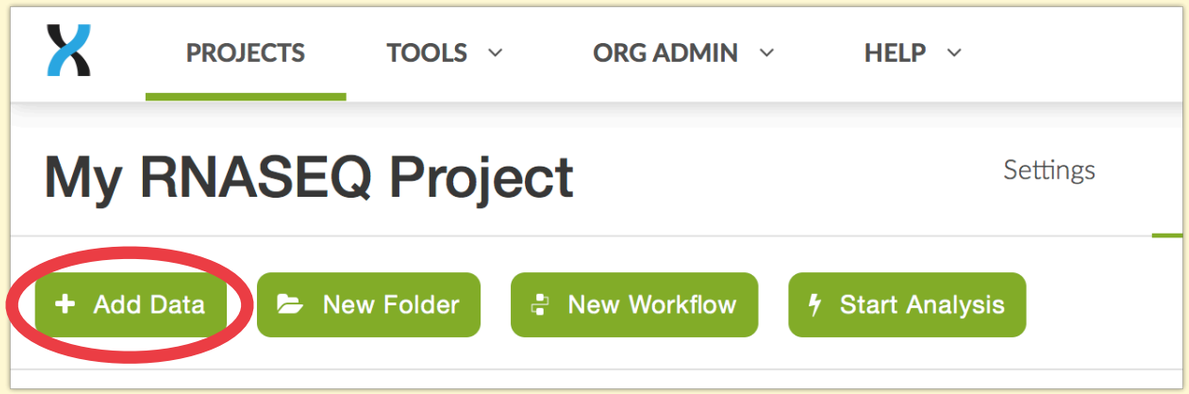
In the pop-up dialogue box fill in the following info:

- 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



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:

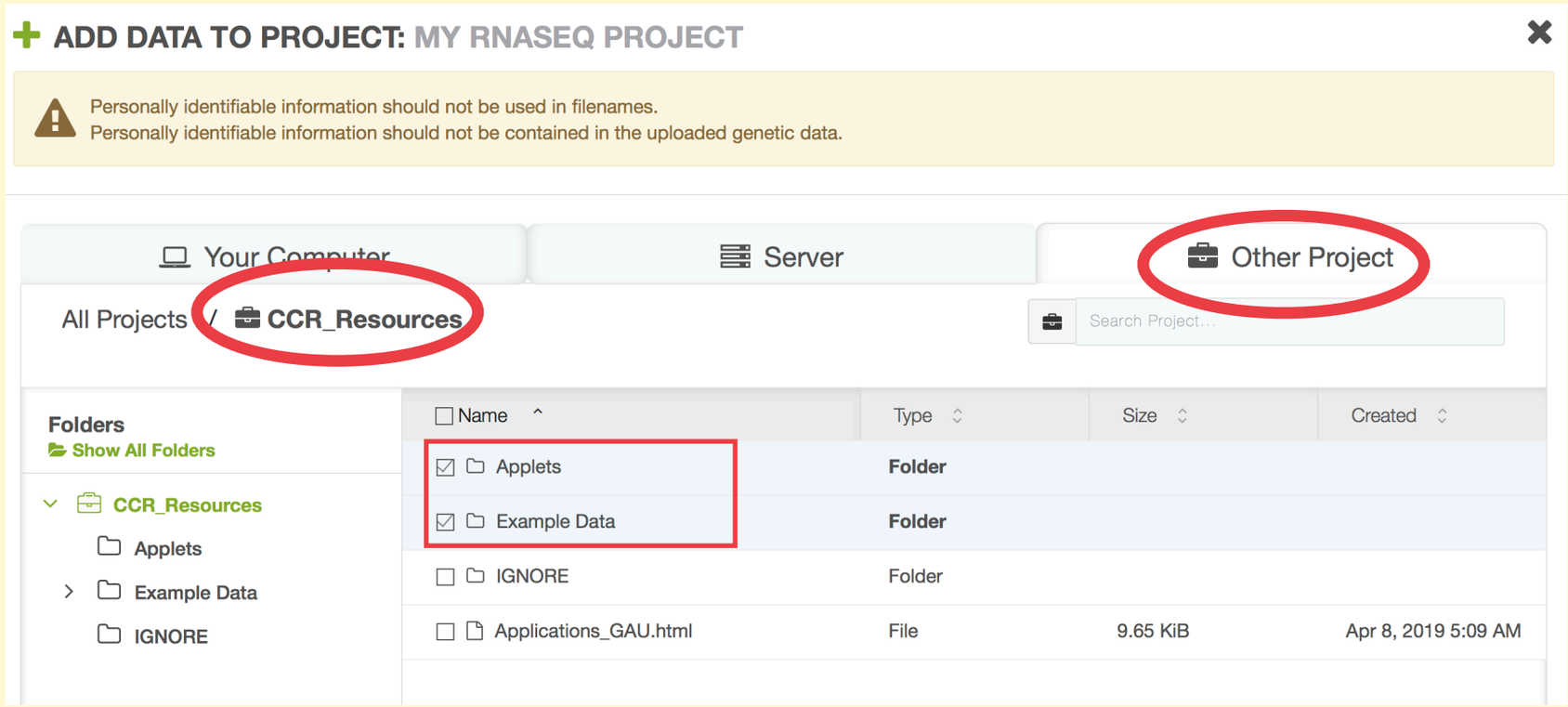
- *Other Project*
- *Project CCR_Resources*
- *Folder CCR_Resources*

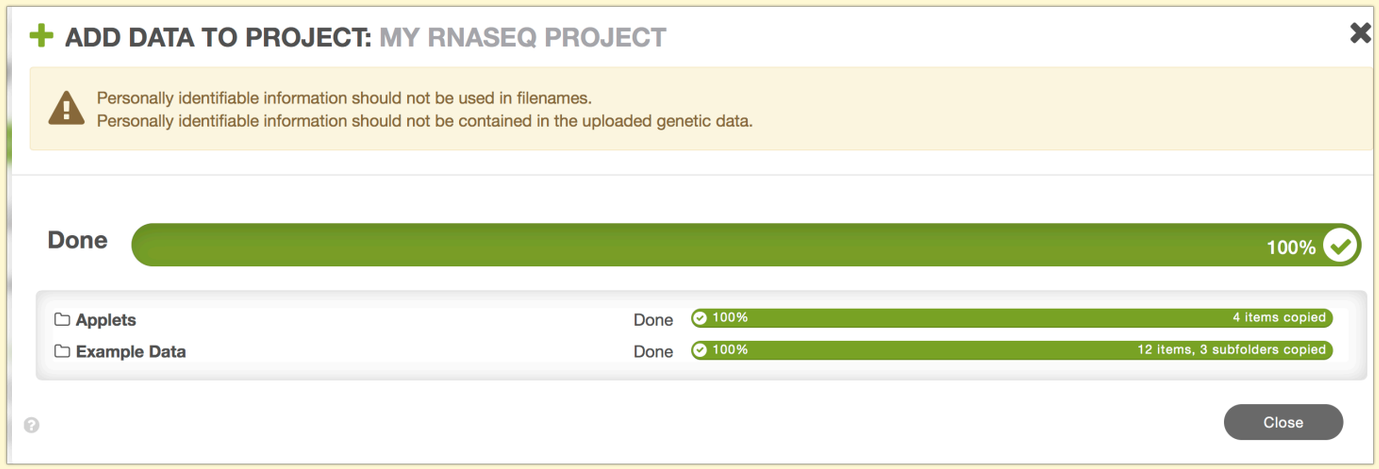
Check the boxes for

- *Applets*
- *Example Data*

Select copy

Close the transfer dialogue

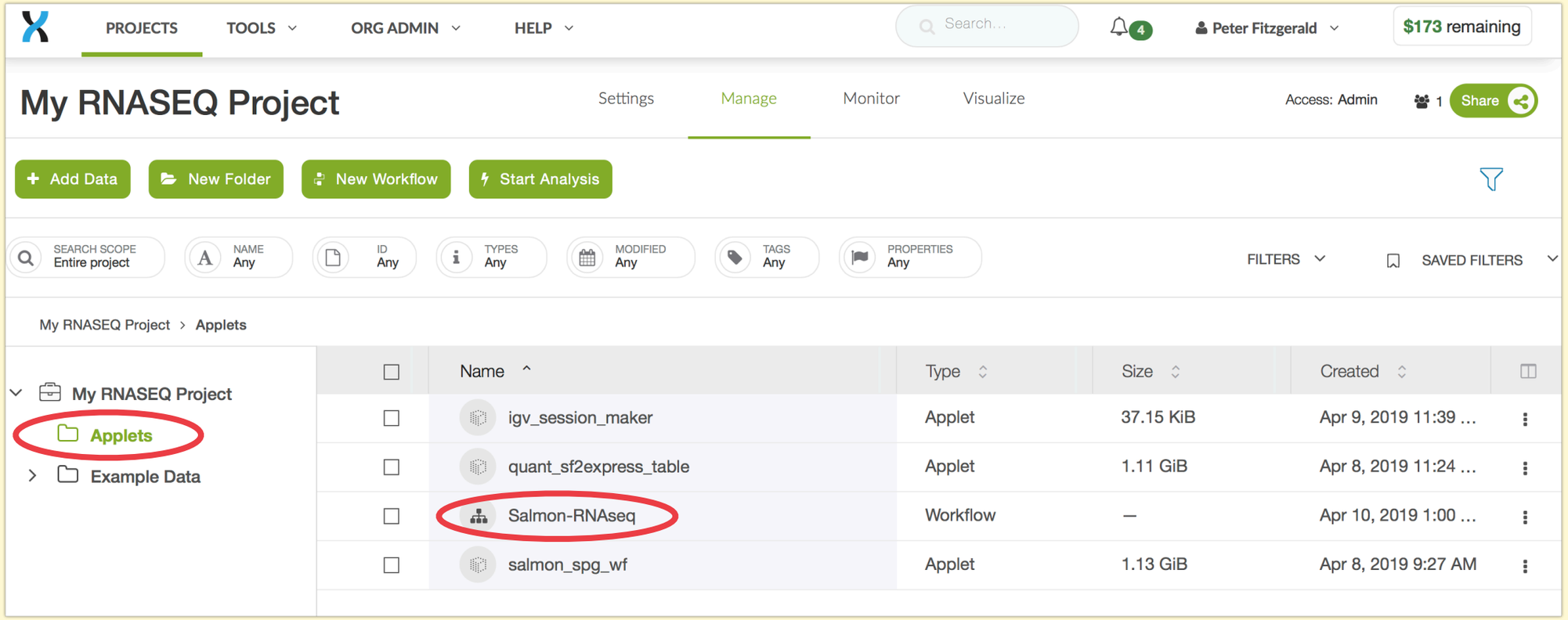




Step 3: Select the workflow and choose files & parameters

The 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



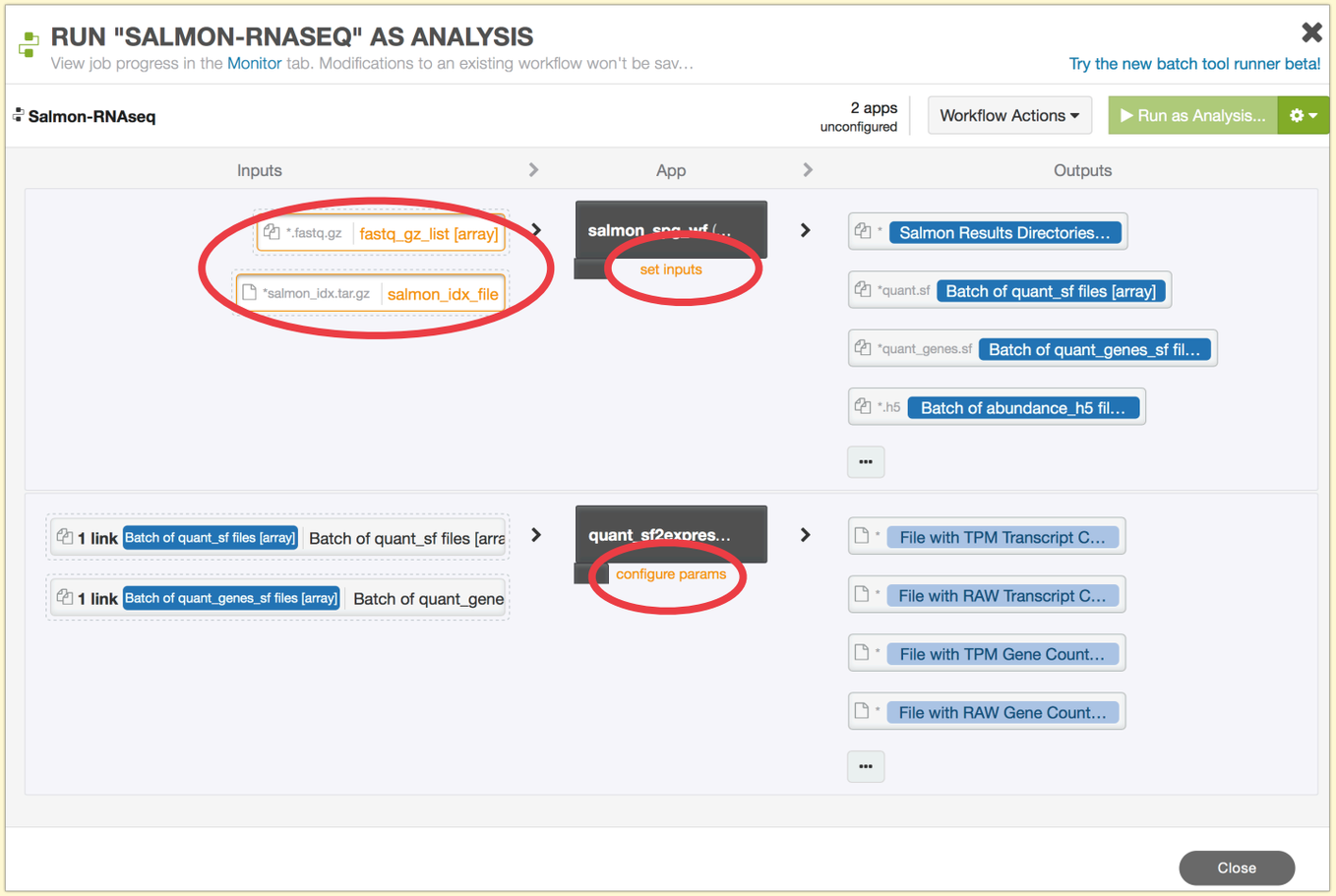
Select the following:

- *Applet folder*
- *Salmon-RNAseq -workflow*

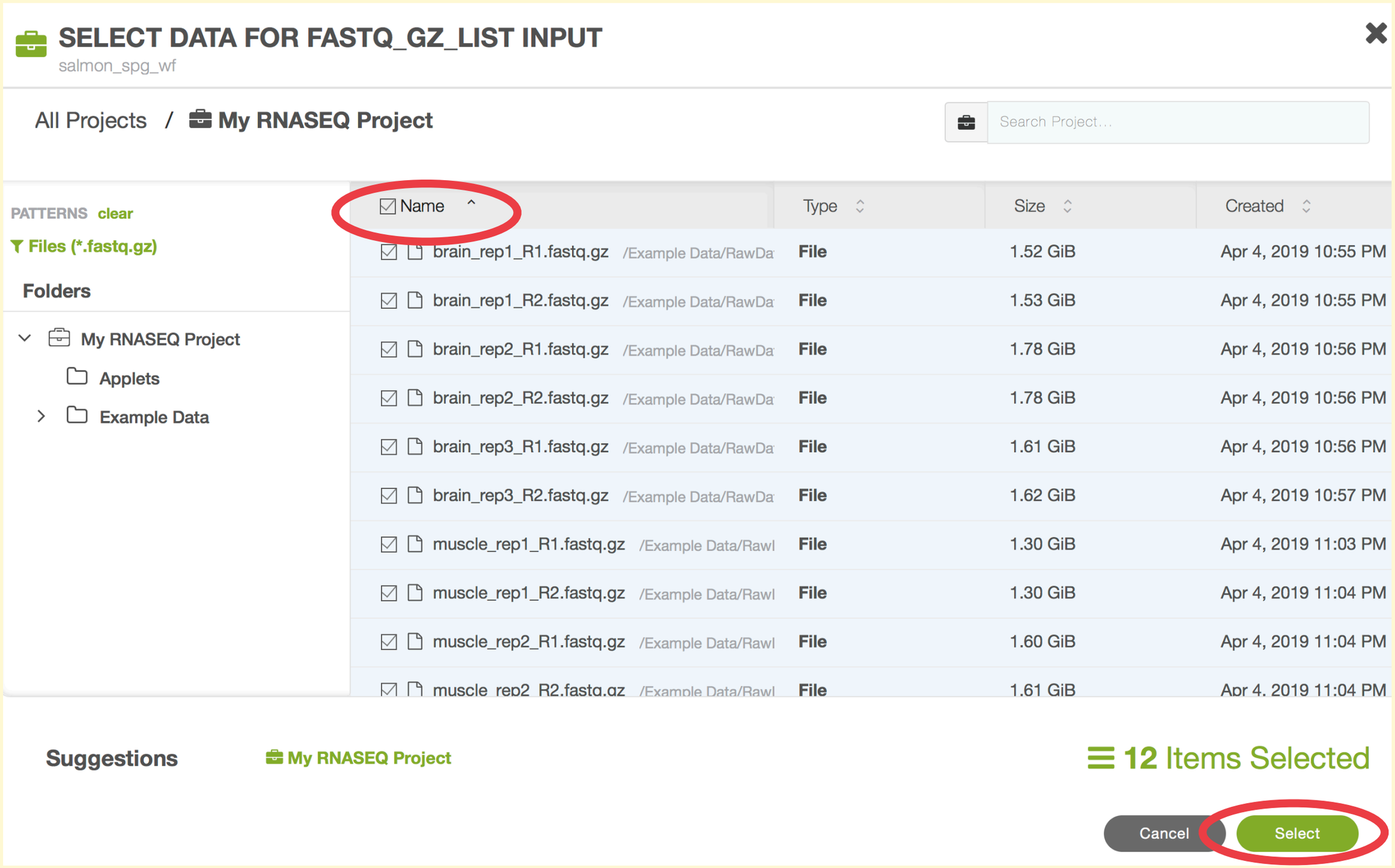
Select the:

- *Example Data*
- *Database*

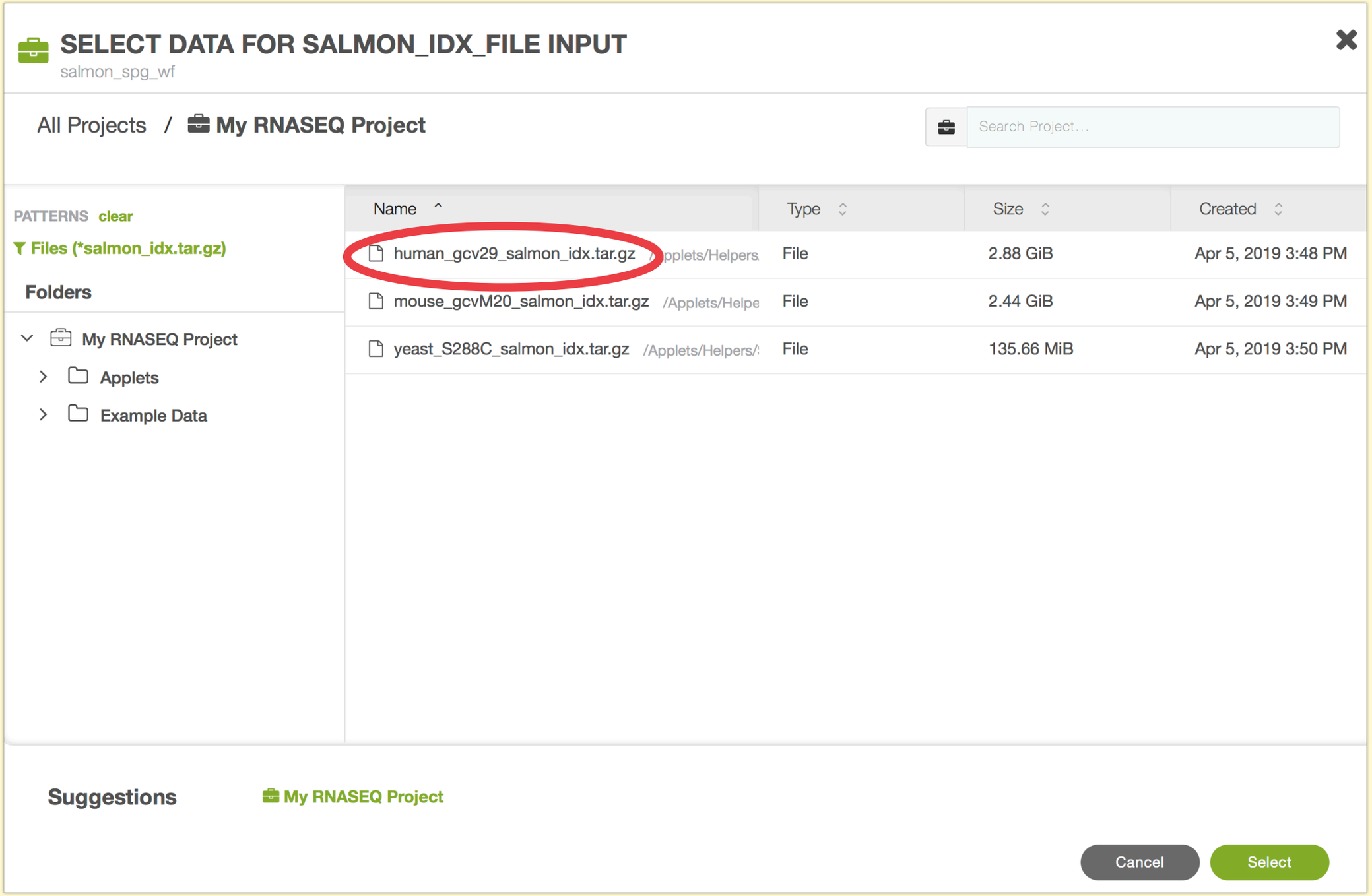
Set the parameters in both applets



Step 4: Selecting the sample data



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



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 developwer'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) and [McIntosh, Carl (NIH/NCI) [E]] (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 salmon_spg_wf *

Output Folder RNASEQ_TEST *

Instance type mem1_ssd1_x4 Select

COMMON

Bootstrap Value 0 *

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) and [McIntosh, Carl (NIH/NCI) [E]] (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, *DNAnexus* 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 quant_sf2express_table *

Output Folder RNASEQ_TEST *

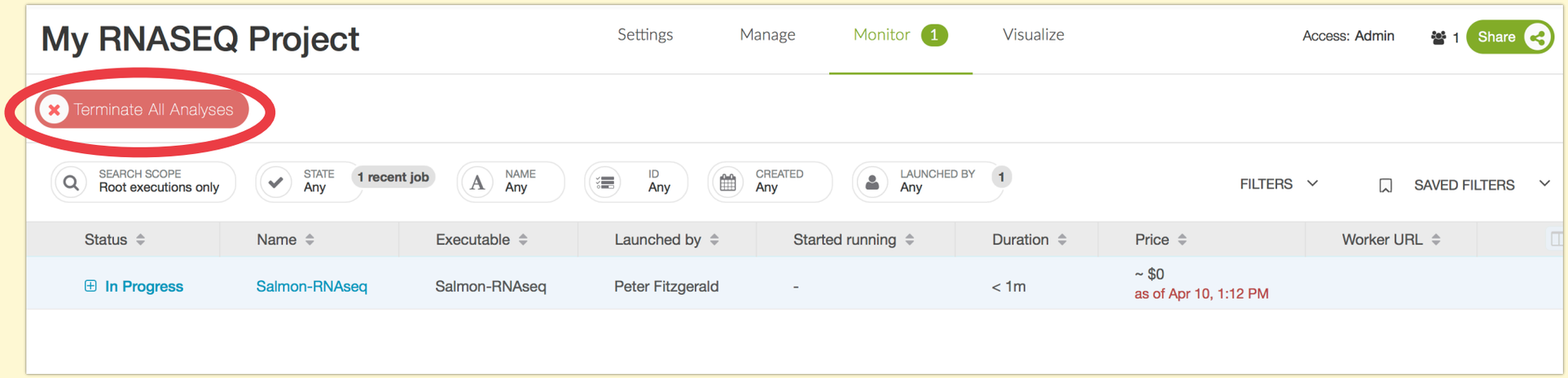
Instance type mem1_ssd1_x4 Select

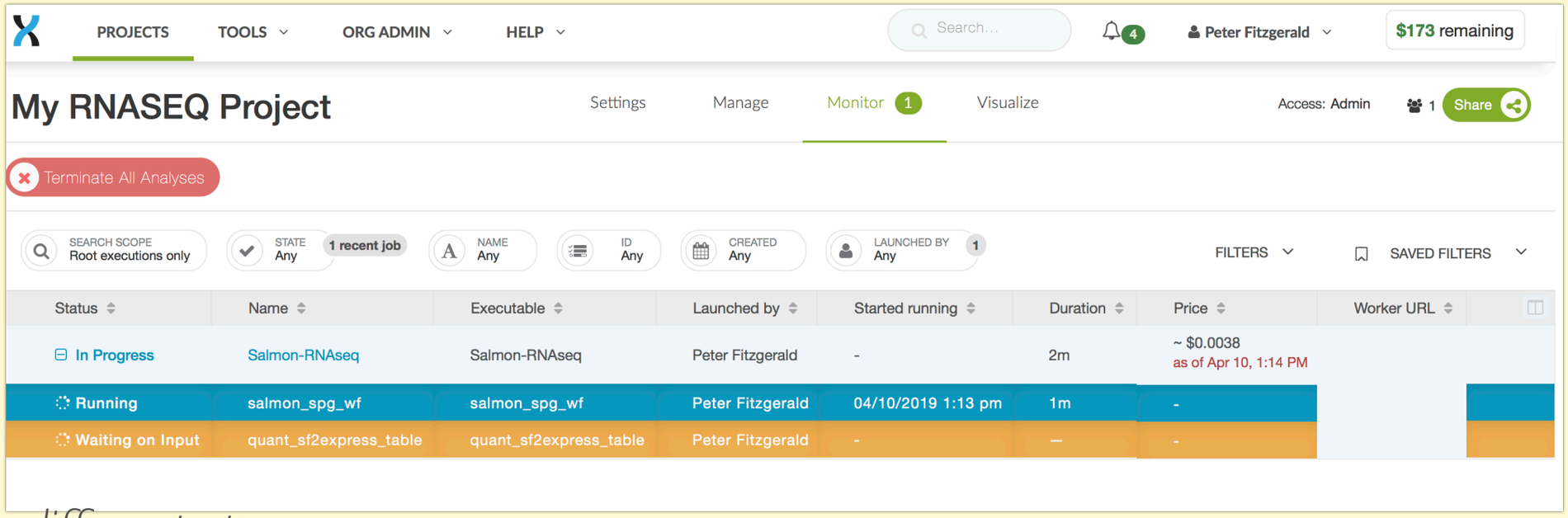
COMMON

prefix brain_muscle *

Reset to applet defaults Save

Step 9: Monitor the Job (or not)





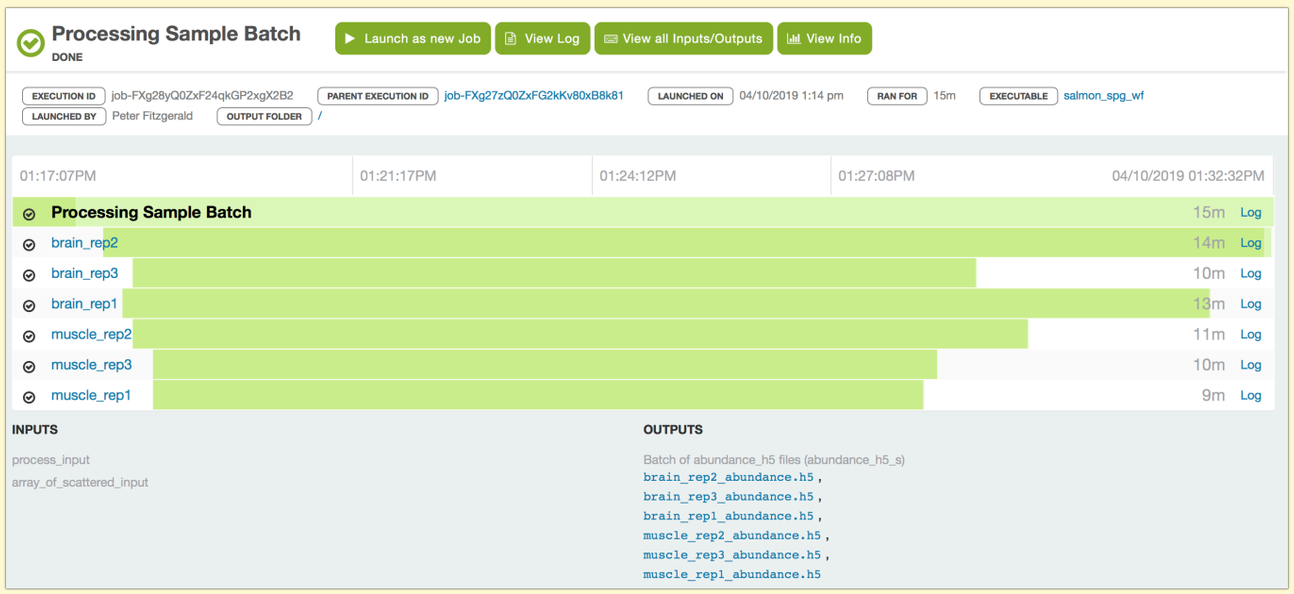
These images show different stages of the process. The job can be terminated at any time by clicking on the Terminate button

States are:

Waiting

Running

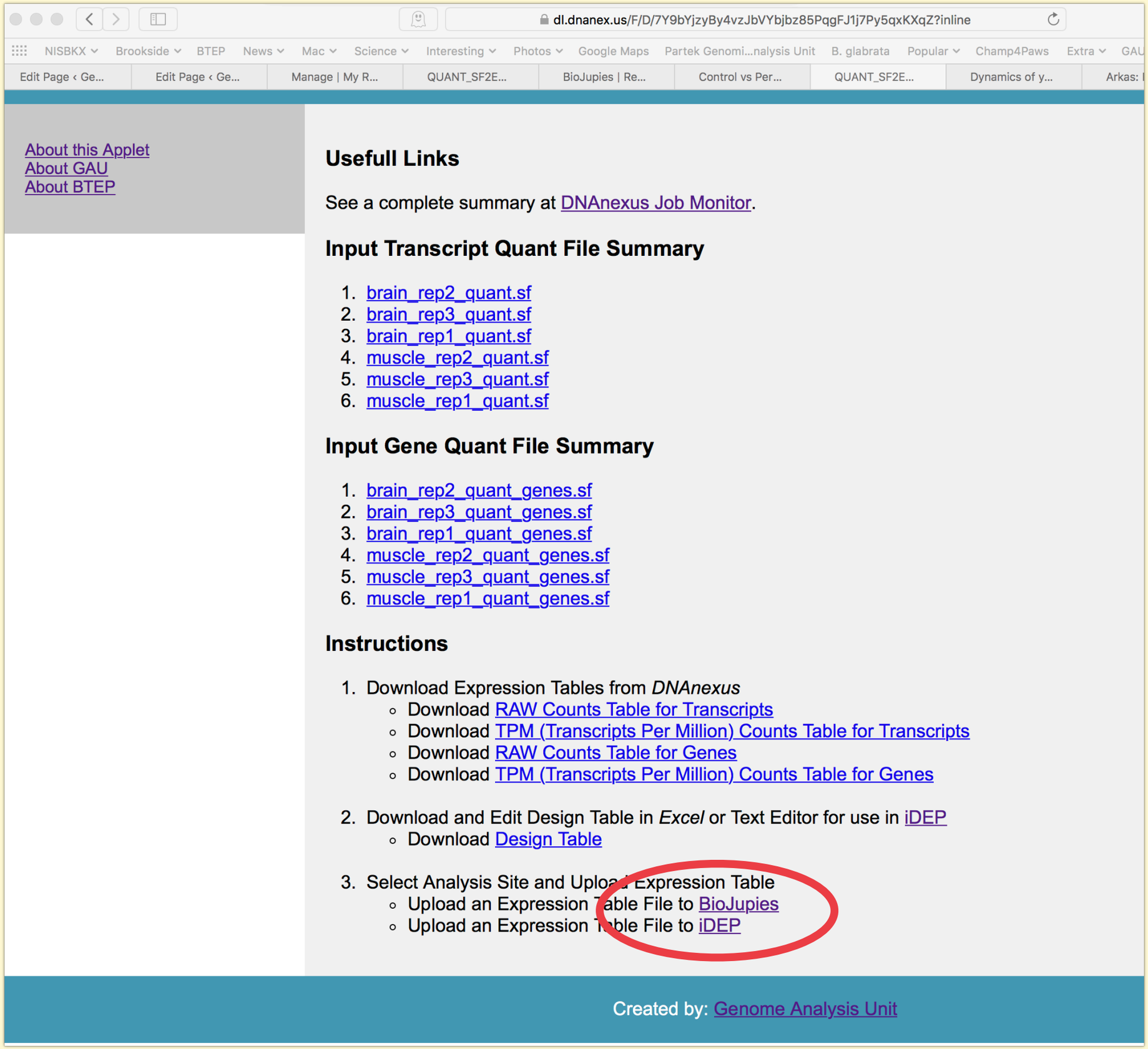
Error



Step 10: The final output from the DNAnexsus 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

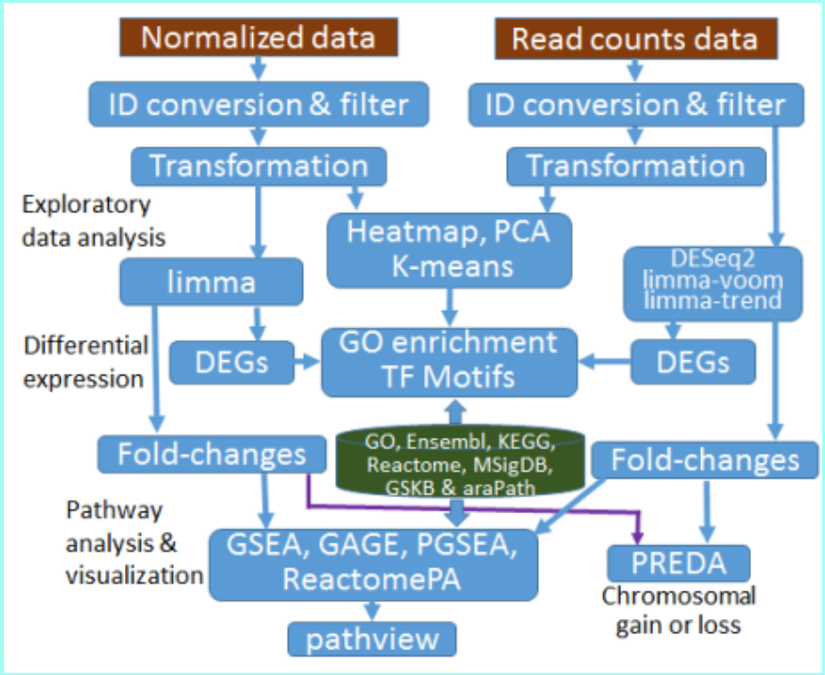


Two excellent Analysis Options - Shiny App Servers (R)

BioJupies

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:

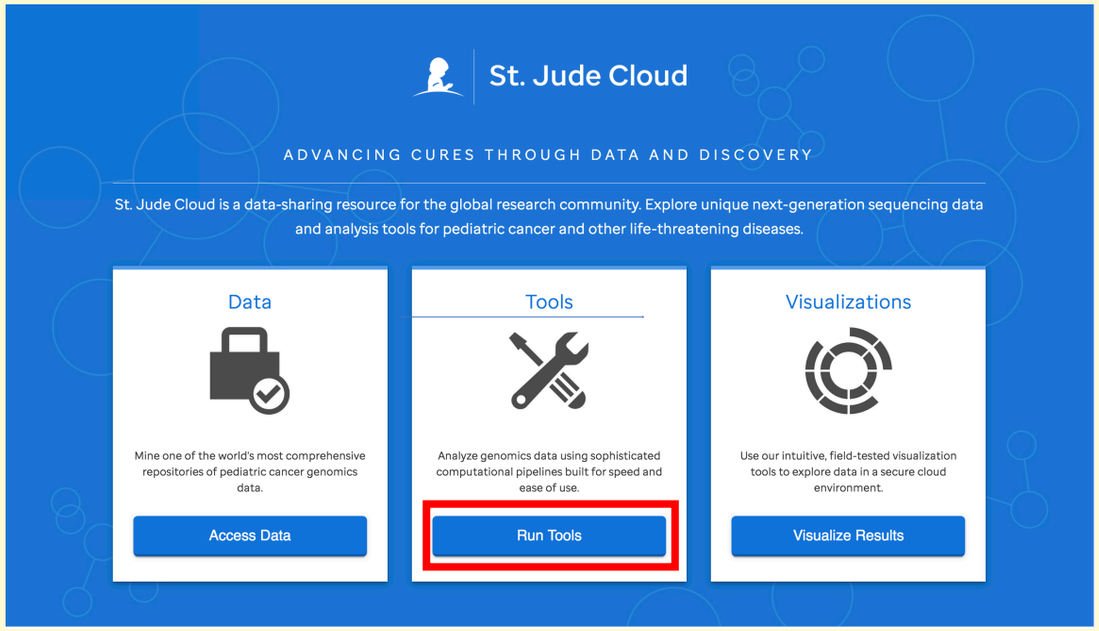
- Peaks as BED (.bed) and big BED (.bb) files.
- Genome coverage files as bigWig (.bw) file for each input fastq.
- QC plots and files:
  - A cross correlation plot (for measuring fragment length and computation of data quality)
  - Sequencing quality metrics (output from fastqc)

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

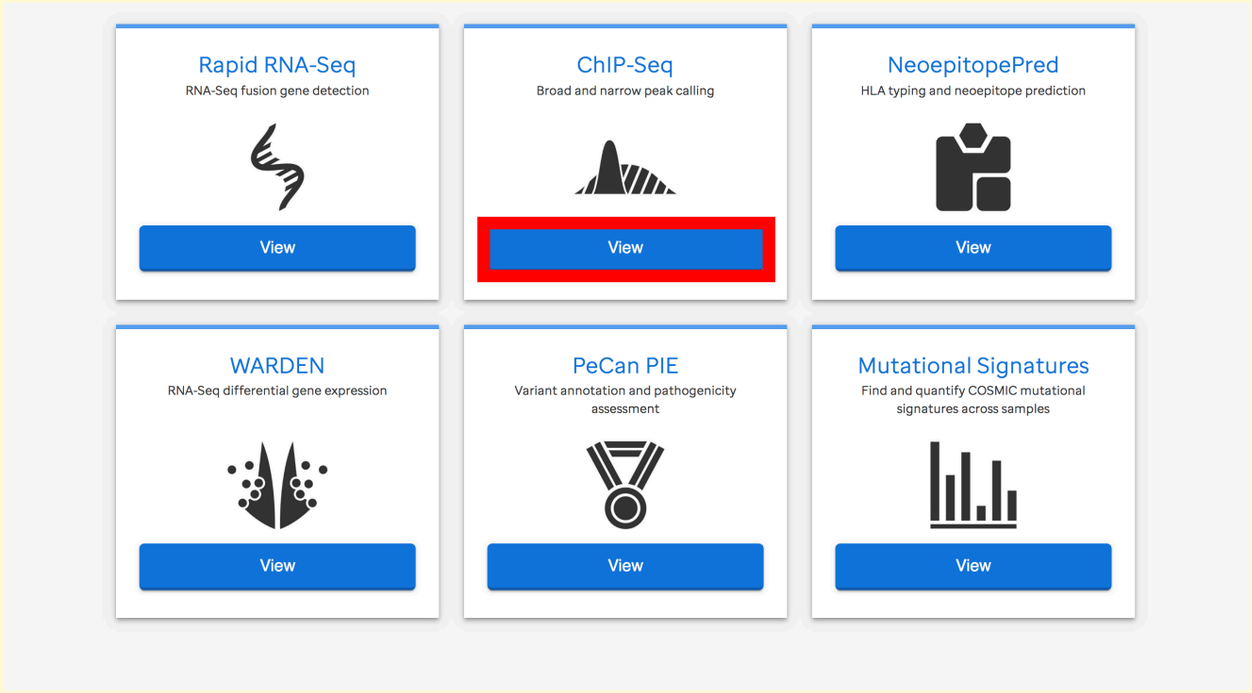
Running ChIP-seq workflows on the St. Jude cloud

Step 1. Set up

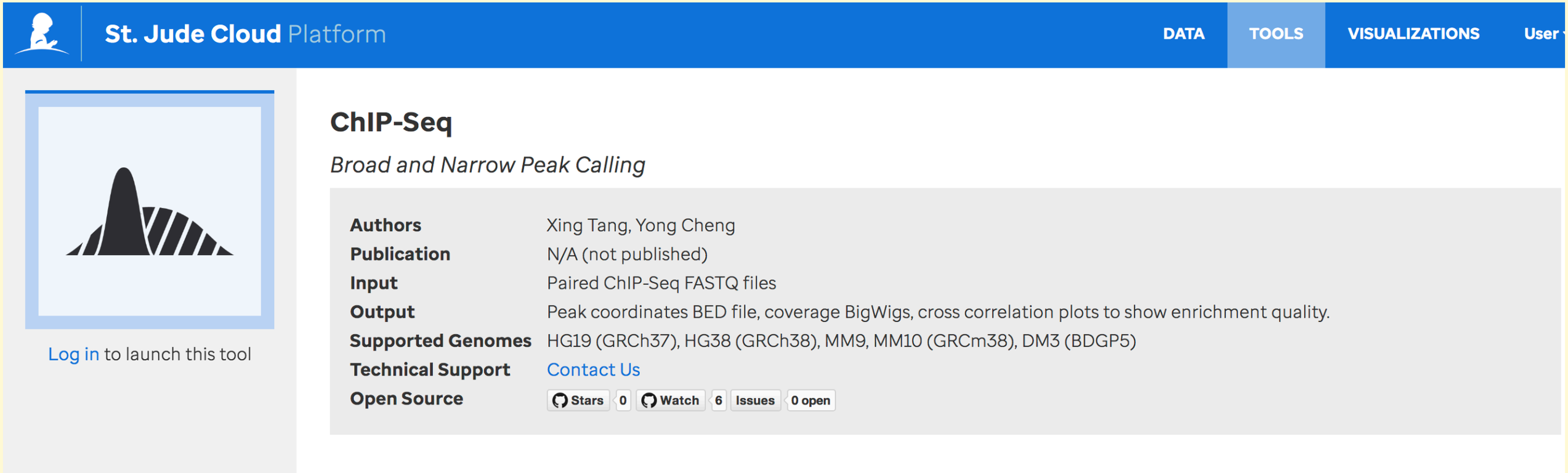
Go to St. Jude Cloud (https://www.stjude.cloud/)
Click on “Run tools”:



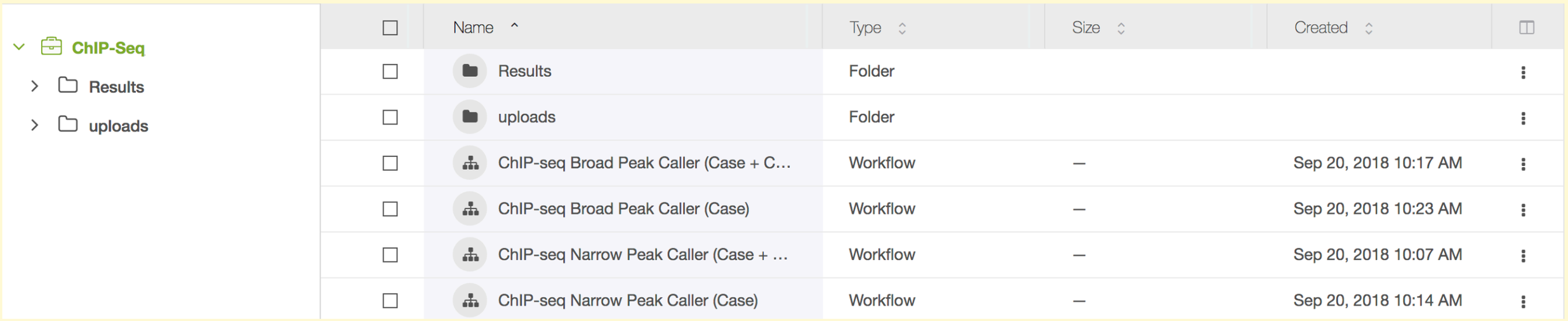
Click on the “View” button under “ChIP-seq”:



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



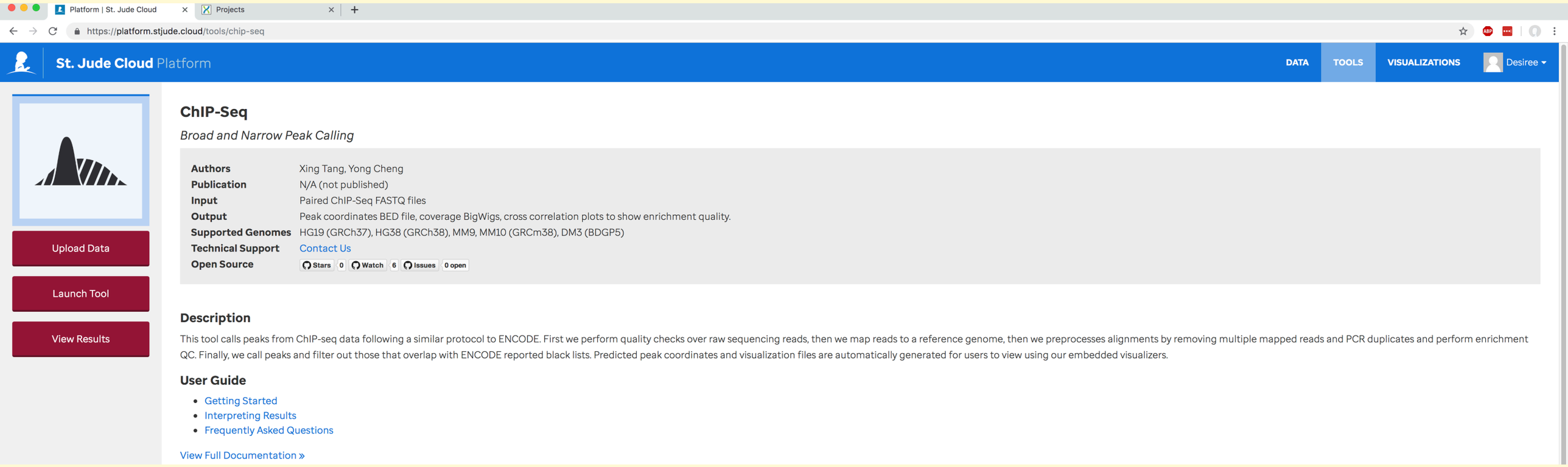
This generates a project called “ChIP-Seq” in your list of projects on your DNAnexus account
In it will be all of the available ChIP-Seq workflows (Broad Peak Caller, Narrow Peak Caller, etc):

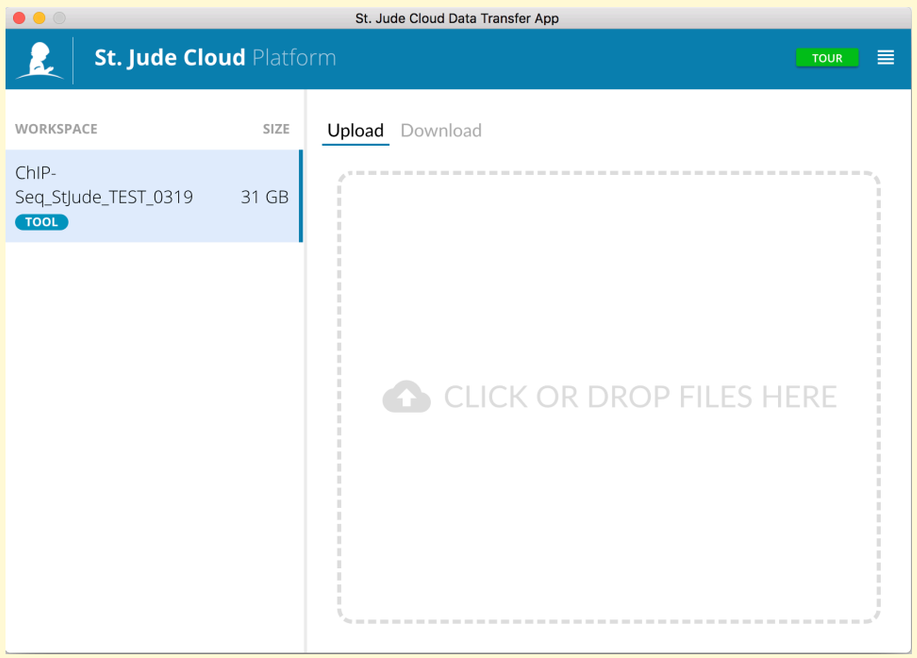


Step 2. Upload your data (.fastq files from a ChIP seq run)

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.



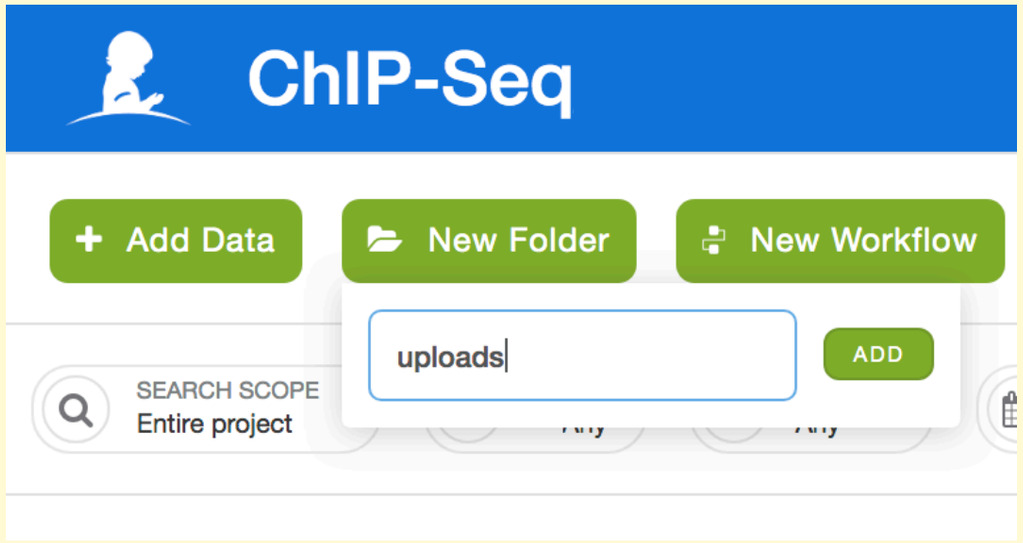


Uploading data using the app will place the data in a folder called “uploads” in your DNAnexus workspace

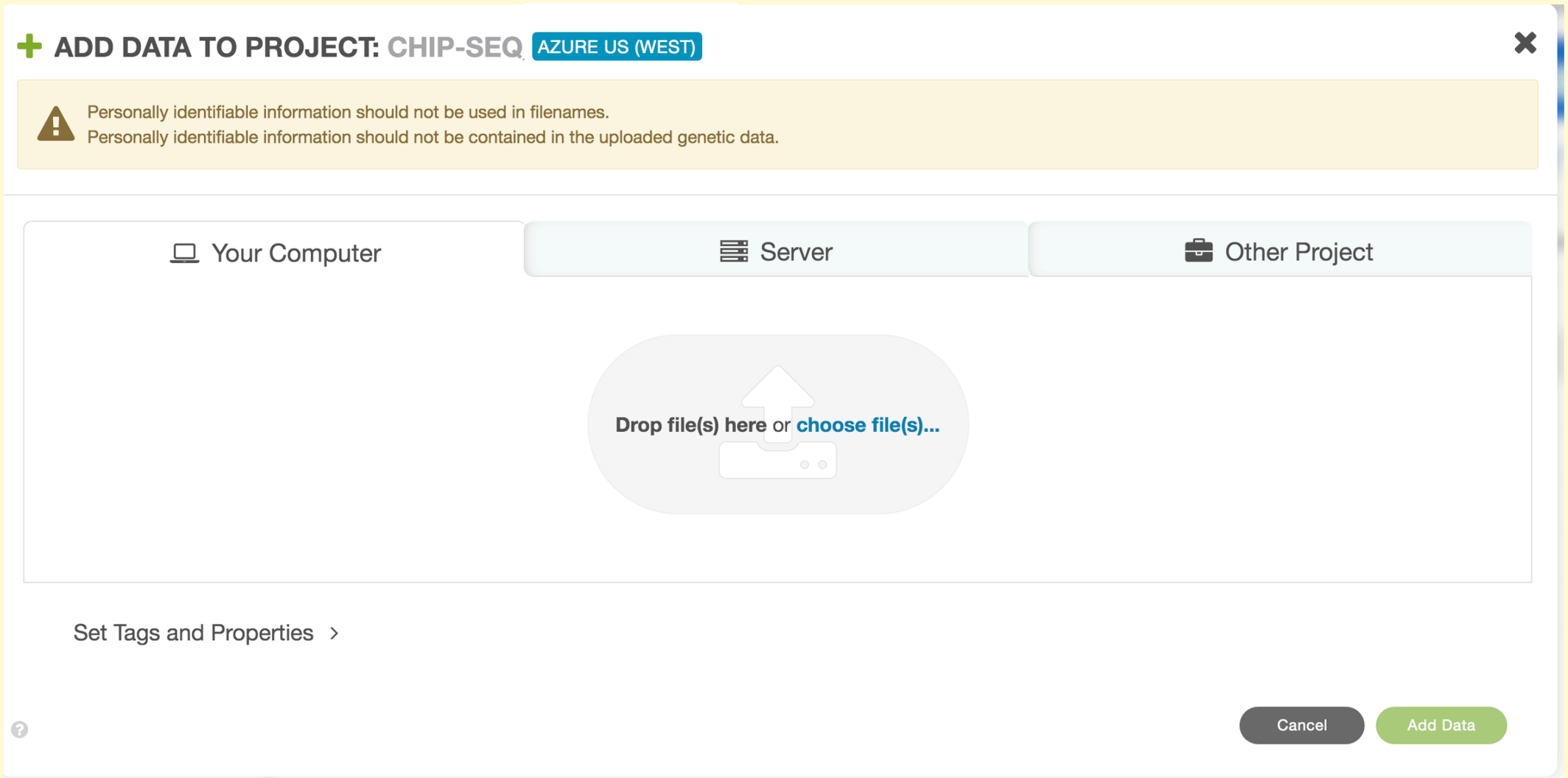
Step 2. Upload your data (.fastq files from a ChIP-seq run)

Option 2: On DNAnexus

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



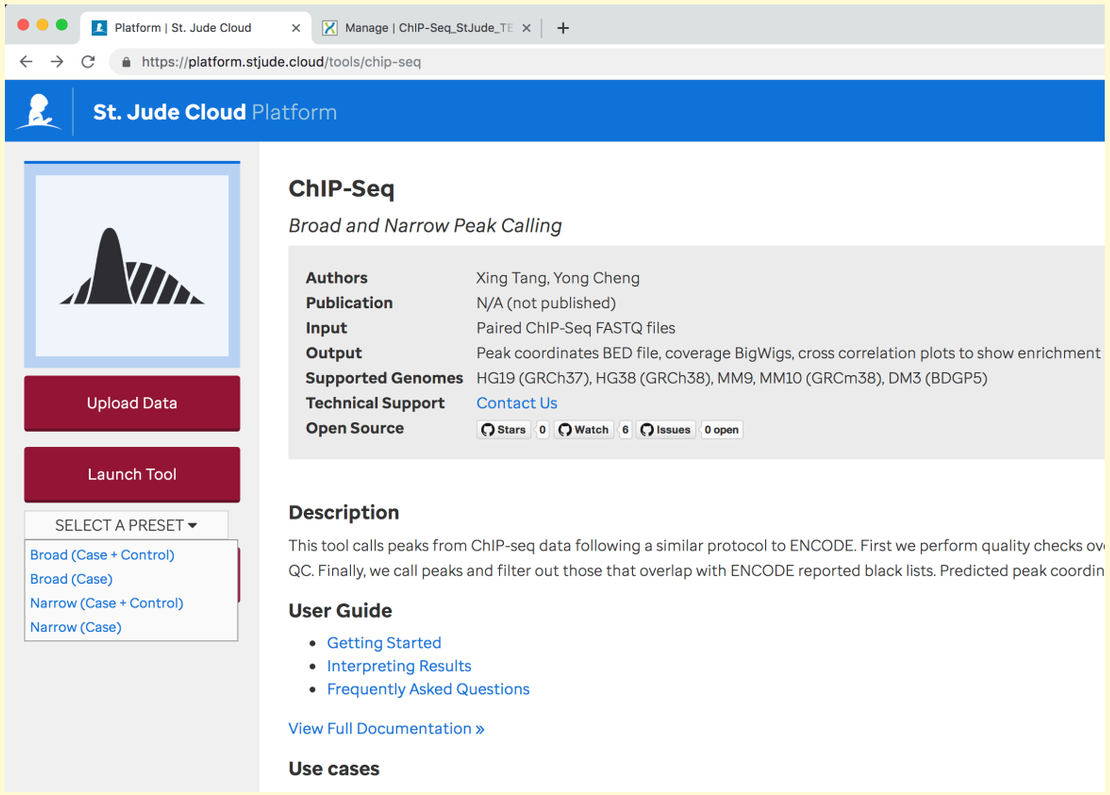
Navigate to your data folder, and click “Add data”



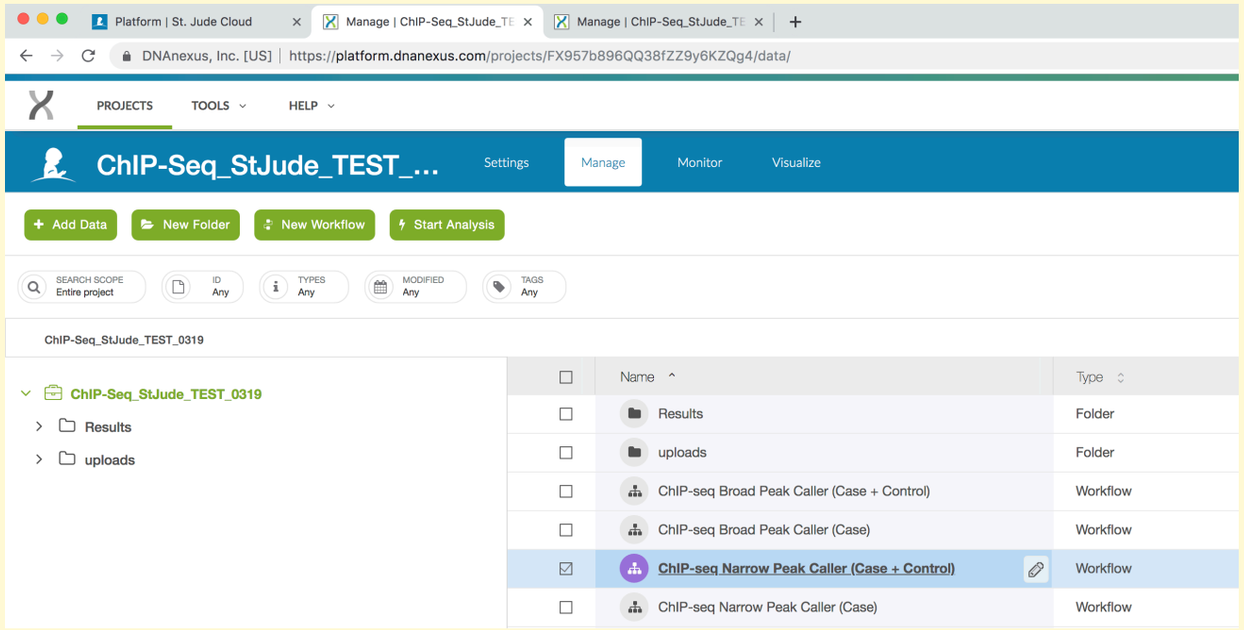
Step 3. Launch the workflow

Option 1: St. Jude portal

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

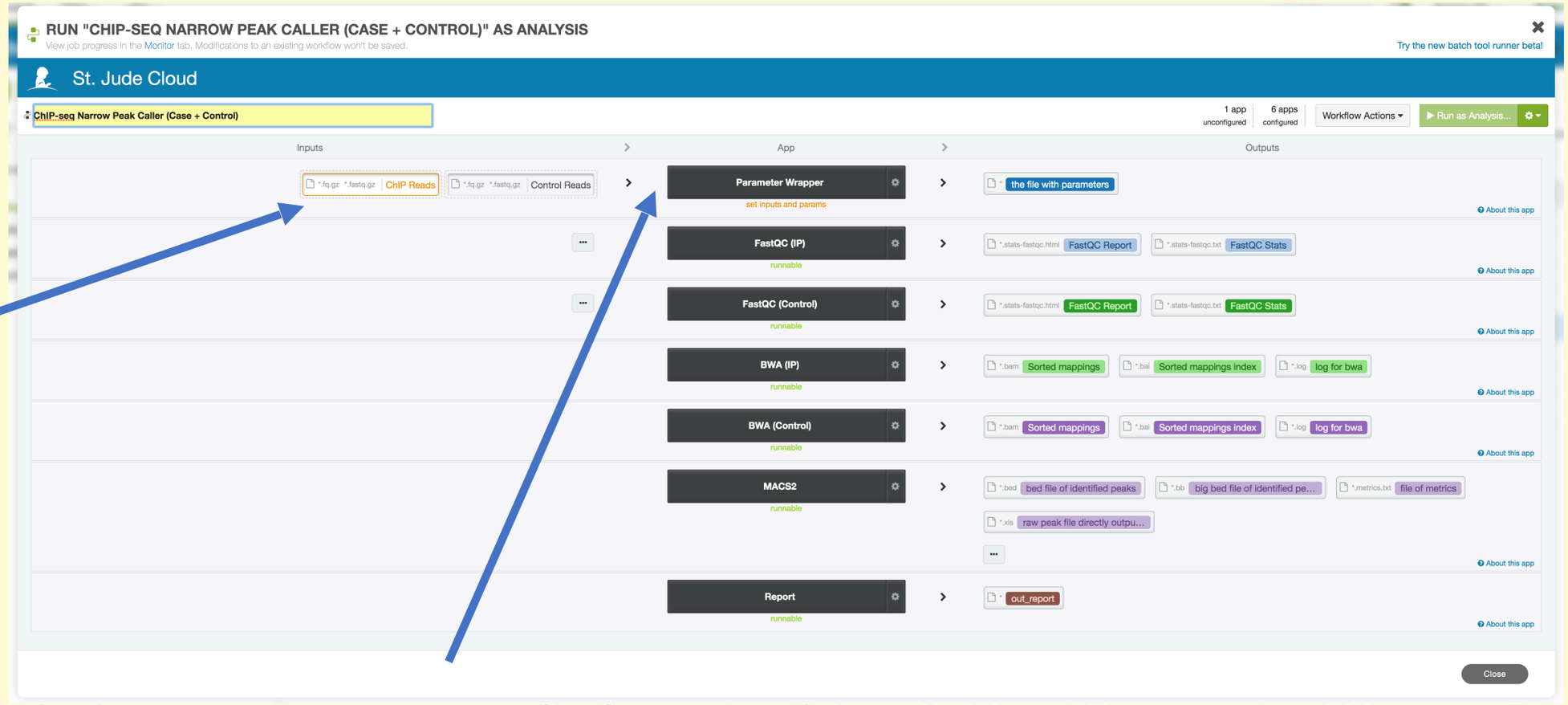


Option 2: Directly on DNAnexus

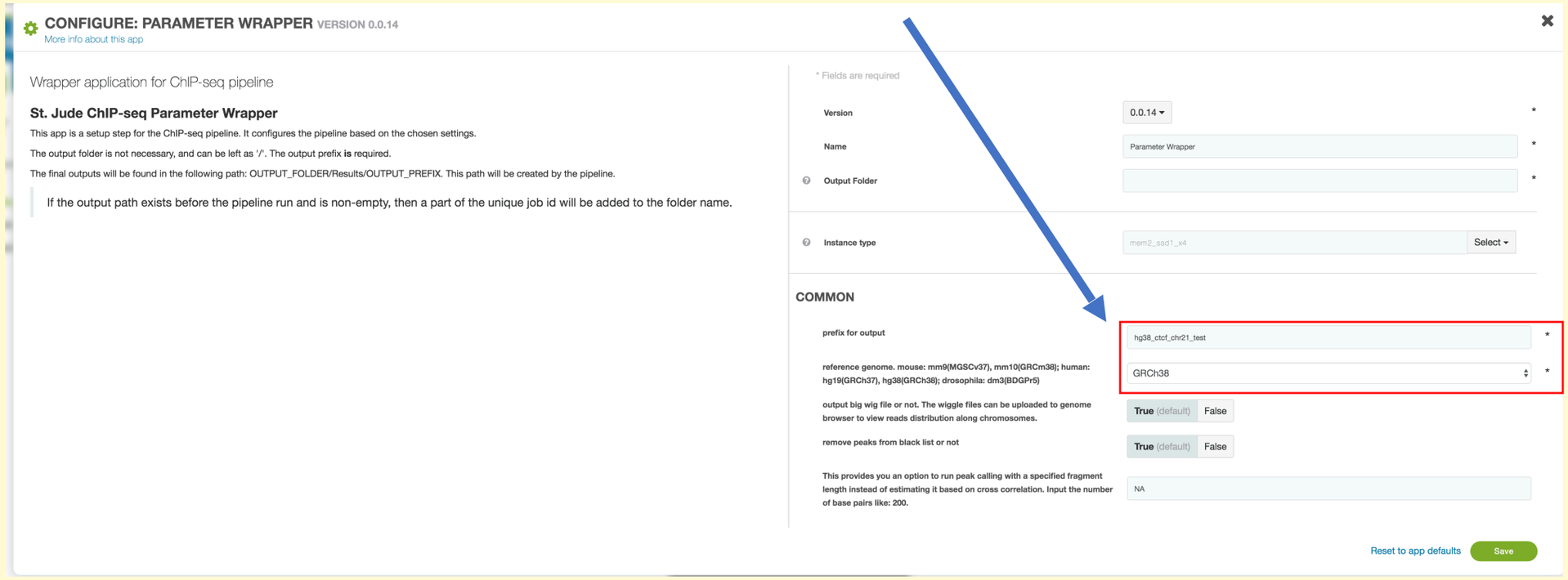


Either method will open the workflow (Option 1 will open a new tab)

Input your uploaded fastq into the input boxes



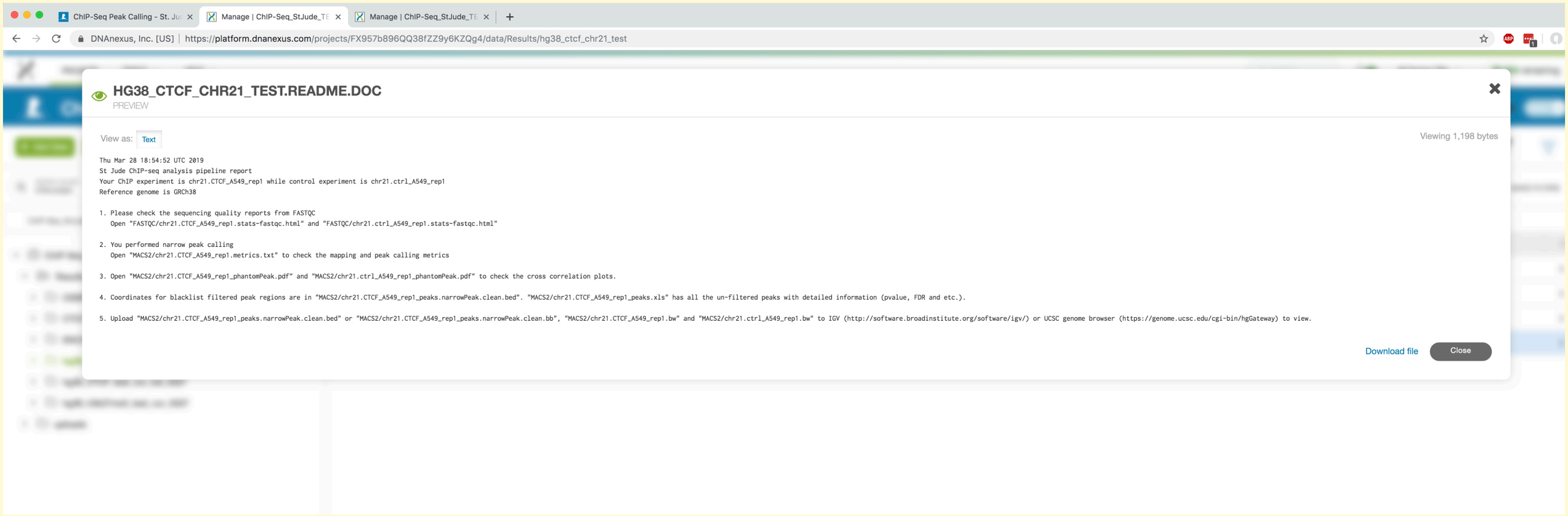
Set parameters (reference genome, prefix for output) by clicking “Parameter Wrapper”:



Hit save, and click “Run as analysis”

Results will be in results/<prefix>

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



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

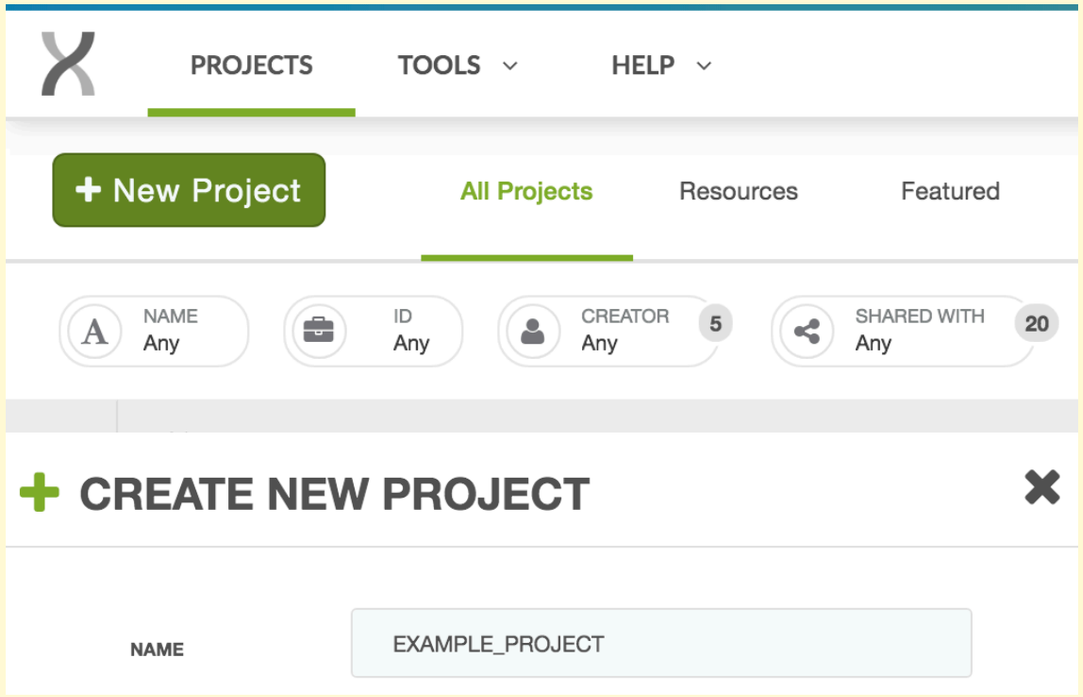
https://www.encodeproject.org/tutorials/chip-pipeline-howto/

Step 1. Set up+Data upload

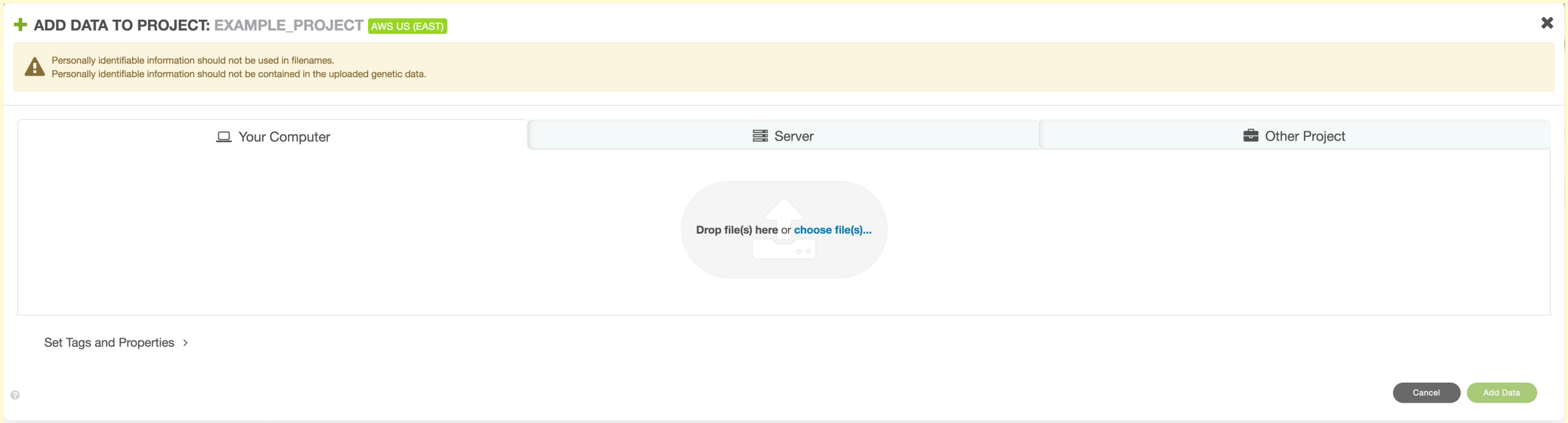
Sign into DNAnexus

Create a project (click on New Project)

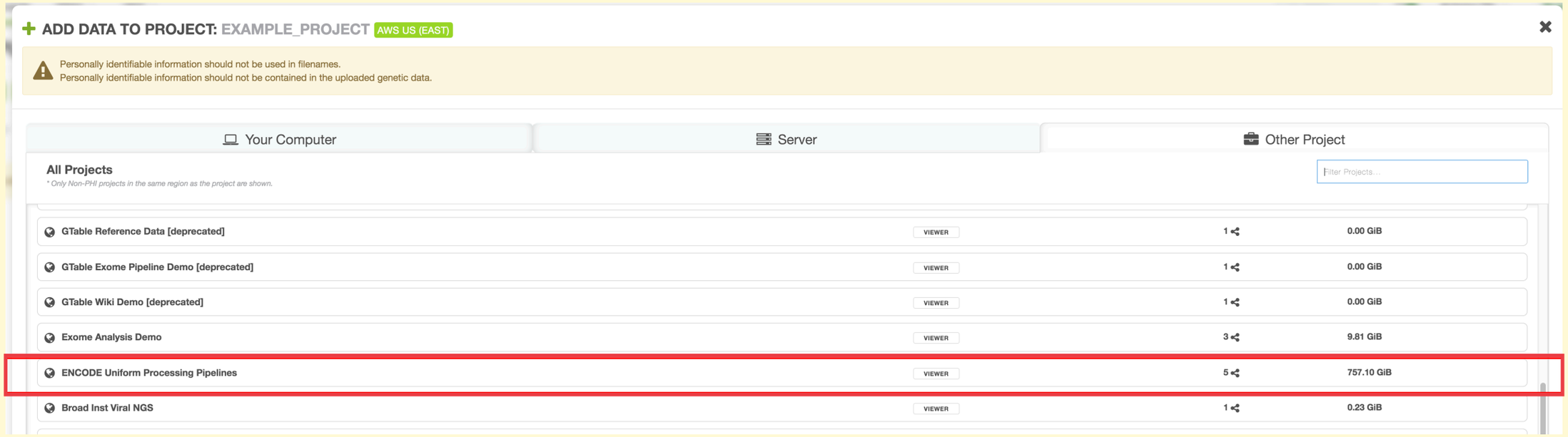
Name your project

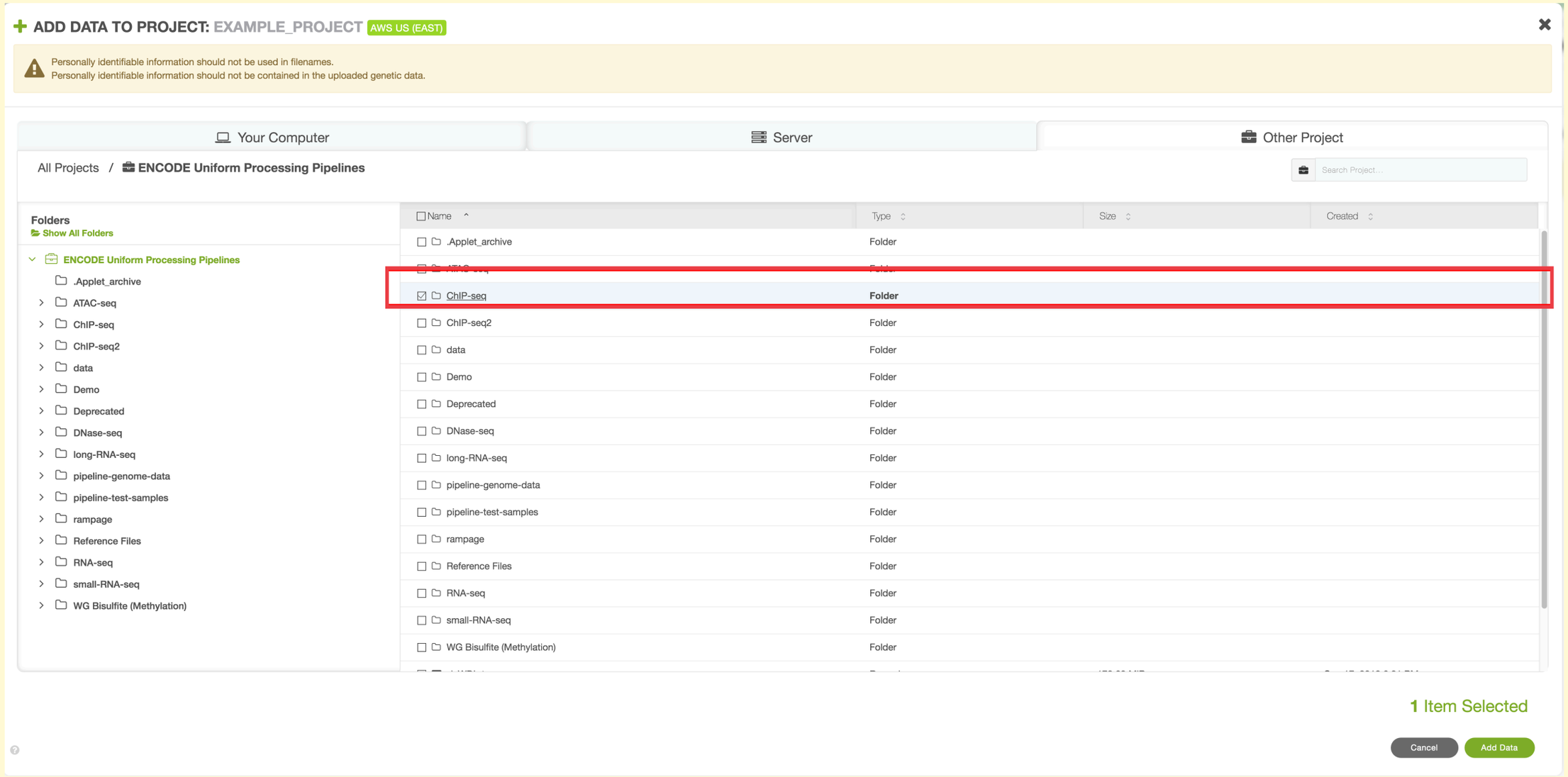


Add your fastq.gz files



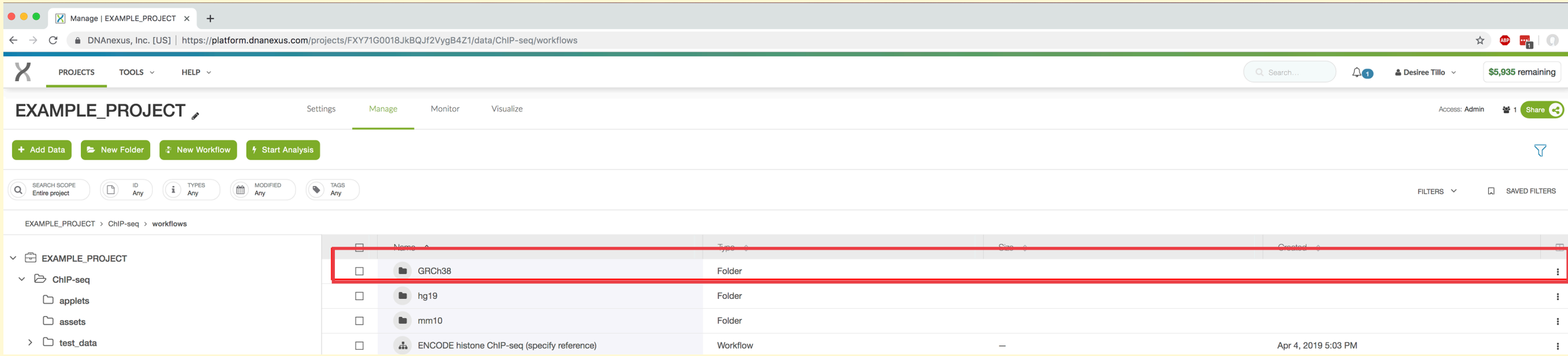
Add the ENCODE ChIP-seq workflows:



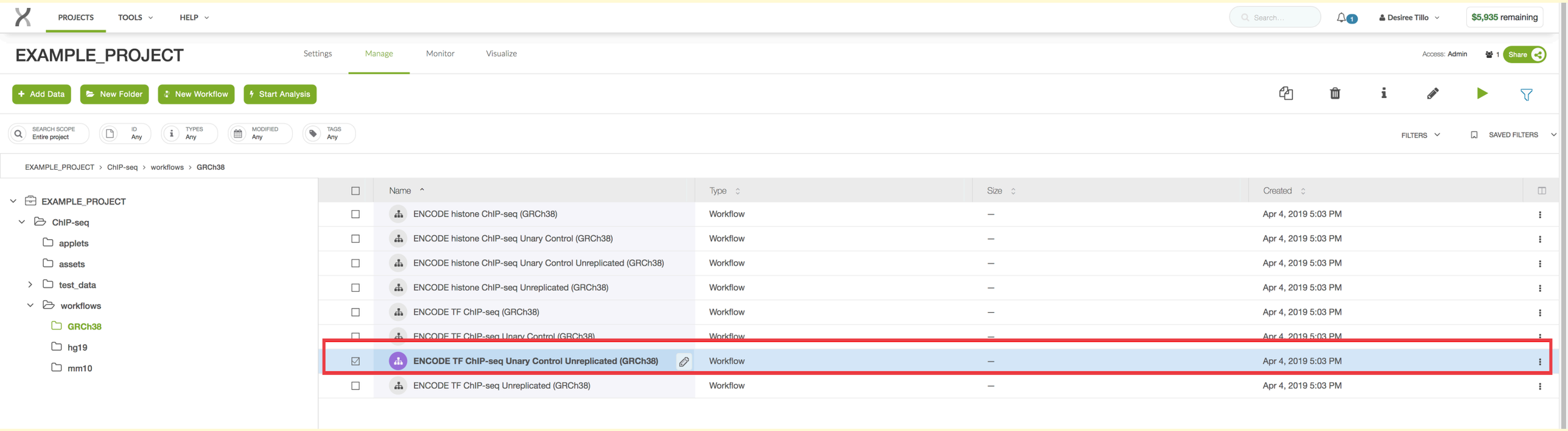


Step 2. Launch the workflow

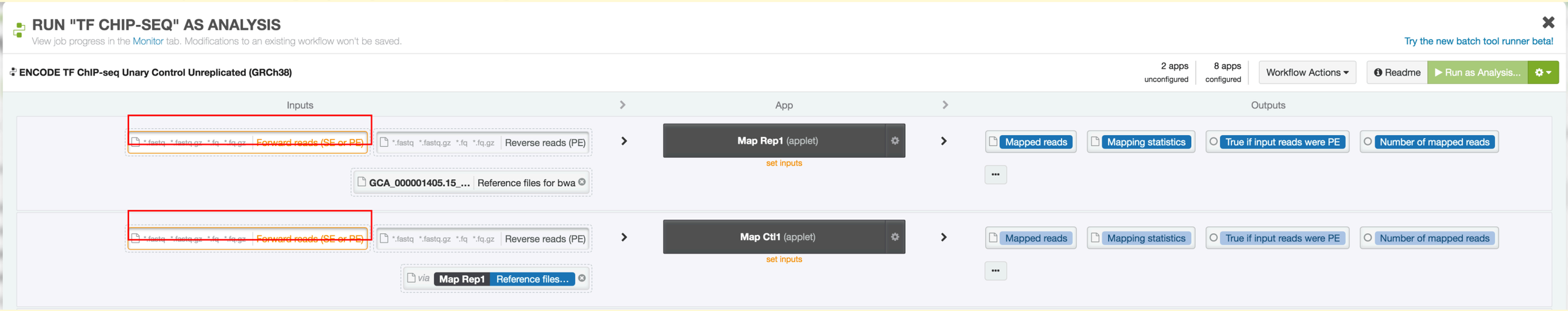
Go to ChIP-seq, workflows, select genome version (e.g. GRCh38)



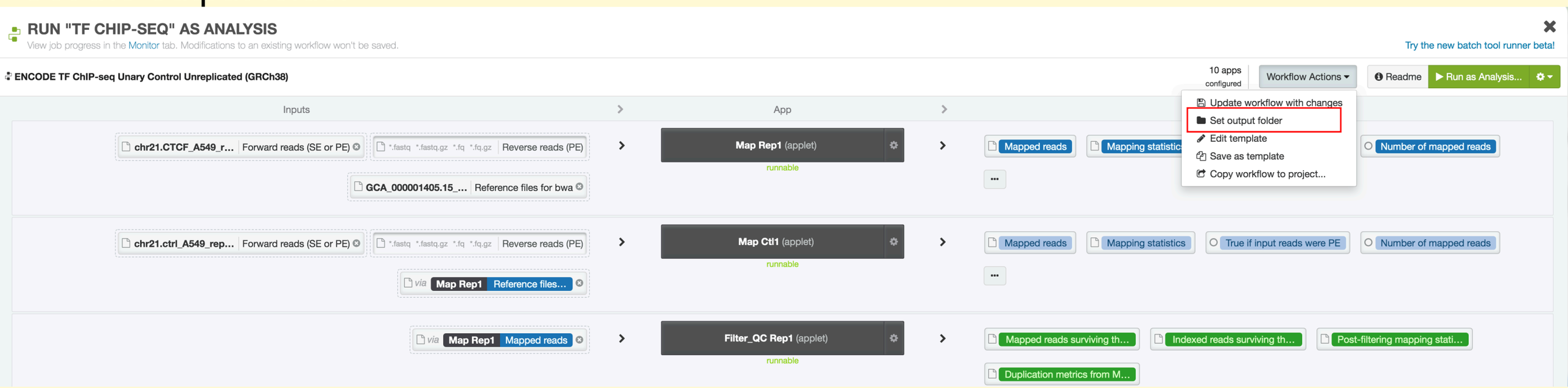
Select the workflow that best suits your experiment
(in this example, we have a single replicate of 1 case, 1 control)



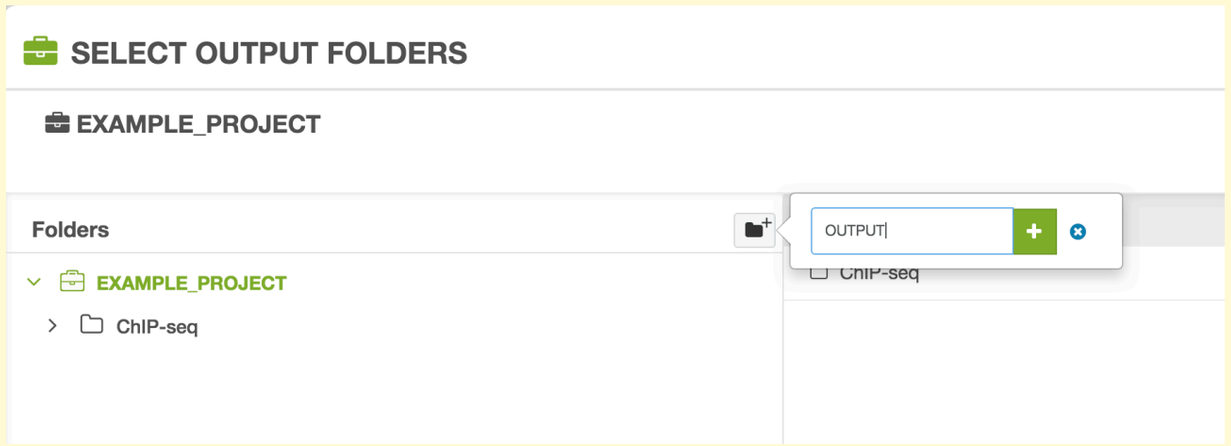
Add data (.fastq.gz) by clicking the appropriate box



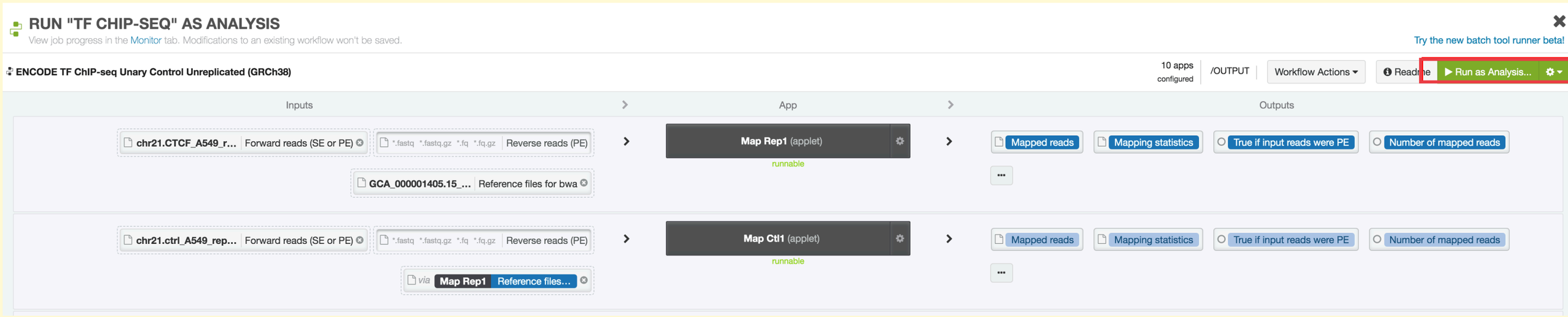
Set the output folder



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

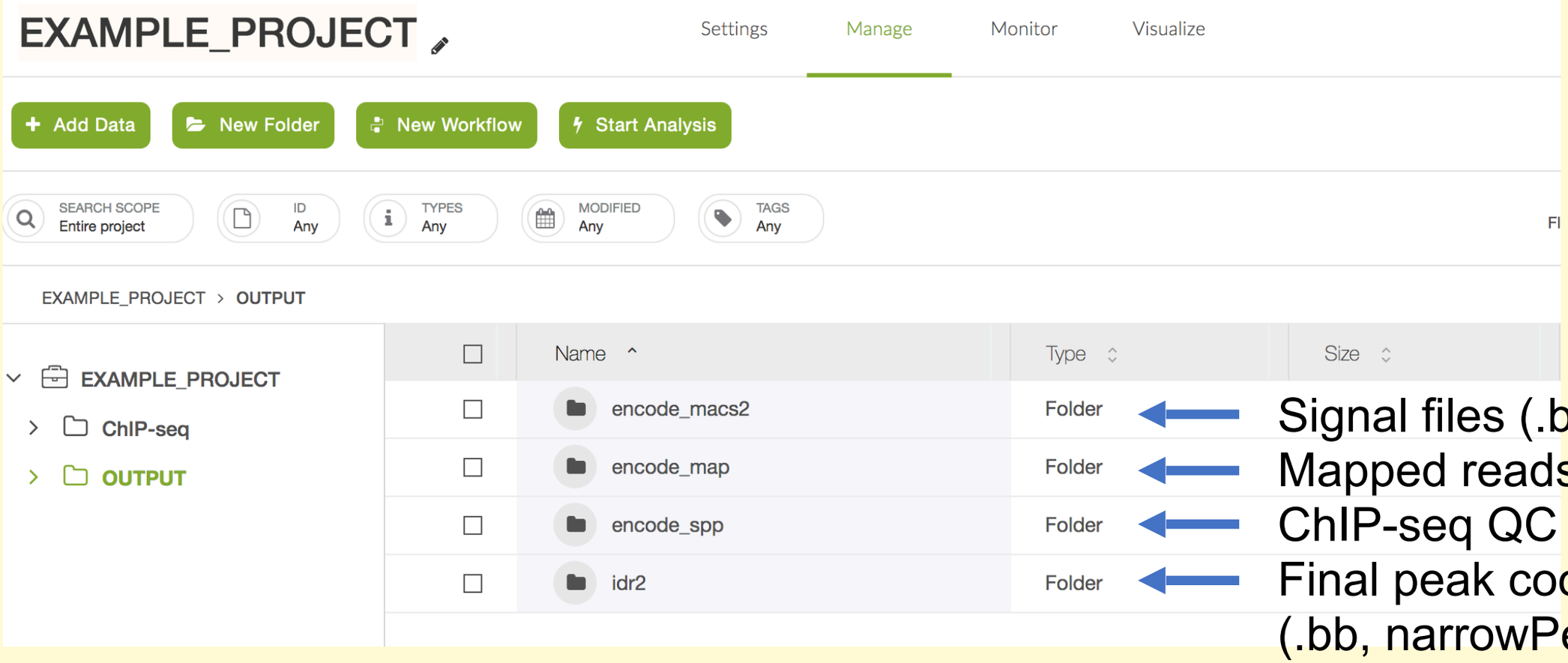


Click “Run as analysis”



Step 3. View results

Results will be in output directory set in Step 2.



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

ENCODE TF ChIP-seq Unary …
DONE

▶ Launch as new Analysis | Save as new Workflow | View Info

EXECUTION ID	LAUNCHED ON	RAN FOR	FINAL COST	EXECUTABLE	LAUNCHED BY	OUTPUT FOLDER
analysis-FXY5gyQ0YK4yQJGp8b7k3y1X	04/04/2019 3:33 pm	6h 43m	$7.0110	ENCODE TF ChIP…	Desiree Tillo	/hg38_ctcf_rep1_…

03:36PM	04:54PM	05:49PM	06:44PM	07:39PM	08:34PM	04/04/2019 10:16:00PM

Map Rep1	47m	Log
Map Ctl1	44m	Log
Filter_QC Rep1	19m	Log
Xcor Rep1	19m	Log
Filter_QC Ctl1	18m	Log
Xcor Ctl1	18m	Log
ENCODE Peaks	1h 8m	Log
SPP Peaks	5h 4m	Log
IDR Rep 1 Self-pseudoreplicates	3m	Log
Final IDR peak calls	4m	Log

St. Jude

ChIP-seq Narrow Peak Caller…
DONE

▶ Launch as new Analysis | Save as new Workflow | View Info

EXECUTION ID	LAUNCHED ON	RAN FOR	FINAL COST	EXECUTABLE	LAUNCHED BY	OUTPUT FOLDER
analysis-FXY5jj896QQ89bx0GypF79G6	04/04/2019 3:34 pm	1h 27m	$0.8409	ChIP-seq Narrow …	Desiree Tillo	/

03:40PM	03:52PM	04:00PM	04:07PM	04:15PM	04:23PM	04:31PM	04:39PM	04:47PM	04/04/2019 05:02:24PM

Parameter Wrapper	< 1m	Log
FastQC (IP)	3m	Log
FastQC (Control)	3m	Log
BWA (IP)	17m	Log
BWA (Control)	17m	Log
MACS2	52m	Log
Report	< 1m	Log

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	**Sample Type**	**Price/Sample**
Sentieon DNAseq FASTQ to VCF (Panel)	Panel (fastq.gz ≤ 5 GiB)	$3.30
Sentieon DNAseq FASTQ to VCF (WES)	WES (5 GiB < fastq.gz ≤ 20 GiB)	$9.90
Sentieon DNAseq FASTQ to VCF (WGS)	WGS (20 GiB < fastq.gz ≤ 90 GiB)	$47.00
Sentieon DNAseq FASTQ to VCF (WGS2)	WGS2 (90 GiB < fastq.gz ≤ 200 GiB)	$70.00 to ~$100.00
Sentieon TNseq FASTQ to VCF	**Sample Type**	**Price/Sample**
Sentieon TNseq FASTQ to VCF (WES)	WGS (40 GiB < fastq.gz ≤ 300GiB)	$80 to ~$130
Sentieon TNseq FASTQ to VCF (WGS)	WES (fastq.gz ≤ 40GiB)	$23 to ~$30

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

PROJECTS TOOLS ORG ADMIN HELP

My RNASEQ Project

Settin

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

SELECT A TOOL

The app or applet will be added to your workflow

Filter tools

*.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

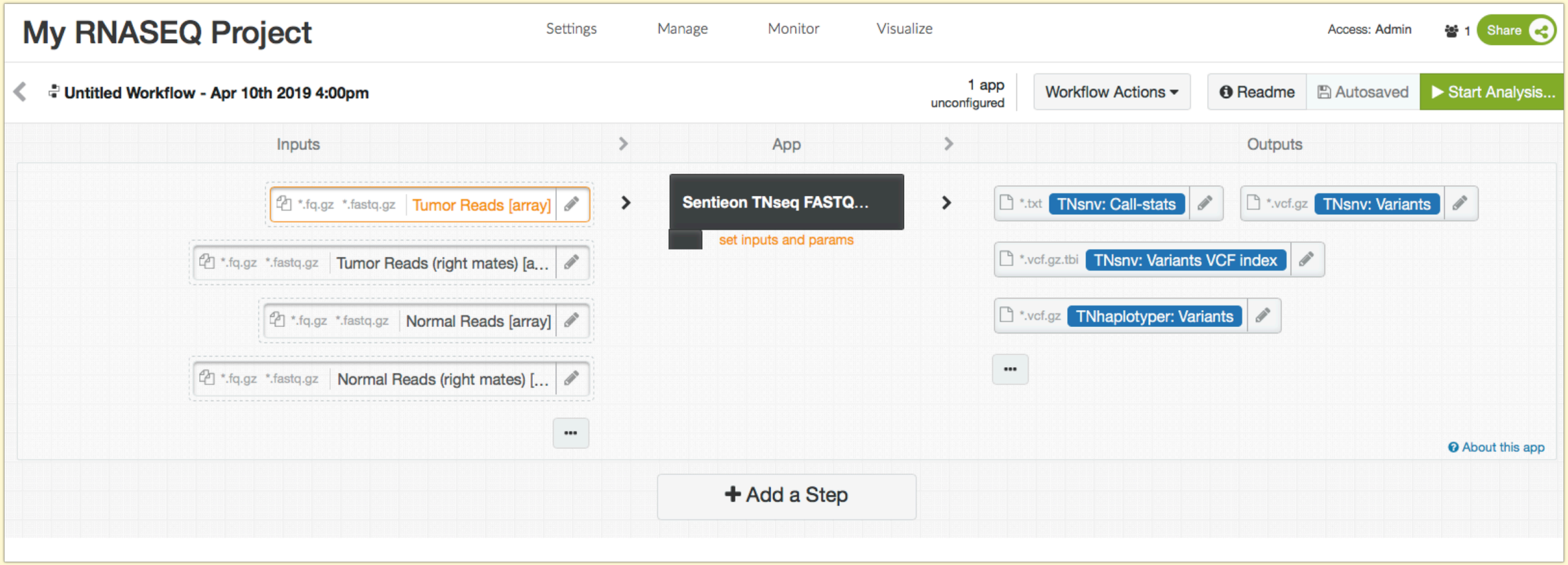
*.FQ.GZ... *.FQ.GZ... *.FQ.GZ... *.FQ.GZ... *.VCF.GZ *.VCF.GZ *.CSV *.BWA-INDEX.TAR.GZ *.FA*.GZ *.RESOURCE.BUNDLE.TAR.GZ *.BED >

*.TXT *.VCF.GZ *.VCF.GZ.TBI *.VCF.GZ *.VCF.GZ.TBI *.VCF.GZ *.VCF.GZ.TBI *.BAM *.BAI *.DUPLICATION_METRICS... *.BAM *.BAI *.TABLE *.BAM *.BAI *.BAM *.BAI *.DUPLICATION_METRICS... *.BAM *.BAI *.BAM *.BAI *.TABLE *.BAM *.BAI

+ STAR Mapping v1.3.3 Read Mapping

*.STAR-INDEX.TAR.GZ *.FASTQ... *.FASTQ... *.VCF... *.TSV.GZ... *.GTF... > *.BAM *.BAI *.LOG.GZ *SJ.OUT.TSV.GZ... *.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]	Sentieon TNseq FA…	*.txt TNsnv: Call-stats
*.fq.gz *.fastq.gz Tumor Reads (right mates) [a…	set inputs and params	*.vcf.gz TNsnv: Variants
*.fq.gz *.fastq.gz Normal Reads [array]		*.vcf.gz.tbi TNsnv: Variants VCF index
*.fq.gz *.fastq.gz Normal Reads (right mates) […		*.vcf.gz TNhaplotyper: Variants
*.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]		

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)**

Search Project...

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*)
  (help, general, development)

- Custom Built Work Flows (RNASEQ workflow, IGV_sesssion_maker, ADAP, *Pausing Peak Aligner*, Tumor Mutation Burden**)

- DNAnexus Applications By Category Page
  (https://dl.dnanex.us/F/D/jpyV1BVZKZJzf811QXfg7X13P8x1Z41P7zKVygpX?inline)

- Management of DNAnexus Account, Funding and cost management