Main pipeline

About

Mammalian RNA is regulated through interactions of RNA-binding proteins (RBPs) with their target transcripts. UV-crosslinking and immunoprecipitation combining high-throughput sequencing (CLIP-seq) is able to profile genome-wide RBP-binding regions accurately and efficiently. However, there is few tool to analyze the data. Here we present PIPECLIP, a web tool which provides a pipeline for both bioinformaticians and biologist to identify the most likely cross-linking sites from PAR-CLIP, HITS-CLIP and iCLIP sequencing data.

Parameters

Input

SAM file from any mapping tool. Make sure the SAM file contains its header, or there will be an error

CLIP type

HITS-CLIP: Deletions, insertions and substitutions will be analyzed separately.

PAR-CLIP (4SU): Only T->C substitution will be analyzed.

PAR-CLIP (6SG): Only G->A substitution will be analyzed.

iCLIP: The 1st nucleotide of each read's 5' end will be analyzed.

Remove PCR duplicate

Look for a representative reads for all the reads which have the same genomic starting location. Two methods are provided:

- Remove by starting location: A represent read will be selected from a bunch of reads with the same start location
- Remove by sequence: A represent read will be selected from a bunch of reads with exactly the same sequence

Shortest matched segment length

This length is the sum of perfectly matched nucleotides number and insertion number. Reads whose matched segment length less than threshold will not be included in further analysis

Maximum mismatch number

Although this can be set during mapping, here we still provide this filter in case users want to try more stringent criteria than what they set during mapping

Distribution

Three distribution models for identifying enriched clusters are provided: Poisson, Negative Binomial (default), Zero-Truncated Poisson.

FDR

There are two FDR thresholds for clusters and mutations selection respectively. Default for enriched clusters and reliable mutations is 0.05

Note: Even if you select to remove PCR duplicates for iCLIP, the program will not do it.

Important output files

1: galaxy_*_crosslinking.pipeclip.txt

FINAL crosslinking sites for the experiment. * can be substitution/deletion/truncation/insertion. Columns are:

field	*Description*
chr	Chromosome of the crosslinking site
start	Start position of crosslinking site
stop	Stop position of crosslinking site
cluster_name	Name of the crosslinking site
reads_count	Reads count for the crosslinking site
strand	Strand of the crosslinking site
cluster_fdr	FDR for the cluster
crosslinking_fisherP	Fisher p value combined by the cluster and mutations
mutation_pos	Reliable mutation positions in the crosslinking site, delimited by comma
mutation_name	Reliable mutation name in the crosslinking site, delimited by comma. This can be used to retrieve detail mutation information from reliableMutation.bed

2: galaxy.enrichedClusters.pipeclip.bed

Enriched cluster. In extended bed format. Each column is explained by column names.

3: galaxy.reliableMutations.pipeclip.bed

Reliable mutation. In extended bed format. Each column is explained by column names.

Contact

If you have any comments, suggestions, questions, bug reports, etc., fell free to contact Beibei.Chen@UTSouthwestern.edu or mins.kim@utsouthwestern.edu

Barcode removal description:

About

This script is exclusively for iCLIP raw fastq data.

For iCLIP, barcodes are added to individual reads before PCR amplification during sample preparation. By doing this, real multiple copies can be identified from PCR duplicates since it is almost impossible for them to have a same barcode.

This script takes fastq as input. Reads have a same barcode and the same sequences are regarded as PCR duplicates and only one of them will be kept in the output file (barcode trimmed). Other reads will be kept after trimming the barcode.

The first n nucleotides of each read will be regarded as barcode. (n is a integer, given by user, default is 3nt) Sequencing error is not considered. If a float is provided, the figures behind decimal will be discarded.

Parameters

Raw fastq file and barcode length (integer).

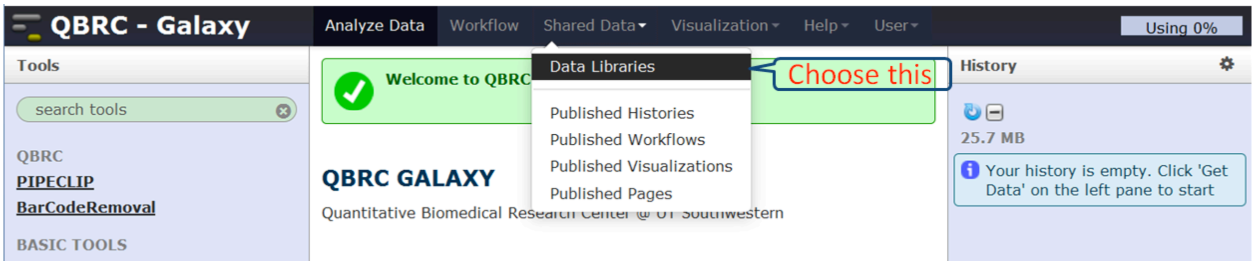
Output:

Fastq file without barcode in each read.

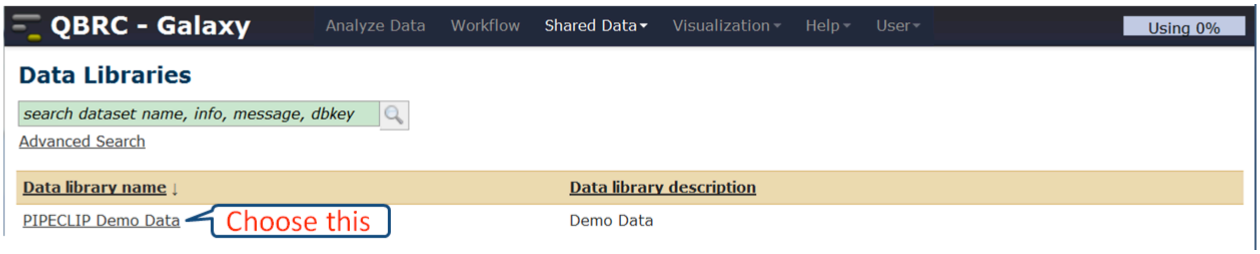
Detail manual

Step 1: Import data

Choose demo data from “Shared Data”:



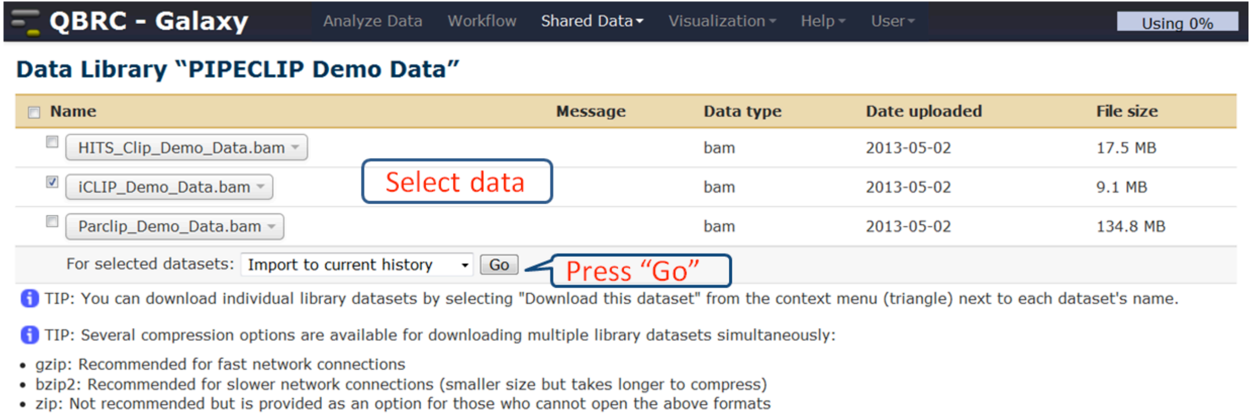
And Click “PIPECLIP Demo Data”:



Data Libraries

Data library name ↓	Data library description
PIPECLIP Demo Data	Demo Data

Select data and “Import to current history”, then press “Go” to



Data Library “PIPECLIP Demo Data”

□ Name	Message	Data type	Date uploaded	File size
□ HITS_Clip_Demo_Data.bam		bam	2013-05-02	17.5 MB
☑ iCLIP_Demo_Data.bam		bam	2013-05-02	9.1 MB
□ Parclip_Demo_Data.bam		bam	2013-05-02	134.8 MB

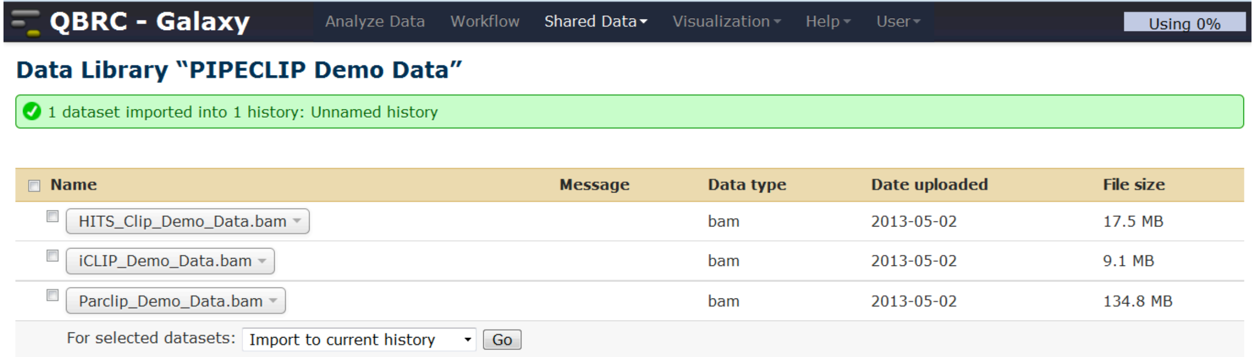
For selected datasets: Import to current history Go

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Screen will show a notice in green background to indicate importing data is successful:



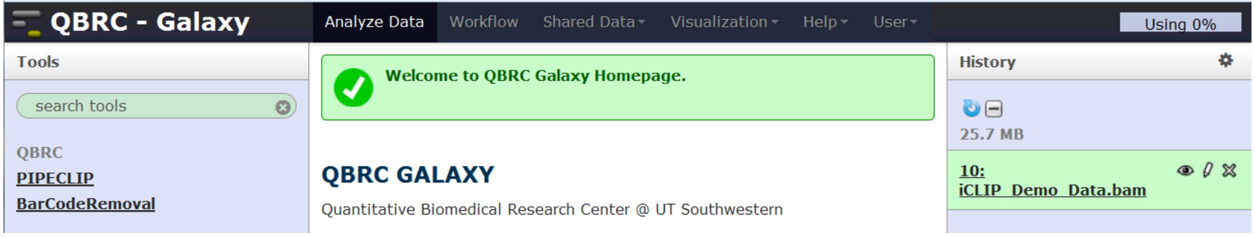
Data Library “PIPECLIP Demo Data”

1 dataset imported into 1 history: Unnamed history

□ Name	Message	Data type	Date uploaded	File size
□ HITS_Clip_Demo_Data.bam		bam	2013-05-02	17.5 MB
□ iCLIP_Demo_Data.bam		bam	2013-05-02	9.1 MB
□ Parclip_Demo_Data.bam		bam	2013-05-02	134.8 MB

For selected datasets: Import to current history Go

Click “QBRC-Galaxy” logo on the top left to return to the main screen. Data just been imported will be shown on the right panel.



Welcome to QBRC Galaxy Homepage.

QBRC GALAXY

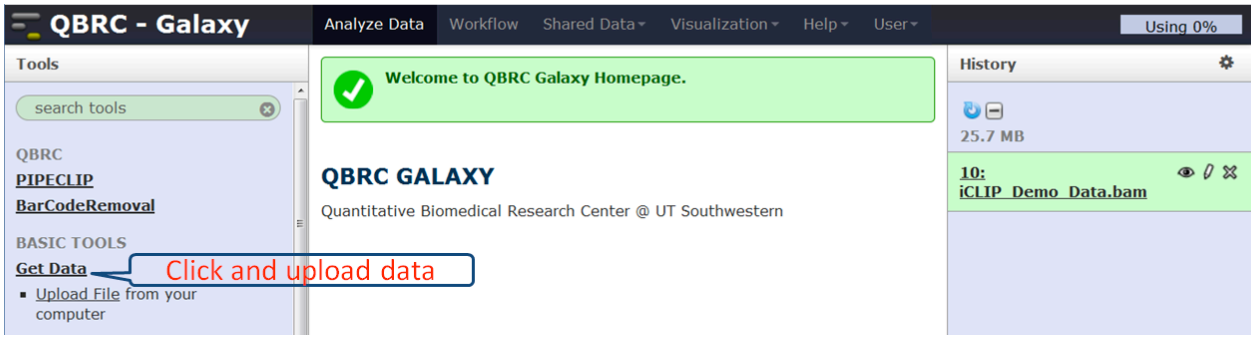
Quantitative Biomedical Research Center @ UT Southwestern

25.7 MB

10: iCLIP_Demo_Data.bam

OR:

You can use “Upload File from your computer” in “Get Data” to upload your own data.



Welcome to QBRC Galaxy Homepage.

QBRC GALAXY

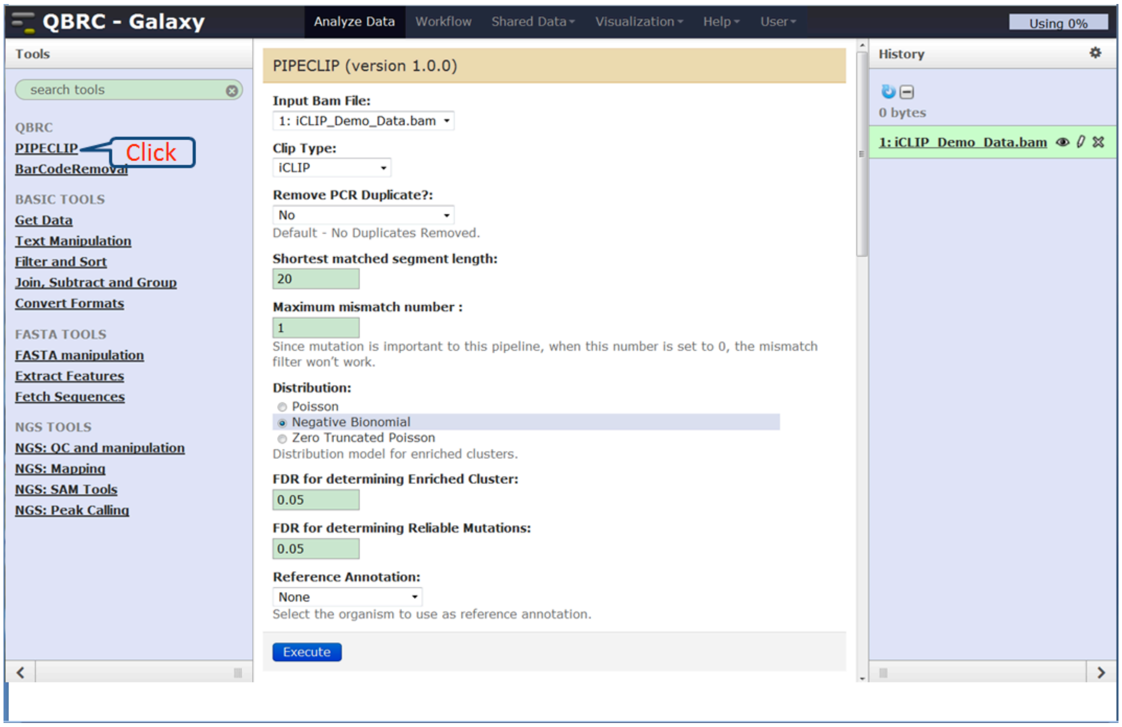
Quantitative Biomedical Research Center @ UT Southwestern

25.7 MB

10: iCLIP_Demo_Data.bam

Step 2: Use PIPECLIP to run the data:

Click “PIPECLIP” on the left panel and the pipeline parameters will appear.



PIPECLIP (version 1.0.0)

Input Bam File: 1: iCLIP_Demo_Data.bam

Clip Type: iCLIP

Remove PCR Duplicate?: No

Default - No Duplicates Removed.

Shortest matched segment length: 20

Maximum mismatch number : 1

Since mutation is important to this pipeline, when this number is set to 0, the mismatch filter won’t work.

Distribution:

○ Poisson
● Negative Bionomial
○ Zero Truncated Poisson

Distribution model for enriched clusters.

FDR for determining Enriched Cluster: 0.05

FDR for determining Reliable Mutations: 0.05

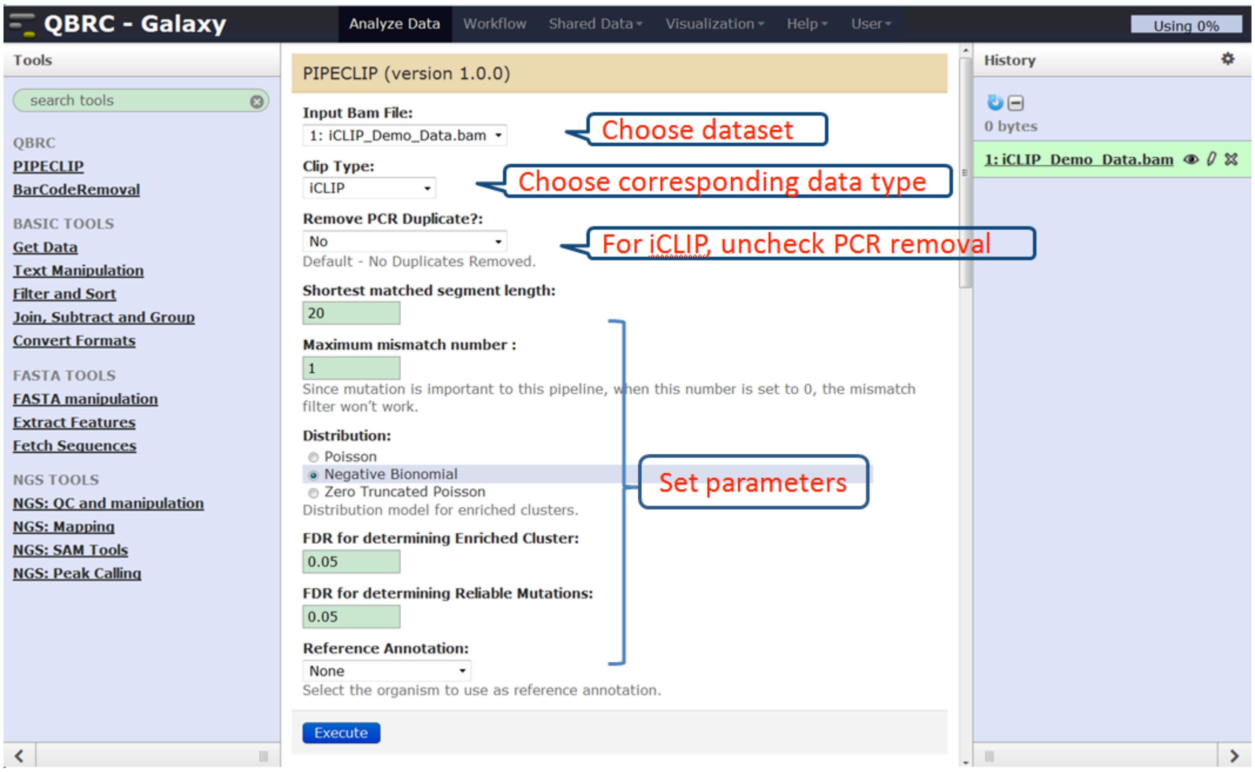
Reference Annotation: None

Select the organism to use as reference annotation.

0 bytes

1: iCLIP_Demo_Data.bam

Choose data and set parameters, then press “Execute”:
For iCLIP, please uncheck PCR removal.



PIPECLIP (version 1.0.0)

Input Bam File: 1: iCLIP_Demo_Data.bam

Clip Type: iCLIP

Remove PCR Duplicate?: No

Default - No Duplicates Removed.

Shortest matched segment length: 20

Maximum mismatch number : 1

Since mutation is important to this pipeline, when this number is set to 0, the mismatch filter won’t work.

Distribution:

○ Poisson
● Negative Bionomial
○ Zero Truncated Poisson

Distribution model for enriched clusters.

FDR for determining Enriched Cluster: 0.05

FDR for determining Reliable Mutations: 0.05

Reference Annotation: None

Select the organism to use as reference annotation.

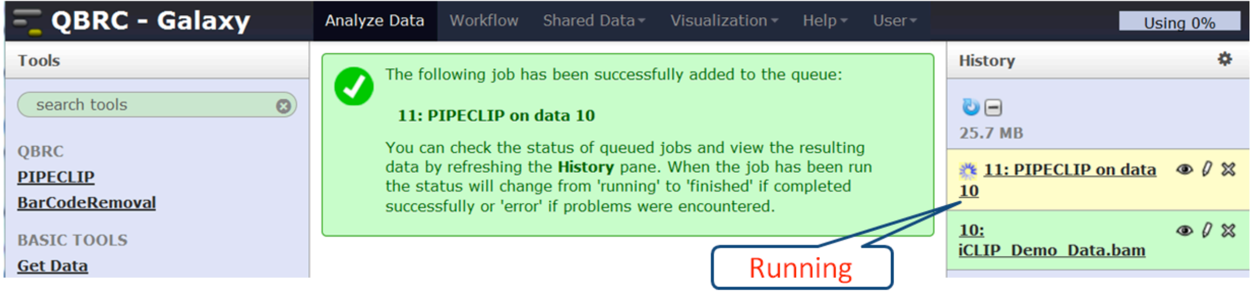
Choose dataset

Choose corresponding data type

For iCLIP, uncheck PCR removal

Set parameters

After press “Execute”, there will be a notice saying the job is submitted successfully and it will appear in right panel. Yellow background means the program is running.



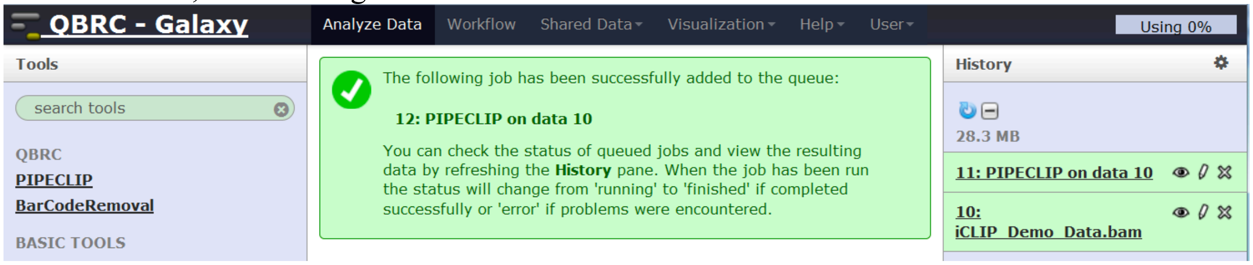
The following job has been successfully added to the queue:

11: PIPECLIP on data 10

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Running

When finished, it will turn green:

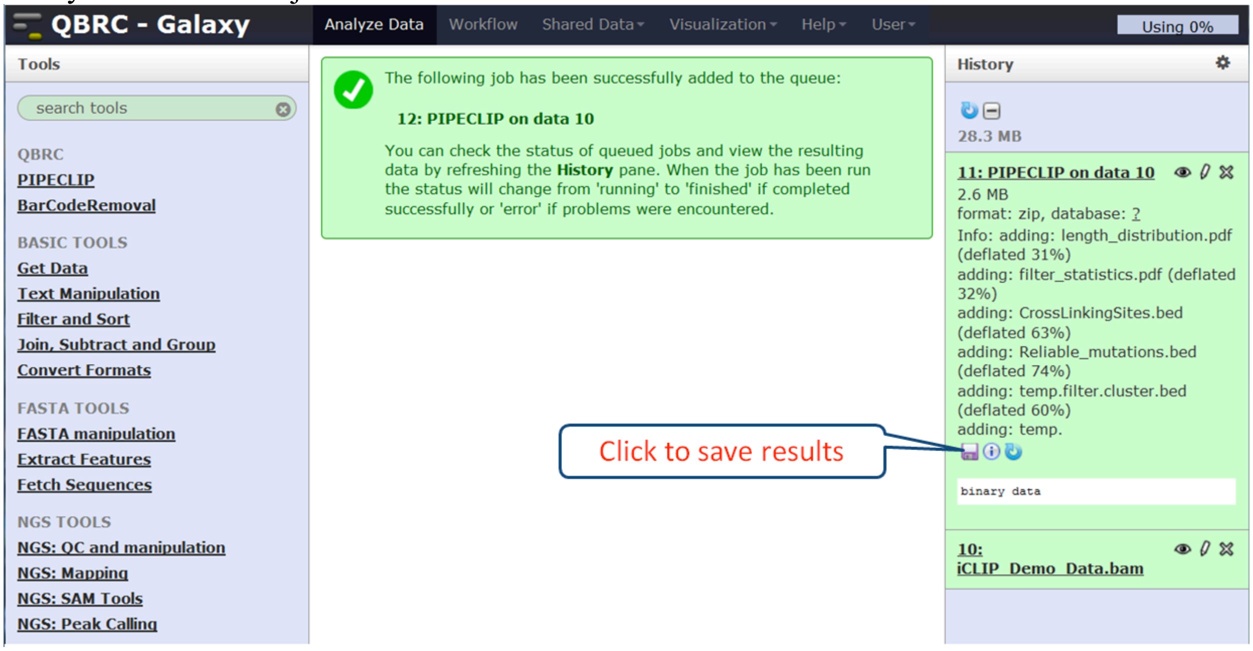


The following job has been successfully added to the queue:

12: PIPECLIP on data 10

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

And you can click the job name to view the detailed information and save results:



The following job has been successfully added to the queue:

12: PIPECLIP on data 10

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

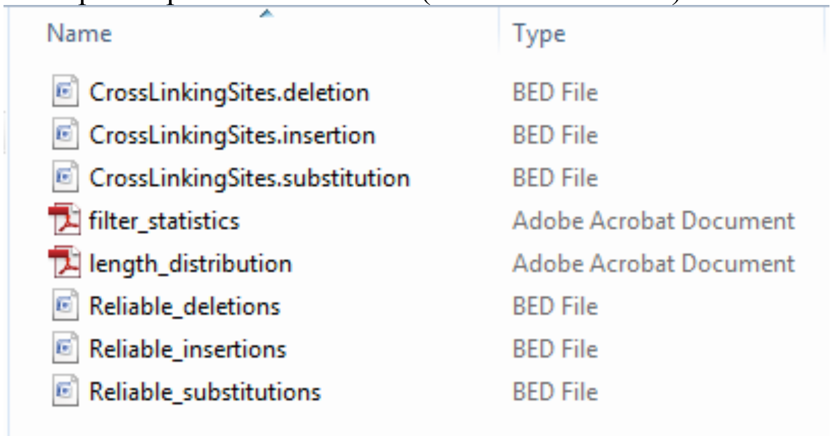
11: PIPECLIP on data 10
2.6 MB
format: zip, database: ?
Info: adding: length_distribution.pdf (deflated 31%)
adding: filter_statistics.pdf (deflated 32%)
adding: CrossLinkingSites.bed (deflated 63%)
adding: Reliable_mutations.bed (deflated 74%)
adding: temp.filter.cluster.bed (deflated 60%)
adding: temp.

binary data

10: iCLIP_Demo_Data.bam

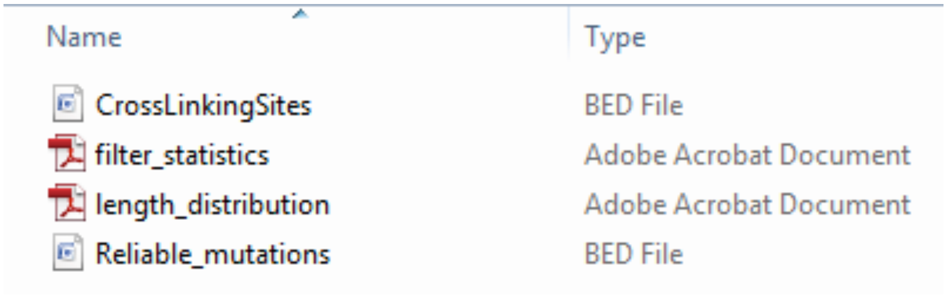
Click to save results

Output:
Example output for HITS-CLIP (without annotation):



Name	Type
CrossLinkingSites.deletion	BED File
CrossLinkingSites.insertion	BED File
CrossLinkingSites.substitution	BED File
filter_statistics	Adobe Acrobat Document
length_distribution	Adobe Acrobat Document
Reliable_deletions	BED File
Reliable_insertions	BED File
Reliable_substitutions	BED File

Example output for iCLIP and PAR-CLIP:



Name	Type
CrossLinkingSites	BED File
filter_statistics	Adobe Acrobat Document
length_distribution	Adobe Acrobat Document
Reliable_mutations	BED File

CrosslinkingSites.*.bed are the final output. For HITS-CLIP, there are 3 files, for PAR-CLIP and iCLIP, there is only one final bed output.
Output columns:
Chr: chromosome
Start: start position of the cross-linking region
End: stop position of the cross-linking region
Name: name of the cross-linking region
ReadsCount: reads number in the cross-linking region
Strand: the strand on which the cross-linking region is on
Mutation_locations: start positions of mutations that fall in the cross-linking region, divided by comma
Mutation_km: (k,m) values for each reliable mutation, divided by comma
-log(q): Fisher's method combined FDR

Filter_statistics.pdf shows the reads number remaining after each SAM/BAM filtering step.
Length_distribution.pdf shows the matched length distribution for the reads remaining after whole filtering step.

Reliable_Mutations/*tions.bed contains detailed information of all the reliable mutations. The ones not within any enriched clusters will also be included in this file.
Output columns:
Chr: chromosome
Start: start position of the mutation
End: stop position of the mutation
Name: name of the mutation
-log(q): p value after BH multiple test correction in –log scale
Strand: the strand on which the mutation is on
Type: type of mutation
K: number of reads covers the position of mutation
M: number of same type mutation at the same position