

Module 1/6: Analyses ADN

- NGS Introduction
- Reads Quality Control
- Reads Cleaning
- Aligning reads on reference
- Alignment parameters
- Reads duplicates

→ *Hélène Touzet*

→ *Hélène*

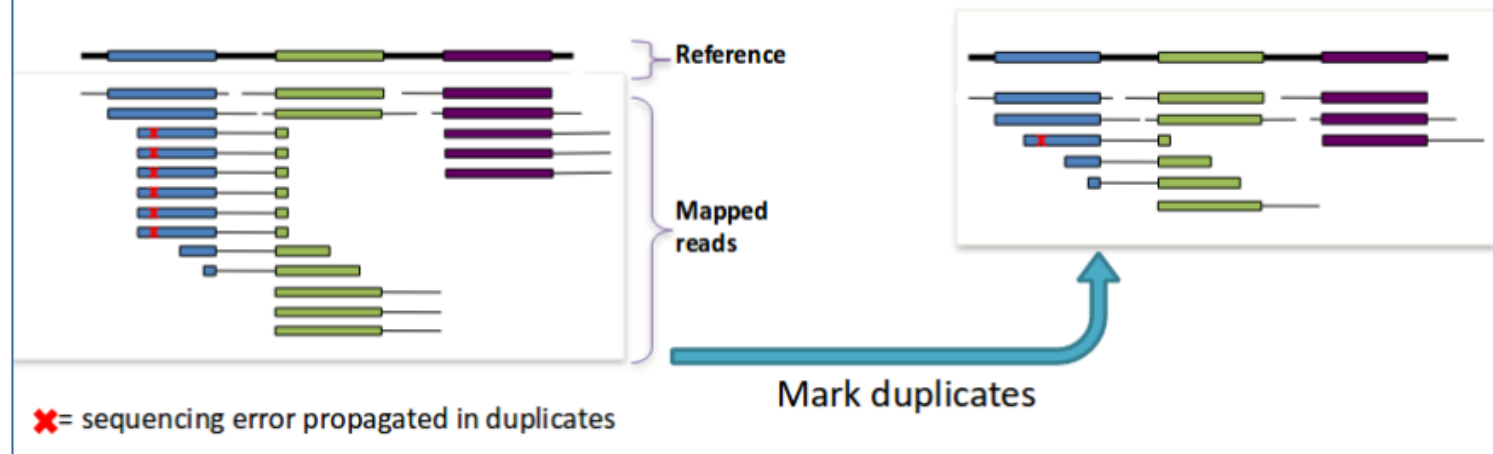
Touzet

→ **Practical #3**

Cleaning duplicated reads

Why mark duplicates?

- Duplicates = sets of reads pairs with same unclipped alignment start and unclipped alignment end
- Suspected to be **non-independent measurements** of a sequence
 - Sampled from the exact same template of DNA
 - Violates assumptions of variant calling
- Errors in sample/library prep will get propagated to *all* the duplicates
 - Just pick the “best” copy – mitigates the effects of errors



Source : GATK Marking duplicates

https://software.broadinstitute.org/gatk/events/slides/1511/Presentations/GATKwh9-3-Marking_duplicates.pdf

Picard / MarkDuplicate

Tools

MarkDuplicates

Conversion and manipulation

MarkDuplicatesWithMateCigar
examine aligned records in BAM datasets to locate duplicate molecules

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

Additional information about Picard tools is available from Picard web site at <http://broadinstitute.github.io/picard/>

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.7.1.0)

Select SAM/BAM dataset or dataset collection

19: HG0101_bowtie2.bam
18: HG0101_BWA.bam

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

If empty, upload or import a SAM/BAM dataset

Comment

+ Insert Comment

You can provide multiple comments

If true do not write duplicates to the output file instead of writing them with appropriate flags set

Yes No

REMOVE_DUPLICATES; default=False

23: MarkDuplicates on data
19: MarkDuplicates BAM
output

22: MarkDuplicates on data
19: MarkDuplicate metrics

21: MarkDuplicates on data
18: MarkDuplicates BAM
output

20: MarkDuplicates on data
18: MarkDuplicate metrics

23: HG0101_bowtie2_MD.bam

22: HG0101_BWA_MD_metrics

21: HG0101_BWA_MD.bam

20: HG0101_bowtie2_MD_metrics

	BWA	Bowtie2
UNPAIRED_READS_EXAMINED	18	19
READ_PAIRS_EXAMINED	4043	4044
SECONDARY_OR_SUPPLEMENTARY_RDS	0	3
UNMAPPED_READS	22	19
UNPAIRED_READ_DUPLICATES	0	0
READ_PAIR_DUPLICATES	12	12
READ_PAIR_OPTICAL_DUPLICATES	0	0
PERCENT_DUPLICATION	0,002962	0,00296
ESTIMATED_LIBRARY_SIZE	679728	680065

Alignment count : *samtools flagstat*

BWA

```
8129 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
3 + 0 supplementary
24 + 0 duplicates
8110 + 0 mapped (99.77%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
7980 + 0 properly paired (98.20%:-nan%)
8088 + 0 with itself and mate mapped
19 + 0 singletons (0.23%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Bowtie 2

```
8126 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
24 + 0 duplicates
8104 + 0 mapped (99.73%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
8074 + 0 properly paired (99.36%:-nan%)
8086 + 0 with itself and mate mapped
18 + 0 singletons (0.22%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Tools

flagstat

Quality Control

Flagstat tabulate descriptive stats for BAM dataset

Flagstat tabulate descriptive stats for BAM dataset (Galaxy Version 2.0) Options

BAM File to Convert

31: HG0101_bowtie2.bam
30: NG0101_BWA.bam

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Execute

33: Flagstat on data 31

32: Flagstat on data 30

33: HG101_bowtie2_flagstat

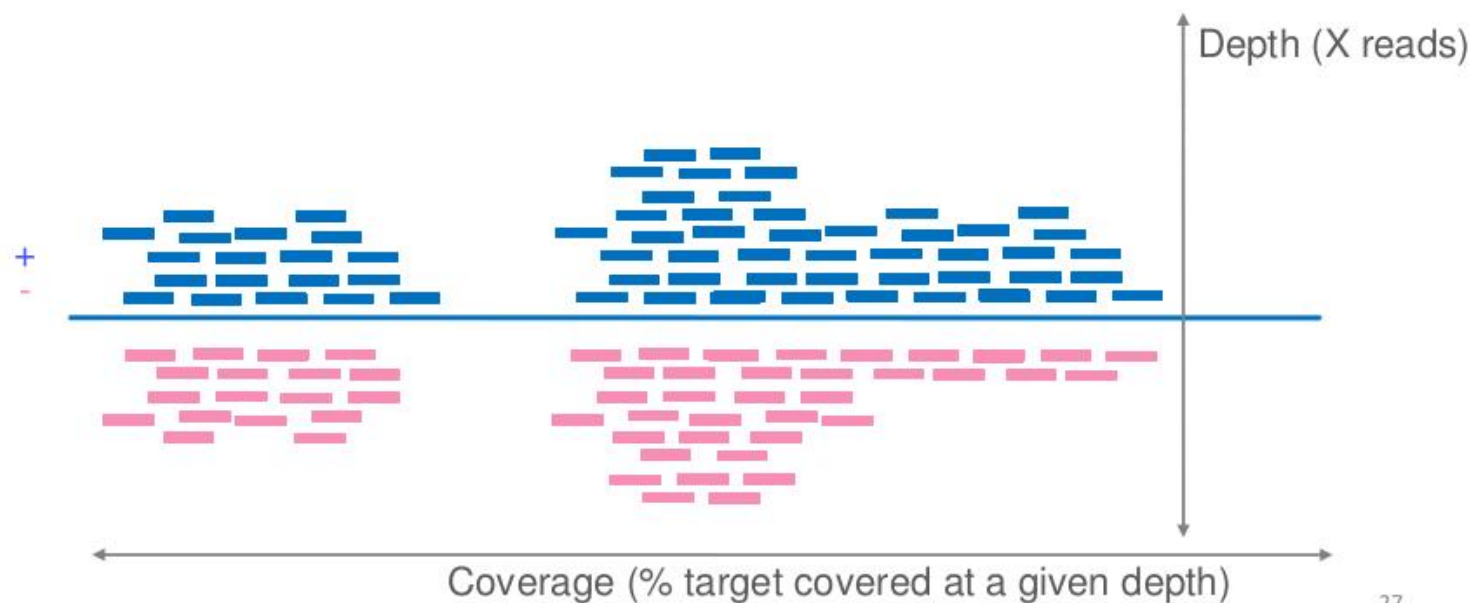
32: HG101_BWA_flagstat

Coverage and depth of coverage



Statistics used as quality control

- **Depth of coverage** = average number of reads covering a base (X)
Example: 30X for normal sample, 100X for tumor sample
- **Coverage** = percentage of the targeted regions covered by at least X read
Example: $\geq 80\%$ of your exome target is covered by 20X for normal sample



27

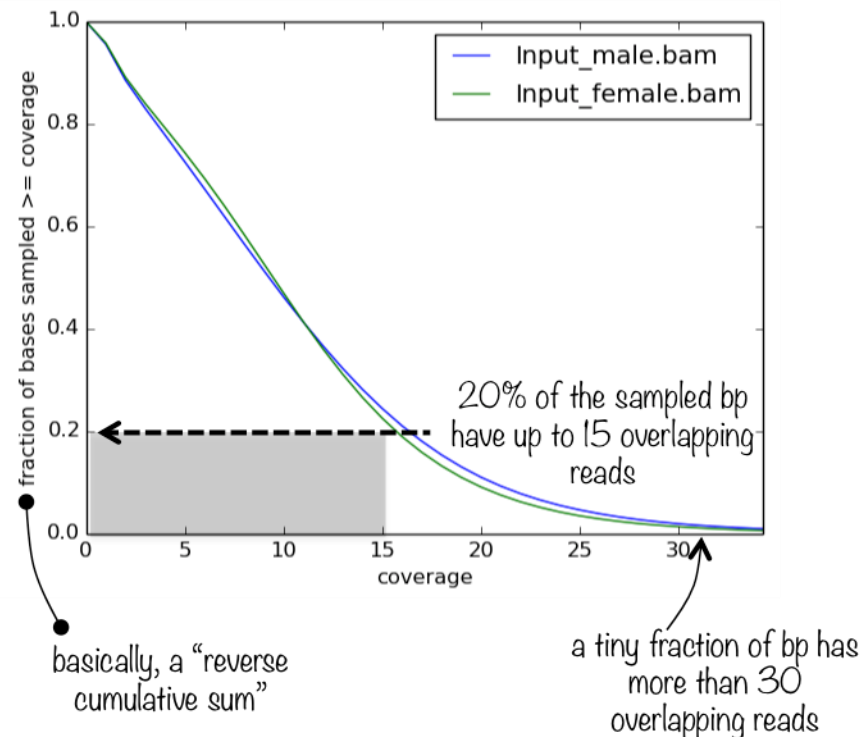
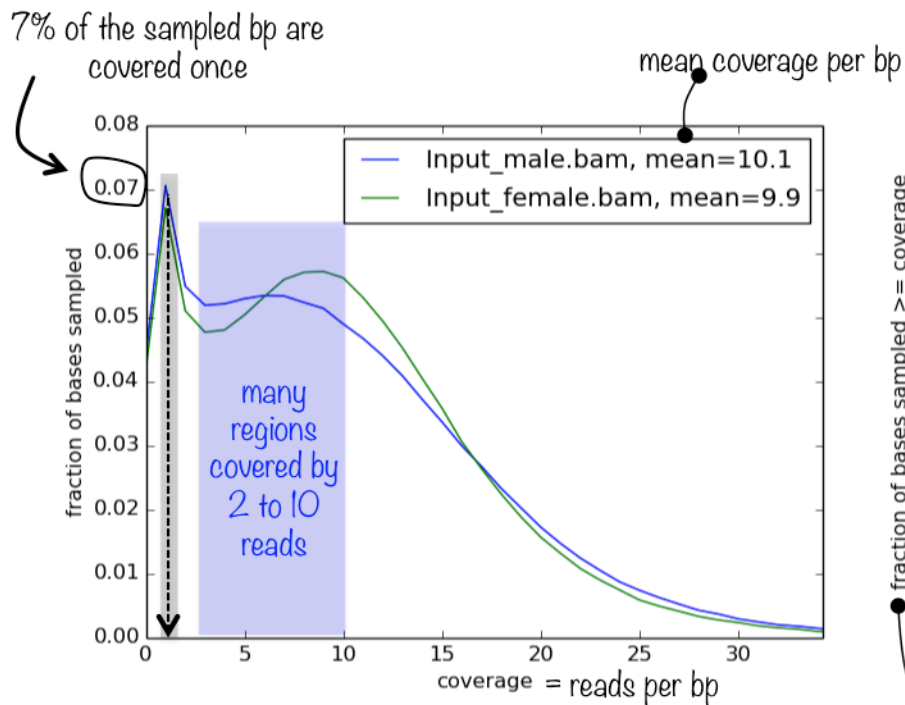
Source : Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016

http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf

Computing coverage and depth of coverage

DeepTools2 / plotCoverage

Ramírez, Fidel and Ryan, Devon P and Grüning, Björn and Bhardwaj, Vivek and Kilpert, Fabian and Richter, Andreas S and Heyne, Steffen and Dündar, Friederike and Manke, Thomas (2016). *deepTools2: a next generation web server for deep-sequencing data analysis*. In *Nucleic Acids Research*, 44 (W1), pp. W160–W165



DeepTools / Plot Coverage

Tools

plotCove

Quality Control

plotCoverage assesses the sequencing depth of BAM files

plotCoverage assesses the sequencing depth of BAM files (Galaxy Version 2.4.2.0)

Options

Sample order matters

No

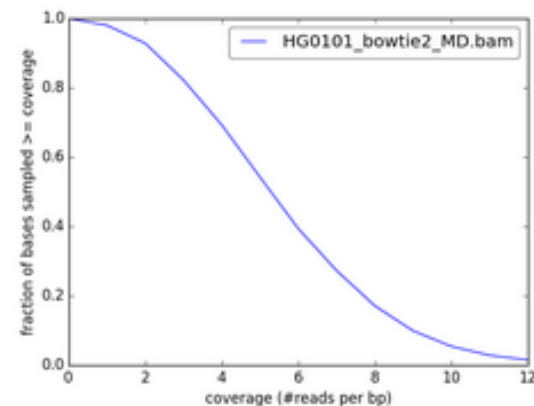
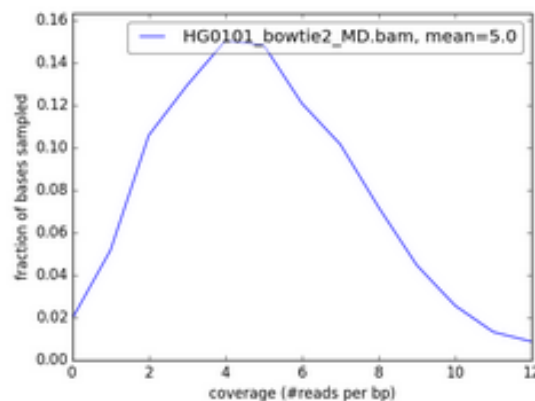
By default, the order of samples given to the program is dependent on their order in your history. If the order of the samples is vital to you, select Yes below.

Bam file

23: HG0101_bowtie2_MD.bam
21: HG0101_BWA_MD.bam
19: HG0101_bowtie2.bam
18: HG0101_BWA.bam

(--bamfiles)

26: plotCoverage image



Galaxy *Workflow*

- Extract *workflow* from an history
- Modify *workflow*
- Execute *workflow* on new data
- Compare results from 2 *workflows* (in 2 histories)

Extract *Workflow* from the history of steps applied to the first sample

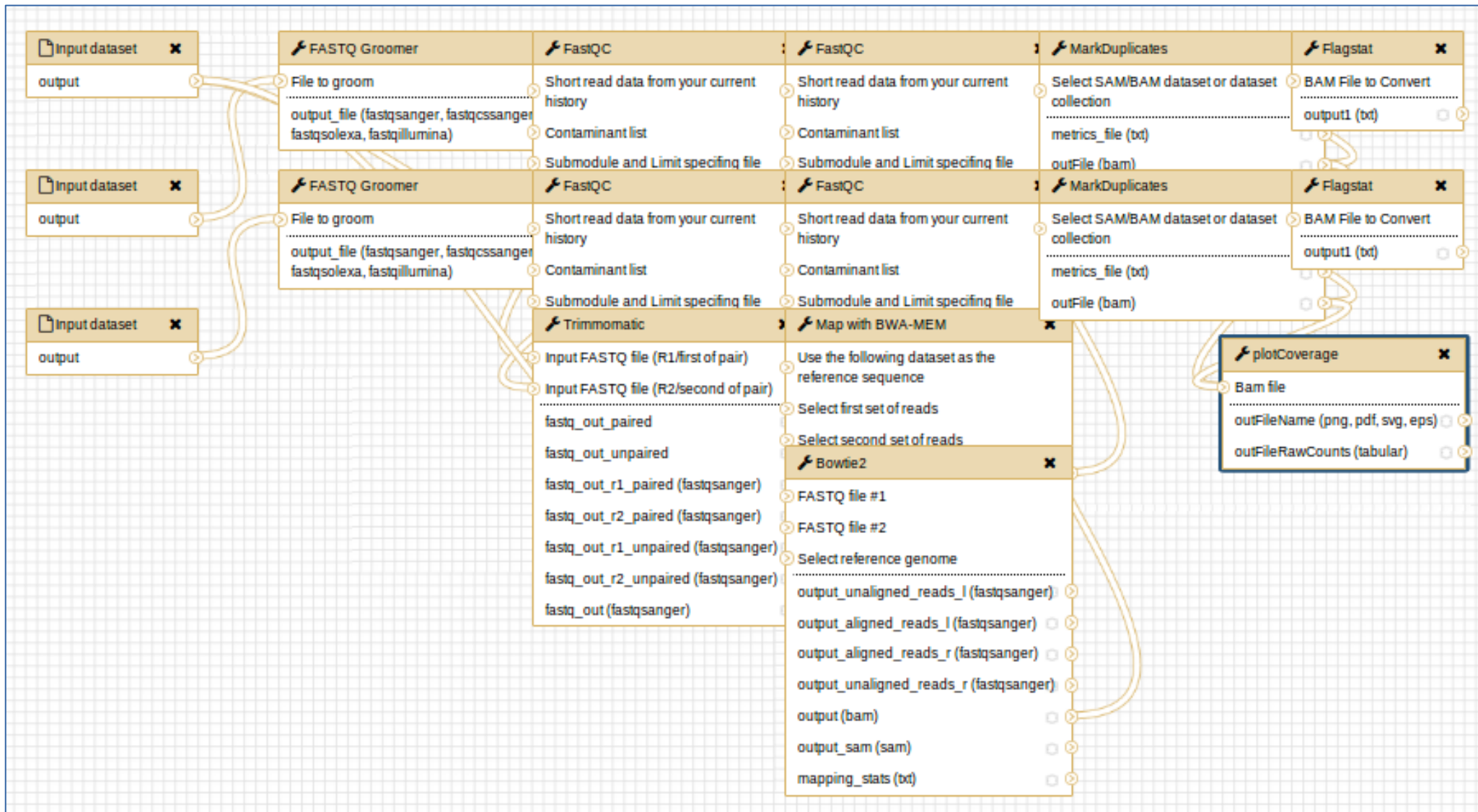
The screenshot shows a sidebar menu titled 'History'. At the top right, there is a gear icon. Below it, the menu is organized into sections: 'HISTORY LISTS' (containing 'Saved Histories' and 'Histories Shared with Me'), 'CURRENT HISTORY' (containing 'Create New', 'Copy History', and 'Share or Publish'), 'DATASET ACTIONS' (containing 'Delete', 'Delete Permanently', 'Copy Datasets', 'Dataset Security', 'Resume Paused Jobs', 'Collapse Expanded Datasets', 'Unhide Hidden Datasets', 'Delete Hidden Datasets', and 'Purge Deleted Datasets'), 'DOWNLOADS' (containing 'Export Tool Citations'), and 'OTHER ACTIONS' (containing 'Export History to File' and 'Import from File'). The 'Extract Workflow' option is highlighted in a dark blue box with a red border.

The dialog box is titled '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.' It includes a 'Workflow name' field with 'TP1_WF1' entered, and 'Create Workflow', 'Check all', and 'Uncheck all' buttons. Below is a table with two columns: 'Tool' and 'History items created'. The 'Tool' column lists 'Upload File' (three times, marked as unusable), 'FASTQ Groomer' (two times, checked), and 'FastQC' (two times, checked). The 'History items created' column lists corresponding files: '1 GRCh37_region1.fasta', '2 HG00101_1.fastq', '3 HG00101_2.fastq', '4 HG00101_OK_1.fastq', '5 HG00101_OK_2.fastq', '6 HG00101_1_QC', and '8 HG00101_2_QC'. Each row has a checkbox to 'Treat as input dataset' which is checked.

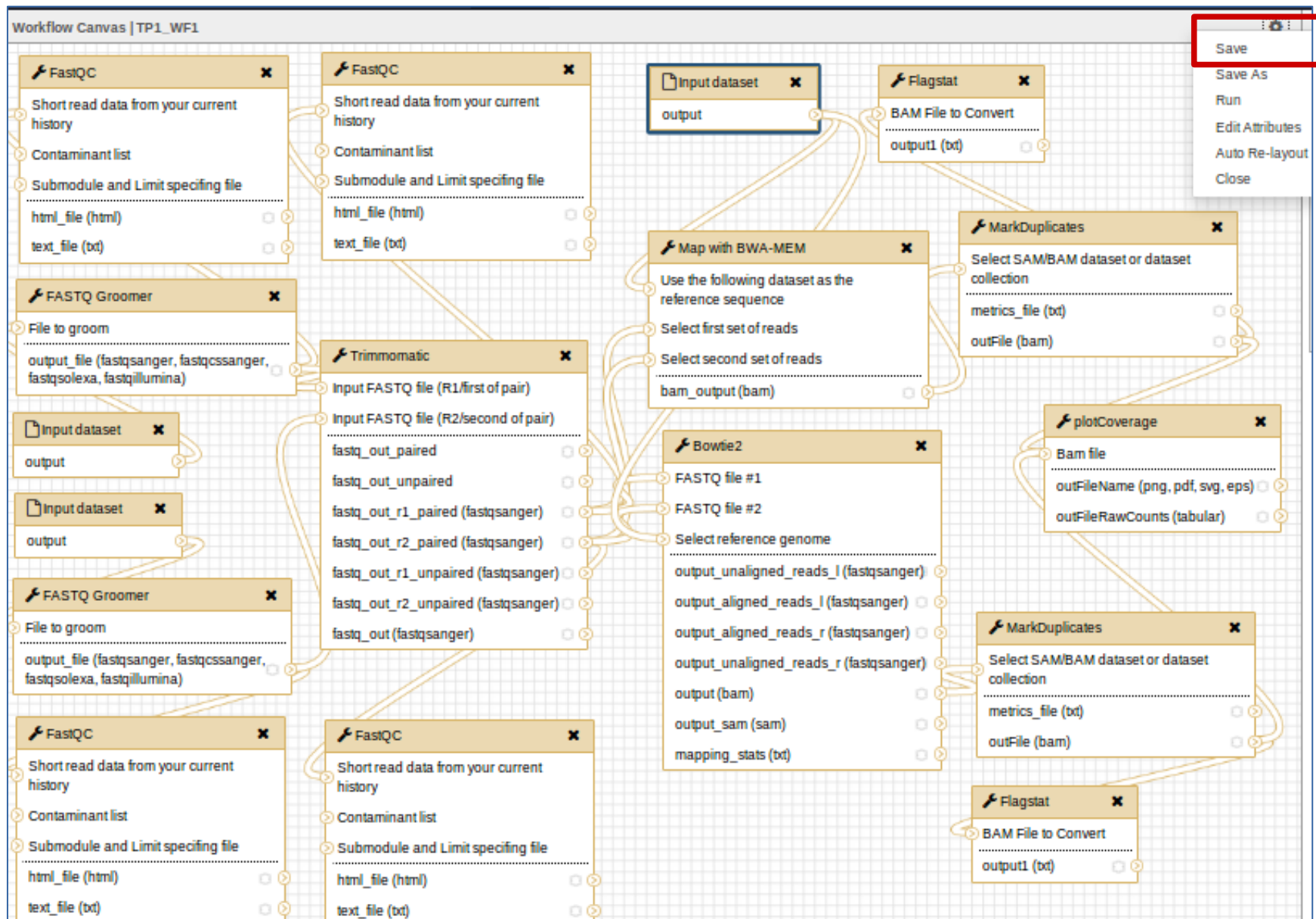
Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	<input checked="" type="checkbox"/> Treat as input dataset 1 GRCh37_region1.fasta
Upload File <i>This tool cannot be used in workflows</i>	<input checked="" type="checkbox"/> Treat as input dataset 2 HG00101_1.fastq
Upload File <i>This tool cannot be used in workflows</i>	<input checked="" type="checkbox"/> Treat as input dataset 3 HG00101_2.fastq
FASTQ Groomer <input checked="" type="checkbox"/> Include "FASTQ Groomer" in workflow	4 HG00101_OK_1.fastq
FASTQ Groomer <input checked="" type="checkbox"/> Include "FASTQ Groomer" in workflow	5 HG00101_OK_2.fastq
FastQC <input checked="" type="checkbox"/> Include "FastQC" in workflow	6 HG00101_1_QC
FastQC <input checked="" type="checkbox"/> Include "FastQC" in workflow	8 HG00101_2_QC

Workflow "TP1_WF1" created from current history. You can [edit](#) or [run](#) the workflow.

Visualize workflow

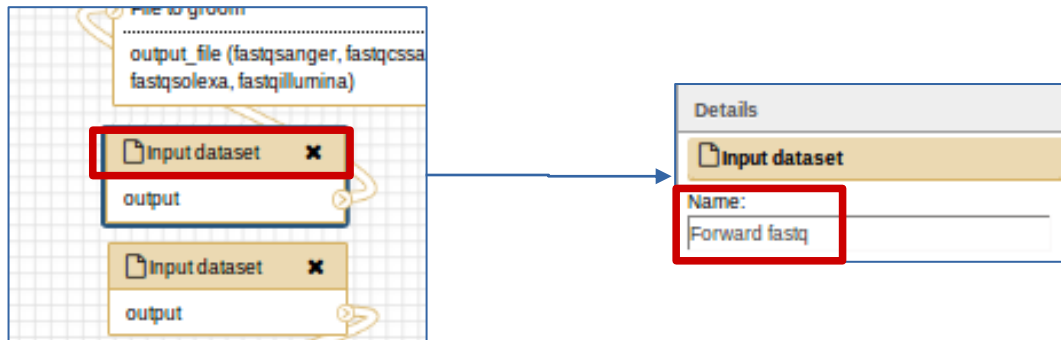


Modify workflow visualisation

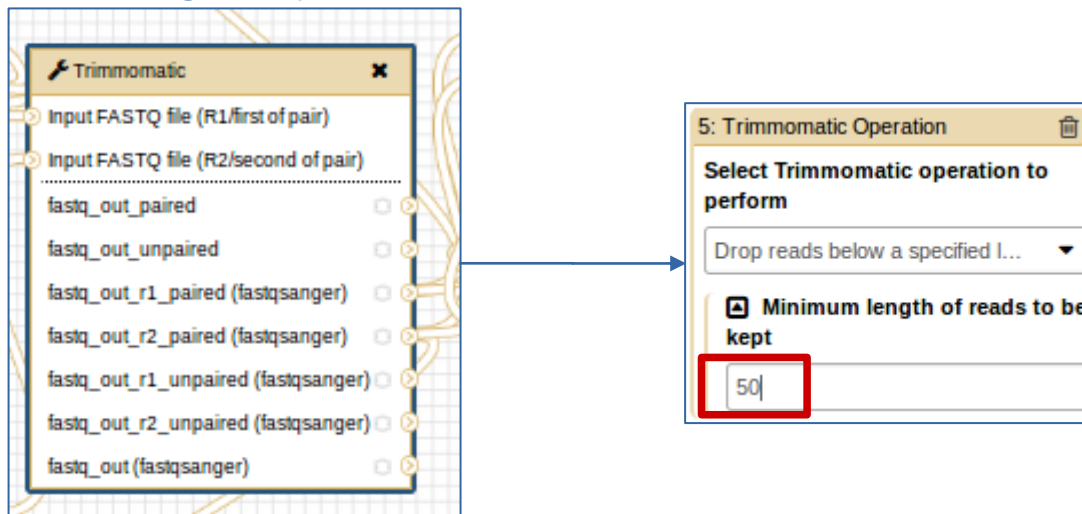


Modify some steps configuration

This WF uses 3 input files. Change box name to describe which data is required for each input : eg *Reference*, *Forward fastq*, *Reverse fastq*



You can also change any parameter for example for *trimmomatic* step.



Enable a parameter to be set at run time

Parameters for each tool will have the predefined values set in the workflow

You can modify this to enable any parameter to be set at run time.

Modify Trimmomatic so that Adapter are set at run time

The image shows a workflow editor interface for the Trimmomatic tool. On the left, a list of parameters is shown with checkboxes and expand/collapse icons. The 'Adapter sequences to use' parameter is highlighted with a red box. An arrow points from this parameter to the 'Details' panel on the right. In the 'Details' panel, the 'Adapter sequences to use' parameter is also highlighted with a red box, showing its current value as 'TruSeq3 (paired-ended, for...)'.

Trimmomatic

- Input FASTQ file (R1/first of pair)
- Input FASTQ file (R2/second of pair)
- fastq_out_paired
- fastq_out_unpaired
- fastq_out_r1_paired (fastqsanger)
- fastq_out_r2_paired (fastqsanger)
- fastq_out_r1_unpaired (fastqsanger)
- fastq_out_r2_unpaired (fastqsanger)
- fastq_out (fastqsanger)

Details

Trimmomatic flexible read
trimming tool for illumina NGS data (Galaxy Version 0.36.1)

Paired end data?
Yes No

Input Type
Pair of datasets

Input FASTQ file (R1/first of pair)
Data input 'fastq_r1_in' (fastqsanger)

Input FASTQ file (R2/second of pair)
Data input 'fastq_r2_in' (fastqsanger)

Perform initial ILLUMINACLIP step?
Yes No
Cut adapter and other illumina-specific sequences from the read

Adapter sequences to use
TruSeq3 (paired-ended, for...)

Adapter sequences to use

Maximum mismatch count which will still allow a full match to be performed

Do'nt forget to save your workflow !







Import new data for sample HG0103

Importe files HG0103_1.fastq and HG_0103_2.fastq

Download from web or upload from disk

[Regular](#) [Composite](#)

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 HG00103_1.fastq	888.7 KB	fastq	TP_ref		
 HG00103_2.fastq	888.7 KB	fastq	TP_ref		

Type (set all): Genome (set all):

History

search datasets

TP1
22 shown, 6 deleted

15.67 MB

28: HG00103_2.fastq

27: HG00103_1.fastq

Analyze these new data with the same workflow

Run the workflow with these new data

Galaxy

Analyze Data Workflow Shared Data Visualization

Your workflows

Create new workflow Upload or import workflow

Name	# of Steps
TP1_WF1_OK	17
	17

Run

Workflow: TP1_WF1_OK

Run workflow

History Options

Send results to a new history

Yes No

History name

HG0103

Step 1: Input dataset

Forward fastq

27: HG00103_1.fastq

Step 2: Input dataset

Reverse fastq

28: HG00103_2.fastq

Step 3: Input dataset

Reference

1: GRCh37_region1.fasta

Step 4: FASTQ Groomer convert between various FASTQ quality formats (Galaxy Version 1.0.4)

Step 5: FASTQ Groomer convert between various FASTQ quality formats (Galaxy Version 1.0.4)

Step 6: FastQC Read Quality reports (Galaxy Version 0.67)

Short read data from your current history

✓ Successfully invoked workflow TP1_WF1_OK.

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.

Browse results

Saved Histories

search history names and tags

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated
<input type="checkbox"/> HG0103	10	6	7	0 Tags	5.5 MB	~33 seconds ago
<input type="checkbox"/> TP1	22	0 Tags		15.7 MB	~4 hours ago	~6 minutes ago

Saved Histories

search history names and tags

Advanced Search

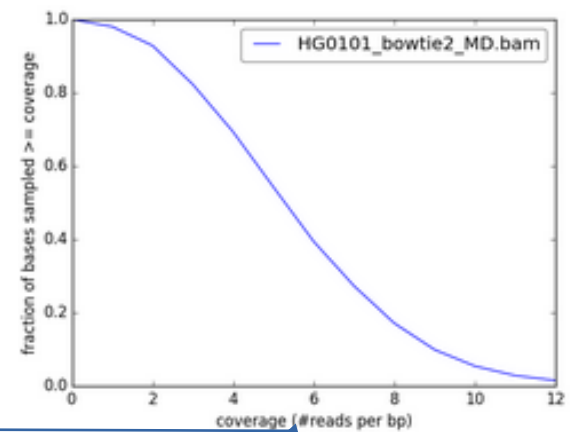
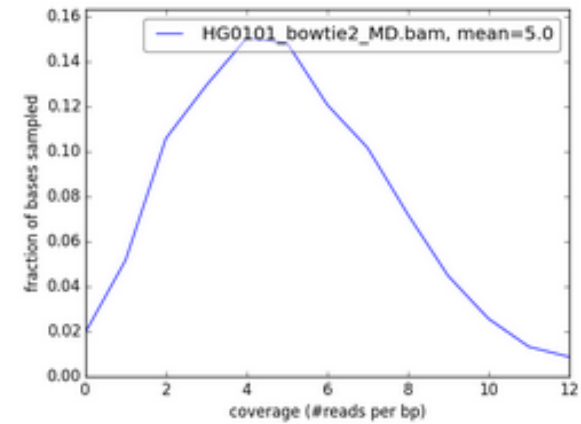
Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated
<input type="checkbox"/> HG0103	23	0 Tags		10.7 MB	~1 minute ago	~1 minute ago
<input type="checkbox"/> View					~4 hours ago	~7 minutes ago
<input type="checkbox"/> Share or Publish		0 Tags		15.7 MB	hours ago	ago
<input type="checkbox"/> Copy						

History

search datasets

HG0103
23 shown
10.73 MB

- 23: Flagstat on data 20
- 22: plotCoverage image
- 21: Flagstat on data 18
- 20: MarkDuplicates on data 16: MarkDuplicates BAM output
- 19: MarkDuplicates on data 16: MarkDuplicate metrics
- 18: MarkDuplicates on data 13: MarkDuplicates BAM output
- 17: MarkDuplicates on data 13: MarkDuplicate metrics
- 16: Bowtie2 on data 1, data 6, and data 5: aligned reads (sorted BAM)
- 15: FastQC on data 6: RawData
- 14: FastQC on data 6: Webpage



Gather BWA alignment results for the 2 samples 1/2

Create a new history named results

From history TP1 : Copy HG0101_BWA_MD.bam *dataset*

The image illustrates the process of copying a dataset from one history to another in a bioinformatics workflow environment. It is divided into three main sections:

- Left Panel (History Lists):** A sidebar menu with sections for 'HISTORY LISTS', 'CURRENT HISTORY', and 'DATASET ACTIONS'. The 'Copy Datasets' option under 'DATASET ACTIONS' is highlighted with a red box.
- Central Panel (Copy History Items Dialog):** A dialog box titled 'Copy any number of history items from one history to another.' It features a 'Source History:' dropdown menu set to '3: TP1' (highlighted with a red box). Below it, a list of datasets is shown with checkboxes: '1: GRCh37_region1.fasta', '2: HG00101_1.fastq', '3: HG00101_2.fastq', and '21: HG0101_BWA_MD.bam' (checked and highlighted with a red box). A 'Copy History Items' button is highlighted with a red box at the bottom right.
- Right Panel (History Panel):** A panel titled 'History' showing a search bar and a list of datasets. The list contains one item: '1: HG0101_BWA_MD.bam' (highlighted with a red box). The size of the dataset is listed as 778.82 KB.

A green notification bar at the top of the right panel states: '1 dataset copied to 1 history: results.' A blue information bar above the dialog also reads: 'Copy any number of history items from one history to another.'

Gather BWA alignment results for the 2 samples 2/2

From history HG0103 : Copy *dataset*

« MarkDuplicates on data 13: MarkDuplicates BAM output »

The screenshot illustrates the process of copying a specific history item from one history to another in the Galaxy environment. The main interface is divided into three main sections:

- Source History:** A dropdown menu is set to '2: HG0103'. Below it, a list of history items is shown, with item 18, '18: MarkDuplicates on data 13: MarkDuplicates BAM output', selected with a red checkmark.
- Destination History:** A dropdown menu is set to '1: results'. Below it, there is a 'Choose multiple histories' link and a 'New history named:' input field.
- Copy History Items:** A button with a red border is located between the source and destination history sections.

On the left side, a sidebar menu is visible with the following categories and items:

- HISTORY LISTS**
 - Saved Histories
 - Histories Shared with Me
- CURRENT HISTORY**
 - Create New
 - Copy History
 - Share or Publish
 - Show Structure
 - Extract Workflow
 - Delete
 - Delete Permanently
- DATASET ACTIONS**
 - Copy Datasets (highlighted with a red box)

On the right side, a 'History' panel is shown with a search bar and a list of results. The results are:

- 2: MarkDuplicates on data 13: MarkDuplicates BAM output (highlighted with a red box)
- 1: HG0101_BWA_MD.bam

Visualize depth of coverage for both samples

Rename *datasets*
Run *plotCoverage*

plotCoverage assesses the sequencing depth of BAM files (Galaxy Version 2.4.2.0) Options

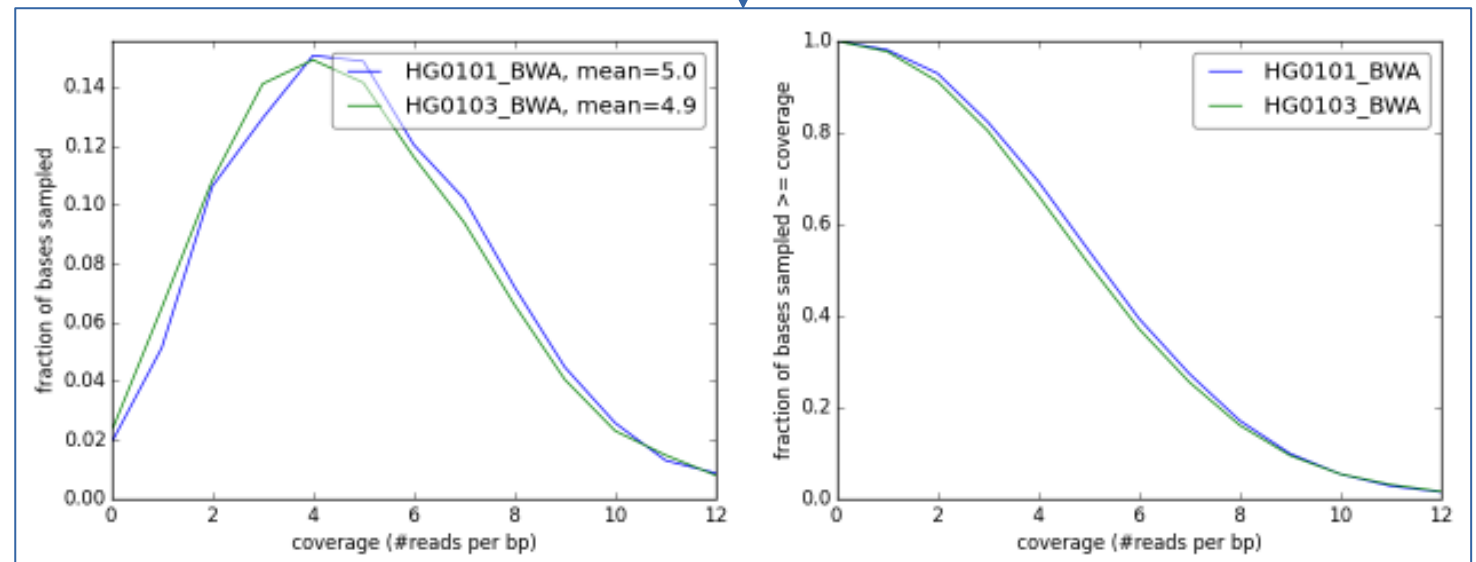
Sample order matters
No

By default, the order of samples given to the program is dependent on their order in your history. If the order of the samples is vital to you, select Yes below.

Bam file

2: HG0103_BWA
1: HG0101_BWA

(--bamfiles)



Galaxy – *Best Practices*

- Manage disk space
- Export analysis results (*datasets and histories*)
- Export / Import analysis protocols (*workflow*)

Manage disk space

User menu options: User, Logged in as user@galaxyjfb.fr, Preferences, Custom Builds, Logout, Saved Histories, Saved Datasets, Saved Pages, API Keys.

Global disk space

User preferences

You are currently logged in as user@galaxyjfb.fr.

- [Manage your information](#) (email, address, etc.)
- [Change your password](#)
- [Change default permissions](#) for new histories
- [Manage your API keys](#)
- [Manage your ToolBox filters](#)
- [Logout of all user sessions](#)

You are using **105.9 MB** of disk space in this Galaxy instance. Is your usage more than expected? See the [documentation for tips](#) on how to find all of the data in your account.

Disk space per dataset

HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New
- Copy History
- Share or Publish
- Show Structure
- Extract Workflow
- Delete
- Delete Permanently

DATASET ACTIONS

- Copy Datasets
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Purge Deleted Datasets

DOWNLOADS

- Export Tool Citations
- Export History to File

OTHER ACTIONS

- Import from File

Disk space per history

Saved Histories

search history names and tags




Advanced Search


<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
<input type="checkbox"/>	TP1	23	0 Tags		15.7 MB	~6 hours ago	~4 seconds ago	
<input type="checkbox"/>	HG0103	23	0 Tags		10.7 MB	~2 hours ago	~15 seconds ago	
<input type="checkbox"/>	results	3	0 Tags		1.6 MB	~2 hours ago	~49 seconds ago	
<input type="checkbox"/>	Unnamed history		0 Tags		0 bytes	~5 minutes ago	~5 minutes ago	current history

For 0 selected histories:

Export analysis results : *datasets*

Image

26: plotCoverage image   

83.9 KB 
format: **png**, database: **TP_ref**

sample mean std min 25% 50% 75% max
MarkDuplicates on data 18:
MarkDuplicates BAM output 5.02 2.63 0 3.0 5.0 7.0 20
MarkDuplicates on data 19:
MarkDuplicates BAM output 5.01 2.63 0 3.0 5.0 7.0 20
Number of non zero bins used:
150001










     







Image in png format

Text

22: MarkDuplicates on data 19: MarkDuplicate metrics   

103 lines
format: **txt**, database: **TP_ref**

Picked up _JAVA_OPTIONS:
-Xmx2048m -Xms256m


```
## htsjdk.samtools.metrics.StringHeader
# picard.sam.markduplicates.MarkDuplicate
ATES=false ASSUME_SORTED=true DUPLICATION_READS_MAP=500000 MAX_FILE_SIZE=500000 RDS_IN_RAM=500000 CREATE_INDEX=false
```

Bam

33: HG0101 BWA.bam   

767.5 KB
format: **bam**, database: **TP_ref**




```
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.04 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forward-only FASTA... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0
```







display with IGV [local](#)
display in IGB [View](#)
display at bam.iobio [bam.iobio.io](#)

Binary bam alignments file


HTML + files


14: FastQC on data 10: Webpage   


229.4 KB
format: **html**, database: **TP_ref**

HTML file

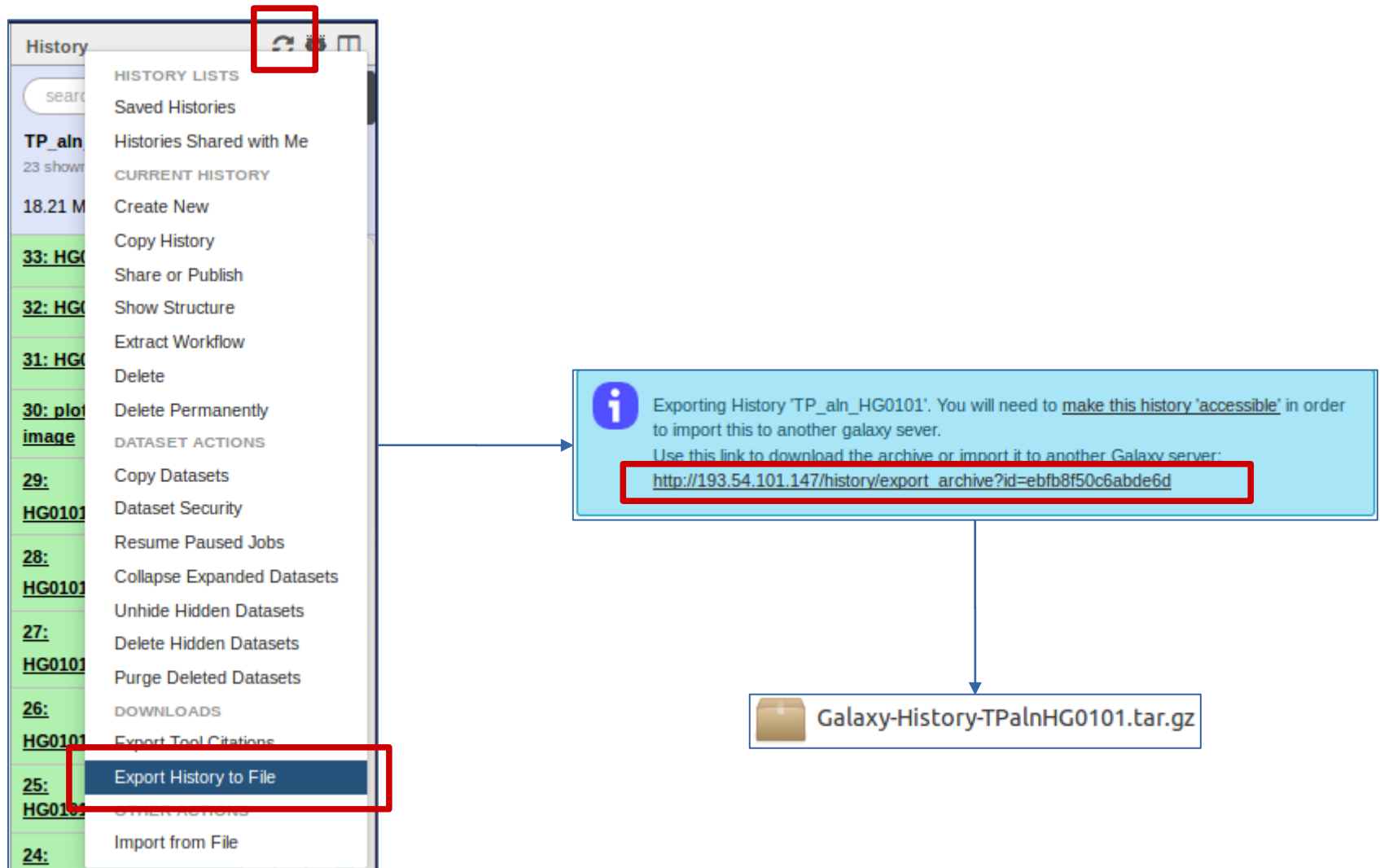
 Galaxy26-[plotCoverage_image].png

 ...MarkDuplicates_on_data_19_MarkDuplicate_metrics].txt

 Galaxy33-[HG0101_BWA.bam].bam

 FastQC_on_data_10_Webpage.zip

Export analysis results : *histories*



Export / import analysis protocols : *workflow*

Export

Your workflows

Name
TP1_WF1_OK ▾

- Edit
- Run
- Share or Download
- Copy
- Rename
- View
- Delete

Import


Galaxy Analyze Data **Workflow** Shared Data ▾ Visualization ▾ Help ▾

Your workflows

[+ Create new workflow](#) [↑ Upload or import workflow](#)

Export

Download workflow as a file so that it can be saved or imported into another Galaxy server.

 Galaxy-Workflow-TP1_WF1_OK.ga