Galaxy Tutorial – Basics of Data Analysis using Galaxy platform, Condensed

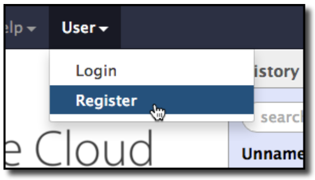
(SW4) The Galaxy Platform for Multi-Omic Data Analysis and Informatics

Dave Clements, Galaxy Project, Johns Hopkins University
8:30-10:00am, Saturday February 20, 2016

ABRF 2016

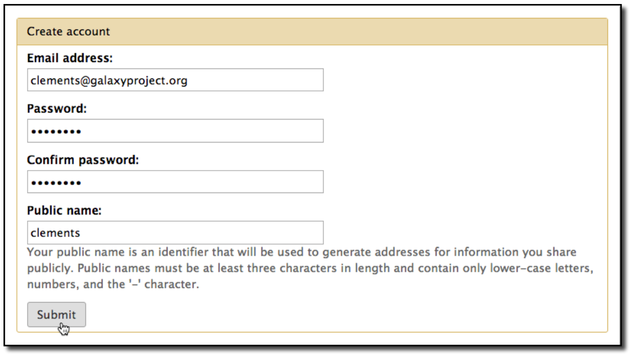
A longer version of the same material is available online at http://bit.ly/gxyabrf16intro-long
This handout is available online at http://bit.ly.gxyabrf16intro

Create an account



To create an account, *click* on the **User** pulldown and *select* **Register**.

Use an email address you can remember, **and a low security password**. Note that these servers use HTTP, not HTPPS, and therefore your password is going out on the net unencrypted. (Galaxy can use HTTPS and most servers do, but the default cloud installation uses HTTP.)



Create account

Email address: clements@galaxyproject.org

Password: ••••••••

Confirm password: ••••••••

Public name: clements

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least three characters in length and contain only lower-case letters, numbers, and the '-' character.

Reload the page by *clicking* **Analyze Data** or **Return to the home page**.

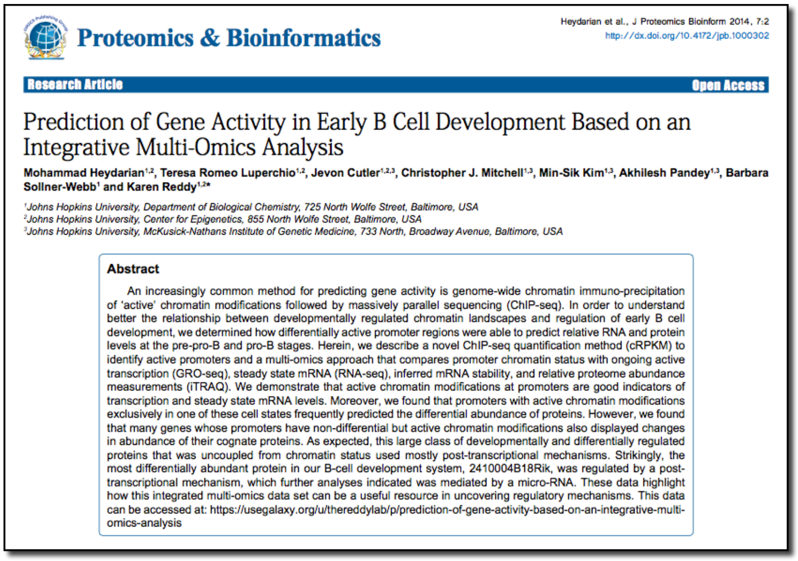
Our goal: Identify transcripts using RNA-Seq in pre-pro-B and pro-B mouse cells

We are going to use RNA-seq data to detect transcripts, possibly novel transcripts, in pre-pro-B and pro-B mouse cells.

This data comes from

> Heydarian *et al.*, (2014) Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis. *J Proteomics Bioinform* 7: 050-063. doi:10.4172/jpb.1000302

The paper uses RNA-Seq, ChIP-Seq, GRO-Seq, and iTRAQ techniques to "demonstrate that active chromatin modifications at promoters are good indicators of transcription and steady state mRNA levels."



Proteomics & Bioinformatics

Heydarian et al., J Proteomics Bioinform 2014, 7:2
http://dx.doi.org/10.4172/jpb.1000302

Research Article — Open Access

Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis

Mohammad Heydarian[1,2], Teresa Romeo Luperchio[1,2], Jevon Cutler[1,2,3], Christopher J. Mitchell[1,3], Min-Sik Kim[1,3], Akhilesh Pandey[1,3], Barbara Sollner-Webb[1] and Karen Reddy[1,2]*

[1]Johns Hopkins University, Department of Biological Chemistry, 725 North Wolfe Street, Baltimore, USA
[2]Johns Hopkins University, Center for Epigenetics, 855 North Wolfe Street, Baltimore, USA
[3]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, 733 North, Broadway Avenue, Baltimore, USA

Abstract

An increasingly common method for predicting gene activity is genome-wide chromatin immuno-precipitation of 'active' chromatin modifications followed by massively parallel sequencing (ChIP-seq). In order to understand better the relationship between developmentally regulated chromatin landscapes and regulation of early B cell development, we determined how differentially active promoter regions were able to predict relative RNA and protein levels at the pre-pro-B and pro-B stages. Herein, we describe a novel ChIP-seq quantification method (cRPKM) to identify active promoters and a multi-omics approach that compares promoter chromatin status with ongoing active transcription (GRO-seq), steady state mRNA (RNA-seq), inferred mRNA stability, and relative proteome abundance measurements (iTRAQ). We demonstrate that active chromatin modifications at promoters are good indicators of transcription and steady state mRNA levels. Moreover, we found that promoters with active chromatin modifications exclusively in one of these cell states frequently predicted the differential abundance of proteins. However, we found that many genes whose promoters have non-differential but active chromatin modifications also displayed changes in abundance of their cognate proteins. As expected, this large class of developmentally and differentially regulated proteins that was uncoupled from chromatin status used mostly post-transcriptional mechanisms. Strikingly, the most differentially abundant protein in our B-cell development system, 2410004B18Rik, was regulated by a post-transcriptional mechanism, which further analyses indicated was mediated by a micro-RNA. These data highlight how this integrated multi-omics data set can be a useful resource in uncovering regulatory mechanisms. This data can be accessed at: https://usegalaxy.org/u/thereddylab/p/prediction-of-gene-activity-based-on-an-integrative-multi-omics-analysis

The data

We have two RNA-Seq datasets, one from mouse pre-pro-B cells, and one from mouse pro-B cells. Each dataset consists of ~90 million single-end reads. Sequencing was done an Illumina HiSeq 2000 and all reads are 97bp long.

We are going to use striiped down datasets consisting of reads that tend to map to the particular regions of the genome we'll focus on today, plus some additional reads, just to make things interesting.

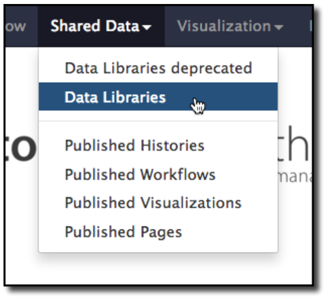
> *Note that while reducing the datasets will make tool execution very fast, it can be a dangerous cheat. Many NGS tools are built on statistical models that assume they are being given data for the entire genome. With our data that is decidedly not the case. Downsampling your data is still a useful technique for early experimentation when building your analysis, but results and assumptions should be checked when scaling up your analysis to complete datasets.*

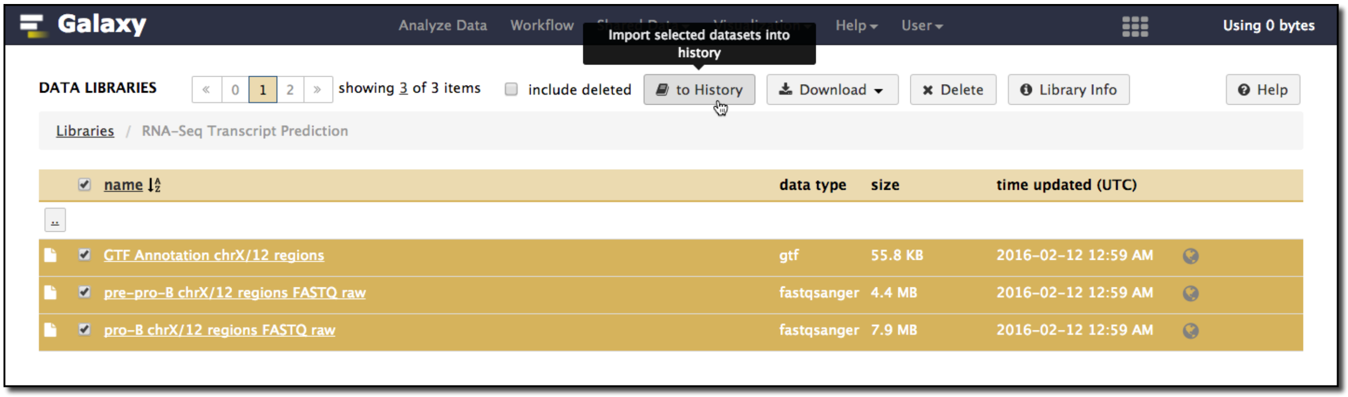
Get the data / Data libraries

Our reduced datasets are available in a *data library* on your server. *Click* on the **Shared Data** pull-down, and then *select* **Data Libraries**.

Then *click* on the **RNA-Seq Transcript Prediction** library and select all 3 datasets in this library and then *click* on **to History** to import them into your current history.

This will ask which history to import these datasets into. So far, you only have one ("Unnamed history"). *Click* **Import** to add them to that history.



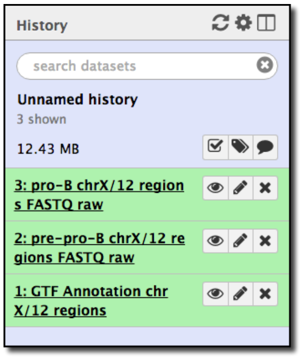


☑ name	data type	size	time updated (UTC)
☑ GTF Annotation chrX/12 regions	gtf	55.8 KB	2016-02-12 12:59 AM
☑ pre-pro-B chrX/12 regions FASTQ raw	fastqsanger	4.4 MB	2016-02-12 12:59 AM
☑ pro-B chrX/12 regions FASTQ raw	fastqsanger	7.9 MB	2016-02-12 12:59 AM

Now *click* **Analyze Data** to see what we have.

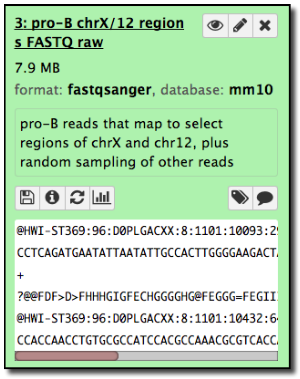
FASTQ datasets

And what we have is a history consisting of three datasets. Two of them are the raw read datasets for the pre-pro-B and pro-B cells. These datasets are in FASTQ format.



Unnamed history
3 shown
12.43 MB

3: pro-B chrX/12 regions FASTQ raw
2: pre-pro-B chrX/12 regions FASTQ raw
1: GTF Annotation chrX/12 regions



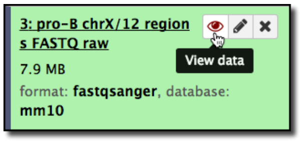
3: pro-B chrX/12 regions FASTQ raw
7.9 MB
format: **fastqsanger**, database: **mm10**

pro-B reads that map to select regions of chrX and chr12, plus random sampling of other reads

```
@HWI-ST369:96:D0PLGACXX:8:1101:10093:2
CCTCAGATGAATATTAATATTGCCACTTGGGGAAGACT
+
?@@FDF>D>FHHHGIGFECHGGGGHG@FEGGG=FEGII
@HWI-ST369:96:D0PLGACXX:8:1101:10432:6
CCACCAACCTGTGCGCCATCCACGCCAAACGCGTCACC
```

To see a preview of any dataset, click on the dataset name. This displays some metadata about the dataset, and if it's in a text-based format, it will also show the first few lines of the dataset.

To view the full dataset, *poke it in the eye*. This will display the content of this FASTQ dataset in the middle panel.

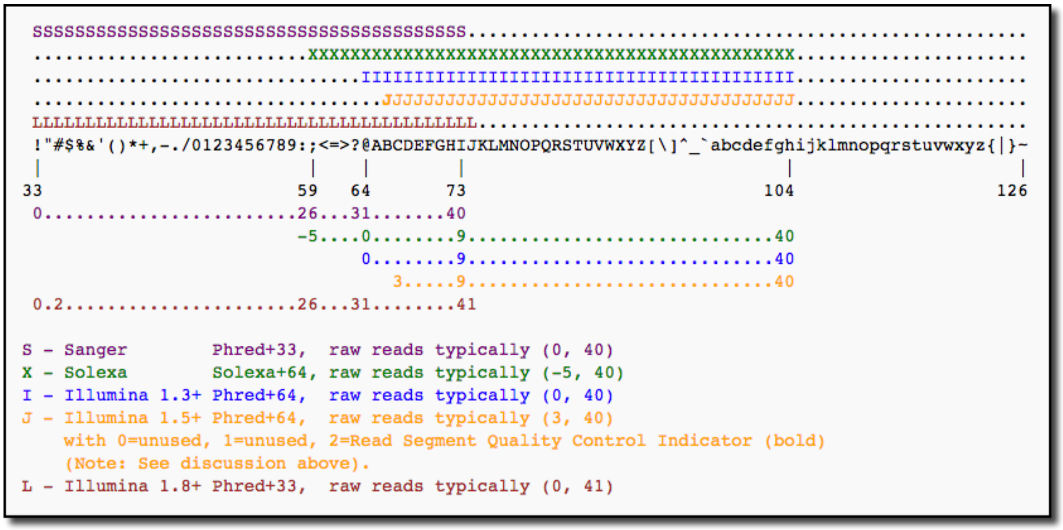


3: pro-B chrX/12 regions FASTQ raw
7.9 MB
format: **fastqsanger**, database: **mm10**

FASTQ Format

Each FASTQ entry is 4 lines long and incorporates three types of information:

- Line 1 starts with an @ and is the read's identifier
- Line 2 shows the called bases in this read
- Line 3 is a separator (the + character).
- Line 4 shows the instrument's confidence in each of the base calls.



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

https://en.wikipedia.org/wiki/FASTQ_format

Phred quality scores scores are assigned to each base call. Phred scores are logarithmic and represent the likelihood that given base call is incorrect. Most datasets use Sanger scoring which typically ranges from 0 to 40. Some reference points in that spectrum:

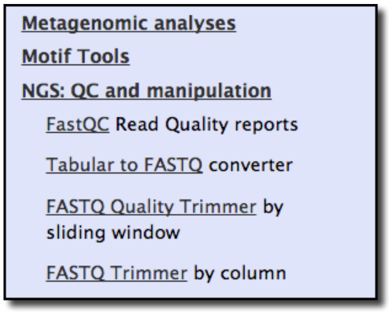
Score	**Instrument thinks it will be wrong**	
10	10% of the time	(1 out of 10 times)
20	1% of the time	(1 out of 100 times)
30	0.1% of the time	(1 out of 1000 times)
40	0.01% of the time	(1 out of 10000 times)

We don't see these scores in the FASTQ files because each score is encoded to a single character according the FASTQSanger convention. FASTQSanger maps scores to a range of adjacent ASCII characters, starting with `!` for 0, and ending with `I` for 40.

Check FASTQ dataset quality

There are two ways to find tools on a Galaxy server. If you know the name of the tool you can enter it in the tool panel search box. You can also try searching for more general terms like "trim" or "quality." Searching with more general terms will be hit or miss.

You can also just browse each Toolbox. In this case we are investigating the quality of an NGS dataset, so the **NGS: QC and Manipulation** toolbox looks the most promising. *Click* on the toolbox name to see the tools in it. In this case **FastQC Read Quality reports** is right at the top, and that sounds exactly like what we are looking for. *Click* on it.



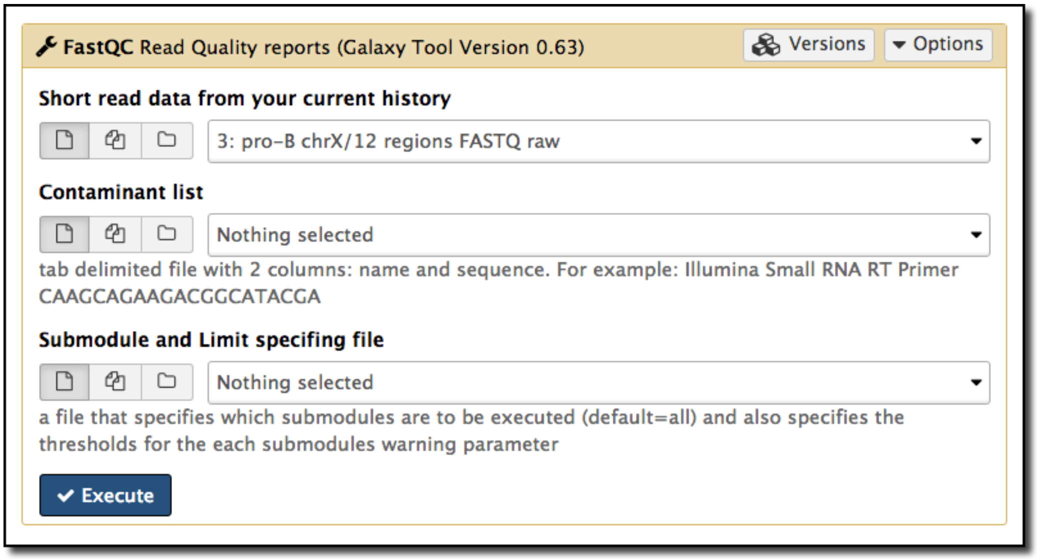
Metagenomic analyses

Motif Tools

NGS: QC and manipulation

- FastQC Read Quality reports
- Tabular to FASTQ converter
- FASTQ Quality Trimmer by sliding window
- FASTQ Trimmer by column

FastQC is a widely used tool for summarizing the quality of FASTQ datasets. It's good at presenting the big picture, and at identifying specific areas that may need to be addressed.



FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

3: pro-B chrX/12 regions FASTQ raw

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

This is a typical, if simple, Galaxy tool form. Tool pages generally have these sections

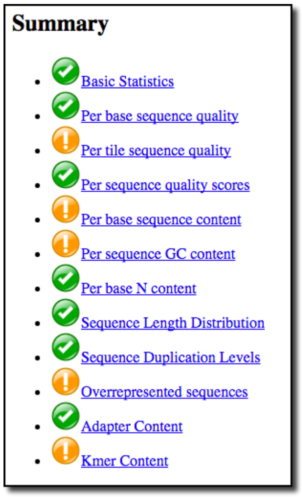
- *Versions and options* - If multiple versions are supported, there will be link to run the others. And the options menu enables you see more about the tool and how it is wrapped in Galaxy.
- *The form* - where you configure and run the tool
- *Documentation* - Describes the tool and provides help. This section often include links to a tool's external documentation.
- *Citation* - A paper, if there is one, for this tool.

Run our first tool: FastQC

FastQC has only three parameters that can be set. Let's set only one, **Short read data from your current history**. *Click* this option's dataset pulldown menu and select **2: pre-pro-B chrX/12 regions FASTQ raw**. We aren't going to set the contaminant list or limit the submodules/tests that FastQC runs. We'll request that it run all of its tests. *Click* **Execute**.

Now a quick succession of things (hopefully) happens

1. A big green box appears in the middle panel and two small gray boxes appear at the top of the history. These mean that the task has been successfully queued: It's ready to run, but isn't running yet.
2. The gray boxes in the history are replaced with yellow boxes. This means the job has started and is actively running
3. The yellow box are replaced with green boxes indicating that the tool finished successfully.



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

Once a tool has finished running you can preview the output datasets (by clicking on the dataset name) or view the data itself (by poking it in the eye). *Poke* the **FastQC Web page** dataset *in the eye*.

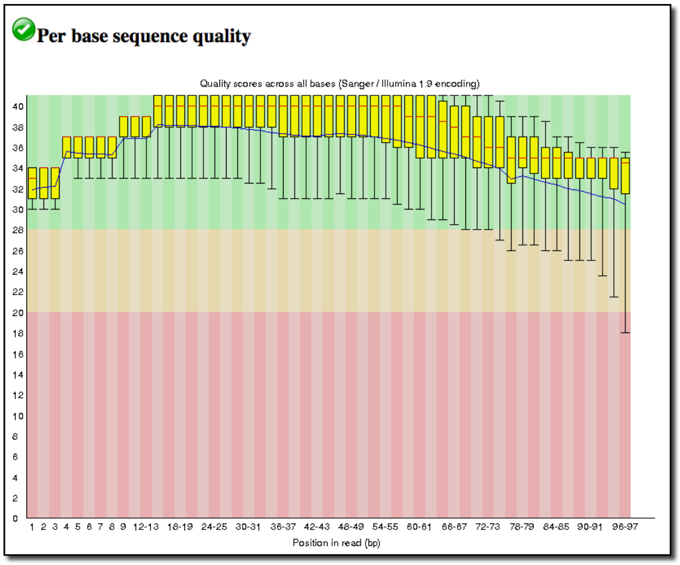
This brings up 12 different reports about the pre-pro-B FASTQ dataset.

For each test it runs, FastQC has predefined thresholds for what constitutes good, not-so-good, and bad. This dataset has passed 7 of the tests and got warning on the rest. Let's take a look at some of these tests.

Per base sequence quality (boxplot)

Boxplots are the most common way to summarize the quality of a FASTQ dataset.

- **X axis is position in the read.** That is, the first column summarizes all the base calls in the first position of all the reads; the second column summarizes the second call in all the reads and so on. Starting at position 10, the information is aggregated for every two bases.
- **Y axis is Phred quality score**, from 0 to 40.
- Each column summarizes the quality scores across all reads at that position
  - **Yellow boxes bound the middle 50%** of quality scores at each position in the read.
  - **Whiskers bound the middle 80%** of quality scores at that position.
  - **Blue line is the average score** at each position
  - **Red lines are the median score** at each position.

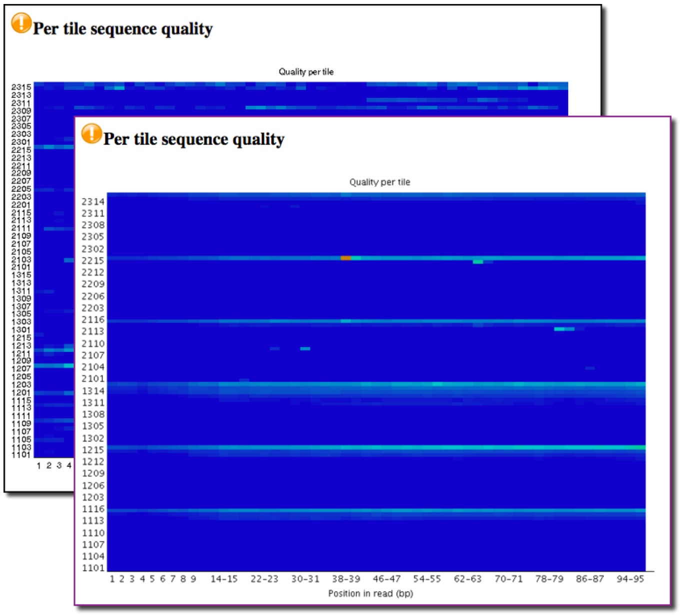


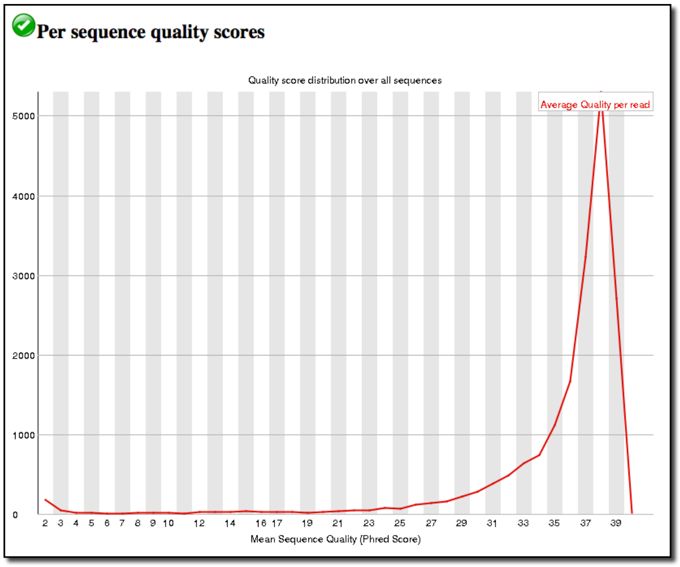
Per tile sequence quality

This graph is particularly significant for this crowd. It highlights where on the slide the low quality reads tended to be. As core staff you can use this information to identify trouble spots with the sequencing run.

For Illumina data, tile information is embedded in the ID line of FASTQ entries.

Note: We're showing the graphic for the complete ~90 million read dataset.





Per sequence quality scores

This graph gives us an idea just how bad (or good) our bottom 10% is. Here, the X axis is the quality score, and the Y axis is the number of times that score occurs across the entire dataset. The scores in this dataset are pretty good and there are only a small number of calls with confidence below 20. However, there is a small spike at 2. We'll want to drop those.

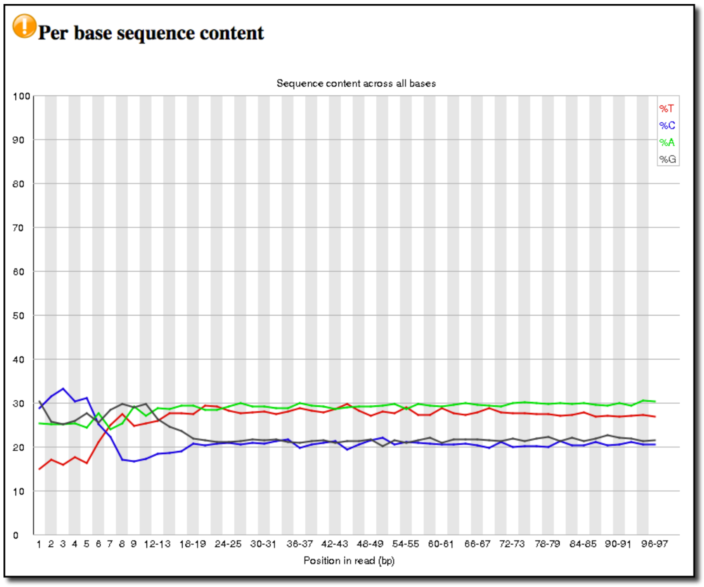
Per base sequence content

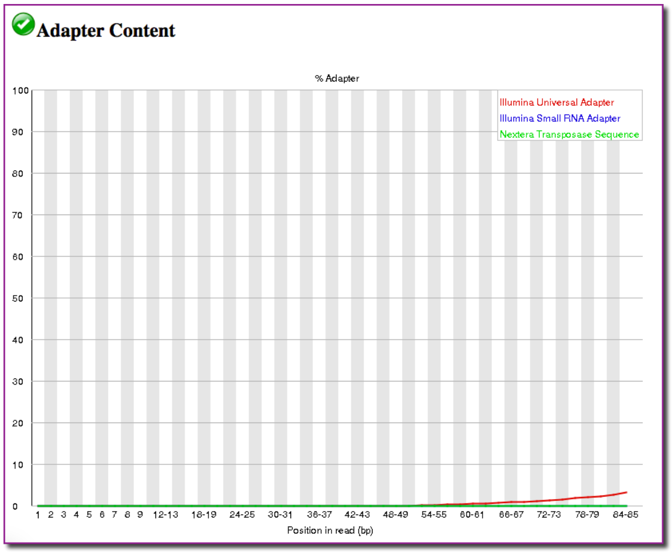
This graph is good at highlighting untrimmed adapter or barcode problems.

Flat lines mean there is no bias for particular bases at particular positions. The turbulence at the front is caused by the use of random hexamers in Illumina library preparation. See

> Hansen, *et al.*, "Biases in Illumina transcriptome sequencing caused by random hexamer priming" *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

The summary is: *It's not our fault.*





Adapter Content

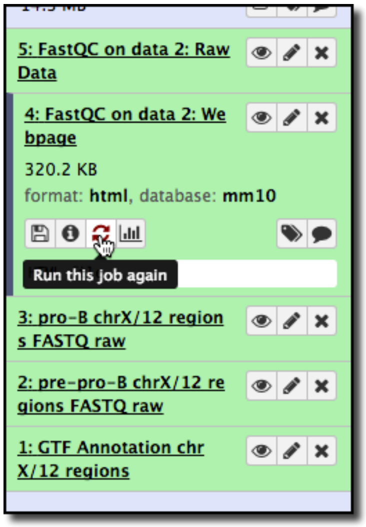
At the end of the reads the dataset is 3 to 4 percent untrimmed adapter. This falls within FastQC's tolerance, but not ours.

And the rest

The remaining tests are either not informative (look at Sequence Length Distribution), flat out good (Per base N content), or look significantly different when the whole dataset is scanned (Per sequence GC content, Sequence Duplication Levels, Kmer Content). We'll ignore all these today.

Rerun FastQC on pro-B dataset

Now let's run FastQC on the other dataset, the pro-B reads. We could rerun FastQC by finding it in the Tool panel again, or, we could rerun FastQC from the History panel. Let's do that.


5: FastQC on data 2: Raw Data

4: FastQC on data 2: Webpage

320.2 KB

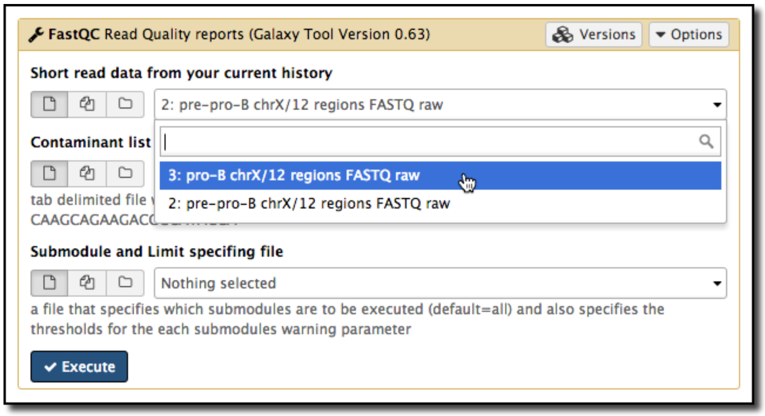
format: **html**, database: **mm10**

3: pro-B chrX/12 regions FASTQ raw

2: pre-pro-B chrX/12 regions FASTQ raw

1: GTF Annotation chrX/12 regions

Preview one of the **FastQC** datasets in your history by *clicking* on its name. In the preview *click* on the **Run this job again** looping arrow icon.
This launches the tool form in the middle panel and pre-populates it with the same settings it was run with before.


FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

2: pre-pro-B chrX/12 regions FASTQ raw

Contaminant list

3: pro-B chrX/12 regions FASTQ raw

2: pre-pro-B chrX/12 regions FASTQ raw

tab delimited file

CAAGCAGAAGACC

Submodule and Limit specifing file

Nothing selected

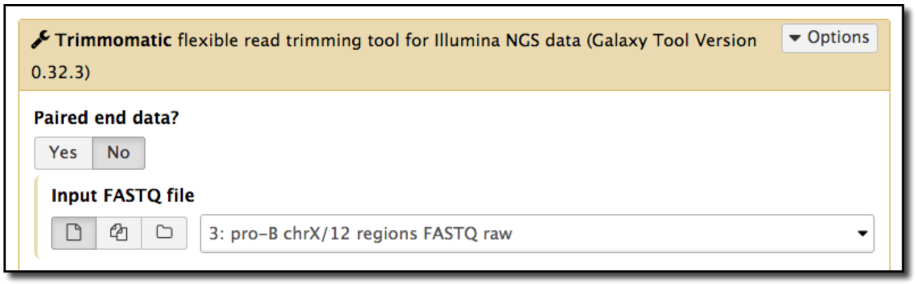
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Click the dataset pulldown menu and *select* **3: pro-B chrX/12 regions FASTQ raw**. Then *click* **Execute**. *Poke* the new **FastQC Webpage** *in the eye* to review the results. The quality is similar.

Improve the quality: Trimmomatic

Use Trimmomatic to separate the good from the bad.

FastQC was the first tool in the **NGS: QC and manipulation** toolbox. **Trimmomatic** is the last. *Click* on it. First, *change* **Paired end data?** to **No**.


Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.32.3)

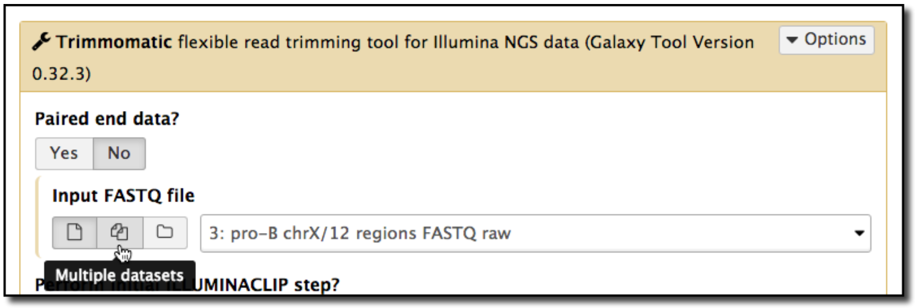
Paired end data?

Input FASTQ file

3: pro-B chrX/12 regions FASTQ raw

Processing Multiple datasets with one submission

We could run Trimmomatic the same way we ran FastQC - by running it on one dataset, and then re-running it on the 2nd.


Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.32.3)

Paired end data?

Input FASTQ file

3: pro-B chrX/12 regions FASTQ raw

UMINACLIP step?

We can achieve the same effect more efficiently by using the **Multiple datasets** feature, another available shortcut.

To enable this *click* on the **Multiple datasets icon** under **Input FASTQ file**.

This changes the parameter selection from a pull-down to a multi-select field. *Select* both datasets.

Select Trimmomatic Filters

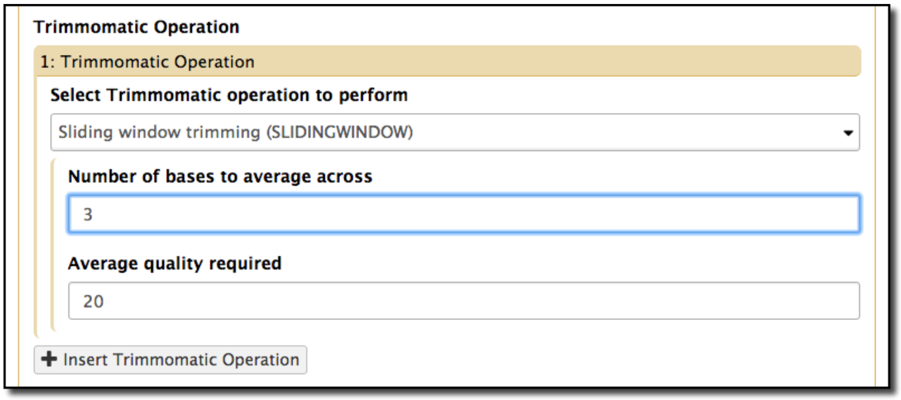
We want to use three different filters in Trimmomatic:

1. **ILLUMINACLIP** - This will trim known Illumina adapters from reads in the dataset.
2. **SLIDINGWINDOW** - Cutting once the average quality within a sliding window falls below a threshold. Trims from both ends
3. **MINLEN** - Drop any reads that are shorter than a given length after trimming.

First, *select* **Yes** under **Perform initial ILLUMINACLIP step?**. Then *change* the **Adapter sequences to use** pulldown to **TruSeq3 (single-ended, for MiSeq and HiSeq)**. Leave other parameters at the default.

Update the default SLIDINGWINDOW step to use the same window definition as the paper. *Change* **Number of bases to average across** to **3**.

To deal with reads that are too short add another operation.



Trimmomatic Operation

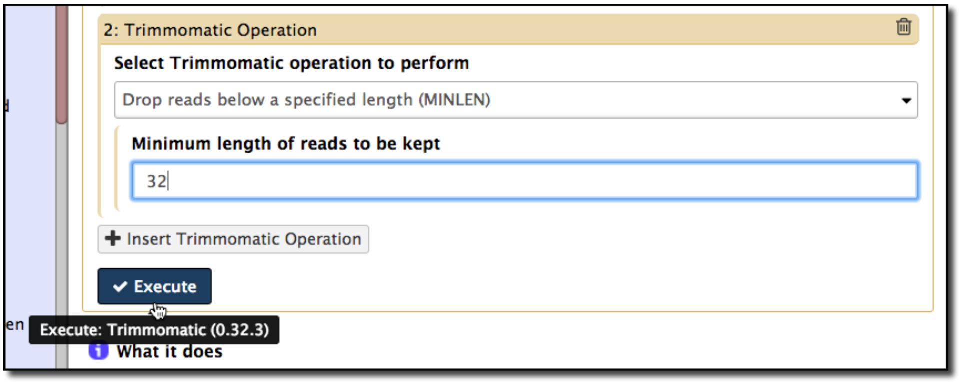
1: Trimmomatic Operation

Select Trimmomatic operation to perform: Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across: 3

Average quality required: 20

Click on **+ Insert Trimmomatic Operation**. *Change* **Select Trimmomatic operation to perform** to **Drop reads below a specified length (MINLEN)** and *set* **Minimum length of reads to be kept** to **32**. All 3 steps are set up. *Click* **Execute**.



2: Trimmomatic Operation

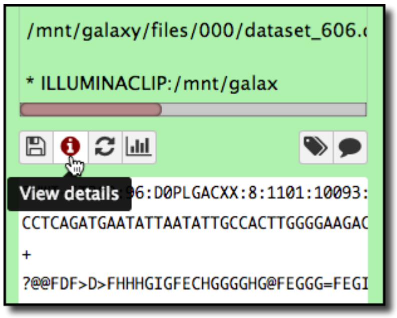
Select Trimmomatic operation to perform: Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept: 32

Trimmomatic Results

To see the net decrease in dataset size, preview the before and after datasets, and compare their size.

To see the full Trimmomatic summary, *click* the ⓘ **icon (View details)** in the dataset preview. This displays the metadata for this dataset.



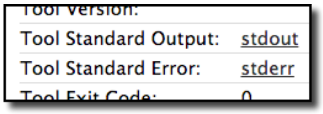
/mnt/galaxy/files/000/dataset_606.

* ILLUMINACLIP:/mnt/galax

```
6:D0PLGACXX:8:1101:10093:
CCTCAGATGAATATTAATATTGCCACTTGGGGAAGAC
+
?@@FDF>D>FHHHGIGFECHGGGGHG@FEGGG=FEGI
```

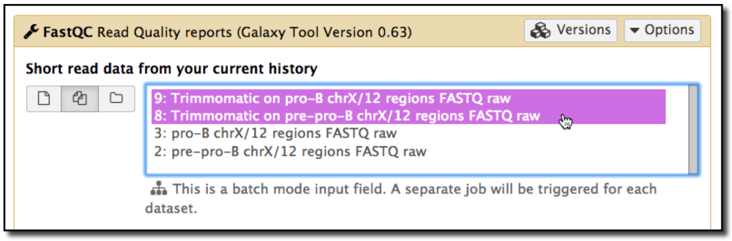
Two links for **stdout** and **stderr** are included with the metadata. These are standard files in Unix to record *standard output* and *standard error* from the tool.

Click **stdout** to see the full summary. This is informative, but let's run FastQC again on both and compare those reports with the original reports.



Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0

Click **FastQC** in the tool panel, *select* the **Multiple datasets icon** under **Short read data from your current history**, and then *select* both **Trimmomatic on … ** datasets. *Click* **Execute.**



FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

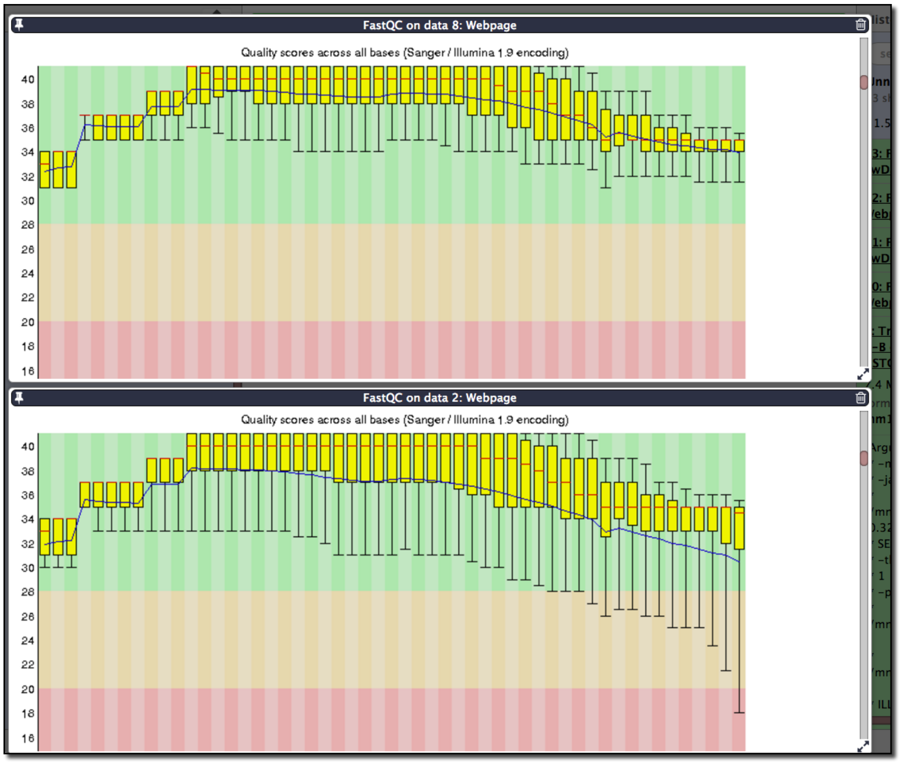
9: Trimmomatic on pro-B chrX/12 regions FASTQ raw
8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw
3: pro-B chrX/12 regions FASTQ raw
2: pre-pro-B chrX/12 regions FASTQ raw

This is a batch mode input field. A separate job will be triggered for each dataset.

Note: We aren't using paired-end data today, but if we were Trimmomatic would automatically maintain read pairings throughout all steps.

Using Galaxy Scratchpad to view multiple datasets

We want to compare pre and post-Trimmomatic FastQC reports. Scratchpad will enable us to do this without going insane. To enable Scratchpad, *click* on the grid icon (▦) in the top menu bar. This turns it yellow (▦)! Now, *scroll down* in your history to first FastQC report on the pre-pro-B data (most likely named **4: FastQC on data 2: Webpage**) and *poke it in the eye*.



FastQC on data 8: Webpage

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

FastQC on data 2: Webpage

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Click somewhere outside the displayed dataset. *Scroll up* the history to the second FastQC report on the pre-pro-B data (most likely the fourth dataset from the top) and *poke it in the eye*.

After some scrolling and resizing you can now compare the two reports at once.. Additional datasets can be added as desired.

We see that the quality has improved.

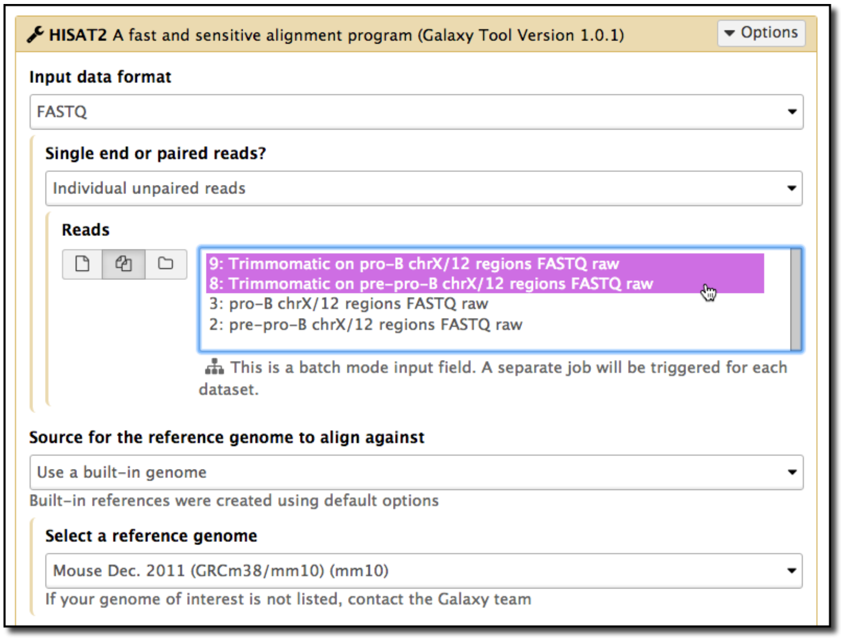
To switch back to the standard view either click outside the displayed datasets or poke the yellow eye (👁). To disable the Scratchbook display, click on the yellow grid icon.

Map the reads

The data has now been cleaned and we are ready to map them against the mouse reference genome.

HISAT2

Bring up the HISAT2 tool form and *change* **Single end or paired reads?** to **Individual unpaired reads** and then *select* the **Multiple datasets icon** under **Reads**. *Select* both **Trimmomatic** datasets.



HISAT2 A fast and sensitive alignment program (Galaxy Tool Version 1.0.1) ▾ Options

Input data format
FASTQ

Single end or paired reads?
Individual unpaired reads

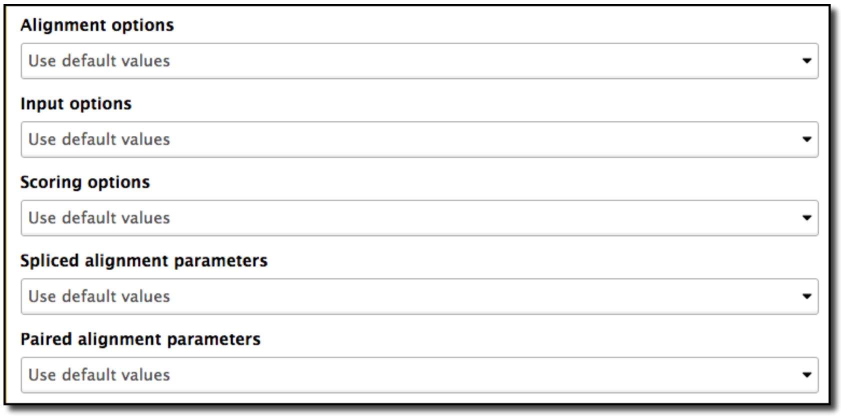
Reads
9: Trimmomatic on pro-B chrX/12 regions FASTQ raw
8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw
3: pro-B chrX/12 regions FASTQ raw
2: pre-pro-B chrX/12 regions FASTQ raw

This is a batch mode input field. A separate job will be triggered for each dataset.

Source for the reference genome to align against
Use a built-in genome
Built-in references were created using default options

Select a reference genome
Mouse Dec. 2011 (GRCm38/mm10) (mm10)
If your genome of interest is not listed, contact the Galaxy team

HISAT is an “option-rich” tool. The HISAT tool wrapper takes a common approach of placing them in sections that can be expanded as needed. HISAT has 5 such sections:



Alignment options
Use default values

Input options
Use default values

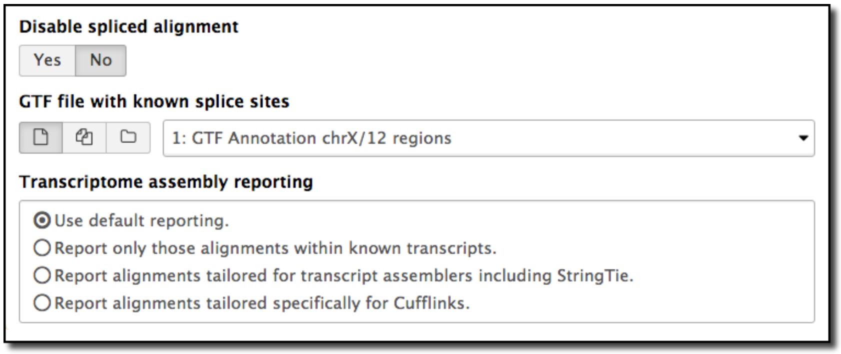
Scoring options
Use default values

Spliced alignment parameters
Use default values

Paired alignment parameters
Use default values

If you expand all sections, HISAT options go from 1 screen to almost 6. Understanding all these options is the right thing to do, but it’s also daunting. One of Galaxy’s strengths is that it allows you to *experiment with tools* and *learn them incrementally*. We are going to start with setting just one parameter.

Change **Spliced alignment parameters** to **Specify spliced alignment parameters**. This causes almost 15 additional options to be displayed, most of them giving you fine-grained control over how reads that are split across multiple exons are scored. Ignore all those and *scroll* down to **GTF file with known splice sites** and *select* **1: GTF Annotation chrX/12 regions** and *click* **Execute**.



Disable spliced alignment
Yes No

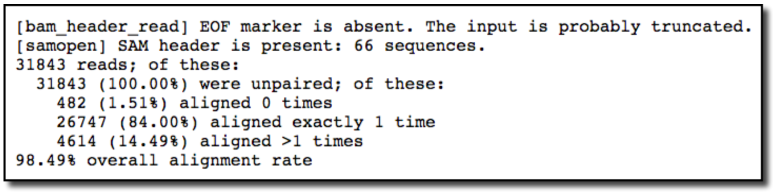
GTF file with known splice sites
1: GTF Annotation chrX/12 regions

Transcriptome assembly reporting
● Use default reporting.
○ Report only those alignments within known transcripts.
○ Report alignments tailored for transcript assemblers including StringTie.
○ Report alignments tailored specifically for Cufflinks.

HISAT2 Results

HISAT2 creates a mapped reads dataset for each of the two inputs. These files are in BAM format, an efficient binary format for representing mapped data.

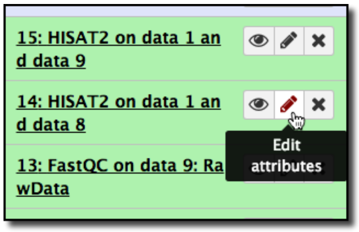
To see the full HISAT2 summary *click* the ⓘ **icon (View details)** in the dataset preview. Then *click* the **stderr** link. Our overall alignment rates are over 98% for both datasets.



```
[bam_header_read] EOF marker is absent. The input is probably truncated.
[samopen] SAM header is present: 66 sequences.
31843 reads; of these:
  31843 (100.00%) were unpaired; of these:
    482 (1.51%) aligned 0 times
    26747 (84.00%) aligned exactly 1 time
    4614 (14.49%) aligned >1 times
98.49% overall alignment rate
```

Updating dataset and history metadata: Renaming

The HISAT2 output datasets have names like **HISAT2 on data 1 and data 8**. Rename them to clearly indicate what's in them.

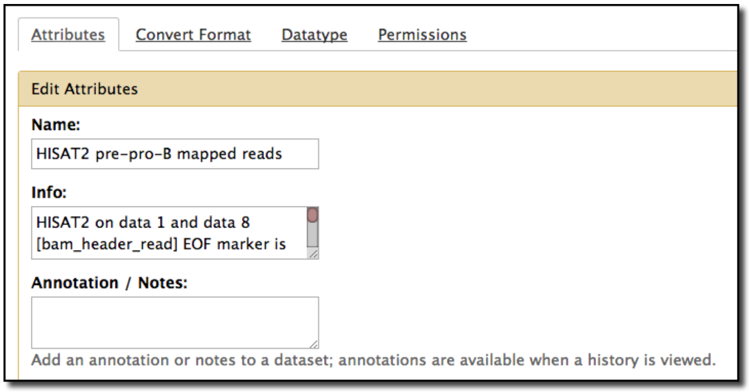


15: HISAT2 on data 1 and data 9

14: HISAT2 on data 1 and data 8

13: FastQC on data 9: RawData

Click on the **pencil (Edit attributes) icon** for the pre-pro-B **HISAT2 dataset**. Name it something like **HISAT2 pre-pro-B mapped reads**. *Click* **Save** and then *repeat* with the **pro-B dataset.**



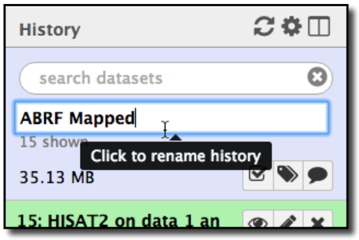
Edit Attributes

Name: HISAT2 pre-pro-B mapped reads

Info: HISAT2 on data 1 and data 8
[bam_header_read] EOF marker is

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.



History

search datasets

ABRF Mapped

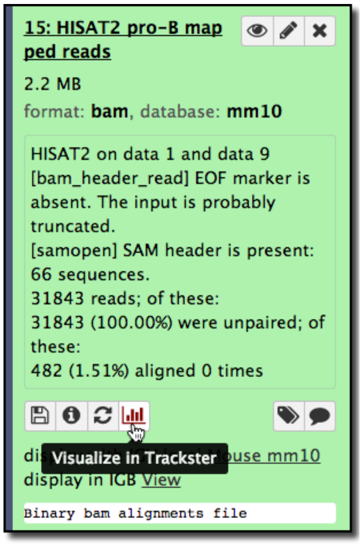
15 shown

35.13 MB

15: HISAT2 on data 1 an

We should also give our history a name. A best practice is to always name your histories. To do this, *click* on **Unnamed history**, *enter* a new and informative history name, and then *press* **return** or **enter**.

But what do the HISAT2 mappings look like?



15: HISAT2 pro-B mapped reads

2.2 MB

format: **bam**, database: **mm10**

HISAT2 on data 1 and data 9
[bam_header_read] EOF marker is absent. The input is probably truncated.
[samopen] SAM header is present: 66 sequences.
31843 reads; of these:
31843 (100.00%) were unpaired; of these:
482 (1.51%) aligned 0 times

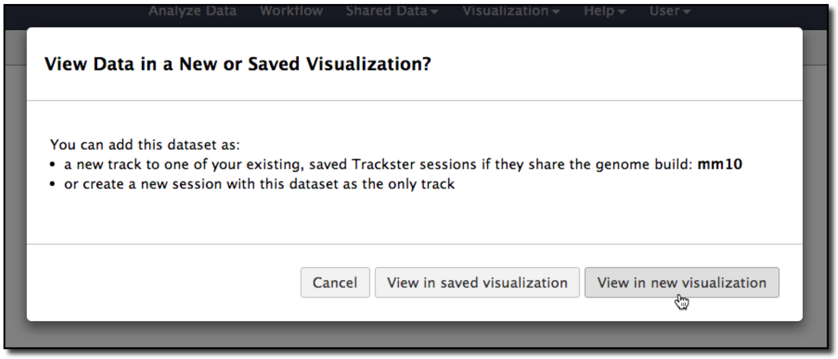
di... use mm10
display in IGB View

Binary bam alignments file

There are summary tools that provide more statistics on things like mapping state (**Flagstat**) and how many reads mapped to each chromosome (**IdxStats**).

We can visually inspect the mappings in the context of the genomic regions they mapped to. Later today you'll use the IGV desktop program to visualize genomic data. Right now, let's learn how to do this inside Galaxy, using the Trackster visualization tool.

To launch Trackster, open a preview of one of the HISAT2 datasets and *click* on the **Visualize in Trackster icon**.

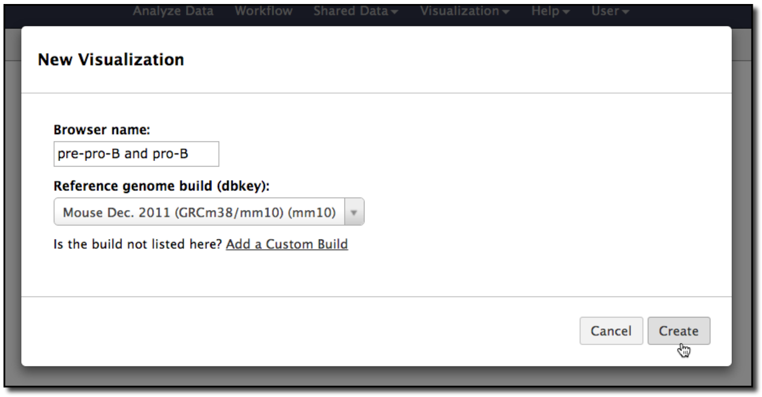


View Data in a New or Saved Visualization?

You can add this dataset as:

- a new track to one of your existing, saved Trackster sessions if they share the genome build: **mm10**
- or create a new session with this dataset as the only track

Select **View in new visualization** and then give the visualization a meaningful name and *click* **Create**.



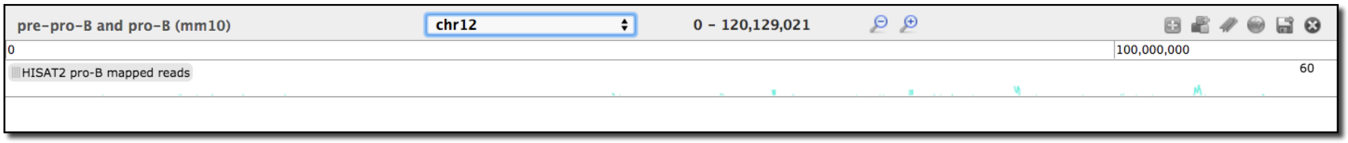
New Visualization

Browser name: pre-pro-B and pro-B

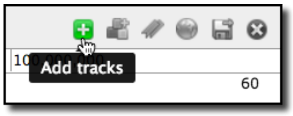
Reference genome build (dbkey): Mouse Dec. 2011 (GRCm38/mm10) (mm10)

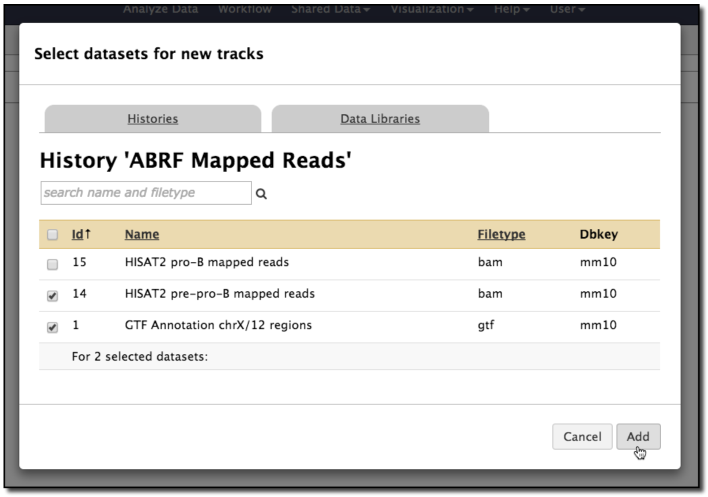
Is the build not listed here? Add a Custom Build

This will bring up a new Trackster browser with a status message saying that it is indexing the dataset for display. *Switch* from chr1 to **chr12** (or chrX), where we have mapped data.



While the track is loading, add other datasets to the display. *Click* on the **+ (Add tracks) icon** in the upper right, and then select the other mapped dataset and the reference annotation.





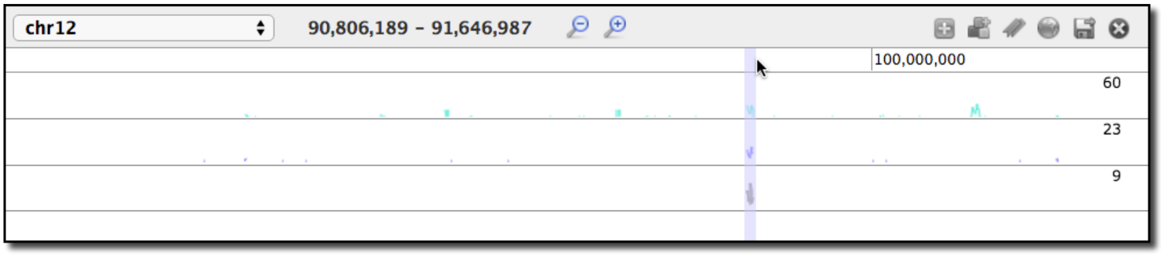
Select datasets for new tracks

History 'ABRF Mapped Reads'

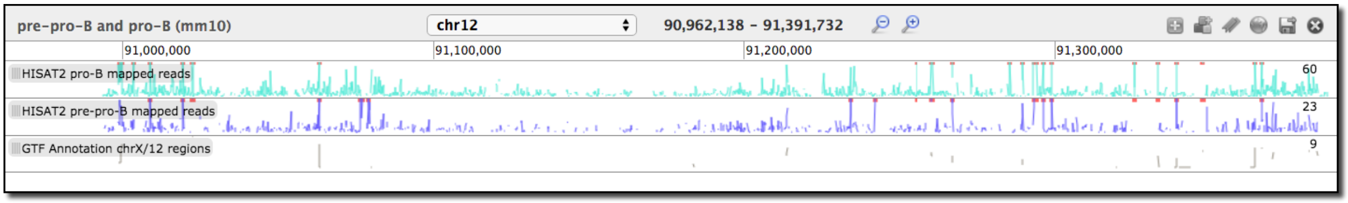
□	Id↑	Name	Filetype	Dbkey
□	15	HISAT2 pro-B mapped reads	bam	mm10
☑	14	HISAT2 pre-pro-B mapped reads	bam	mm10
☑	1	GTF Annotation chrX/12 regions	gtf	mm10

For 2 selected datasets:

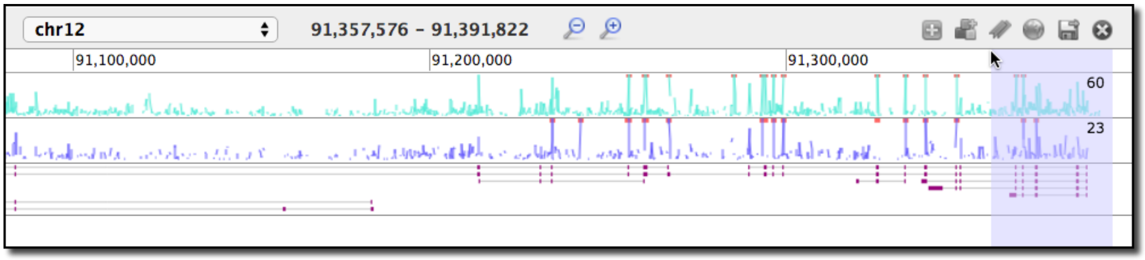
Once those new tracks have loaded, zoom in on the region of interest - the area with the annotations: *Click and drag* on the coordinates track (near the top) and include the entire region with annotation.

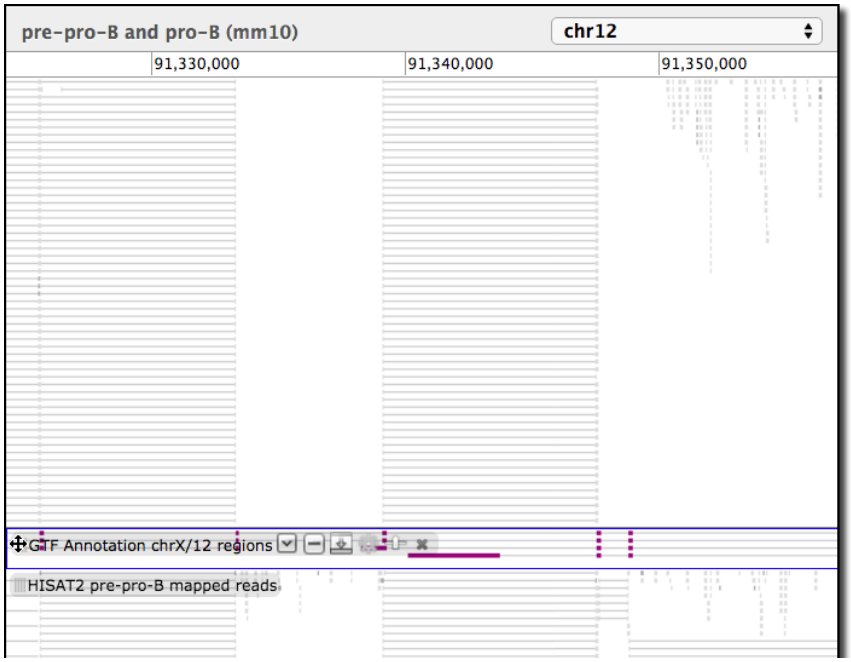


You may have to repeat that a few times to get the region of interest:


pre-pro-B and pro-B (mm10)
chr12
90,962,138 - 91,391,732
91,000,000
91,100,000
91,200,000
91,300,000
HISAT2 pro-B mapped reads
60
HISAT2 pre-pro-B mapped reads
23
GTF Annotation chrX/12 regions
9

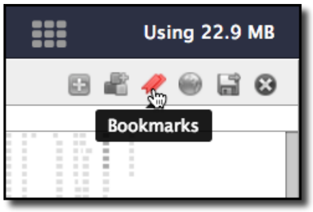
The top 2 tracks show the read depth for any position in this region, and the bottom track shows the annotation (but not very well). *Hover* over the **GTF Annotation chrX/12 regions** track and *click* the ⌄ **Set display mode icon** and *select* **Squish**. To get a better idea of what the aligned reads look like, zoom in the last 5 or so introns on the far right.

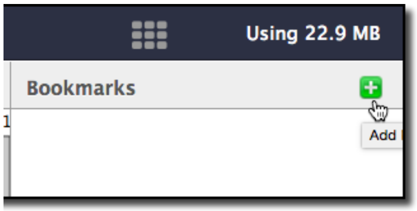


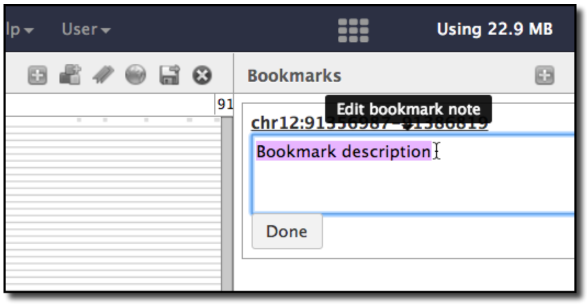

pre-pro-B and pro-B (mm10)
chr12
91,330,000
91,340,000
91,350,000
GTF Annotation chrX/12 regions
HISAT2 pre-pro-B mapped reads

This changes the display mode for all tracks and now we can see individual reads, including where splice junction were introduced to map the reads. Drag the annotation track to the middle by *clicking and dragging* the **|||| vertical lines icon** at the left of the track title upwards to between the two BAM tracks.

We'll want to come back to this location once we have predicted transcripts for these datasets. Bookmark this location by *clicking* on the **bookmark ribbon icon** in the upper right corner. Then *click* the **+ (Add bookmark) icon** and give the bookmark a meaningful description. Finally, *click* the **save icon** to save your visualization.







chr12:91

Bookmark description

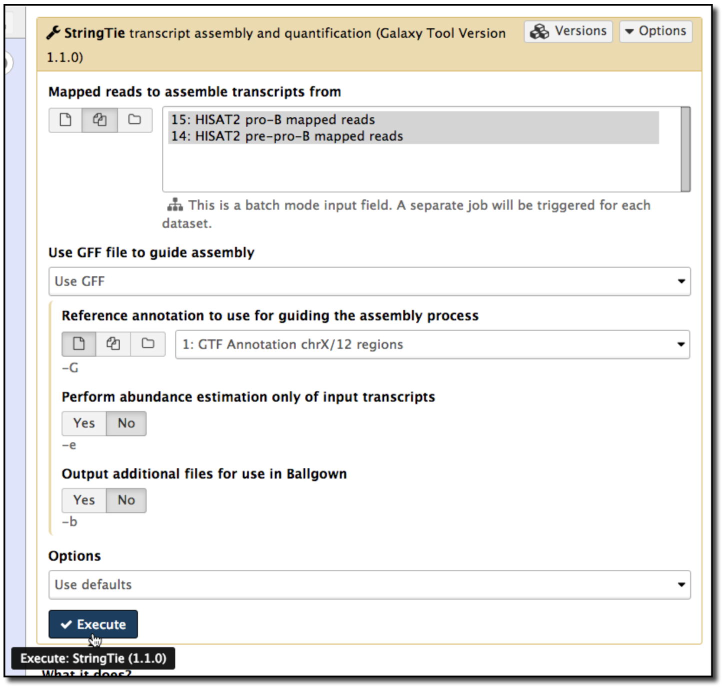
Transcript Prediction with StringTie

HISAT2 makes no prediction about how those reads might assemble into transcripts. That job is handled by StringTie.

Search for **StringTie** in the tool panel and launch the tool form. *Select* **Multiple datasets** under **Mapped reads to assemble transcripts from**, and then *select* both **HISAT2** datasets. *Change* **Use GFF file to guide assembly** to **Use GFF**. *Click* **Execute**.

StringTie generates 3 output datasets for each input dataset:

- Coverage - All the transcripts in the reference GTF that are fully covered by reads.
- Gene abundance estimates - This is created but it's empty because we did not ask for it in the advanced options section.
- Assembled transcripts - This is the one we care most about.



StringTie transcript assembly and quantification (Galaxy Tool Version 1.1.0)

Mapped reads to assemble transcripts from

15: HISAT2 pro-B mapped reads
14: HISAT2 pre-pro-B mapped reads

This is a batch mode input field. A separate job will be triggered for each dataset.

Use GFF file to guide assembly: Use GFF

Reference annotation to use for guiding the assembly process: 1: GTF Annotation chrX/12 regions
-G

Perform abundance estimation only of input transcripts
-e

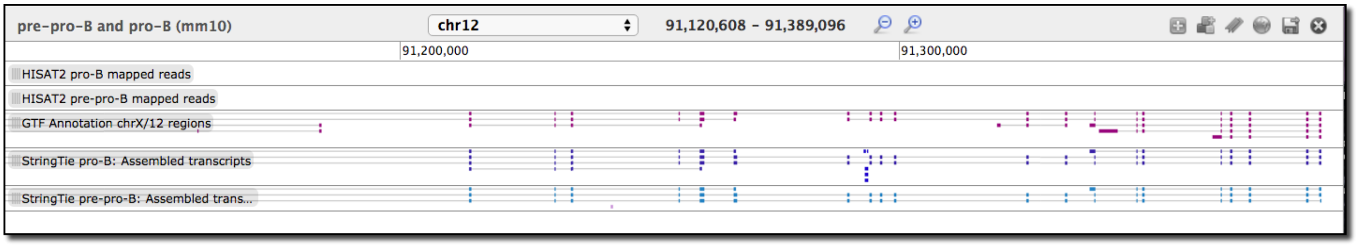
Output additional files for use in Ballgown
-b

Options: Use defaults

What it does?

Rename the two **assembled transcripts** datasets to include **pre-pro-B** or **pro-B** in the dataset name. Then add the assembled transcripts datasets to the visualization you created earlier.

Once in Trackster, add the other assembled transcripts dataset and hide the content of the two mapped reads datasets, and zoom out a few times. Set the display mode for both new datasets to Squish.


pre–pro–B and pro–B (mm10)
chr12
91,120,608 – 91,389,096
91,200,000
91,300,000
HISAT2 pro-B mapped reads
HISAT2 pre-pro-B mapped reads
GTF Annotation chrX/12 regions
StringTie pro-B: Assembled transcripts
StringTie pre-pro-B: Assembled trans...

A couple of things to note about this region:

- the pro-B assembled transcripts have some transcripts that are not in the reference, and even some new exons
- the pre-pro-b assembled transcripts are largely a subset of the pro-B predicted transcripts.
- several annotated transcripts and exons are not detected at all in these datasets.

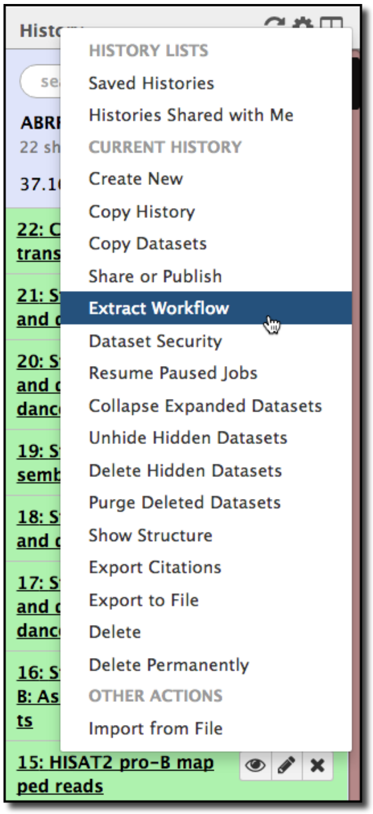
Save the visualization and go back to **Analyze Data**.

And there is more!

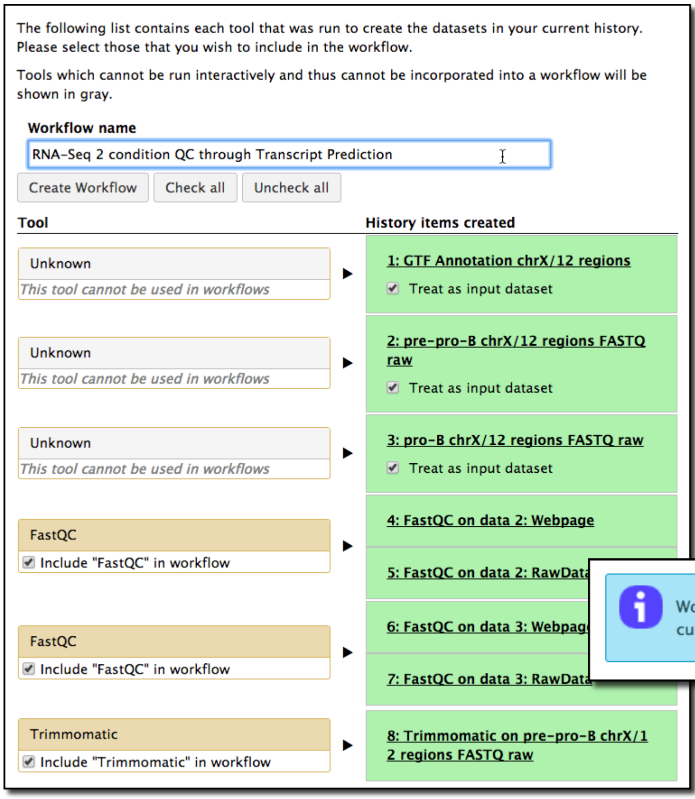
Workflows: repeating an analysis

We'll want to apply the same analysis to our full datasets, and maybe with other similar experiments as well. To do this create a workflow - a repeatable recipe for doing an analysis.

Galaxy supports *de novo* workflow creation, and creating them from histories. To create a workflow from our history, *click* the **cog** at the top of the history panel, and then *select* **Extract Workflow.**



This takes you to the create workflow page, where you can specify which of the steps and input in your current history you wish to include in the new workflow.



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

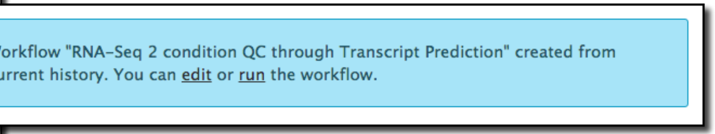
Workflow name

RNA-Seq 2 condition QC through Transcript Prediction

Create Workflow | Check all | Uncheck all

Tool		History items created
Unknown *This tool cannot be used in workflows*	▶	1: GTF Annotation chrX/12 regions ☑ Treat as input dataset
Unknown *This tool cannot be used in workflows*	▶	2: pre-pro-B chrX/12 regions FASTQ raw ☑ Treat as input dataset
Unknown *This tool cannot be used in workflows*	▶	3: pro-B chrX/12 regions FASTQ raw ☑ Treat as input dataset
FastQC ☑ Include "FastQC" in workflow	▶	4: FastQC on data 2: Webpage 5: FastQC on data 2: RawData
FastQC ☑ Include "FastQC" in workflow	▶	6: FastQC on data 3: Webpag 7: FastQC on data 3: RawData
Trimmomatic ☑ Include "Trimmomatic" in workflow	▶	8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw

Give the new workflow an informative name and *click* **Create Workflow**. This message appears:

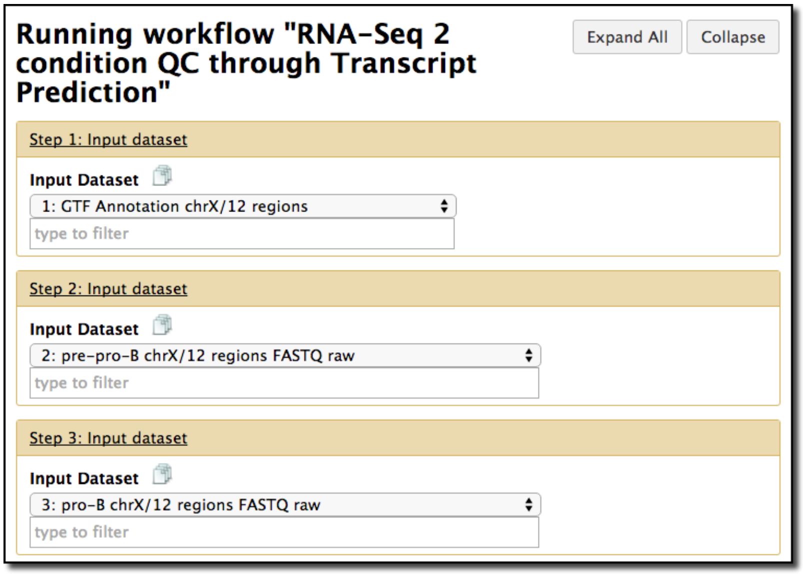


Workflow "RNA-Seq 2 condition QC through Transcript Prediction" created from current history. You can edit or run the workflow.

Test the new workflow on the exact same inputs that we just analyzed manually

Click the **run** link in the message (or you can also get there by clicking Workflow in the top bar). This brings up a form asking you to define inputs to the workflow. *Set* the first three input datasets for this run to the first three datasets from your current history:
Scroll down to the bottom of the form and *check* **Send results to a new history**. Name the workflow and *click* **Run workflow.**

This displays an enormous green box in the middle panel. *Click* the link near the top of the box to get to the new history.



Running workflow "RNA-Seq 2 condition QC through Transcript Prediction"

Step 1: Input dataset

Input Dataset: 1: GTF Annotation chrX/12 regions

type to filter

Step 2: Input dataset

Input Dataset: 2: pre-pro-B chrX/12 regions FASTQ raw

type to filter

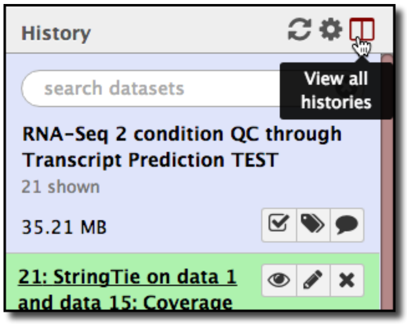
Step 3: Input dataset

Input Dataset: 3: pro-B chrX/12 regions FASTQ raw

type to filter

Working with multiple histories

Running the workflow and sending results to a new history means that we now have two histories in this Galaxy instance. There are a couple of ways to see all your existing histories and to switch between them. The most recent and easiest to use is the all histories view. This can be accessed by *clicking* on the **table icon (View all histories)** at the top of the history panel.



RNA-Seq 2 condition QC through Transcript Prediction TEST

21 shown

35.21 MB

21: StringTie on data 1 and data 15: Coverage

The all histories view presents all your saved histories (we only have two), with the current one pinned at the left, and your other histories listed in reverse chronological order, from left to right.

Sharing and publishing

Galaxy histories, workflows, and visualization can all be shared and published with Galaxy. Sharing in Galaxy means sharing something with someone else either directly with their Galaxy account, or by creating a URL that can be shared. In addition Galaxy objects can be published, making it easy for anyone to discover the object. Anything that is published will be listed in the appropriate Shared Data section.

Galaxy pages

The datasets and analysis used in the paper (and in this tutorial) are available here:
https://usegalaxy.org/u/thereddylab/p/prediction-of-gene-activity-in-early-b-cell-development-based-on-an-integrative-multi-omics-analysis

This is a *Galaxy Page*, a document integrated into a Galaxy server that embeds and provides direct links to Galaxy objects such as histories, workflows, and datasets. Galaxy Pages are a way to bundle together all related analysis and to describe the semantics of your analysis.